Why is English relative clause extraposed?:
A discourse-based statistical approach*

Seung Han Lee
(Chodang University)

Lee, Seung Han. 2019. Why is English relative clause extraposed?: A discourse-based statistical approach. *Linguistic Research* 36(2), 213-240. This study aims to investigate to what extent grammatical variables are related to each other in order to predict the discourse-based classification of English relative clause extraposition (RCE) construction with the help of a total 597 corpus data. In doing so, an extraposed relative pronoun is assumed to function as a cohesive device, thereby classifying RCE construction into four types according to discourse manner: Discourse-Old-Old-RCE, Discourse-Old-New-RCE, Discourse-New-Old-RCE, and Discourse-New-New-RCE. This discourse-based classification tells us which information is prevalent within an extraposed relative clause as well as a subject NP. Multinomial logistic regression is used for determining which one among the indefiniteness/definiteness of a subject NP, the predicate type (i.e., passive, presentative intransitive, and predicative complement of copula *be*), and/or grammatical weight (i.e., subject NP-to-extraposed RC length ratio) primarily has the inclination for each type of RCE construction. The definiteness of a subject NP, passive voice, and positive value of the length ratio are critical predictors to bring about Discourse-Old-Old-RCE and Discourse-Old-New-RCE, whereas the indefiniteness of a subject NP is a main predictor of Discourse-New-Old-RCE and Discourse-New-New- RCE. Interestingly, only Discourse-New-Old-RCE taking up the largest portion in this corpus study has a negative correlation with the length ratio. In other words, both a small number of constituents within a subject NP and heavy information within an extraposed relative clause are strongly associated with the occurrence of Discourse-New-Old-RCE. Therefore, the conclusive statement is drawn that the grammatical heaviness within a restrictive relative clause becomes a powerful trigger for the extraposition. **(Chodang University)**

Keywords   RCE, definiteness, indefiniteness, predicate, length ratio, extraposition

1. Introduction

English allows considerable variability in the ordering of constituents, particularly in the area following a main verb (Wasow 2002). The language seems to give more freedom of ordering than is sometimes identified. One such variation of English structures is that only part of an element is postposed at the end of a sentence (Quirk et al. 1985). The most frequently occurring variation is the postmodification of a noun phrase, resulting in a discontinuous noun phrase:

(1) a. A rumor circulated widely [that he was secretly engaged to the Marchioness].
　　b. The problem then arose [(of) what contribution the public should pay].

　　　　　　　　　　　　　　　　　　　　　　　　　(Quirk et al. 1985)

As shown in (1), English allows the extraposition from NP, which extracts a clause or phrase from inside an NP and places it at the end of a sentence. This phenomenon is also witnessed from a restrictive relative. The relative clause has an anaphoric element whose interpretation is determined by its antecedent (Huddleston and Pullum 2002). The anaphoric item is overtly or covertly marked by the presence of one of the relative pronouns: *whom*, *who*, *which*, *whose*, etc. Interestingly, when a subject is modified by a restrictive relative clause, its modifier often moves to the end of a sentence:

(2) a. A gun [which I had cleaned] went off.
　　b. A gun went off [which I had cleaned]. (Ross 1967)

(3) [NP   S]NP   VP   OPT
　　　1   2   3   →
　　　1   0   3+2

In a canonical structure like (2a), a head noun is immediately modified by a relative clause, but the relative clause is extraposed from NP as exemplified in (2b). Ross (1967) postulates that the sentence like (2b) is derived from the one like (2a) by a grammar rule in (3). Considering all of these aspects, this paper

introduces the so-called 'English relative clause extraposition (hereafter RCE)' as illustrated by the attested corpus examples in (4):

(4) a. [NP Several articles] [VP have been [VP[*en*] written]] [which focus primarily on identifying specific psychological factors related to athlete alcohol use]. (COCA, 1995, ACAD)

   b. [NP A syllabus] [VP is [VP[*en*] made]] [AP available] [that outlines the rules and procedures]. (COCA, 2000, ACAD)

   c. [NP An agreement] [VP was [VP[*en*] reached] [AdvP yesterday] [which represents a sensible balance on maternity pay]. (BYU-BNC, 1992, W_HHV)

   d. However, the law does not normally allow for [NP any order] [VP[*inf*] to be made] [which will directly affect pension entitlements]. (BYU-BNC, 1992, W_FST)

   e. It is also concerned to avoid. [NP any precedent] [VP[*inf*] being established] [AdvP locally] [which might have serious and deleterious implications for the rural environment generally]. (BYU-BNC, 1985-1994, W_ GXG)

In all of these examples, non-canonical order appears when a grammatical relative clause is postposed to the position preceded by a main predicate. This shows discontinuous dependency, violating a typical X-bar structure, which implies that a modifier or complement should occur within the same maximal projection with its head.[1] A subject-modifying relative clause is detached from its head noun and a main predicate intervenes between two positions. Other grammatical categories such as AP, adverbial noun, and AdvP also occur between the main predicate and the extraposed clause as shown in (4b), (4c), and (4e). All of these characteristics of RCE construction can be described as a template given in (5):

---

1   The Complex Noun Phrase Constraint (CNPC): no element contained in a sentence dominated by a noun phrase with a lexical head noun may be out of the noun phrase by a transformation (Kim and Sells 2008).

(5) NP*i*[subject] + VP[*finite* or *infinite*] + (XP[Adjunct]) + [Extraposed Relative
Clause[GAP <NP*i*>]]

The feature of VP encompasses *to*-infinitive and gerundive as well as finite tense.
The type of XP is classified into AP and AdvP which function as an adjunct.
The gap NP*i* within an extraposed relative clause needs to be filled with a
subject NP*i*. This schema can explain the difference between RCE construction
and other similar constructions:

(6) a. The message had been made [that the woman had caused the
outburstand the tragic outcome]. (BYU-BNC, 1985-1994, CS1)
b. [A syllabus that outlines the rules and procedures] is made available.
(Lee 2017b)
c. Meanwhile, this objective, experimental mode dispels any qualms we
might feel about the fact that the research subjects [are] [humans] who
may have their own ideas and opinions about how they want to live.
(Lee 2017b)

These are all different from RCE construction for several reasons. In (6a) there is
no omission in a complement clause, whereas a main predicate in (6b) is not
followed by a relative clause. An NP in (6c), *humans*, does not function as an
adjunct. The template in (5) guarantees only the subject NP and the gap NP
within the extraposed relative clause to be co-referent to each other, but *humans*
in (6c) as a complement of copular *be* is entitled to have co-indexation with a
gap NP within an extraposed relative clause. Therefore, the sentence in (6c) is
not licensed as RCE construction.

