



A Corpus-based Study on the Use of English Relative Clauses by Korean EFL Learners of Different Proficiency

Sooyeon Kang

Seoul National University

Sun-Young Oh

Seoul National University

The current study examined Korean EFL learners' use of English relative clauses (RCs) in essays with a special focus on proficiency-specific characteristics. The overall frequency distribution of RCs, RC structural types, and RC-related errors were examined by comparing two different levels of learner corpora and the native speaker corpus. The findings reveal native speakers' more frequent use of RCs compared with learners and the overall proficiency effect on the number of RC uses. In terms of RC structural types, native speakers and learners were more alike than different, and regardless of proficiency level, learners' use of RC structural types showed supporting evidence for Noun Phrase Accessibility Hierarchy (Keenan & Comrie, 1977). The numbers and types of RC-related errors were different depending on learners' proficiency levels. From these findings, some useful pedagogical implications are drawn and discussed for level-specific English RC instructions.

Keywords: English relative clauses, second language processing, error analysis, proficiency, learner corpora

Introduction

A relative clause (hereafter RC) is a grammatical feature that has caught great attention in a large body of literature due to its grammatical complexity, high frequency, and universality across languages (Wiechmann, 2014). In English, it has intrigued researchers in both first (L1) and second language (L2) acquisition and sentence processing for decades (DeDe, 2015; Diessel & Tomasello, 2005; Eckman et al., 1988; Gennari & MacDonald, 2008; Hwang, 2003; Izumi, 2003; Kidd, et al., 2007; Macdonald et al., 2020; Traxler et al., 2005; Wells et al., 2009). The primary interest of these experimental studies was in revealing the relative ease and difficulty of processing different RC structures.

Several accounts exist concerning the processing of RCs. Earlier syntax-based accounts such as Noun Phrase Accessibility Hierarchy (Keenan & Comrie, 1977) and Parallel Function Hypothesis (Sheldon, 1974) explained the difference in processing RCs by dividing them into several different structures. More recent accounts (Gennari & MacDonald, 2008; Traxler et al., 2005; Wells et al., 2009) presented more elaborate explanations on the difference by considering modulating factors (e.g., the referential properties or the animacy of the head nouns) for such difference.

In line with the latest view that corpus linguistic techniques can methodologically complement experimental studies to understand human language processing (Gablasova et al., 2017; Wiechmann,



2014), RCs have also been a core subject of corpus-based analysis. L1 corpora and L2 learner corpora have been investigated with a particular interest in how English RCs are used by L1 and L2 speakers. Corpus analyses of English RCs used by L1 speakers revealed the distributional patterns of different RC structures concerning the characteristics of the head nouns (Kidd et al., 2007; Reali & Christiansen, 2007; Roland et al., 2007; Shin, 2020). Taking a frequency-based viewpoint, some have also provided evidence on the role of frequency of occurrence in establishing processing ease for certain RCs (Kidd et al., 2007; Reali & Christiansen, 2007).

The interests of corpus-based analysis of English RCs used by L2 learners (Baek & Guo, 2014; Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Shin, 2018; Sung, 2014; Takashima, 2000) do not significantly differ from those of L1 studies in that many of them focused on revealing distributional patterns of English RC usage by learners. Among those studies, some also paid attention to the errors displayed in learners' use of English RCs (Baek & Guo, 2014; Lee & Shin, 2017a). Except for one study (Bae & Lee, 2018), however, previous corpus-based studies on L2 learners' English RC use did not refer to a native corpus in discussing the results drawn from learner corpora. Frequency information regarding features in learner corpora itself is valuable evidence about learners' language use and development, but interpreting the evidence is difficult without a reference point (Gablasova et al., 2017). This study thus attempts to conduct a detailed examination of a Korean EFL (English as a foreign language) learner corpus compared to a native speaker corpus as a reference point as to their use of English RCs. To shed light on the possible difference in the use of English RCs according to learners' proficiency level, a topic that was dealt with only in a limited number of studies with a small size corpus (Baek & Guo, 2014; Sung, 2014), we pay special attention to proficiency-specific characteristics in RC uses by dividing the Korean learner corpus into two sub-corpora in terms of the learners' proficiency level. Given the error-prone nature of the target structure for L2 learners (Cowan, 2008; Larsen-Freeman & Celce-Murcia, 2016; Lee et al., 2011), the study also examines the English RC-related errors displayed in the learner corpus to provide pedagogical implications on what to focus on in teaching English RCs to different levels of learners. The followings are the research questions that are addressed in the current study:

1. How often do Korean EFL learners use English RCs? To what extent does the proficiency level of the learners affect the frequency of RCs?
2. What types of English RC structures are produced by the different levels of the learners?
3. What types of errors do the learners produce in the use of RCs? Do the frequency and types of errors vary depending on their proficiency levels?

Literature Review

L1 Studies of English RCs

Along with the acquisition order of grammatical morphemes, the developmental sequence of some key grammatical structures has long been a topic for research on language acquisition. As one of those structures, English RCs have been focused on in the field of psycholinguistics with a particular interest in how they are processed. Various theoretical accounts have been proposed to explain RC processing difficulties. Among them, Keenan and Comrie's (1977) Noun Phrase Accessibility Hierarchy (NPAH) has received much attention. Based on the extensive comparative study of RC structures over more than fifty typologically different languages, the researchers suggested that the grammatical roles that the head of RCs serves within the RCs are hierarchically ordered, and the relative accessibility to these grammatical roles reflects psychological ease of comprehension. According to the hypothesis, RCs are divided into six different types: subject RC, direct object RC, indirect object RC, object of preposition RC, genitive RC, and object of comparison RC, with the order of mention reflecting the degree of accessibility from the highest to the lowest. Another well-known early syntax-based account is Sheldon's (1974)

Parallel Function Hypothesis (PFH), which is based on the order of difficulty in comprehending RCs by L1 children. Unlike the NPAH, the PFH centers both on the function of the head noun in the matrix clause and of the relative pronoun in the RCs. The main claim is that processing ease occurs when the grammatical function of the head noun is identical with that of its coreferential relative pronoun.

