Linguistic Research 32(3), 749-771
DOI: 10.17250/khisli.32.3.201512.009

Information structure, topic predictability and gradients in Korean case ellipsis: A probabilistic account*

Hanjung Lee
(Sungkyunkwan University)

Lee, Hanjung. 2015. Information structure, topic predictability and gradients in Korean case ellipsis: A probabilistic account. *Linguistic Research* 32(3), 749-771. This paper examines previous characterizations of the phenomenon of case ellipsis in Korean in terms of the information status of argument NPs and provides empirical evidence for a probability-based account. Previous information structural accounts of case ellipsis tended to place exclusive importance on information structure and make a categorical distinction between case-marked and unmarked NPs based on their information status. However, evidence from experimental and conversation data demonstrates that judgments on case ellipsis are not categorical but gradient and that the frequency of argument ellipsis and case ellipsis for topic subjects and objects is correlated with the degree of topic predictability. This evidence supports the conclusion that both the effect of information structure and the gradient acceptability of case ellipsis can be explained in terms of probabilistic properties associated with argument NPs. **(Sungkyunkwan University)**

Keywords case ellipsis, focus, gradient acceptability, information status, subject-object asymmetry, topic predictability, usage probability

1. Introduction

Korean is a language that marks nominative case and accusative case for subjects and objects of agentive transitive predicates, respectively and also allows optional case-marking in which the case marker can be omitted in certain contexts. The following sentence in (1) illustrates the phenomenon of case ellipsis (henceforth CE):

(1) ecey Minswu-ka chinkwu manna-ss-ta.
yesterday Minsoo-NOM friend(-ACC) meet-PAST-DECL

* I would like to thank two anonymous reviewers for their useful comments. Any errors and misinterpretation are my own.

In (1), the object *chinkwu* 'friend' appears without the following accusative case marker *−lul*, which normally indicates the object of the verb. In colloquial Korean, this kind of ellipsis is extremely frequent. The same phenomenon has been observed in Japanese and studied extensively in the literature (e.g., Hinds 1983; Tsutsui 1984; Masunaga 1988; Matsuda 1996; Minashima 2001; Fry 2001; D. Lee 2002; Shimojo 2005, among others).

Korean CE has often been described in terms of the information structure status of the NP. Among notions relevant to the information status of nominal arguments, contrastive focus has been claimed to be one of the important factors affecting the naturalness of CE in Japanese and Korean. A number of previous studies have suggested that case markers in Japanese and Korean cannot be omitted when the argument they mark is contrastively focused (Tsutsui 1984; Masunaga 1988; Ko 2000; D. Lee 2002; Kwon and Zribi-Hertz 2008, among others). The claim about the unacceptability of CE on contrastively focused arguments raises the following questions that remain unanswered: Do CE on subjects and objects exhibit the same pattern with respect to topic and focus? Are judgments on CE for topic and focus arguments categorical or gradient?

The present paper aims to provide answers to these questions by closely inspecting previous claims made about the (un)acceptability of CE on focused arguments and presenting new empirical data. Section 2 provides a review of Kwon and Zribi-Hertz's (2008) information structural account of Korean CE, which claims that a single factor of focus-structure visibility determines categorical distinction between case-marked vs. unmarked NPs. Based on analyses of empirical data, I will show that their account is flawed because judgments on CE are not categorical but gradient statistical preferences that are dependent upon multiple factors. Section 3 first introduces the rationale behind our probability-based account of CE and then presents a probabilistic account of CE that can account for both the effect of information structure and the gradient acceptability of CE. Section 4 presents new evidence from an analysis of argument encoding patterns in conversation data, demonstrating that the frequency of argument ellipsis and CE for topic subjects and objects is correlated with the degree of topic predictability. This finding provides strong support to the view that native speakers' knowledge of grammar includes not only some degree of knowledge of probabilistic information, as suggested by prior work, but even access to fine-grained probabilities.

2. Information structure and case ellipsis

Previous research on CE in Japanese and Korean has identified a number of linguistic and non-linguistic factors that favor CE. Recently, Kwon and Zribi-Hertz (2008) proposes an information structure-based account that derives a range of properties of CE described above from a single information structural property of arguments. They put forward a proposal that case marking in Korean correlates with interpretive contrasts that should be captured in terms of information structure (focus structure). Their leading descriptive assumptions are summarized in (2):

(2) a. NPs that support functional markers indicating structural positions in syntax are visible in f(ocus)-structure.
b. NPs that fail to support such markers are not visible in f-structure, unless some other type of marking guarantees their visibility as f-structure constituents.

Adopting Erteschik-Shir's (1997, 2007) framework to represent f-structure, Kwon and Zribi-Hertz (2008) argue that unmarked subjects and objects in Korean fail to be visible at this level. Consequently, they may be construed neither as active topics nor foci, and thus must either be left out of f-structure or incorporated within larger f-structure constituents in order to be interpreted: Unmarked objects are never construed as topics or foci and always exhibit a form of semantic incorporation, while case-marked objects always stand as f-structure constituents as focused at some level. They illustrate the interpretive contrast between case-marked objects and unmarked objects with the following example:

(3) A: Minswu-nun mwe-l sa-ss-ni?
Minsoo-TOP what-ACC buy-PAST-INT
'And Minsoo, what (kind of thing) did he buy?'
B: Minswu-nun sakwa*(-lul)sa-ss-ta.
Minsoo-TOP apple-ACC buy-PAST-DECL
'Minsoo(, he) bought apples.' (Kwon and Zribi-Hertz 2008: 263)

In (3B), the object *sakwa* ‘apple’ is the non-contrastive, informational focus as it simply marks new information in the sentence without explicitly contrasting it with something in the discourse. Kwon and Zribi-Hertz (2008) claim that the object *sakwa* ‘apple’ must be case-marked by virtue of the fact that it is focused.