This study aims to reveal the significance between syntactic properties and
discourse-based classification of RCE construction with the help of statistical
techniques. In doing so, we first describe the previous researches on RCE
construction in section 2. We also introduce the corpus findings of RCE
construction as well as its grammatical properties, thereby assuming the role of
an extraposed relative pronoun as a cohesive device, and asserting that RCE
construction is classified into four categories according to discourse manner in
section 3. The challenge for this classification will be a cornerstone to shed light

on the introduction to proper use of RCE construction; it tells us which information starts in a subject NP and ends in an extraposed relative clause. Section 4 provides multinominal logistic regression to check whether the discourse-based classification has a significant relation with the grammatical behavior of RCE construction. Finally we will conclude this study in section 5.

2. Literature review

There is an argument that old information tends to precede new information, and this generalization is called the GIVEN BEFORE NEW PRINCIPLE (Arnold et al. 2000). Given and old information is more accessible than new information from an angle of a continuum of accessibility, so the givenness precedes the newness. This continuity lends easy accessibility to the meaning of a sentence, thereby promoting the comprehension. Thus, speakers tend to postpone a constituent that refer to newly introduced entities. Francis (2010) asserts that RCE construction has a general inclination that old information is followed by a focused and new constituent. In other words, RCE construction plays a role in conveying new or important information. Song (2009) also points that the main function of non-RCE construction is to provide background information on discourse, whereas RCE construction emphasizes and highlights a new statement within an extraposed relative clause that a speaker intends to convey strongly. Interestingly, this corpus study proves that such a typical tendency cannot be generalized in RCE construction as follows:

(7)   It takes food much longer (a week or two, rather than half a day) to spoil, or **stops bacterial action** altogether. Once a product is thawed, however, the bacteria can become active again, multiplying under the right conditions. Enzyme activity is slowed down but not stopped during freezing. Canning: Provides a way to store foods for extremely long periods of time. Food is first boiled to **destroy bacteria** and inactivate enzymes. It is then placed in an airtight container. As the food cools, **a vacuum seal** is formed that prevents **any new bacteria** from getting in.

(COCA, 2002, MAG)

As shown in (7), this attested data tells us that RCE construction allows discourse-new information to be positioned in a subject NP, *a vacuum seal*, and simultaneously the information within the extraposed relatives is derived from prior background. This unique information structure occupies 61.31% (n=366) of a total of this corpus study.

Huck and Na (1990) delves into the discourse function of the extraposition from a subject NP. They assert that a definite subject NP of RCE construction can exist only when an extraposed relative provides a contrastive context:

(8) [The guy] just came in [that I met at TRENO's yesterday].

(Huck and Na 1990)

A sentence like (8) contrasts the guy from Treno's with the other at Andrea's if two guys have already been introduced in the previous context. They claim that if this situation does not occur, the definite subject NP only functions semantically as a proper noun, thereby not being able to be followed by an extraposed relative clause:

(9) *[John Smith] is here [who has three ears]. (Huck and Na 1990)

They insist that a sentence like (9) is pragmatically unacceptable because *John Smith* as a proper noun is a complete referring expression. However, this logic also cannot explain all of corpus data:

(10) a. [The HELPS intervention] was implemented [which caused an overall upward trend]. (COCA, 2015, ACAD)
    b. [Her granddaughter Charlotte] was born [who had thanked her for "fighting for paid family leave"]. (COCA, 2015, MAG)

Even if a definite subject NP in (10a) is not guaranteed to be compared with another *HELPS intervention*, the relative clause is still extraposed. The definite antecedent NP *Charlotte* in (10b) is also a proper noun, but it is still modified by the extraposed relative clause.

Grammatical complexity is critical factor for extraposing a heavy constituent,

thereby promoting the process of planning and production in language (Arnold et al. 2000; Al Khalaf 2018). In other words, a heavier constituent tends to be placed later than a lighter one:

(11) Heavy NP Shift (HNPS)
 a. The waiter brought [the wine we had ordered] to the table.
 b. The waiter brought to the table [the wine we had ordered]. (Arnold
  et al. 2000)

A heavy NP complement in (11a) moves to the place preceded by an oblique argument PP as illustrated in (11b). In the same line with this part, Wasow (2002) suggests Principle of End Weight (PEW):

(12) Principle of End Weight (PEW):
 Phrases are presented in order of increasing weight. (Wasow 2002)

(13) a. [A rumor] circulated widely [that he was secretly engaged to the
  Marchoiness].
 b. [A steering committee] had been formed, [consisting of Messrs Ogawa,
  Schultz, and Robinson]. (Quirk et al. 1985)

The heavy constituents such as a subordinate clause and infinitival VP, in (13a) and (13b) respectively, are placed at the end of sentences. Extraposing heavy elements to the final position serves to augment the probability of satisfying PEW. Understandably, postposing a heavy element at the end of a sentence helps people to understand both speakers' needs and promotes the process of comprehension (Arnold et al. 2000; Hawkins 2004; Kim 2018). In other words, the speakers tend to produce shorter and more accessible information in the front of a sentence, which enables them to have enough time for producing longer and less accessible information at the end of a sentence (Francis and Michaelis 2017):

(14) a. The informativeness account shares [with the processing account PP]
  [the claim that a subjectless sentence is not a grammatical option for

the child, and that the omission is due to some aspect of performance NP].

b. The informativeness account shares [the claim that a subjectless sentence is not a grammatical option for the child, and that the omission is due to some aspect of performance NP] [with the processing account PP]. (Arnold et al. 2000)

The sentence like (14b), which is an alternative order of (14a), is almost impossible to process. The extraposition of a long and complex constituent that is hard to produce provides speakers with more time to reformulate it. Interestingly, Strunk (2004) also claim that the average length of an extraposed relative clause within RCE construction in Low Saxon is longer than that of its canonical relative clause. Furthermore, he insists that the grammatical weight of an XP intervening between a subject NP and an extraposed relative clause can be crucial for extraposing a constituent.

