

# Annotations for Zero Pronoun Resolution in Korean Using the Penn Korean Treebank

Sun-Hee Lee, Donna K. Byron, & Whitney Gegg-Harrison

Computer Science and Engineering at the Ohio State University

## 1. The Problem of Null Elements

In topic prominent languages like Korean and Japanese, a repeated element has no surface realization, called a *zero pronoun*, in contexts where an explicit pronoun would be used in English. Compare example (1) and (2):

- (1) a. John watched the children on the street.  
b. Mary watched *them*, too.
- (2) a. John-i kil-eyse aitul-ul po-ass-ta.  
John-NOM street-on children-ACC see-PAST-END  
'John watched the children on the street.'  
b. Mary-to Ø po-ass-ta.  
Mary-also OBJ see-PAST-END  
'Mary also watched (zero=them).'

This property of Korean creates an issue for developers of Treebanks and other annotated language resources: when and how should these unrealized elements be explicitly introduced into the linguistic material being developed? Although this question could be asked when working with any language, in so-called pro-drop languages such as Korean, Japanese, Spanish and Portuguese, the problem is more immediate, because basic units of sentence structure, such as subjects of matrix clauses, are frequently unrealized.

Although researchers, especially the anaphora resolution community, have been eager to study the distribution of zero pronouns and their link to other discourse properties, there has historically been a lack of annotated material available to the wider research community that would allow us to investigate these questions. Researchers in the past worked mainly with small amounts of hand-constructed data rather than being able to do large-scale corpus analysis. This lack has been recently addressed by the release of the Penn Korean Treebank (Han et. al. [5]; henceforth PKT), which includes annotations indicating the position of zero pronouns. Our initial attempts to utilize the PKT as test data for a zero pronoun resolution algorithm have unfortunately revealed several problems in the construction of the trees and the positions of zero pronouns in the PKT. These problems need to be resolved before the PKT can fulfill its potential as a source of linguistic knowledge about zero pronouns, and they should be taken into consideration by other teams developing similar resources in other languages.

The work described here is primarily motivated by our efforts to develop an annotation scheme for zero pronouns in Korean that can be used to develop a gold standard for testing an anaphora resolution algorithm. Previous authors have pointed out that the antecedents of zero anaphors can often be determined by using various grammatical properties such as topicality, agreement, tense, and aspect as well as subcategorization information (Walker et al. [12]; Iida [8]; Hong [7], etc.). However, in order for these factors to be useful in developing anaphora resolution algorithms, they must be reliably and consistently annotated into the source data. In this paper, we will examine the necessity for a new annotation scheme that provides relevant semantic and discourse factors for reference resolution in the Korean Treebank. Our analysis can be extended to naturally occurring discourse that shows more complicated discourse properties and structures. The revised annotation scheme will be adopted for our new software handling zero anaphors in languages like Korean.

## 2. Identifying Positions of Zero Pronouns in the Penn Korean Treebank

Why mark zero pronouns in the PKT? Unrealized arguments are important for tracking the attentional state of a discourse in topic-oriented languages like Korean and Japanese. Within the framework of centering theory, e.g. Walker et al. [12], Iida [8], Hong [7], etc. it has been shown that a salient entity recoverable by inference from the context is frequently omitted, and therefore interpreting these zero anaphors allows one to follow the center of the attentional state. Walker et al. [12] applied the centering model, developed for pronoun resolution in English, to zero anaphor resolution in Japanese. They argue that interpretation of a zero anaphor is determined by discourse factors. This suggests that identifying occurrences of zero anaphors and retrieving their antecedents are important in developing a computational model of discourse interpretations as well as syntactic and semantic analyses.

### 2.1 Obligatory vs. Optional Arguments

The first crucial step for zero anaphor resolution is to identify the exact positions zero pronouns. According to the guidelines of the PKT, only a missing obligatory argument should be annotated as an empty element. Missing optional arguments are not. Thus, missing subject or object elements were marked as zero pronouns while missing locative arguments, such as *hakkyo-ey* in (2), were not marked when they were omitted.