Kwon and Zribi-Hertz contend that unmarked subjects, like unmarked objects, can be construed as neither active topics nor as foci, and always occur in tense-deficient clauses construed as thetic and anchored to speech time. They illustrate the interpretive contrast between case-marked subjects and unmarked subjects with the following example:

(4) a. Sangsa-nun i-pen-ey kipwukum manhi nay-ss-e.
boss-TOP this-time-LOC donation much pay-PAST-DECL
‘(As for) the boss, he made a big donation this time.’
b. Sangsa-ka i-pen-ey kipwukum manhi nay-ss-e.
boss-NOM this-time-LOC donation much pay-PAST-DECL
‘This time a/the boss made a big donation.’
c. Sangsa i-pen-ey kipwukum manhi nay-ss-e.
boss(-NOM) this-time-LOC donation much pay-PAST-DECL
‘This time the boss has made a big donation.’ (Kwon and Zribi-Hertz 2008: 289)

According to Kwon and Zribi-Hertz, while the nominative-marked subject in (4b) can be construed as indefinite, the unmarked subject in (4c) can only be construed as discourse-accessible. This claim, however, is problematic, as we will explain later in this section.

Kwon and Zribi-Hertz account for the correlation between the definiteness effect of bare subjects[1] and the tense deficiency of bare subject clauses in terms of f-structure visibility. Due to their tense deficiency, clauses containing unmarked subjects are pragmatically anchored to speech time, standing as a default option. Assuming that tense anchoring provides an existential quantifier, Kwon and Zribi-Hertz can derive the definiteness of unmarked subjects from tense deficiency. Since unmarked subjects are invisible in f-structure and hence are neither active

[1] Kwon and Zribi-Hertz use the term ‘bare’ arguments to refer to those occur without the functional particles signaling structural case.

topics nor foci, their referents may be construed only as discourse-accessible, that is, as definite, or as rigid designators.

Kwon and Zribi-Hertz's information structure-based account is theoretically attractive in that it derives a range of observed properties of CE from a single f-structure property of arguments, i.e., f-structure visibility. However, not all of their claims are supported when they are tested empirically. Let us take the focus effect of object case marking first. H. Lee (2006a) conducted an elicitation experiment with 132 native speakers of Korean. Each participant was asked to fill in a questionnaire, which contained short conversations between two speakers, providing contexts for the choice of case-marked and unmarked forms of an object. The participants had to choose as spontaneously as possible between the two object forms in the given contexts. Table 1 below shows the relative frequency of case-marked and unmarked objects according to the factor 'focus type':

Table 1. Interaction between focus type and object form

Focus type	Accusative-marked object	Unmarked object
Contrastive	3109 (59%)	1531 (29%)
Non-contrastive	2161 (41%)	3749 (71%)

As shown in Table 1, the majority of the contrastively-focused objects in the data (59%) are case-marked. Nevertheless, the data do not provide a support for Kwon and Zribi-Hertz's claim that case marking is obligatory for focus objects. Note that more than 70% of the non-contrastive focus objects are not case-marked and that nearly 30% of the contrastively focused objects are not case-marked. This pattern of results suggests that it would be premature to offer any categorical generalization about the correlation between focality and CE.

Let us now look at subject marking. When we consider subject marking, Korean presents a serious problem for the account of CE along the lines of f-structure visibility. Although transitive subjects generally cannot be unmarked when they are under narrow focus, they may be unmarked when they are active topics. This has been demonstrated convincingly by Oh (2009), who argues that the topic marker *-(n)un* and zero pronouns/unaccented unmarked NPs tend to encode different types of topic. According to Lambrecht and Michaelis (1998), topics can be classified into ratified and unratified topics. Topics are ratified when their topic role is assumed to be taken for granted or readily expected by the addressee at the time of utterance;

topics are unratified when their topic role is not taken for granted or is not readily expected by the addressee at the time of utterance. (5) below is an example that contains ratified topics:

(5) Context: A is calling his mother (B) and asking about his son Minho, who is staying at his grandmother's place during summer vacation.

A: Emma, Minho-nun ettay-yo? ∅ pap cal meke-yo?
mom, Mihno-top how.is he meal well eat
'Mom, how is Minho? Is he eating well?'
B: Ung. ku nyesek yekise cal mek-ko cal cinay.
yes that kid here well eat-conj well get.along
'Yes, He is eating well and having a good time here.' (Oh 2009: 608-609)

Here A's second sentence has a zero pronoun which refers to *Minho*. Because the referent of this zero pronoun was already introduced in the previous sentence and is now active, it easily qualifies as a ratified or expected topic at the time of utterance. In B's utterance, the same referent is expressed as an unmarked NP. Here again, the referent is active and ratified: its topic status is expected and taken for granted by the addressee. The sentences in (5) thus show that active topic subjects can naturally occur as unmarked NPs in Korean.

