Linguistic Research 38(1): 75-98
DOI: 10.17250/khisli.38.1.202103.004

A multifactorial approach to Korean -(*u*)*m* and -*ki**

Yong-hun Lee · Gihyun Joh*****
(Chungnam National University · Kunsan National University)

Lee, Yong-hun and Gihyun Joh. 2021. A multifactorial approach to Korean -(*u*)*m* and -*ki*. *Linguistic Research* 38(1): 75-98. This paper takes a corpus-based approach and examines the linguistic properties of two Korean nominalizers -(*u*)*m* and -*ki*. From the Sejong Treebank corpus, all the sentences with -(*u*)*m* and -*ki* are extracted. Twenty linguistic factors are manually encoded into the extracted sentences. Then, all the encoded data are statistically analyzed with (binary) logistic regression. Although we take a monofactorial analysis, we obtain a good statistical model whose *C* value is 0.956. Through the analysis, the followings are observed: (i) -(*u*)*m* and -*ki* are used with the ratio of 1:9 in Korean, (ii) among twenty linguistic factors, only ten factors are statistically significant, and (iii) not only the verbs which take -(*u*)*m* and -*ki* as a complement but also the verbs which merge with these two nominalizers also play important roles in the determination of nominalizers. **(Chungnam National University·Kunsan National University)**

Keywords Korean nominalizer, Sejong corpus, treebank, logistic regression, effect plots

1. Introduction

It is known that -(*u*)*m* and -*ki* are two nominalizers in Korean (Nam and Ko 1993; Suh 1996; Sohn 1999; Song 2005) but that their syntactic distributions and semantic properties are known to be different. Several previous studies have stated the differences of these two nominalizers and linguistic factors which influence the differences (Choe 1937; Chang 1966; Kim 1973; Im 1974; Yang 1975; Chae 1979; Sim 1980; Hong 1983; Kim 1984; Woo 1987). However, there are few studies which examine them with corpus data.

Recently, as corpus and corpus linguistics develop, there have been several trials to examine syntactic phenomena using corpus-based approaches and their statistical analyses. For example, Deshors (2015) and Deshors and Gries (2016) investigate the alternations

* 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

of English gerunds and *to*-infinitives, which are similar to -(*u*)*m* and -*ki* in Korean respectively.[1]

This paper takes similar methods in Deshors (2015) and Deshers and Gries (2016) and scrutinizes the syntactic and semantic properties of -(*u*)*m* and -*ki*, using a corpus-based approach and its statistical analysis. The goal of the current study is (i) to construct a statistical model which is based on corpus data on -(*u*)*m* and -*ki* and (ii) to investigate which linguistic factors are involved in the determination of alternations. For this purpose, the Sejong Treebank corpus is chosen. In this corpus, all the sentences with -(*u*)*m* and -*ki* are extracted. Then, twenty linguistic factors are manually encoded. The encoded data are statistically analyzed with a (binary) logistic regression

This paper is organized as follows. Section 2 introduces previous studies (i) on -(*u*)*m* and -*ki* and (ii) on corpus-based studies on English gerunds and *to*-infinitives. Section 3 provides accounts on corpus data and research method. Section 4 enumerates the analysis results with tables and plots. Section 5 includes discussions, and Section 6 summarizes this paper.

2. Previous studies

2.1 On Korean -(*u*)*m* and -*ki*

It is known that Korean verbal elements (verbs and adjectives) can merge with either -(*u*)*m* or -*ki*.[2] The following examples illustrate this fact (Suh 1996: 1318).[3]

---

1 As mentioned in Quirk et al. (1985), Celce-Murcia et al. (1999), and Huddlestone and Pullum (2003), gerunds have a past- oriented meaning and *to*-infinitives have a future-oriented meaning in English. As you will find in (1), -(*u*)*m* has a past-oriented meaning and -*ki* has a future-oriented meaning in Korean. These semantic properties are similar to (English) gerunds and *to*-infinitives respectively. Many previous studies in Table 1 pointed out that -(*u*)*m* is similar to gerunds and -*ki* is similar to *to*-infinitives.

2 Unlike English, not only a verb but also an adjective can independently be a predicate in Korean. Accordingly, the term *verbal* is used in this paper, which includes both a verb and an adjective.

3 As usual, Yale Romanization system is used for Korean sentences, and the following abbreviations are used here: ACC for accusative, DECL for declarative, GEN for genitive, LOC for locative, NMN for nominalizer, NOM for nominative, PASS for passive, PAST for past tense, and TOP for the topic marker.

(1) a. *Ce oykwukin-i* *kimchi-lul* *cal* ***mek-um-i***
that foreigner.NOM kimchi.ACC very-well eat.NMN.NOM
ali-eci-ess-ta.
know.PASS.PAST.DECL
'It was known that the foreigner ate kimchi very well.'
b. *Oykwukin-i* *kmichi-lul* ***mek.ki-ka*** *elyep-ta*.
foreigners.NOM kinchi.ACC eat.NMN.NOM be-difficult.DECL
'It is difficult that foreigners eat kimchi.'

In these two sentences, Korean nominalizer -(*u*)*m* and -*ki* are attached to the same verb *mek-ta* 'eat', both in (1a) and in (1b). The nominative Case marker -*i*/-*ka* shows that both *mek-um* and *mek-ki* are nouns. Note that the nominalizer -(*u*)*m* and -*ki* can be attached to the same verb *mek-ta* 'eat' in (1a) and (1b) respectively and that both sentences are grammatical.

However, not all the verbal elements can take both -(*u*)*m* and -*ki* in Korean. The following sentences demonstrate this fact (Suh 1996: 1324).