Last, Francis and Michaelis (2014) provides an in-depth insight into RCE construction. However, they do properly not handle a predicative complement of copula *be* as well as infinitival RCE construction. They did also not provide the property of the passive voice which is definitely associated with the occurrence of RCE construction. This may be due to their small sample; their study is based on only 53 RCE examples extracted from ICE-GB.[2] Nevertheless, their research is in accordance with our finding that the length of an extraposed relative clause is almost always longer than that of a main predicate, thus supporting Francis (2010) and Francis and Michaelis (2017). This is also supported by Rasekh-Mahand et al. (2016)'s study that in Persian a longer main predicate increases the probability of non-RCE construction. From the perspective of these previous literatures, in this study, we adopt a subject NP-to-extraposed RC length ratio in order to predict RCE construction. Last of all, Rasekh-Mahand et al. (2016)'s study tells us that the passive voice is negatively associated with the occurrence of Persian RCE construction, but in this corpus study, most of English RCE construction is positively under the influence of passive

---

2   The International Corpus of English Great Britain (ICE-GB) includes one million words of British English in a variety of genres of both speech and writing. All examples are tagged with part of speech and grammatical functions.

construction (84.09%). In addition, in Persian the definiteness of a subject NP and the extraposition of a relative clause is strongly interrelated, but in English corpus data the indefiniteness of a subject NP significantly amounts to 89.45%.

3. Grammatical properties and discourse-based classification

3.1 Distributional properties: Corpus findings

To better grasp the grammatical properties of RCE construction, we employed corpus data rather than the data fabricated from linguists' intuition. The corpus we used here is COCA (Corpus of Contemporary American English) and BYU-BNC (Brigham Young University-British National Corpus), containing 410 million words of American English and 100 million words of British English, respectively.[3] We have first examined all of RCE construction by putting 99 combination patterns of part of speech (POS) into the search boxes of BYU-BNC and COCA corpora, and then sorted out RCE examples.[4] A total of 597 instances were extracted: 341 ones from COCA corpora and 256 ones from BYU-BNC.[5]

Indefiniteness (89.45%) of a subject NP is predominant in RCE construction, whereas a definite NP occupies a small portion (10.55%). Interestingly, *a/an* (53.37%) is the most common determiner in the indefiniteness category while *the*

---

3   This large amount of material introduces the state of spoken and written English from a wide range of sources, designed to represent a wide cross-section of English. They are freely-available online at http://corpus.byu.edu/coca/ and http://corpus.byu.edu.

4   BYU-BNC and COCA corpora provide parts of speech (POS) for an easy search, so we came up with all possible combinations of POS to extract examples of RCE construction.

5   The main reason why the data is not classified according to the genre is the limited examples within COCA and BYU-BNC. Thus, RCE structure is mainly analyzed as formal (9.21%) and written (90.79%) styles. RCE construction is more likely to be associated with a careful production, expository and informational purpose, and formal style rather than a spoken, interpersonal, casual and informal conversation. All examples of RCE construction are centered from 1970s to 2010s; most of the RCE corpus data stays on the top in the 1990s. Zhang (2015) asserts that the tremendous explosion of scientific popularization has made academic writing become an important register. In 1985 the Royal Society reported that the public understanding of science is highly associated with the enrichment of individual life and promotion of national prosperity, thereby leading scientists to pour a more massive effort on how well their works are spread over the public. This movement became an important factor for the explosion of RCE structure in written style.

(58.73%) is also the most frequently used in a definite subject. One interesting point is that the most remarkable type of a main predicate is passive voice (89.04%), followed by the secondary status of presentative intransitive (11.89%) where a verb contains an active intransitive, or denotes a concept related to existence or appearance, and introduces the referent of a subject NP (Francis and Michaelis 2014). On the other hand, adjectival predicate is relatively rare (4.02%) in RCE construction. In addition, the average number of constituents within the main predicate is 2.64 words. RCE construction does not allow any main predicate to have more than five words, implying that an extraposed relative clause can easily be traced or linked with its modifying subject NP.[6] The average number of words within the extraposed relative clause is 11.27 in comparison with that of subject NP, which is 2.26. The difference value D (number of the subject NP words minus number of the extraposed relative clause words) is about nine, suggesting one idea that the extraposition is related to the grammatical weight within a restrictive clause relative to a subject NP constituent. REC construction is preferred over its canonical (i.e., non-RCE) construction when the constituent length of a subject NP is shorter than that of the extraposed relative. This assumption is also supported by Huddleston and Pullum (2002)'s study that RCE construction is most likely to happen when the informational content of a restrictive relative clause is greater than that of its antecedent. Therefore, we assume that a subject NP length divided by an extraposed relative clause (RC) length (i.e., length ratio) can be a trigger to result in RCE structure.

3.2 Discourse and information structure

3.2.1 Information structure

Not all information presented in a discourse has equal status (Wasow 2002). Speakers often have main points they want to pass on, and they wish to highlight its core in some way. Some information tends to be used simply as

---

6   Francis and Michaelis (2014) assert that there is a strong preference for RCE construction when a relative clause is at least five times longer than the verb phrase.

background so as to make the main points approachable. The logic can be applied to individual sentences. A sentence can have partial portion critically contributing to that sentence, whereas other portions function as an informationally secondary role. In relation to this structure, information structure partitions a sentence into several categories such as focus, background, topic, and comment (Büring 2007). What is more, the information status of elements within a sentence can affect their syntactic structure, thereby making speakers tend to use particular structures to convey their thoughts in part. This leads to a more coherent discourse, promoting comprehension in language. Therefore, it is quite reasonable that the background information about what is being talked about should be positioned first. This ordering makes listeners understand the speakers' main points at the utterance. Consequently, the speakers usually lay the groundwork first and postpone the highlight of the utterance until later.