- (2) wuli-nun eycey hakkyo-ey ka-ss-ta.  
 we-Top yesterday school-to go-Past-Decl  
 ‘We went to school yesterday.’

However, the annotation method based on an obligatory vs. optional argument may result in the loss of crucial information needed at later stage of retrieving an antecedent of a zero element. For example, the locative argument, 제45 사단에 ‘the 45<sup>th</sup> division-in’ has not been marked up as a zero anaphor in a tagged sentence of (3) B.

(3) A: 제 45 사단은 또 무엇으로 구성되어 있는가 ?  
 the 45 division again with what composed be  
 ‘What is the 45<sup>th</sup> Division composed of?’  
 (S (NP-SBJ 제/XPF+45/NUU  
 사단/NNC+은/PAU)  
 (VP (VP (ADVP 또/ADV)  
 (VP (NP-COMP 무엇/NPN+으로/PAD)  
 (VV 구성/NNC+되/XSV+어/EAU)))  
 있/VX+는가/EFN)  
 ?/SFN)

B: 사단 지휘부가 있습니다 .  
 division head-Nom exist  
 ‘The head division is (there).’  
 (S (NP-SBJ 사단/NNC  
 지휘부/NNC+가/PCA)  
 (ADJP 있/VJ+습니다/EFN)  
 ./SFN)

In the given discourse segments, the adjective *있다* *issta* requires a locative argument which has been treated as an optional argument in the PKT. Thus, there is no information on the empty element in (3B). However, this sentence is uninterpretable unless the missing optional element is recovered.

Due to some unique properties of the PKT, we believe that the importance of missing optional arguments has been underestimated. The annotated Treebank consists of texts from military language training manuals with 54,000 words (Han et al. [5]), where most sentences are composed of a question and an answer between a military personnel and a captive. Thus, the discourse structure and conversational flow of the corpus are rather artificial and simpler than natural dialogs would be. In the Korean Treebank, occurrences of zero anaphors generally depend on the local contexts of the preceding utterance. However, in naturally occurring discourse, zero pronouns are licensed not only in local contexts but also in global contexts. In example 4, the last zero anaphor corresponding to a missing optional argument refers to ‘Jeju island’ that appears in the first utterance.

(4) A: kot Jejudo-ey-nun yuchaykkoch-i phil kesita.  
 soon Jeju island-in-Top yellow rape flowers-Nom bloom will  
 ‘Soon, yellow rape flowers will bloom in Jeju island’  
 B: manhun kwankwangkayk-i Ø molyetul-kess-ci.  
 many tourists gether-Fut-Decl  
 ‘Many tourists will flock (into there).’

- A: Na-to kulehkey sayngkakhay.  
 I-also so think  
 ‘I also think so.’
- B: owel-ey-nun yehayngkayktul-i tewuk Ø pwumpinta.  
 May-in-Top travelers-Nom more crowd  
 ‘In May, more travelers crowd (there).’

In order to maintain consistent annotations of zero anaphors and develop a reference resolution system, we argue that missing optional arguments need to be marked as zero pronouns. For this process, it is essential to check interannotator agreement and to use constant subcategorization frames of predicates. Dictionaries with specific subcategorization information can be used here, such as the Yonsei Korean dictionary, where different subcategorization frames are listed according to semantically disambiguated senses for each predicate.

## 2.2 Errors in the Penn Korean Treebank

We examined the current annotations of PKT and found some significant problems with respect to zero elements. For this study, we examined only 100 sentences from the Treebank. Those 100 sentences are composed of dialogue sequences and contain 133 occurrences of zeros. Among 133 zeros, 29 tokens should not have been marked as zero pronouns. In 15 cases, the sentence was marked as having a zero subject although subject NPs overtly appear in the same sentence. The problem arises due to an ambiguous case marking on nouns referring to collectives or groups. While the nominative case markers *ka* and *i* normally indicates subject nominals, groups or organizations appearing in subject position use the case marker *eyse* as in (5)<sup>1</sup>, thus the case marker *eyse* is treated as a nominative marker in Korean grammar. In the 15 cases, subject nominals referring to a group or organization were mistakenly parsed as NP adverbials rather than as subjects.

