The Role of Prosody and the Uniform Information Density Effect in Korean Particle Omission

Haseon Park(201401381)

1. Introduction

Particle omission in the Korean language is an abstruse phenomenon, for which decades of attempts in scientific communities, including semantic and pragmatic approaches, have not discovered any single explanation without exceptions concerning the variation within a register. In this thesis, we examine by corpus analysis the validity of two non-semantic and non-pragmatic hypotheses that account for it: Kiaer & Sin (2012)'s prosodic hypothesis on Korean particle omission and Levy & Jaeger (2007)'s Uniform Information Density hypothesis, which is based on the "surprisal" theory of sentence processing. We also argue that none of the phonological, semantic, pragmatic, registerial, or processing motivations can be the sole factors involved in Korean particle omission, and thus we need a quantitative model combining all the effects from several different motivations in order to properly understand the phenomenon by interaction among them.

2. Previous Studies and Theoretical Backgrounds

2.1. Descriptive corpus researches on Korean Particle Omission

While the functions and motivations of Korean particle omission is not abundantly researched in a quantitative manner, several descriptive corpus studies about the phenomenon can be enumerated. Ross, Dickinson & Sun-Hee Lee(2013) analyzed Korean learners' particle omission errors with Learner Korean Corpus and built a particle presence predictor (as a component of their error detector/corrector) using a Conditional Random Fields methods, showing a result implying that particle presence/omission is predictable at some extent given words and POS tags in the surrounding context. Hong (2010) discovered that particled-omitted NPs are asymmetrically distributed in the sense that the frequency of particle omission is considerably lower in nominative NPs than in accusative ones. In Sun-Hee Lee & Song (2012)'s paper, an annotation scheme for the realization and ellipsis of Korean particles is suggested and followed by a corpus analysis and linguistic discussion. The most notable outcome from their contribution in our research's point of view is that they measured and statistically tested the effect of registers and genres in Korean particle omission, demonstrating the significant, but not absolute, difference between spoken (12% particle-omitted) and written (2% particle-omitted) language. They also discussed and reviewed several possible linguistic properties and motivations of particle omission other than registers or genres in a qualitative manner but did not provide systematic evidence for them by corpus analysis or experiments.

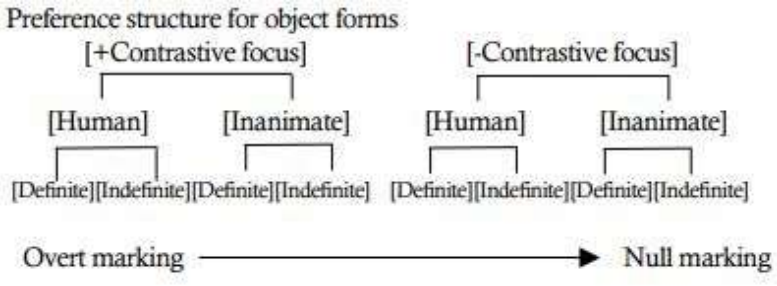
2.2. Semantic/Pragmatic Analyses

2.2.1. Ko (2000)'s Contrastive Focus Hypothesis

Ko (2000) argue that the realization of the accusative case marker has a function of marking contrastive focus, based on several theoretical backgrounds and statistics. According to her explanation, for example, *os*("clothes") in (1a) is an entity that is already shared by two interlocutors' minds, with no contrastive focus, and therefore its particle is omitted whereas (1a)'s *os* has alternative options that can be expected, causing it to be contrastively focused and therefore its case marker is not dropped. However, Kiaer & Shin (2012) suggested (2) and pointed out that Ko (2001)'s theory cannot account for a number of particle-omitted cases with obvious contrastive focus.

(1) a. A: Nana-hanthey os ponayssni? [Did you sent clothes to Nana?]
B: mian. kkamppakhaysse. [I'm sorry. I forgot to.]
b. A: sopho an-ey mwe issni? [What is in the packet?]
B: Nana-ka OS-ul ponaysse. [Nana sent CLOTHES.]
(2) A: sakwa mekullay? pay mekullay? [Do you wanna eat an apple or a pear?]
B: SAKWA mekullay. [I wanna eat an APPLE.]

