Annotating Particle Realization and Ellipsis in Korean

Sun-Hee Lee
Wellesley College
Wellesley, MA 02481, U.S.A.
slee6@wellesley.edu

Jae-young Song
Yonsei University
Seoul, Korea
jysong@yonsei.ac.kr

Abstract

We present a novel scheme for annotating the realization and ellipsis of Korean particles. Annotated data include 100,128 *Ecel* (a space-based word unit) in spoken and written corpora composed of four different genres in order to evaluate how register variation contributes to Korean particle ellipsis. Identifying the grammatical functions of particles and zero particles is critical for deriving a valid linguistic analysis of argument realization, semantic and discourse analysis, and computational processes of parsing. The primary challenge is to design a reliable scheme for classifying particles while making a clear distinction between ellipsis and non-occurrences. We determine in detail issues involving particle annotation and present solutions. In addition to providing a statistical analysis and outcomes, we briefly discuss linguistic factors involving particle ellipsis.

1 Introduction

In Korean, the grammatical function of a nominal is represented by a morphologically-attached postpositional particle. Particles involve a wide range of linguistic information such as grammatical relations (subject, object), semantic roles (Agent, Patient, Location, Instrument, etc.), discourse/pragmatic properties, such as topic markers, delimiters and auxiliary particles, as well as conjunctions. Due to their complex linguistic functions, particles are one of the most rigorously investigated topics in Korean linguistics.

In example (1), the particle *ka* indicates subjecthood and *ul* refers to objecthood.[1]

(1) onul-**un** Mina-**ka** kyosil-**eyse** cemsim-**ul** mek-e.
today-**TOP** M-**SUBJ** classroom-**in** lunch-**OBJ** eat-ENG
'Mina eats lunch in the classroom.'

The subject particle *ka* also marks Agent (semantic role); the locative particle *eyse* combines with a nominal referring to Location; *un* marks topichood in the given discourse, etc.

In spite of their linguistic function (representing the grammatical relations of subject and object), these particles frequently disappear, particularly in spoken Korean (Hong et al., 1998; Kim and Kwon, 2004; Lee and Thompson, 1985; Lee 2006, 2008). Previous studies have mainly focused on case particles and suggested that register variation is the key factor in particle ellipsis. However, few studies have comprehensively examined both spoken and written data with specific annotation features and guidelines. By using balanced spoken and written data, this paper explores the realization of all particles and ellipsis of case particles including subject and object case. In order to test the effect of register variation on particle realization, we designed a balanced corpora to include four different styles. The spoken corpora include everyday conversations, informal monologues (story-telling), TV debates, and lectures/speeches; the written corpora include personal essays, novels, news articles, and academic papers.

Categorizing particles requires a well-articulated classification. Particles have complex grammatical

[1] The subject and object particles have the phonological variants *i* and *lul*, respectively.

Proceedings of the 6th Linguistic Annotation Workshop, pages 175–183,
Jeju, Republic of Korea, 12-13 July 2012.

functions, and it is difficult to determine if a missing particle is a case of ellipsis or non-occurrence. We discuss these challenges in the context of developing a novel annotation scheme and guidelines. We examine particle ellipsis patterns across registers, as well as semantic and pragmatic factors triggering particle ellipsis.

2 Relevant Background

Within theoretical linguistics, Korean particles have been classified according to three distinct linguistic functions: case particles, auxiliary particles, and conjunctive particles (Nam, 2000; Lee, 2006)[2]. A case particle combines with an argument or adjunct nominal and specifies the grammatical relation and semantic role of the nominal within the argument structure of a predicate. In contrast to case particles, auxiliary particles are not based on the grammatical relation of a nominal and a predicate; they introduce extra semantic and discourse interpretations. This category includes topic markers and delimiters, as well as other particles with diverse lexical meanings. In addition, there are conjunctive particles that attach to nominals and connect them to the following ones.

Identifying the diverse functions of particles is important for syntactic, semantic, and discourse analyses in Korean. When a particle is elided, recovering the information behind a missing particle is essential for determining accurate grammatical relations, which is a prerequisite for computational processes of parsing, discourse analysis, machine translation, etc. However, the recovery process for missing particles does not include auxiliary particles as candidates due to their unpredictable distributions; auxiliary particles have their own discourse and pragmatic meanings, and their distributions over nominals are not restricted by grammatical relations with predicates.

