Enriching the Syntactic Annotation of Korean Treebanks for Higher-level Processing: A Comparative Study of the Penn Korean Treebank and the 21st Sejong Korean Treebank

Sun-Hee Lee
Ohio State University

Seok Bae Jang
Georgetown University

Enriching the Syntactic Annotation of Korean Treebanks for Higher-level Processing: A Comparative Study of the Penn Korean Treebank and the 21st Sejong Korean Treebank

Enriching the Syntactic Annotation of Korean Treebanks for Higher-Level Processing: A Comparative Study of the Penn Korean Treebank and the 21st Sejong Korean Treebank

Sun-Hee Lee & Seok Bae Jang

1 Introduction

This paper explores several important issues in developing syntactically annotated Korean corpora for higher-level language processing, including semantic-discourse parsing, question-answering, machine translation, information retrieval, etc. In particular, we compare the Penn Korean Treebank (PKT) and the Korean Treebank of the 21st Century Sejong Project (ST) and discuss four critical issues in syntactic annotation. We argue for the use of more sophisticated morphosyntactic information, and based on our comparative study, we propose revisions in the syntactic annotation schemes of the existing Korean Treebanks in order to improve the quality of annotated corpora and their usability both for conducting theoretical research and for developing computational tools.

The results of our comparative study reveal four significant issues in syntactic annotations: the syntactic analysis of verbal complexes, the hierarchical structure of noun phrases, the representation of traces, and the marking of zero elements. These factors may trigger erroneous syntactic representations for certain linguistic phenomena, and they may increase difficulties in data search and lessen reliability in computational processing. Thus, evaluating and improving the syntactic annotation of Treebanks is an important task for aspects of both theoretical and computational linguistics.

2 Properties of the PKT and ST

In this section, we briefly summarize the organization and characteristics of the PKT and ST. Basically, the formation of parsed sentences in the PKT and ST is not much different in that they both use phrase structure annotation for syntactic bracketing and similar phrasal categories and lexical categories. In addition, both treebanks also provide information regarding morphological combinations. However, they have distinct properties in terms of their content and their analyses of particular constructions.

The corpus on which the PKT is based is composed of texts from artificial military training manuals. The corpus is one part of English and

Korean bilingual corpora developed for a machine translation project at the University of Pennsylvania. The corpus includes dialogues between a member of the military and a captive. Thus, it mostly contains question and answer pairs. The content reflects information about the military such as troop movement, intelligence gathering, and equipment supplies. The PKT includes about 54,000 words and 5,000 sentences according to Han et al. (2002). An example from the PKT is given as follows.

(1) A;01:4:대 대 에 서 는 어 떤 구 분 대 들 이 지 휘 망 에 들 어 가 지 ?
(S (NP-ADV대 대 /NNC+에 서 /PAD+는 /PAU)
(S (NP-SBJ어 떤 /DAN
구 분 대/NNC+들 /XSF+이 /PCA)
(VP (NP-COMP지 휘 망 /NNC+에 /PAD)
들 어 가 /VV+지 /EFN)) ?/SFN)
B;01:5:보 병 중 대 들 뿐 이 지 요 .
(S (NP-SBJ *pro*)
(VP (NP보 병 /NNC
중 대 /NNC+들 /XSF
뿐 /NNX+이 /CO+지 요 /EFN)) ./SFN)

The ST has been under construction as part of the 21st Century Sejong Project, which was launched in 1998 to build huge national Korean corpora. The guideline for the ST and some parsed outputs are currently available. The version of the ST that we cite in this paper includes eleven documents, including news articles and books in the humanities. It includes about 127,000 words and 10,600 sentences as of 2003. An example from the ST is given below.

(2) 고 향 처 럼 깊 은 꿈 을 안 겨 주 는 말 은 없 다 .
(S- (S (NP_SBJ (VP_MOD (NP_AJT고 향 /NNG+처 럼 /JKB)
(VP_MOD (NP_OBJ (VP_MOD
깊 /VA+은 /ETM)
(NP_OBJ꿈 /NNG+을 /JKO))
(VP_MOD (VP안 기 /VV+어 /EC)
(VP_MOD주 /VX+ 는 /ETM))))
(NP_SBJ말 /NNG+은 /JX))
(VP없 /VA+다 EF)) (S- + ./SF))

In terms of content, the corpus on which the ST is based is more balanced than the one on which the PKT is based. The ST includes various texts such as novels, newspaper articles, fairy tales, etc., while the PKT consists only of spoken dialog from military training materials. The structural analysis in the ST includes fewer embeddings than the PKT, because the ST does not assume an empty category in the position of a trace or a missing element. Rather, the PKT marks the positions of a trace or a missing argument. Thus, the structural representations in the two treebanks are significantly different although they use similar phrasal categories and morphological analyses.