English introduces such a tendency to order 'given' information to be followed by 'new' information in an utterance (Ward and Birner 2001). 'Given' information is defined as the knowledge which exists in the consciousness of a listener at the time of an utterance, whereas 'new' information is interpreted as the information, being entered into the listener's consciousness by a speaker's utterance. We can check this relation from the perspective of topic-comment structure. In a sentence with topic-comment articulation, topic is the entity that an utterance is primarily about, or the entity of discourse that connects one part of discourse to other ones through 'given' information that runs through the whole discourse (Celce-Murcia and Olshtain 2000):

(15) a. [Your daughter] just killed [a bear]. (Dooley and Levinsohn 2000)
   b. [Rona] was the youngest of three sisters. [She] liked music and literature. Being the youngest sister was in some ways a blessing an in others a curse... (Celce-Murcia and Olshtain 2000)

As shown in (15a), *your daughter* is the topic and 'given' information, and *just killed a bear* is comment and 'new' information; *a bear* is focused. Comment is what is said about the topic and is new or added information. On the other hand, anaphoric in (15b) is topic since they link the topic of subsequent sentences across the discourse back to the initial mention of *Rona*. In most

languages, the topic regularly precedes the comment in sentences with topic-comment articulation (Dooley and Levinsohn 2000).

On the other hand, we can also examine the sentence structure on the basis of discourse-old, hearer-old, and discourse-new information. Ward and Birner (2001) mentions that discourse-old information is what has been evoked in the prior discourse, while hearer-old information is what a speaker believes to be present within a hearer's knowledge. This means that what is new to the discourse needs not to be the hearer-new:

(16) Last night [the moon] was so pretty that I called [a friend] on the phone and told [him] to go outside and look. (Ward and Birner 2001)

The moon in (16) is discourse-new but hearer-old, implying that the item has not been evoked in the prior discourse, but it can be assumed to be known to the hearer. *A friend* is simultaneously discourse-new and hearer-new, having not been previously evoked and also unknown to the hearer. *Him* represents discourse-old and hearer-old information, having been explicitly evoked in the previous clause (as *a friend*). Considering this fact, we can define that discourse-old information has been evoked in the prior discourse while discourse-new information has not been evoked. The ordering of these two is that discourse-old information generally precedes discourse-new one. For example, in passive voice construction, a syntactic subject must not represent newer information than does an NP within *by*-phrase (Ward and Birner 2001). The syntactic subject of such passive consistently represents discourse-old information within the discourse while an NP within *by*-phrase represents discourse-new information. In other words, passivization places familiar information before relatively unfamiliar one.

However, there is no consensus on how to categorize the information structure and how to identify the categories (Büring 2007). This implies that idiosyncratic instances occur which are not fitted into the 'discourse-old' & 'discourse-new' relation. It is not the case that a preposed constituent is always discourse-old and a postposed constituent always needs to be discourse-new. In Ward and Birner (2001)'s corpus study, left-dislocation introduces entities that are discourse-new, whereas a right-dislocated NP is discourse-old, which is given within discourse:

(17) a. Two of my sisters were living together on 18th Street. They had gone to bed, and a man, their girlfriend's husband, came in. He started fussing with one sister and she started to scream. [The landlady], [she] went up and he laid her out. (Ward and Birner 2001)

b. Below the waterfall (and this was the most astonishing sight of all), a whole mass of enormous glass pipes were dangling down into the river from somewhere high up in the ceiling! [They] really were enormous, [those pipes]. (Ward and Birner 2001)

The landlady in (17a) is new to the discourse, as being introduced in a subject NP position; the left-dislocated NP creates discourse-new information. On the other hand, in the right-dislocation which places an argument of a head verb in postverbal position, the sentence-final constituent in (17b) represents the information that has been evoked in a prior discourse. If a sentence-final constituent of right-dislocation is discourse-new, it becomes infelicitous. This unsymmetrical balance on the information structure can also be detectable from other sources. English grammar prevents a definite NP from introducing a new element in discourse, and it keeps an indefinite NP from referring to an element already introduced into the discourse (Huck and Na 1990; Kim 2017). The definite NP only refers to an individual to which an index has already been assigned, whereas the indefinite NP has no prior existential commitment or no established reference prior to the statement. In other words, the indefinite NP is typically used to express discourse-new entities, bearing informational focus and stress as well. Opposed to this facet, the definite NP can also receive focus in contrastive contexts:

(18) [The guy] just came in [that I met at TRENO's yesterday].

(Huck and Na 1990)

As already introduced earlier, if two guys have already been mentioned in a prior context, one guy from Treno's is contrasted with the other at Andrea's; the definite subject NP becomes focused. In the same vein with this point, Francis and Michaelis (2014)'s corpus findings reveal that a discourse-new subject NP can also be definite. Only 54.2% of a discourse-given subject NP was featured

with definiteness while only 17% of definite subject NPs was characterized with discourse-given information.

Therefore, with reference to the above explication about the information structure, we encode the discourse status of a subject NP and an extraposed relative clause. We do not follow a typical relation between 'givenness & definiteness' and 'newness & indefiniteness', but we interpret RCE construction on the basis of prior background information and context. In other words, we assume two information labels for an antecedent NP and its modifier respectively: 'Discourse-Old' (i.e., prior background), and 'Discourse-New' (i.e., no background information). The items identified within prior background of preceding lines are named as 'Discourse-Old', whereas the constituents with no prior mention of referents themselves are classified as 'Discourse-New'.

3.2.2 Extraposed relative pronoun as a cohesive device

We assume that an extraposed relative pronoun functions as a cohesive device. According to Halliday and Hasan (1976), among cohesive devices 'reference' includes pronouns, demonstratives, and comparatives, mentioning something outside the text, and referring to the person(s) or items(s) described within a previous (anaphoric) and succeeding (cataphoric) context:

(19) a. Look at [the Sun]. [It]'s going down quickly. (anaphoric)
   b. [It]'s going down quickly, [the Sun]. (cataphoric)

As shown in (19a), a pronoun *it* refers to the item *the Sun*, talked about within prior background while a demonstrative in (19b) subsequently indicates an item at the end of the sentence. At this point, we suggest that an extraposed relative pronoun also plays a role in connecting a subject NP with background information within an extraposed relative clause:[7]

7   Kiss (2005) also assumes that extraposition is a type of anaphoric relation. He insists that particularly restrictive relative clauses semantically choose the antecedent even if modifier and modified element are not adjacent. If the index of the antecedent NP is identified with the index of the relative clause, the content of the relative clause is integrated with the information of the antecedent NP, resulting in an intersective integration. In addition, the relative pronoun obligatorily needs an antecedent to receive an interpretation, so the antecedent and the relative

(20) a. [The back of the case] is raised slightly [which] [tilts the tablet to form
           a convenient drawing surface]. (BYU-BNC, 1985-1994, W_Goo)
      b. [The back of the case] is raised slightly [Ø] [tilts the tablet to form a
           convenient drawing surface].