(2) a. *Kicha-ka* *pangkum* ***cinaka-m-ul***/****cinaka-ki-lul*** *po-ass-ta*.
train.NOM right-now pass.NMN.ACC see-PAST.DECL
'I saw the train's passing right now.'
b. *Kutul-un* *selo* ?***manna-m-ul***/***manna-ki-lul*** *coaha-(y)ss-ta*.
they.TOP one-another meet.NMN.ACC like.PAST.DECL
'They like to meet each other/one another.'

In (2a), *cinaka-m* is natural, whereas *cinaka-ki* is impossible. In (2b), *manna-m* may possible in some specific situations, but it seems weird. On the other hand, *manna-ki* is perfectly possible. As you can find in these sentences, the distributions of -(*u*)*m* and -*ki* are not identical.

Many studies have mentioned the syntactic distributions and semantic differences between these two nominalizers, and they are summarized in the following table.

Table 1. Previous studies on Korean -(*u*)*m* and -*ki*

	-(*u*)*m*	-*ki*
Choe (1937)	conceptual	physical
Chang (1966)	abstract, qualitative	concrete, quantitative
Kim (1973)	internal, implicit	external, explicit
Im (1974)	[+existence], [+target]	[-existence], [-target]
Yang (1975)	factual	expectational
Chae (1979)	only once, individual	repeated, general
Sim (1980)	[+decisive], [+substance]	[-decisive], [-substance]
Hong (1983)	instantaneous, past-oriented	eternal, future-oriented
Kim (1984)	factual	nonfactual
Woo (1987)	individual, substantial	general, hypothetical

That is, the verbals with the meanings in the second column prefer the nominalizer -(*u*)*m* and the verbals with the meanings in the third column prefer to take -*ki*.

As you can observe in this table, Korean -(*u*)*m* and -*ki* are semantically similar to gerunds and *to*-infinities in English respectively. For example, many books in traditional grammar mentioned that gerunds have a past-oriented meaning but that *to*-infinitives have a future-oriented meaning (Quirk et al. 1985; Celce-Murcia and Larsen-Freeman 1999; Huddleston and Pullum 2003; Cowan 2008). Let's see the following sentence (Cowan 2008: 506).

(3) a. John remembered *mailing* the letter.
b. John remembered *to mail* the letter.

In (3a), the action of *mailing* occurred before remembering. In (3b), however, the action of *to mail* was remembered before it was carried out. Accordingly, we can say that (3a) is past-oriented whereas (3b) is future-oriented. These kinds of properties are very similar to -(*u*)*m* and -*ki* in Korean, as Hong (1983) has already pointed out. For example, in (2a), the event of *manna-m* 'meet' has already occurred before the action of *po-ta* 'see'. In (2b), however, the event of *manna-ki* 'meet' does not occur before the action of *coaha-ta* 'like'. Consequently, it can be said that (2a) has a past-oriented meaning but that (2b) has a future-oriented reading.

2.2 Corpus-based approaches to English gerunds and *to*-infinitives

Recently, as corpus and corpus linguistics develop, there have been several trials to examine syntactic phenomena using corpus-based approaches and their statistical analyses. Among them, this section introduces two studies which are directly related to the current study.

Deshors (2015) employed a multifactorial analysis and compared natives' English with non-natives'. She used two components of the International Corpus of English (ICE; Greenbaum 1996) to analyze gerunds and *to*-infinitives in English. One is the USA component of ICE (ICE-USA) for natives, and the other is the Hong Kong components (ICE-HK) for non-natives. From these corpora, she extracted 3,119 sentences with these two constructions and manually encoded the following nine linguistic factors (Deshors 2015: 218).

Table 2. Encoded linguistic factors in Deshors (2015)

Factor	Factor levels
ComplPattern (complementation pattern; dependent factor)	gerund, infinitive
Country (English variety)	usa, hong kong
MatrixVerbForm (form of the matrix verb)	finite (active), finite (passive), non-finite (passive), non-finite (active)
MatrixVerbsem (semantics of the matrix's lexical verb)	abstract, action, communication/informational, copula, cognitive/emotional, perception
CompVerbsem (semantics of the complement's lexical verb)	abstract, action, communication/informational, copula, cognitive/emotional, perception
MatrixVerbtype (type of the matrix verb)	state, accomplishment, achievement, process
CompVerbType (type of the complement verb)	state, accomplishment, achievement, process
Neg	neg, affirm
ObjectForm (form of the object)	PP (prepositional phrase), NP (noun phrase), DO (double object), PR (pronoun), NO (no object)

Then, the study applied a generalized linear model with logistic regression to identify which factors played a statistically significant role in the alternations of gerunds and *to*-infinitives. She also utilized a bootstrapping analysis to assess the reliability of the statistical model. Through the analyses, the study found that both groups of native and

non-native speakers made their decisions based on complex grammatical contexts rather than isolated syntactic/semantic environments. She also observed that different linguistic factors played roles in a particular complement type.

In Deshors and Gries (2016), on the other hand, they expanded the scope of the study to the Englishes in several Asian countries. They adopted five ICE components in their analysis. ICE-GB and ICE-USA were for natives, and the Hong Kong, India, and Singapore components of ICE (ICE-HK, ICE-IND, and ICE-SIN respectively) were for non-natives. In this study, instead of nine linguistic factors in Deshors (2015), twelve factors were manually encoded, and they were analyzed with logistic regression. Through the analysis, they showed that there were a variety of differences between the Asian and the native Englishes. They also found that the Asian Englishes were more similar to American English (ICE-USA), rather than British English (ICE-GB). They also employed a Multifactorial Prediction and Deviation Analysis (MuPDAR) to examine how much non-native speakers' English deviated from the natives. Through the analysis, they demonstrated that MuPDAR was a valuable analytical tool which could be utilized to the messy and skewed corpus data.