On the one hand, the validity of recovering a missing particle into its original form itself can be questionable; it has been argued in the literature that zero marking is the unmarked option and there is no ellipsis or deletion of particles (Lee and Thompson, 1989; Fujii and Ono, 2000 inter alia). However, whether a particle is deleted or originates as a zero form, it is important that a missing particle corresponds to a particular case particle and identification of it is crucial for determining the grammatical and semantic function of the bare nominal.

With respect to particle ellipsis in Korean and also Japanese, most previous research has focused on subject and object particles. There have been contradictory reports on the dropping rates of these particles. Whereas Kwon (1989) and Hong et al. (1998) report a higher dropping rate for subject particles, Kim and Kwon (2004) and Lee (2006) argue for a higher dropping rate for object case markers in colloquial Korean. Among these studies, Hong et al. (1998) analyzes different radio shows with a total time span of 60 minutes and Lee (2006) analyzes the Call Friend Korean (CFK) corpus of telephone speech. Even disregarding the small data size (the former with fewer than 2000 noun phrases and the latter with 1956 overtly expressed subject and object NPs), the statistical results are less than convincing given the lack of a specific annotation scheme and guidelines. For example, Hong et al. (1998) include nominals with some topic markers or delimiters as tokens of case marker ellipsis. However, as mentioned in Lee (2008), these cases need to be excluded from the list of case ellipsis because the subject or object particles are morphologically restricted from co-occurring with auxiliary particles in Korean. Although Lee (2008) excludes optional occurrences of object particles in light verb constructions, it is not quite clear how non-occurrences of particles are separated from ellipsis of particles in the corpus study without specific guidelines. In order to develop a more comprehensive analysis of case ellipsis, it is necessary to employ large data sets with different registers across spoken and written Korean and a well-established annotation scheme and guidelines.

3 The Data and Annotation Scheme

3.1 Data

We extracted 100,128 *Ecel* with morphological tagging from the Sejong Corpora to create spoken and written balanced corpora composed of four different registers with different degrees of formality. Approximately 2000 *Ecel* were each

[2] Although particles combine with nominals, they sometimes follow a verbal phrase or a sentence adding semantic and pragmatic meanings of honorification, focus, etc. Some researchers assign these particles to a special category (Nam, 2000). In this study, we only examine particles combining with nominals and not with phrasal or sentential categories.

selected from 49 files to build balanced corpora. Table 1 summarizes the composition of the data.

Type	Registers		# of Files	Size
Spoken	Private	Everyday Conversations (E)	7	12,504
		Monologues (M)	6	12,502
	Public	TV Debates & Discussions (D)	6	12, 547
		Lectures & Speeches (L)	6	12, 526
Written	Personal Essays (PE)		6	12, 510
	Novels (N)		6	12, 505
	Newspaper Articles (P)		6	12, 511
	Academic Textbooks (A)		6	12, 505

Table 1. Composition of Balanced Corpora

3.2 Annotation Scheme

In agglutinative languages like Korean, particles are attached to preceding nominals without spaces, and identifying the position of a particle requires accurate segmentation. Although we extracted data with morphological tags, the tags sometimes reflected errors in spacing, morpheme identification, segmentation, etc. Therefore, we manually corrected relevant errors in segmentation and morpheme tags before performing annotation. Using morpheme tags, we identified all the nominal categories in the corpora that can combine with particles, including all the nominals with and without particles. We annotated realized particles and determined their categories using the tag set in Figure 1. In addition, we selected four annotation features to mark up particle realization and ellipsis. The given tag set has been used to annotate both realized particles and missing particles. However, annotating missing particles presents challenges and requires a new annotation scheme. Elided particles are recovered using the case particles based upon grammatical relations between a nominal and a predicate. The details are presented in the next section.

