Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

A Corpus-Based Study on Korean High School Students' Common Errors in English Writing

Myung Ho, Park[1], Dong Ju Lee[2]*
[1]Incheon Shinhyeon High School
[2]Korea National University of Education

한국 고등학생 영작문에 나타난 코퍼스 기반 오류 분석

박명호[1], 이동주[2]*
[1]인천신현고등학교, [2]한국교원대학교

*Corresponding Author: Dong Ju Lee, maydjlee@knue.ac.kr

ABSTRACT

The purpose of this study is to investigate the common errors in writing production made by Korean high school students. It aims to provide information on the areas that Korean high school students need to develop and to inform their educational needs. For this, high school students' writing samples were collected and built into a learner corpus named Shinhyeon Corpus of High school Learners' English Essays (SCOHLEE). The study identified frequent errors produced by the high school students adopting a computer-aided error analysis (CEA) technique using the *Louvain Error Tagging Manual*, and quantitative and qualitative analyses were conducted by using corpus tools (*AntConc 4.0*). The results revealed that a total of 3,368 errors were categorized into 7 major categories and 47 sub-categories, with article errors (GA) produced the most. Regarding English proficiency, mid-level learners produced a higher rate of errors and a wider range of error types. Additionally, the study showed a significant increase in the frequency of punctuation errors in the second semester. Overall, this study has pedagogical implications for lexis, grammar and writing instruction and the provision of error correction.

Key words: Learner corpus, error tagging, CEA, frequent errors, writing instruction

Brain, Digital, & Learning
2023, Vol. 13, No. 4, 397-414
https://doi.org/10.31216/BDL.20230025

Received: October 18, 2023
Revised: December 04, 2023
Accepted: December 04, 2023

Introduction

The errors of language learners are evidence of the language system that they learned or used at certain stages of their language acquisition (Corder, 1967). It means learners' errors can be meaningful data for researchers and instructors because their errors in the process of learning a target language index how much the learners have learned or acquired the target language system. To analyze data in Korean high school learner corpus from a different perspective, this study

carried out computer-aided error analysis (CEA). It is necessary to identify common weaknesses and strengths that high school learners showed in writing production. Raising learners' awareness and preventing common errors in writing can be facilitated by providing learners with information regarding the errors their peers frequently and universally produce. Furthermore, when teachers provide feedback on learners' writing, they can utilize systematic examples as a reference, rather than relying solely on their intuitions. This approach allows for a more structured and objective assessment of learners' writing, promoting effective error correction and language development.

In addition, Dagneaux, Denness and Granger (1998) suggested that the CEA approach provides one way of discovering valuable features of learner writing, in particular areas of persistent grammatical difficulties and error-prone items. However, there is a noticeable research gap regarding Korean high school learners' errors in writing within the broader context of holistic grammatical categories (e.g., Dagneaux, Dennes, Granger, & Meunier, 1996), with just a few studies conducted after Lee (2008). Thus, the present study is to enhance the understanding of Korean high school students' nature of errors in writing relating to comprehensive lexical and grammatical categories. The followings are research questions:

1. What types of errors are produced by Korean high school students?
2. How does the distribution of errors differ across students' English proficiency levels?
3. How does the distribution of errors differ by school semester?

Theoretical Background

Computer-aided Error Analysis (CEA)

Many researchers have investigated learners' linguistic developmental stages from various angles such as error analysis (Corder, 1967) and interlanguage development (Selinker, 1972). Researchers' increasing interest in errors connected to linguistic development has led to the developmental sequence in second language acquisition.

Corder (1967) argues that the errors of language learners are evidence of the language system that they learned or used at certain stages of their language acquisition. It means learners' errors can be meaningful data for researchers and instructors because their errors in the process of learning a target language index how much the learners have learned or acquired the target language system.

In line with Corder (1967), Selinker (1972) saw interlanguage as learners' unstable, yet systematic, continually developing second language system. Many researchers in this strand have attempted to investigate L2 learners' language development in terms of complexity, accuracy, and fluency (Ellis, 1989; Norris & Ortega, 2009; Ortega, 2003).

Gass et al. (2020) claimed that errors are to be viewed as learners' attempts to figure out some systems and they are evidence of an underlying rule-governed system. That is, if a regular pattern of errors is found in learners' written production, and if the errors reveal changes across learners' proficiency levels, errors become an index to measure achievement in language learning (Ellis, 1994). Likewise, error analysis is a crucial tool for investigating the process of second language acquisition, and provides fundamental information about learners' language acquisition process.

Computer-aided Error Analysis (CEA) has its roots in 'Error Analysis (EA),' which was both popular and criticized in the 1970s. Criticism of the traditional EA was mainly due to the use of data collected through practice and compiled a list

of errors. It was unnatural and mixed data regardless of task, learner, or language variables (Ellis, 1994). There were also criticisms that EA had problems with unobservable items, subjective characteristics of items, and overlapping error items (Dulay et al. 1982). Computer-aided Error Analysis (CEA), on the other hand, can overcome the criticisms for EA by using a vast collection of computerized/digitalized learner data and providing information on the elements that learners need to develop, and informing their educational needs (Granger, 2002; Lee, 2007, 2008).

Previous research on error analysis of Korean college learners' language use (Hwang, 2012; Kim, 1998; Kim, 2020; Kim, 2016; Lee, 2016) has analyzed learners' errors in their writing based on the five main categories and 16 subcategories proposed by Ferris (2002). The only research that focused on high school learners (Song & Park, 2012) also followed Ferris's (2002) error categories.

A small body of research has employed CEA techniques with comprehensive error tagging to investigate EFL learners' errors in their writing (Chuang & Nesi, 2006; Granger, 1998; MacDonald et al., 2013; Lee, 2008; Thewissen, 2013). Lee (2004) conducted a study in which 33,000 words of learner corpus were compiled from the writing outputs of Korean high school students. The study extracted sentences in which learners tend to make errors (*because, have, many, much, if,* and *a lot of*) and analyzed learner errors using corpus analysis tool (WordSmith Tools). The study then compared learner expressions with textbook corpus to investigate how closely learners use the expressions as they learned. The results showed that despite the fact that *because* is mainly used as a subordinating conjunction in most cases in the textbook corpus, learners made errors in 72.9% of cases in the learner corpus. Considering the fact that the learners were frequently exposed to the conjunction, but showed high frequency errors in the subordinating conjunction, Lee proposed form-focused instruction and consciousness-raising activities for learning the use of *because*.