As mentioned in Table 1, many previous studies investigated the linguistic properties of -(*u*)*m* and -*ki*, and the properties of these two nominalizers are similar to gerunds and *to*-infinitives respectively. Since the linguistic factors in Table 2 were used to examine distributions of gerunds and *to*-infinitives in English, it is reasonable to test if the factors in Table 2 can be applied to investigate the distributions of -(*u*)*m* and -*ki* in Korean. When the factors in Table 2 are applied to corpus data in Korean, however, language-specific properties also have to be considered.

3. Research method

3.1 Corpus

In the studies of the Korean language, one of the most frequently-used corpora is the *Sejong* Corpus (National Institute of Korean Language 2007), which is sometimes called the Korean National Corpus. The corpus is composed of several subcorpora which include raw text corpora, POS-tagged corpora, syntactically-parsed corpora, historically-annotated corpora, and so on.

Originally, we chose the Sejong POS-tagged corpus to extract all of the sentences with -(*u*)*m* and -*ki*, whose corpus size is about 800,000 *ejeol* (word tokens in English corpora). However, when all the relevant sentences were extracted from this corpus, the number of extracted sentences was too big to be handled. It was very difficult to manually weed out the false positives from the extracted data. Thus, we moved to the Sejong Treebank corpus instead, which was smaller than the Sejong Morphologically-annotated corpus.[4] The Sejong Treebank corpus includes 433,839 *ejeol* and 43,828 sentences, and it is a syntactically-parsed corpus. Thus, it is also possible to use morphological information. In order to extract all of the sentences with -(*u*)*m* and -*ki* from the corpus, we have to utilize the morphological or syntactic information.

3.2 Analysis procedure

Our analysis proceeded as follows. From the Sejong Treebank corpus, all of the sentences with -(*u*)*m* and -*ki* were extracted using the tag ETN, where ETM was a tag for nominalizer. Then, the extracted sentences were manually encoded with twenty linguistic factors in Table 3. After the encoding process, all the data were statistically analyzed with a (binary) logistic regression analysis.

The following schema was applied during the analysis.

(4) Basic Schema of -(*u*)*m* and -*ki*

a. -(*u*)*m*

[…] (S_1) MVerb […] (S_2) CVerb-(*u*)*m*

[…] (S_2) CVerb-(*u*)*m* […] (S_1) MVerb

b. -*ki*

[…] (S_1) MVerb […] (S_2) CVerb-*ki*

[…] (S_2) CVerb-*ki* […] (S_1) MVerb

Here, *MVerb* refers to the matrix verb which takes -(*u*)*m* or -*ki* as a complement, and CVerb to the complement verb which merges with the nominalizer -(*u*)*m* or -*ki*. Likewise, S_1 indicates the subject of the matrix verb, and S_2 is the subject of the complement verb.

4 Note that it is nearly impossible to extract the sentences with -(*u*)*m* and -*ki* from the raw text subcorpus of the Sejong corpus, since (i) the number of extracted sentences is too big and (ii) the raw texts are neither morphologically nor syntactically annotated.

If the schema is applied to the two sentences in (1), the results are as follows. In (1a), *ali-eci-ta* 'be known' becomes an MVerb and *mek-ta* 'eat' is a CVerb. Likewise, in (1b), *elyep-ta* 'be difficult' is an MVerb and *mek-ta* 'eat' is a CVerb.

3.3 Encoded factors

Although the linguistic factors in Table 2 were also useful to Korean, some factors were added or deleted as in Table 3.

Table 3. Encoded linguistic factors

Type	Variable	Comment	Levels
Morphological	MVPOS	POS of Matrix Verbal Expression	adjective, verb
	CVPOS	POS of Complement Verbal Expression	adjective, verb
	MVComplex	Complexity of Matrix Verbal Expression	no, yes
	CVComplex	Complexity of Complement Verbal Expression	no, yes
	MVSinoKorean	Sino-Korean of Matrix Verbal Expression	no, yes
	CVSinoKorean	Sino-Korean of Complement Verbal Expression	no, yes
	MVHaToy	*-Ha*/*-Toy* of Matrix Verbal Expression	ha, no, toy
	CVHaToy	*-Ha*/*-Toy* of Complement Verbal Expression	ha, no, toy
	MVVerbal	Verbal Noun of Matrix Verbal Expression	no, yes
	CVVerbal	Verbal Noun of Complement Verbal Expression	no, yes
	MVVoice	Voice of the Matrix Verbal Expression	active, passive
	CVVoice	Voice of the Comp. Verbal Expression	active, passive
Syntactic	MVTransitivity	Valency of the Matrix	ditransitive, intransitive,

		Verbal Expression	transitive
	CVTransitivity	Valency of the Complement Verbal Expression	ditransitive, intransitive, transitive
	MVCompForm	Complement Form of the Matrix Verbal Expression	cp, do (double object), no (no object), np, np-np, pp, pp-np, pr (pronoun), tp
	CVCompForm	Complement Form of the Complement Verbal Expression	cp, do (double object), no (no object), np, np-np, pp, pp-np, pr (pronoun), tp
Semantic	MVType	Vendler's Classification	accomplishment, achievement, process, state
	CVType	Vendler's Classification	accomplishment, achievement, process, state
	MVSem	Matrix Verbal Semantics	abstract, concrete
	CVSem	Complement Verbal Semantics	abstract, concrete

