Why Is Zero Marking Important in Korean?

Sun-Hee Lee[1,*], Donna K. Byron[2], and Seok Bae Jang[3]

[1,2] 395, Dreese Lab., 2015, Neil Avenue,Columbus, OH 43210
shlee@ling.ohio-state edu
dbyron@cse.ohio-state.edu
[3] 37th and O Sts., NW, Washington, D.C, 20057
sbj3@georgetown.edu

Abstract. This paper argues for the necessity of zero pronoun annotations in Korean treebanks and provides an annotation scheme that can be used to develop a gold standard for testing different anaphor resolution algorithms. Relevant issues of pronoun annotation will be discussed by comparing the Penn Korean Treebank with zero pronoun mark-up and the newly developing Sejong Teebank without zero pronoun mark-up. In addition to supportive evidence for zero marking, necessary morphosyntactic and semantic features will be suggested for zero annotation in Korean treebanks.

1 Introduction

This paper discusses the importance of zero pronoun marking in treebanks and investigates what kind of linguistic features are needed for treebank annotation in order to increase the usability of annotated corpora. Zero pronouns refer to empty pronouns without phonological realization, which work in a similar manner as English pronouns.

In the recent decade, there has been remarkable progress in the realm of building large corpora in Korean and applying them for linguistic research and natural language processing. Based on the broad acknowledgement of the importance of corpus and applicative tools, the 21st century Sejong project was launched in 1998 and has been developing various database and relevant computational tools including electronic dictionaries, annotation tools, morphological analyzers, parsers, etc. As a part of the Sejong project, the syntactically annotated treebank of Korean has been under construction. In addition to the Sejong Treebank (henceforth, ST), the Penn Korean Treebank (Han et al.[4] henceforth, PKT) has already been released and continues to be expanded. These treebanks with abundant linguistic information are expected to fulfill a function as informative databases in broad domains of theoretical linguistics and computational linguistics such as statistical approaches, machine learning, etc.

A notable point is that there is a critical difference between annotations of ST and PKT with respect to marking zero elements including traces, zero pronouns, etc. The most current guidelines of ST specify that zeros are dropped in order to maintain the

* This work was supported by the Korea Research Foundation Grant (KRF-2004-037-A00098) for the author.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 588 – 599, 2005.

consistency and efficiency of the treebank. In contrast, PKT advocates for representing zero elements. According to different approaches to zero marking, the structure of the following sentence (1a) is differently analyzed as in (1b) and (1c); in (1b) ST does not contain any missing subject while PKT marks the missing subject and object as *pros* in (1c)[1]

(1) a. 어제밤 12시-에 받-았-습니다.
eceypam 12 si-ey pat-ass-supnita.
last night 12 o'clock-at receive-Past-E
'Last night at 12 o'clock, (I/(s)he/they) received (it)'.
b. ST: (VP (AP 어젯밤/MAG)
(VP (NP_AJT 12/SN + 시/NNB + 에/JKB)
(VP 받/VV+았/EP+습니다/EF.+ ./SF)))
c. PKT :(S (NP-SUBJ *pro*)
(VP (NP-ADV 어젯밤/NNC)
(NP-ADV 12/NNU 시/NNX+에/PAD)
(VP (NP-OBJ *pro*)
받/VV+았/EPF+습니다/EFN))) ./SFN).

The sentence representation of (1b) does not fully present the subject-predicate relation in contrast with (1c). In this paper, we argue that failure to mark zeros may cause a loss of valuable linguistic information such as filler-gap dependencies, argument-predication relations, semantic and discourse interpretations of sentences, etc. The ST style zero-less annotation will impose the burden of zero marking on the post-annotation tasks, which utilize treebank resources for developing computational tools. This, however, is inconsistent with the purpose of developing treebanks. As pointed out in Dickinson & Meurers [3], treebanks have major usage for two types of linguists; one is for theoretical linguists who search through the corpora in order to identify certain linguistic patterns. The other is for computational linguists who use computational technology and develop statistical models from the annotated corpora in order to develop parsers and question-answer systems and to extract information such as subcategorization frames of predicates, event nouns, complex predicates, etc. In general, treebanks are manually or semi-manually annotated by humans. This guarantees more sophisticated representations of sentence structure and reliable mark-ups for ambiguous morphosyntactic units. While focusing on the usability of treebanks, we propose an argument against dropping zero mark-ups in treebanks and investigate the empirical necessity of zero annotation in Korean treebanks.

