Linguistic Research 39(3): 467-497
DOI: 10.17250/khisli.39.3.202212.003

Honorific agreement and plural copying revisited: Experimental and deep learning approach*

Yong-hun Lee · Ji-Hye Kim*****
(Chungnam National University · Korea National University of Education)

Lee, Yong-hun and Ji-Hye Kim. 2022. Honorific agreement and plural copying revisited: Experimental and deep learning approach. *Linguistic Research* 39(3): 467-497. The study investigated two types of subjecthood diagnostics in Korean using two different kinds of approaches. Based on the previous studies on subjecthood diagnostics in Korean, this paper examined the two subjecthood diagnostics: Honorific Agreement (HA) and Plural Copying (PC). For the experimental analysis, this study adopted the analysis results of Kim et al. (2017). For the deep learning analysis, this paper employed the KR-BERT for the deep learning model and three sources of data sets: the Korean version of the Corpus of Linguistic Acceptability (K-CoLA), the Sejong Morphologically-Analyzed Corpus, and the extended target sentences of Kim et al. (2017). Two separate experiments were conducted in the deep learning analysis. In the first experiment, the KR-BERT was trained only with the K-CoLA, and the target sentences were analyzed. In the second experiment, the KR-BERT was trained with the K-CoLA and the sentences from the Sejong corpus, and the acceptability scores of the target sentences were measured. The acceptability scores were measured with the numeric scores using the algorithms in Lee (2021). After the experiments with the deep learning models, the scores were normalized and were statistically analyzed with Generalized Linear Models (GLMs). Through the two experiments, the following fact was observed: both HA and PC did not show similar tendencies of experimental results with human participants in the first experiment, but they did in the second experiment. The analysis results demonstrated that both HA and PC could be used as subjecthood diagnostics but that they played significant roles only when native speakers were exposed to enough examples. **(Chungnam National University · Korea National University of Education)**

Keywords honorific agreement, plural copying, deep learning, KR-BERT, K-CoLA

* We wish to thank two anonymous reviewers of this journal for their helpful comments and suggestions. All remaining errors, however, are ours.
** First author
*** Corresponding author

1. Introduction

Many previous studies have investigated subjecthood diagnostics in Korean (Yoon 1986; Youn 1990; Hong 1991; Kang 2002; Yoon 2009; Hong 2014; Yoon 2015). Crosslinguistically, constructions with shared subject properties (like Case-marking or agreement) do not converge on only a single nominal NP (as in the Dative Subject Constructions), and they have been crucial in the study of the subject properties in many languages. Double/Multiple Subject Constructions (DSCs/MSCs), where more than one Nominative-marked NP is represented, are also present in Korean, and have been the focus of the inquiry into the subject properties. A few typical examples of Korean MSCs are illustrated in (1).

(1) a. *Cheli-ka* *Yenghi-lul* *salanghan-ta.*
Cheli-NOM Yenghi-ACC love-PRST.DECL
'Cheli loves Yenghi.'
b. *Yenghi-lul* *Cheli-ka* *salanghan-ta.*
Yenghi-ACC Cheli-NOM love-PRST-DECL
'Cheli loves Yenghi.'
c. *Cheli-ka* *kho-ka* *khu-ta.*
Cheli-NOM nose-NOM big-COPULA-DECL
'It is Cheli whose nose is big (= Cheli's nose is big).'
d. *Cheli-ka* *kho-ka* *alay-pwupwun-i* *khu-ta.*
Cheli-NOM nose-NOM bottom-part-NOM big-DECL
'The bottom part of Cheli's nose is big.'

Regardless of its position within the sentence, the NP *Cheli* in (1a) and (1b), which has the Nominative Case (NOM) marker '*-ka*,' is the Grammatical Subject. However, which NP serves as a subject has been debated in the literature on MSCs that have many NOM-marked NPs in a single sentence, such as (1c) or (1d).

If more than one subject position exists in MSCs, several scholars have come to different conclusions (Park 1973; Yoon 1986; Yoon 1989; Hong 1991; Park 1995; Lee 1997; Schütze 2001; Kang 2002; Yoon 2009, 2015; etc.). According to a recent claim supported by Yoon (2009, 2015), there are multiple subject positions in MSCs; and the leftmost NOM-marked NP functions as a Grammatical Subject (GS), whereas the

outer NOM-marked NPs are Major Subjects (MS) which are in combined with a Sentential Predicate (SP). (Lee 1985; Heycock 1993; etc.)

Following is a summary of the subject diagnostics that have been suggested so far. (Yoon 1986; Youn 1990; Hong 1991)

(2) Proposed Subject Diagnostics in Korean
   a. Nominative Case-marking
   b. Controller of optional plural-marking (i.e., Plural Copying)
   c. Controller of subject honorification (i.e., Honorific Agreement)
   d. Target of Subject-to-Object raising
   e. Target of Control
   f. Controller of PRO in complement (obligatory) control
   g. Controller of PRO in adjunct control
   h. Controller of coordinated deletion
   i. Antecedent of (subject-oriented) anaphors
   j. Exhaustive-listing interpretation of '*-ka*/*-i*'

Yoon (2009) says that the debate on whether all of these characteristics in (2) or just some of them diagnose subjecthood in Korean is misguided because earlier studies failed to distinguish between two different kinds of subjects (subject as prominent argument vs. subject as pivot, in the sense of Falk 2006). Even though they are all subject diagnoses, some of them choose the GS as a subject whereas others are controlled by the MS. The diagnostics of GS and MS frequently converge on a single nominal in single subject constructions (SSCs or Non-MSCs), but in MSCs (and a few other constructions), the two kinds of subjects (GS or MS) may diverge. Yoon (2008, 2009) specifically suggested that in MSCs, (2d), (2g), and (2h) lean toward MS while (2b), (2c), (2f), and (2i) are characteristics of GS. (Youn 1990, Hong 1991, etc.).

However, it is necessary to confirm that the subjecthood diagnostics suggested so far are empirically validated before the issues of the subjecthood of the NOM-marked NPs are addressed in MSCs. Since most diagnostics are developed based on just the intuitions of linguists who are Korean native speakers, their validity cannot be taken for granted without any verification. Recently, some experimental studies on a few of the proposed subjecthood diagnostics ((2b), (2c), (2f), (2g), (2h), and (2i)) have shown that these diagnostics need to be re-evaluated (Kim et al. 2015; Lee et al. 2015a; Kim et al. 2017).

This study reports an experimental investigation using the deep learning models for the following two subjecthood diagnostics: Honorific Agreement (HA) and Plural Copying (PC). This study utilized the analysis results of Kim et al. (2017) for the experiments with human participants. For the deep learning experiments, this paper adopted the KR-BERT for the deep learning model and three sources of data sets: the Korean version of the Corpus of Linguistic Acceptability (K-CoLA), the Sejong Morphologically-Analyzed Corpus, and the expanded target sentences in Kim et al. (2017). In the deep learning analysis, two different experiments were carried out. In the first experiment, the target sentences were examined after the KR-BERT model had been trained only with the K-CoLA dataset. In the second experiment, however, the acceptability of the target sentences was measured after the KR-BERT had been trained using the K-CoLA and the sentences which were extracted from the Sejong Morphologically-Analyzed corpus. The algorithm in Lee (2021) was used to measure the acceptability scores of the input sentences. Upon completion of the deep learning experiments, all the acceptability scores were collected for the target sentences, and they were normalized using the *z*-scores and statistically analyzed with Generalized Linear Models (GLMs).

This paper is constructed as follows. Section 2 is an introduction to previous studies on subjecthood diagnostics as well as an experimental approach to this topic which was focused on HA and PC. Section 3 tells about the research method. This section provides explanations of the data sets, the deep learning models, and the algorithm of how the acceptability scores are measured in the model. Section 4 enumerates the analysis results of the experiments with deep learning models. Section 5 is for the discussions, where the differences between the two kinds of approaches are discussed in a more detailed way. Section 6 is a summary of this paper.

