The Journal of Studies in Language

The Journal of Studies in Language 35(3), 361-373 (2019)

Analyzing Subjecthood Tests in Korean Using a Cross-Validation

Lee, Yong-hun* and Kim, Ji-Hye**

Chungnam National University
Korea National University of Education

***First Author / **Corresponding Author**

ABSTRACT

The Journal of Studies in Language 35.3, 361-373. Many diagnostics have been proposed to examine subjecthood in Korean, some of which were investigated in terms of their validity as subject diagnostics through a series of experimental studies. This paper takes the experimental data from three experimental studies (Kim et al., 2015; Lee et al., 2015a; Kim et al., 2017) that were previously conducted, and examines the validity of six diagnostic tests using cross-validation and a machine learning method. The experimental data were evenly divided into ten parts (10-fold cross-validation): Nine of them were used for machine learning (training data set), and the remaining one was for the testing (test data set). The analysis was conducted ten times, and the average values were taken. A simple linear regression was conducted for machine learning. Through the analysis, the following results were observed: (i) The accuracy of the diagnostics for Single Subject Construction (SSC) was much higher than that of Multiple Subject Construction (MSC) (SSC: 81,78, MSC: 68.66), (ii) Adjunct Control was the most accurate for SSC, whereas Coordinated Deletion for MSC, (iii) HA and PL Showed low performance. The results also seem to imply that some parts cannot be explained by a simple linear model but can be analyzed with non-linear modeling of language. **(Chungnam National University · Korea National University of Education)**

Keywords: subjecthood diagnostics, multiple subject constructions, magnitude estimation, cross-validation, machine learning

https://doi.org/10.18627/jslg.35.3.201911.361

pISSN : 1225-4770
eISSN : 2671-6151

Received: October 14, 2019

Revised: November 10, 2019

Accepted: November 15, 2019

1. Introduction

It has been well-known that Subject and other grammatical roles (GRs) play an important role in Korean - especially, through a line of approach where Subjecthood is considered a defined/derived notion (i.e., defined in terms of structure) (Andrews, 1985; Falk, 2006). Although several scholars have proposed identifying Subjects through diagnostic properties (Yoon, 1986; Hong, 1991; Youn, 1990, etc.), there are still some controversies, since scholars do not agree as to what could be the exact set of subject properties.

In addition, those proposed diagnostics are based on the intuitions of linguists, which can be biased to represent native speakers' intuitions. Recently, as computer techniques and statistics develop, some studies took experimental approaches and examined the validity of the diagnostics (Kim et al., 2015; Lee et al., 2015a; Kim et al. 2017, among others.). However, each experiment was conducted only once in a limited setting - for a few subjecthood tests only out of the extensive list of possible subjecthood diagnostics, which means that there is a possibility that further experiments could yield different results in other times and other situations - or with other participants. This problem motivated the current study, which started from the question as to how to get more reliable analysis results from a few discrete experiments. In order to answer the question, the current study attempted to examine the validity of various subjecthood tests using a cross-validation and a machine learning method, which are based on the analysis results of several previous experiments.

For the present study, the experimental data of three previous studies (Kim et al., 2015; Lee et al., 2015a; Kim et al. 2017) were taken, which covered six types of subjecthood diagnostics in Korean. The experimental data were evenly divided into 10 parts (10-fold cross-validation), nine of which were used for machine learning (training data set) and the remaining one for the testing (test data set). The analysis was conducted 10 times, and the average values were taken. A simple linear regression was conducted for machine learning.

2. Previous Studies

2.1 Subject Properties in Korean

2.1.1 Subjecthood Tests in Korean

Among the previous studies on what is valid subjecthood diagnostics in Korean, the following subjecthood diagnostics were some of those proposed, which have been examined in some recent experimental studies (Kim et al., 2015, 2017, Kim et al., 2016, Lee et al., 2015a, etc.).

(1) Proposed Subject Diagnostics in Korean (Yoon, 1986; Hong 1991; Youn 1990, etc.)
   a. Nominative Case marking
   b. Controller of optional plural-marking
   c. Controller of subject honorification
   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 coordinate deletion
   i. Antecedent of (subject-oriented) anaphors
   j. Exhaustive-listing interpretation of '*-ka*/*-i*'

Though all the listed subjecthood tests in (1) are all valid in dealing with Single Subject Constructions (SSCs) in Korean as in (2a), in case of Korean Multiple Subject Constructions (MSCs) as shown in (2b) where there are more than one Nominative-marked NPs, the theoretical opinions diverge and questions arise as to which NP takes the given Subject property.