Other syntax-based accounts that have caught attention include Perceptual Difficulty Hypothesis (PDH) (Kuno, 1974) and Subject-Object Hierarchy Hypothesis (SOHH) (Hamilton, 1994). The PDH explains RC difficulty order with the concepts of center-embedding and right-embedding of the RC. Regardless of RC types, a center-embedded RC interrupts the processing of the matrix sentence and is thus perceptually more challenging to process, whereas a right-embedded RC does not cause such interruption, being perceptually easier to process. Putting forward the SOHH, Hamilton (1994) divided RCs into four different types, considering both the position of RCs in matrix sentences and the role of the NP relativized in the RCs. According to this classification, OS is the easiest to process (with a single discontinuous S), followed by OO (with a discontinuous S and a discontinuous VP) and SS (with a discontinuous S and center embedding) with equal difficulty, and the most difficult RC type is SO with one discontinuity by center embedding and two discontinuities by two phrasal discontinuities (i.e., S-node, and VP) within the RC (see Table 2 for examples).¹

The aforementioned accounts for RC processing difficulty have been tested by experimental studies using diverse tasks with both L1 children and adults. By examining the frequency of four RC structures in native speakers' discourse, Stauble (1978) found supporting evidence for the PDH. Meanwhile, other studies yielded stronger support for the NPAH (Bates et al., 1984; Tavakolian, 1981). The NPAH's claim that subject RC (SR) is more accessible than object RC (OR) has also been discussed extensively as one of the underlying reasons for SR-OR asymmetry in processing English RCs (Baek, 2012). Recently, taking a usage-based approach, which claims the critical role of frequency of occurrence in establishing such preference for particular RC structures, an increasing number of studies has encompassed corpus-based analysis on English RCs (Kidd et al., 2007; Reali & Christiansen, 2007). The studies have shown that the distributional patterns of English RCs generally mirror processing difficulties found in the experiments, thereby reinforcing the role of speakers' linguistic experience in processing sentences (e.g., Roland et al., 2007).

L2 Studies of English RCs

Similar to L1 studies of English RCs, earlier studies on English RC acquisition by L2 speakers were also conducted to evaluate the competing hypotheses in the field. In the examination of the production data from ESL learners, for example, Schumann (1980) and Wong (1991) found empirical evidence for the PDH as the learners produced RCs in the frequency order of OS, OO, SS, and SO. The diversity in measurement, however, led to somewhat different results across studies. In other L2 studies, more robust support for the NPAH was detected (Hwang, 2003; Izumi, 2003; Kim, 2014; Lee et al., 2011) as in L1 studies. In a similar vein, sentence processing literature revealed that L2 learners' English RC processing was not dissimilar to that of L1 speakers, showing a preference for SR over OR (Baek, 2012).

Recently, an increasing number of studies have begun to utilize learner corpora in examining how English RCs are used by L2 learners with a typologically different L1, for example, Chinese (Baek & Guo, 2014), Japanese (Takashima, 2000), or Korean (Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Shin, 2018; Sung, 2014). Many of these studies, despite variations in particular focus in analysis, examined the distributional patterns of RC structures, discussing the findings in view of the RC acquisition hypotheses. They discovered that learners of English produce more SR than OR, which can be interpreted as supporting evidence for the NPAH (Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Sung, 2014). When the structural types of RCs were further analyzed, however, considering the relationships between the head noun of RCs in the matrix clause and the relative pronoun of the embedded clause (e.g., OS, OO, SS, and

¹ SS: Subject head NP, Subject target of relativization; SO: Subject head NP, Object target of relativization; OS: Object head NP, Subject target of relativization; OO: Object head NP, Object target of relativization

SO), the studies found a different frequency order, thereby presenting different arguments concerning the hypotheses. Some apparently endorsed the NPAH by revealing learners' more frequent use of OS and SS type of RC structures than the others (Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Takashima, 2000) while others showed learners' greater use of OS and OO types, which was then discussed in relation to the processing discontinuity (Baek & Guo, 2014).

Among the studies mentioned above, two, in particular, deserve a separate mention in the context of the current study. Baek and Guo (2014) analyzed a spoken corpus of Chinese learners of English, which was divided into two sub-corpora according to proficiency level, in terms of the correct versus incorrect use of RCs. They showed that learners' proficiency level affects the overall frequency and correct use of English RCs and that incorrect use is attributable to the learners' L1. As acknowledged by the researchers themselves, however, the finding is based on a rather small corpus consisting of 207 Chinese learners' oral narratives. An examination of a learner corpus that is of larger size and/or different modality and compiled from learners with a different mother tongue may increase our understanding of L2 learners' use of English RCs.