- (5) hakyō-eyse John-eykey sang-ul cwuessta.  
 school-NOM John-to prize-ACC awarded  
 ‘The school awarded a prize to John.’

This error significantly increases the number of zeroes in PKT, since it is composed of military training materials with frequent mention of group nominals such as *taytay* ‘squadron’, *sotay* ‘platoon’, *yentay* ‘regiment’, etc.

In addition, missing subjects have been assumed for some conventional fixed expressions that are similar to semi-idiomatic phrases. For example, the usages of *-ey tayhayse* ‘regarding on’, *-ey uyhayse* ‘by’ and *-ul wihayse* ‘for the sake of’ are conventional in that their meanings refer to something other than the compositional sum of their parts. In traditional Korean

---

<sup>1</sup> Refer to Yonsei Korean Dictionary, and Nam & Ko [10] etc.

grammar, the verbs like *tayhata*, *uyhata* and *wihata* are classified as incomplete verbs that do not require a subject and have restricted inflectional forms. We found 14 examples of missing subjects in the subject positions of these incomplete predicates. Similar to subjectless incomplete verbs, it is well known that there exists a class of subjectless verbs in Korean which include modal predicates such as *sayngkita*, *moluta*, *toyta*.<sup>2</sup> However, empty subject positions have been assumed for these subjectless predicates in the parsed outputs of the PKT. Therefore, unnecessary zero elements appear in the Treebank and may lead to inaccurate conclusions about the distribution of zero pronouns in Korean.

### 3. Classification of Zero Pronouns

Once the positions of zero pronouns are identified, further information can be specified to indicate the antecedent, when there is one, and other properties of the reference. With respect to anaphor resolution, Han [6] proposed a classification of zeros and an annotation scheme based on Korean text corpora including the PKT. In this section, we will discuss some problems that we found in the current scheme and argue for a rather revised classification for annotating zeros.

#### 3.1. Han [6]’s Classification

In order to build an annotated corpus that can be used for developing a pronoun resolution system, Han [6] proposes a classification of zero elements in Korean as in Table 1.<sup>3</sup>

text dependent use	discourse anaphoric	Propositional
		NP anaphoric
text independent use	zero deictic and zero indexical	
	indefinite null pronoun	Generic (nonspecific)
		Specific
	general situational null pronoun	

Table 1: Han’s Classification of Korean Zeros

Furthermore, she proposes the coding scheme for annotation as follows.

- (i) NP-anaphoric elements are linked to their antecedents by the addition of a numeric index indicating the coreference class.

<sup>2</sup> Kim [11] analyzes some modal verbs as subjectless verbs.

<sup>3</sup> Han’s classification is similar to Kameyama [9]’s classification of Japanese zero anaphors except that Han classifies discourse anaphoric zeros into propositional anaphoric zeros and NP anaphoric zeros while dividing indefinite personal zero anaphors into indefinite generic zeros and indefinite specific zeros.

(ii) Other categories are marked with an alphabetic index:

- |                                |                             |
|--------------------------------|-----------------------------|
| -deictic speaker: i            | -deictic hearer: y          |
| -deictic speaker and hearer: w |                             |
| -indefinite generic: g         | -indefinite specific: s     |
| -situational: x                | -anaphoric propositional: p |

One of advantages in Han's coding system is that it allows dual markings for zeros that can be interpreted as deictic and anaphoric at the same time. Thus, the following occurrence of a zero element can be marked as deictic and also anaphoric.

(6) *nai<sub>i5</sub> nun kakkum ku secem-ey kanta. Ø<sub>i5</sub> hoksi ku-lul mannako sipheseita.*  
*I<sub>i5</sub> occasionally stop by at the bookstore. It's because (I<sub>i5</sub>) want to see him.*

Han's system, however, introduces some unnecessary complexity in classifying zeros, which may complicate the annotation scheme and interannotator agreement. We will discuss relevant problems in the next section and will propose more simplified version of zero classification.

### 3.2. Our Proposed Zero Anaphor Classification

We classify Korean zeros into three different classes as shown in Table 2; discourse anaphoric zeros, deictic and indexical zeros, and indefinite zeros.