As shown in (20a), an extraposed relative pronoun drives readers into
associating a subject NP with an extraposed relative clause whereas a null
extraposed relative pronoun in (20b) makes readers distracted about which verb
among *is* and *tilts* is connected with a subject NP, and further which noun
between *the back* and *the case* within the subject NP is interpreted as a subject for
each finite verb. In addition, the event process of (20a) is ordered in the manner
of 'raising' event preceded by 'tilting' event, but in the (20b) the 'tilting' event
follows the 'raising' event. Therefore, the extraposed relative pronoun is a
cohesive device which gives a right direction into event interpretation and offsets
the distance between the subject NP and the postposed relative clause. However,
one thing to make sure is we must presume that a relative pronoun of non-RCE
construction does not have to be treated as a cohesive device:

(21) a. [The back of the case] [which] [tilts the tablet to form a convenient
           drawing surface] is raised slightly.
      b. [*That* [The back of the case] [Ø] [tilts the tablet to form a convenient
           drawing surface]] is raised slightly.

As exemplified in (21a), a relative pronoun intervenes between an antecedent
and its restrictive relative clause without any distance. At this point, even if we
delete the relative pronoun, the antecedent and the relative clause are both
conjoined together and form a noun clause as a subject. It can be interpreted as
if *that*-complementizer is omitted in the subject position. Furthermore, there is no
difference in the interpretation between (21a) and (21b). In other words, both
have same event order that the 'tilting' event occurs before the 'raising' event. In
conclusion, we suppose that an extraposed relative pronoun is employed as a
cohesive device in RCE construction, which connects the discourse information

---

pronoun should comply with respect to identification of index.

within a subject NP with that within an extraposed restrictive relative clause.

3.2.3 Discourse-based classification of RCE

As introduced earlier, we suggested that both a subject NP and its extraposed relative clause are characterized as 'Discourse-New' or 'Discourse-Old' according to prior context. Two grammatical functions are connected by a cohesive device (i.e., an extraposed relative pronoun). 'Discourse-Old' and 'Discourse-New' are related to 'prior' background and 'no background' information respectively. In other words, a subject NP with prior mention or with no prior mention of a referent itself is combined with an extraposed relative clause identified or not identified within prior background of preceding lines. For reference, in order to determine 'Discourse-New' or 'Discourse-Old', we looked at over than at least three lines of prior discourse for each example. If there is a prior mention about a subject NP or an extraposed relative clause, we considered it as 'Discourse-Old'.

■ Type I: Discourse-Old-Old-RCE

The first type of RCE construction is that both a subject NP and an extraposed relative clause are Discourse-Old, so readers can fully trace the information that is evoked from the prior background:

(22) Opponents generally focused upon the great cost of the railroad. Proponents of the project, such as Thomas Butler King of Glynn County, argued that the initial investment of several million dollars would come back four fold in the revenue generated. King also heralded the benefits the line would have for farmers when he wrote: "**The cheap transportation of iron and salt**, **those necessary articles**. would greatly **reduce the annual expenses** of the farmer. The merchant would be enabled to purchase **his goods** on shorter credit and **the freight** would be reduced that would **lessen their costs** to the consumer. (COCA, 2007, ACAD)

As exemplified in (22), a definite subject NP *the freight* is associated with *transportation of iron and salt, those necessary articles,* and *his goods*, identified within preceding lines. The information on *lessen their costs* within an extraposed relative clause can also be fully traced from *reduce the annual expenses* described in the prior context.

■ **Type Ⅱ: Discourse-Old-New-RCE**

Another is when a subject NP is Discourse-Old, but its extraposed relative clause has the content that cannot be evoked:

(23) When these emotional strategies cannot be used, then the FEASP-approach does not make any sense in educational practice and related research. Is there a relation between the application of the **FEASP-strategies** and the experience of certain types of emotions during instruction? .. The effectiveness of **the FEASP-strategies** is an important issue for instructional designers, because in instructional practice **only those strategies** will be used that significantly help to solve practical problems. (COCA, 2001, ACAD)

A definite NP *only those strategies* in (23) contains the information described in prior background. On the other hand, an extraposed relative clause holds new information with no prior mention of how to solve practical problems.

■ **Type Ⅲ: Discourse-New-Old-RCE**

A subject NP conveys Discourse-New information and simultaneously the information within its extraposed relative is derived from prior context:

(24) It takes food much longer (a week or two, rather than half a day) to spoil, or **stops bacterial action** altogether. Once a product is thawed, however, the bacteria can become active again, multiplying under the right conditions. Enzyme activity is slowed down but not stopped during freezing. Canning: Provides a way to store foods for extremely long

periods of time. Food is first boiled to **destroy bacteria** and inactivate enzymes. It is then placed in an airtight container. As the food cools, **a vacuum seal** is formed that prevents **any new bacteria** from getting in.

(COCA, 2002, MAG)

A subject NP *a vacuum seal* in (24) is not identified within preceding texts, so it is judged as Discourse-New information. On the other hand, its restrictive relative clause contains the information on *bacteria* written in preceding lines. For reference, this study also reflects a Discourse-New subject NP with a definite determiner.

■ **Type Ⅳ: Discourse-New-New-RCE**

The last category is when both a subject NP and its extraposed relative clause do only contain Discourse-New information:

(25) It's not like Germany, it's not organized activity, " said Lennart Rohdin, undersecretary of state for immigration and refugees. " But the reasons for the incidents are familiar. People are asking why all these refugees should come here when we have a recession, when we have such high unemployment. The moral barriers are being lowered, and **the atmosphere** is being created that allows **people to commit acts of violence**. (COCA, 1992, NEWS)

As seen in (25), a definite NP *the atmosphere* is discourse-new information that is not evoked from preceding context, and even we cannot find any clue to facilitate the interpretation of the information within a postposed clause: *that allows people to commit acts of violence*.