In this paper, we will discuss some significant problems of zero-less treebank annotation and explain why zero annotation is important in languages like Korean. Then we will present a general annotation scheme and features of zero pronouns that can be used to develop a gold standard for testing an anaphor resolution algorithm. Adding zero mark-up will solidify accurate syntactic representation and increase the usability of treebanks even though it takes strenuous efforts and time for development.

[1] The tagsets of ST and PKT are somewhat different (i.e., MAG represents an adverb, AP, an adverbial phrase, and NP-ADV , a nominal functioning as an adverbial modifier in ST).

2 Necessity of Zero Annotation in Korean Treebank

In contrast with English where a repeated element tends to appear as a pronoun, in topic prominent languages like Korean, a repeated element has no surface realization. Thus, Korean zero elements are often called zero pronouns.

(2) a. John met Mary yesterday.
b. Kim met her, too.
(3) a. John-i eycey Mary-lul mannassta.
John-Nom yesterday Mary-Acc met
'John met with Mary yesterday. .
b. Kim-to Ø ,mannaassta.
Kim also OBJ met
'Kim also met (zero=her).'

The discrepancy between ST and PKT with respect to zero annotation brings us two different values with respect to corpus annotation; economy vs. usability. At the stage of annotating corpora, excluding all the missing subjects from the Korean treebanks may reduce the burden of annotation tasks such as classifying zeros, sorting markable zeros, training annotators and maintaining the legitimate level of inter-annotator agreement. However, at the later stage zero marked treebanks have higher usability by higher level processing including anaphor resolution, extracting subcategorization frames of predicates, discourse analysis, etc.

More specifically, our arguments against zero-less treebanks can be presented as follows. First, zero-less treebanks may provide misleading representations with respect to the general patterns of sentence realization. In so-called pro-drop languages such as Korean, Japanese, Spanish and Portuguese, basic units of sentence structure, such as subjects of matrix clauses, are frequently unrealized. Although missing subject information in languages like Spanish and Portuguese is recoverable from verb morphology, interpretations of missing arguments do not correspond to specific verb morphology in Korean. Thus, marking the place of a zero element is an inevitable process not only for structural representation but for processing the meaning of a sentence. Zero-less treebanks license various VP or S nodes without capturing correct argument-predicate relations. For example, the following sentence is simply represented as VP, which is inconsistent with the subcategorization frame of the main verb.

(4) 그냥 잠자코 운동장-만 내다보-고 있-었-습니다.
kunyang camcakho wuntongcang-man naytapo-ko iss-ess-supnita.
just silently playground-only look down-PreP-Past-E
'(I/you/he/she/they) was only looking down the playground just silently.'

```
(VP (AP 그냥/MAG)
    (VP (AP 잠자코/MAG)
        (VP (NP_OBJ 운동장/NNG + 만/JX)
            (VP (VP 내다보/VV + 고/EC)
                (VP 있/VX + 었/EP + 습니다/EF + ./SF)))))
```

According to Hong [6], the rate of subject drop is 57% in spoken Korean, which is higher than other elements. In particular, when the subject refers to a nominal entity mentioned in the previous utterance, it naturally disappears in speech rather than ap-

pearing as a pronoun. This suggests that the number of VPs lacking subjects will be significantly high in the spoken corpora. We extracted only 100 sentences from the ST corpus containing natural spoken conversations and found that 81 sentences are represented as VPs or VNPs (predicate nominal phrases). However, it may derive a misleading generalization such that canonical sentence patterns in the given corpus are VPs or VNPs. In line with this, semantic interpretations of those incomplete VPs or VNPs subsume the meaning of the zero pronouns whose antecedents appear in the previous utterances. However, zero-less mark-up poses a difficulty in retrieving the complete sentential meaning from the given phrasal categories of VPs or VNPs.