2.2.2. Hanjung Lee (2006)'s Hierarchy of Object Case Ellipsis



<Figure 1> Hanjung Lee (2006)'s Hierarchy of Object Case Ellipsis

Hanjung Lee(2006, 2010, 2011) conducted experimental and corpus studies in order to identify the semantic motivation of Korean object case ellipsis and conclusionally theorized a hierarchy of object case ellipsis by the features [±Contrastive Focus], [±Human/Animate] and [±Definite]. According to this theory, accusative nouns with the features [+Contrastive Focus], [+Human], [+Definite] have only slim chances of having its case marker dropped, whereas nouns of [-Contrastive Focus], [-Animate], [-Definite] are highly likely to have its accusative case marker omitted. Even with this empirically well-researched theory, however, Kiaer & Shin (2012) suggested (3) as a counterexample. *tangsin*("you") in (3) has [+Contrastive Focus, +Human, +Definite], which corresponds to the minimum probability of particle omission according to the theoretical prediction. Nevertheless, *tangsin* without the case marker is actually preferred.

(3) [+Contrastive Focus, +Human, +Definite]
A: kunal kekin Sohuy mannale wassten kecyo?
[On that day, you went there to meet Sohuy, right?]
B: tangsin-ul/tangsin mannale kan keci, Sohuy-nun musun.

[It was you that I went to meet, not Sohuy]

2.2.3. Sungbom Lee (2006)'s Neo-Gricean Hypothesis

Sungbom Lee (2006) proposed that, from the perspective of neo-Gricean pragmatics, the case-marker deletion of accusative NPs strongly reveals metapragmatic functions and prefer to be interpreted with marked, habitual or repetitive meanings, on the assumption that Korean's accusative case marker is unmarkedly realized. The following table is from Sungbom Lee (2006), supplemented with the translation.

son-ul pota (repair)	son pota (chastise)	Higher ↑ Level of conventionality ↓ Lower
khun kho-rul tachita (break one's nose)	khun kho tachita (make failure)	
kho-rul peyekata (cut a nose)	kho peyekata (damage)	
pay-rul thata (ride a boat)	pay thata (be a sailor)	
ppyam-ul chita (slap cheek)	ppyam chita (measure up to)	
kwutwu-rul takkta (polish shoes)	kwutwu takkta (be a shoe shiner)	
noray-rul puruta (sing a song)	noray puruta (sing habitually)	

<Table 1> Unmarked/Marked interpretations induced by realization/ellipsis of the case marker

Kiaer & Shin (2012) put a counterexample on this hypothesis as well; that is, all the nouns in the sentences of (4) having no case markers do not lead to any marked interpretations.

(4) a. mwe hay? [What are you doing?]
b. pap meke. [I'm eating the meal.]
c. os pese. [I'm putting off the cloth.]
d. ramyen meke. [I'm eating ramyeon.]
e. meyil cheykhuhay. [I'm checking emails.]

2.3. Kiaer & Shin (2012)'s Prosodic Hypothesis

After observing the previously mentioned counterexamples to each hypothesis, Kiaer & Shin (2012) maintained that the previous discourses' over-concentration on semantic or pragmatic approaches had left some issues fundamentally unsolvable and, as an alternative point of view, one also need to pay attention to the prosodic effects caused by presence or omission of the accusative case marker. They speculated that whether an accusative case marker is dropped or not is also related to the tendency to build prosodic units with more natural lengths. According to Shin (2011)'s statistic, the average length of accentual phrases in Korean is 3.28. Kiaer & Shin (2012) argues that, because an accentual phrase with an overly short or long length raises prosodic cumbersomeness, there is a tendency to avoid it by utilizing optional realization/omission of case markers. For example, (5a) and (6b) are preferred over (5b) and (6a), respectively, and their preferences are explained on the observation that the numbers of syllables in (5a) and (6b)'s corresponding accentual phrases are closer to 3.28 than the ones in (5b) and (6b)'s.