For the frequency rate for four types of RCE construction, Discourse-New-Old-RCE (i.e., 61.31% (n=366)) stays on the top in frequency rate, and Discourse-New-New-RCE (i.e., 28.64% (n=171)) occupies the secondary portion. This result implicates that RCE construction is more appropriate to employ when a writer or speaker wants to convey discourse-new information within a subject NP, and puts more given information in an extraposed relative clause.

4. Statistical analysis for RCE

This section focuses on complex interplay among several different factors contributing to the categorization of four types of RCE construction. In other words, we aim to reveal any significance in the relation between the syntactic properties and the discourse-based classification of RCE construction with the help of statistical techniques. As witnessed above, the indefiniteness/ definiteness of a subject NP, the passive/presentative intransitive/predicative complement of copula *be*, and the subject NP-to-extraposed RC length ratio are likely to be associated with a high likelihood of RCE, so it is worthwhile to ask to what extent these variables are related to each other in order to predict the discourse-based classification of RCE construction. We assume that discourse functions are correlated with syntactic constructions in a principled way, so the relationship among syntactic variables can be used for understanding the functional relationships among various discourse-related properties of RCE construction. In doing so, we introduce multinominal logistic regression to check the determiners which influence the preference for four types of RCE structure.

4.1 Multinominal logistic regression

We adopted multinomial logistic regression to predict the occurrence of RCE classification, which tells us the relative power of each independent factor.[8] In other words, we investigate which factor among independent variables is more negatively or positively critical than others in order to predict RCE construction. To begin, we provide the significance for Discourse-Old-Old-RCE:

---

8   The independent variables included were encoded as follows: the definiteness/indefiniteness of a subject NP (definite=0, indefinite=1) and the predicate type (predicative complement of copula *be*=0, passive voice=1, presentative intransitive=2). The subject NP-to-extraposed RC length ratio was treated as covariate because it is a continuous variable. The dependent variable was also given a status: Discourse-New-New-RCE=0, Discourse-Old-Old-RCE=1, Discourse-Old-New-RCE =2, and Discourse-New-Old-RCE=3. Discourse-New-New-RCE was treated as a reference group for criterion.

Table 1. Regression for Discourse-Old-Old-RCE

Discourse-Old-Old		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	-7.294	1.855	15.466	1	.000			
Subject NP	Definite	8.754	1.622	29.122	1	.000	6337.392	263.690	152309.8

Table 1 tells us that the definiteness of a subject NP among three independent variables is found to be a significant predictor of Discourse-Old-Old-RCE at an alpha level of p <.001 (Wald=29.122***). The positive regression coefficient (8.754) of a subject NP indicates that when the definiteness of a subject NP becomes more distinctive, the discourse status within the subject and extraposed relative clause tends to be older. This implies that the regression coefficient increases the likelihood of Discourse-Old-Old-RCE construction. In addition, Exp(B) value (6337.392) tells us that a definite subject NP overwhelmingly predicts Discourse-Old-Old-RCE much better than an indefinite subject NP. In conclusion, a powerful tool to anticipate Discourse-Old-Old-RCE is the definiteness of a subject NP.

Discourse-Old-New-RCE construction also has similar characteristics with Discourse-Old-Old-RCE. The definiteness of a subject NP (Wald=49.350***, p<.001) is a significant predictor of Discourse-Old-New-RCE; p values of other independent variables are higher than 0.05:

Table 2. Regression for Discourse-Old-New-RCE

Discourse-Old-New		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	-4.939	1.151	18.405	1	.000			
Subject NP	Definite	6.647	.946	49.350	1	.000	770.796	120.637	4924.885

The positive regression coefficient (6.647) of a subject NP in Table 2 indicates the probability that Discourse-Old-New-RCE occurs. Exp(B) value (770.796) indicates that the definiteness of a subject NP overwhelmingly explicates Discourse-Old-New-RCE 771 more times than the indefiniteness of a subject NP. Consequently, a predictive power of the occurrence of Discourse-Old-Old-RCE is the definiteness of a subject NP. Interestingly, we can notice another interesting point as follows:[9]

Table 3. Regression for Discourse-Old-Old-RCE

Discourse-Old-Old		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	-8.202	1.854	19.569	1	.000			
Length Ratio	Length Ratio	3.209	1.058	9.192	1	.002	24.758	3.110	197.111
Subject NP	Definite	9.324	1.606	33.705		.000	11208.4	481.318	261007.5

Table 4. Regression for Discourse-Old-New-RCE

Discourse-Old-New		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	-5.847	1.150	25.852	1	.000			
Length Ratio	Length Ratio	2.825	.868	10.595	1	.001	16.863	3.077	92.411
Subject NP	Definite	7.218	.909	63.003	1	.000	1363.24	229.383	8101.788

As shown in Table 3 and 4, Discourse-Old-Old-RCE and Discourse-Old-New-RCE have positive regression coefficients in length ratio at the significant value respectively: 3.209 and 2.825. This implies that both heavy information within a subject NP and the decreasing number of constituents within an extraposed restrictive relative clause contribute to determining the occurrence of these two types of RCE construction.

The information in the following Table 5 also conveys a meaningful result:

Table 5. Regression for Discourse-New-Old-RCE

Discourse-New-Old		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	8.202	1.854	19.569	1	.000			
Length Ratio	Length Ratio	-3.209	1.058	9.192	1	.002	.040	.005	.322
Subject NP	Definite	-9.324	1.606	33.705	1	.000	8.922E-005	3.831E-006	.002

The subject NP-to-extraposed RC length ratio of three independent variables is found to be a significant predictor of Discourse-New-Old-RCE construction (Wald=9.192***, p<.05). The negative regression coefficient (-3.209) of the length ratio tells us that the more the length ratio increases, the less Discourse-New-Old-RCE tends to occur (i.e., negative length ratio is more desirable). In other words, a larger number of constituents within an extraposed relative clause is

---

9   We set Discourse-New-Old-RCE as a reference group.

strongly correlated with Discourse-New-Old-RCE type, but heavy information within a subject NP cannot guarantee Discourse-New-Old-RCE. Therefore, the relatively powerful factor to predict Discourse-New-Old-RCE is the large number of words within an extraposed relative clause (i.e., negative subject NP-to-extraposed RC length ratio). This finding says in effect that the grammatical heaviness within an extraposed relative is a significant and plausible trigger to bring about the extraposition of a constituent; it is because Discourse-New-Old RCE takes up the most in this corpus study. In addition, a negative regression coefficient (-9.324***) for the definiteness of a subject NP implies that the indefiniteness of a subject NP positively predicts Discourse-New-Old-RCE.