Second, zero-less treebanks make it difficult to extract certain constructions that linguists want to identify. For example, one of the most frequently discussed topics in Korean grammar is formation of Double Subject Constructions (DSCs), which license two subjects. However, zero-less treebanks do not correctly represent Double Subject Constructions and represent (5) and (6) differently in spite of their similarity in argument realization.

(5) 햇밤-이 맛-이 썩 좋-았-습니다.
hayspam-i mas-i ssek choh-ass-supnita.
new chestnut-Nom taste-Nom quite good-Past-End
'New chestnuts had pretty good taste.'
(S (NP_SUB 햇밤/NNG + 이/JKS)
(S (NP_SBJ 맛/NNG + 이/JKS)
(VP (AP 썩/MAG)
(VP 좋/VA + 았/EP + 습니다/EF + ./SF))))

(6) 유난히 맛-이 썩 좋-았-습니다.
yunanhi mas-i ssek choh-ass-supnita.
particularly taste-Nom quite good-Past-End
'Particularly, the taste of (it) was pretty good.'
(S (AP 유난히/MAG)
(S (NP_SBJ 맛/NNG + 이/JKS)
(VP (AP 썩/MAG)
(VP 좋/VA + 았/EP + 습니다/EF + ./SF))))

According to the analysis of ST, (5) is represented as a DSC that licenses two subjects, *hayspam* and *mas*. In contrast, (6) is represented as a complex clause that only licenses a single matrix subject, *mas* and the first zero subject referring to the same nominal entity in the preceding phrase has been ignored in the sentential representation. It is difficult to extract certain syntactic patterns from the zero-less treebanks because their structural representations do not reflex the accurate argument-predicate realization. It is because they focus on surface realization of arguments instead of considering lexical constraints of argument-predicate relations.

The third critical problem of zero-less treebanks is related to discourse analysis. 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. [9], Iida [7]), Hong [6], etc. it has been shown that a salient entity recoverable by inference from the context is frequently omitted, and therefore interpreting these zero pronouns allows one to follow the center of the attentional state. Walker et al. [9] applied the centering model, developed for pronoun

resolution in English, to zero pronoun resolution in Japanese. They argue that interpretation of a zero pronoun is determined by discourse factors. This suggests that identifying occurrences of zero pronouns and retrieving their antecedents are important in developing a computational model of discourse interpretations as well as syntactic and semantic analyses. When it comes to topic information retrieval, the salient element under the discussion of the given discourse is realized as a zero. Grammatical roles and semantic restrictions provide crucial cues for the interpretations of them. However, without specifying the argument positions of these zeros, discourse processing of the given utterances is impossible.

3 Relevant Issues of Zero Annotation

Zero marked treebanks function as useful resource for researchers, especially the anaphora resolution community. For developing computational tools of anaphor resolution, it is necessary to determine 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 pointed out by Lee et al. [8] evaluating the Penn Korean Treebank (Han et. al. [4]), which includes annotations indicating the position of zero pronouns. In PKT, annotations of zeros are problematic due to inconsistent mark-up for zero pronouns and structural representation of trees. Inconsistent annotation of zero pronouns in PTK brings an imminent 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? Unless these questions are resolved, treebanks cannot fulfill their potential as a source of linguistic knowledge about zero pronouns. Also, the same question should be taken into consideration by other teams developing similar resources in other languages.