The asymmetry between subjects and objects in the acceptability and frequency of CE presents another serious problem for Kwon and Zribi-Hertz's account, where bare subjects and objects in Korean equally fail to be visible at f-structure. However, analysis of naturally occurring data and acceptability judgment data clearly shows that there exists case ellipsis asymmetry between subject and object in speakers' production and intuitive judgments: in general, nominative case markers are harder to omit than accusative case markers (T. Kim 2008; H. Lee 2010), and this general asymmetry extends to argument NPs that are in focus. CE on contrastively focused direct objects occurs naturally, whereas CE on contrastively focused transitive subjects is unnatural whether the subject is contrastively focused or not (H. Lee 2010).

The claims of Kwon and Zribi-Hertz's (2008) analysis are further disputed by Chung (2010), who examined native judgments of their data such as (3) and (4) using an acceptability judgment task. Native speakers of Korean of varying age groups not only allowed the interpretation that should not be possible under the focus-structure analysis but also sometimes rejected the interpretations that should be acceptable in the focus-structure analysis. As such, Chung argues that information structure seems to be an important factor in CE, but it cannot be the only factor that determines CE.

In this section, I have shown that Kwon and Zribi-Hertz's information structure-based account of CE is inadequate to account for (i) the variability of the correspondence between discourse functions and argument expressions, and (ii) the asymmetry between subjects and objects in the acceptability and frequency of case ellipsis.

3. Usage probability and case ellipsis

In this section, I will discuss an alternative account of case ellipsis in terms of frequency/probability of use. In probabilistic models of grammar (Boersma and Hayes 2001; Bod, Hay and Jannedy 2003; Bresnan and Ford 2010), grammatical constraints are defined in terms of graded preferences, weights or rankings, rather than categorical or discrete levels of grammaticality. These models are well-suited to account for CE because they can describe syntactic phenomena in terms of grammaticality that emerges from preferences that develop over phrases and constructions. In turn, such preferences can be linked to factors that affect processing difficulty, e.g., frequency/probability of use, prototypicality, etc.

Haspelmath (2008) argues that any efficient sign system in which costs correlate with signal length follows the Zipfian principles in (5) (see also Bybee and Hopper (2001) and Hawkins (2004)).

(5) a. The more predictable a sign is, the shorter it is.
b. The more frequent a sign is, the shorter it is.

Jaeger (2006) and Levy and Jaeger (2007) propose the principle of Uniform Information Density (UID) as a possible theoretical explanation of the effect of the principles in (5) on syntactic reduction. In accordance with information theory

(Shannon 1948), information is measured in such a way that the more probable an item in context, the less informative it is, and conversely the less probable, the more informative. If the rate at which information is conveyed in the speech stream is roughly constant, then more predictable words, which carry less information, should take less time to produce during production than less predictable words. The efficiency of this strategy for communication over a speech channel lies in the fact that it allows utterances to be shorter and easier to produce without reducing the less predictable words the hearer would have the most difficulty reconstructing.

As a general computational strategy that derives language use at all levels of linguistic representation, UID holds that speakers are approximating optimal production by aiming to produce utterances with uniform information density. Thus, UID predicts that the choices speakers have to make when they encode an intended message into an utterance are at least partially determined by information density: if one way to convey a message leads to more uniform information density than another way to convey the same message, the variant with a more uniform distribution of information is preferred (Jaeger 2006; Levy and Jaeger 2007).

Jaeger (2010) argues that omission of case markers for more predictable phrases and using case markers to mark less probable phrases have a processing advantage: when speakers use case markers to mark less probable phrases, they can buy more time to produce syntactic elements that are difficult to process and spread information on the phrase's grammatical and discourse function over a longer time, thereby leading to more uniform information density compared to leaving it unmarked. Thus, from the perspective of usage probability, the presence of case markers can be interpreted as a signal to expect the unexpected, a rational exchange of time for reduced information density or a meaningful delay.

The sentence processor's preference to uniformly distribute information across linguistic signals for increased processing efficiency (by using an extra morpheme or word to mark less probable phrases) is likely to have been grammaticalized as probabilistic linguistic constraints that penalize zero marking for rare types of argument. This view of case marking can also account for the fact that in general, CE for subject occurs less frequently and is also less acceptable than CE for object (T. Kim 2008; H. Lee 2010). It has long been observed that explicit NPs occur more frequently as objects than as transitive subjects, whereas argument NPs are omitted more frequently when they are transitive subjects. The tendency to pronominalize or to omit (i.e., ‘zero- pronominalize’) transitive subjects more often than intransitive subjects and objects seems to be universal across human languages, especially in conversational speech (Givón 1983; DuBois 1987; Comrie 1989; Lambrecht 1994).

Given the high frequency of overt realization of object NPs and the rarity of overt realization of subject NPs, it is not surprising that CE is more acceptable for the more frequent type of explicit NPs, i.e., overt objects, whereas case marking is more acceptable for the rare type of explicit NPs, i.e., overt subjects.

4. Topic predictability and case ellipsis

This section presents new evidence for the probability-based account of CE from an analysis of encoding patterns of topical arguments in conversation data.

The conversation data for this study come from sixty-two hours of audio-recorded conversation between four pairs of native speakers of Korean, who were born and raised in Korea, and enrolled in the university at the time of the recording. The paired participants agreed to have tape recorders placed in their apartments for periods ranging from four to ten days in July 2011. Because they operated the recorders manually, uninterrupted recording lengths ran from one to four hours. The apartments were all small one-bedroom or two-bedroom units, and the recorders picked up all kitchen and living room conversation as well as louder talk from the bedroom and bath. The participants — four females and four males — were between the ages of 21 and 29, and were mutual friends. The casual nature of the conversation is clearly indicated by the predicate form which they used in their conversation; they used the plain (i.e., casual) form of predicates consistently, except for one pair, in which the distal (i.e., polite) forms were mixed with the plain forms.