For the analysis of Discourse-New-New-RCE, we set Discourse-Old-Old-RCE as a reference category:

Table 6. Regression for Discourse-New-New-RCE

Discourse-New-New		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
	Intercept	7.294	1.855	15.466	1	.000			
Length Ratio	Length Ratio	1.246	.347	12.873	1	.001	3.478	1.760	6.870
Subject NP	Definite	-8.754	1.622	29.122	1	.000	.000	6.566E-006	0.004

Table 6 tells us that the definiteness of a subject NP is found to be a powerful predictor of Discourse-New-New-RCE (Wald=29.122***, p<.001). The interesting point here is that a negative regression coefficient of a subject NP is the same as the positive regression coefficient (8.754) in Table 1 under the absolute value. This negative coefficient explicates that the definiteness of a subject NP decreases the possibility that Discourse-New-New-RCE occurs. In other words, the indefiniteness of a subject NP functions as a powerful tool to predict Discourse-New-New-RCE. In addition, a positive regression coefficient (1.246) implies that high value of the length ratio plays a role in causing Discourse-New-New-RCE to exist.

We also focus on the interplay between independent variables to predict four types:

Table 7. Interaction between two independent variables for RCE construction

Discourse-New-New		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
Discourse-Old-Old	Definite* Passive	6.889	2.730	6.367	1	.012	981.571	4.656	206943.546
Discourse-Old-New	Definite* Passive	5.423	2.354	5.307	1	.021	226.584	2.247	22853.252
Discourse-New-Old	Definite* Passive	-5.423	2.354	5.307	1	.021	.004	4.376E-005	.445
Discourse-New-Old	Length Ratio* Passive	-2.050	1.008	4.140	1	.042	.129	0.018	.927
Discourse-New-New	Length Ratio* Passive	2.050	1.008	4.140	1	.042	7.771	1.078	56.005

Positive regression coefficients (6.889 & 5.423) of the interplay between the definiteness of a subject NP and passive voice imply that the two independent variables collaboratively predict Discourse-Old-Old-RCE and Discourse-Old-New-RCE. Interestingly, Discourse-New-Old-RCE had negative regression coefficients about the definiteness of a subject NP and the length ratio in Table 5, so subsequently the negative values also influence the interplay between the passive voice and the definiteness of a subject NP as well as that between the passive voice and the length ratio (i.e., coefficient -5.423 and -2.050 respectively). The passive voice positively influences the occurrence of Discourse-New-Old-RCE. In addition, the negative value about length ratio still requires an extraposed relative clause to load more information in order to influence the occurrence of Discourse-New-Old-RCE. Finally, Discourse-New-New-RCE is positively predicted by the interaction between the positive length ratio and passive voice. Last, one thing to notice in Table 7 is that only passive voice has a mutual relationship with other independent variables except the other predicates: presentative intransitive and predicative complement of copula *be*. This finding implies that it is reasonable to employ more passive voice in order to predict the occurrence of RCE construction.

We also embark on the interaction between three independent variables in order to predict each type of RCE construction as follows:

Table 8. Interaction between three independent variables for RCE construction

Discourse-New-New		B.	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval about Exp(B)	
Discourse-Old-Old	Definite*Passive *Length Ratio	8.062	2.683	9.027	1	.003	3171.943	16.49	610125.65
Discourse-Old-New	Definite*Passive *Length Ratio	9.908	2.488	15.861	1	.000	20084.65	153.2	2633164.67
Discourse-New-Old	Definite*Passive *Length Ratio	-12.577	4.921	6.531	1	.011	3.451E-006	2.234E-010	.053
Discourse-New-Old	Indefinite*Passive *Length Ratio	-1.196	.364	10.781	1	.001	.302	.148	.617
Discourse-New-New	Indefinite*Passive *Length Ratio	1.196	.364	10.781	1	.001	3.308	1.620	6.758

Positive regression coefficients (8.062 & 9.908) of the interplay between the definiteness of a subject NP, passive voice, and length ratio tell us that three independent variables are combined to positively predict Discourse-Old-Old-RCE and Discourse-Old-New-RCE. At this point, we conclude that the definiteness of a subject NP, the passive voice, and the positive length ratio are crucial factors to predict RCE construction especially when a subject NP contains discourse-old information. On the other hand, we can still witness that the interaction of Discourse-New-Old-RCE is described as negative regression coefficients (i.e., -12.577 and -1.196). Unexpectedly, the former value is derived from the interplay among the definiteness of a subject NP (i.e., negative regression coefficient), the passive voice, and the low value of the length ratio (i.e., negative regression coefficient). In earlier time, we concluded that the passive voice positively predicts Discourse-New-Old-RCE. However, at this point, when three independent variables are combined, the passive voice seems to affect negatively the occurrence of Discourse-New-Old-RCE, thus resulting in negative regression coefficient (i.e., -12.577). On the other hand, the latter value (i.e., -1.196) results from the fact that only the high value of the length ratio negatively influences the prediction of Discourse-New-Old-RCE, but the indefiniteness of a subject NP and the passive voice play a positive role in the occurrence of Discourse-New-Old-RCE. In other words, the passive voice has both negative and positive influence on Discourse-New-Old-RCE when other two independent variables are added. Once again, we come to the conclusion that the short length of a subject

NP and the increasing number of constituents within an extraposed relative clause are generally crucial factors to trigger RCE construction; Discourse-New-Old-RCE is the most frequently occurring type. Last, Discourse-New-New-RCE is positively predicted by the interaction between indefiniteness of a subject NP, passive voice, and the high value of the length ratio.

