# *Argumentness* and Probabilistic Case Structures

**Dan-Hee Yang**
Department of Computer Science, Pyongtaek Univ.
111 Yongyi-Dong, Pyongtaek,
Kyungki-Do, 450-701, KOREA
dhyang@ptuniv.ac.kr

**Ik-Hwan Lee**
Department of English, Yonsei University
134 Shinchon-Dong, Seodaemun-Gu,
Seoul 120-749, KOREA
ihlee@yonsei.ac.kr

## Abstract

This paper proposes that the argument structures be stated in a way that uses probabilities derived from a corpus to replace a Boolean-value system of subcategorization. To do this, we make a cognitive model from a situation to an utterance to explain the phenomena of arguments' ellipsis, though the traditional term *ellipsis* is not suitable under our new concepts. We claim that the binary distinction is neither rational nor suitable for a real syntactic analysis. To solve this problem, we propose two new concepts *argumentness* and *probabilistic Case structures* by adapting the *prototype theory*. We believe that these concepts are effective in the syntactic analysis of NLP.

## 1 Introduction

Researches on building Korean Case frames have been done by several organizations such as the National Academy of the Korean Language, Language Research Institute of Seoul National University, and the Center for Linguistic Informatics Research of Yonsei University, including the SEJONG 21 National Project. For English, there are several comprehensive Case frames such as FrameNet, COMLEX corpus, and LDOCE (Longman Dictionary of Contemporary English), etc. All these researches distinguish optional arguments from obligatory ones in the logic of black or white, assuming that arguments are the participants *minimally* involved in the activity or state expressed by the predicate, and they discuss only obligatory arguments (Dusan 1998, Hong 1997, Lee 1997, Song 1999). We note, however, that the term *minimally* is very vague. Furthermore, it is very dubious that such a simple binary distinction of obligatory vs. optional arguments is objectively well-grounded (Nam 1993).

Concerning this issue, the present study recognizes that there is a significant difference in native speakers' intuition on whether or not an argument can be omitted. By establishing a cognitive model from a situation to an utterance, we explain why arguments' ellipsis occurs, though the traditional term ellipsis is not suitable under our new concepts. Here we devise two filters: an individual cognitive filter and an individual linguistic filter. Then, we claim that the binary distinction is not appropriate. Instead, we propose two types of new concepts, namely argumentness and probabilistic Case structures by adapting the prototype theory. Finally, we show that these concepts have several merits for NLP.

## 2 Cognitive Process of Arguments' Ellipsis

### 2.1 Ellipsis of Arguments

It has been noted that there are two rather different kinds of ellipses of arguments: namely, syntactic ellipsis and pragmatic ellipsis. A syntactic ellipsis can occur when we consider the semantic property of the predicate. A pragmatic ellipsis can occur when the omitted constituent can be naturally recovered from the context. Therefore, in a given sentence, an obligatory argument has been understood to undergo the pragmatic rather than the syntactic ellipsis. On the other hand, an optional

argument has been understood to undergo the syntactic/pragmatic ellipsis. However, we observe some delicate differences in intuition on whether or not the underlined constituents in (1-6) can be omitted:

(1) 이 도회지에 굉음이 새벽을 뒤흔들었다.
*i dohoeji-e goengeum-i saebyeg-eul dwiheundeul-eossda.*
In this urban area, a roaring sound shook dawn violently

(2) 그는 사춘기에 집을 나와 버렸다.
*geu-neun sachungi-e jib-eul nawa beolyessda.*
He ran away from home at adolescence.

(3) 그는 그 도시에 머물렀다.
*geu-neun geu dosi-e meomulleossda.*
He stayed in the city.

(4) (a) 그분이 저 사람을 의사로 만들었다.
*geubun-i jeo salam-eul uisa-lo mandeul-eossda.*
The gentleman made a doctor of that man.

(b) 그분이 저 사람을 만들었어.
*geubun-i jeo salam-eul mandeul-eoss-seo.*
*The gentleman made of that man.

(5) (a) 기쁨은 두려움으로 변했다.
*gippeum-eun dulyeoum-eulo byenhaessda.*
Pleasure turned into fear.