2. Previous studies

2.1 Honorific agreement as a subjecthood diagnostic test

Hong (1991, 1994) mentioned that the grammatical subject might be clearly distinguished using the '*-si*' marking in simple sentences (i.e., Non-MSC sentences). According to Hong (1994), there are several ways for Korean speakers to convey *respect*

in grammar, one of which is by adding the verbal suffix '*-si*' to a verb stem, as in (3) (Hong, 1994).

(3) a. *Halapeci-ka cikum o-si-n-ta.*
Grandfather-NOM now come-HON-PRES-DECL
'Grandfather comes now.'
b. **Minswu-ka sensayngnim-ul manna-si-ess-ta.*
Minawu-NOM teacher-ACC meet-HON-PAST-DECL
'Minswu met the teacher.'
c. **Sensayngnim-un Minswu-ka manna-si-ess-ta.*
Teacher-TOP Minswu-NOM meet-HON-PAST-DECL
'As for the teacher, Minswu met him.'
d. *Sensayngnim-uy malssum-i olhu-si-ta.*
Teacher-GEN speech-NOM be.correct-HON-DECL
'The teacher's speech is correct.'
e. **Sensayngnim-uy ankyeng-i kum-i-si-ta.*
Teacher-GEN glasses-NOM gold-be-HON-DECL
'The teacher's glasses are made of gold.'

When the subject NP *halapeci* 'grandfather' agrees with '*-si*', the sentence becomes grammatical as in (3a). However, the non-honorific subject NP cannot get the honorific marking as in (3b). Furthermore, (3c) shows that non-subject NP (i.e., object) cannot trigger '*-si*' marking even when topicalized. Finally, the subject NP that is in a metonymic relationship (cf. 3d) with the honorified entity can receive the honorific '*-si*', but it cannot be in a non-metonymic relationship with an inanimate object (cf. 3e).

While some studies (Suh 1977; Yoon 1987) asserted that '*-si*' marking might be pragmatically rather than grammatically motivated, Hong (1994) contended that '*-si*' marking could be stimulated by the grammatical subject, in agreement with other researchers (Han 1990; Lee 1990; Youn 1990), although semantic and/or pragmatic factors also played a significant role.

Yoon (2008) asserted that, in MSC settings, GS rather than MS was in charge of subject honorification. According to Yoon (2008, 2009), GS was seen as a competent controller of honorification in MSC (cf. 4a); however, the sentence was unacceptable when MS attempted to regulate honorification when GS was present (cf. 4b).

(4) a. *Kim-kyoswunim-i*$_{i}$ (MS) *haksayngtul-i* (GS) *e*$_{i}$
K-professor.HON-NOM students-NOM
yocum manhi chacaka-n-ta.
now a.lot visit-PRES-DECL
'It is Professor Kim who students visit a lot these days.'
b. **Kim-kyoswunim-kkeyse*$_{i}$ (MS) *haksayngtul-i* (GS) *e*$_{i}$
K-professor-HON-NOM students-NOM
yocum manhi chacaka-si-n-ta.
now a.lot visit-HON-PRES-DECL
'It is Professor Kim who students visit a lot these days.'

Additionally, Yoon (2008) identified a few instances when the GS did not appear to always have control over the honorific as demonstrated in (5a) and (5b).

(5) a. **Kim-sensayngnim-i*$_{i}$ *haksayngtul-eykey-n*$_{j}$ [*PRO*$_{j}$
K-teacher-HON-NOM students-DAT-TOP
e$_{i}$ *manna-ki*]-*ka* *himtu-si-ta*.
meet-NMZ-NOM difficult-HON-DECL
'Professor Kim is difficult for the students to meet.'
b. *?*Kim-sensayngnim-i*$_{i}$ *chwucongcatul-eykey-n*$_{j}$ [*e*$_{i}$
K-teacher-HON-NOM followers-by-TOP
chencay-lako] *sayngkaktoy-si-nun kes kathta*.
genius-COMP be.thought-HON-PRES seem.DECL
'Professor Kim must seem like a genius to his followers.'

He also mentioned that the Experiencer DP might be the GS if one argued that the NOM-marked NP (*Kim sensayngnim* 'Professor Kim' in (5a) and (5b)) was MS. This could be another example demonstrating that MS does not control honorification in the presence of GS (Yoon 2004).

2.2 Plural copying as a subjecthood diagnostic test

According to Song (1975) and Kuh (1987), when the subject represented a plural

entity, the plural suffix '*-tul*' might be appended to the other elements of a sentence, such as an object, the infinitive form of a verb, an oblique argument (cf. 6a), or even adverbials (cf. 6b).

(6) a. *Nehi-tul ku chayk-tul ilke-tul po-ass-ni?*
You-PL the.book-PL read-PL try-PAST-INT
'Have you (PL) tried to read the book?'
b. *Haksayng-tul-i hakkyo-lo-tul kuphi-tul ka-n-ta.*
Student-PL-NOM school-to-PL hurriedly-PL go-PRES-DECL
'The students go to school hurriedly.'

Plural Copying (PC) denotes the plurality of events that the predicate denotes. For instance, in (6b), the following circumstance should exist: The students who arrive at school in a hurry are student A, student B, student C, and so on.

The arrangement of suffixes in PC differs from the standard plural marking, according to Song (1975) and Kuh (1987). For instance, the derivational suffix '*-tul*' appears before all the other suffixes if the noun is plural, as in (7a). However, the plural suffix '*-tul*' follows semantic Case markers when the noun is singular and has the plural '*-tul*' pasted onto it (cf. 7b).

(7) a. *hakkyo-tul-ey*
school-PL-to
'to the schools'
b. *hakkyo-ey-tul*
school-to-PL
'to the school'

Later, Hong (1994) found that even when non-subject arguments were treated as the subject of the PREDICATION, plural '*-tul*' marker copying might still take place (Williams 1980). According to Hong (1994), PC was a diagnostic that could pick out both the topic and another thing in addition to the subject.

As shown in the following examples, Yoon (2008) countered that, in MSCs, GS but not MS could govern PC on adverbs.

(8) a. *Ku tayhak-i* (MS) *kyoswutul-i kongpwu-lul*
that university-NOM professors-NOM study-ACC
yelsimhi-tul ha-n-ta.
diligently-PL do-PRES-DECL
'As for that university, the professors diligently engage in research.'
b. **Cheli-wa Yenghi-ka* (MS) *kathi sa-nun*
Cheli-CONJ Yenghi-NOM together live-REL
cip-i acwu-tul nalk-ass-ta.
house-NOM very-PL run.down-PERF-DECL
'As for Cheli and Yenghi, the house they live in together is really rundown.'

As shown in (8), the MS *Cheli-wa Yenghi* in (8b), which signifies plural entities, cannot be deemed an appropriate controller because the local subject (GS) is single; while the GS *kyoswutul* 'faculty members' in (8a), which denotes numerous entities, can be a controller of PC.

2.3 Experimental studies in Kim et al. (2017)

Based on the theoretical discussions on HA and PC as subjecthood diagnostics, Kim et al. (2017) conducted the experiments using the method in experimental syntax (Bard et al. 1996; Schütze 1996; Cowart 1997; Keller 2000; and so on). The test materials consisted of 80 sentences in total, 40 of which tested the HA and 40 of which assessed the PC. Twenty filler sentences were constructed in addition to the 20 target sentences (4 type conditions × 5 tokens) in each diagnostic test. The examined subject diagnostic features were not included during the process of creating the fillers which were built utilizing the structure of the target sentences. For the HA diagnostic, for instance, we constructed non-target sentences lacking the '*-si*' morpheme so that participants might assess the acceptability of the sentences without taking honorific issues into account. For the purpose of creating the fillers for the PC, the plural morpheme '*-tul*' was also dropped.