Lee (2008) constructed a learner corpus consisting of 19,610 vocabulary items based on the writing activities of 290 10th-grade high school learners from five high schools. Using the error analysis tool developed by Dagneaux et al. (1998), Lee tagged learner errors into five main categories and 56 subcategories using Error Tagging Manuals 1.1 and 1.2, and analyzed the frequency of error codes using the corpus analysis tool, WordSmith Tools. The study found a total of 4,479 errors in the learner corpus, and 48 errors of the 56 subcategories in Error Tagging Manual version 1.2 were identified. In terms of frequency, grammatical errors (G) were the highest at 37.5%, followed by punctuation errors (Q, 14.5%), morphology errors (F, 12.4%), and vocabulary errors (L, 11.5%). Based on the results, Lee suggested that CEA can contribute to the development of learner-customized learning materials and teaching activities that reflect learners' needs.

Chuang and Nesi (2006) invented linguistic category taxonomy and investigated Chinese L1 learners' errors in a learner corpus comprising 50 essays with 88,000 running words. The result showed that the relative frequency of grammatical, lexico-grammatical, and lexical errors was 85.9%, 5.0%, and 9.1% respectively. The ten most problematic linguistic features made by L1 Chinese learners were determiners (27.6%), nouns (17.8%), verbs (8.9%), prepositions (8.1%), punctuation (5.9%), sentence parts (4.7%), tense/aspect (4.4%), modals (4.1%), conjunctions (3.9%), and pronouns (3.9%). Of lexico-grammatical errors, erroneous use of prepositions with a verb, a noun, or an adjective was found the most frequent error, followed by misuse of uncountable nouns, and erroneous use of syntactic pattern. Based on the findings, they prioritized article errors for correction and developed online self-study materials. The identification of specific linguistic features and their frequencies can inform the development of targeted materials for error correction

and self-study resources for Korean high school learners.

MacDonald et al. (2013) conducted a study using a learner corpus consisting of synchronous and asynchronous online conferences and emails from 126 university learners who use Spanish, German, Latvian, Norwegian, and French as their first language. They classified errors according to Error Tagging Manual 1.2 (Dagneaux et al., 1998) and used the computer-aided error analysis tool, the Universite Catholique de Louvain Error Editor (UCLEE) to analyze the data. The results showed that overall 5.6% of errors were found in synchronous communication, with morphological and grammatical errors occurring most frequently. In contrast, 4.4% of errors were found in asynchronous communication, with vocabulary and style-related errors occurring most frequently. The study also found that the types of errors varied depending on the learners' first language.

Despite systematic error tagging shed lights on learners' use of English language, and their development, a handful of research has been carried out. The only reported research on Korean high school learners' errors employing a technique of CEA (Lee, 2008) analyzed high school learners' errors solely from a quantitative perspective, without providing any data on errors related to learners' different proficiency levels. Therefore, there is a clear necessity for further research on Korean high school learners' errors that incorporates both quantitative and qualitative analysis and considers learners' proficiency levels. By analyzing frequent errors made by students at different proficiency levels and examining them both quantitatively and qualitatively, this research aims to provide a more comprehensive understanding of the errors made by Korean high school learners. It will contribute to a deeper insight into the language development of this previously unexamined group of learners.

Methods

Learners' English Proficiency

To analyze whether there is a difference in error production according to learners' English proficiency, it was necessary to classify learners' English proficiency. Since the learners do not have any official English score, the grades of the 2020 - 2022 high school first-year Geongook Scholastic Ability Test (GSAT) were used as a criterion for classifying learners' English proficiency. Learners who achieved 80 points or more in level 1 and 2 were classified as "high-level learners" (High), learners who achieved level 3 to 4 (over 60 points) were classified as "mid-level learners" (Mid), and the rest of the learners were classified as "lower-level learners" (Low). Table 1 shows learners' proficiency information.

Table 1. The number and proportion of learners by English proficiency N(%)

Year	High	Mid	Low	Total
2020	38(15.7)	98(40.5)	106(43.8)	242(100)
2021	39(19.8)	59(29.9)	99(50.3)	197(100)
2022	38(16.4)	80(34.5)	114(49.1)	232(100)

As shown in Table 1, high-level learners accounted for less than 20% of all learners during 2020 - 2022 school years. Mid-level learners accounted for 35.5% of all learners, and low-level learners made up the majority at 47.5%.

Table 2 shows the results of the one-way analysis of variance (ANOVA), which analyzed the difference in average English scores between different proficiency groups. The results of the one-way ANOVA, displayed in Table 3, indicate

a significant difference in mean values among the three groups (F = 1138, $p < .001$).

Table 2. Descriptions of levels based on GSAT

Group	*N*	*Mean*	SD	SE	F	df	*p*
High	115	86.7	5.16	0.481	1138	2	< .001
Mid	237	68.4	5.37	0.349			
Low	319	44.1	11.59	0.649			

The Learner Corpus: SCOHLEE

The learner corpus of this study consists of 671 high school first year students' writing samples. The learners showed academic achievement levels between levels 1 and 8 in English of the GSAT. In the regular curriculum, they take four hours of common English classes per week. The writing products of the learners were collected and compiled into a learner corpus, called Shinhyoen Corpus of High school Learners' English Essays (SCOHLEE), the details of which are presented in Table 3. SCOHLEE consists of six semester-by-semester corpora of first year of high school learners from 2020 to 2022.

Table 3. Information on the learner corpus (SCOHLEE)

Test	2020	2021	2022	Total
No. of texts	242	197	232	671
Tokens	44,824	43,175	47,241	135,240
Types	3,616	3,576	3,798	10,990
Type / Token Ratio	8.07%	8.28%	8.04%	
STTR	36.2%	36.5%	35.1%	
Average Word Length*	4.38	4.18	4.31	
Sentences	3,992	3,699	4,201	11,892

*Average Word Length is mean word length in characters.

As seen in Table 3, the STTR (Standardized Type-Token Ratio) of the learners in 2020 was 36.2%. In 2021, the STTR was 36.5%, and the STTR was 35.1% in 2022. The total number of vocabulary items that appeared in the learner corpus was 135,240 words, with a total of 10,990 distinct types of words.