(b) 철수 목소리가 두려움으로 변했다.
*Cheolsu mogsoli-ga dulyeoum-eulo byenhaessda.*
The voice of Cheolsu turned into fear.

(6) 저 방에 전구 좀 새것으로 갈아 주세요.
*jeo bang-e jeongu jom saegeos-eulo gala juseyo.*
Please change the light bulb of that room into a new one.

Try to understand the above sentences, omitting the underlined constituents. If someone omits 이 도회지에 *i dohoeji-e* 'in this urban area' in sentence (1), the resulting sentence probably makes us to be anxious to know where the event happened, though we do not think that the sentence is wrong (i.e. ungrammatical). Example (with the underlined part omitted) (2) might lead us to ask "When? or Why?", though it is grammatical. As for example (3), we are likely to use it pragmatically, that is, supposing that we already know where he stayed. Example (4a) is similar to (3). It is more natural if it is colloquially uttered like (4b). Example (5b) seems to be more natural than (5a), though both have the same verb 변했다 *byenhaessda* 'turned'. Readers are likely to understand (5b), guessing diverse situations such as Cheolsu got a cold, he was at the age of the voice change, or he got hoarse because he used vocal chords too much. In contrast, example (5a) is very unnatural without an appropriate context. Example (6) seems to be natural because we can obviously infer the omitted constituent 새것으로 *saegeos-eulo* 'into a new one'. One may not even recognize the ellipsis itself in (6).

In distinguishing optional arguments from obligatory ones, Ki-Shim Nam (1993) notes the following:

> "In many cases, it is not easy to decide whether NP-로 *lo* 'to' is obligatory or not in a given sentence. There are no formal criteria. There is only a heuristic method such as: The given NP-로 *lo* 'to' is obligatory in a sentence if and only if the elliptical structure becomes ungrammatical. To decide the possibility of its ellipsis in a sentence, we do nothing but consider the semantic property of the predicate, totally depending on our linguistic intuition."

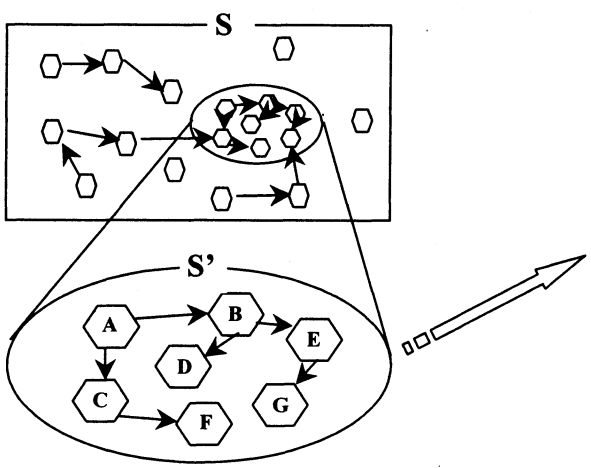
The problem is that native speakers are split into various groups on whether or not the examples in (1-6), without the underlined parts are right (i.e. natural or grammatical): total agreement, a little bit

agreement, total disagreement, and so on. This implies that native speakers are different from each other in linguistic intuition. It is not the case that this difference was observed only in several peculiar sentences. The difference is not due to our grammatical ignorance, either. We may sometimes observe it even in linguistics papers. If so, why does the syntactic/pragmatic ellipsis occur? Why does the linguistic intuition differ from each other?

## 2.2 Individual Cognitive/Linguistic Filter

There are diverse objects in the real world, where various events always occur among such objects. This is the way that this world is working. This implies that there are objects intrinsically related to the events.

Figure 1 below is a cognitive model, which delineates the process that a real situation S is uttered linguistically through its projected individual situation S'. In Figure 1, hexagons represent objects in the real world. The arrows among hexagons indicate some kinds of interaction among the related objects. For instance, the arrow from object A to object B indicates that object A has some sort of influence on object B. In this study, we define the term *situation* as referring to both an event and its related objects.