The target sentences were divided into four different patterns, which were created by combining two sentence types (i.e., Non-MSC vs. MSC) and two types of NPs that met the required property: NP1 (i.e., possessor in Non-MSC/MS in MSC) vs. NP2 (i.e., GS

in both conditions). All the sentence patterns were displayed below.

(9) Design of the Target Sentences
a. Type 1: $[NP1]_{gen}$ $[NP2]_{nom}$ [Non-MSC, agreement with NP2 (subject)]
b. Type 2: $[NP1]_{gen}$ $[NP2]_{nom}$ [Non-MSC, agreement with NP1 (possessor)]
c. Type 3: $[NP1]_{nom}$ $[NP2]_{nom}$ [MSC, agreement with NP2 (GS)]
d. Type 4: $[NP1]_{nom}$ $[NP2]_{nom}$ [MSC, agreement with NP1 (MS)]

Based on these patterns, all of the target sentences were constructed. (10) and (11) demonstrate examples of the target sentences in HA and PC respectively.

(10) a. *Cheli-uy apeci-ka pwuca-i-si-ta.*
Cheli-GEN father-NOM rich-COP-HON-DECL
'Cheli's father is rich.'
b. **Kimkyoswu-nim-uy cengwon-i alumtawu-si-ta.*
Professor.Kim-GEN garden-NOM beautiful-HON-DECL
'Professor Kim's garden is beautiful.'
c. *Cheli-ka apeci-ka pwuca-i-si-ta.*
Cheli-NOM father-NOM rich-COP-HON-DECL
'It is Cheli whose father is rich.'
d. **Kimkyoswu-nim-i cengwon-i alumtawu-si-ta.*
Professor.Kim-NOM garden-NOM beautiful-HON-DECL
'It is Professor Kim's garden that is beautiful.'

(11) a. *Chicago-uy kenmwul-tul-i acwu-tul nop-ta.*
Chicago-GEN building-PL-NOM very-PL high-DECL
'Chicago's buildings are high.'
b. **Namhan-kwa pwukhan-uy kyengkyey-ka*
S.Korea-and N.Korea-GEN boundary-NOM
maywu-tulsakmakha-ta.
very-PL desolate-DECL
'The boundary between South and North Korea is very desolate.'
c. *Chicago-ka kenmwul-tul-i acwu-tul nop-ta.*
Chicago-NOM building-PL-NOM very-PL high-DECL
'Chicago's buildings are high.'

d. **Namhan-kwa pwukhan-ka kyengkyey-ka*
S.Korea-and N.Korea-NOM boundary-NOM
maywu-tul sakmakha-ta.
very-PL desolate--DECL
'The boundary between South and North Korea is very desolate.'

These target sentences were mixed with the same number of filler sentences, and all of the sentences were presented to the participants. In the experiments, a total of 70 Korean native speakers participated. In the main task, they were instructed to draw a line for each given sentence, depending on the degree of acceptability/naturalness of the sentence.

The degree of acceptability in each of the four HA circumstances was shown in the following plot. In this plot, *Multiple* denoted the sentence type (i.e., Non-MSC vs. MSC), and *Factor* indicated the agreement (i.e., *Presence*: the honorific marker '*-si*' agrees with NP2; *Absence*: the honorific marker '*-si*' agrees with NP1). 95% confidence intervals (CIs) were depicted by the I-shaped lines in the bars.

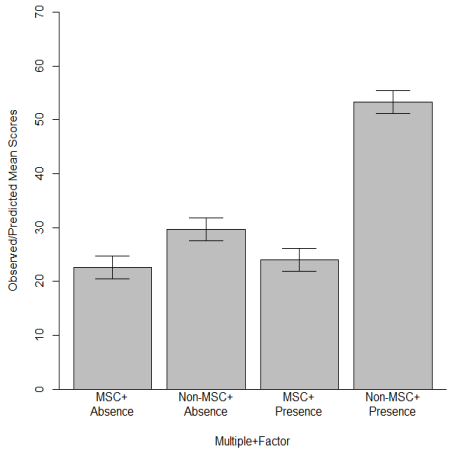

Figure 1. Bar plot (Experimental, HA)

As you could see from this plot, the combination of 'Non-MSC+Presence (agreement with NP2)' had noticeably higher acceptability than the other three combinations.

The results of the Generalized Linear Model (GLM) analysis were shown in the table below.[1] Here, the terms *Multiple1* and *Factor1* corresponded to the agreement pattern

(Non-MSC vs. MSC) and the sentence types (NP1 vs. NP2) respectively.

Table 1. Analysis results of GLM (Experimental, HA)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	32.4333	0.5414	59.91	.000	***
Multiple1	-9.0976	0.5414	-16.80	.000	***
Factor1	-6.2738	0.5414	-11.59	.000	***
Multiple1:Factor1	5.5809	0.5414	10.31	.000	***

The acceptability scores of the sentences were strongly influenced by both variables (*Multiple1* and *Factor1*), as illustrated in this table (p<.001). The p-value of the interaction (*Multiple1*:*Factor1*) suggested that there was also an interaction between the two factors (p<.001).

The level of acceptability scores for each of the four sentence patterns in the PC data set were shown in the following plot. Similar to the HA analysis, the terms *Presence* and *Absence* in *Factor1* referred to whether or not the plural marker '*-tul*' accorded with NP2 and NP1, respectively.

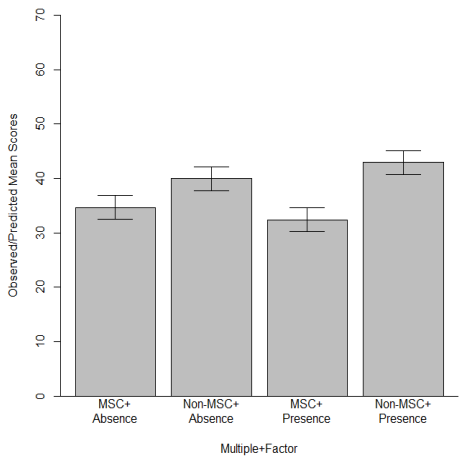

Figure 2. Bar plot (Experimental, PC)

1 As mentioned in Kim et al. (2017), the distributions of acceptability scores did not follow the normal distributions. That's why the GLM analysis was adopted here, following the criteria in Gries (2013) and Lee (2016).

The two Non-MSC conditions were more acceptable than the two MSC conditions, as shown by this plot. Also notice that the CIs in Non-MSCs might overlap and that MSCs exhibited the same tendency.

A GLM analysis was conducted to look at how the explanatory factors affected the acceptability scores of sentences. The results were shown in the following table. The agreement variable was *Factor1* (NP1 vs. NP2), while *Multiple1* was the sentence type variable (Non-MSC vs. MSC).

Table 2. Analysis results of GLM (Experimental, PC)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	37.4977	0.5616	66.770	.000	***
Multiple1	-3.953	0.5616	-7.039	.000	***
Factor1	-0.181	0.5616	-0.323	.747	
Multiple1:Factor1	1.323	0.5616	2.356	.019	*

Table 2 showed that, whereas agreement type had no effect on sentence acceptability (p=.747), sentence type did (p<.001). There was also a significant interaction between the two parameters (p<.05).