Lee and Shin (2017a), on the other hand, analyzed the use of English RCs in the writing of Korean high school and college freshman learners and compared the findings with the corpus of Korean high school English textbooks. The analysis demonstrated that the two groups of learners differed significantly in the use of the relative pronoun *which*, with high school learners overusing it with more grammatical errors. In comparison with the textbook corpus, the learner corpora were found to show a significant overuse of restrictive RCs with a rare attempt to use nonrestrictive counterparts. The researchers argued that their corpora demonstrated supporting evidence for the NPAH as learners produced OS and SS types most frequently in their writing. Errors also occurred most frequently in the OS and SS types, with the most frequent type being a disagreement between the head of RCs and the verb of the embedded clause. Although it was a large-scale study involving three types of corpora, the scope of analysis was limited to RCs using the relative pronouns *who* and *which*, and the focus of the analysis was mainly upon the distributional patterns of restrictive versus nonrestrictive RCs in addition to RC-related errors.

The previous learner-corpus-based research on English RCs suggests that L2 learners may use RCs in a similar way to L1 speakers showing more frequent use of SR than OR. Nevertheless, a large-scale learner corpus study is still lacking which systematically examines how L2 learners use a full range of English RCs. In particular, it is an underexplored topic in the field whether and how learners use RCs differently according to their L2 proficiency level. Thus, to fill this gap, the present study seeks to investigate how Korean learners of English with different proficiency use English RCs by examining both the distributional patterns of different RC structures and RC-related errors in a large corpus.

Method

Data

The present study used the Yonsei English Learners Corpus (YELC) as the main corpus of analysis and the Louvain Corpus of Native English Essays (LOCNESS) as a reference corpus to which learners' texts were compared. YELC consists of two types of essays collected from over 3,000 Korean high school graduates or equivalents who were admitted to Yonsei University (Rhee & Jung, 2014). One portion of YELC is comprised of personal essays on familiar topics, and the other section, argumentative essays in which learners express their opinions on a given topic. All the texts in YELC are divided into nine proficiency levels based on the Common European Framework of Reference (CEFR), ranging from A1 (the lowest) to C2 (the highest). To compare high- and low-level learners' use of English RCs, the present study categorized YELC into two sub-corpora, YELC HP (B2, B2+, C1, C2) and YELC LP (A1, A1+, A2), excluding B1 and B1+ for a clearer distinction.

LOCNESS contains American and British college students' argumentative and literary essays and is widely used as a reference corpus in learner corpus research (Granger, 2015), including those that used YELC (Choi, 2019; Lee & Oh, 2018). To exclude the possible differences in the use of RCs between American and British English, and to make the size of the three corpora more comparable, the present study analyzed only the sub-corpus of essays written by American students. A detailed description of the three corpora is presented in Table 1.

TABLE 1
The Three Corpora

	YELC LP	YELC HP	LOCNESS
Number of text files	1,820	996	232
Tokens	272,717	222,047	188,396
Types	235,074	196,027	167,015

Data Analysis Procedure

For data analysis, the study used Sketch Engine, a powerful online text analysis tool available from <http://www.sketchengine.eu> (Kilgarriff et al., 2014). Specifically, data analysis involved the following procedures. First, five relative pronouns (*that*, *who*, *which*, *whom*, and *whose*) were searched for using the built-in Corpus Query Language (CQL) function in the 'Concordance' setting.²

The second step involved inspecting the concordances manually to identify the occurrences of *that*, *who*, *which*, *whom*, and *whose* as restrictive relative pronouns. Considering the dissimilar functions of restrictive and nonrestrictive RCs concerning the preceding noun phrases (Larsen-Freeman & Celce-Murcia, 2016) and a much higher frequency of restrictive RCs than the nonrestrictive counterpart in the written register (Biber et al., 1999), the present study focuses only on the use of restrictive English RCs. Thus, the use of *who* and *which* as an interrogative and nonrestrictive relative pronoun, and the use of *that* as a demonstrative, conjunction, and complementizer were sorted out and deleted. Also excluded are the uses of *who* and *that* in cleft construction, the instances of *whom* as an interrogative pronoun, and *whose* as a nonrestrictive relative pronoun. The following data were further eliminated from the analysis: clauses consisting of the same consecutive words (e.g., *object that has special meaning*), which apparently come from the prompts of the essay, and clauses of which the meaning was hard to interpret (e.g., *Smoker is smoking around nonsmoker who their body is crash too*). After all the irrelevant data were removed, the frequencies of occurrence of each relative pronoun were calculated in the three corpora.

In the next step, the target RCs were classified into one of the four structural types, i.e., OS, OO, SS, and SO, according to the relationships between the head of RCs in the matrix clause and the relative pronoun of the embedded clause. For the categorization, we referred to the examples for each RC structure presented in *The Grammar Book* (Larsen-Freeman & Celce-Murcia, 2016). Table 2 shows the specific examples for each structure type adapted from the book.

² A CQL function in Sketch Engine is used to search for a (or series of) token(s) in a corpus (Jakubíček et al., 2010). For instance, as the first step to search for *that* used as a relative pronoun, sentences containing *that* with no verb preceding it were extracted by using a query of [tag!="V.*|IN.*"] [lc="that"].

TABLE 2
RC Structure Types and Examples

RC Structure Types	Examples
OS	I know the boy <i>who</i> speaks Spanish. Mrs. Thomas is a teacher <i>who</i> prepares her lessons. He gave the girl <i>who</i> broke the door a warning. She talked with the boy <i>who</i> speaks Spanish.
OO	I like the place <i>that</i> you recommended. I know the place <i>which</i> he spoke about. I work for the woman <i>that</i> you met. Tom knows about the girl <i>that</i> I gave the pen to. She knows of the place <i>which</i> Mary spoke about.
SS	The boy <i>who</i> speaks Spanish is my nephew.
SO	The man <i>who(m)</i> you met is my cousin. The woman <i>that</i> I gave the umbrella to is over there. The place <i>which</i> he talked about is New York.