3.1 Argument vs. Adjunct

Previous authors have pointed out that the antecedents of zero pronouns can often be determined by using various grammatical properties such as topicality, agreement, tense, and aspect as well as subcategorization information (Walker et al. [9]; Iida [7]; Hong [6], 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. Thus, the first crucial step for zero pronoun resolution is identifying the exact positions of zero pronouns. Determining the positions of invisible zeros is a difficult task. This process needs to refer to the argument realization in a given utterance and the previous utterances of the same discourse unit. The argument realization of a sentence is based upon argument structure of a predicate.

(7) a. John-i . kesil-eyse swi-ko iss-ess-ta?
John-Nom living room-in rest-Pres Prog-Past-E
'John was resting in the living room.'
b. sakwa-lul mek-ess-ta.
apple-Acc eat-past-E
'(He) ate an apple.'

In (7b), the argument structure of *mekta* 'eat' suggests that the subject is missing in a sentence *sakwalul mekesse* 'ate an apple'. However, do we need to mark the adjunct *kesileyse* 'in the living room' in (7b)? In the given utterances, it seems to be possible for John to have eaten the apple in the living room but it is not necessarily true. The combinations of adjunct and predicate are not predictable by using argument structure of a predicate. With no specific guideline, identifying missing adjuncts complicates the annotation process. Thus, we argue that only missing arguments must be marked.

As for zero argument annotation, the current annotation in PKT is somewhat problematic due to unclear distinction of obligatory argument vs. optional argument. According to the guidelines of PKT, only a missing obligatory argument should be annotated as an empty element. Missing optional arguments and adjuncts are not. Thus, in PKT, missing subject or object elements were marked as zeros while missing locative arguments were not marked when they were omitted. 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, 'the 45th division-in' has not been marked up as a zero pronoun in the tagged sentence of (8b)

(8) a: 제 45 사단은 또 무엇-으로 구성되어 있는가 ?
the 45 division again what-with composed be
'What is the 45th Division composed of ?'
(S (NP-SBJ 제/XPF+45/NNU
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 지휘부-가 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 PKT. Thus, the information of the missing locative has not been represented even though it is crucial for retrieving the meaning of the sentence. Another concern with respect to distinction of argument and adjunct is that ST classifies only subject and object as arguments and excludes other case-marked nominals as adjunct[2]. This classification may cause problems when zero pronouns are added in their treebanks.

In identifying missing zero arguments, maintaining consistency is crucial. For this task, we can rely on a dictionary containing constant argument structure of predicates. Dictionaries with specific argument structure information can be used here, such as

[2] In addition to subject and object, nominals in front of predicates, *toyta*, *anita* and quotation clause have been included as arguments.

the Yonsei Korean dictionary, where different subcategorization frames are listed according to semantically disambiguated senses for each predicate. For correct identification of a zero pronoun in the given utterance, annotators need to examine the relevant previous utterances in the same discourse unit and determine the exact verb sense of the relevant predicate by using the dictionary. In addition, checking inter-annotator agreement is also an essential task (Carletta [2]).

3.2 Language Specific Properties of Korean

Another notable point is that the developers need to pay attention to language specific properties of Korean. There are some notable morphosyntactic properties of Korean with respect to zero pronoun annotation. In order to maintain constant annotation of zero pronouns, it is important to carefully represent specific features related to zero pronouns. In this section, we will discuss specific properties that can be added for zero pronoun annotation. It will increase the applicability of the treebank to both theoretical research on anaphors and computational modeling of anaphor resolution.

[1] CASE MARKING

In determining an antecedent of a zero pronoun, the existence of topics plays an important role in Korean. In the previous theoretical literature, it has been commonly assumed that the topic marked elements appear at the higher phrasal level than the phrasal combination of subject-predicate. At the discourse level, Walker et al. [9] and Iida [7] provide evidence that topic marked elements function as antecedents of zero pronouns in Japanese. The similar property has been also observed with Korean by Hong [6]. As seen in the following examples, the sentence-initial topic functions as the antecedent of zero pronouns that appear in the latter utterances. This phenomenon suggests 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.