Data Collection and Processing

Writing samples from a process-oriented writing task were collected every semester in 2020 - 2022 school year. Learner variables such as language background, age, and learning situation, as well as linguistic variables such as task type (e.g., descriptive writing) and topic (e.g., my favorite thing), were controlled to build a homogeneous learner corpus. Writing activities were conducted under the same conditions, such as writing prompts and test time and procedures. Overall 680 first year learners participated in writing activities. but 671 learners' final drafts were compiled to construct a learner corpus (SCOHLEE). 9 learners' writing samples were excluded from the corpus construction process. Excluded writing samples were writing samples consisting of only one or two sentences, writings from students who transferred schools, or writings from students who were unable to produce final drafts due to sick leaves or absences. By excluding these samples, the resulting corpus was refined to ensure the inclusion of high-quality and representative data for analysis.

The writing samples collected from each semester from 2020 to 2022 underwent a data collection process, beginning

with scanning and conversion into PDF files. Subsequently, a word processing procedure was conducted using Google Documents to enable further analysis. To facilitate the use of corpus analysis tools (i.e. AntConc), the writing samples were converted from Google Docs into text files. Overall 6 corpora of first year learners were constructed in the learner corpus (SCOHLEE). These corpora served as the basis for analyzing the overall errors produced by Korean high school learners.

To facilitate the use of CEA, the six corpora of the first year of high school learners were manually annotated following the Louvain Error Tagging Manual 2.0 (Granger et al., 2022). This annotation process involved the identification and tagging of errors, and annotating possible suggestions in writing samples with the help of Grammarly (Kim, 2020).

Tools

Grammarly

This study followed the Error Tagging Manual Version 2.0 (Granger et al., 2022). Unfortunately, since the software, UCLEE that aids annotating learners' errors was not available, error tagging for this study was conducted manually with the help of English editing software, Grammarly (Kim, 2020). Fig. 1 presents a screenshot illustrating the annotation of learners' errors in a corpus from the second semester of 2022.

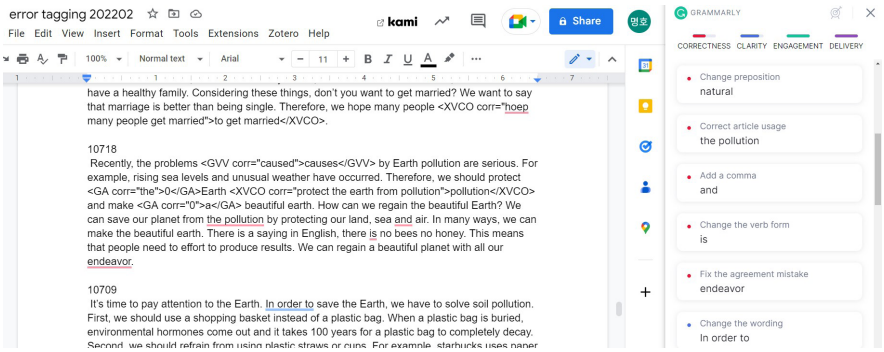

Fig. 1. Error annotation with grammarly

The screenshot in Fig. 1 shows the use of Grammarly, a language checking tool, which displays the ungrammatical part of the text underlined in red. On the right-side panel, Grammarly provides possible suggestions to make them correct.

Error Tagging Manual

Granger et al. (2022) categorized seven main error categories. They are Formal errors, Grammatical errors, Lexico-grammatical errors, Lexical errors, Word-related errors, and Infelicities, which are divided into 54 subcategories. This study includes one more tag for the erroneous use and/or wrong formation of comparatives and superlatives of adverbs (GADVCS). In the case of adverb, when a learner produces errors by misplacing an adverb (e.g., "they see only other criminals"), the code of <GADVO> was used in the Louvain Tagging Manual. There is no other error code for adverb errors. However, a wrong formation of comparatives of adverbs was detected in the SCOHLEE. To annotate this kind of errors (i.e., misusing comparatives or using a wrong formation of comparative or superlative of adverbs), the code

<GADVCS> was coined. The first letter of the error tag (e.g., GADVCS) indicates the error category; G for grammar. The second letter indicates the sub-category; ADV for adverbs, and the following letter indicates precision about the error type; CS for misuse of comparatives and superlatives. Error annotation includes inserting error codes and corrections into learner text. This study follows the format of <error code ="suggestion"> an erroneous word or phrase </error code>. An example of the coding is <GNN corr= "movies">movie</GNN>, which indicates a learner makes a noun number error (use of the singular *movie* instead of the plural *movies*). Table 4 shows 7 major error types and 55 error codes for error annotation.

As Table 4 shows, this study categorized learners' errors into 7 major types and 55 sub-categories. This study manually annotated learners' errors with the help of a computerized application, Grammarly. The error annotation will always contain an element of subjectivity (Granger et al, 2022). To increase the reliability of the error annotation and minimize researcher's subjectivity, two teachers who have been teaching English in high schools for over 15 years and both have MA in English Education were consulted. With 25 learners' texts, the teachers and the researcher categorized almost 90% of errors into the same code. The main disagreement among the learner's errors was shown in the <LP> code which indicates that lexical errors affect word combinations, i.e., compound units, phrasal verbs, and idioms (e.g., The singer delivers <LP corr="a lot of">a lot</LP> emotions and ideas through his songs.). In the next round, all three reached an agreement on error categorization in 36 different learners' texts. The rest of learners' written texts in 6 sub-corpora of SCOHLEE were annotated by the researcher himself.

Table 4. Categories of errors (Granger et al., 2022)(continued)

Category	Subcategory	Errors	N
<F>, Form	Spelling <FS>	misuse or omission of capital letters	2
		issues concerning the doubling of consonants/vowels, letter substitution or misordering, misuse/omission of hypens between words or parts of words	
	Morphology <FM>	errors on derivational affixes	
<G>, Grammar	Determiners <GD>	<GDD> Demonstrative determiners	4
		<GDO> Possessive determiners	
		<GDI> Indefinite determiners	
		<GDT> Other types of determiners	
	Articles <GA>	<GA> all problems of definite, indefinite or zero articles	1
	Nouns <GN>	<GNC> Noun case	2
		<GNN> Noun numbers	
	Pronouns <GP>	<GPD> Possessive pronouns	7
		<GPP> Personal pronouns	
		<GPO> Possessive pronouns	
		<GPI> Indefinite pronouns	
		<GPF> Reflexive and reciprocal pronouns	
		<GPR> Relative and interrogative pronouns	
		<GPU> Unclear pronominal reference	
	Adjective <GADJ>	<GADJO> Adjective order	3
		<GADJN> Adjective number	
		<GADJCS> Adjective comparative/ Superlative	
	Adverbs <GADV>	<GADVO> Adverb order	2
		*<GADVCS> Adverb comparative/Superlative	
	Verbs <GV>	<GVN> Verb number	6
		<GVM> Verb morphology	
		<GVNF> Non-finite/finite verb forms	
		<GVV> Verb voice	
		<GVT> Verb tense	
		<GVAUX> Auxiliaries	