Among the factors, some factors were specific to Korean. For example, MVComplex and CVComplex indicated whether the matrix verb or the complement verb was a single verb (such as *o-ta* 'come' or *ka-ta* 'go') or a complex verb (such as *o-ka-ta* 'come and go'). MVSinoKorean and CVSinoKorean implied whether the verbals had a Sino-Korean origin or not. MVHaToy and CVHaToy stated whether the matrix verb or the complement verb was merged with *ha-ta* 'do' or *toy-ta* 'become'. MVVerbal and CVVerbal said whether the verbal elements were originated from the verbal nouns or not. All of these linguistic factors were not necessary to the study of English gerunds and *to*-infinitives, but they were necessary to describe the linguistic phenomena in Korean.[5]

3.4 On the statistical analysis

In the actual statistical analysis, this paper primarily employed a binary logistic regression using R (R Core Team 2020). For regression analysis, Deshors (2014: 11) stated that "[b]inary logistic regression is a confirmatory statistical technique that allows the analyst to identify possible correlations between the dependent and the independent factor/variables." In this paper, since the dependent variable had one of two values (i.e., -(*u*)*m* or -*ki*), a binary logistic regression was taken.

5 As you may find, all the factors in Table 3 are related to the verbal elements. There may be some linguistic factors which are related to the subject S_1 or S_2. However, they are not included in Table 3 for analysis convenience.

A random forest analysis is both a statistical and machine-learning method, which can be used for various tasks including classification, regression, and other types of statistical analyses. This analysis usually starts by constructing a multitude of decision trees at training time and produces as an output the class which is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Breiman 2001; Hastie et al. 2008). The reason why this analysis was taken was that the importance of each variable could be calculated during the analysis, and the outputs of variable importance were utilized.

4. Analysis results

4.1 Descriptive statistics

The following plot shows us the basic distributions of -(*u*)*m* and -*ki*.

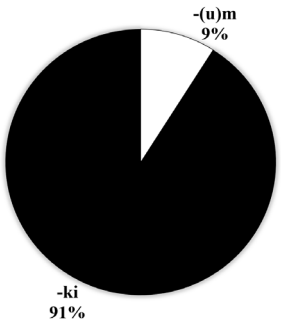

Figure 1. Distributions of -(*u*)*m* and -*ki* in Korean

As you can see, the nominals with -*ki* occupied more than 90%, but the nominals with -(*u*)*m* occupied slightly less than 10%. It implied most nominals in Korean were merged with the nominalizer -*ki*, rather than -(*u*)*m*.

4.2 Analysis with logistic regression

The first step for the logistic regression is to construct an initial model. After a milticollinearity test, the initial model was constructed as follows.[6]

Table 4. Initial model

CPattern~CVPOS+CVComplex+CVSinoKorean+CVHaToy+CVVerbal+CVVoice+CVTransitivity +CVCompForm+CVType+CVSem+MVPOS+MVComplex+MVSinoKorean+MVHaToy+MVVer bal+MVVoice+MVTransitivity+MVCompForm+MVType+MVSem

As you could notice, this model contained no interaction among the twenty linguistic factors. It implied that only the effects of the main factors were considered in the statistical analysis.

After the construction of this model, we applied a backward-selection process and obtained the following final model.

Table 5. Final (optimal) model

CPattern~CVTransitivity+CVCompForm+CVType+MVPOS+MVComplex+MVSinoKorean+MV HaToy+MVCompForm+MVType+MVSem

Among the twenty linguistic factors, only ten factors survived in the final model. Among these ten factors, only three factors were related to CVerbs, and the other seven factors belonged to MVerbs.

The following table demonstrates the result of the model selection process with detailed statistics.

6 A multicollinearity test is a statistical test to examine a state of very high intercorrelation or inter-associations among the independent variables (Baayen 2008). That is, the test checks whether (or not) the effect/influence of a certain factor overlap with those of other factors. Because this paper employed twenty factors, a multicollinearity test is a prerequisite. Usually, Variance Inflation Factors (VIF) is adopted in the test, and there was no factor whose VIF value was more than 5.

Table 6. Result of model selection

	df	Deviance	AIC	LRT	*p*	
<none>	201.41	267.41				
CVPOS	1	201.88	265.88	0.470	0.493	
CVComplex	1	201.41	265.41	0.000	0.996	
CVSinoKorean	1	201.60	265.60	0.188	0.665	
CVHaToy	2	203.79	265.79	2.376	0.305	
CVVerbal	1	201.70	265.70	0.287	0.592	
CVVoice	1	201.66	265.66	0.246	0.620	
CVTransitivity	2	207.80	269.80	6.386	0.041	*
CVCompForm	4	212.88	270.88	11.463	0.022	*
CVType	3	215.14	275.14	13.731	0.003	**
CVSem	2	203.71	265.71	2.297	0.317	
MVPOS	1	208.48	272.48	7.073	0.008	**
MVComplex	1	225.62	289.62	24.208	<0.001	***
MVSinoKorean	1	214.20	278.20	12.789	<0.001	***
MVHaToy	1	208.32	272.32	6.911	0.009	**
MVVerbal	1	202.48	266.48	1.072	0.301	
MVVoice	1	201.41	265.41	0.000	0.999	
MVTransitivity	1	201.43	265.43	0.022	0.881	
MVCompForm	3	211.80	271.80	10.388	0.016	*
MVType	2	251.30	313.30	49.887	<0.001	***
MVSem	1	216.06	280.06	14.650	<0.001	***

The upper part was for CVerbs, and the lower part was for MVerbs. The linguistic factors with * were the factors with statistical significance.