Discourse Anaphoric Zeros	Individual Entities
	Eventualities
	Propositions
Deictic and Indexical Zeros	
Indefinite Zeros	

Table 2. Categories of Korean Zeros

The discourse anaphoric zeros take their reference from antecedents in the previous utterances in the given discourse. This class is the main one that anaphor resolution systems aim to handle. As for discourse anaphoric zeros, there are three subclasses. The first refers to individual domain entities, the second, eventualities, and the third, propositions. The first and the third subclasses correspond to Han [6]'s NP anaphoric zeros and propositional zeros. The zeros of individual entities refer to entities that were introduced into the discourse via noun phrases. The zeros of propositions refer to propositions introduced in the previous utterance. Relevant examples are provided in (7) and (8).

(7) A: *i taytay-uy yepi hochwul penho-nun mwunka?*  
 2 squadron-GEN provisional watchword-TOP what  
 'What is the watchword of the 2<sup>nd</sup> squadron?'

- B: Ø “chwutong”-ipnimta  
 SUBJ chwutong-COP  
 ‘It is Chwutong.’  
 (Ø = ‘the watchword of the 2<sup>nd</sup> squadron’)
- (8) A: 108 yentay cihwipwu-nun hyencay eti-ey wichihako issnun-ka?  
 108 regiment headquarter-TOP now where-at locate being-Q  
 ‘Where is the headquarter of the 108<sup>th</sup> regiment located?’
- B: Ø<sub>1</sub> Ø<sub>2</sub> molukeyss-supnita.  
 SUBJ OBJ not know-END  
 ‘I don’t know.’  
 (Ø<sub>1</sub> = ‘B’, Ø<sub>2</sub> = ‘Where the headquarter of the 10<sup>th</sup> regiment is located’)

While Han distinguish only two kinds of discourse anaphoric zeros, there exist another kind that does not belong to either of her two categories. Let us consider the following examples:

- (9) A: aitul-i sihem-ul chi-ko sipheha-ci anha.  
 children-NOM exam-ACC take-END want-END don’t  
 ‘Children don’t want to take an exam.’
- B: na-to Ø silhe.  
 I-also hate  
 ‘I also hate it.’ (Ø = the action of taking the exam)
- (10) A: kyothong-i wenhwalha-myen wuli-nun nuc-ci ahulkeya.  
 traffic-NOM smooth-if we-TOP late-END not  
 ‘We won’t be late if the traffic is smooth.’
- B: kulssey, i sikan-ey-nun Ø himtul-kel.  
 well this hour-in-TOP tough-will  
 ‘Well, it will be tough at this time.’  
 (Ø = ‘the event that the traffic is smooth’)

In (9) and (10), zeros do not refer to an individual or a proposition. They refer to eventualities, i.e. action and event (Asher [2]). Thus, we categorize non-entity referring zeros as zeros of eventualities.

The second class of zero anaphors includes deictic and indexical zeros that directly refer to entities that can be determined in the given spatiotemporal context, which generally include a speaker and an addressee. The third class includes indefinite zeros referring to general people, which corresponds to *they*, *one*, and *you* in English.

Our classification of zeros, however, does not contain categories of situational zeros that have been assumed both in Kameyama [9] and Han [6]. Han provides (11) as an example of situational zeros.

- (11) Ø pelsse yel-si-ta  
 SUBJ already 10 o’clock-COP  
 ‘It is 10 o’clock already.’

As mentioned in the previous section, Korean allows subjectless constructions. While a dummy subject, *it* or *there* is required for sentences referring to time, weather, or the situation in English, the subject position does not need to be filled in Korean.<sup>4</sup> Therefore, there should not be a zero element in the subject position of (11).

Another classification that we do not adopt is Han's distinction between indefinite specific zeros vs. indefinite generic zeros.<sup>5</sup> We instead categorize indefinite zeros as one category. This is because it is not clear that indefinite zero elements themselves are ambiguous. The indefiniteness of zeros is closely tied to the semantic interpretation of a sentence as Han also admits. For bare plurals in English, Carlson [4] derives generic interpretations by using a generalization operator, *Gn*, without assuming ambiguity of bare plurals. A similar account can be applied to Korean zeros. In addition, Amaral [1] shows that generic vs. specific interpretations of zeros in European Portuguese, a *pro*-drop language, can be constrained by semantic-pragmatic factors. With respect to empirical issues, it is difficult to distinguish

---

<sup>4</sup> Han argues that the following example supports her category of situational zeros.