3. Research method

3.1 Dataset of the syntactic experiments with human participants

This paper used the same data sets of Kim et al. (2017) for the experimental results with human participants.[2] As mentioned in Section 2.3, a total of 80 sentences were constructed (4 type conditions × 5 tokens × target/ filler × HA/PC). Among them, half of them (40 sentences) were target sentences. 20 sentences were for HA, and the other 20 sentences were for PC. A total of 70 native speakers participated in the experiments,

2 One of the reviewers pointed out there were some more recent studies on honorification in Korean, such as Choe (2004), Kim and Sells (2007), Song et al. (2019), Kim et al. (2022), and so on. Of course, we identified these studies. However, note that the primary goal of this study is to compare the analysis results in the experiments with human participants and those of deep learning models. Even though the above studies included more updated discussions, it was impossible or very difficult to obtain the raw data in these studies. That is why we chose the data sets in Kim et al. (2017).

and the analysis results were enumerated in Section 2.3.

3.2 Dataset of the deep learning model

This paper used the extended data sets of Kim et al. (2017). The number of original target sentences in Kim et al. (2017) included 40 sentences (20 sentences for each diagnostic). They were extended to 400 sentences (200 sentences for each diagnostic) by the change of nouns in the examples. For instance, *Cheli* and *apeci* 'father' in (10) could be changed into other nouns such as *Younghee* and *e.me.ni* 'mother'. Likewise, *Chicago* and *kenmwul* 'building' in (11) could be changed into *New York* and *mulka* 'price' respectively. Then, the Case markers were adjusted to the NPs (*New York-i* not **New York-ka*). Then, the filler sentences were constructed with the K-CoLA data set. Here, the K-CoLA data set refers to the Corpus of Grammaticality Judgment which is distributed by the National Institute of the Korean Language.[3,4] Because it is common that the number of fillers to be three or five times the target sentences, a total of 2,000 sentences (400 target sentences × 5 times) were randomly selected from the K-CoLA database.[5] After a total of 2,400 sentences (400 targets and 2,000 fillers) were prepared, they were randomized and used as input sentences to the deep learning models (the KR-BERT models). The filler sentences were used for the evaluation of the model in Section 3.6.

3.3 Deep learning model

There were a few deep learning models that were available for Korean: mBERT (Pires et al. 2019)[6], KoBERT (SK Telecom 2019)[7], KorBERT (ETRI 2019)[8], KR- BERT

3 https://corpus.korean.go.kr/main.do

4 Originally, the Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2018) was included in the General Language Understanding Evaluation (GLUE) benchmark (Wang et al. 2019). On the other hand, the K-CoLA dataset was constructed independently, and it was not included in the KLUE benchmark (Park et al. 2021).

5 The K-CoLA data sets contained the following three files: `NIKI_CoLA_in_domain_tran.tsv` (7,957 grammatical and 8,083 ungrammatical sentences), `NIKI_CoLA_in_domain_dev.tsv` (569 grammatical and 508 ungrammatical sentences), `NIKI_CoLA_out_of_domain_dev.tsv` (510 grammatical and 472 ungrammatical sentences).

6 https://github.com/google-research/bert/blob/master/multilingual.md

7 https://github.com/SKTBrain/KoBERT

(Lee et al. 2020)[9], KcBERT (Lee 2020)[10], and so on. Among them, this paper took the KR-BERT, because it showed the best performance. The reason seemed to come from the tokenizing with Byte Pair Encoding (BPE; Senrich et al. 2016).

As mentioned in Section 1, two different experiments were conducted in the experiments with deep learning models. In the first experiment, the target sentences were examined after the KR-BERT had been trained only with the K-CoLA data set. As you will be shown in Section 3.6, however, the performance was not so good. In addition, the K-CoLA data set did not contain enough data for learning the HA and the PC as subjecthood diagnostics. The second experiment was designed to solve this problem. In the second experiment, the acceptability scores of the target sentences were measured after the KR-BERT had been trained using both the K-CoLA data set and the sentences which came from the Sejong Morphologically-Analyzed corpus (National Institute of Korean Language, 2007). The corpus contained 17,645 occurrences of '*-si*' and 62,845 occurrences of '*-tul*'. Among the sentences in the corpus, 1,000 sentences were randomly extracted from the corpus per each diagnostic test, which contained '*-si*' or '*-tul*' respectively. Since the sentences in the training data set were constructed to be the pair of acceptable/grammatical and unacceptable/ungrammatical sentences (Park et al. 2021), 1,000 additional sentences were made per the diagnostic based on the extracted sentences which were unacceptable/ungrammatical. For the HA, the sentences with and without the morpheme '*-si*' were labelled with either TRUE (acceptable/grammatical) or FALSE (unacceptable/ ungrammatical) respectively. For the sentences in PC, the sentences with the plural morpheme '*-tul*' both in the subject NPs and adverbials were labelled with TRUE, and the sentences with the plural morpheme '*-tul*' only in the adverbials (with sg. number of NPs) were labelled with FALSE respectively.[11] A total of 2,000 sentences

8 https://aiopen.etri.re.kr/service_dataset.php

9 https://github.com/snunlp/KR-BERT

10 https://github.com/Beomi/KcBERT

11 There were the following possible combinations on the plural suffix '*-tul*'.

(i) a. ... Subject NP [without '*-tul*'] ... Adverbial [without '*-tul*'] ...
b. *... Subject NP [without '*-tul*'] ... Adverbial [with '*-tul*'] ...
c. ... Subject NP [with '*-tul*'] ... Adverbial [without '*-tul*'] ...
d. ... Subject NP [with '*-tul*'] ... Adverbial [with '*-tul*'] ...

Among these four combinations, the patterns (ia) and (ic) appeared frequently in the corpus data. Therefore, the additional sentences were constructed based on the patterns (ib) and (id) so that (ib) and (id) made a pair and that each pattern contained exactly 1,000 sentences.

were constructed for HA and PC respectively, and they were added to the training data set. Then, the KR-BERT was trained with the data set that was the combination of the original K-CoLA and the 4,000 sentences that were extracted from the Sejong Morphologically-Analyzed corpus.

Several deep learning models in previous studies measured the acceptability scores with *surprisal* (Wilcox et al. 2019) or TRUE/FALSE (acceptable/unacceptable; Wang et al. 2019). However, it was difficult to compare the analysis results of these deep learning model(s) with those of the syntactic experiments in these analyses. Accordingly, Lee (2021) developed a new type of deep learning model where the acceptability scores could be measured with numeric scales, which ranged from 0 to 100 and were similar to the magnitude estimation in experimental syntax.

Even though the deep learning model (the KR-BERT model) was selected, it was necessary to re-train the model with the K-CoLA database, since the original BERT model was pre-trained with *unlabeled* data. Therefore, we downloaded the KR-BERT model from the Hugging Face and trained it with the K-CoLA database.[12]

3.4 Procedure

The procedure of experiments with deep learning models proceeded as follows. First, a data set was prepared which contained a total of 2,400 sentences (400 targets and 2,000 fillers), and a pre-trained KR-BERT model was downloaded from the Hugging Face site. Second, the dataset was used as input into the KR-BERT model and the acceptability scores were calculated for all the sentences in the target data set (whether they were targets or fillers), using the algorithm in Section 3.5. Third, after all the acceptability scores for the fillers (2,000 sentences) were extracted, the validity of the trained deep model was assessed by the procedure in Section 3.6. Fourth, after all the acceptability scores for the target sentences (400 sentences for HA/PC) were extracted, the scores were normalized using the *z*-scores.[13] Fifth, statistical analyses were applied to the normalized scores typically using R (R Core Team 2022), and the analysis results of the deep

12 https://huggingface.co/snunlp/KR-BERT-char16424

13 According to Gravetter and Wallnau (2013), there would be two different kinds of normalization methods. One was just to use the *z*-scores, and the other option was to re-convert the *z*-scores to the numeric scores which were similar to the original scores. Kim et al. (2017) and this paper took the second option for better graphic representations. However, remember that the mean score of each data set was adjusted to be 50, where the data set contained both target and filler sentences.

learning model were compared with those of the experiments in Section 2.3.