4.3 Analysis with effect plots

In this section, we closely examine the significant factors with effect plots. The following is the effect plot for CVTransivity. Here, the *x*-axis enumerates the value of each factor (i.e., the values under 'Level' in Table 3), and the *y*-axis shows the predicted probability of using -(*u*)*m* in the given situation. The I-shaped lines above and below the dot indicate the 95% confidence intervals.[7]

7 For example, the predicted probability for *intransitive* was about 0.2 (slightly above 0.2). It said that about 20% of nominals were constructed with -(*u*)*m* when the CVerb was an *intransitive* verb. It implied that about 80% of nominals were constructed with -*ki* in the same situation (when the CVerb was an *intransitive* verb). That is, in each plot, if the predicted probability of -(*u*)*m* was *p*, then the predicted probability of -*ki* would be 1-*p*.

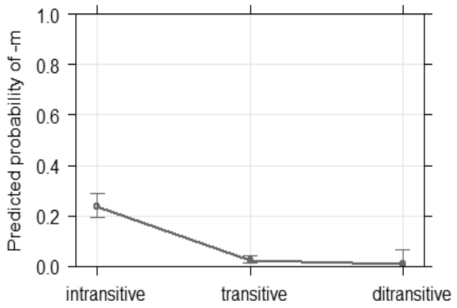

Figure 2. Effect plot for CVTransivity

When CVerb was a *ditransitive* or a *transitive* verb, -*ki* was applied in most cases. When CVerb was an *intransitive* verb, -*ki* and -(*u*)*m* were utilized with the ratio of 4:1, and the differences were statistically significant. It implied that -(*u*)*m* were applied more when CVerb was an *intransitive* verb.

The following plot demonstrates the plot for CVCompForm.[8]

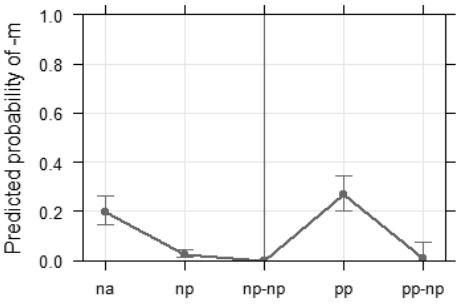

Figure 3. Effect plot for CVCompForm

When CVerb took *np*, *np-np*, or *pp-np* as a complement, -*ki* was employed in most cases. When CVerb took *na* or a *pp* complement, -*ki* and -(*u*)*m* were used with the ratio of 4:1, and the differences were statistically significant. It implied that -(*u*)*m* were employed more when CVerb took *na* or a *pp* complement.[9]

---

8 Here, *na* implies that the verbs do not take any complement (i.e., when the verb is an intransitive verb). Also note that the 95% confidence interval for *np-np* goes from 0 to 1. The reason is that there are only a few cases where the verb takes *np-np* complements.

9 The factor CVCompForm seemed related to the factor CVTransitivity. However, there was no

The following plot illustrates the plot for CVType.

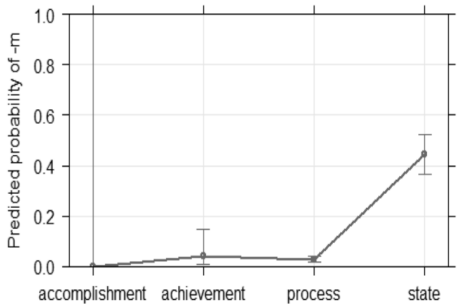

Figure 4. Effect plot for CVType

When CVerb was an *accomplishment*, an *achievement*, or a *process* verb, *-ki* was hired in most cases. When CVerb was a *state* verb, the ratio of *-ki* and *-(u)m* were close to 1:1, and the differences were statistically significant. It implied that *-(u)m* were hired more when CVerb was a *state* verb.

The following plot presents the plot for MVPOS.

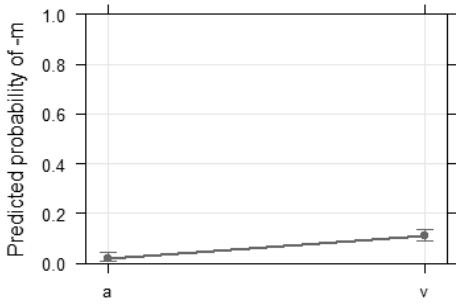

Figure 5. Effect plot for MVPOS

When MVerb was an *adjective*, *-ki* was used in most cases. When CVerb was a *verb*, *-ki* and *-(u)m* were employed with the ratio of 9:1, and the differences were statistically significant. It implied that *-(u)m* were used more when CVerb was a *verb*.

multicollinearity between these two factors. That is why the factor CVCompForm was included in the final model.

The following shows the plot for MVComplex.

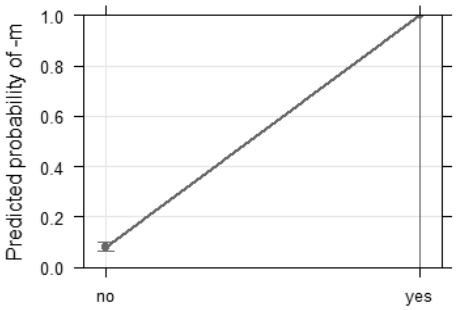

Figure 6. Effect plot for MVComplex

When MVerb was a complex verb (*yes*), -(*u*)*m* was utilized in most cases. When MVerb was not a complex verb (*no*), -*ki* were applied in most cases, and the differences were statistically significant. It implied that the distributions of -(*u*)*m* and -*ki* were clearly different depending on the complexity of MVerb.

The following plot is the plot for MVSinoKorean.