(a) Real situation S and its individual situation S' recognized by a speaker

A B C D E F G V.
A B C D E F V.
...
A C D E F V.
...
V.

(b) Illustrative sentences that can be generated to depict S'

Figure 1. Cognitive model from a situation to an utterance

Everyone has a different set of social, scientific knowledge and belief, and he faces a different situation. Hence, his point of view is necessarily biased. After all, when people observe an event, they cannot truly recognize the real situation. Suppose that a real situation S in Figure 1 (a) represents a full view of a certain event T. Each observer can recognize his individual situation S', which is a part of S. Here each observer may establish S' differently (i.e. a different part of S). We define this projector as *an individual cognitive filter.* To express S' linguistically, the observer can generate various sentences such as given in (b), applying his *individual linguistic filter*, which reflects the cultural, syntactic, and pragmatic factors. If he recognizes S' as large as possible and lexically realizes most objects in S', the resulting sentence has few ellipses. On the contrary, the more he depends on non-linguistic means such as gestures and contexts, the more ellipses may occur in the resulting sentence.

People can naturally communicate with each other, though everyone has a different cognitive/linguistic filter. It is because these filters are standardized to some degree thanks to the universal human cognitive faculty, linguistic education, common sense, and so on. Of course, the individual cognitive filter is most likely to interact with the individual linguistic one. This implies that

if the former changes, the latter will be possibly affected by the change. Since the individual cognitive/linguistic filter may be different according to region, culture, age, and so on, the intuition may also be different on whether an argument is obligatory or not. Therefore, talking about whether a sentence is right or not on the basis of the binary distinction may be meaningful only for establishing a prescriptive grammar, which is inclined to assume a universal linguistic filter. However, the binary distinction is not adequate for the syntactic analysis of sentences (i.e. a corpus) that are really uttered by various people, who have different individual cognitive/linguistic filters.

As mentioned earlier, Chomsky defines arguments as the participants minimally involved in the activity or state expressed by the predicate (Haegeman 1994). In fact, since his definition is to claim that everyone picks out the same participants in a certain event, he seems to bear a universal cognitive filter in mind, assuming that everyone has the same universal cognitive faculty. However, we claim that the linguistic/non-linguistic distinction itself is not useful for practical natural language processing and understanding utterances. Grammaticality is mostly useful for teaching standard languages and smoothly communicating with each other in the same linguistic environment.

Yonsei Korean Dictionary (Dusan 1998) also defines arguments similarly, probably because it should present a prescriptive grammar. However, notice that there may be diverse opinions on examples (1-6) in Section 2.1. This fact demonstrates that an individual cognitive/linguistic filter does not coincide with the corresponding universal one. So, if we attempt to develop a syntactic analyzer based on such a universal filter, we may not get a satisfactory result. The binary concept cannot analyze real sentences in a corpus. The important factor is the matter of acceptability rather than grammaticality.

### 2.3 Reason for Ellipsis

Noun phrases manifesting time or place are most likely to be optional ones. If such constituents are important to a listener, the listener will require the speaker to fill the omitted constituents. Receiving the constituents, he will construct his own mental sentence that he expected. Therefore, if most listeners require that a constituent should be filled, the sentence may be ungrammatical. In this case, the constituent is possibly obligatory. For example, see the sentences in (7-10) below:

(7) 그는 집을 나와 버렸다. *gu-neun jib-eul nawa beolyeossda.* 'He ran away from home.'
(8) 언제? *eonje*? 'When?'
(9) 사춘기에 *sachungi-e* 'at adolescence'
(10) 그는 사춘기에 집을 나와 버렸다.
*geu-neun sachungi-e jib-eul nawa beolyeossda.*
He ran away from home at adolescence.

Uttering sentence (7) may bring about question (8). Hence, the answer (9). In this case, the speaker is regarded to deliver the information represented by sentence (10). Then, why do people utter sentence (7) instead of (10)? We can find the answer in the *economy of speech* principle; *economy* reducing redundancy and *definitude of discrimination* for understanding are the most important principles that constrain the speech (Lee 1990). Based on this *economy of speech*, we think that the syntactic ellipsis of an argument is never possible unless the pragmatic ellipsis frequently occurs in a normal situation. In other words, a syntactic ellipsis may be interpreted as a consequence of a pragmatic ellipsis.