- **Tag Set of Particles**
  - **Case Particles**[3]**:**

Subject (S): *ka/i* Subject Honorific (SH): *keyse*
Object (O): *ul/lul* Genitive (G): *uy*
Dative (D): *ey/eykey* (‘to’), *hanthey* (‘to’)
Dative Honorific (DH): *kkey* (‘to’)
Complement (C): *ka/i*
Adverbial Case (B):
Time (BT): *ey* (‘in, at’)
Location (BL): *ey* (‘to’), *eyse* (‘from’)
Instrument (BI): *lo/ulo* (‘with’)
Direction (BD): *lo/ulo* (‘to, as’)
Source (BS): eyse (‘from’), *eykey*(*se*) (‘from’), *hanthey*(*se*) (‘from’) , *pwuthe* (‘from’), *ulopwuthe* (‘from’), *eysepwuthe* (‘from’),
Goal (BG): *ey* (‘to’), kkaci (‘to’)
Accompany (BA): *wa/kwa* (‘with’), *hako* (‘with’), *ilang/lang* (‘with’)
Vocative (V): *a/ya*
Comparative (R): *pota* ('than'), *mankhum* ('as~as'), etc.

  - **Discourse/Modal**:

Topic (T): *un/nun/n*
Auxiliary (A): *to* (‘also’), *man* (‘only), *mata* (‘each’), *pakkey* (‘only’), chelem (‘like’), *mankhum* (‘as much as’), etc.

  - **Conjunction** (J): *wa/kwa* (‘and’), *hako* (‘and’), *ina/na* (‘or’), *itunci/tunci* (‘or’), *ilang/lang* (‘and’), etc.

- **Annotation Features**
  Realized Particle, Realized Particle Type
  Missing Particle, Missing Particle Type

Figure 1. Annotation Scheme of Particles

3.3. Ellipsis vs. Non-Occurrence of Particles

As defined in Fry (2001), ellipsis is the phenomenon whereby a speaker omits an obligatory element of syntactic structure. However, there are at least three morpho-syntactic constructions in Korean where a particle does not need to be recovered because it is not obligatory in the given position. Our annotation distinguishes these optional non-occurrences from the particle ellipsis phenomenon and marks them separately.

First, the occurrence of the genitive case *uy* is optional depending on various syntactic and semantic relation between two nominals in Korean. For example, the genitive *uy* tends to disappear after a complement nominal of a verbal noun, e.g., *yenghwa-uy/Ø chwalyeng* (movie-GEN + filming) 'filming of a movie', whereas it appears after a subject nominal of a verbal noun, e.g., John-uy/*?*Ø wusung* (John-GEN + winning) 'John's winning'. Due to complex linguistic factors, there is still controversy regarding how to predict occurrences of the genitive case in Korean (Lee, 2005; Hong,

[3] We focused on particles that directly follow nominals. Thus, particles that appear after verb phrases or sentences have been excluded from our tag set, including the direct quotation particle *lako* and *hako.*

2009), and native speakers' intuitions on the positions of the dropped genitive particle and its recoverability vary.[4] Therefore, we chose not to annotate the genitive particle *uy* when it does not occur and we do not count particle ellipsis within a nominal phrase.

Second, particles are optional in light verb constructions, as mentioned in previous research (e.g., Lee and Thompson, 1989; Lee and Park, 2008). In Korean, the morphological formation of a Sino-Korean (or foreign-borrowed) verbal noun and the light verbs (LV) *hata* 'do', *toyta* 'become', and *sikhita* 'make' is very frequent, e.g., *silhyen* (accomplishment)+*hata*/*toyta*/*sikhita*to 'accomplish /to be accomplished/to make it accomplish', *stheti* (study) +*hata*, 'to study' etc. In these light verb constructions, the subject particle *i/ka* or the object particle *ul/lul* can appear after the verbal nouns as in *silhyen-**ul** hata* (accomplishment-OBJ do), *silhyen-**i** toyta* (accomplishment-SBJ become), *silhyen-**ul** sikhita* (accomplishment-OBJ make), *stheti-**lul** hata* (study-OBJ do), etc. Realization of these case particles, however, is not mandatory and even unnatural when the argument of a verbal noun appears in the same sentence, as in the following example.

(3) ?*John-i kkum-**ul** silhyen-**ul** hayssta.
J-nom dream-**OBJ** accomplishment-**OBJ** did
'John accomplished his dream.'

In considering the morpho-syntactic unity of N+LV combinations as single predicates and the awkwardness of a realized particle after a verbal noun, we conclude that N + LV combinations do not involve case ellipsis.[5] However, when these LV combinations include negation, the negative adverb intervenes between a verbal noun and the LV, and the particle *i/ka* or *ul/lul* follows the verbal noun. In those constructions, we exceptionally assume particle ellipsis. This decision affects the result of our corpus analysis due to the high frequency of LV combinations, particularly with respect to object particle ellipsis. In contrast, Lee and Thompson (1989) assume particle ellipsis in N+LV combinations unless there is another nominal with an object particle licensed in front of the verbal noun. Although we exclude particle ellipsis in light verb constructions, we separately mark up possible case realizations of LV combinations in order to measure the extent to which they affect the statistical results.