3.5 Measuring acceptability scores in the deep learning model

As mentioned in Section 3.3, since it was very difficult to compare the analysis results of the deep learning model(s) with those of the syntactic experiments, Lee (2021) developed a new deep learning model where the acceptability scores could be measured with numeric scales (0~100).

The acceptability scores in this model were measured as follows. The algorithm started with the basic architecture of the BERT model. (Devlin et al. 2019: 15)

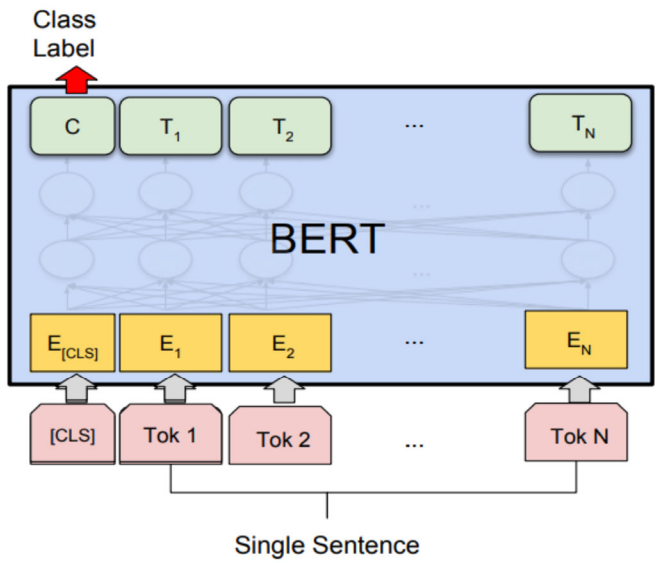

Figure 3. BERT model with a single sentence

After the original BERT model analyzed the input sentence, the model made a class label which was TRUE or FALSE. In Lee (2021), on the other hand, the last step of the output was revised so that the deep learning model produced both (i) a class label [CLS] and (ii) the probability which the given input sentence would be TRUE or FALSE. The outputs of the BERT model were basically *logit* values, and the *logit* values were converted to probabilities using the *inverse logit function*. After the probability was

computed for each input sentence, the values were normalized with both the minimal scores and the maximal scores in the input data set. Because the data set contained both perfectly acceptable sentences and perfectly unacceptable sentences, all the acceptability scores were located between 0 and 1.[14] Finally, these normalized probabilities were converted to acceptability scores by multiplying 100, which as a result ranged from 0 to 100.

3.6 Evaluation

After the acceptability scores were measured for all the sentences in the data set using the KR-BERT model (whether they were target sentences or fillers) of the first experiment, the validity of the model was evaluated with the filler sentences (2,000 sentences). The process proceeded in two steps.

In the first step, the performance of the KR-BERT model was tested with class labels [CLS]. Since the K-CoLA dataset contained a class label (i.e., correct answers of TRUE/FALSE) for all sentences, the class labels for all the fillers which were produced by the KR-BERT model were compared with those in the K-CoLA data set. 70.25% accuracy was obtained for this step. It implied that the KR-BERT model predicted the acceptability scores of filler sentences with 70.25% accuracy.

In the second step, the predicted acceptability scores (0 ~ 100) were divided into two groups with the following algorithm. If the score was equal to or greater than 50, the given sentence had the TRUE label. If not, the sentence had a FALSE label. Then, the classified labels were compared to the predicted labels in the KR-BERT model. 98.05% accuracy was obtained in this step. It suggested that the KR-BERT model predicted class labels (TRUE or FALSE; acceptable or unacceptable) with 98.05% of accuracy when the acceptability scores were converted into two labels. 1.95% of errors might come from the normalization process.

From these two steps of evaluation procedures, it could be said that the predicted acceptability scores in the KR-BERT would be about 69% of accuracy for the target sentences ($0.7025 \times 0.9805 = 0.6888$).[15]

---

14 For the reason why this normalization process was necessary, see Lee (2021).

15 For the same data set, mBERT showed about 57.2% of accuracy and KoBERT had about 67.3% of accuracy. These values were low than the accuracy of the English BERT on the CoLA data set. As mentioned in Lee (2021), the $\text{BERT}_{\text{LARGE}}$ model showed about 96.0% of accuracy. The reason for the

4. Analysis results

4.1 Experiment 1: K-CoLA

In the first experiment with the deep learning models, after the KR-BERT was trained only with the K-CoLA data set, the acceptability scores were measured with the trained deep learning model. The following bar plot showed the analysis results of the HA in this experiment.

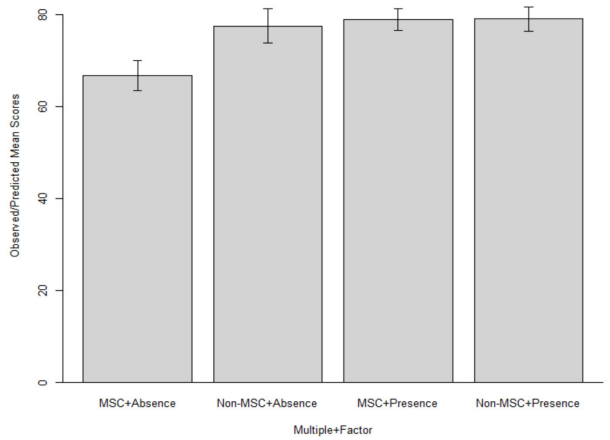

Figure 4. Bar plot (K-CoLA, HA)

As you could observe, the scores in Figure 4 became higher than those in Figure 1. Especially, the scores of the first three groups (MSC+Absence, Non-MSC+Absence, and MSC+Presence) were much higher than those in Figure 1. As for the general patterns, it was hard to say that Figure 1 and Figure 4 had similar patterns. Note that the acceptability scores of the last three groups (Non-MSC+Absence, MSC+ Presence, and Non-MSC+Presence) were similar in Figure 4 but the scores in Figure 1 demonstrated clear differences.

To examine the statistical significance of two factors and their interaction, the GLM analysis was conducted.[16] The following table illustrated the results of GLM analysis.

discrepancies seemed to come from the characteristics of the Korean language (such as *josa* and *e.mi*), even though further studies are necessary.

16 As in Kim et al. (2017), the distributions of acceptability scores did not follow the normal

Table 3. Analysis results of GLM (K-CoLA, HA)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	67.768	1.439	47.094	<.001	***
Multiple1	10.194	2.035	5.009	<.001	***
Factor1	11.494	2.035	5.648	<.001	***
Multiple1:Factor1	-10.064	2.878	-3.497	<.001	***

Unlike our expectation, not only the two factors (*Multiple1* and *Factor1*) but also the interaction (*Multiple1*:*Factor1*) was statistically significant (p<.001). The reason was that the mean value of the first and the third group (i.e., MSC+Absence and MSC+ Presence) was lower than that of the second and the fourth group (i.e., Non-MSC+ Absence and Non-MSC+Presence), which was the influence of *Multiple1*. Likewise, the mean of the first two groups (i.e., MSC+Absence and Non-MSC+Absence) was lower than that of the second two groups (i.e., MSC+Presence and Non-MSC+ Presence), that was the influence of *Factor1*. That's why the two factors (*Multiple1* and *Factor1*) were statistically significant.

The following bar plot showed the analysis results of the PC in this experiment.