## 3 Induction of Probabilistic Case Structures

### 3.1 Ellipsis of Arguments and Prototype Theory

The more frequently an argument is realized by most native speakers, the more the argument is considered as necessary to the predicate. If it becomes all the more common, we can consider the

argument as syntactically required. Hence, if few native speakers omit the argument, though the argument can be pragmatically inferred from the context, they will think that the elliptical sentence is unnatural. In the extreme case, they will think that it is syntactically wrong. On the contrary, an argument may be frequently omitted for some other reasons. Here, we introduce the *prototype theory* (Taylor 1995) to devise new concepts that can accommodate this phenomenon. For example, in the prototype theory, a bird may be defined in a structural representation as follows:

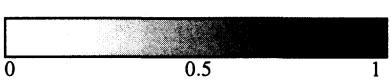
bird = { two wings, two legs, small head, sharp bill, ... }

Here, a bird is compared to a sentence. The attributes of the bird correspond to the arguments in the sentence. If "two wings" are removed from a bird or if a bird was born without "two wings", can we call it a bird any longer? What if without "two legs"? If we proceed to remove one by one in this way, we will reach the state that we cannot call it a bird any longer. Someone may claim that if an object does not have "two wings", it can never be a bird. Eventually, every people may have a different definition on whether an object is a bird or not. One may claim that since the evidence was an experimental finding that there is no significant difference between individuals in what are typical for a category, when the prototype concept is introduced, it will be natural that everyone shares a typical definition of birds. This position does not conflict with our claim. Notice, however, that we do not mention a *typical* definition on birds, but an *exact* and *unique* one.

### 3.2 *Argumentness* and Probabilistic Case Structures

Considering the prototype theory, we propose a new concept *argumentness,* which can be depicted on a spectrum such as Figure 2 (b), instead of the binary classification such as Figure 2 (a). In Figure 2 (b), the value 0 of *argumentness* means that an argument need never occur. The value 1 means that it is obligatory. The values between 0 and 1 mean optional. The higher in number this *argumentness* is, the stronger is the requirement of the argument of the verb.


(a) Binary classification




(b) Argumentness

Figure 2. Alternative concept of obligatory/optional arguments

$$S_{k1} = (V_{k1}, p_{k1}),\ S_{k2} = (V_{k2}, p_{k2}),\ \ldots,\ S_{kj} = (V_{kj}, p_{kj})$$

$$V_{k1} = C_{11}\ C_{12}\ C_{13} \ldots C_{1i},\quad V_{k2} = C_{21}\ C_{22}\ C_{23} \ldots C_{2i},\quad \ldots,\quad V_{kj} = C_{j1}\ C_{j2}\ C_{j3}\ \ldots C_{ji}$$

Where, $S_{kj}$ is the $j^{th}$ probabilistic Case structure of the verb k, $V_{kj}$ is the $j^{th}$ Case structure of the verb k, $p_{kj}$ is a relative frequency, $f(V_{kj}) / \sum_{t=1} f(V_{kt})$, $f(V)$ is the frequency of $V$ in a corpus, and $C_{ji}$ is a Case particle.

Figure 3. The definition of probabilistic Case structures

If we define a Case structure as the syntactic realization of arguments, it should be represented probabilistically as in Figure 3, reflecting the concept of *argumentness*. Hence, we have *probabilistic case structures*. Based on the characteristics of Korean, this study represents a Case structure by the sequential enumeration of Case particles, which manifest the syntactic Case of arguments. According to the definition of Figure 3, for example, the case structures of verb 걸다 *geolda* 'link' are *probabilistically* represented as in Figure 4.