Third, optional particles frequently appear with bound nouns (or defective nouns) in Korean. Bound nouns refer to nominals that do not occur without being preceded by a demonstrative, an adnoun clause, or another noun, which includes *tey* 'place', *ttay* 'time' *swu* 'way', *ke*(*s*) 'thing', *cwul* 'way', *check* 'pretense', etc.

(4) hakkyo-eyse kongpwuha-l swu(**-ka)** issta.
school-at study-REL way (**-NOM**) exist
'It is possible to go to study at school.'

Bound nouns are functionally limited with respect to neighboring constituents. For instance, a bound noun *ttay* 'time' only combines with a clause ending with the adnominal ending *-(u)l*, whereas *hwu* 'after' combines with a clause ending with *-(u)n*.[6] In addition to morpho-syntactic reliance on the preceding clause, many bound nouns form formulaic expressions with the following predicates (i.e., the bound noun *swu* 'way' only combines with existential predicates, *issta* 'exist' and *epsta* 'do not exist'). Considering that particles in bound nouns are frequently dropped and do not represent grammatical relations of bound nouns with respect to the predicate, we also exclude them as cases of ellipsis.[7]

[4] Although semantic change and lexical insertion can be used for identifying morphological compounds, it is still very difficult to distinguish nominal compounds and syntactic nominal complexes. Therefore, school grammars present some inconsistent distinctions. For example, *wuli nala* (we country) 'our country' is considered a single lexical word, a compound nominal, whereas the similar combination, *wuli kacok* (we family) 'our family' is a complex NP composed of two separate nouns.

[5] It is also arguable whether the realization of a particle after a verbal noun is based on the subcategorization feature of the light verb *hata* or *toyta*. Through personal conversations, some scholars suggest that the realization of a particle after a verbal noun may be a case of insertion. When adopting this argument, particle omission is not even possible for the LV constructions. This needs to be more thoroughly investigated through examining historical corpus data.

[6] For bound nouns in Korean, refer to Sohn (1999).

[7] Classifiers belonging to bound nouns show interesting patterns of case particle realization in Korean; classifiers form morphosyntactic combinations such as [Noun + Number + Classifier], e.g., *sakwa han kay* (apple one thing) 'one apple'. Normally, a case particle appears on the initial content noun or the final classifier (e.g. [sakwa-**ka/lul** han kay][[sakwa han kay-**ka/lul**]) or there is a copy of the case particle from the content noun (e.g.[sakwa-**ka/lul** han kay-**ka/lul**]). In this study,

Spoken Corpora		E	M	D	L	Total
Particle Realization		**2081**	**2853**	**3334**	**3672**	**11940**
Predicate Nominals (P)		741	590	742	757	2830
Zero Particles	Ellipsis	**843**	**395**	**237**	**185**	**1660**
	Compounds (N)	320	297	350	411	1378
	Optional (E)	796	735	841	802	3174
	Light Verb (L)	308	190	482	410	1390
	Vocative (V)	24	3	6	20	53
Errors		82	36	41	43	202
Written Corpora		PE	N	P	A	Total
Particle Realization		**4707**	**4715**	**4603**	**4928**	**18953**
Predicate Nominals (P)		593	600	393	612	2197
Zero Particles	Ellipsis	**98**	**86**	**165**	**12**	**361**
	Compounds (N)	406	104	1941	728	3179
	Optional (E)	996	1125	1492	712	4325
	Light Verb (L)	361	437	965	917	2680

Table 2. Grammatical Realization of the Nominal Category[8]

In addition to optional particles, we also note that some constructions mandatorily require non-occurrence of particles. We have already seen that the genitive particle is not allowed within nominal compounds, e.g. [*palcen+Ø*(**-uy*) *keyhwoyk+Ø*(**-uy*) *pokose*] 'development plan report'. In addition, some bound nouns form formulaic (or idiomatic) expressions with their neighboring words and do not combine with particles, e.g., *kes-(*kwa)+ kathta* (thing-(*with) + similar) 'seem', *ke-Ø* + *aniya* (thing + isn't) 'isn't it?', N-*Ø* + *ttaymwun* (N + reason), etc.