(2) SSC vs. MSC

a. *Cheli-ka* *Yenghi-lul* *ttayliessta.*
Cheli-NOM Yenghi-ACC hit-PAST-DECL
'Cheli hit Yenghi.'

b. *Cheli-ka* *apeci-ka* *pwuca-i-ta.*
Cheli-NOM father-NOM rich-COP-DECL
'It is Cheli whose father is rich.'

One line of approach regarding sentences like (2b) claims that the outer nominative NP in the construction are not a Subject, but a Topic or a Focus (Yoon, 1986; J-Y Yoon, 1989; K-S Hong, 1991; K-S Park, 1995; Schütze, 2001, etc.), while others assume that all those Nom-marked NPs can be Subjects in different ways (B-S Park, 1973; Teng, 1974; I-H Lee, 1997; Heycok, 1993; B-M Kang, 2002; Yoon, 2004, 2007, 2008, etc.) or in different layers of derivations (C. Youn, 1990; S-E Cho, 2000, etc.). The question of whether there is more than one Subject position in MSCs depends upon whether the putative subject properties can be controlled by more than one Nom-marked NP in MSCs. In case of SSCs where there is only one Subject position, the subject properties shown in (1) would be controlled by the single NP (which has a combination of Major/Grammatical Subject properties in MSCs). However, the properties of Subject may diverge in MSCs; and Yoon (2008, 2009) specifically propose that (1b), (1c), (1f), and (1i) are properties of Grammatical Subject (GS) but that (1d), (1g), and (1h) lean towards Major Subject (MS) in MSCs.

Given the previous discussions of subjecthood in MSCs, there are several problems in terms of validity of the discussions. First of all, most of the previous theoretical studies on Subjecthood in Korean have been based primarily on the intuitions of Korean linguists, which turned out to be different from non-linguist native speakers of Korean through some experimental studies (Kim et al., 2015; Lee et al., 2015a; Kim et al., 2016; Kim et al., 2017, etc.). Only a handful of studies based on experimental data have attempted to address the question of subjecthood properties in different NPs in SSCs and MSCs, by testing non-linguist Korean native speakers. Those experimental studies investigated some of the proposed subjecthood tests out of the list mentioned earlier in (1): Kim et al. (2017) for the diagnostics (1b) and (1c); Kim et al. (2015) for (1f) and (1g); Lee et al. (2015a) for (1h) and (1i), etc. These experimental studies demonstrated that what have been claimed only with the intuition of Korean linguists did not converge on the performance of non-linguist native speakers.

Another problem of the previous studies which motivated the present study is that those experimental results come from one-time experiment for each time, which needs to be generalized in broader population and situations. Since the experiments in the previous studies have been performed only once for each subjecthood test, the question arises as to how to get more reliable results from those previous experiments. Therefore, the goal of the current paper is to statistically examine the validity of subjecthood tests using a cross-validation and a machine learning method.

3. Research Methods

3.1 Designs of the Previous Experimental Studies

The current study is based on the results of three previous experimental syntactic studies (Kim et al., 2015; Lee et al., 2015; Kim et al., 2017). The three studies were based on the same methodology and tested the same population but with different diagnostic properties. For example, Kim et al. (2015) focused on Obligatory Control (OC) and Coordinated Deletion (CD), while Lee et al. (2015a) tested Adjunct Control (AC) and Reflexive Binding (RB). Later, Kim et al. (2017) applied the similar methodology for examining the other two diagnostics - Honorific Agreement (HA) and Plural Copying (PC). Since the current study is built up through their experimental results and attempted to incorporate all of the six diagnostics together for analysis, this subsection first introduces overall methodology used in all six studies.

3.1.1 Participants

Seventy Korean native speakers (age range: 21~45; *m*=27.53, *sd*=9.71) participated in the experiments who resided in and near Seoul, South Korea. They were either current university students or graduates. They were monolingually raised in Korea and have not lived in non-native language-speaking countries more than 3 months. Most of the participants were recruited from the college of Liberal Art and Social Science.

3.1.2 Task, Materials, and Procedure

The main task used in the experiments was an acceptability judgment task using the Magnitude Estimation (ME), in which the participants were asked to draw different lengths of lines to indicate the naturalness (acceptability) of a given sentence (after reading the sentence). The test materials consisted of 40 sentences: 20 target sentences (4 type conditions × 5 tokens) and 20 filler sentences. Accordingly, a total of 240 sentences were used in the experiments.