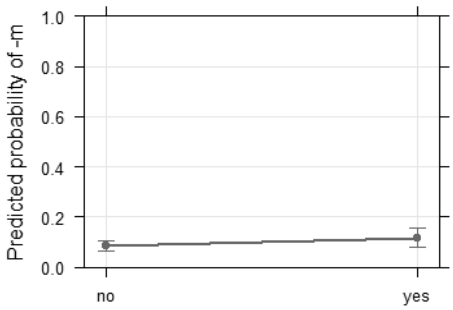

Figure 7. Effect plot for MVSinoKorean

The factor MVSinoKorean did not have a massive influence on the distributions of -*ki* and -(*u*)*m*. However, -(*u*)*m* were utilized more when MVerb had a Sino-Korean origin (*yes*), and the differences were statistically significant. It implied that -(*u*)*m* were applied more when MVerb had a Sino-Korean origin.

The following plot demonstrates the plot for MVHaToy.

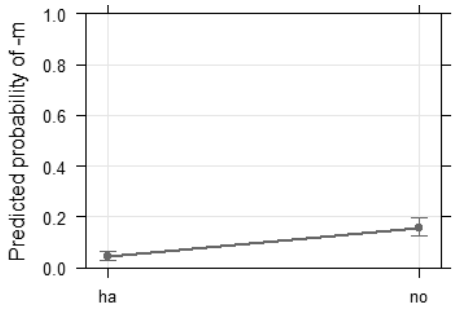

Figure 8. Effect plot for MVHaToy

The factor MVSinoKorean also did not have a massive influence on the distributions of *-ki* and -(*u*)*m*. However, -(*u*)*m* were hired more when MVerb did not merge with *ha-ta* 'do' or *toy-ta* 'become', and the differences were statistically significant. It implied that -(*u*)*m* were applied more when MVerb did not merge with *ha-ta* or *toy-ta*.

The following plot illustrates the plot for MVCompForm.[10]

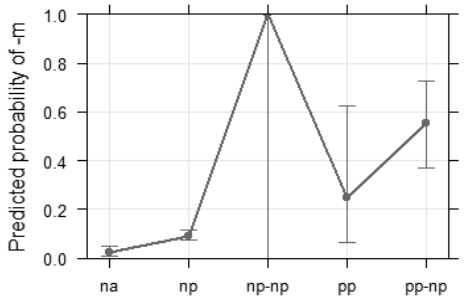

Figure 9. Effect plot for MVCompForm

When MVerb took *na* or an *np* complement, *-ki* was employed in most cases. When MVerb took a *pp* complement, *-ki* and -(*u*)*m* were used with the ratio of 7:3. When MVerb took a *pp-np* complement, *-ki* and -(*u*)*m* were used with the ratio of 4:6. When MVerb took *np-np* complement, -(*u*)*m* was employed in most cases.

The following plot presents the plot for MVType.[11]

10 Note that the nominals with -(*u*)*m* or *-ki* can be an NP. Therefore, the *np*s in this plot included not only the general NPs bot also the nominals with -(*u*)*m* or *-ki*.

11 There was no verb whose MVType was an *accomplishment* in our data.

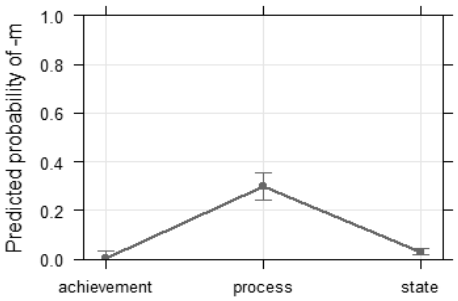

Figure 10. Effect plot for MVType

When MVerb was an *achievement* or a *state* verb, -*ki* was used in most cases. When MVerb was a *process* verb, the ratio of -*ki* and -(*u*)*m* was 7:3, and the differences were statistically significant. It implied that -(*u*)*m* were hired more when MVerb had a *process* meaning.

The following plot shows the plot for MVSem.

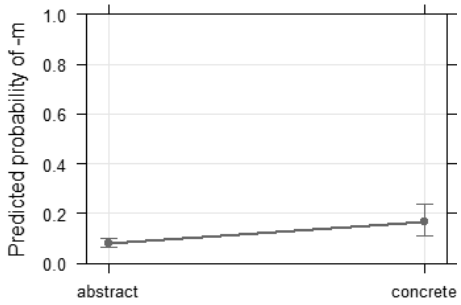

Figure 11. Effect plot for MVSem

The factor MVSem did not have a massive influence on the distributions of -*ki* and -(*u*)*m*. However, -(*u*)*m* were utilized more when MVerb had a *concrete* meaning, and the differences were statistically significant. It implied that -(*u*)*m* were applied more when MVerb did not merge with *concrete* meaning.

4.4 Goodness of fit

After we got the final model, we verified the goodness of fit of the model, and the results were as follows.

Table 7. Goodness of fit statistics

Statistics	Value
Log-likelihood Ratio	340.01
p-value	<0.0001
R^2	0.623
Classification Accuracy	91.51
C	0.956
D_{xy}	0.911

For the *C*-values, Harrell (2001: 248) mentioned that "*C*-values range from 0.5 to 1 and the higher the value, the better a regression model is at classifying or predicting the dependent variable; *C*-values ≥0.8 are generally considered good." Note that the *C*-value for our final model is 0.956. This suggested that our statistical model was excellent for explaining the similarities and differences between -(*u*)*m* and -*ki*, even though our model included no interaction among the linguistic factors.

4.5 Analysis with random forest

The following plot shows us the variable importance of ten linguistic factors, which had a statistical significance in Table 6. This plot was obtained from random forest analysis.