The recorded conversations were transcribed by two Korean graduate students who majored in linguistics and prepared in terms of clausal units. Utterance boundaries were identified according to pause and the transcribed text was divided into units accordingly. Pause could be identified fairly clearly and was found to be consistent index to identify utterance boundaries. However, a pause caused morphosyntactically unnatural divisions occasionally, for example, dividing a sentence final particle and the predicate to which it is attached, and in these cases the pause was ignored. After utterance units were identified by pause, they were further divided into clausal units. Complex sentences are divided into clausal units, regardless of subordination types. Hence, an adverbial subordinate clause, a noun complement clause, and a nominalized clause were all considered as separate clausal units.

The procedures outlined above produced about 1000-2000 clausal units for each conversation pair, and 7890 clausal units for all pairs in total. These clausal units were manually coded for predicate type by the same graduate students who transcribed the recorded conversations and then checked by the author. Predicate types were classified as (i) two-place or higher predicates which encompass transitives[2] with direct object NPs, transitives with sentential complements and ditransitive predicates and (ii) one-place predicates which include intransitives, non-verbal predicates (adjectival and nominal) and passives (with or without the *by*-phrase). For the purposes of this study, the main interest was in transitive predicates with direct object NPs, which is why we made a decision not to include transitives with sentential complements, ditransitives and sentence fragments such as bare NPs and interjections in the analysis.

Table 2 shows all argument types in terms of encoding and argument types. In terms of argument type, the subject of one-place predicates is by far the most frequent argument type (4384 tokens; 44.2% of the total). In terms of the encoding type, argument ellipsis is most frequent (5483 tokens; 50.1% of the total), followed by particle ellipsis (2533 tokens; 23.1% of the total). In other words, the encoding types which are more or less attenuated in form are by far dominant over the three explicit encoding types; the two types of ellipsis add up to 73.2% of the total.

[2] When coding Korean clauses, nominative-marked NPs were coded as subjects, but constructions that arguably contained nominative-marked objects were tallied. They were of three types: (i) existential-possessive constructions with the possessor in the dative or nominative case and the possession in the nominative, (ii) psych predicates with the experiencer in the dative or nominative case and the theme in the nominative, and (iii) locative predicates with the location in the dative or nominative case and the theme in the nominative.

Table 2. All argument tokens by encoding and argument types

Encoding type	Subject of Transitives (A)		Subject of one-place predicates (S)		Object of transitives (O)		Total	
	No.	%	No.	%	No.	%	No.	%
∅	2380	77.9	2370	49	733	24	5483	50.1
NP-∅	214	7	822	17	1497	49	2533	23.1
NP-NOM	255	8.3	1204	24.9	102	3.3	1561	14.3
NP-ACC	-		-		464	15.2	464	4.2
NP-*(n)un*	173	5.7	218	4.5	92	3	483	4.4
NP-other Particle	34	1.1	220	4.6	168	5.5	422	3.9
Total	3056	100	4834	100	3056	100	10946	100

We also observe that there is asymmetry in the token frequency of argument ellipsis in terms of argument types: it is most frequent with the subject of transitive predicates than with other two types. Transitive subjects exhibit an ellipsis rate that is almost 30% higher than that of the subjects of one-place predicates and almost 54% higher than that of direct objects. This pattern is consistent with previous observations that argument ellipsis in Japanese and Korean is most commonly associated with transitive subjects (Fry 2001; M. Kim 2001; Shimojo 2005; T. Kim 2008; Ueno and Polinsky 2009). In fact, the tendency to pronominalize or to omit (i.e., 'zero-pronominalize') subjects more often than objects seems to be universal across human languages, especially in conversational speech (Givón 1983; Du Bois 1987; Lambrecht 1994; Ueno and Polinsky 2009).

On the other hand, the majority of direct objects (76%) are overtly expressed. In terms of the token frequency of encoding type, particle ellipsis is the most frequent encoding type for direct objects (49%), followed by argument ellipsis (24%) and explicit case marking (18.5%).

Table 2 also shows that the nominative case markers *-i/-ka* occur more frequently with the subject of one-place predicates than with the subject of transitive predicates. Among the 1561 NPs marked with *-i/-ka*, 77.1% occur with one-place predicates, and only 16.3% are transitive subjects. The very low frequency of case marking in transitive subjects can be attributed to a phenomenon referred to as 'preferred argument structure' in the literature (Du Bois 1987; Ashby and Bentivoglio 1993; Kärkkäinen 1996). Crosslinguistically and in Korean, A (transitive subject) is a slot for given information. Such information is typically unexpressed in Korean (so-called 'argument ellipsis' or 'zero anaphora'), resulting in very few As that are overtly realized (M. Kim 2001).