(5) a. //os peseyo//. [AP composed of 4 syllables]
b. //os-ul peseyo//. [AP composed of 5 syllables]
(6) a. //os// pese posikeysssupnikka? [AP composed of 1 syllable]
b. //os-ul// pese posikeysssupnikka? [AP composed of 2 syllables]

They also proposed that Korean case marker omission is induced not solely by semantic/pragmatic motivations but by the combination of several different motivations, including Phonology. However, they only presented few example sentences that may be related to prosodic cumbersomeness and, on the other hand, be explainable by other interpretations, for example, by the differences in registers, and did not provide any experimental or quantitative evidence that supports their hypothesis. Therefore, there remains a scientific need to statistically testing their proposal.

2.4. Jaeger & Levy (2007)'s Uniform Information Density Hypothesis

Levy & Jaeger(2007) formulated the "Uniform Information Density" hypothesis that accounts for the optional syntactic reduction in general from the perspective of Shannon information content or "surprisals," and empirically proved its validity in the case of *that*-omission in the English language. Surprisal is proposed by Hale(2001) as a quantification of the cognitive effort required to process a word in a sentence and defined by the following formula:

$$Surprisal = \log\frac{1}{P(x_t|x_{t-1})} = -\log P(x_t|x_{t-1}) \approx -\log P(w_t|w_{t-1}, w_{t-2}, \ldots)$$

<Formula 1> Definition of surprisal

In this formula, w_t stands for the next word at time t, w_{t-i} for the word that is i word before w_t. $P(x_t|x_{t-1})$ is the probability that w_t appears as the next word when one has heard $w_1, w_2, \ldots, w_{t-2}, w_{t-1}$ at time t. The lower this probability is, the more quantity of information is conveyed. Sentence processing difficulty at a point of time is known to be proportional to the amount of information conveyed at the word, which is also known as "surprisal." This is because human brains optimize the efficiency and reliability of sentence processing by always predicting what will come as the next word and preparing for it. The lower the probability of a given next word, the less prepared the human mind is for, and therefore the more difficulty arises in processing it, which intuitively resembles the delay effects of surprisal from unexpected events. Surprisal is settling as one of the main theoretical concepts in Quantitative linguistics that plays a key role in explaining many psycholinguistic phenomena, such as garden path sentences (Levy, 2008), and typological tendencies, such as SOV and SVO's dominant appearance across the languages in the world (Maurits et al., 2010; Gibson et al., 2019; Hahn et al., 2020).

Levy & Jaeger (2007) claims that sentences in natural languages prefer uniform information density if possible in order to facilitate sentence processing and to utilize the channel capacity of language efficiently; that is because, in terms of channel coding, the larger variance the information density over time has, the more part of channel capacity is being wasted. From this general hypothesis,

one can derive a specific hypothesis that, in order to minimize the variance in the information density over time, there should be a tendency of inserting an optional low-information word as a buffer in the context where the next word has much surprisal and conversely, deleting an optional low-information word in the context where the next word has only little amount of information. This theoretical point of view was applied to various languages' syntactic reduction such as that-omission in English (Levy & Jaeger, 2007), article omission in German (Asr & Demberg, 2015), omission of discursive connectives in English (Horch & Reich, 2016), and case-marking and word choices in Hindi (Ranjan et al., 2019).

Considering these results, "Uniform Information Density" hypothesis seems to be fruitfully applicable for our research of Korean particle omission, which may be viewed as an case of optional syntactic reduction. Thus, this thesis will also cover the test of the UID effect as it is a previously unexamined factor that account for Korean particle omission and need to be statistically tested.

3. Materials and Methods

3.1. Data

The collection of speech transcriptions (morphologically annotated) in the Sejong corpus is used. It contains 175 files, total 302,173 tokens(eojeol). The particle-realized cases are already tagged in the corpus and able to be automatically identified. However, since particle-omitted cases are not tagged, their very presences and grammatical cases cannot be identified with such convenience. Therefore, we manually classified and collected particle-omitted cases until the sample size 1,000 is achieved (by repetitive random sampling with replacement). The tagging work process is described in 3.5.