Next, errors in the use of RCs were identified in the two learner corpora and classified manually. As in Kim and Yoo (2015) who used both pre-determined and post-determined categories in analysing learner errors, the study used the RC error classification suggested in previous studies (Baek & Guo, 2014; Lee & Shin, 2017a; Lee et al., 2011) with necessary modifications. Errors that are not directly related to forming RCs, such as spelling, tense, articles, verb usages in terms of semantic features, and subject-verb agreement, were not considered.

Lastly, to evaluate the statistical significance of the differences in frequency across types of RCs and errors within and across corpora, chi-square tests were carried out.

Results and Discussion

Overall Frequencies of RCs

This subsection deals with the question of how often the learners at different proficiency levels use RCs in essays as compared with the native speakers of English. Table 3 presents both the raw and normalized (per 100,000 words) frequencies of occurrence of RCs using restrictive relative pronouns *that*, *who*, *which*, *whom*, and *whose* in each of the three corpora.

TABLE 3
Overall Frequencies of *That*, *Who*, *Which*, *Whom*, and *Whose*

Corpus	<i>That</i>	<i>Who</i>	<i>Which</i>	<i>Whom</i>	<i>Whose</i>	Total
YELC LP	622 (228)	568 (209)	216 (80)	7 (3)	10 (4)	1423 (522)
YELC HP	530 (239)	696 (314)	242 (109)	15 (7)	13 (6)	1496 (674)
LOCNESS	846 (449)	372 (198)	234 (125)	7 (4)	7 (4)	1466 (779)

(Frequencies per 100,000 words)

TABLE 4

Results of Chi-square Tests Comparing Overall Frequencies in YELC LP/YELC HP/LOCNESS

	<i>That</i>	<i>Who</i>	<i>Which</i>	<i>Whom</i>	<i>Whose</i>	Total
χ^2	219.21***	75.90***	24.23***	5.23	1.61	120.06***

***statistically significant ($p < .001$)

As shown in Table 4, there was a statistically significant difference among the three groups in the total number of RCs used ($\chi^2 = 120.06$, $p < .001$). The overall frequencies in Table 3 reveal that the native speakers of English use RCs more frequently in comparison to both groups of learners. Among the two learner groups, the YELC HP was found to employ more of RCs in their writing than the YELC LP. The finding that learners with higher proficiency use RCs more often is in line with a previous study with Chinese learners of English (Baek & Guo, 2014), suggesting that the ability to utilize RCs appears to be related to learners' proficiency.

Distribution of Different Types of RC structures

In this subsection, the types of RC structures that learners and native speakers produced in their writing are presented and discussed. See Table 5 for the frequencies of the five relative pronouns across four RC structure types in each corpus.

TABLE 5

Overall Frequencies of RC Structure Types

Corpus		RC structure types				Total
		OS	OO	SS	SO	
YELC LP	Frequency	658(242)	200(74)	376(138)	47(18)	1281(472)
	Percentage	51.3%	15.7%	29.3%	3.8%	100%
YELC HP	Frequency	735(331)	197(89)	308(139)	46(21)	1286(580)
	Percentage	57.1%	15.4%	24.0%	3.6%	100%
LOCNESS	Frequency	809(430)	255(136)	255(136)	61(33)	1380(735)
	Percentage	58.5%	18.5%	18.5%	4.5%	100%

(Frequencies per 100,000 words)

As shown in the table, the percentages of different RC structures for each group revealed both similar and disparate pictures. First, the three groups were comparable in demonstrating more frequent use of SRs than ORs (80.6%, 81.1%, and 77.0% for YELC LP, YELC HP, and LOCNESS, respectively). This finding is consistent with previous corpus-based studies that targeted speakers of English as L1 (Ronald et al., 2007) and L2 (Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Shin, 2018). According to Biber et al. (1999), the subject gaps that occur in SRs are more comfortable to process than the object gaps and thus are more common in the native speakers' written production. This argument is supported by the result of the present study, which at the same time ratifies the finding that learners' processing of English RCs is not dissimilar to native speakers' (Baek, 2012) as the learners in the current study also used SRs more frequently in their essays regardless of proficiency level.

When the relative percentages of the four RC structures were considered, learners and native speakers were again more alike than different. Both of the learner groups used OS most frequently, followed by SS, OO, and SO while the native speaker group produced the OS structure most often, followed by SS and OO with an equal rate, and then by SO. The lowest occurrence of the SO type in all three groups seems to derive from the fact that it imposes a heavier processing burden (Hamilton, 1994; Kuno, 1974; Sheldon, 1974) than the rest of the structures. In the same vein, the most frequent application of the OS type by all the groups may be interpreted as the structure being relatively easier to process compared to the others (Hamilton, 1994; Kuno, 1974; O'Grady, 1999) due to the least processing discontinuity required. More application of OS than SO can also be accounted for by an information structure principle (Larsen-

Learners' Errors in the Use of RCs

Overall frequencies of errors

The close inspection of the learner corpora produced a total of eight types of RC-related errors. Table 6 presents each type of error with examples taken from the YELC LP. Types 4, 5, 7, and 8 are new additions to the others which have been adopted from the previous studies (Baek & Guo, 2014; Lee & Shin, 2017a; Lee et al., 2011).