Having laid out the overview of the distribution of argument types in terms of encoding and argument types, the remainder of this section presents the token counts by types of topic. Of the 10946 argument tokens, 9567 were analyzable in the sense that they were clearly referential (i.e., not grammatical items, question words, etc.) They were then coded for their topic status. For this purpose we have adopted the distinction between ratified and unratified topic, originally proposed by Lambrecht and Michaelis (1998). A topic NP is classified as a ratified or expected topic when its referent has been under discussion throughout the conversation or when it is saliently present in the speech situation (e.g., as a speaker and a hearer). A topic NP whose referent is simply accessible (not previously occupied the topic role or has competitors for topic status) is classified as an unratified topic. The example below, repeated here as (6), is an example that contains ratified topics:

(6) Context: A is calling his mother (B) and asking about his son Minho, who is staying at his grandmother's place during summer vacation.

A: Emma, Minho-nun ettay-yo? ∅ pap cal meke-yo?
mom, Mihno-TOP how.is he meal well eat
'Mom, how is Minho? Is he eating well?'

B: Ung. ku nyesek yekise cal mek-ko cal cinay.
yes that kid here well eat-CONJ well get.along
'Yes, He is eating well and having a good time here.'

Here A's second sentence has a zero pronoun which refers to *Minho*. Because the referent of this zero pronoun was already introduced in the previous sentence and is now active, it qualifies as a ratified or expected topic at the time of utterance.

A topic expression which has a low degree of predictability often bears a linguistic mark, which may be viewed as a topic-establishing device. In Korean, such topic expressions are typically marked with −*(n)un*; they can also be realized as a case-marked NP or an unmarked NP, as illustrated in (7):

(7) A: ∅ Minswu kiek hay?
you Minsoo remember
'Do you remember Minsoo?'
B: Ung. pwuin-un/pwuin-i/??pwuin acwu chakhay poi-ess-ci.
yes. wife-TOP/wife-NOM/wife very be.nice look-PAST-DECL
'Yes, his wife looked very nice.'

In this response, the VP *acwu chakhay poi-ess-ci* 'looked very nice' expresses the focus. The subject *pwuin* 'wife' is as yet an unratified topic: Minsoo's wife has not been mentioned before (but is accessible as a topic via the marriage frame).

The B response of (8) illustrates a different kind of unratified topic.

(8) A: ∅ Minswu kiek hay?
you Minsoo remember
'Do you remember Minsoo?'
B: Ung. ku saram/∅/??ku saram-un/??ku saram-i acwu chakhay
yes. he/he/he-TOP/he-NOM very be.nice
poi-ess-ci.
look-PAST-DECL
'Yes, he looked very nice.'

In this response, the subject *ku saram* 'that person' is an unratified topic: it was activated and introduced in focus position in the previous utterance, but is not yet taken to be an approved topic of discussion. Unlike inactive unratified topics illustrated in (7B), active unratified topics can be naturally omitted or occur as an unmarked NP.

As topic predictability can be thought of as scalar, the account of argument encoding based on this makes a clear prediction about the rates of CE for different subtypes of topic NPs. Due to economy motivation, more readily predictable or expected entities tend to be referred to by shorter or less complex form: Thus, preference for reduced forms (argument ellipsis and CE) are expected to increase relative to the degree of ratification or predictability of the topic NPs:

(9) Prediction of topic predictability:
Form reduction ←——————— Non-reduction
Ratified topic > Unratified active topic > Unratified inactive topic
High predictability ←——————— Low predictability

In contrast, Kwon and Zribi-Hertz's account of CE, where a single factor of focus-structure visibility determines categorical distinction between case-marked vs. unmarked NPs, does not predict any difference in the frequency of CE for different subtypes of topic arguments, for under their analysis, both unmarked subjects and objects can be construed as neither active topics nor foci. Instead, their account predicts that active topics do not occur as unmarked NPs.

We further predict that there is asymmetry in the encoding of topic subjects and objects. Because topic is a rare property for objects, topic objects are expected to show a greater preference for non-reduction (i.e., overt realization and explicit case marking) even when their topic status is ratified. In contrast, accounts of CE such as Kwon and Zribi-Hertz's, which do not distinguish between subject CE and object CE, do not predict any difference in the frequency of CE for (topic) subjects and objects.

Tables 3 and 4 show the distribution of tokens of ratified and unratified topics over the subject (As and Ss) and the direct object (O) in the data.

Table 3. Tokens of topic subjects by encoding and topic types

Encoding type	Ratified, active topic		Unratified, active topic		Unratified, inactive topic		Total	
	No.	%	No.	%	No.	%	No.	%
∅	1714	59	-	-	-	-	1714	46.6
NP-∅	610	21	12	2.2	-	-	622	16.9
NP-NOM	378	7	16	2.3	12	5.4	406	11
NP-*(n)un*	203	13	524	94.9	209	94.6	936	25.5
Total	2905	100	552	100	221	100	3678	100

Table 4. Tokens of topic objects by encoding and topic types

Encoding type	Ratified, active topic		Unratified, active topic		Unratified, inactive topic		Total	
	No.	%	No.	%	No.	%	No.	%
∅	109	30.8	-	-	-	-	109	17
NP-∅	103	29.1	-	-	-	-	103	16
NP-ACC	60	16.9	2	6.5	11	4.3	73	11.4
NP-*(n)un*	82	23.2	29	93.5	246	95.7	357	55.6
Total	354	100	31	100	257	100	642	100

We can see that of all 4,320 tokens of topic expressions, the great majority (85.1%) goes into the subject, and the great majority of topic subjects are a ratified topic (80%). In the case of topic objects, however, the proportion of ratified vs. unratified topics was more balanced (ratified: 55.2%; unratified: 44.8%). Hence, it seems possible to regard the subject as the main locus for ratified topics.