Also, particle omission is required by the lexical properties of nominals. For example, numbers belonging to the nominal category combine with subject or object particles as well as with other auxiliary and discourse particles (e.g., *tases-**un/-i/-ul*** 'five-**TOP**/**SBJ**/**OBJ**'). However, they cannot take any particle when followed by count bound nouns, e.g., *tases-**Ø*** + *kay/salam/pen/kaci*/... (five + items/people/sorts, etc.). Similarly, time nominals such as *onul* 'today', *ecey* 'yesterday', *nayil* 'tomorrow' stand alone without particles as adverbial phrases even though they combine with other particles in different syntactic positions. In contrast, time nominals such as *onul achim* 'this morning' and *2000 nyen* 'year 2000' can stand alone but also combine with the time particle *ey*. These temporal *ey*s are considered to be optional.

In summary, optional and mandatory non-occurrence of particles restricted by morpho-syntactic and lexical constraints needs to be distinguished from the omission of obligatory particles. Therefore, we include the following features to annotate bare nominals that do not mandate recovery of particles.

E - Non-occurrence of a particle based upon lexical or morpho-syntactic constraints.

N - Non-occurrence of a particle after a nominal that forms a compound with the following nominal

L - Non-occurrence of a particle in light verb constructions

In addition, nominals can be combined with copula *ita* or appear at the end of a phrase or a sentence without the copula in Korean. These predicate nominals have been annotated separately from other nominals. When a nominal is repeated by mistake with or without a particle, these erroneous nominals are separately marked and excluded from counts of particle realization and ellipsis. Separate features are given to handle these cases.

P- Predicate nominals combining with copula *ita*. It also marks a nominal standing alone without *ita*, as answering utterance.

ER - Errors including a repeated nominal by mistake or an incomplete utterance

as long as there is one particle realized in either the content noun or the classifier, we do not count it as case ellipsis.

[8] E: Everyday Conversations; M: Monologues, D: Debates; L: Lectures; PE: Personal Essays; N:Novels; P:Newspapers, A:Academic Texts

3.4 Principles of Annotating Particle Omission and Inter-Annotator Agreement

Our annotation principles of missing particles are presented as follows:

- With respect to missing particles, we annotate only obligatory case particles and conjunctive particles while excluding discourse/modal particles. This captures the minimum needed for a particle prediction system.
- In the process of recovering elided forms, there are cases in which more than one particle could be correct. Instead of selecting a single best particle, we present a set of multiple candidates without preference ranking.
- Particle stacking is allowed in Korean. We annotate stacked particles as single units without separating them into smaller particles. However, their segmentation is specified under the annotation feature of realized particle type. Missing particles, however, exclude stacked particles. Most particle stacking includes a discourse/modal particle that adds its specific meaning to the attached nominals.

Based on our annotation scheme and guidelines, two experienced annotators manually annotated realized particles, missing particles, and their types on the spoken and written corpora separately and cross-examined each other's annotation. Difficult cases were picked out and discussed with each other to reach an agreement. In order not to overly inflate the values with words that do not take particles, we removed words that do not belong to the nominal categories (nouns, pronouns, bound nouns, and numbers). The realized particles were provided to the annotators with the morphological analysis. Thus, we decided to compute the inter-annotator agreement on only 466 nominals with no particles within 5000 *Ecels* (before cross-examination). The kappa statistic on the case ellipsis by the two annotators is 91.23% for the specific particles. The agreement rate is much higher than we expected, but can be attributed to the annotation guidelines, which were clear and limited recovery of particles to case particles not including auxiliary and discourse particles. The two annotators were highly trained, having over two years of experience with particle annotation tasks.

4 Corpus Analysis

Table 2 summarizes the results of particle annotation of all the nominals, and Table 3 focuses on particle realization and ellipsis. Table 2 shows all nominal realizations with particles and without. Zero particles include both bare particle ellipsis and bare nominals including nominals that do not require particles as a component of compound nominals (N) and nominals that appear without particles in the corpora although they may optionally (E). In addition, the spoken corpora include bare nominals used as vocative phrases without particles. These cases have been counted separately. Erroneous usage of nominals only appears in the spoken corpora. Light verb combinations here only include cases that may allow realization of subject or object case particles, whose numbers are significantly high both in the spoken corpora and the written corpora.