The target sentence types had a 2×2 design, according to sentence type (i.e., SSC vs. MSC) and the NP that functions as the controller of a given subjecthood diagnostic (NP1 vs. NP2). In MSCs, the two NPs of interest are the initial NP, NP1 (i.e., Major Subject) and the second/final NP, NP2 (i.e., Grammatical Subject). We also had SSCs that closely matched MSCs where NP1 was a Possessor of NP2 (i.e., the unique Subject nominal). In this way, the 4 conditions for the target sentence types were constructed which were applied to all 6 diagnostics, as illustrated below.

(3) Target Sentence Types

a. Type 1: $[\text{NP1}]_{\text{poss}}$ $\mathbf{[NP2]_{nom}}$ $\text{Predicate}_{[\text{controlled by NP2}]}$ (SSC+NP2 controller)

b. Type 2: $\mathbf{[NP1]_{poss}}$ $[\text{NP2}]_{\text{nom}}$ $\text{Predicate}_{[\text{controlled by NP1}]}$ (SSC+NP1 controller)

c. Type 3: $[\text{NP1}]_{\text{nom}}$ $\mathbf{[NP2]_{nom}}$ $\text{Predicate}_{[\text{controlled by NP2(GS)}]}$ (MSC+NP2 controller)

d. Type 4: $\mathbf{[NP1]_{nom}}$ $[\text{NP2}]_{\text{nom}}$ $\text{Predicate}_{[\text{controlled by NP1(MS)}]}$ (MSC+NP1 controller)

The sample target sentences in the forms shown in (3) are illustrated in (4) below, in the format used in the diagnostics of Honorific Agreement (HA) among others. For instance, the honorific morpheme *-si* in (4a) is controlled by (or agrees with) the Subject 'Cheli's father', but in (4b) by the possessor of Subject 'Professor Kim', and in (4c) representing MSC - by Grammatical Subject (GS) 'Cheli's father', and Major Subject (MS) 'Professor Kim' in (4d), respectively.

(4) Target Sentences for HA

a. *Cheli-uy* ***apeci-ka*** *pwuca-i-**si**-ta.*
Cheli-GEN father-NOM be.rich-COP-HON-DECL
'Cheli's father is rich.' (Subject + SSC)

b. ***Kim kyoswu-nim-uy*** *cengwen-i* *alumtawu-**si**-ta.*
Professor Kim-GEN garden-NOM be.beautiful-HON-DECL
'Professor Kim's garden is beautiful.' (Non-subject (possessor) + SSC)

c. *Cheli-ka* ***apeci-ka*** *pwuca-i-**si**-ta.*
Cheli-NOM father-NOM be.rich-COP-HON-DECL
'Cheli's father is rich.' (GS + MSC)

d. ***Kim kyoswu-nim-i*** *cengwen-i* *alumtawu-**si**-ta.*
Professor Kim-NOM garden-NOM be.beautiful-HON-DECL
'As for Professor Kim, his garden is beautiful.' (MS + MSC)

Likewise, the examples of the target sentences from each subjecthood diagnostics - with the format of (4a) above - are shown in (5). In (5a) representing Plural Copying (PC), plural morpheme *-tul* is copied to adverbials and is controlled by the Subject 'buildings.' Also, in (5b), the PRO is interpreted as its controller 'Cheli's brother', which is Subject of the sentence. In (5c), the reflexive *cakicasin* is interpreted to be 'Cheli's father' as its antecedent. Likewise, in (5d) and (5e), the deleted Subject for the coordinated structure, and the PRO of the adjunct clauses, are interpreted as 'Cheli' (cf. 5d) and 'President Kim' (cf. 5e), respectively, both of which are the main clause Subject.

(5) Target Sentences for Each Diagnostics

a. *Chicago-uy* *kenmwul-**tul**-i* *acwu-**tul*** *nop-ta.*
Chicago-GEN building-PL-NOM very-PL be.high-DECL
'Chicago's buildings are high.' **(Plural Copying: PC)**

b. *Cheli-uy* *tongsayng-i* [***PRO*** *yuhak-ul* *ka-ko*] *sipheha-n-ta.*
Cheli-GEN brother-NOM study.abroad-ACC go-COMP want-PRES-DECL
'Cheli's brother wants to study abroad.' **(Obligatory Control: OC)**

c. *Cheli-uy* *apeci-ka* ***cakisasin**-ul* *nul* *cinachihey* *piphanha-n-ta.*
Cheli-GEN father-NOM self-ACC always severely criticize-PRES-DEC
'Cheli's father always criticizes self severely.' **(Reflexive Binding: RB)**

d. [*Cheli-ka* *Dongswu-lul* *cohaha-ko*], [***e*** *Yengi-lul* *acwu* *silheha-n-ta.*]
Cheli-NOM Dongswu-ACC like-CONJ Yenghi-ACC very hate-PRES-DECL
'Cheli likes Dongswu and hates Yenghi.' **(Coordinated Deletion: CD)**

e. [*Pilok* ***PRO*** *nalk-ass-ci-man*], *Kim.sacang-uy* *samulsil-i* *calcengtontoy-e* *iss-ta.*
Although old-PAST-COMP-but Kim.president-GEN office-NOM well-cleaned-COMP be-DECL
'Although it is old, President Kim's office is clean.' **(Adjunct Control: AC)**