- (i) cikum sikan-i yel-si-ta.  
 now time-NOM 10 o'clock-COP  
 'Now, it is 10 o'clock.'

However, in Korean multiple subjects can be licensed based on various semantic relations between nominative NPs such as possessor-possessed, part-whole, class-member, etc. For example, the sentence (i) can be extended into (ii).

- (ii) New York-i cikum-i sikan-i yel-si-ta.  
 New York-NOM now-NOM time-NOM 10 o'clock-END  
 'Now it is 10 o'clock in New York.'

Thus, the existence of *sikan-i* 'time-NOM' does not support the existence of a zero subject for the predicate *yel-si-ta*, which does not require that the subject position be filled.

<sup>5</sup> Examples for indefinite specific zeros and indefinite generic zeros have been provided by Han [6].

- (i) Ø kyupu-nun kongsangwahakyenhwa-uy thul-ul pilie milo-uy  
 Cube- TOP SF movie-GEN frame-ACC adop labyrinth-GEN  
 talchwul-ilanun sinhwa-lul yenghwa-lo mantu-n cakpwum-i-ta.  
 escape-QUOTE mith-ACC movie-as make-REL work-COP-END  
 'Cube is a work that (someone) made the myth of an escape from a labyrinth  
 into a movie adopting a SF movie framework.'
- (ii) paopap-namwu-nun Ø cachic nuckey son-ul ssu-myen kuttayn  
 paopap-tree-TOP SUBJ possibly late hand-ACC use-if finally  
 cengmal Ø Ø chechihal swu epskey toynta.  
 really SUBJ OBJ manage way impossible become  
 'As for Baobab trees, they become impossible (for one) to manage (them) if  
 (one/he) treat them late.'

While the zero subject in (i) refers to someone specific, who is the movie maker, the zero subject in (ii) has an arbitrary reference. One notable thing is that in (i) the movie title, *Cube* and possibly the whole situation or script triggers a sort of bridging inference (Clark [3]) and restricts the reference of the zero subject. In contrast, there exist no clues for the referent of the zero subject in (ii).

interpretations of indefinite specific zeros vs. generic zeros, which will increase interannotator disagreement. Thus, we decide not to distinguish indefinite specific and generic zeros while leaving a more theoretical discussion as a separate study.

## 4. Some Useful Features for Anaphor Annotations

Based on our coding system for Korean zeros, we now argue that more sophisticated information relating to semantic and discourse properties need to be added to the annotated corpora like the PKT. This will increase the applicability of the annotated corpora to both theoretical research on anaphors and computational modeling of anaphor resolution.

### 4.1. Topic Information

One crucial discourse factor that has an effect on determining an antecedent of a zero anaphor is the existence of topics. In Korean, a topic appears in a sentence initial position with the topic marker  $\text{ㄴ} \text{ nun}$ . While the marker *nun* functions as a topic marker in a sentence initial position, it also works as an auxiliary postposition in a non-initial position of a sentence. The first is classified as a grammatical topic marker while the latter is a contrastive topic marker. According to the current annotation scheme of the PKT, the two kinds of topic marker *nun* are treated as the same auxiliary postposition, which is similar to other postpositions *man* ‘only’, *to* ‘also’, and *mace* ‘even’. Structurally, a subject NP with a topic marker has been analyzed as the subject, while a topic marked object is treated as a scrambled argument out of its canonical position. Although grammatical topics are not independently annotated in the PKT, they are closely associated with interpretations of zero anaphors as in (12).

- (12) 1. A: mwucenkiyoung chwukencci-nun pothong elmana olay kanun-ka?  
 radio batteries-TOP usually how long last-Q  
 ‘As for batteries of radios, how long do they last usually?’  
 2. B: kuken Ø elmana manhi ssununka-ey tallye iss-ci-yo.  
 that how much use-On depend exit-END-HON  
 ‘It depends on how much (we) use.’  
 3. B: pothong Ø han ilcwuil cengto kap-nita.  
 usually approximately a week about last-HON  
 ‘Usually, **(they)** last for about one week.’