걸다$_1$ *geolda*$_1$ = ($V_{11}$, 0.8), 걸다$_2$ *geolda*$_2$= ($V_{12}$, 0.1), 걸다$_3$ *geolda*$_3$= ($V_{13}$, 0.07), 걸다$_4$ *geolda*$_4$= ($V_{14}$, 0.03)

$V_{11}$ = -가 -와 - 을 *-ga – wa – eul* ' NOM COM ACC' ;
ex) 갑돌이가 갑순이와 손가락을 걸었다.
*Gabdoli-ga Gabsuni-wa songalag-eul geol-eossda.* ' Gabdoli linked his finger in gabsuni' s.'

$V_{12}$ = -가 – 을 *-ga – eul* ' NOM COM ACC' ;
ex) 갑돌이가 손가락을 걸었다.
*Gabdoli-ga songalag-eul geol-eossda.* ' Gabdoli linked his finger.'

$V_{13}$ = -가 – 와 *-ga – wa* ' NOM COM' ;
ex) 갑돌이가 갑순이와 걸었다.
*Gabdoli-neun Gabsuni-wa geol-eossda.* ' Gabdoli linked in gabsuni' s.'

$V_{14}$ = -가 *-ga* ' NOM' ;
ex) 갑돌이가 걸었다.
*Gabdoli-ga geol-eossda.* 'Gabdoli linked.'

Figure 4. Examples of the probabilistic Case structures of the verb 걸다 *geolda* 'link'

In Figure 4, verb 걸다 *geolda* 'link' has four possible probabilistic Case structures: 걸다$_1$ *geolda*$_1$, 걸다$_2$ *geolda*$_2$, 걸다$_3$ *geolda*$_3$, 걸다$_4$ *geolda*$_4$. We can see that the probabilistic Case structure of 걸다$_1$ *geolda*$_1$ is used most frequently. Its Case structure is $V_{11}$ = -가($C_{11}$) -와($C_{12}$) -을($C_{13}$) *-ga –wa –eul* 'NOM(nominative Case) COM(comitative Case) ACC(accusative Case)' and has the probability of 0.8. Here the probability 0.8 is obtained by (1) below, where f($V_{1t}$) ($1 \le t \le 4$) is a function that calculates the frequency of each $V_{1t}$ like f(걸다$_1$) when t = 1 from a corpus.

$$f(V_{11}) / \sum_{t=1}^{4} f(V_{1t}) = f(\text{걸다}_1) / (f(\text{걸다}_1)+f(\text{걸다}_2)+f(\text{걸다}_3)+f(\text{걸다}_4)) \quad (1)$$

### 3.3 Use of Probabilistic Case Structures

Probabilistic Case structures have several merits for NLP. First of all, the existing Case structures stick only to one structure for one meaning of a verb. In contrast, probabilistic Case structures allow different structures $V_{kj}$, which is the j$^{th}$ Case structure of the verb k, as in Figure 3. Hence, it is much easier to build them automatically from a corpus merely because we need not contemplate whether one argument is obligatory or not. Second, we can build them statistically from a corpus rather than from linguists' subjective intuition. Hence, the resulting case structures are statistically objective. Finally, in Korean, all arguments outside of the normal governing domain of the verb. In other words, they can be placed before the subject. In a normal case, the configuration is: S = SUBJ VP, VP = NPs V. However, in the present model the free structure is possible: S = NPs SUBJ NPs VP, where VP consists of V and NPs. Hence, this raises syntactic ambiguity. To *probabilistically* solve such

ambiguity, probabilistic Case structures are very effective. See the illustrative sentences (11-14) below:

(11) 나는 갑순이와 손가락을 갑돌이가 거는 것을 보았다.
*na-neun Gabsuni-wa songalag-eul Gabdoli-ga geoneun geos-eul boassda.*
I-NOM gabsuni-COM finger-ACC gabdoli-NOM linking that-ACC saw.

(12) (a) 나는 보았다.
*na-neun boassda.* 'I saw.'
(b) 갑돌이가 갑순이와 손가락을 걸었다.
*Gabdoli-ga Gabsuni-wa songalag-eul geoleossda.* 'Gabdoli linked his finger in gabsuni's.'

(13) (a) 나는 갑순이와 보았다.
*na-neun Gabsuni-wa boassda.* 'I saw with gabsuni.'
(b) 갑돌이가 손가락을 걸었다.
*Gabdoli-ga songalag-eul geoleossda.* 'Gabdoli linked his finger.'