The experiments were conducted with these target sentences, along with six subjecthood diagnostics. Each target sentence has its counterparts for non-subject (possessor) controller in SSCs and GS vs. MS controller in MSCs, as illustrated in the format in (3) and examples with HA diagnostic shown in (4). The acceptability scores of four different types of sentences in each diagnostics were measured by through Magnitude Estimation (ME). Based on the acceptability scores of the participants, the results were examined with two distinct factors: *Multiple* (SSC vs. MSC) and *Factor* (agreement with or controlled by NP1 or NP2); the same method was applied to all of six subjecthood diagnostics (HA, PC, OC, RB, CD, and AC).

3.1.3 Regression Analysis

After the acceptability scores were collected in the experiments, the collected scores were statistically analyzed. First, the normality tests (especially, Shapiro-Wilks test) were conducted to the data set of each diagnostic so that the distributions of the scores were examined. The normality tests revealed that all of the six data sets did not follow the normal distribution.[1] Therefore, a generalized linear regression (GLM, a non-parametric test) was adopted to check how the two factors (*Multiple* and *Factor*) behaved and how they interacted with each other.[2] Table 1 summarizes the results of the analysis.[3]

Table 1. Results of Regression Analysis

	HA	PC	OC	RB	CD	AC
(Intercept)	***	***	***	***	***	***
Multiple (SSC vs. MSC)	***	***	***	***	***	***
Factor (NP1 vs. NP2)	***		***	***	***	
Multiple:Factor	***	*	***	***	***	

As is shown from the above table, (Intercept) and *Multiple* (i.e., the distinction between SSC and MSC) were statistically significant in all of the six diagnostics. *Factor* (i.e., agreement with or controlled by NP1 or NP2) were statistically significant in the four tests except PC and AC. The interactions between these two factors (i.e., *Multiple*:*Factor*) were statistically significant in five diagnostics except AC. These results indicated that some diagnostics were unable to pick up the subject of the given sentence(s).

In order to scrutinize which subjecthood tests had problems, each data sets (six groups) were divided into two groups (SSCs and MSCs) and the effects of each diagnostic were re-examined. The idea was straightforward: if the differences in terms of acceptability between Type 1 and Type 2 (represented in (3)) were statistically significant ($p<.05$), the

1) This results might be embarrassing to some people, since most books on statistics (including Gravetter and Wallnauc (2013) and Gries (2013)) mention that the data sets follow the normal distribution as the number of data increase. Though it is a general tendency, this tendency is rarely shown in the grammaticality judgment tasks (intuition tests) - because grammaticality/acceptability scores is usually negatively or positively skewed.

2) Someone might ask why ANOVA (ANalysis Of VAriance) tests were not applied here. The reason is simple. ANOVA is a parametric test which can be applied to the data set(s) with normal distributions; however, our experimental data did not follow the normal distribution. In the same reason, an ordinary linear regression was not applied which can be applied to the data set(s) with normal distributions: Instead, a GLM was applied with Gaussian distributions.

3) Here, * indicates $p<.05$, ** means $p<.01$, and *** implies $p<.001$.

diagnostics can be used for identifying the subject of SSCs. If not (.05<*p*), the diagnostics cannot be used for identifying the subject of SSCs. The same logic was also applied for MSCs: if the differences of acceptability between Type 3 and Type 4 were statistically significant (*p*<.05), the diagnostics can be used for identifying the subject of MSCs. If not (.05<*p*), the diagnostics cannot be used for identifying the subject of MSCs. The results from the analyses are as follows. In the table below, '○' indicates that the diagnostics can be used for identifying the subject of SSCs or MSCs, whereas '×' indicates that the diagnostics can be used for identifying the subject.

Table 2. SSCs vs. MSCs

Diagnostics	SSCs	MSCs
Honorific Agreement (HA)	○	×
Plural Copying (PC)	×	×
Obligatroy Control (OC)	○	○
Reflexive Binding (RB)	○	○
Coordinated Deletion (CD)	○	×
Adjunct Control (AC)	○	×

As you can observe, five diagnostics (except PC) correctly picked up the subject of SSCs, while only two (OC and RB) correctly picked up the subject of MSCs.[4)]

3.2 Cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis can be generalized to an independent data set (Geisser, 1993; Kohavi, 1995). It is mainly used in the settings where the goal is prediction, and is used to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (i.e., training data set), and a dataset of unknown data (or first-seen data set) against which the model is tested (i.e., testing data set).