As we see in Table 3, the overall case ellipsis rates are not that high across the two registers, but the difference between the spoken and the written corpora is significant (χ^2=851.78, p <.001).

Spoken	E	M	D	L	Total
Realized	71%	88%	93%	95%	88%
Ellipsis	29%	12%	7%	5%	12%
Written	PE	N	P	A	Total
Realized	98%	98%	97%	99.7%	98%
Ellipsis	2%	2%	3%	0.3%	2%

Table 3. Particle Realization vs. Ellipsis

Furthermore, genre plays an even more significant role within the spoken corpora. Particle ellipsis in everyday conversations is significantly more frequent than in monologues, debates, or lectures using a Bonferroni adjusted alpha level of .008 per comparison (.05/6). ($\chi^2(1)$=266.64, p<.001; $\chi^2(1)$=571.19, p<.001; $\chi^2(1)$=746.93, p<.001). Particle ellipsis in monologues is significantly more frequent with debates or lectures ($\chi^2(1)$=61.66, p<.001; $\chi^2(1)$=126.59, p<.001). In contrast, particle ellipsis between debates and lectures shows a lower chi-square value than the other cases, although the value is still significant. ($\chi^2(1)$==11.72, p<.001).

Table 4 presents the annotation results of case particle realization and ellipsis including subject and object particles.

Particles	Spoken					Written				
	E	M	D	L	Total	PE	N	P	A	Total
SUBJ +	63% (539)	88% (776)	93% (927)	95% (848)	85% (3090)	97% (743)	97% (840)	92% (635)	99.7% (588)	98% (2806)
SUBJ −	37% (318)	11% (97)	7% (67)	5% (48)	15% (530)	3% (25)	3% (24)	3% (18)	0.3% (2)	2% (69)
OBJ +	51% (398)	73% (535)	85% (698)	89% (771)	75% (2402)	94% (967)	95% (1066)	99% (1050)	99% (1026)	97% (4109)
OBJ −	49% (389)	27% (198)	15% (121)	11% (92)	25% (800)	5% (56)	5% (53)	1% (13)	1% (9)	3% (131)
CONJ +	92% (57)	68% (54)	90% (89)	98% (137)	88% (337)	100% (133)	100% (113)	97% (226)	99.7% (276)	99% (748)
CONJ −	8% (5)	32% (26)	10% (10)	2% (3)	12% (44)	0% 0	0% (0)	3% (7)	0.3% (1)	1% (8)
OTHERS +	81% (549)	90% (634)	95% (859)	97% (1174)	92% (3213)	99% (1778)	99.5% (1739)	93% (1680)	100% (2173)	98% (7370)
OTHERS −	19% (131)	10% (74)	4% (39)	3% (42)	8% (286)	1% (17)	0.5% (9)	7% (127)	0% (0)	2% (153)

Table 4. Realization and Ellipsis of Case Particles

Overall dropping rates of subject particles and object particles show a difference between the spoken and the written corpora. Object particle dropping is significantly more frequent in the spoken corpora than in the written corpora (χ^2 =797.03, p<.001). Within the spoken corpora, there is also some variation according to genre. Both subject and object dropping rates increase as the genres become less formal. In everyday conversations, the dropping rate of object particles reaches 49% and the dropping rate of subject particles is 37%. While the dropping rates of both particles decrease in the formal registers of the spoken corpora, the dropping rate of the object particles is consistently higher than the dropping rate of the subject particles at each register. In parallel, conjunctive particles and other case particles are more frequently dropped in the spoken corpora than in the written corpora.[9]

Our findings can be summarized as follows:

- In Korean, particle ellipsis is not very frequent. The particle dropping rate for subjects is 12% in the spoken corpora and 2% in the written corpora.
- The effect of register variation on particle ellipsis (everyday conversations vs. debates & lectures) demonstrates that particle dropping is less preferred in formal contexts. However, formality per se is not the deciding factor, but a partially related factor.[10]
- Across the spoken corpora, object particles drop more frequently than subject particles. (χ^2 =115.17, p <.001)
- Other case and connective particles are also more frequently elided in the spoken corpora.

5 Linguistic Properties in Particle Ellipsis

The frequent case particle ellipsis in the spoken corpora suggests that discourse need to be further investigated. This implies that discourse factors contribute to particle ellipsis, as suggested in Lee and Thompson (1989). Using the corpus annotation, we can explore linguistic properties involving in particle ellipsis.