Turning now to encoding types for the three subtypes of topic which differ in the degree of topic predictability, we can observe the following. The majority of the active and ratified topics in our data are encoded with forms that are more or less reduced or attenuated: the great majority of active and ratified topic subjects are omitted (59%) or unmarked (21%), and the majority of active and ratified topic objects are so (argument ellipsis: 30.8%; CE: 29.1%). The high frequency of CE for active topic subjects and objects contrasts with the prediction of Kwon and Zribi-Hertz (2008) because on their account, active topics are not predicted to occur as unmarked NPs. In contrast, all tokens of unratified topics are overtly realized. For both subjects and objects, -*(n)un*-marking is the most frequent encoding type for unratified topics, followed by explicit case marking and CE. When collapsing argument ellipsis and CE into a single category and -*(n)un*-marking and explicit case-marking into another category, the relationship of the two variables ―topic predictability and form ―reaches significance at the 0.05 level for both argument roles (subject: X^2 = 1621.3, df = 1, $p < 0.05$; object: X^2 = 257.5, df = 1, $p < 0.05$). Hence, it seems possible to regard argument ellipsis as the most preferred encoding type for the most predictable type of topic, i.e., active and ratified topic, and -*(n)un*-marking as the most preferred encoding type for the topics that are less predictable.

These results confirm our prediction that preference for form reduction will increase relative to the degree of topic predictability. The data also provide empirical evidence in support of our second major prediction that there is asymmetry in the encoding of topic subjects and objects. Active and ratified topic objects show a stronger overall preference for non-reduction (i.e., overt realization and explicit case marking), compared to active and ratified topic subjects. As shown in Tables 3 and 4, 40.1% of all tokens of active and ratified topic objects are overtly realized or explicitly case-marked, whereas only 20% of active and ratified topic subjects are so. The difference between the two cells in Table 5 is significant with the Chi-Square test at the 0.05 level (X^2 = 265.3, df = 1, p < 0.001). As pointed out earlier, the stronger overall preference for non-reduction exhibited by topic objects naturally follows from the relative infrequency of topic as objects, but does not follow from Kwon and Zribi-Hertz's account, which does not distinguish between subject CE and object CE.

Table 5. Tokens of non-reduced topic subjects and objects

Encoding type	Subject (No.)	Object (No.)	Total
Non-reduction	581	141	722

However, the distribution of CE shows a considerable overlap with that of argument ellipsis, and this raises the question whether and how the status of non-case-marked topic arguments and arguments encoded in zero anaphor can be characterized by degree of topic predictability. In order to find an answer to this question, I have measured the referential distance and potential interference of the topic subjects and objects under discussion here. Referential distance (RD) is one of the quantitative measurements introduced by Givón (1983) for assessing topic continuity/predictability, and used frequently in the subsequent text analyses in a variety of languages. RD is a linguistic distance in clausal units measured backward to the most recent coreferential expression, including that encoded in zero anaphor. For example, an RD of 1 indicates that a referent in question was represented in the immediately preceding clausal unit, and an RD of 2 indicates that there is one intervening clausal unit between the current representation of the referent and the antecedent. Other things being equal, a referent whose RD is 1 may be considered to be more activated than a referent whose RD is greater than 1 (Shimojo 2005).

A second measure I will use to assess the referential predictability of topic subjects and objects in this study is potential interference (PI). This measure assesses the effect of other semantically compatible NPs within the immediately preceding discourse environment on the referential predictability of an NP (Givón 1983). Following Watanabe (1986), for each of the NPs, I counted the number of other NPs that are semantically compatible with their predicate in question and appeared between the present and last mention of the same referent in the discourse. The results scaled between 0 (no semantically compatible NP) and 4 (four such NPs), and the mean are counted as the PI value for the NP.

The results of the RD measurement for the encoding types are given in Tables 6 and 7. As Table 6 shows, the mean values of RD of subject referents separate elided arguments from the other three encoding types in the predictable direction. Argument ellipsis and nominative marking represent the two extremes of the saliency scale −argument ellipsis on the salient side and nominative-marking on the nonsalient side. The mean RD of unmarked topic subjects and −*(n)un*-marked topic subjects is approximately in the middle of the saliency scale.

Table 6. Encoding types in terms of RD of ratified topic subjects

RD	∅		NP-∅		NP-*(n)un*		NP-NOM	
	No.	%	No.	%	No.	%	No.	%
1	1389	81	301	49.3	93	45.8	64	16.9
2	214	12.5	82	13.4	28	13.8	131	34.6
3	91	5.3	83	13.6	29	14.3	63	16.7
4-6	20	1.2	73	12	37	18.2	62	16.4
7-9	0		71	11.7	12	5.9	28	7.4
10-20	0		0		3	1.5	23	6.1
20+	0		0		1	0.5	7	1.9
mean	1.28		2.69		2.86		4.07	
Total	1714	100	610	100	203	100	378	100

Despite the overall higher RD mean values of objects, the results of the RD measurements for objects are strikingly similar to those of subjects. As shown in Table 7, argument ellipsis is on the salient side and accusative-marking on the nonsalient side. Also, unmarked topic objects and −*(n)un*-marked topic objects are situated between the other two encoding types. For both argument roles, the relationship of the two variables—encoding type and RD—is significant at the 0.05 level (subject: X^2 = 458.9, df = 1, p = 0.00001; object: X^2 = 35.8, df = 1, $p <$ 0.00001), demonstrating that the overall distribution is significant.