Among the many cross-validation methods, 10-fold cross-validations are the most frequently-used one, whose schematic representation is shown in Figure 1.

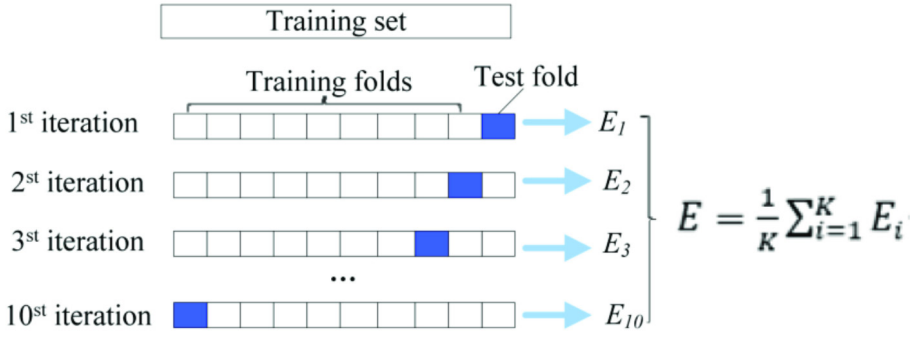

Fig. 1. 10-fold Cross-validation

4) For detailed comparisons of six subjecthood diagnostics, see Lee et al. (2015b).

The 10-fold cross-validation works as follows: First of all, the whole data set is divided into 10 parts. We will call each data set D_1, D_2, ..., D_{10} respectively. In the 1st iteration, nine data sets (D_1, D_2, ..., D_9) are used for training, and the other one (D_{10}) is used for testing, whose evaluation result is E_1 in Figure 1. In the 2nd iteration, nine data sets (D_1, D_2, ..., D_8, D_{10}) are used for training, and the other one (D_9) is used for testing, whose evaluation result is E_2. The iteration is conducted 10 times, whose evaluation results are E_1, E_2, ..., E_{10} respectively. Then, the overall performances are measured by taking mean of the 10 iterations (i.e., $E = (E_1, E_2, ..., E_{10})/10$).

In our experiments, since 20 target sentences were used for each subjecthood diagnostics and 70 participants were involved in the experiments, 1,400 data points (= 20 × 70) were obtained for each diagnostics. Then, each group of data were divided into two groups: SSCs and MSCs. Finally, 12 groups of data were prepared, where each group had 700 data points. Then, these 700 data points were divided 10 parts, which formed D_1, D_2, ..., D_{10} respectively: From this, 630 data points are used for training, and the other 70 data points are used for testing. Then, the overall performances are measured by taking mean of the 10 iterations (i.e., $E = (E_1, E_2, ..., E_{10})/10$). Based on this cross-validation analysis, we will use machine learning method to discuss the validity of each subjecthood diagnostic as the final results.

3.3 Machine Learning[5)]

Machine learning refers to "fields of study that gives computers the ability to learn without being explicitly programmed" (Samuel 1959). Mitchell (1997:2) defined it as follows: "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." Summarizing these two famous definitions, "machine learning can be defined as one field of Artificial Intelligence (AI) which makes a machine (here, a computer) automatically learn from the data (which can be referred to *E*) and perform some class of tasks *T* with the performance measure of *P*".

In the current paper, *E* corresponds to 9/10 of the data set, *T* to identifying the subject of the given sentence, and *P* to the mean accuracy of the *T* in the testing sets (1/10 of the data set). That is, in our paper, the computer learned how to identify the subject of the given sentence (*T*) based on 9/10 of the data set (*E*), and the performance was measured by the mean accuracy in 1/10 of the data set (*P*).

There are several different types of methods in machine learning: *k* nearest neighbors (*k*nn), naive Bayes, decision trees, (linear and non-linear) regressions, artificial neural networks (ANN), support vector machines (SVM), association rules, and so on. Each machine learning method has its pros and cons, and the choice of the method depends on the problems which must be solved. Among the several different machine learning methods, we chose a linear regression analysis for the current study. That is, a statistical model was constructed by applying a linear regression to the training data sets, and its accuracy was measured with the testing data sets using the prediction functions in a linear regression.