3.2. The Logit Model and the Analysis Procedures

Logistic regression is the most widely used statistical method when building and testing a model whose response variable is categorical. It models values of $P(Y|X_1, X_2, ..., X_n)$ as follows:

$$p = P(Y|X_1, X_2, ..., X_n) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}}$$

<Formula 2> General equation of an logit model

where X_1, X_2, ..., X_n are the continuous or categorical explanatory variables and Y is the binary categorical response variable, and $0 \leq p \leq 1$ as p is interpreted as a probability. A logistic regression model is also called a 'logit model' and predicts probability of whether Y is 0 or 1 given the designated set of explanatory variables X_1, X_2, ..., X_n.

Below is our specific model equation that will be used in this research.

$$\log \frac{P(x = \text{particled})}{P(x = \text{unparticled})} = \beta_{\text{intercept}} + \beta_{\text{surp}} x_{\text{surp}} + \beta_{\text{prsd}} x_{\text{prsd}}$$

<Formula 3> The model equation of this study

$x = \text{particled}$ is true when the case of x has its case marker realized, and $x = \text{unparticled}$ is

true when its case marker is omitted. And x_{surp} is the estimated surprisal and x_{prsd} is the estimated prosodic pressure. Their estimation methods are described in 3.3 and 3.4, respectively.e.

3.3. Estimation of Surprisals

The practical estimation of surprisals basically uses trigram model, formulated as follows:

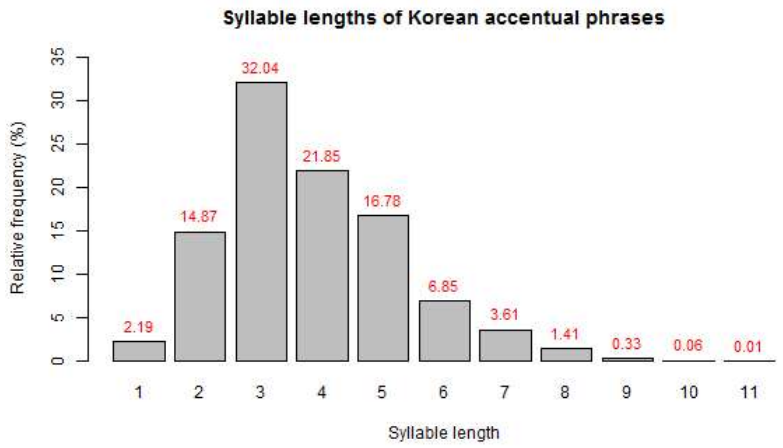
$$Surp = -\log P(w_t|w_{t-1}, w_{t-2}) = -\log\left(\frac{N(w_t, w_{t-1}, w_{t-2})}{N(w_{t-1}, w_{t-2})}\right)$$

<Formula 2 > Trigram model of surprisals

Note that we deleted all particles in w_{t-1} to prevent self-prediction, before constructing the trigram model. And, as Levy & Jaeger (2007) did, we used "hold-one-out" estimation, which means that we ruled out the very data point from the training data when calculating the prediction of the data point itself. This is also for preventing self-prediction bias, especially when the identical (n)-gram sequences are rare in the data. In cases where the other identical (n)-gram sequences are not found, we used (n-1)-gram instead. In cases where the other identical unigrams(tokens) are not found, we dropped the cases from the data, following the methodology of Levy & Jaeger (2007). Punctuation marks were regarded as words. When w_t is the second word of the sentence, bigram was used to predict it, instead of holding trigram by inserting a padding item or using the previous sentence's last punctuation marks in the place of w_{t-2}.

There is another issue that arises from Korean's morphological richness; that is, if we just use the inflected words in the places of w_t, w_{t-1} and w_{t-2}, it shows poorer results than in English due to the sparsity of data originating from the rich inflection. As a solution, we decided to use only the first morphemes in words when building the n-gram model because it showed the best prediction performance on our research data.