TABLE 6
RC Error Types and Examples

No.	Type	Examples
1	Agreement	*Allowing nickname makes <i>many rumors that is</i> not certified.
2	Antecedent omission	*You also have been trusted yourself that you will never meet accidents and <i>be the who</i> get wounded.
3	Antecedent use	* <i>They who</i> use harm words attach not only children and youths among diverse people.
4	Verb errors within RC	*I and my friend decided to get over stress <i>that be made</i> while we studied for 3 years. *The most interesting thing ... was the camp <i>that hold</i> in Songdo.
5	Incomplete RC	*We can find some medicines <i>that quite good</i> for animals but terribly bad for human beings. *But it is only thing <i>that I have a fun</i> in tough school life.
6	Resumption	*They are not the only things <i>that</i> we can use <i>them</i> by our thought.
7	RC as a teddy bear	*We have <i>rights that say their opinion</i> . *This will naturally decrease the time <i>that being spent</i> for smoking.
8	Mixed	*Students remember physical punishment <i>that don't forget every day</i> .

Type 1 errors refer to those in the number agreement between the head NP and the verb within an RC, and Type 2 demonstrates the error of omitting the head noun where it is required. Type 3 is the error in the use of antecedent, which includes the wrong form of a word (e.g., *they* instead of *those*) in the antecedent position. Learners also made errors with verbs within an RC, as in the examples of the Type 4 error, especially regarding inflections and the choice of the correct tense/aspect or voice. Type 5 errors concern the missing of a necessary syntactic element within an RC. As seen in the examples, there were instances of missing verbs and/or prepositions within RCs, which creates a syntactically incomplete clause. Type 6 errors indicate “the restatement of the coreferential noun” (Quintero, 1992, p. 47), and Type 7, which was detected predominantly among LP learners, is the errors caused by misusing relative pronouns where other noun postmodifiers such as *-ing/-ed*, infinitive or preposition plus *-ing* clauses should have been used. Lastly, there was an instance of more than two kinds of errors, all mixed (i.e., type 8).

Table 7 below shows the overall frequency of the errors made by the two learner groups in the use of the three relative pronouns, *that*, *who*, and *which* by (no errors were found in the use of *whom* or *whose*). It is seen that both groups made less frequent errors in the use of *who* (22.75% for HP, 8.72% for LP) compared to *which* (24.75% for HP, 11.16% for LP) or *that* (31.90% for HP, 14.07% for LP). This result may be interpreted with respect to the RC structures where *who* occurs. Compared to *which* or *that*, *who* prefers the OS and SS structures, which require relatively less processing burden on the learners. This processing ease, in turn, may have provided learners with more opportunities to make error-free RCs. When the two groups were compared, a significant difference was found across all types of relative pronouns (see Table 8), meeting the expectation that learners with higher proficiency would perform better in producing error-free RCs.

TABLE 7
Overall Error Frequency

Corpus	That	Who	Which
YELC LP	185 (31.90%)	111 (22.75%)	49 (24.75%)
YELC HP	64 (14.07%)	51 (8.72%)	25 (11.16%)

Note. Numeric figures in parentheses are the percentages of errors out of the frequency of each relative pronoun

TABLE 8
Results of Chi-square Tests Comparing Error Frequencies in YELC LP and YELC HP

	That	Who	Which
χ^2	43.37***	40.85***	13.42***

***Statistically significant ($p < .001$)

Distribution of types of errors

In order to see which error types are more common among the learners, the frequencies of each RC error type were calculated for each relative pronoun across the corpora. The result is shown in Table 9.

TABLE 9
Distribution of Error Types

Corpus		1	2	3	4	5	6	7	8	Total
That	LP	86	0	0	25	50	5	18	1	185
	HP	44	0	0	3	11	1	5	0	64
Who	LP	77	7	2	6	9	1	9	0	111
	HP	42	2	1	0	6	0	0	0	51
Which	LP	30	0	0	8	5	1	5	0	49
	HP	16	0	0	2	6	0	1	0	25
Total		295	9	3	44	87	8	38	1	485

First, Type 1, which is the error in number agreement between the antecedent NP and the verb within an RC, is found to be prevalent regardless of group. This result supports a previous finding where the most frequent RC-related error among both Korean high school and college-level English learners was the number disagreement between the antecedent and verb (Lee & Shin, 2017a). This is likely to be an L1-induced error since the subject-verb agreement is not a grammatical feature that exists in the learners' L1, Korean. Note also that among the fewer errors made in the use of *who*, most of them occurred in the number agreement (73.45%). When such erroneous sentences were scrutinized in detail, it was found that the antecedent was a general noun referring to humans (i.e., *people* or *person*). The error was especially noticeable because *people* was the most common antecedent of *who*, and *person*, the third. This type of error, however, was relatively rare (i.e., only one instance for HP and two instances for LP) when the antecedent was a regular plural noun such as *students*, the second most frequent antecedent of *who*. It thus seems to be the case that learners are more vulnerable to number agreement error when a particular class of words serves as antecedents of *who*. This finding, in fact, is in line with Lee and Shin's (2017a) study, according to which indefinite pronouns such as *someone* or *one* also belong to this special class of antecedents. In the present study, the LP learner group, but not the HP, showed the same tendency with the pronoun *someone*. By incorporating learner proficiency as a variable, this study thus complements the previous finding and provides a pedagogical suggestion targeted specifically for low-level L2 learners.

Other commonly problematic areas for both groups were the use of a correct verb form within an RC (i.e., Type 4 errors) and producing an incomplete RC (i.e., Type 5 errors). The qualitative analysis of the two types of errors demonstrated that both groups of learners struggle with selecting an appropriate verb form for a given antecedent NP. As shown in examples (3) and (4) below, in many cases, learners chose an active voice where the passive form of the verb was necessary.