The reasons we chose the linear regression analysis were the following: First, a linear regression is the simplest and the most fundamental machine learning method: it is thus easy to computationally implement. Second, since it is a 'linear' regression, not a 'non-linear' regression, if the subjecthood diagnostics are statistically modelled with a linear regression, we can numerically measure how much a linear model can explain the given data. Then, the measurement

5) For machine learning and deep learning in the linguistic environments, see Lee et al. (2017).

implies that the differences between 100% and the mean accuracy of each diagnostics must be explained with a non-linear model. In a nutshell, if the subjecthood diagnostics are statistically modeled with a linear regression model, it is possible to measure the complexity of each subjecthood diagnostic through a machine learning method.

4. Results

The results of 10-fold cross-validation with a machine learning method (especially, a linear regression analysis) are as follows. Table 3 enumerates the numerical values of analysis results (accuracy, %) for different diagnostics.

Table 3. Numerical Results with 6 Diagnostics

Diagnostics	SSCs	MSCs
Honorific Agreement (HA)	68.04	49.38
Plural Copying (PC)	62.68	63.50
Obligatroy Control (OC)	85.13	70.89
Reflexive Binding (RB)	92.59	73.79
Coordinated Deletion (CD)	86.76	78.69
Adjunct Control (AC)	95.50	75.68

The mean accuracy in SSCs was 81.78%, and the value in MSC was 68.66%. From the above table, we can see that diagnostics such as HA, OC, RB, CD and AC show higher accuracy value with SSCs compared to MSCs, whereas PC show not much difference between the two conditions. As for the accuracy values of the 6 diagnostics with SSCs, AC and RB shows the highest accuracy values; and then, CD and OC, then HA and PC followed. On the other hand, with MSC conditions, CD and AC reported the highest two values; then RB, OC, PC come next in that order, and finally, HA reported the lower score compared to the other diagnostics. For the easy comparisons of the 6 diagnostics, the data in Table 3 were divided into the two groups (SSCs, MSCs), and were graphically represented as in Figure 2 and Figure 3.

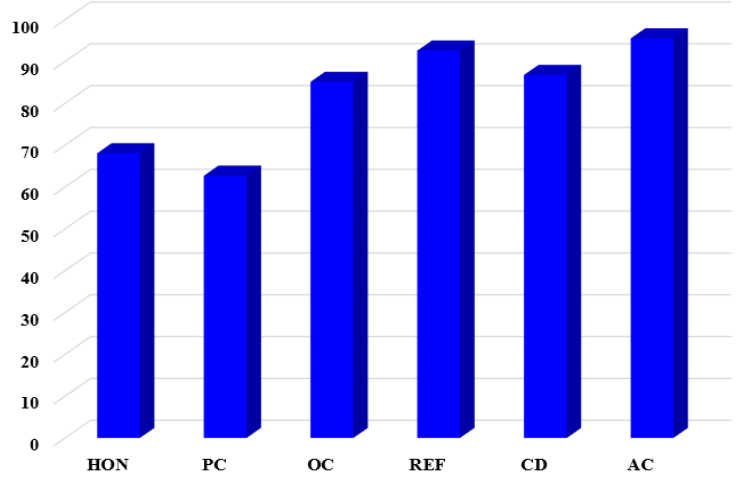

Fig. 2. Single Subject Constructions (SSCs)

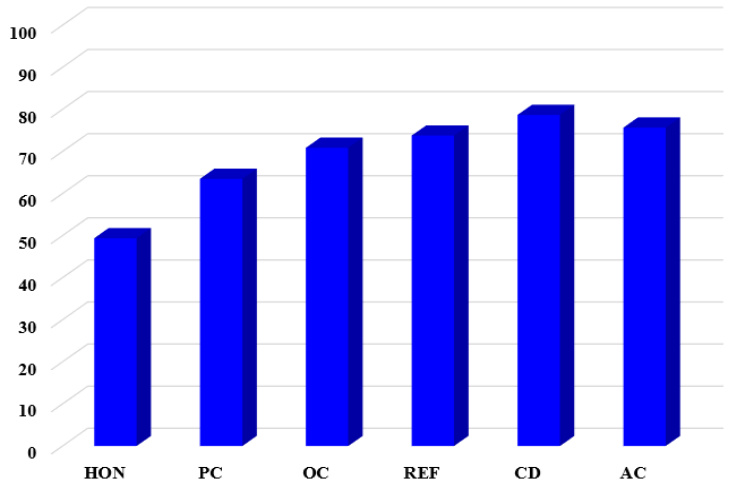

Fig. 3. Multiple Subject Constructions (MSCs)

As you can observe in Table 3 and Figure 2 & Figure 3, the performance rankings are AC>RB>CD>OC>HA>PC in case of SSCs and CD>AC>RB>OC>PC>HA in MSCs. In summary, AC in SSCs and CD in MSCs were ranked higher than the others, while PC in SSCs and HA in MSCs showed lower performance and accuracy compared to the other diagnostics.