Table 7. Encoding types in terms of RD of ratified topic objects

RD	∅		NP-∅		NP-*(n)un*		NP-ACC	
	No.	%	No.	%	No.	%	No.	%
1	76	69.7	46	44.7	36	43.9	10	16.7
2	10	9.2	20	19.4	11	13.4	11	18.3
3	17	15.6	18	17.5	15	18.4	13	21.6
4-6	6	5.5	8	7.8	13	15.9	10	16.7
7-9	0		6	5.8	4	4.9	10	16.7
10-20	0		5	4.9	1	1.2	3	5
20+	0		0		2	2.4	3	5
mean	1.62		2.94		3.07		4.83	
Total	109	100	103	100	82	100	60	100

The results of the PI measurements are shown in Tables 8 and 9. For both subjects and objects, the mean PI values clearly group unmarked arguments with elided arguments, separating both of these from case-marked arguments as well as from *-(n)un*-marked ones. When collapsing the PI values of 0-1 into a single category and 2-4 into another category, the variables of encoding type and PI show a significant association at the 0.05 level for both argument roles (subject: X^2 = 662.6, df = 1, p = 0.00001; object: X^2 = 75.4, df = 1, p = 0.00001).

Table 8. Encoding types in terms of PI of ratified topic subjects

PI	∅		NP-∅		NP-*(n)un*		NP-NOM	
	No.	%	No.	%	No.	%	No.	%
0	1371	80	456	74.8	50	24.6	75	19.8
1	257	15	83	13.6	61	30.1	113	29.9
2	34	2	43	7	61	30.1	132	34.9
3	32	1.9	2	3.3	19	9.4	49	13
4	20	1.1	8	1.3	12	5.9	9	2.4
mean	0.29		0.43		1.42		1.48	
Total	1714	100	610	100	203	100	378	100

Table 9. Encoding types in terms of PI of ratified topic objects

PI	∅		NP-∅		NP-*(n)un*		NP-ACC	
	No.	%	No.	%	No.	%	No.	%
0	82	75.2	74	71.8	12	14.6	9	15
1	15	13.8	15	14.6	24	29.2	18	30
2	9	8.3	7	6.8	27	32.9	18	30
3	2	1.8	5	4.9	16	19.5	10	16.7
4	1	0.9	3	2.9	3	3.7	5	8.3
mean	0.39		0.54		1.68		1.73	
Total	109	100	103	100	82	100	60	100

The finding of the quantitative analysis of the topic predictability of subjects and objects is summarized in (10):

(10) The index of topic predictability and encoding types:
a. RD
∅ > NP-∅ > NP-*(n)un* > NP-CASE
Salient ←————————→ Nonsalient
b. PI
∅, NP-∅ > NP-*(n)un* > NP-CASE
Salient ←————————→ Nonsalient

The two indices collectively point to the contrast between {∅, NP-∅} and {NP-*(n)un,* NP-CASE}, the former two encoding types are associated with high topic predictability, as suggested by their lower RD and PI values, whereas the latter two are associated with low topic predictability, having higher RD and PI values. Clustering of the tokens of topic arguments in the middle range of the RD scale further suggests that CE is associated with medium topic predictability in Korean. This result is also consistent with previous studies on Japanese CE including Watanabe (1986) and Suzuki (1995), who find that the tokens of unmarked object NPs cluster in the midrange of the RD scale rather than clustering at both ends of the scale.

In summary, this section has presented new empirical evidence for the probability-based account of CE. The results of the quantitative analysis of the topic predictability of subjects and objects in conversation data confirm the predictions of the current probability-based account, showing that preference for form reduction (i.e., argument ellipsis and CE) increases relative to the degree of topic predictability: argument ellipsis is the most preferred encoding type for the most predictable type, and *-(n)un*-marking is for the least predictable type; CE is associated with medium topic predictability. The results, however, are not consistent with Kwon and Zribi-Hertz's account, which predicts CE for topic arguments to be unacceptable.

The results further show that there is asymmetry in the encoding of topic subjects and objects in the predicted direction: The data also provide empirical evidence in support of our second major prediction that there is asymmetry in the encoding of topic subjects and objects. Active and ratified topic objects show a stronger overall preference for non-reduction (i.e., overt realization and explicit case marking), compared to active and ratified topic subjects. As pointed out earlier, the stronger overall preference for non-reduction exhibited by topic objects naturally follows from the relative infrequency of topic as objects.

5. Conclusion

The present study has examined previous characterizations of Korean CE in terms of the information status of argument NPs and has provided empirical evidence for a probability-based account. Previous information structural accounts of case ellipsis tended to place exclusive importance on information structure and make a categorical distinction between case-marked and unmarked NPs based on their information status without differentiating between subject and object CE. However, evidence from our experimental and conversation data demonstrates that judgments on case ellipsis are not categorical but gradient and that the frequency of argument ellipsis and case ellipsis for topic subjects and objects is correlated with the degree of topic predictability. The data also provide empirical evidence that there is asymmetry in the encoding of topic subjects and objects. Active and ratified topic objects show a stronger overall preference for non-reduction (i.e., overt realization and explicit case marking), compared to active and ratified topic subjects. The stronger overall preference for non-reduction exhibited by topic objects naturally follows from the relative infrequency of topic as objects. This evidence supports the conclusion that both the effect of including information structure and the gradient acceptability of case ellipsis can be explained in terms of probabilistic properties associated with argument NPs.