The zero subject in (12.3) refers to the topicalized subject of (12.1). Walker et al. (1994) provides evidence that topic marked elements function as antecedents of zero anaphors in Japanese. Besides having similar morphosyntactic properties and sentence structure, Korean and Japanese also share the property that interpretations of zero anaphors are connected to grammatical topics. Therefore, we argue that the topic marker needs to be

differentiated from other postpositions and that grammatical topics are to be differentiated from other grammatical arguments like subjects and objects.

## 4.2. Morphosyntactic Information of Speech Acts

To provide a potential source for zero anaphor resolution, verbal suffixes representing sentence types are useful to add to the annotated corpora. This is because they are associated with certain speech acts such as declaration, request, question, promise, etc. and information of a missing subject can be retrieved from verbal morphology. There exist five different types of verbal inflections respectively representing a different sentence type; declaratives, interrogatives, imperatives, propositives, and exclamatives.<sup>6</sup> Among them, the imperative verbal endings suggest that a missing subject tends to refer to the hearer while promising verbal endings imply that a missing subject is the speaker. For example, the missing subjects of the following examples are respectively interpreted as *you*, *I*, and *we* based on the verbal suffixes representing a particular speech act.

- (13) a. Ø ca-ni. (Question)  
sleep-Q  
'Are (you) sleeping?'
- b. Ø ca-llyay. (Declaration)  
sleep-will  
'(I) will sleep.'
- c. Ø ca-ca. (Request)  
sleep-let's  
'Let's sleep.'

More specific classification of verbal endings could enhance the process of determining an antecedent of a zero anaphor subject. In the current annotations of the PKT, verbal suffixes are simply categorized together as final endings although the five subclasses according to sentence types have been recognized in the tagging guidelines. We propose to mark the five classes of verbal suffixes differently for developing an anaphor resolution system.

## 4.3. *Wh*-pronoun Information

Also useful for anaphor resolution is *wh*-element information. *Wh*-elements

---

<sup>6</sup> Some examples of final suffixes matching each sentence type are listed as follows.

1. Declaratives: *-ta*, *-ney*, *-o*, *-pnita*, *-nunta*, *-ci*, etc.
2. Interrogatives: *-ni*, *-kka*, *-yo*, *-nunka*, *-pnikka*, *-e*, etc.
3. Imperatives: *-la*, *-psiyo*, *-ela*, *-key*, *-o*, etc.
4. Propositives: *-ca*, *-psita*, *-cakkwuna*, *-sey*, etc.
5. Exclamatives: *-kwuna*, *-ney*, *-kwun*, *-ela*, etc.

in Korean include *nwuka* ‘who’, *mwues* ‘what’, *encey* ‘when’, *etise* ‘where’, *way* ‘why’, *ettehkey* ‘how’, etc. Answering utterances for the *wh*-questions generally contain zero elements, whose antecedents can be found in the preceding questioning utterances. In general, a fragment directly related to a *wh*-element while non-*wh*-elements previously mentioned easily drop as shown in (14).

- (14) A: John-i            Min-ul            mwe-la-ko            mitko iss-ni?  
           John-NOM   Min-ACC        what-COP-COMP   believe being-Q  
           ‘What does John believe Min to be?’  
       B: Ø                    Ø        kyoswu-la-ko            mitko iss-nuntey.  
           SUBJ                    professor-COP-COMP   believe   being-END  
           ‘(He) believes (her) to be a professor.’

Since question-answer pairs have similar argument structure, the referents of the subject and the object in (14) can be easily retrieved from the preceding question sentence. However, *wh*-elements are not distinctly tagged from other pronouns in the PKT.

Based on our classification of zero anaphors and relevant properties for anaphor resolution in Korean, an annotation example with a parsed syntactic structure of the Korean Treebank is given as follows (\* indicates newly introduced annotations).