3 Issues in Syntactic Annotations

In this section, we focus on features of syntactic annotations that are necessary in Korean treebanks by comparing the PKT and ST. Four kinds of grammatical factors are discussed, including information in or regarding verbal complexes, the hierarchical structure of NPs, the representation of traces, and zero anaphor marking.

3.1 Verbal Complexes

The first point of comparison involves syntactic structures of verbal complexes. The PTK separates each component of a verbal complex and allows each auxiliary verb in that complex to project to a VP, as in (3a). This contrasts with the syntactic analysis of the ST, which combines verbal complexes under the same phrasal category, as in (3b).

(3) a.대 대장 -의 허가 없이 -는 쓰지 못하게 되어 있습니다.
Taytaycang-uy heka epsi-nun ssuci moshakey toye isssupnita.
Commander's permission not-Top use cannot become be
'(It) is not supposed to be used without permission from the battalion commander.'

(VP (VP (VP (ADVP (NP-COMP (NP대 대장 /NNC+의 /PCA)
(NP허가 /NNC))
없 이 /ADV+는 /PAU)
(VP (NP-OBJ *T*-1)
쓰 /VV+지 /EAU))
못하 /VX+ 게 /EAU)
되 /VX+어 /EAU)
있 /VX+습 니 다 /EFN))

b. 악몽 -의 순간 -을 되새기 -고 싶어하 -지 않았다 .
akmong-uy swunkan-ul toysayki-ko sipeha-ci anhassta
nigthmare-GENmoments-Acc remember-End want-END didn't
'(I) didn't want to remember the moments of the bad dream.'
(VP (NP_OBJ (NP_MOD악 몽 /NNG+의 /JKG)
(NP_OBJ 순간 /NNG+을 /JKO))
(VP (VP (VP되새기 /VV+ 고 /EC)
(VP싶 /VX+어 /EC+ 하 /VX+지 /EC))
(VP않 /VX+았 /EP+다 /EF)))

In (3), each component of a verbal complex is separated and each auxiliary verb in that complex projects to a VP. While the main verb first combines with an object and then combines with the following auxiliary VPs in (3a), in (3b) the main verb first combines with the following auxiliary VPs and then the whole verbal complex combines with an object NP.

Given the agglutinative properties of Korean, which license strong morphosyntactic dependencies among multiple verbal elements, we argue for a unified syntactic annotation for verbal clusters in Korean. This argument is supported by previous research on Korean verbal complexes, including Choi (1971), Kang (1996), Nam & Ko (2002), etc. In verbal complexes, morphological inflections of agreement, aspect and tense, such as the past tense morpheme *-ass-* in (3b), appear on auxiliary verbs but not on the main verb. This shows that verbal complexes have a morphological dependency. In addition, lexical insertion between verbal clusters is not allowed as in (4).

(4) a. *쓰 지 대대장 -의 허가 없이 못하게 되어 있습니다 .
ssuci taytaycang-uy heka epsi moshakey toye isssupnita.
use commander's permission not cannot become be
b. *쓰 지 못하게 대대장 -의 허가 없이 되어 있습니다 .
ssuci moshakey taytaycang-uy heka epsi toye isssupnita.
use cannot commander's permission not become be
c. *쓰 지 못하게 되어 대대장 -의 허가 없이 있습니다 .
ssuci moshakey toye taytaycang-uy heka epsi isssupnita.
use cannot become commander's permission not be

It is also not possible to move a single verbal element out of a verbal complex as shown in the following example.

(5) a. *쓰 지 대대장 -의 허가 없이 _못하게 되어 있습니다.
ssuci taytaycang-uy heka epsi moshakey toye isssupnita.
use commander's permission not cannot become be
b. *못하게 대대장 -의 허가 없이 쓰지 _되어 있습니다.
moshakey taytaycang-uy heka epsi ssuci toye isssupnita.
cannot commander's permission not use become be
c. *되어 대대장 -의 허가 없이 쓰지 못하게 _있습니다.
toye taytaycang-uy heka epsi ssuci moshakey isssupnita.
become commander's permission not use cannot be

Splitting the elements of a verbal cluster as in the PKT expands the VP structure at the highest level of the sentence. Thus, it is difficult to extract verbal complexes that are composed of only verbal elements. For example, an object NP and other complements, as well as adjuncts, combine with the main predicate first and form a VP, which later combines with the following auxiliary verb and forms another VP at the higher level. Because the embedded VP includes extra elements like the object NP, extracting pure verbal combinations is a difficult task in spite of the clear morphological dependencies among verbal elements. Therefore, we argue that analyzing a verbal complex as a single unit is better because it correctly captures morphosyntactic properties of Korean and makes it easy to extract verbal combinations.

3.2 The Hierarchical Structure of NPs

The second critical issue in treebank design relates to the hierarchical structure of NPs. In the PTK, all nouns appearing in an NP are licensed in a flat structure. According to the PKT, the structure of a noun cluster will be represented as follows.