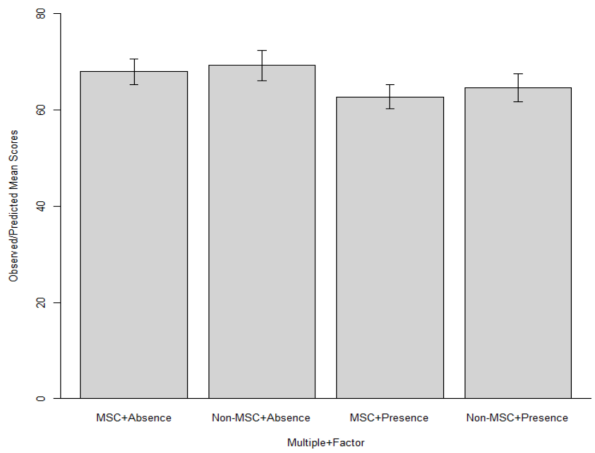

Figure 5. Bar plot (K-CoLA, PC)

As in Figure 4, the overall scores in this plot were also higher than those in Figure 2.

distributions. That's why the GLM analysis was adopted here.

As for the general patterns, it was also hard to say that Figure 2 and Figure 5 had similar patterns, since Figure 2 and Figure 5 demonstrated clearly different patterns.

The following table provided the results of GLM analysis.

Table 4. Analysis results of GLM (K-CoLA, PC)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	68.273	1.318	51.815	<.001	***
Multiple1	1.173	1.863	0.630	0.530	
Factor1	-4.830	1.863	-2.592	0.010	*
Multiple1:Factor1	0.559	2.635	0.212	0.823	

As shown in this table, only *Factor1* was significant (p<.05). The reason was that the mean of the first two groups (i.e., MSC+Absence and Non-MSC+Absence) was lower than that of the second two groups (i.e., MSC+Presence and Non-MSC+ Presence). The patterns in Table 4 were also very different from those in Table 2.

4.2 Experiment 2: K-CoLA + Sejong Corpus

In the second experiment, the KR-BERT was trained with the K-CoLA plus the sentences which were extracted from the Sejong Morphologically-Analyzed Corpus, and the acceptability scores were measured with the trained KR-BERT model.

The analysis results of the HA in this experiment are displayed in the following bar plot.

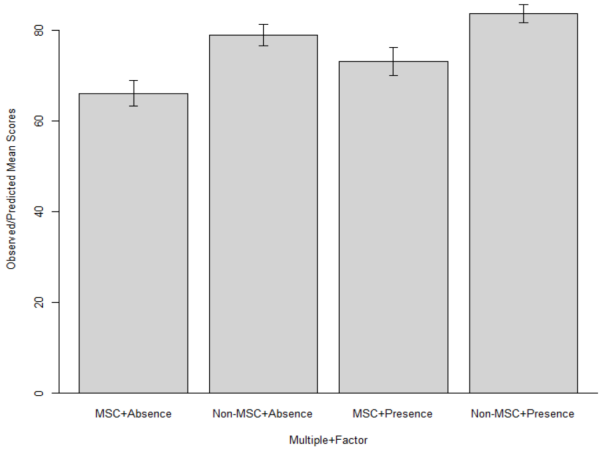

Figure 6. Bar plot (K-CoLA + Sejong, HA)

As you could observe, the overall patterns in this figure were very similar to those in Figure 1. The mean values of the first and the third group (i.e., MSC+Absence and MSC+Presence) were lower than those of the second and the fourth group (i.e., Non-MSC+Absence and Non-MSC+Presence), and the means of the first two groups (i.e., MSC+Absence and Non-MSC+Absence) were lower than those of the second two groups (i.e., MSC+Presence and Non-MSC+Presence). These patterns were also observed in the analysis results of the experimental approach in Figure 1.

The results of the GLM analysis were shown in the following table.

Table 5. Analysis results of GLM (K-CoLA + Sejong, HA)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	73.890	1.285	57.500	<.001	***
Multiple1	10.468	1.817	5.760	<.001	***
Factor1	-6.967	1.817	-3.834	<.001	***
Multiple1:Factor1	2.290	2.570	0.891	0.374	

This table mentioned that the two factors (*Multiple1* and *Factor1*) were statistically significant (p<.001) but that the interaction (*Multiple1*:*Factor1*) was not. The reason seemed to be that the scores of the first three groups (MSC+Absence, Non-MSC+Absence, and MSC+Presence) were significantly higher in Figure 6 (than those in Figure

1).

The analysis results from the PC in the second experiment were presented in the following bar plot.

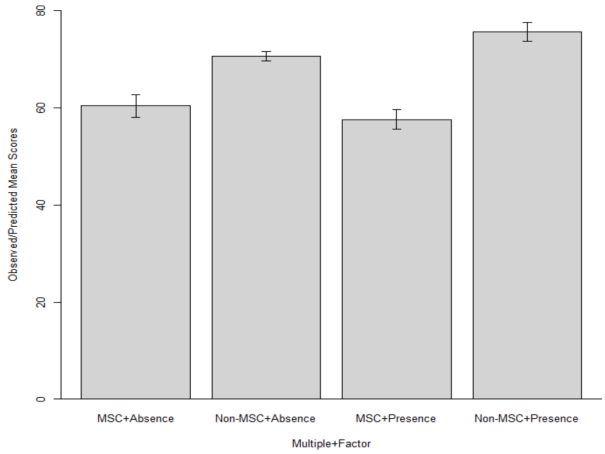

Figure 7. Bar plot (K-CoLA + Sejong, PC)

As you could see, the overall patterns in this figure closely resembled those in Figure 2. The differences/distances between the first two groups (i.e., MSC+Absence and Non-MSC+Absence) were smaller than those between the second two groups (i.e., MSC+Presence and Non-MSC+Presence), and the mean values of the first and third groups (i.e., MSC+Presence and MSC+Absence) were lower than those of the second and fourth groups (Non-MSC+Absence and Non-MSC+Presence). These patterns were also observed in the analysis results of Figure 2.

The results of the GLM analysis were displayed in the following table.

Table 6. Analysis results of GLM (K-CoLA + Sejong, PC)

	Estimate	Standard Error	*t*	*p*	
(Intercept)	58.113	0.927	62.708	<.001	***
Multiple1	17.918	1.311	13.672	<.001	***
Factor1	2.736	1.311	2.087	0.038	*
Multiple1:Factor1	-7.722	1.853	-4.166	<.001	***

This table illustrated that the interaction between the two factors (*Multiple1*:*Factor1*), as well as the two factors themselves (*Multiple1* and *Factor1*), were statistically significant (p<.001). These patterns closely matched the analysis results in Table 2 of the experimental approach.

5. Discussion

5.1 Implications on the subjecthood tests

Diagnostic tests for the subjecthood have long been one of the hottest topics in Korean syntax because Korean has MSCs as well as Non-MSCs (or Single Subject Constructions). In the MSCs (and other syntactic constructions such as Dative Subject Constructions), the questions have been how many subjects exist in Korean and how the subject(s) can be identified.

There have been lots of theoretical approaches to the subjecthood tests in Korean (Yoon 1986; Yoon 1989; Hong 1991; Park 1995; Schütze 2001; Park 1973; Lee 1997; Kang 2002; Yoon 2009, 2015; etc.), and they were summarized in (2). After the advent of experimental syntax (Bard et al. 1996; Schütze 1996; Cowart 1997; Keller 2000), there were a few trials to examine the validity of the subjecthood tests with syntactic experiments. Kim et al. (2015) investigated Obligatory Control (OC, 2f) and Coordinated Deletion (CD, 2h), Lee et al. (2015a) examined Reflexive Binding (RB, 2i) and Adjunct Control (AC, 2g), and Kim et al. (2017) scrutinized HA (2c) and PC (2b). In the comparative analysis of Lee et al. (2015b), however, OC and RB showed the best performance, and HA and PC demonstrated the worst results, which was against the predictions of Hong (1991, 1994).