- (3) *So military service *that should complete* in youth has different influence to each men.
(YELC HP group)
- (4) *The most interesting thing that happened this past vacation was the camp *that hold* in Songdo.
(YELC LP group)

The type 4 error is particularly noteworthy in that it is not the kind of RC-related errors that have been traditionally mentioned in the literature. Errors such as resumptive pronouns and inappropriate choice of relative pronouns for antecedents have been discussed widely as common RC-related errors by L2 learners (Cowan, 2008; Larsen-Freeman & Celce-Murcia, 2016). The use of an incorrect verb form within RCs, however, was found to be the second most problematic area for the learners of the current study. While this error may be constrained to the EFL learners whose L1 is Korean, the finding attests to the presence of distinctive types of RC-related errors that present challenges for a particular group of learners and need to be carefully considered in teaching them English RCs.

Meanwhile, for Type 5 errors, both groups omitted an obligatory preposition within an RC when they used *that* (e.g., She was a girl who went to a same high school *that I graduated.*). With the LP group, in particular, the missing of a be-verb within RCs was prevalent across all types of relative pronouns (e.g., We can find some medicines *that quite* good for animals but terribly bad for human beings.). The scrutiny of the two types of errors thus suggests that teachers need to direct their instructional efforts into helping learners to choose a grammatically correct form of verbs within an RC for the antecedent NP and not to miss any complements within an RC such as subjects, verbs, objects, and prepositions.

Lastly, it can be seen in Table 9 that the YELC LP produced the most diverse types of errors in using the relative pronoun *that*. In particular, the instances of Type 7 errors (i.e., RC as a teddy bear) were widespread among LP learners. The LP group misused the relative pronoun in the position where other postnominal modifiers such as *-ing/-ed*, infinitive or prepositional clauses are to be used instead. This may be a teaching-induced error in that restrictive relative pronouns are heavily emphasized in English classrooms as postnominal modifiers. Students who are given few opportunities to try out their languages through writing may thus overuse RCs to the degree that they inadequately alternate it with other types of postnominal modifying structures. Type 7 errors may also be interpreted with respect to the developmental stages for noun modifiers suggested by Biber et al. (2011). According to Biber et al.'s index of writing development covering a total of 28 linguistic features, there are 11 types of noun modifiers that are distributed from stage 2 to stage 5 (Lan & Sun, 2019).⁴ Among these diverse noun modifiers, *-ing/-ed* clauses and infinitive or prepositional clauses belong to stage 4 and 5 (i.e., the advanced stages), respectively, whereas relative clauses are put in a lower stage, i.e., stage 3. LP learners, who are at an early stage of writing development, may thus have incorrectly used relatively more basic noun modifiers (e.g., relative clauses) where advanced ones such as *-ing/-ed*, infinitive or prepositional clauses should be used. The overall broad spectrum of errors found among the LP learners in the use of *that* also suggests that despite the seeming ease of use, its correct use is not easy to develop, warranting more pedagogical efforts towards its successful acquisition.

To sum up, the investigation of learner errors in the use of RCs reveals that the learners produce various types of errors including those that have not been reported in the previous studies. The results also show a clear discrepancy in the number and types of errors depending on learners' proficiency level, which calls for level-specific instruction in teaching RCs.

⁴ Stage 1 includes no noun modifiers, and stage 2 has attributive adjectives as pre-modifiers. Stage 3 includes relative clauses and nouns as modifiers, and stage 4, prepositional phrases and *-ing* or *-ed* clauses. Finally, stage 5 is indexed by infinitive clauses, preposition + *-ing* clauses, noun complement clauses, and appositive noun phrases.

Conclusions

The current study aimed to explore the use of English RCs by Korean EFL learners as compared with native speakers, with a special focus on learners' proficiency level. It investigated various structural types of RCs used in texts written by two different levels of learners and the natives in an attempt to understand the relative ease or difficulty of RC processing. Another purpose of the study was to examine the types and frequencies of RC-related errors according to learners' proficiency level to provide pedagogical implications on level-specific English RC instructions. The research thus compared the overall frequency of RCs, four different RC structures, and RC-related errors by comparing the two different levels of learner corpora and the native speaker corpus.

The analysis of RC structural types used by the learners revealed supporting evidence for NPAH (Keenan & Comrie, 1979) as the learners produced OS and SS types more often than any others. This result is in accord with the previous literature that explored distributional patterns of English RC usage through learner corpora (Bae & Lee, 2018; Lee & Shin, 2017a, 2017b; Sung, 2014). Regarding the discussion on the processing ease or difficulty of RCs using the four RC structures (i.e., OS, OO, SS, SO) (Hamilton, 1994; Kuno, 1974, O'Grady, 1999; Sheldon, 1974), the frequencies of each type in the learner corpora showed clear processing ease for OS but difficulty for SO. What the present study contributes most to this line of research is the discovery of L2 proficiency effect on the use of English RCs. Not only did high-level learners employ RCs more frequently than low-level learners, but they produced fewer numbers and types of RC-related errors as well. In particular, the study revealed that Korean EFL learners produce more diverse types of errors in using English RCs than previously discussed in the literature, which in turn highlights the importance of considering learner characteristics such as L1 in English RC pedagogy. The results of the present study also disclosed both convergences and divergences between the learner groups and the natives. First, the native speakers turned out to use more RCs than both learner groups in writing. Concerning the RC structure types, the native speakers and the learners were found to be more alike than different, producing OS type most often and SO type the least.