5. Conclusion

This paper has identified the corpus findings of 597 RCE instances and then discussed its grammatical properties such as the indefiniteness/ definiteness of a subject NP, the predicate type (i.e., passive, presentative intransitive, and predicative complement of copula *be*), and/or the grammatical weight (i.e., subject NP-to-extraposed RC length ratio). An extraposed relative pronoun was assumed to function as a cohesive device, thereby enabling RCE construction to be classified into four types according to discourse manner: Discourse-Old-Old-RCE, Discourse-Old-New-RCE, Discourse-New-Old-RCE, and Discourse-New-New-RCE.

We introduced multinominal logistic regression to determine whether the strong inclination for each type of RCE construction is primarily predicted by the indefiniteness/definiteness of a subject NP, the predicate type, and/or the grammatical weight. As witnessed before, the definiteness of a subject NP and the positive value of the length ratio are crucial factors to bring about Discourse-Old-Old-RCE and Discourse-Old-New-RCE, whereas the indefiniteness of a subject NP is a main predictor of Discourse-New-Old-RCE and Discourse-New-New-RCE. This result implicates that Discourse-Old-Old-RCE and Discourse-Old-New-RCE are more suitable when a writer or speaker wants to convey discourse-old information within a subject NP, whereas an indefinite subject NP is useful for Discourse-New-Old-RCE and Discourse-New-New-RCE when putting new or focused information in a subject position. Interestingly, only Discourse-New-Old-RCE taking up the largest portion in this corpus data has a negative correlation with subject NP-to-extraposed RC length ratio. This means that a small number of constituents within a subject position and longer information within an extraposed relative clause are preferred for Discourse-

New-Old-RCE (i.e., negative length ratio). This meaningful statement leads us to a plausible conclusion that the heavy grammatical weight within a restrictive relative clause becomes a decisive factor to trigger the extraposition. Furthermore, we observed that the definiteness of a subject NP, passive voice, and the positive value of the length ratio predict Discourse-Old-Old-RCE and Discourse-Old-New-RCE. On the other hand, the definiteness of a subject NP and the high value of the length ratio have a negative correlation with Discourse-New-Old-RCE. The passive voice predicts negatively or positively Discourse-New-Old-RCE when three independent variables are combined. Last, Discourse-New-New-RCE is predicted by the interplay among the indefiniteness of a subject NP, the passive voice, and the positive value of the length ratio.

References

Al Khalaf, Eman. 2018. Remarks on the syntax and semantics of so-called comitative coordination. *Linguistic Research* 35(2): 253-273.

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1): 28-55.

Büring, Daniel. 2007. Semantics, intonation and information structure. In Gillian Ramchand and Charles Reiss (eds.), *The Oxford handbook of linguistic interfaces*, 445-474. Oxford: Oxford University Press.

Celce-Murcia, Marianne and Elite Olshtain. 2000. *Discourse and context in language teaching*. New York, NY: Cambridge University Press.

Dooley, Robert A. and Stephen H. Levinsohn. 2000. *Analyzing discourse: A manual of basic concepts*. Dallas: SIL International.

Francis, Elaine J. 2010. Grammatical weight and relative clause extraposition in English. *Cognitive Linguistics* 21(1): 35-74.

Francis, Elaine J. and Laura A. Michaelis. 2014. Why move? How weight and discourse factors combine to predict relative clause extraposition in English. In Edith Moravcsik and Andrej Malchukov (eds.), *Competing motivation in grammar and usage*, 70-87. Oxford: Oxford University Press.

Francis, Elaine J. and Laura A. Michaelis. 2017. When relative clause extraposition is the right choice, it's easier. *Language and Cognition* 9(2): 332-370.

Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan. 1976. *Cohesion in English*.

London: Longman.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. New York, NY: Oxford University Press.

Huck, Geoffrey J. and Younghee Na. 1990. Extraposition and focus. *Linguistic Society of America* 66: 51-77

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. New York, NY: Cambridge University Press.

Kim, Jungsoo. 2018. Acquisition of verb representations: *Get*'s gotta get us somewhere. *Linguistic Research* 35(Special Edition): 207-232.

Kim, Jong-Bok and Peter Sells. 2008. *English syntax: An introduction*. Stanford: CSLI Publications.

Kim, Jong-Bok. 2017. Mixed properties and matching effects in English free relatives: A construction-based perspective. *Linguistic Research* 34(3): 361-385.

Kiss, Tibor. 2005. Semantic constraints on relative clause extraposition. *Natural Language and Linguistic Theory* 23: 281-334.

Lee, Seung Han. 2017a. *Factors at play in extraposing English relative clauses*. PhD Dissertation. Chonnam National University.

Lee, Seung Han. 2017b. English relative clause extraposition: A pedagogical approach. *Linguistic Research* 34(Special Edition): 49-68.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English literature*. London: Longman.

Rasekh-Mahand, Mohammad, Mojtaba Alizadeh-Sahraie, and Raheleh Izadifar. 2016. A corpus-based analysis of relative clause extraposition in Persian. *Ampersand* 3: 21-31.

Ross, John Robert. 1967. *Constraints on variables in syntax*. PhD Dissertation. MIT.

Song, Bu-Seon. 2009. A discourse-pragmatic explanation for extraposed relative clauses in English. *Journal of English Language and Literature* 53(4): 303-315.

Strunk, Jan. 2004. Relative clause extraposition in Low Saxon. Ms. Stanford University.

Ward, Gregory and Betty J. Birner. 2001. Discourse and information structure. In Deborah Schiffrin, Deborah Tennen, and Heidi E. Hamilton(eds.), *Handbook of discourse analysis*, 119-136. Oxford: Blackwell.

Wasow, Thomas. 2002. *Postverbal behavior*. Stanford: CSLI Publications.

Zhang, Guiping. 2015. It is suggested that⋯ or it is better to⋯? Forms and meanings of subject it-extraposition in academic and popular writing. *Journal of English for Academic Purposes* 20: 1-13.

Seung Han Lee
Assistant Professor
Dept. of Culture Teaching

Chodang University
380 Muanro, Muaneup, Muangun
Jeollanamdo 58530, Rep. of Korea
E-mail: veritas0889@cdu.ac.kr