(9) a **Seyho-nun** apeci-eykey chingchan-ul pat-ca ekkay-ka ussukhayci-pnita.
Seyho-Top father-to praise-Acc receive-E be proud of
'As for **Seyho**, he felt pride when he received praise from father.'
b. Ø 20 ilman-ey tut-nun chingchan-ila kippum-un hankyel tehaysssupnita
20days-in hear-Rel praise-since pleasure-Top far more
'Since it was the praise (he) heard for the first time in 20 days, his pleasure was much more.'
c. Ø emeni-eykey-nun nul kkwucilam-kwa cansoli-man tulesssupnita.
mother-to-Top always scolding-and lecture-only heard
'From his mother, (he) always heard only scolding and lecture.'

In general, 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 in Korean. The first is classified as a grammatical topic marker while the latter is a contrastive topic marker in traditional Korean grammar. However, the current annotations of PKT and ST treat topic marker *nun* as the same auxiliary postposition, which is similar to other postpositions *man* 'only', *to* 'also', and *mace* 'even'. In particular, PKT and ST represent a subject NP with a topic marker as the subject, while

an object with a topic marker is treated as a scrambled argument out of its canonical position. With respect to zero pronouns, the sentence initial topic marker needs to be distinctly marked from other postpositions. In addition, we claim that the structural position of a topicalized subject needs to be differentiated from a normal subject position in parallel with a topicalized object and other element, which leave zero traces in their original positions.

Another problem with case markers is subject marker *-eyse*, which only combines with nominals referring to a group or organization. In Korean, these group nominals do not take the nominative case *i/ka* but the case marker *-eyse* as in (12)

(10) wuli hakkyo-eyse wusung-ul hayssta.
our school-Nom winning-Acc did
'Our school won.'

Although *-eyse* has been treated as a nominative case marker in traditional Korean grammar, both PKT and ST do not treat *-eyse* as a nominative marker. Instead, group or organization nominals with the case marker-*eyse* are analyzed as NP adverbial phrases. This, however, mistakenly licenses a zero subject in the following example of PKT even though the subject with case marker *-eyse* exists. In order to eliminate redundant marking for the zero subjects, it is better to analyze the case marker, *-eyse* as a nominative case marker in Korean.

(11) 2 대대-에서 어떤 무전망-을 운용하-고 있-지?
2 taytay-eyse etten mwucenmang-ul wunyonha-ko iss-ci?
2 squadron-Nom which radio network-Acc use-PresP-Q
'What kind of radio network is the 2nd squadron using?'
(S (NP-ADV 2/NNU
대대/NNC+에서/PAD)
(S (NP-SBJ ***pro***)
(VP (VP (ADVP 또/ADV)
(VP (NP-OBJ 어떤/DAN
무전망/NNC+을/PCA)
(VV 운용/NNC+하/XSV+고/EAU)))
있/VX+지/EFN) ?/SFN)

[2] SUBJECTLESS CONSTRUCTIONS

Unlike English having expletive pronouns *it* or *there*, certain predicate constructions do not license subject positions at all in Korean. Some examples are presented in (12),which include incomplete predicates with few inflectional forms.

(12) *-ey tayhaye*, 'regarding on'*-ey kwanhaye*, 'about' *-ey uyhaye,* 'in terms of'
-lo inhayse 'due to~', *-wa tepwule* 'with ~' etc.

In addition, some modal auxiliary verbs like *sangkita*, *poita*, *toyta*, etc. do not license subject positions and have already been classified as subjectless constructions in Korean grammar. While ST treats these modal verbs to be included in the preceding verbal clusters, PKT separates them from the preceding verbs and assigns zero subjects for these verbs. Thus, the PKT approach redundantly assigns zero subjects for subjectless predicates.