Pedagogically, the current study provides some useful insights and implications for Korean EFL contexts. First of all, in designing instructions on English RCs, the relative ease or difficulty of processing different RC structures needs to be considered carefully. In terms of the sequence of instruction, the presumably more accessible RC structures such as OS and SS types should be presented first, followed by OO and SO types, with particular attention paid to the relatively tricky RC structures. Given our finding that learners' proficiency interacts with their use of English RCs in the overall application and the numbers and types of errors, instructional efforts need to be made that take into account learners' proficiency. Significantly more frequent errors made by low-level learners suggest that special attention and constructive feedback are needed to help these learners to acquire the appropriate use of RCs. As applying corpora has been shown to be useful in making learners aware of and self-correcting their errors on English RCs (Sahragard et al., 2014), the results of corpus analysis presented here may serve as a useful resource in developing teaching materials and/or providing instructions for Korean EFL learners.

While the present research contributes to a deeper understanding of English RC uses by L2 learners, it is not without limitations. The corpus analysis of the present study was limited to written production data and excluded object RCs with a zero relativizer. In order to fully understand how L2 learners use English RCs, future research may need to be conducted with both spoken and written data, including RCs where relative pronouns are not used. Besides, given the observation that different genres may depict different pictures of RCs and the RC structural distributions are related to the discourse roles that the structures play (Roland et al., 2007), learner corpus studies would benefit from including texts of diverse genres with further focus on the discourse functions that RC structures serve in the text.

The Authors

Sooyeon Kang is a Ph.D. candidate in the department of English Language Education at Seoul National University in Korea. She has been teaching English as a foreign language at a public middle school in Korea. Her current research interests include instructed second language acquisition, pragmatics, conversation analysis, task-based language teaching, and corpus linguistics.

Dept. of English Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 08826, Korea
Email: englishsk329@naver.com

Sun-Young Oh received her doctoral degree in applied linguistics at the University of California, Los Angeles. Currently, she is professor in the department of English Language Education at Seoul National University in Korea, where she teaches courses in applied linguistics, corpus linguistics, and discourse analysis. She has published widely in the areas of learner corpus research, corpus-based linguistics, second language pedagogy, and conversation analysis.

Dept. of English Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 08826, Korea
Tel: +82-2-880-7675
Email: sunoh@snu.ac.kr

References

- Bae, S. Y., & Lee, E. J. (2018). A corpus-based analysis of relative pronouns used in academic essays by Korean and American college students. *Studies in English Education*, 23(1), 203-229. <http://dx.doi.org/10.22275/SEE.23.1.08>
- Baek, S. (2012). *Processing of English relative clauses by adult L2 learners* [Unpublished doctoral dissertation]. The University of Illinois at Urbana-Champaign.
- Baek, S., & Guo, M. (2014). An analysis of relative clause acquisition in the spoken corpus of Chinese EFL learners. *English21*, 27(4), 483-508. <https://doi.org/10.35771/engdoi.2014.27.4.022>
- Bates, E., Devescovi, A., & D'Amico, S. (1999). Processing complex sentences: A cross-linguistic study. *Language and Cognitive Processes*, 14(1), 69-123. <https://doi.org/10.1080/016909699386383>
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5-35. <https://doi.org/10.5054/tq.2011.244483>
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Longman.
- Cho, D. W. (2009). A corpus-based study on the frequency and usage of English relative clauses in two science journals. *Foreign Language Education*, 16(3), 1-20.
- Choe, J.-W., & Song, J.-Y. (2013). The topical classification of essays by college student English learners using hierarchical clustering. *Language Information*, 17, 93-115
- Choi, W. (2019). A corpus-based study on “Delexical Verb + Noun” collocations made by Korean learners of English. *The Journal of Asia TEFL*, 16(1), 279-293. <http://dx.doi.org/10.18823/asiatefl.2019.16.1.18.279>

- Cowan, R. (2008). *The teacher's grammar of English*. Cambridge University Press
- Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, 882-906.
- DeDe, G. (2015). Effects of animacy on processing relative clauses in older and younger adults. *The Quarterly Journal of Experimental Psychology*, 68(3), 487-498. <https://doi.org/10.1080%2F17470218.2014.956766>
- Eckman, F. R., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied linguistics*, 9(1), 1-20. <https://doi.org/10.1093/applin/9.1.1>
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130-154. <https://doi.org/10.1111/lang.12226>
- Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, 58, 161-187. <https://doi.org/10.1016/j.jml.2007.07.004>
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 41-61. <https://doi.org/10.1075/ijlcr.1.1.01gra>
- Hamilton, R. (1994). Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning*, 44(1), 123-157. <https://doi.org/10.1111/j.1467-1770.1994.tb01451.x>
- Hwang, J. B. (2003). Korean EFL learners' acquisition of English relatives: PFP, PDH, or AHH? *English Teaching*, 58(3), 189-208.
- Izumi, S. (2003). Processing difficulty in comprehension and production of relative clauses by learners of English as a second language. *Language Learning*, 53(2), 285-323. <https://doi.org/10.1111/1467-9922.00218>
- Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. *PACLIC*, 741-747.
- Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1), 63-99.
- Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes*, 22(6), 860-897. <https://doi.org/10.1080/01690960601155284>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7-36. <https://doi.org/10.1007/s40607-014-0009-9>
- Kim, J. E., & Yoo, I. W. (2015). A corpus-based study of to-infinitive errors in Korean college freshmen's writing. *The Journal of Asia TEFL*, 12(4), 37-60. <http://dx.doi.org/10.18823/asiatefl.2015.12.4.2.37>
- Kim, N. G. (2014). A study on the acquisition of the English relative pronoun by Korean college students. *The Journal of Mirae English Language and Literature*, 19(4), 181-207.
- Kim, W. Y. (2018). *Hedging in Korean EFL high school graduates' writing* [Unpublished master's thesis]. Seoul National University, Seoul.
- Kuno, S. (1974). The position of relative clauses and conjunctions. *Linguistic Inquiry*, 5(1), 117-136.
- Lan, G., & Sun, Y. (2019). A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course. *Journal of English for Academic Purposes*, 38, 14-24. <https://doi.org/10.1016/j.jeap.2018.12.001>
- Larsen-Freeman, D., & Celce-Murcia, M. (2016). *The grammar book: Form, meaning and use for English language teachers*. National Geographic Learning with Heinle Cengage Learning.
- Lee, E. Y., & Oh, S. Y. (2018). Korean EFL learners' use of I in English argumentative writing: Focusing on genre-specific and proficiency-specific characteristics. *Language Research*, 54(2), 357-381. <https://doi.org/10.30961/lr.2018.54.2.357>