5.1 Definiteness and Specificity

A case particle is likely to be dropped when the preceding noun is definite or specific (Kim, 1991). The definite NP *ku haksayng* 'that student' can drop subject case. This contrasts with the fact that the indefinite expression *etten haksayng* 'some student' cannot appear without the subject particle.

[9] Unexpectedly, conjunctive particles drop more frequently in monologues than in everyday conversations.

[10] This can be supported by the fact that register variation does not affect particle dropping in the written corpus.

(5) a. ku haksayng-i/-Ø na-lul chacawa-ss-e.
that s tudent-SBJ/Ø I-OBJ visit-PAST-END
'That student visited me.'
b. etten haksayng-i/*Ø na-lul chacawa-ss-e.
some student-SBJ /Ø I-OBJ visit-PAST-END
'Some student visited me.'

In our annotated corpus, the particles that are attached to personal pronouns and *wh*-pronouns are frequently dropped. This implies that definiteness is a crucial factor for licensing particle dropping.[11]

5.2 Familiarity and Salience in Discourse

Particle ellipsis is also based on discourse properties of familiarity (background).[12] In the following example, it is more natural to drop the object particle from *tampay* 'cigarette' when speaking in a convenience store. This is because selling cigarettes is already familiar knowledge shared among the discourse participants.

(6) tampay-**ʔlul/-Ø** cwu-seyyo.
cigarette-OBJ-Ø give-IMPERATIVE
'Please give me cigarette.'

However, when the object particle is used in (6), the object cigarette is exclusively designated or highlighted. This contrasts with the fact that the speaker commonly uses a nominal referring to discourse participants such as *you* and *I*, proper names, or titles without a particle in order to catch the attention of the listener(s). Also, when a subject or object nominal is scrambled out of its original position and appears at the sentence initial or final position, the particle disappears to emphasize the salience of the nominal element, as in (7).

(7) a. philyohan-n kel hanato mos tule, na-Ø.
necessary- REL thing anything not take I-Ø
'I cannot take anything that is necessary.'
b. saylo o-n sensayng-**Ø (ul)**, ne alla?
newly come-REL teacher-Ø (OBJ) you know
'Do you know the new teacher?'

Examination of our annotated corpora strongly suggests that particle ellipsis is associated with two contrastive discourse properties, familiarity and salience, and also that it interacts with other grammatical mechanisms such as word order, lexical category, and possibly prosody.[13]

6 Final Remarks

In this study, we presented our annotation work on particle realization and ellipsis using spoken and written corpora in Korean. A new annotation scheme and principles were presented, along with challenging issues and solutions, such as the recovery of missing particles and the distinction between ellipsis and non-occurrence of particles. In order to evaluate the effect of register variation on particle ellipsis, we incorporated four different genres. Our major finding is that the rate of particle ellipsis in Korean is not as high as generally assumed and register variation is a significant factor only in spoken corpora. The more informal dialogs are, the more often particles are elided. Our corpus annotation suggests that particle ellipsis is related to activated semantic/pragmatic constraints among discourse participants, which include definiteness, specificity, familiarity and salience.

The implication of these findings is significant not only for linguistic theory, but also for language processing, Korean language teaching, and translation. Particle ellipsis will be a more serious issue for computational modeling that incorporates informal spoken dialogs than for computational processing on written texts. In language teaching, particles need to be emphasized more for formal writing and formal speaking based on their frequency in the given register (Lee et al., this volume). Next, we plan to run error detection software on our corpus to verify the consistency of our annotation (Dickinson and Meurers, 2003), to prepare for releasing the data with guidelines, to further analyze the results of the annotation, and to address more elaborate linguistic implications in the annotated data.

[11] Lee (2006, 2010) argues that case ellipsis of subjects and objects interacts with the definiteness of nominals. The rate of case ellipsis for strongly definite subject NPs is significantly higher than the rate for weakly definite NPs. However, object case ellipsis works in the opposite direction. It is difficult to identify definiteness of a nominal in Korean, where definite and indefinite articles do not exist. We have not annotated definiteness features in our corpora, but intend to as part of future work.

[12] Similarly, Lee and Thompson (1989) propose that "sharedness between communicators" is the pragmatic factor determining object particle ellipsis in discourse.