3.4. Estimation of Prosodic Pressures

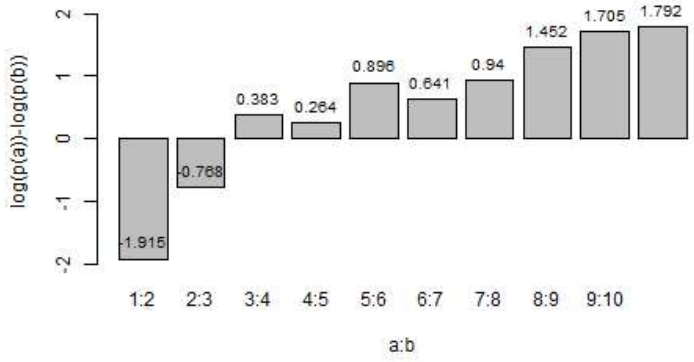


<Figure 2> Relative frequencies of Korean AP's syllable lengths (Shin, 2011)

While AP boundaries data themselves are not directly accessible in public so far, some important information from them such as relative frequencies of syllable lengths of AP is provided in Shin (2011). On the assumption that Kiaer & Shin (2012)'s hypothesis is correct, we can expect that Korean speakers' general preference on AP lengths is proportional to their probabilities, which are empirically measured by relative frequency. Thus, the theoretical pressure that works on a choice between two options is also expected to be proportional to the ratio of the two options'probability. In addition, we apply logarithm to it for computational convenience. This notion of prosodic pressure can be expressed as <Formula 3>. And exemplary calculations of prosodic pressure between two adjacent classes are suggested in <Figure 3>.

$$\mathrm{Prsd}(\mathrm{a},\mathrm{b}) = \log\left(\frac{\mathrm{p}(\mathrm{a})}{\mathrm{p}(\mathrm{b})}\right) = \log(\mathrm{p}(\mathrm{a})) - \log(\mathrm{p}(\mathrm{b}))$$

<Formula 3> Prosodic pressure between two options' AP length, *a* and *b*



<Figure 3> Prosodic pressure between two options' AP lengths

To compute the prosodic pressure of a given case, we need to know the length of the AP. However, corpus data in which the AP boundaries are labelled are currently neither available nor accessible. Despite this obstruction, indirect methods of estimating prosodic pressures could be developed on certain assumptions, whose results are never expected to be exact, but still meaningfully interpretable.

$$\begin{aligned}\mathrm{Prsd}_{\mathrm{AP}} &= \max_{A_r \in \Omega_r} \log\left(\prod_{a_r \in A_r} \mathrm{p}(\mathrm{a_r})\right) - \max_{A_o \in \Omega_o}\left(\prod_{a_o \in A_o} \mathrm{p}(\mathrm{a_o})\right) \\ &= \max_{A_r \in \Omega_r} \log \sum_{a_r \in A_r} \log(\mathrm{p}(\mathrm{a_r})) - \max_{A_o \in \Omega_o} \sum_{a_o \in A_o} \log(\mathrm{p}(\mathrm{a_o}))\end{aligned}$$

<Formula 4> Estimation of prosodic pressure based on the most probable AP assumption

On the assumption that Kiaer & Shin (2012)'s hypothesis is true once again, the most likely AP boundaries should be the ones that construct the optimal allocation of APs, in which the probabilities of each AP length are maximized and the overall prosodic cumbersomeness is minimized. In this way, we roughly estimate prosodic pressure between two options as <Formula 4>, where Ω_r is the

set of all possible allocations of APs in the scenario that the particle is realized, Ω_o, the set of all possible allocations of APs in the scenario that the particle is dropped and a_x represents each AP's number of syllables in those allocations. However, this is based on so many assumptions that it might be not feasible to be used in this research.

On the other hand, the length of a phonological word is fairly straightforward; in most cases, it is equivalent to the number of syllables in a orthographic word. Therefore, to see if the prosodic effect occurs in units of phonological words, we computed $\mathrm{Prsd}_{\mathrm{PW}}$ using only the base forms of the words and built a model based on it instead of $\mathrm{Prsd}_{\mathrm{AP}}$.