(14) (a) 나는 갑순이와 손가락을 보았다.
*na-neun Gabsuni-wa songalag-eul boassda.* 'I saw a finger with gabsuni.'
(b) 갑돌이가 걸었다.
*Gabdoli-ga geoleossda.* 'Gabdoli linked.'

In (11), the arguments 갑순이와 *Gasuni-wa* 'with Gasuni' and 손가락을 *songalag-eul* 'finger' can be linked to either verb 걸다 *geolda* 'link' or 보다 *boda* 'saw', hence there are three possible analyses as in (12-14). Generally, this kind of syntactic ambiguity cannot be solved without contexts or common sense. With the probabilistic case structures, however, we can select the biggest one by adding the relative frequency $p_{ki}$ of (a) and the one $p_{mj}$ of (b) for (12-14) each.

## 4 Conclusion

This study observed that there is a difference in native speakers' intuition on whether an argument can be omitted. To explain this difference, we devised two filters, namely an *individual cognitive filter* and an *individual linguistic filter*. In addition, we explained the phenomena of argument ellipsis by the *economy of speech*. We introduced two concepts *argumentness* and *probabilistic Case structures* to solve the intrinsic defects of the binary classification (obligatory and optional arguments). Then, we showed that these concepts are very useful for the statistical processing of the Korean language. Accordingly, we claim that verbs have alternative Case frames that vary in strength or probability.

There are several issues that we need to pursue in the future. The present study needs to devise the testing method of the probability assignment. We also need to establish the theoretical relationship between the cognitive model and the computational one. Furthermore, we need to do some real empirical experiments in order to show the utility of the new concepts introduced in this study. Nonetheless, we are confident that our new concepts work more effectively in the syntactic analysis of NLP than the existing ones.

## Acknowledgments

This research was funded by grant No. 2001-2-52200-001-2 from the Basic Research Program of the Korea Science & Engineering Foundation.

## References

Brent, Michael R.. 1991. "Automatic Acquisition of Subcategorization Frames from Untagged Text", In the *Proceedings of the 29th ACL.*

Dusan Press. 1998. *Yonsei Korean Dictionary.*

Haegeman, Liliane. 1994. *Introduction to Government & Binding Theory*, 2nd Edition, p. 7, Blackwell Publishers, Oxford.

Hong, Jae-Sung, et al.. 1997. *Syntactic Dictionary of Modern Korean Verbs*. Dusan Donga.

Lee, Chungmin, Bae Young-Nam. 1990. *Linguistic Dictionary*. Bak Young Sa.

Lee, Chungmin, Kang Beom-Mo, Nam Seung-Ho. 1997. "A Study on Semantic Structures of Korean Predicates," In the *2th Workshop of Soft Science.*

Li, Hang, Naoki Abe. 1995. "Generalizing Case Frames Using Thesaurus and the MDL Principle", In the *Proceedings of the 33th ACL.*

Manning, Christopher D.. 1992. "Automatic Acquisition of a Large Subcategorization Dictionary from Corpora", In the *Proceedings of the 30th ACL.*

Nam, Ki-Shim. 1993. *The Usage of Korean Particle*, pp. 9-117, 219-368,. Seoul: Seogwang Academic Press.

Oflazer, Kemal, Okan Yilmaz. 1996. "A Constraint-based Case Frame Lexicon Architecture", In the *Proceedings of COLING.*

Song, Mansuk, Ki-Shim Nam, Dan-Hee Yang, et al.. 1999. *A Research on Automatic Construction of Case Frames for Korean Language Processing – Centered on the Case Particle 'ulo'* -, Research Report of the Ministry of Information and Communication.

Taylor, John R. 1995. *Linguistic Categorization: Prototypes in Linguistic Theory*, pp. 60-94. Oxford U. Press.

Yang, Dan-Hee. 1999. *Representation of Word Meanings for Understanding Thematic Roles and Their Acquisition by Machine Learning*, pp. 36-42, Ph. D. Dissertation, Department of Computer Science, Yonsei University.