[3] VERBAL MORPHOLOGY OF SPEECH ACT

As for zero pronoun resolution, verbal suffixes representing speech act can be a useful source. Thus, we argue for adding these morphosyntactic features in treebanks. It has been well known that in Korean certain speech acts such as declaration, request, question, promise, etc. are associated with verb morphology; five different types of verbal inflections are used to indicate declaratives, interrogatives, imperatives, propositives, and exclamatives. Information of a missing subject can be retrieved from verbal morphology. For example, the imperative verbal endings suggest that a missing subject refers to the hearer while promising verbal endings imply that a missing subject is the speaker. Thus, the missing subjects of the following examples are respectively interpreted as I, you and we based on the verbal suffixes representing a particular speech act.

(13) a. Ø ka-llay. (Question)
go-Q
'Do (you) want to go?'
b. Ø ka-llay. (Declaration)
go-will
'(I) will go.'
c. Ø ka-ca. (Request)
go-let's
'Let's go.'

Verbal endings of speech acts can be used to enhance the process of determining an antecedent of a zero pronoun subject. In the current annotations of PKT and ST, verbal suffixes do not subclassify the final endings. We argue that annotating the five classes of verbal suffixes differently will facilitate application of anaphor resolution algorithms on treebanks.

[4] WH-PRONOUN TAGGING

Wh-pronouns in Korean include *nwuka* 'who', *mwues* 'what', *encey* 'when', *etise* 'where', *way* 'why', *ettehkey* 'how', etc. Unfortunately, *wh*-pronouns are not distinctly tagged from other pronouns in the PKT and the ST. The information of *wh*-pronouns can be useful for resolving the meaning of zero pronouns in the next answering utterance As seen in (14b), a fragment directly related to a *wh*-pronoun necessarily appears in the answering utterance while non-*wh*-elements previously mentioned are easily dropped. This is because pairs of *wh*-question-answer tend to have the same predicates with the same argument structure. Therefore, answering utterances of the *wh*-questions generally contain zero pronouns, whose antecedents appear in the preceding questioning utterances.

(14) A: John-i Min-ul mwe-la-ko mitko iss-ni?
John-Nom Min-Acc what-Co-Comp believe being-Q
'What does John believe Min to be?'
B: Ø Ø kyoswu-la-ko mitko iss-nuntey.
SUBJ OBJ professor-Co-Comp believe being-END
'(He) believes (her) to be a professor.'

4 New Annotation Scheme of Korean Zero Pronouns

Once zeros are identified by argument structure information of predicates and the previous utterances in the given discourse, the additional reference information can be added in treebanks to support anaphor resolution. Zeros in Korean can be classified into different classes according to properties of their reference. Anaphor resolution algorithms can be applied for certain types of pronouns. For example, in order to retrieve the meaning of a zero pronoun referring to a nominal entity in the previous utterance, the resolution algorithm will search nominal entities that appear in the previous utterance by making a list of antecedent candidates and selecting the most appropriate candidate. In contrast, the searching algorithm does not need to apply for a zero element referring to an indefinite entity as in (15).

(15) Ø holangi kwul-ey ka-ya holangi-lul capnunta.
tiger den-to go-E tiger-Acc catch
(Lit.) 'One should go to the tiger's den in order to catch a tiger.
(Trans.)'Don't hesitate but pursue what you need to do.'

According to reference relation between a zero pronoun and its antecedent, zero pronouns in Korean can be divided into three classes as in Table 1; discourse anaphoric zeros, deictic and indexical zeros, and indefinite zeros.

In the given classification, 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. The discourse anaphoric zeros can be divided into three subclasses according to the semantic properties of their antecedents.

Table 1. Classification of Korean Zero Pronouns

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

The first subclass of discourse anaphoric zeros refers to individual domain entities, the second, eventualities, and the third, propositions. The zeros of individual entities refer to entities that were introduced into the discourse via noun phrases. Most examples presented in the previous sections correspond to this class. The zeros of propositions refer to propositions introduced in the previous utterance as in (16).