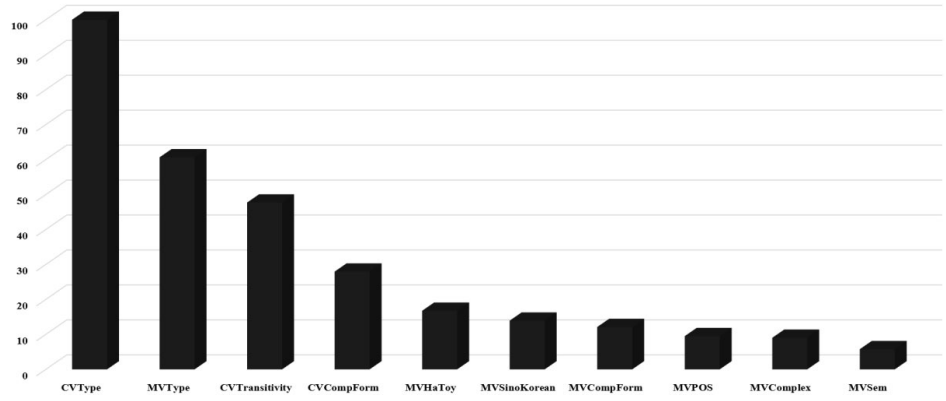

Figure 12. Variable importance of ten linguistic factors

In this plot, the importance of CVType was set to be 100 (i.e., 100%), and the importance of the other nine factors was calculated relative to the values of the factor CVType.

5. Discussion

This paper took a corpus-based method and investigated the distributions of -(*u*)*m* and -*ki* in the actual corpus data. This paper also took a statistical method and examined which linguistic factors played important roles in the determinations of -(*u*)*m* and -*ki* in Korean.

In the distributional properties, the ratio of -*ki* and -(*u*)*m* was about 9:1, which was an asymmetric distribution. This was an unexpected result. When we think of linguistic alternations (such as *can* vs. *may* in English; Deshors 2014), we usually suppose that the ratio of the alternations will be similar to each other or one another, though the alternations do not occur with the exact same ratio. However, the ratio of 9:1 in the distributions of -*ki* and -(*u*)*m* was extremely asymmetric. This ratio of distributions may imply that -*ki* is an unmarked and a default form in the Korean norminalization, although we cannot say that -(*u*)*m* is derived from -*ki*.

From the statistical analysis of -(*u*)*m* and -*ki* in Section 4, the following things were revealed.

First, the statistical analysis in Table 5 and Table 6 demonstrated that not only MVerbs but also CVerbs played an important role in the determination of -(*u*)*m* and -*ki* in Korean. It implied that the studies on the CVerbs were also necessary, in addition to the investigations on the MVerbs, to uncover the exact properties of -(*u*)*m* and -*ki*.

Second, even though both MVerbs and CVerbs were important in the choice of alternations, MVerbs had more factors which had statistical significance. Note that, in Table 5 and Table 6, three factors were statistically significant in the CVerbs (CVTransitivity, CVCompForm, and CVType) but that the other seven three factors were significant in the MVerbs (MVPOS, MVComplex, MVSinoKorean, MVHaToy, MVCompForm, MVType, and MVSem). It implied that MVerbs played more important roles than CVerbs in the choice of alternations.

Third, the linguistic properties which were specific to the Korean language also played important roles in the determination of -(*u*)*m* and -*ki*. Note that, in Table 5 and Table 6, three factors (MVComplex, MVSinoKorean, and MVHaToy) had a statistical significance. These factors were not included in the encoding of Table 1, since English has no property of Sino-Korean or *ha-ta*/*toy-ta*. This paper had a contribution in that the current studies statistically examined how these kinds of Korean-specific factors played roles in the choice of alternations.

The factor MVSem needed more close examination. In Table 1, there was a contradiction in the theoretical properties of -(*u*)*m* and -*ki*. Chang (1966) mentioned that -(*u*)*m* merges with the verbs which have an abstract reading. whereas -*ki* merges with the verbs which have a concrete meaning. On the other hand, Woo (1987) said that -(*u*)*m* merges with the verbs which have a substantial meaning. whereas -*ki* merges with the verbs which have a hypothetical reading. Figure 11 demonstrated that, although the factor MVSem also did not have a massive influence on the distributions of -*ki* and -(*u*)*m*, -(*u*)*m* were utilized more when MVerb had a *concrete* meaning. That is, our corpus data supported that Woo's observations (1996) were closer to the actual data, rather than Chang's claim (1966).

The *C* statistics of our optimal model was 0.956, which was much higher than the threshold value in Harrel (2001). It implied that our statistical model was excellent for explaining the similarities and differences between -(*u*)*m* and -*ki*, even though our model included no interaction among the linguistic factors. Also note that the classification accuracy was 91.51. It indicated that our statistical model could classify -(*u*)*m* and -*ki* with more than 90% of accuracy, along with a 'linear' regression model. It implied that

the distributions of -(*u*)*m* and -*ki* might be a linear problem which could be handled and solved with a linear model.

In Figure 12, the following things could be observed. First, the linguistic factors associated with CVerbs were generally more important than the factors related to MVerbs. Note that CVType, CVTransivity, and CVCompForm were located in the left part of the plot, while the others were in the right part (excepting MVType). This finding may have an interesting implication. In English, it has usually been mentioned and taught that the MVerb chose the complement type (*to*-infinitive vs. gerunds). The analysis results in Figure 12, however, might open the possibility that the CVerb (not the MVerb) chose the constructions in Korean.

Second, semantic factors had more importance values than morphological or syntactic factors in the determinations of -(*u*)*m* and -*ki* in Korean. In Table 3, four linguistic factors (i.e., MVType, CVType, MVSem, and CVSem) were semantically related factors, and three of them were statistically significant. Furthermore, CVType and MVType were located at the first and second in the variable importance. Thus, it can be said that semantic factors were more important than morphological or syntactic factors in the choice of -(*u*)*m* and -*ki* in Korean.