Table 4. Categories of errors (Granger et al., 2022)

Category	Subcategory	Errors	N
<G>, Grammar	Word Class <GW>	<GWC> Word class	1
	C omplementation <X*CO>	<XADJCO> Adjective complementation <XNCO> Noun complementation <XPRCO> Preposition complementation <XVCO> Verb complementation	4
	Dependent prepositions <X*PR>	<XADJPR> Adjective dependent preposition <XADVPR> Adverbs dependent preposition <XNPR> Noun dependent preposition <XVPR> Verb dependent preposition	4
<X>,	Nouns: Uncountable <XNUC>	<XNUC> Uncountable/countable noun	1
<X>, Lexico -Grammar	C omplementation <X*CO>	<XADJCO> Adjective complementation <XNCO> Noun complementation <XPRCO> Preposition complementation <XVCO> Verb complementation	4
	Dependent prepositions <X*PR>	<XADJPR> Adjective dependent preposition <XADVPR> Adverbs dependent preposition <XNPR> Noun dependent preposition <XVPR> Verb dependent preposition	4
	Nouns: Uncountable <XNUC>	<XNUC> Uncountable/countable noun	1
<L>, Lexis	Lexical single <LS>	<LSADJ> Lexical errors affecting adjectives <LSADV> Lexical errors affecting adverbs <LSN> Lexical errors affecting nouns <LSPR> Lexical errors affecting single or complex prepositions <LSV> Lexical errors affecting verbs	5
	Lexical phrase <LP>	<LP> Lexical errors affecting fixed word combinations	1
	Lexis, Connectors <LCL>	<LCLS> Single Logical Connector <LCLC> Complex Logical Connectors <LCC> Coordinating conjunctions <LCS> Subordinating conjunctions	4
<W>, Word	Word Redundant, Word missing, Word order <W>	<WR> Word redundant <WM> Word missing <WO> Word order	3
<Q>, Pun ctuation	Punctuation <Q>	<QM> missing punctuation <QR> redundant punctuation <QC> confusion of punctuation marks <QL> punctuation mark instead of lexical item	4
<Z>, Infe licities	Infelicities <Z>	Register problems, questions of political correctness and stylistic problems, unclear sequences of words or sentences	1
Total			55

* Lexico-grammar (X) errors result from the learner's lack of knowledge not of sentence grammar, but of word grammar (Granger et al. 2022). Granger et al. (2022) used <G> for annotating the violation of general sentence grammar. Word grammar means the lexico-grammatical properties that a particular word has (e.g., the fact that the adjective *interested* is followed by the preposition *in* rather than *about* or that the verb *decide* is followed by a to-infinitive rather than a gerund form).

Results

Overall Distribution of Learners' Errors

With respect to the overall distribution of errors, errors in every category were analyzed by using the frequency of each type of error, and in terms of its percentage. Figure 2 illustrates the distribution of errors in the major categories.

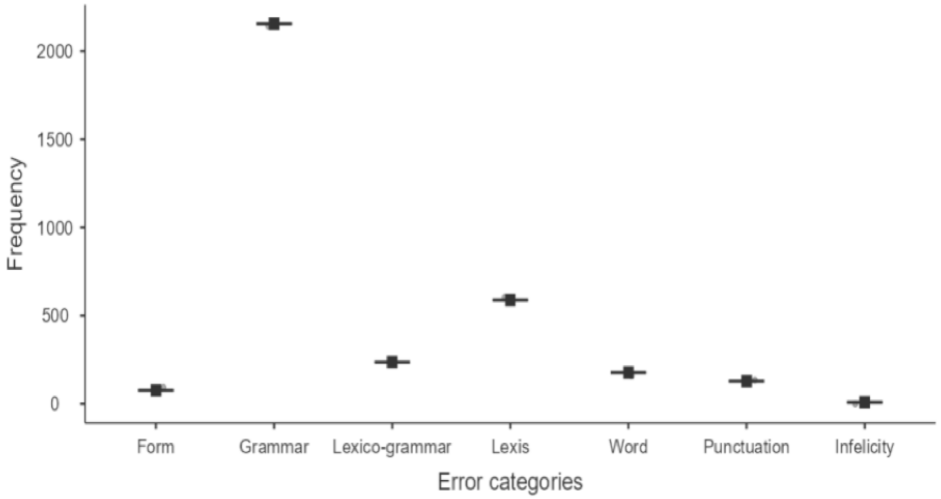

Fig. 2. Distribution of errors in major categories

Of the seven major categories, the learners produced the most frequent errors regarding grammar (G*, N = 2,155, 63.9%). Lexis (L*, N = 588, 17.4%) was the second most frequent major error type, and Lexico-grammar (X, N = 236, 7.0%) was the third highest frequent error type, followed by Word-related error (W, N = 177, 5.2%), Punctuation (Q, N = 128, 3.8%), Form (F, N = 76, 2.25%), and Infelicities (Z, N = 8, 0.2%). Despite grammar instruction being the primary focus of reading classes in many Korean high schools (Cho, 2004; Lee, 2008), learners frequently produced grammatical errors. The learners also displayed difficulties in selecting incorrect prepositions, verbs, and verb complements. These results can be attributed to a lack of practice in applying the grammatical principles in their writing.

As shown in Table 5, a further analysis of subcategories of <G> revealed that GA (N = 1005, 29.8%, e.g., Dori is a good housekeeper and <GA corr="a">0</GA>sociable dog.) was the highest frequency error in <G>. The second highest frequency was GNN (N = 357, 10.6%, e.g., The second reason is each <GNN corr="book">books</GNN> has distinctive stories.) and the third highest frequency error was GVN (N = 196, 5.8%, e.g., Here <GVN corr="are">is</GVN> three ways you can do.) followed by GWC (N = 124, 3.7%, e.g., Obesity curses many <GWC corr="health">healthy</GWC> problems.), GVNF(N = 106, 3.4%, e.g., I started <GVNF corr="to listen to">to listening to</GVNF> music.) and GPP (N = 70, 2.1%, e.g., My parents are focusing their attention on <GPP corr="me">my</GPP>.). The least frequency error in the category <G> was GADVCS (e.g., So I can sympathize with their emotions <GADVCS corr="more deeply">deeplier</GADVCS>.) with one occurrence and GADJO (e.g., I want to see<GADJO corr="my current team">current my team</GADJO> members playing the games) with two occurrence. It is worth noting that GA, GNN, GDO (e.g., Why don't you listen to the song and find <GDO corr="its">their</GDO> attractiveness?), and GPO (e.g., I was amazed by <GPO corr="its">it's</GPO> beauty.) errors are all related to nouns. A careful examination of the subcategories of <G> revealed that Korean high school learners have grammatical difficulty in three main areas: articles, nouns, and verbs.