- (16) A: 제45 사단은 어느 부대에 예속하는가 ?  
           ‘Which army does the 45<sup>th</sup> division belong?’  
       (S (NP-SBJ 제/XPF+45/NUU  
           사단/NNC+은/\*TOP\*)  
           (VP (NP-COMP 어느/DAN/\*WH\*  
           부대/NNC+에/PAD)  
           (VV 예속/NNC+하/XSV+는가/\*Q\*))  
           ?/SFN)  
       B: 제6 군단-에 예속합니다.  
           ‘(It) belongs to the 6<sup>th</sup> corps.’  
       (S (NP-SBJ \*pro-individual\*)  
           (VP (NP-COMP 제/XPF+6/NUU  
           군단/NNC+에/PAD)  
           (VV 예속/NNC+하/XSV+버니다/\*DEC\*))  
           ./SFN)  
       A: 6 군단을 또 어떻게 부르는가?  
           ‘How do you call the 6<sup>th</sup> corps?’  
       (S (NP-SBJ \*pro-deictic/individual\*)  
           (VP (NP-OBJ 6/NUU  
           군단/NNC+을/PCA)  
           (VP (ADVP 또/ADV)  
           (ADVP 어떻게/\*WH-ADV\*)  
           (VP 부르/VV+는가/\*Q\*)))  
           ?/SFN)

B: 모르겠습니다 .  
 (S (NP-SBJ \***pro-deictic**\*)  
 (VP (NP-OBJ \***pro-proposition**\*)  
 모르/VV+겠/EPF+습니다/\***DEC**\*)  
 ./SFN)

## 5. Conclusion

In this paper, we examined a classification system of zeros in Korean and an annotation scheme applicable to zero anaphor resolution systems. We evaluated the current annotation system of the PKT and Han [6]'s proposal for an annotation scheme based on the Korean Treebank. Specific problems related to zero annotations and morphosyntactic analysis in the PKT have been discussed. We showed what kind of morphosyntactic information needs to be added to a newly developed annotation scheme. This newly developed annotation scheme will be adopted for our new software handling zero anaphors in languages like Korean.

### References

- [1] Amaral, P. 2004. Inferrables with pronominal subjects in European Portuguese: implications for theories of discourse anaphora. *Proceedings of the 5<sup>th</sup> Discourse Anaphora and Anaphor Resolution Colloquium*, 1-7.
- [2] Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.
- [3] Clark, H. 1977. Bridging: In P.N. Johnson-Lard & P.C. Wason Eds. *Thinking: Readings in Cognitive Science*, 411-420. Cambridge: Cambridge University Press.
- [4] Carlson, G. 1997. A unified analysis of the English bare plural. *Linguistics and Philosophy*, 1:3, 413-458.
- [5] Han, C-H., N-R Han, and M. Palmer. 2002. Development and Evaluation of a Korean Treebank and its Application to NLP. *Proceedings of the 3<sup>rd</sup> International Conference on Language Resources and Evaluation (LREC, 2002)*
- [6] Han, N-R. 2004. Korean null pronouns: classification and annotation. *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, 33-50.
- [7] Hong M. 2000. Centering theory and argument deletion in spoken Korean. *The Korean Journal of Cognitive Science*, 11-1: 9-24.
- [8] Iida, M. 1998. Discourse coherence and shifting centers in Japanese texts. In Walker, M.A., A.K. Joshi, and E.F. Prince, (eds.), *Centering theory in discourse*. Oxford University Press, Oxford: UK.
- [9] Kameyama, M. 1985. *Zero Anaphora: the Case of Japanese*. Stanford University Doctoral Dissertation.
- [10] Kim, J-E. 1999. *A Study of Korean Modal Predicates ( Wulimal Yangthey Yongen Yenkwu)*. Thayhaksa. Seoul: Korea
- [11] Nam, K-S., and Y-G. Ko. 2002. *Standard Korean Grammar (Phyocwun Kwuke Mwonpeplon)*. Top Publishing Co.
- [12] Walker, M.A., M. Iida, and S. Cotes. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2): 193-232
- [13] Dictionary *Yonsei Korean Dictionary*. 1999. Dong-A Publishing Co.