Third, though the language-specific factors significantly played some roles in the distributions of -(*u*)*m* and -*ki*, their importance was not so great in Korean. In Figure 12, note that MVHaToy and MVSinoKorean were located in the middle part of the plot. It implied that these two factors surely layed some roles in the distributions of -(*u*)*m* and -*ki*, their importance was not so great in Korean.

6. Conclusion

This paper took a corpus-based and statistical method to examine the distributions of -(*u*)*m* and -*ki* in Korean. For this purpose, the Sejong Treebank corpus was selected, and all the sentences with these two nominalizers were extracted. Then, twenty linguistic factors were manually encoded, and the encoded sentences were statistically analyzed with logistic regression.

Through the analysis, the followings were observed: (i) -(*u*)*m* and -*ki* were used with the ratio of 1:9 in Korean, (ii) among twenty-five linguistic factors, only ten factors were statistically significant, and (iii) not only the verbs which took -(*u*)*m* and -*ki* as a

complement but also the verbs which merged with these two nominalizers also played important roles in the determination of nominalizers. Although we took a monofactorial analysis where no interaction was included, we got an excellent model whose *C* value was 0.956.

Although we had a satisfactory model for the distribution of -(*u*)*m* and -*ki*, this paper had some limitations. First, the analysis in this paper included the linguistic factor for verbs only. However, it was also necessary to include the factors on the subjects (S_1 and S_2 in (4)) and to scrutinize the behaviors of those factors, although their influence might not be severe. Second, the data in this paper were extracted from the Sejong Treebank corpus, which had a relatively small size. It is necessary to study the distribution of -(*u*)*m* and -*ki* based on the corpus data with the larger size of balanced corpus. Third, although the genre differences were not considered in this paper, it is necessary to re-examine the distributions of these two nominalizers, since the frequency of words and morphemes are heavily influenced by the genre differences. These topics have to be considered in future research. Notwithstanding, we think that the analysis results in this paper can be a good guide to future research.

References

Baayen, Herald. 2008. *Analyzing linguistic data*: *A practical introduction to statistics using R*. Cambridge, MA: Cambridge University Press.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1): 5-32.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book* (2nd Edition). Boston, MA: Heinle and Heinle Publishers.

Chae, Wan. 1979. On the Korean nominalizer -*ki*. *Korean Linguistics* 8: 95-107.

Chang, Suk-Jin. 1966. Some remarks on Korean nominalization. *Language Research* 2(1): 18-31.

Choe, Hyeonbae. 1937/1975. *Urimalbon*. Seoul: Jeongumsa.

Cowan, Ron. 2008. *The teacher's grammar of English*. Cambridge, MA: Cambridge University Press.

Deshors, Sandra. 2014. Constructing meaning in L2 discourse: The case of modal verbs and sequential dependencies. In Dylan Glynn and Mette Sjölin (eds.), *Subjectivity and epistenicity*: *Stance strategies in discourse and narration*, 329-348. Lund: Lund University Press.

Deshors, Sandra. 2015. A multifactorial approach to gerundial and *to*-infinitival verb-complementation patterns in native and non-native English. *English Text Construction* 8(2): 207-235.

Deshors, Sandra and Stefan Gries. 2016. Profiling verb complementation constructions across New

Englishes: A two-step random forests analysis of -*ing* vs. *to* complements. *International Journal of Corpus Linguistics* 21(2): 192-218.

Greenbaum, Sidney 1996. *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon Press.

Harrell, Frank. 2001. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. Berlin: Springer.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. *The elements of statistical learning* (2nd edition). Berlin: Springer.

Hong, Chong-Sun. 1983. A study on the historical change of the nominalized endings. *Korean Language and Literature* 89: 31-51.

Huddleston, Rodney and Geoffrey Pullum. 2003. *The cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Im, Hong-Pin. 1974. On the semantic characteristics of nominalization. *Korean Linguistics* 2: 83-104.

Kim, Jae-Yoon. 1973. A study on the derivational suffix -*m* and -*ki*. *Study on the Korean Language Education* 2: 1-20.

Kim, Nam-Kil. 1984. *The grammar of Korean complementation*. Honolulu, HI: Center for Korean Studies, University of Hawaii.

Nam, Ki-sim and Young-kun Ko. 1993. *The standard Korean grammar*. Seoul: Top Publishers.

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

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

R Core Team. 2020. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Sim, Jae-Gi 1980. Semantic function of nominalizer. *Korean Journal of Linguistics* 5(1): 79-102.

Sohn, Ho-min. 1999. *The Korean language*. Cambridge, MA: Cambridge University Press.

Song, Jae Jung. 2005. *The Korean language*: *Structure, use, and context*. London: Routledge.

Suh, Jeong-soo. 1996. *Kukemunpep* (*The Korean grammar*). Seoul: Hanyang University Press.

Woo, Hyeong Shik. 1987. Distributions and semantic function of Korean nominalizer -*m* and -*ki*. *Language* (Yonsei University) 12: 119-160.

Yang, Dong-whee. 1975. On complementizers in Korean. *Korean Journal of Linguistics* 1(2): 18-46.

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

Gihyun Joh
Assistant Professor
Department of English Language and Literature
Kunsan National University
558 Daehak-ro, Gunsan-si, Jellabuk-do 54150, Korea
E-mail: johgihyun@kunsan.ac.kr

Received: 2020. 09. 12.
Revised: 2021. 01. 10.
Accepted: 2021. 01. 17.