Table 5. Distribution of grammatical errors

No.	Error	Freq(%)	No.	Code	Freq(%)
1	GDD	15(0.4)	14	GPU	3(0.1)
2	GDO	45(1.3)	15	GADJCS	12(0.4)
3	GDI	24(0.7)	16	GADJN	0(0)
4	GDT	0(0)	17	GADJO	2(0.1)
5	GA	1,005(29.8)	18	GADVO	11(0.3)
6	GNC	32(1)	19	GADVCS	1(0)
7	GNN	357(10.6)	20	GVN	196(5.8)
8	GPD	5(0.1)	21	GVM	11(0.3)
9	GPP	70(2.1)	22	GVNF	106(3.1)
10	GPO	19(0.6)	23	GVV	39(1.2)
11	GPI	10(0.3)	24	GVT	23(0.7)
12	GPF	3(0.1)	25	GVAUX	9(0.3)
13	GPR	12(0.4)	26	GWC	124(3.7)

A detailed examination of subcategories of each major error type gives a more detailed picture of the learners' grammatical difficulties (Subcategories of lexis errors <L> are in Fig. 3).

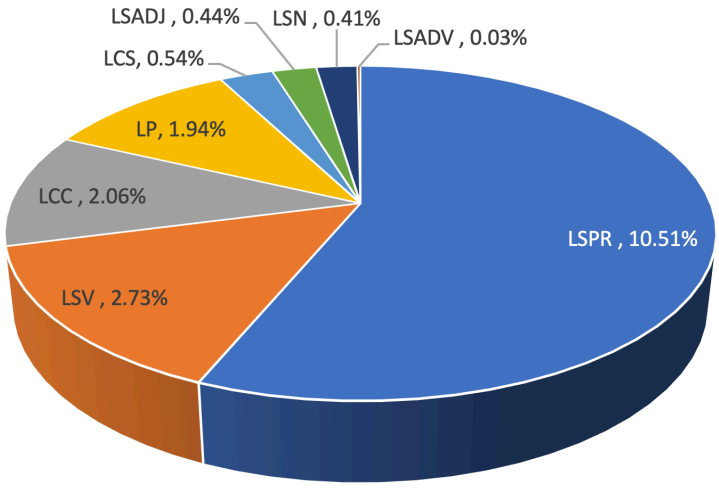

Fig. 3. Breakdown of errors in lexis categories

LSPR: Prepositions, LSV: Verbs, LCC: Coordinating Conjunctions, LP: Phrases, LCS: Subordinating Conjunctions, LSADJ: Adjectives, LSN: Nouns, LSADV: Adverbs

Fig. 3 reveals that of lexical errors, the highest frequent lexical errors were related to misusing prepositions (LSPR, N = 331, 10.5%, e.g., Some people may think that the internet is not good <LSPR corr="for">to</LSPR> us.). The learners also omitted prepositions or inserted superfluous prepositions in the sentences. The second highest frequent errors were related to verbs (LSV, N = 86, 2.7%, e.g., It was very tough at first but my body <LSV corr="adapted">adopted</LSV> to the routine). The third highest frequent lexical errors were LP (N = 61, 1.9%, e.g., When I played my favorite song with my guitar <LP corr="for the first time">first time</LP>). LP errors affect fixed word combinations (i.e., phrasal

verbs, idioms and compound units). Erroneous use of adjectives (LSADJ, N = 14, 0.44%, e.g., There are thrilling, and <LSADJ corr="scary">scared</LSADJ> stories in this book.) was the fourth frequent error, followed by LSN (N = 13, 0.41%, e.g., Eating ice cream removes <LSN corr="h

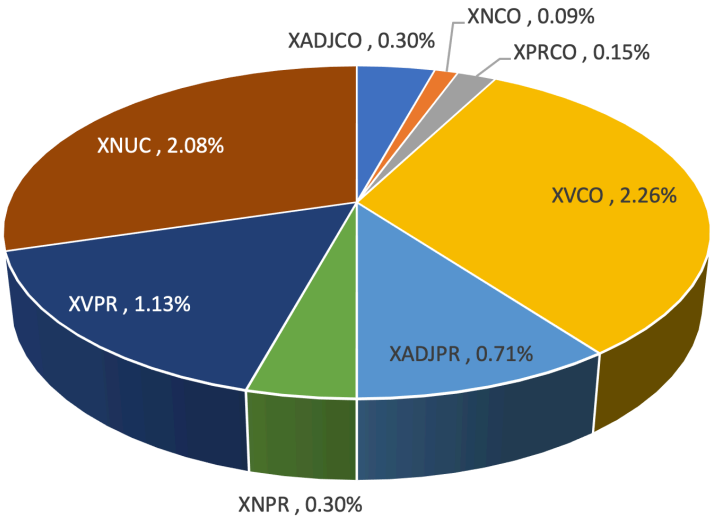

Fig. 4. Breakdown of errors in lexico-grammar category

XVPR: Verb-dependent Prepositions, XNPR: Noun-dependent Prepositions, XADJPR: Adverb-dependent Prepositions, XVCO: Verb Complements, XPRCO: Preposition Complements, XNCO: Noun Complements, XADJCO: Adjective Complements, XNUC: Noun UnCountable

The breakdown of the Lexico-grammar <X> category in Fig. 4 indicates that the learners had difficulty in complementation. Of complementation errors, verb complementation (XVCO, N = 76, 2.41%, e.g., what I <XVCO corr="want to say">want say</XVCO> is that let's try to play an instrument.). Other erroneous complements were found including XADJCO, adjective complements, (N = 10, 0.3% e.g., when you <XADJCO corr="considerate of">considerate to</XADJCO>others), and XNCO, noun complements, (N = 3, 0.09%, e.g., we can have the <XNCO corr="possibility of winning">possibility to win</XNCO> a prize.). The second most common Lexico-grammar errors observed in the learners' writing were related to the use of uncountable nouns. Specifically, 2.1% of the errors (N = 70) were classified as XNUC errors, such as the use of **informations* instead of *information* in the sentence (e.g., it is easy to find <XNUC corr="information"> informations</XNUC>). As seen in Figure 4, Korean high school learners also produced errors in using dependent prepositions such as XVPR, XNPR, and XADJPR. Among X*PR errors, XVPR were found the most frequent errors. (N = 38, 1.1%, e.g., by exchanging opinions <XVPR corr="with">0<XVPR> friends), followed by XADJPR (N = 24, 0.76%, e.g., people are obsessive <XADJPR corr="with">of</XADJPR> appearance), and XNPR (N = 10, 0.3%, e.g., you can have <XNPR corr="confidence in">confidence about</XNPR> yourself).