Two experiments with the deep learning models in this paper revealed interesting facts about HA and PC as subjecthood diagnostic tests. As shown in Section 4, the analysis results of the first experiment (Section 4.1) were a little far from those of the experiments with human participants (Section 2.3). Note that only the K-CoLA data set was employed in the training of the KR-BERT model. In the second experiment, however, the analysis results of the deep learning model (Section 4.2) were quite similar to those of the syntactic experiments (Section 2.3). Remember that 2,000 sentences were extracted from the Sejong Morphologically-Analyzed Corpus and that another 2,000

sentences were constructed as unacceptable sentences, following the design in Park et al. (2021). Those 4,000 sentences were combined with the original K-CoLA data set, and all of them were used in the training process.

The fact that the analysis results in Section 4.2 (Figure 6 and Figure 7) were quite similar to those in Section 2.3 (Figure 1 and Figure 2) implied that both HA and PC could be used as subjecthood diagnostics, as Hong (1991, 1994) proposed. In addition, the GLM analysis (Table 5 and Table 6) demonstrated that both factors *Multiple1* and *Factor1* were statistically significant. This implied that the agreement features (i.e., '*-si*' and '*-tul*' respectively) significantly played important roles in the determination of subjecthood in Korean sentences, whose effect could be measured with *Factor1*.

Then, why did HA and PC show the worst performance in the comparisons with other subjecthood tests in Lee et al. (2015b)? The reason could be found in the comparison between the analysis results in the first experiment and those in the second experiment with deep learning models. Note that only the K-CoLA data set was used in the first experiment but that 4,000 more sentences were added in the second experiment. The important thing to be noticed here was that the additional 4,000 sentences were oriented to HA and PC. That is, all of them were designed for testing HA and PC as subjecthood tests. And, as expected, the KR-BERT model became close to the experiment with human participants.

Here, the question was whether the environments of daily language life were close to those of Experiment 2 or to those of Experiment 1? Maybe, the answer was 'Experiment 1'. This implied that, in the language use of daily life, there was not enough data for the HA and PC to play important roles as subjecthood tests. That's why HA and PC showed the worst performance in the comparisons with other subjecthood tests in Lee et al. (2015b). Experiment 2 showed that both HA and PC could be used as important diagnostic tests for the subjecthood in Korean, only when native speakers were provided with enough data.

One more interesting fact was that PC showed worse performance than HA in Experiment 1. The reason seemed to be found in the frequency effects of '*-si*' and '*-tul*'. Actually, the K-CoLA data set contained 1,407 occurrences of '*-si*' and 1,605 occurrences of '*-tul*'. From the perspective of raw frequencies, the frequency of '*-tul*' was slightly higher than that of '*-si*'. However, there were many cases where '*-si*' co-occurred with the subject NPs with honorific markers, but there were few cases where '*-tul*' occurred in the adverbials or predicates. That is, in most of the examples in '*-si*', they agreed with

the honorific markers '*-kkeyse*'. In the sentences with '*-tul*', however, the morpheme occurred in the subjects NPs, but there were only a few cases where '*-tul*' appeared with the adverbial or predicates. This implied that the deep learning model had little chance to learn the distributions of '*-tul*' where the model had to learn '*-tul*' as an agreement morpheme. This also indicated that the frequency effects (the *entrenchment* effects in Cognitive Linguistics) were important not only for human beings but also for deep learning models.

5.2 Implications on the deep learning models

McCoy et al. (2020) mentioned that an important research question in language modeling was to design the models which possibly had an appropriate inductive bias such that their internal linguistic representations and capabilities could resemble as much as possible the ones of human language learners after they were exposed with an as little volume of raw training data as the ones humans learners are exposed to. In this sense, this study demonstrated how important the training data were in language modeling with deep learning techniques. In the first experiment with only the K-CoLA data set, the language model was a little far from our expectations (from the intuition of native speakers). In the second experiment with more data, however, the language model became close to the intuitions of native speakers. This fact clearly illustrated how important enough and relevant training data were for the correct language modelling of human intuition. In addition, this study also showed that the language models could be fine-tuned more accurately with the relevant data.

The experiments in this study also demonstrated how useful the deep learning models could be employed when there were some cases where the experiments with human participants were impossible or hard to be conducted. Educating Korean native speakers on the use of '*-si*' and '*-tul*' morphemes would be either impossible or hard to be conducted, even though it could theoretically be possible. However, such trials were possible in the experiments with deep learning models by the change of the training data sets. Accordingly, this characteristic could be another advantage of using the deep learning models in the study of syntax, in addition to many advantages mentioned in Lee (2021) and Lee (2022).

However, remember that there are some discrepancies between the analysis results of

deep learning models and those of the experiments with human participants. Note that the acceptability scores in the deep learning models became higher than those of the experiments with human participants. Especially, the acceptability scores of the first three groups (MSC+Absence, Non-MSC+Absence, and MSC+Presence) were much higher than those in the deep learning models. These discrepancies were due to the characteristics of deep learning models; and this was why the deep learning experiments could not substitute the experiments with human participants, as Lee (2021) mentioned. Remember that the experiments with deep learning models would assist but could not substitute the experiments with human participants.

6. Conclusion

The study investigated two types of subjecthood diagnostics in Korean using deep learning models. For the experimental analysis, this study employed the analysis results of Kim et al. (2017). For the deep learning analysis, this paper adopted three types of data sets (the K-CoLA, the Sejong Morphologically-Analyzed Corpus, and the same target sentences in Kim et al. (2017). This paper also utilized the KR-BERT as a deep learning model.

There were two independent experiments in the deep learning analysis. In the first experiment, the KR-BERT was trained only with the K-CoLA data set, and the target sentences were analyzed with the trained model. In the second experiment, however, the KR-BERT was trained with both the K-CoLA and the 4,000 sentences from the Sejong Morphologically-Analyzed corpus. The acceptability scores in the KR-BERT were measured with the numeric scores with the algorithms in Lee (2021). After the experiments with the deep learning models, the acceptability scores were normalized and statistically analyzed with the GLMs. Through the two experiments, the following facts were observed: The language model in the first experiment was a little far from those of the experiments with human participants but the deep learning model in the second experiment was close to those of the syntactic experiments.

The analysis results implied that both HA and PC could be used as subjecthood diagnostics. However, the reason why HA and PC showed the worst performance in the comparisons with other subjecthood tests in Lee et al. (2015b) was that there were not enough data for the HA and PC to play an important role as subjecthood tests in the

language use of daily life. This paper also had an implication on deep learning literature in that it showed how important enough and relevant training data were for the correct language modelling of human language and how useful the deep learning models could be used when the experiments with human participants were impossible or hard to be conducted.

This paper demonstrated that deep learning models could be used for the study of syntactic phenomena and how deep learning models could be used to assist the syntactic experiments with human participants. We hope that the deep learning techniques can provide a perspective on the study of language.

References

Bard, Ellen, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1): 32-68.

Choe, Jae-Woong. 2004. Obligatory honorification and the honorific feature. *Studies in Generative Grammar* 14(4): 545-559.

Cowart, Wayne. 1997. *Experimental syntax*: *Applying objective methods to sentence judgments*. Thousands Oaks, CA: Sage Publications.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Falk, Yehuda. 2006. *Subjects and universal grammar*. Cambridge: Cambridge University Press.

Gravetter, Fedrick and Larry Wallnau. 2013. *Statistics for the behavioral sciences*. Belmont, CA: Wadsorth.

Gries, Stefan Th. 2013. *Statistics for linguistics with R*: *A practical introduction*. Berlin: Gruyter.

Han, Eunjoo. 1990. *Honorification in Korean*. Manuscript. Stanford University.

Heyock, Caroline. 1993. Syntactic predication in Japanese. *Journal of East Asian Linguistics* 2(2): 167-211.

Hong, Ki-Sun. 1991. *Argument selection and Case-marking in Korean*. PhD Dissertation. Stanford University.

Hong, Ki-Sun. 1994. Subjecthood tests in Korean. *Language Research* 30(1): 99-136.