- Lee, K. E., & Shin, J. A. (2017a). A corpus-based study on the use of relative pronouns by Korean learners of English. *Korea Journal of English Language and Linguistics*, 17(4), 817-837. <https://doi.org/10.15738/kjell.17.4.201712.817>
- Lee, J. & Shin, K. Y. (2017b). A syntactic analysis of English relative clauses used by Korean EFL speakers: A corpus-based study. *English21*, 30(3), 219-245. <https://doi.org/10.35771/engdoi.2017.30.3.011>
- Lee, J., Lee, G., & Kim, K. J. (2011). Teaching English relative pronouns for Korean EFL high school students. *English21*, 24(1), 181-201. <https://doi.org/10.35771/engdoi.2011.24.1.008>
- Macdonald, R., Brandt, S., Theakston, A., Lieven, E., & Serratrice, L. (2020). The role of animacy in children's interpretation of relative clauses in English: Evidence from sentence–picture matching and eye movements. *Cognitive Science*, 44(8). <https://doi.org/10.1111/cogs.12874>
- O'Grady, W. (1999). Toward a new nativism. *Studies in Second Language Acquisition*, 21(4), 621-633. <https://doi.org/10.1017/S0272263199004040>
- Quintero, K. W. (1992). Learnability and the acquisition of extraction in relative clauses and wh-questions. *Studies in Second Language Acquisition*, 14(1), 39-70. <https://doi.org/10.1017/S0272263100010469>
- Real, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57(1), 1-23. <https://doi.org/10.1016/j.jml.2006.08.014>
- Rhee, S. C. & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Journal of the Korean Contents Association*, 14(1), 1019-1029. <https://doi.org/10.5392/JKCA.2014.14.11.1019>
- Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3), 348-379. <https://doi.org/10.1016/j.jml.2007.03.002>
- Sahragard, R., Kushk, A., & Ansari-pour, E. (2014). The application of corpora in teaching grammar: The case of English relative clause. *Pan-Pacific Association of Applied Linguistics*, 17(1), 79-93.
- Schumann, J. (1980). The acquisition of English relative clauses by second language learners. In R. C. Scarcella, & S. D. Krashen (Eds.), *Research in Second Language Acquisition* (pp. 118-131). Newbury House.
- Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior*, 13(3), 272-281. [https://doi.org/10.1016/S0022-5371\(74\)80064-2](https://doi.org/10.1016/S0022-5371(74)80064-2)
- Shin, K. Y. (2018). Relation between frequency and processing difficulty of English relative clauses by L2 learners: A learner corpus analysis. *The Journal of Studies in Language* 34(3), 491-504. <https://doi.org/10.18627/jslg.34.3.201811.491>
- Shin, K. Y. (2020). Non-linear interactions of factors influencing relative clause distribution and their implications on relative clause processing. *Korean Journal of Linguistics*, 45(4), 919-940. <https://doi.org/10.18855/lisoko.2020.45.4.009>
- Stauble, A. (1978). *A frequency study of restrictive relative clause types and relative pronoun usage in English*. Unpublished manuscript.
- Sung, H. M. (2014). A study on the use of English relative clauses in Korean EFL learners' writing. *English Language Teaching*, 26(1), 65-83. <https://doi.org/10.17936/pkelt.2014.26.1.004>
- Takashima, H. (2000). Another look at the order of difficulty of relative clauses from corpus linguistics – statistical procedures, analysis and results. *International Review of Applied Linguistics in Language Teaching*, 38(3), 313-329. <https://doi.org/10.1515/iral.2000.38.3-4.313>
- Tavakolian, S. (1981). The conjoined-clause analysis of relative clauses. In S. Tavakolian (Ed.), *Language Acquisition and Linguistic Theory* (pp. 167-187). MIT Press.

- Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, 53(2), 204-224. <https://doi.org/10.1016/j.jml.2005.02.010>
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58, 250-271. <https://doi.org/10.1016/j.cogpsych.2008.08.002>
- Wiechmann, D. (2014). *Understanding relative clauses: A usage-based view on the processing of complex constructions*. Walter de Gruyter.
- Wong, J. (1991). Learnability of relative clauses: A Hong Kong case. *Working Papers of the Department of English*, 3(1), 108-117.

(Received July 29, 2021; Revised February 20, 2022; Accepted March 18, 2022)