Distribution of Errors' across Proficiency Levels

Learners with different proficiency levels experienced difficulty in different areas of learning. Table 6 shows the distribution of ten most frequent errors across proficiency levels.

Table 6. Distribution of learners' errors across proficiency

No.	High		Mid		Low	
	Error	%(Freq)	Error	%(Freq)	Error	%(Freq)
1	GA	21.8(193)	GA	46.6(364)	GA	30.2(448)
2	LSPR	8.1(72)	GNN	18.1(141)	GNN	11.1(165)
3	GNN	5.8(51)	LSPR	15.6(122)	LSPR	9.2(137)
4	GVN	3.8(34)	GVN	7(55)	GVN	7.2(107)
5	QM	3.4(30)	GWC	5.4(42)	GVNF	4.2(62)
6	GWC	3.1(27)	QM	5.4(42)	WM	3.8(56)
7	FS	2.1(19)	GVNF	4.6(36)	GWC	3.7(55)
8	WM	2(18)	LSV	4.6(36)	QM	3.4(50)
9	XNUC	1.8(16)	XVCO	4.5(35)	WR	2.6(39)
10	LSV	1.6(14)	WM	3.7(29)	FS	2.5(37)

As seen in Table 6, the study revealed that mid-level learners made errors at the highest proportion among the three proficiency levels, which is consistent with previous research (e.g., Lee, 2016). High-level learners had the highest rate of GA errors (21.8%), followed by LSPR (8.1%), GNN (5.8%), GVN (3.8%), QM (3.4%), and GWC (3.1%). Similarly, mid-level learners produced article-related errors (GA) at the highest rate (46.6%), followed by GNN (18.1%), LSPR (15.6%), GVN (7%), and QM (5.4%). Meanwhile, low-level learners produced GA errors at the highest rate (30.2%), followed by GNN (11.1%), LSPR (9.2%), GVN (7.2%), GVNF (4.2%), and WM (3.8%).

Notably, low-level learners showed different error tendencies compared to the other groups. GVNF and WM errors were produced at the sixth and seventh most frequent errors in this group. It is worth noting that there was a significant difference in the rate of GVNF errors among the low-level (4.0%), high-level (1.3%), and mid-level (2.9%) learners. However, for WM errors, the difference between high-level and low-level was negligible (3.0% and 3.6%, respectively), while there was a slight difference between mid-level and low-level (2.4% and 3.6%, respectively).

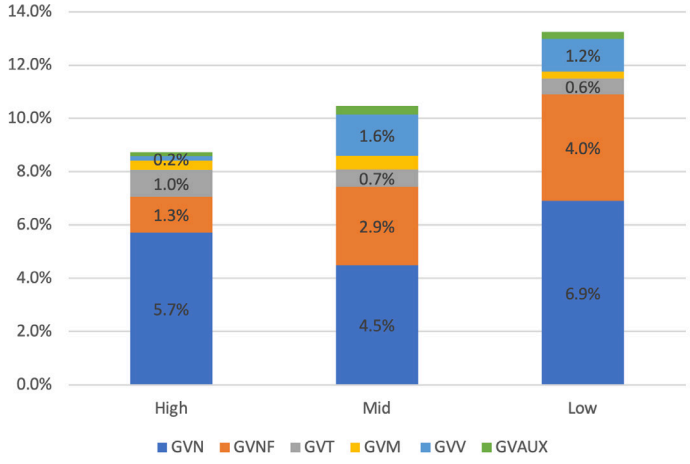

Fig. 5. Distribution of verb-related errors by learners' proficiency
GVN: Verb Number, GVNF: Verb Finite/Non-finite, GVM: Verb Morphology, GVV: Verb Voice, GVAUX: Verb Auxiliary

The most distinct errors that correlate with learners' proficiency was verb-related errors (GV*). Figure 5 presents the details of verb-related errors across different learners' proficiency. Figure 5 illustrates that low-level learners produced verb-related errors (GV*) at the highest proportion (N = 205, 13.2%) and high-level learners produced the GV* errors at the lowest proportion (N = 53, 8.7%). In terms of subcategory, errors of subject and verb agreement (GVN) were the most common types of errors at all levels, followed by misusing finite/non-finite verb forms (GVNF, e.g., It's fun <GVNF corr="to spend">spend</GVNF> time), and erroneous use of passive/active voice (GVV, e.g., most people <GVV corr="have seen">have been seen</GVV> TV dramas).

It is worth noting that the proportion of GVNF increased as learners' proficiency decreased (1.3% for high-level, 2.9% for mid-level, and 4% for low-level). In the case of GVN, mid-level was the lowest among the three levels, whereas mid-level learners produced the highest rate of GVV errors. This suggests that mid and lower–level learners struggle more with selecting and applying the appropriate form of the verb in their writing.

Distribution of Errors across Semesters

This study analyzed learner errors on a semester-by-semester basis to investigate whether the patterns of learner errors change. Table 7 shows Top 10 sub-categories of errors showing a decrease in an error rate over time.

Table 7. Top 10 errors showing a decrease in the second semester

No.	Error	1st term		2nd term		
		Freq (a)	% (b)	Freq (c)	% (d)	d-b(c-a)
1	LSPR	175	10.9	156	8.8	-2.1(-19)
2	LSV	53	3.3	33	1.9	-1.4(-20)
3	GA	486	30.4	519	29.3	-1.1(33)
4	GPP	42	2.6	28	1.6	-1(-14)
5	GVT	19	1.2	4	0.2	-1(-15)
6	GVN	101	6.3	95	5.4	-0.9(-6)
7	GDO	29	1.8	16	0.9	-0.9(-13)
8	XVCO	43	2.7	33	1.9	-0.8(-10)
9	LCS	13	0.8	4	0.2	-0.6(-9)
10	Z	8	0.5	0	0.0	-0.5(-8)

As presented in Table 7, the sub-category LSPR (prepositions) exhibited the most substantial decrease in the rate among the 55 sub-categories analyzed, followed by LSV (verb choice), GA (articles), GPP (personal pronouns), and GVT (verb tense). Notably, with regard to GA (articles), while the overall frequency increased, its proportion in the second semester decreased by 1.1%. Furthermore, in the case of Z (infelicities), there were no occurrences of sentence fragment errors in the second semester.