5. Discussion

There have been many different types of subjecthood diagnostics proposed in the literature of Korean syntax. They explained how each subjecthood diagnostic picked up the subject in the sentence(s) and enumerated the properties of each subjecthood diagnostic. Although numerous studies were done on the subjecthood tests, only a few studies examined the properties of each subjecthood test with experimental approaches: Kim et al. (2015), Lee et al. (2015a), and Kim et al. (2017) took six subjecthood diagnostics (HA, PC, OC, RB, CD, and AC) and experimentally examined the properties of each subjecthood test and we used the data for those previous experiments for our current study of cross-validation with a machine learning.

In the current paper, a 10-fold cross-validation was taken with a machine learning method (especially, a linear regression analysis), and the validity of the experiments were examined. The results were shown in Table 3 and Figure 2 & Figure 3 in the earlier sections, which showed the performance rankings AC>RB>CD>OC>HA>PC in SSCs and CD>AC>RB>OC>PC>HA in MSCs. In summary, AC and CD showed relatively high performance, while HA and PC had relatively low performance.

An interesting fact is that HA and PC have been the most frequently mentioned subjecthood tests in some previous studies including Hong (1991, 1994), despite their low performance/accuracy in our results. The reason may be that honorific and plural markers are a few of the easily detected (morphological) properties in Korean. In addition, the examples used in HA and PC (such as (4) and (5a)) are typical enough to demonstrate the properties of these two diagnostics. However, following the experimental results in Kim et al. (2017), the results in the current study again revealed that HA and PC had low performance and low accuracy.[6)]

It is embarrassing that these two diagnostics HA and PC had low performance and low accuracy, since they were the most frequently mentioned subjecthood tests in previous studies. However, there is a (theoretical) possibility that the diagnostics work well for some typical examples but they become fuzzy in other sentences, which made some scholars argue that HA and PC were not valid subject diagnostics (Hong, 1994, etc.). Accordingly, the analysis results of this paper implies that a further close examination of the proposed subjecthood diagnostics is necessary .

What is also notable is that a linear model of subjecthood properties may not be enough to explain all the data related to subjecthood tests: It is natural that the accuracy of subjecthood diagnostics in MSCs is lower than that of SSCs. However, the coverage is too low: Especially, in the case of HA in MSCs, its mean accuracy was only 49.38, which means that HA can correctly identify only 49.38% of subjects in Korean sentences with MSCs. If we accept that this accuracy is almost close to the actual performance, it would be better to use coin-flipping in the subject test.

Then, what implies the difference between the ceiling (100%) and this value (49.38%)? In Section 3.3, it is mentioned that the differences between the ceiling and the mean accuracy of each diagnostic point to the area which must be explained with a non-linear model. That is, 50.62% of data can be explained with a non-linear model, in case of HA in MSCs. The same can be applied to the other diagnostics: The differences between the ceiling of the graph and the mean accuracy of each diagnostics imply that they should be treated in the area which must be accounted for with non-linear modeling of phenomena. Then, the subjecthood tests are supposed to be much more complex tasks than we originally supposed, which cannot be explained only with a simple linear model.

6. Conclusion

The current study took the experimental data from the three experimental studies (Kim et al., 2015; Lee et al., 2015; Kim et al., 2017) focusing on 6 subjecthood diagnostics in Korean proposed so far and re-examines the validity of six diagnostic tests - using a cross-validation and a machine learning method. The experimental data in three experiments were evenly divided into 10 parts (10-fold cross-validation). Nine of them were used for machine learning (training data set) and the remaining one for the testing (test data set). The analysis was conducted 10 times, and the mean prediction accuracies were taken. Among many machine learning techniques, a simple linear regression is taken. Through the machine learning and its analysis, the followings were observed: (i) The accuracy of the diagnostics for SSCs is much higher than that of MSCs (SSC: 81,78, MSC: 68.66), (ii) Adjunct Control is the most accurate for SSCs whereas Coordinated Deletion is for MSCs, (iii) HA and PC had low performance. The analysis results also seem to imply that there are some parts which cannot be explained by a simple linear model only, and they should be analyzed with non-linear modeling of language.

References

Andrews, A. 1985. The Major Functions of the Noun Phrases. *In T. Shoepen (ed.), Language Typology and Syntactic Description.* Cambridge, MA: Cambridge University Press. 132-223.

6) The same problem was also pointed out in Lee et al. (2015b).

Falk, Y. 2006. *Subjects and Universal Grammar*. Cambridge, MA: Cambridge University Press.