Hong, Ki-Sun. 2014. Hankwuke-uy kyekcwungchwul kwumwun-kwa tamhwakwuco (Multiple case construction in Korean and the discourse structure). In *Proceedings of 2014 Korean Society for Language and Information (KSLI) Annual Conference*, 101-109.

Kang, Beom-Mo. 2002. *Pemcwu mwunpep*: *Hankwuke-uy hyengthaylon, thongsalon, thaipnonlicek uymilon* (*Categorial grammar*: *The morphology, syntax, and type-logical semantics of Korean*). Seoul: Korea University Press.

Keller, Frank. 2000. *Gradient in grammar*: *Experimental and computational aspects of degrees of grammaticality*. PhD Dissertation. University of Edinburgh.

Kim, Jeong-Seok, Jin Hyung Lee, and Jee Young Lee. 2022. A note on subject honorification in Korean. *Studies in Generative Grammar* 32(1): 153-167.

Kim, Ji-Hye, Yong-hun Lee, and Eunah Kim. 2015. Obligatory control and coordinated deletion as Korean subject diagnostics: An experimental approach. *Language and Information* 19(1): 75-101.

Kim, Ji-Hye, Yong-hun Lee, and Eunah Kim. 2017. Honorific agreement and plural copying as Korean subject diagnostics: An experimental approach. *Studies in Modern Grammar* 93: 119-144.

Kim, Jong-Bok and Peter Sells. 2007. Korean honorification: A kind of expressive meaning. *Journal of East Asian Linguistics* 16(4): 303-336.

Kuh, Hakan. 1987. Plural copying in Korean, In Susumu Kuno (eds.), *Harvard studies in Korean linguistics II*, 239-250. Cambridge, MA: Department of Linguistics, Harvard University.

Lee, Chungmin. 1990. Grammatical constraints on honorific expressions in Korean. Manuscript. Seoul National University.

Lee, Hyo-Sang. 1985. Causatives in Korean and the binding hierarchy. In William Eilfort, Paul Kroeber, and Karen Peterson (eds.), *Papers from the parasession on causatives and agentivity at the 21st regional meeting of Chicago Linguistic Society*, 138-153. Chicago, IL: University of Chicago Press.

Lee, Ik-Hwan. 1997. Double subject constructions in GPSG. In Susumu Kuno (eds.), *Harvard Studies in Korean Linguistics VII*, 287-296. Cambridge, MA: Department of Linguistics, Harvard University.

Lee, Junbum. 2020. KcBERT: Korean comments BERT. *Proceedings of the 32nd Annual Conference on Human & Cognitive Language Technology*, 437-440.

Lee, Sangah, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. KR-BERT: A small-scale Korean-specific language model. arXiv preprint arXiv:2008.03979.

Lee, Yong-hun. 2016. *Corpus linguistics and statistics with R*. Seoul: Hankook Publishing Co.

Lee, Yong-hun. 2021. English island constraints revisited: Experimental vs. deep learning approach. *English Language and Linguistics* 27(3): 21-45.

Lee, Yong-hun. 2022. Lexical effects in island constraints: A deep learning approach. *The Linguistic Association of Korea Journal* 30(1): 179-201.

Lee, Yong-hun, Eunah Kim, and Ji-Hye Kim. 2015a. Reflexive binding and adjunct control as subject diagnostics in Korean: An experimental approach. *Studies in Language* 31(2): 427-449.

Lee, Yong-hun, Yeonkyung Park, and Eunah Kim. 2015b. A multi-level analysis of subjecthood diagnostics in Korean. *Linguistic Research* 32(3): 671-691.

McCoy, Thomas, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics* 8: 125-140.

National Institute of Korean Language. 2007. *The final report of 21th century Sejong project*. Seoul: National Institute of Korean Language.

Park, Byung-Soo. 1973. On the multiple subject constructions in Korean. *Linguistics* 100: 63-76.

Park, Ki-Seong. 1995. *The semantics and pragmatics of Case marking in Korean: A role and reference grammar account*. PhD Dissertation. State University of New York at Buffalo.

Park, Kwonsik, Seongtae Kim, and Sanghoun Song. 2021. Verification of Korean pre- trained models' feasibility of syntactic research using pairwise sentences. *Language and Information* 25(3): 1-21.

Park, Sungjoon, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? arXiv preprint arXiv:1906.01502.

R Core Team. 2022. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Schütze, Carson. 1996. *The empirical base of linguistics*: *grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Schütze, Carson. 2001. On Korean 'Case stacking': The varied functions of the particles *-ka* and *-lul*. *The Linguistic Review* 18: 193-232.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909

Song, Sanghoun, Jae-Woong Choe, and Eunjeong Oh. 2019. An empirical study of honorific mismatches in Korean. *Language Sciences* 75: 47-71.

Song, Suk-Choong. 1975. Bare Plural Marking and Ubiquitous Plural Marker in Korean. In Robin Grossman and James San (eds.), *Proceedings of the 11th regional meeting of Chicago Linguistic Society*, 536-546. Chicago, IL: University of Chicago Press.

Suh, Cheong-Soo. 1977. Remarks on subject honorification, In Chin-Woo Kim (ed.), *Papers in Korean Linguistics*, 297-304. Columbia, SC: Hornbeam Press.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv: 1804.07461.

Warstadt, Alex, Amanpreet Singh, and Samuel Bowman. 2019. Neural network acceptability Judgments. arXiv preprint arXiv:1805.12471.

Wilcox, Ethan, Roger Levy, and Richard Futrell. 2019. What syntactic structures block dependencies in RNN language models? arXiv preprint arXiv:1905.10431.

Williams, Edwin. 1980. Predication. *Linguistic Inquiry* 11(1): 203-238.

Yoon, James. 1986. Some queries concerning the syntax of multiple subject constructions in Korean. *Studies in the Linguistic Sciences* 16(3): 215-236.

Yoon, James. 1987. Some queries concerning the syntax of multiple subject constructions in Korean. In Susumu Kuno (ed.), *Harvard Studies in Korean Linguistics II*, 138-162. Cambridge, MA: Department of Linguistics, Harvard University.

Yoon, James. 2004. Non-nominative (major) subjects and case-stacking in Korean. In Peri Bhaskararao and Karumuri Venkata Subbarao (eds.), *Non-nominative subjects* 2, 265-314. Berlin: Mouton.

Yoon, James. 2008. Subjecthood and subject properties in multiple subject constructions. Handout provided in East Asian Linguistics Seminar, February 26. Oxford, UK: Oxford University.

Yoon, James. 2009. The distribution of subject properties in multiple subject constructions. *Japanese/Korean Linguistics* 19: 64-83.

Yoon, James. 2015. Double nominative and double accusative constructions. In Lucien Brown and Jaehoon Yeon (eds.), *The handbook of Korean linguistics*, 79-97. Chichester: John Wiley & Sons, Inc.

Yoon, Jong-yurl. 1989. On the multiple *-ka* and *-lul* constructions in Korean. In Susumu Kuno (ed.), *Harvard Studies in Korean Linguistics III*, 383-394. Cambridge, MA: Department of Linguistics, Harvard University.

Youn, Cheong. 1990. *A relational analysis of Korean multiple nominative constructions*. PhD Dissertation. State University of New York at Buffalo.

Yong-hun Lee
Lecturer
Department of Linguistics
Chungnam National University
99 Daehak-ro, Yuseng-gu
Daejeon 34134, Korea
E-mail: yleeuiuc@hanmail.net

Ji-Hye Kim
Associate Professor
Department of English Education
Korean National University of Education
250 Taeseongtapyeon-ro
Cheongju, Chungcheongbuk-do 28173, Korea
E-mail: psychlg@gmail.com

Received: 2022. 09. 30.

Revised: 2022. 11. 21.
Accepted: 2022. 11. 27.