Table 8. Top 10 errors showing an increase in the second semester

No.	Error	1st term		2nd term		
		Freq (a)	% (b)	Freq (c)	% (d)	d-b(c-a)
1	QM	38	2.4%	84	4.7%	2.4(46)
2	LCC	19	1.2%	46	2.6%	1.4(27)
3	GWC	49	3.1%	75	4.2%	1.2(26)
4	GNN	161	10.1%	196	11.1%	1(35)
5	WR	28	1.8%	46	2.6%	0.8(18)
6	GVNF	44	2.8%	62	3.5%	0.8(18)
7	FS	28	1.8%	44	2.5%	0.7(16)
8	XNPR	0	0.0%	10	0.6%	0.6(10)
9	LSN	2	0.1%	11	0.6%	0.5(9)
10	LP	25	1.6%	36	2.0%	0.5(11)

Based on the findings presented in Table 8, it is evident that different learners experienced difficulty in different categories of errors in the second semester. The error categories of QM (omission of punctuations), LCC (logical connectors), GWC (word choice), GNN (noun numbers), and WR (word addition) showed an increase in both the rate and frequency of errors. Specifically, QM exhibited the highest rate of increase, followed by LCC, GWC, GNN, and WR. The sixth highest increase was observed in the category of GVNF (finite/non-finite verb) errors, followed by FS (spelling), XNPR (noun dependent prepositions), LSN (misuse of nouns), and LP (combinations). It is noteworthy that XNPR errors were exclusively observed in the second semester, with a frequency of 10. This increase can be attributed to the increased use of NP and PP in the second semester. This trend of increased use of NP was also found in the higher rates of LSN (noun choice) and GNN (noun numbers) errors in the second semester. To see a more clear picture of the change in learners' errors, this study conducted qualitative analyses of learners' errors.

As seen in the example (1), in the first semester, a high-level learner produced a <GVN> error, misusing the noun *news* as plural nouns which in turn leads to the misuse of the incorrect determiner *many*. From the learner's perspective, *many news* should be in agreement with its verb *are*. In the second semester, the learner produced a <GVN> error by misusing *have* for *has*. It is interesting that the learner used the plural form *have* correctly in the second sentence in (2).

(1) Since there <GVN corr="is">are</GVN> so <GDI corr="much">many</GDI> fake internet news these days, books by specialists are much more reliable to learn something new. 2021H10209

(2) We can find that foreigners are really interested in Hallyu culture. Our one and only traditional culture, and K-pop have all influenced on this Hallyu boom. Our own traditional culture <GVN corr="has">have</GVN> attracted the eyes of the whole world. We should all preserve and develop our sole and beautiful Hallyu culture. 2021H20209

Discussion and Conclusion

A total of 3,368 errors were identified in the Korean high school learners' corpus (SCOHLEE). Among 55 sub-categories, GDT, GADJN, XNCO, LCLC, LCLS, QC, and QL errors were not found in the corpus, SCOHLEE. It is interesting to note that GADJN errors (e.g., the <GADJN corr="poor">poors</GADJN> people) were not found as in Chuang and Nesi's (2006) study. The most frequent errors produced by Korean first year high school learners were GA

(29.8%). The result is consistent with the previous studies on Korean high school learners' errors (Lee, 2008), and those on Korean college learners' (Ahn & Kang, 2015; Hwang, 2012; Kim, 2020; Lee, 2016).

The present study examined the types of errors made by Korean learners of English, with a particular focus on frequency and proportion. The results of the analysis revealed that grammatical errors (G*) were the most common type of error, accounting for a significant proportion (63.9%) of the total errors. G was followed by Lexis (17.4%), Lexico-grammatical errors (X*, 7%), and Word-related errors (W*, 5.2%). Punctuation errors were found at a rate of 3.8% of total errors, followed by Form (2.3%), and infelicities were found at the lowest rate (0.2%). These findings are consistent with the previous studies (Granger et al., 1998; Hwang, 2012; Lee, 2008; McDonald et al., 2013). However, the present study also found that the most frequent errors differed from those reported by Song and Park (2012), who found that punctuation errors were the most common type of error among Korean foreign language high school learners. This disparity can be attributed to the feedback and correction process employed in the current study, which may have influenced the types of errors made by the participants. In addition, the feedback and correction process may have contributed to the lower rate of infelicities in the learners' writing, as only a small number of sentence fragments (0.2%) were identified in the present study, all of which were produced by mid-level learners using the subordinator *because* (e.g., Because the lyrics of the song are touching.).

A closer look at the distribution of grammatical errors <G> showed that article errors (GA) were the most problematic grammatical elements for Korean high school learners. This finding is in support of previous research. Murakami and Alexopoulou (2016) argued that grammatical morphemes that exist in the learners' L1 are more likely to be used accurately compared to those that do not. This is because these morphemes are already established in the learners' language system, and can be easily transferred to the L2. MacDonald et al. (2013) further explored the influence of L1 on the types of errors made by learners, showing that learners with different L1 backgrounds demonstrate distinct patterns of grammatical errors. For instance, learners with a Latvian L1 exhibited a notably higher frequency of GA compared to learners with other L1 backgrounds. Especially, Latvian L1 made notably more GA errors than other participants. The overall findings of the present study corroborate this argument. In addition to the impact on grammatical morphemes, the present study also found evidence of L1 influence on two specific types of grammatical errors produced by Korean high school learners: noun number (GNN) and subject-verb agreement (GVN) errors. These errors reflect differences in the linguistic systems of Korean and English, which may lead to difficulties in acquiring and using these grammatical structures in the L2. The present study's identification of noun number (GNN) and subject-verb agreement (GVN) errors as particularly frequent among Korean high school learners aligns with previous research on L2 English acquisition by Chinese speakers. Specifically, Chuang and Nesi (2006) found that Chinese L1 English learners made GNN and GVN errors at the second and third highest frequency, respectively. Chuang and Nesi (2006) developed learning materials for English articles based on their research results, and Lee (2007) also developed 'Data-Driven Learning (DDL)' materials to prevent and treat article errors in the English writing of Korean high school students. In line with these, the findings of the present study could contribute to the development of grammar learning materials.