Geisser, S. 1993. *Predictive Inference*. New York: Chapman and Hall.

Gravetter, F. and L. Wallnau. 2013. *Statistics for the Behavioral Sciences*. Belmont, CA: Wadsworth.

Gries, S. Th. 2013. *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Guyter.

Heycock, C. and Y.-S. Lee. 1989. Subjects and Predication in Korean and Japanese. *Language Research* 25.4, 755-792.

Hong, K.-S. 1991. *Argument Selection and Case-Marking in Korean*. Doctoral dissertation, Stanford University.

Hong, K.-S. 1994. Subjecthood Tests in Korean. *Language Research* 30, 99-136.

Kang, B.-M. 2002. *Pemcwu Mwunpep: Hankwuke-uy Hyengthaylon, Thongsalon, Thaipnonlicek Uymilon* (*Categorial Grammar: The Morphology, Syntax, and Type-Logical Semantics of Korean*). Seoul: Korea University Press.

Kim, J.-H., E. Kim, and H.-S. J. Yoon. 2016. An Experimental Study of Subject Properties in Korean Multiple Subject Constructions (MSCs). In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation* (*PACLIC 2016*), 183-190. Seoul: Kyung Hee University.

Kim, J.-H., Y.-H. Lee, and E. Kim. 2015. Obligatory Control and Coordinated Deletion as Korean Subject Diagnostics: An Experimental Approach. *Language and Information* 19.1, 75-101.

Kim, J.-H., Y.-H. Lee, and E. Kim. 2017. Honorific Agreement and Plural Copying as Korean Subject Diagnostics: An Experimental Approach. *Modern Grammar Studies* 93, 119-144.

Kohavi, R. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2.12, 1137-1143. San Mateo, CA: Morgan Kaufmann.

Lee, I.-H. 1987. Double Subject Constructions in GPSG. *Harvard Studies in Korean Linguistics* II, 287-296.

Lee, Y.-H., E. Kim, and J.-H. Kim. 2015a. Reflexive Binding and Adjunct Control as Subject Diagnostics in Korean: An Experimental Approach. *Studies in Language* 31.2, 427-449.

Lee, Y.-H., J. H. Yu, and T.-J. Yoon. 2017. Predicting the Occurrence of the English Modals *Can* and *May* using Deep Neural Networks. *Studies in Modern Grammar* 96, 167-189.

Lee, Y.-H., Y. Park, and E. Kim. 2015b. A Multi-level Analysis of Subjecthood Diagnostics in Korean. *Linguistic Research* 32.3, 671-691.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill.

Park, K.-S. 1995. The semantics and pragmatics of case marking in Korean: A role and reference grammar account. Ph.D. Dissertation, State University of New York at Buffalo.

Park, B.-S. 1973. On the Multiple Subject Constructions in Korean. *Linguistics* 100, 63-76.

Samuel, A. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal* 3.3, 210-229.

Schütze, C. 2001. On Korean 'Case Stacking': The Varied Functions of the Particles *-ka* and *-lul*. *The Linguistic Review* 18, 193-232.

Teng, S-H. 1974. Double Nominatives in Chinese. *Language* 50, 455-473.

Yoon, J. H.-S. 1986. Some Queries Concerning the Syntax of Multiple Subject Constructions in Korean. *Studies in the Linguistic Sciences* 16, 215-236, Department of Linguistics, University of Illinois, Urbana-Champaign.

Yoon, J. H.-S. 2004. Non-nominative (Major) Subjects and Case-stacking in Korean. In Peri Bhaskararao and Karumuri VenkataSubbarao (eds.), *Non-nominative Subjects*, volume 2, 265-314. Mouton de Gruyter, Berlin.

Yoon, J. H.-S. 2007. Raising of Major Arguments in Korean and Japanese. *Natural Language and Linguistic Theory* 25, 615-653.

Yoon, J. H.-S. 2008. Subjecthood and Subject Properties in Multiple Subject Constructions. *Talk presented at the East Asian Linguistics Seminar.* Oxford: Oxford University.

Yoon, J.-Y. 1989. On the Multiple -*ka* and -*lul* Constructions in Korean. *Harvard Studies in Korean Linguistics* III, 383-394.

Youn, C. 1990. *A Relational Analysis of Korean Multiple Nominative Constructions*. Doctoral dissertation, State University of New York at Buffalo.

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

Kim, Ji-Hye, Professor
250 Taeseongtapyeon-ro, Heungdeok-gu, Cheongju 28173, Republic of Korea
Department of English Education, Korea National University of Education
E-mail: jkim@knue.ac.kr