References

Ashby, William and Paola Bentivoglio. 1993. Preferred argument structure in spoken French and Spanish. *Language variation and change* 5: 61-76.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy. (eds.). 2003. *Probabilistic linguistics.*

Cambridge: MIT Press.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32: 45-86.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86: 168-213.

Bybee, Joan, and Paul Hopper, eds. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Chung, Eun Seon. 2012. Challenging a single-factor analysis of case ellipsis in Korean. Paper presented *at the Seoul International Conference on Linguistics (SICOL 2010).* Korea University, Seoul.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. Chicago: University of Chicago Press.

DuBois, John W. 1987. The discourse basis of ergativity. *Language* 63: 805-855.

Erteschik-Shir, Nomi. 1997. *The dynamics of focus structure*. Cambridge: Cambridge University Press.

Erteschik-Shir, Nomi. 2007. *The syntax/discourse interface: Information structure*. Oxford: Oxford University Press.

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

Givón, Talmy. (ed.) 1983. *Topic continuity in discourse*. Amsterdam: John Benjamins.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19: 1-33.

Hawkins, John A. 2004. *Efficiency and complexity in grammar*. Oxford: Oxford University Press.

Hinds, John. 1983. Topic continuity in Japanese. In Talmy Givón (ed.), *Topic continuity in discourse*, 43-93. Amsterdam: John Benjamins

Jaeger, T. Florian. 2006. *Redundancy and syntactic reduction in spontaneous speech*, PhD Dissertation, Stanford University.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61: 23-62.

Kärkkäinen, Elise. 1996. Preferred argument structure and subject role in American English conversational discourse. *Journal of Pragmatics* 25: 675-701.

Kim, Mi-Kyung. 2001. Preferred information structure in conversational Korean. *Discourse and Cognition* 8: 21-41.

Kim, Taeho. 2008. *Subject and object markings in conversational Korean.* PhD Dissertation, State University of New York at Buffalo.

Ko, Eon-Suk. 2000. A discourse analysis of the realization of objects in Korean. *Japanese/Korean Linguistics* 9: 195-208. Stanford: CSLI Publications.

Kwon, Song-Nim, and Anne Zribi-Hertz. 2008. Differential functional marking, case, and information structure: Evidence from Korean. *Language* 84: 258-299.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Lambrecht, Knud, and Laura Michaelis. 1998. Sentence accent in information questions: Default and projection. *Linguistics and Philosophy* 21: 477-544.

Lee, Duck-Young. 2002. The function of the zero particle with special reference to spoken Japanese. *Journal of Pragmatics* 34: 645-682.

Lee, Hanjung. 2006a. Iconicity and variation in the choice of object forms in Korean. *Language Research* 42: 323-355.

Lee, Hanjung. 2006b. Parallel optimization in case systems: Evidence from case ellipsis in Korean. *Journal of East Asian Linguistics* 15: 69-96.

Lee, Hanjung. 2010. Explaining variation in Korean case ellipsis: Economy versus iconicity. *Journal of East Asian Linguistics* 19: 291-318.

Levy, Roger, and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt, and Thomas Hoffman (eds.), *Advances in Neural Information Processing Systems (NIPS)* 19: 849-856. Cambridge: MIT Press.

Masunaga, Kiyoko. 1988. Case deletion and discourse context. In Willliam Poser (ed.), *Papers from International Workshop on Japanese Syntax*, 145-156. Stanford: CSLI Publications.

Matsuda, Kenjiro. 1996. *Variable zero-marking of o in Tokyo Japanese*. PhD Dissertation, University of Pennsylvania.

Minashima, Hiroshi. 2001. On the deletion of accusative case markers in Japanese. *Studia Linguistica* 55: 175-190.

Oh, Chisung. 2009. -*Un/nun* as an unratified topic marker in Korean. *Korean Journal of Linguistics* 34: 603-624.

Shannon, Claude. 1948. A mathematical theory of communications. *Bell Systems Technical Journal* 27: 623-656.

Shimojo, Mitsuaki. 2005. Argument encoding in Japanese conversation. New York: Palgrave Macmillan.

Suzuki, Satoko. 1995. The function of topic-encoding zero-marked phrases: A study of the interaction among topic-encoding expressions in Japanese. *Journal of Pragmatics* 23: 607-626.

Tsutsui, Michio. 1984. *Particle ellipsis in Japanese*. PhD Dissertation, University of Illinois at Urbana-Champaign.

Ueno, Mieko and Maria Polinsky. 2009. Does headedness affect processing? A new look at the VO-OV contrast. *Journal of Linguistics* 45: 675-710.

Vallduví, Enric and Maria Vilkuna. 1998. On rheme and kontrast. In Peter Culicover and Louise McNally (eds.), *The Limits of Syntax (Syntax and Semantics 29)*, 79-108. San

Diego: Academic Press.

Watanabe, Yasuko. 1986. Two kinds of ellipsis in Japanese discourse: A quantitative text study. *Studies in Language* 10: 337-351.

Hanjung Lee
Department of English Language and Literature
Sungkyunkwan University
25-2 Sungkyunkwan-ro, Jongro-gu, Seoul 03063, Korea
E-mail: hanjung@skku.edu

Received: 2015. 10. 23.
Revised: 2015. 12. 07.
Accepted: 2015. 12. 07.