It is the most intriguing finding that errors related to verbs (GV*) show a correlation with learners' English proficiency. Few studies have reported similar results to the present study. High-level learners produced GV* errors at the lowest rate

(N = 53, 8.9%), followed by mid-level (N = 126, 10.3%), and low-level learners (N = 205, 13.2%). Specifically, GVNF (N = 8, 1.3%) showed a marked margin between high and low-level group (1.34% for high and 4% for low-level group).

This study is not exempt from limitations. Firstly, the size of the learner corpus utilized in this study is relatively small, consisting of 135,240 running words. Second, the formal English test scores (e.g., TOEIC, TOEFL or CEFR levels) were not available to decide learners' English proficiency levels. Lastly, the present study relied on Grammarly to annotate learners' erroneous use of English. However, it should be noted that Grammarly occasionally suggested the incorrect form of verbs in complex sentences (e.g., wrong subject-verb agreements, finite verbs instead of non-finite ones).

Despite some limitations, based on the findings of this study, some pedagogical implications can be suggested. To reduce learners' errors, it is important to design learning activities adopting Data-Driven Learning (DDL) approach that present both accurate and erroneous expressions, thereby enabling learners to make informed choices regarding the appropriate usage. It is also suggested that, based on the interconnected nature of lexico-grammatical errors (e.g., LSV and XVCO), a holistic and integrated language instruction could be implemented to address multiple error categories simultaneously. Moreover, research in Second Language Acquisition (SLA) has revealed that form-focused instruction has a positive effect on acquiring L2 structures by drawing learners' attention to them (Doughty, 1991; Doughty & Williams, 1998; Ellis, 2001).

Drawn from the limitations, future research may consider the investigation of language development by categorizing error types according to learners' school grade and their proficiency level. This would give a more detailed analysis of how English writing skills evolve among Korean high school students over time. Such a study could provide valuable insights into the specific difficulties faced by different proficiency groups of Korean high school learners at different stages of their development, leading to more targeted strategies.

Despite some limitations, this study contributes to the existing body of knowledge by shedding light on the correct and incorrect use of lexis, grammar, and lexico-grammar among Korean high school learners. The present study also offers empirical evidence, serving as a foundational resource for error correction. Both teachers and textbook writers can develop teaching-learning materials for diagnosing their learners' most frequent errors, and adopt Data-Driven Learning (DDL) approaches that present both accurate and erroneous expressions, thereby enabling learners to make informed choices regarding the appropriate usage.

References

Ahn & Kang (2015). An analysis of written errors of Korean adult learners of English. Modern English Education, 16, 67-89.

Chuang, F. Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. Corpora, 1, 251-271.

Corder, S. P. (1967). The significance of learner's errors. International Review of Applied Linguistics 5, 162-170.

Dagneaux, E., Dennes, S., & Granger, S. (1998). Computer-aided error analysis. System, 26, 163-174.

Dagneaux, E., Dennes, S., Granger, S., & Meunier, F. (1996). Error tagging manual version 1.1 & 1.2. Louvain-la-Neuve: Center for English corpus Linguistics, Universite Catholique de Louvain.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. Studies in Second Language Acquisition, 13, 431-469.

Doughty, C., & Williams, J. (1998). Focus on Form in Classroom Second Language Acquisition. Cambridge University Press.

Dulay, H., Burt., & Krashen, S. (1982). Language Two. Oxford University Press.

Ellis, R. (1989). Are classroom and naturalistic acquisition the same? Studies in Second Language Acquisition, 11, 305-328.

Ellis, R. (1994). The Study of Second Language Acquisition. Oxford University Press.

Ellis, R. (2001). Investigating form-focused instruction. In R. Ellis (Ed.), Form-focused Instruction and Second Language Learning (pp. 1-46). Blackwell.

Ferris, D. R. (2002). Treatment of Errors in Second Language Student Writing. The University of Michigan Press.

Gass, S., Behney. J., & Plonsky. L. (2020). Second Language Acquisition: An Introductory Course (5th ed.). Routledge.

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

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. (pp. 3-33). John Benjamins.

Granger, S., Swallow, H., & Thewissen, J. (2022). The Louvain Error Tagging Manual Version 2.0. Louvain-la-Neuve: Center for English corpus Linguistics, Universite Catholique de Louvain.

Hwang, E. K. (2012). Korean EFL learner's language development across proficient levels in written productions. English Teaching, 67, 27-50.

Kim, J. Y. (1998). Error analysis: A study of written errors of Korean EFL learners. Journal of the Applied Linguistics Association of Korea, 14, 21-58.

Kim, J. Y. (2020). A corpus-based analysis of syntactic/semantic errors in university level EFL learners' writing. Lingua Humanities, 22, 135-156.

Kim, S. G. (2016). Revisiting causes of grammar errors: Commonly confused lexical category dyads in Korean EFL student writing. English Teaching, 71, 61-85.

Lee, D. J. (2004). A computer-aided research into error analysis of Korean secondary school learners' writing. Foreign Languages Education, 11, 133-160.

Lee, D. J. (2007). Corpora and the classroom: A computer-aided error analysis of Korean students' writing and the design and evaluation of data-driven learning materials (Unpublished Doctoral Dissertation). University of Essex, Colchester, United Kingdom.

Lee, D. J. (2008). A Computer-Aided Error Analysis (CEA) of a Korean learner corpus. Modern English Education, 9, 73-89.

Lee, S. K. (2016). Effects of language proficiency on error patterns of Korean EFL university Students' writing. English Language Teaching, 28, 95-116.

MacDonald, P., Garcia-Carbonell, A., & Carit-Sierra, J. (2013). Computer learner corpora: Analyzing interlanguage errors in synchronous and asynchronous communication. Language Learning & Technology, 17, 36-56.

Murakami, A., & Alexopoulou, T. (2016). L1 influence of the acquisition order of English grammatical morphemes: A learner corpus study. Studies in Second Language Acquisition, 38, 365-401.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30, 555-578.
Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics, 24, 492-518.
Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics, 10, 209-231.
Song, H. Y., & Park, E. S. (2012). An error analysis of foreign language high school students' English. Korean Journal of English Language and Linguistics, 12, 425-449.
Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. The Modern Language Journal, 97, 77-101.

Authors Information

Park, Myungho: https://orcid.org/0000-0002-0770-1657
Lee, Dong Ju: https://orcid.org/0000-0002-9906-835X