4. Results

	Estimate	Std. Error	z-value	Pr(>|z|)	
$\beta_{\mathrm{intercept}}$	−0.706791	0.009654	−73.21	<2e−16	***
β_{surp}	3.081300	0.031953	96.43	<2e−16	***
β_{Prsd}	0.163690	0.006542	25.02	<2e−16	***

<Table 2> Regression table

The regression coefficients of both surprisal and prosodic pressure are highly significant. This suggests that there is a clear positive correlation between Korean particle realization, surprisal and prosodic pressure, in the sense that the larger surprisal and the stronger prosodic pressure of a given speech condition, the more likely the case particle is realized.

5. Discussion

From the observation above, particle omission in the Korean language seems to be one of the intersecting points where various levels of linguistic motivations cross and meet (semantic, pragmatic motivations, prosodic motivations, processing motivations etc.). This suggests a new possibility of investigating these types of interaction more by further experimentation and mathematical modelling.

References

Asr, F. T. & Demberg, V. (2015), "Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission," Poceedings of the 11th International Conference on Computational Semantics, 118-128, London, UK, April 15-17 2015.

Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L. & Levy, R. (2019). "How Efficiency Shapes Human Language,"Trends in Cognitive Sciences 23(5), 389-407.

Hale, J. (2001), "A probabilistic Earley parser as a psycholinguistic model," Proceedings of NAACL, 2, 159-166.

Hahn, M., Jurafsky, D. & Futrell, R. (2020), "Universals of word order reflect optimization of grammars for efficient communication,"Proceedings of the National Academy of Sciences, 117(5), 2347-2353.

Hong, J. (2010), "무조사구의 주어-목적어 비대칭 분포와 의미: 형태적 실현성" [Subject-Object Asymmetries of NPs without Markers and their Meaning: Morphological Realization], 언어정보, 0(11), 119-138.

Horch, E. & Reich, I. (2016). "On "Article Omission" in German and the "Uniform Information Density Hypothesis,"" Proceedings of the 13th Conference on Natural Language Processing.

Kiaer, J. & Shin, J. (2012), "목적격조사 생략 현상에 대한 운율적 해석" [Prosodic Interpretation of Object Particle Omission in Korean], 한국어학, 57, 331-355.

Ko, E. (2000), "A discourse analysis of the realization of object NP forms in Korean," 어학 연구, 36(1), 43-62.

Ko, E. (2001), "A discourse analysis of the realization of object NP forms in Korean," In M. Nakayama and C. Quinn (eds.) 2001, Japanese Korean Linguistics IX, CSLI.

Maurits, L., Perfors. A. & Navarro, D., (2010),"Why are some word orders more common than others? A uniform information density account," Advances in neural information processing systems 23, 1585-1593.

Lee, Hanjung (2006), "Effects of focus and markedness hierarchies on object case ellipsis in Korean," 담화와 인지, 13(2), 205-231.

Lee, Hanjung (2010), "Explaining variation in Korean case ellipsis: Economy versus iconicity," Journal of East Asian Linguistics, 9, 291-318.

Lee, Hanjung (2011), "Gradients in Korean case ellipsis: An experimental investigation," Lingua, 121(1), 20-34.

Lee, Sungbom (2006), "대격 조사 생략의 화용적 분석" [A Pragmatic Analysis of Accusative Case-Marker Deletion], 담화와 인지, 13(3), 69-89.

Lee, Sun-Hee & Song, J. (2012), "Annotating Particle Realization and Ellipsis in Korean," Proceedings of the Sixth Linguistic Annotation Workshop.

Levy, R. & Jaeger, T. F. (2007), "Speakers Optimize Information Density Through Syntactic Reduction," Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS).

Levy, R. (2008),"Expectation-based syntactic comprehension,"Cognition, 106(3), 1126-1177.

Ranjan, S., Agarwal, S. & Rajkumar, R. (2019),"Surprisal and Interference Effects of Case Markers in Hindi Word Order,"Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics", 30-42.

Ross, I., Dickinson, M. & Lee, Sun-Hee. (2013), "Detecting and Correcting Learner Korean Particle Omission Errors," International Joint Conference on Natural Language Processing, 1419- 1427.

Shin, J. (2011),「한국어의 말소리」[The Speech Sound of Korean], 지식과 교양.