[13] Case ellipsis and realization have been also examined within information structure-based analyses such as Lee (2006, 2010) and Kwon and Zribi-Hertz (2008)

References

Markus Dickinson and Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. Proceedings of the 10th Conference of European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, Hungary.

John Fry. 2001. Ellipsis and 'wa'-marking in Japanese conversation. Doctoral Dissertation. Stanford University.

Noriko Fujii and Tsuyoshi Ono. 2000. The Occurrence and Non-Occurrence of the Japanese Direct Object Marker *O* in Conversation. Studies in Language, 24(1): 1-39.

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. Language, 69: 274-307.

Young-joo Hong. 2009. Syntactic Relation between Two Nominals in NP and Non-occurrence of の/uy (*Myengsakwu Nay-uy Cenhang Myengsa-wa Hwuhang Myengsa-uy Thongsacek Kwankey-wa の /uy-uy Pisilhyen*, In Korean). Japanese Study (*Ilbon Yengoo*), 40: 639-653. The Institute of Japanese Studies. Seoul.

Paul Hopper and Sandra A. Thompson. 1984. The Discourse Basis for Lexical Categories in Universal Grammar. Language, 60: 703-752.

Ji-Eun Kim. 1991. A Study on the Condition in Realizing Subject without Case Marker in Korean, Hangul, 212.

Kun-hee Kim and Jae-il Kwon. 2004. Korean Particles in Spoken Discourse-A Statistical Analysis for the Unification of Grammar. Hanmal Yenku, 15: 1-22.

Eon-Suk Ko. 2000. A Discourse Analysis of the Realization of Objects in Korean. Japaenese/Korean Linguistics, 9: 195-208. Stanford: CSLI Publication.

Jae-il Kwon. 1989. Characteristic of Case and the Methodology of the Case Ellipsis, Language Research, 25(1): 129-139.

Song-Nim Kwon and Anne Zribi-Hertz. 2008. Differential function marking, case, and information Structure: Evidence from Korean. Language, 84(2): 258-99.

Hyo Sang Lee and Sandra A. Thompson. 1989. A discourse account of the Korean accusative marker. Studies in Language, 13: 105-128.

Hanjung Lee. 2006. Parallel Optimization in Case Systems: Evidence from Case Ellipsis in Korean. Journal of East Asian Linguistics, 15: 69-96.

Song-Nim Kwon and Anne Zribi-Hertz. 2008. Differential Function Marking, Case, and Information Structure: Evidence from Korean. Language, 84:2:258-299

Hanjung Lee. 2010. Explaining Variation in Korean Case Ellipsis: Economy versus Iconicity. Journal of East Asian Linguistics, 19: 292-318.

Seon-woong Lee. 2005. A Study on Realization of Nominal Arguments (*Myengsa-uy Nonhang Silhyen Yangsang*, In Korean).

Sun-Hee Lee. 2006. Particles (*Cosa*). Why Do We Need to Reinvestigate Part of Speeches? (in Korean): 302-346.

Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2012. Developing Learner Corpus Annotation for Korean Particle Errors. In Proceedings of the Sixth Linguistic Annotation Workshop (this volume). Jeju, Korea

Minpyo Hong, Kyongjae Park, Inkie Chung, and Jiyoung Kim. 1998. Elided Postpositions in Spoken Korean and their Implications on Center Management, Korean Journal of Cognitive Science, 9(3): 35-45.

Yoon-jin Nam. 2000. A Statistical Analysis of Mondern Korean Particles (*Hyentay Hankwuke-ey tayhan Kyelyang Enehakcek Yenkwu*). Thayhaksa.

Ho-Min Sohn. 1999. The Korean Language. Cambridge University Press. Cambridge, UK.

Yongkyoon No. 1991. A Centering Approach to the *[CASE][TOPIC] Restriction in Korean. Linguistics, 29: 653-668.

Yu-hyun Park. 2006. A Study on the Particle '-ka"s Non-Realization in Modern Korean Spoken Language. Emwunlonchong, 45: 211-260.

EnricVallduv íand Maria Vilkuna, M. 1998. On Rheme and Kontrast. The Limits of Syntax, eds. Peter Culicover and Louise McNally, 79-109. New York: Academic Press.

Shuichi Yatabe. 1999. Particle Ellipsis and Focus Projection in Japanese. Language, Information, Text, 6: 79-104.