(16) 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 108th regiment located?'
B: Ø1 Ø2 molukeyss-supnita.
SUBJ OBJ not know-END
'I don't know.'
(Ø1 = 'B', Ø2= 'Where the headquarter of the 108th regiment is located ')

The third class of zero anaphors referring to eventualities, i.e. action and event as in (17) (Asher [1]).

(17) A: Mary-ka cip-ey - ka-ko sipheha-ci anha.
Mary-Nom home-to go-E want-END don't
'Children don't want to go home.'
B: na-to Ø silhe.
I-also hate
'I also hate to go home.' (Ø = the action of going home)

The second class of zero pronouns 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.

Given the classification of zero pronouns, different coding systems can be provided for each class for annotating these elements. According to different classes of zeros, the resolution process varies. Zero anaphors of discourse anaphoric entities will be marked the same as their antecedents in the previous utterances. Anaphor resolution algorithms determine the antecedent of a zero anaphor by searching through the antecedent candidates in different orders. Deictic and indexical zeros are dependent on discourse participants. In general, a zero anaphor can also refer to the speaker or the hearer. Overlapping mark-up for these zeros need to be allowed although resolution mechanisms for deictic and indexical zeros are different from those for anaphors. Indefinite zeros need to be marked but anaphor resolution algorithms do not need to be applied to them.

5 Conclusion

In this paper, we discussed why zero marking is necessary for Korean treebanks and how invisible zeros can be consistently marked in annotated corpora like treebanks. The importance of zero mark-up in Korean treebanks has been discussed with respect to correct linguistic analysis and efficient application of computational process. We also claimed that only missing arguments are marked as zeros and a dictionary like Yonsei Dictionary with full specification of argument-predicate relations can be a useful source for the annotation task. By examining PKT and the newly developing ST, we determined four linguistic features that are useful for anaphor resolution in Korean; case marking, subjectless construction, verb morphology of speech acts and *wh*-pronoun tagging. In addition, we provided a new annotation scheme that can be utilized for annotating treebanks and testing anaphor resolution algorithms with annotated corpora.

References

1. Asher, N.:. Reference to Abstract Objects in Discourse. Kluwer Academic Publishers. (1993).
2. Carletta, J. Assessing Agreement on Classification Tasks: the Kappa Statistic, Computational Linguistics 22(2) (1996) 249-254.
3. Dickinson, M. and Meurers, D.: Detecting Inconsistencies in Treebanks in Proceedings of the Second Workshop on Treebanks and Linguistics Theories.(TLT 2003)..Växjö. Sweden. (2003)

4. Han, C-H., Han, N-R., Ko, E-S.and Palmer, M.: Development and Evaluation of a Korean Treebank and Its Application to NLP.in Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC).(2002)
5. Han, N-R.: Korean Null Pronouns: Classification and Annotation in Proceedings of the ACL 2004 Workshop on Discourse Annotation. (2004) 33-40.
6. Hong M.: Centering Theory and Argument Deletion in Spoken Korean. The Korean Journal Cognitive Science. Vol. 11-1 (2000) 9-24.
7. Iida, M.: Discourse Coherence and Shifting Centers in Japanese texts in Walker, M., Joshi A.K., Prince E.F. (Eds.) Centering Theory in Discourse. Oxford University Press, Oxford: UK..(1998) 161-182.
8. Lee, S., Byron, D., and Gegg-Harrison, W.: Annotations of Zero Pronoun Resolution in Korean Using the Penn Korean Treebank in the 3rd Worksop on Treebanks and Linguistics Theories (TLT 2004). Tübingen. Germany. (2004) 75-88.
9. Walker, M., Iida, M., .Cotes, S.: Japanese Discourse and the Process of Centering in Computational Linguistics, Vol. 20-2.: (1994.) 193-232
10. 10. Dictionary Yonsei Korean Dictionary. (1999) Dong-A Publishing Co.
11. Guidelines of the Sejong Treebank. Korea University