(6) (NP미 국 /NPR 사우스웨스트 /NPR 항공 /NNC 소속 NNC
(mikwuk sauswest hangkong sosok
America Southwest airline belonging to
여객기 /NNC 1/NNU 대 /NNX)
yekaykki han tay)
airplane 1 CLF
'One airplane that belongs to Southwest Airline'

In contrast, the ST brackets nouns that appear in a semantically close relation as we see in the following example.

(7) 우리 나라 국보 8만 대장경 -이
wuli nala kwukpo phalmantaycangkyeng-i
we country national treasure phalmantaycangkyeng-NOM
'the national treasure of our country, phalmantaycangkyeng'
NP_SBJ (NP (NP (NP우 리 /NP)
(NP나 라 /NNG))
(NP국 보 /NNG))
(NP_SBJ 8만 대장경 NNP +이 /JKS))

The flat structure approach is potentially problematic because it may assign incorrect modification relations to examples like (6). For example, we can add a modifying phrase to (6) as follows.

(8) (최 근 설립된)
((choykun selliptoy-n)
recently establish-REL
(NP미 국 /NPR 사우스웨스트 /NPR 항 공 /NNC 소 속 /NNC
(mikwuk sauswest hangkong sosok
America Southwest airline belonging to
여 객기 /NNC 1/NNU 대 /NNX))
yekaykki han tay))
airplane 1 CLF
'One airplane that belongs to Southwest Airline, which has been recently established'

In (8), *choykun-ey selliptoyn* 'recently established' modifies *sauswest hangkong* 'Southwest Airline' and not the entire NP corresponding to 'one airplane that belongs to Southwest Airline'. However, the flat structure representation used in the PKT does not capture the exact modification relation. Instead it introduces unnecessary ambiguity to the given example.

Furthermore, the flat structure analysis tends to increase computational complexity by allowing too many tokens of noun complexes. For example, it is possible to provide multiple analyses for the unambiguous example in (9a). While the correct analysis should be (9b), the flat structure analysis also allows (9 c-e).

(9) a. (NP 우리 엄마 가죽 지갑 속)
wuli emma kacwuk cikap sok
we mother leather wallet inside
'the inside of our mother's leather wallet'

b. (NP 우리 엄마) ((가죽 지갑) 속))
we mother leather wallet inside
c. *(NP 우리 ((엄마 가죽) 지갑) 속))
we mother leather wallet inside
d. *(NP (우리 엄마) 가죽) (지갑 속)
we mother leather wallet inside
e. *(NP 우리 엄마) (가죽 (지갑 속))
we mother leather wallet inside etc.

Taking into consideration correct modification relations and computational complexity, we argue for a hierarchical representation of NPs. In particular, the structure of NPs in Korean can be easily expanded by adding more nouns, which contrasts with English. This is partially due to the existence of sino-Korean nominals originating from Chinese and a language-specific phenomenon of case marker dropping. In most nominal complexes, certain nouns show more intimate semantic relations. Thus, it is more efficient and reasonable to provide hierarchical structures for noun clusters by combining clusters with clear semantic relations.

3.3 The Representation of Traces

Another crucial issue with respect to Korean treebanks involves the representation of traces. While the PTK assumes traces for certain long-distance dependency constructions, the ST simply does not assume traces at all. The former approach overgenerates trace constructions by assigning empty *wh*-operators to relative clauses as in (10). In contrast, the latter undergenerates traces and fails to capture the syntactic and semantic dependency between a trace and its filler as in (11).

(10) 5중대 -에서 사용하 -는 R-116 무전기 -의 주파수
5cwungday-eyse sayongha-nun R-116 mwucenki-uy cwuphaswu
5th company-in use-Rel R-116 radio-Gen frequency
범위 -는 얼만 -가 ?
pemwy-nun elman-ka?
range-Top what-Q
'What is the frequency range of the R-116 radio which the 5th company uses.'
(S (NP-SBJ (NP (S (**WHNP-1 *op***
(S (NP-SBJ *pro*)
(VP (NP-ADV 5/NNU

중 대 /NNC+에 서 /PAD)
(VP (**NP-OBJ *T*-1**)
(VV사 용 /NNC+하 /XSV+ 는 /EAN)))))
(NP R-116/NFW
무 전 기 /NNC+의 /PCA))
(NP 주파수 /NNC
범 위 /NNC+는 /PAU))
(VP (NP얼마 /NPN+ 이 /CO+ㄴ 가 /EFN)) ?/SFN)

(11) 그것 -은 **영재 아버지 -가 생전 -에 써둔** 묵은
Kukes-un **Yengcay apeci-ka sayngcen-ey ssetwu-n** mwukun
that-Top **Yengcay apeci-Nom while alive-in write-Rel** old
일기책이었다 .
ilkichayk-iessta.
diary-Cop
'It was the old diary **that Yengjay's father wrote while he was alive**.'
(S- (S (NP_SBJ그 것 /NP +은 /JX)
(VP (S_MOD (NP_SBJ (NP**영 재** /**NNP**)
(NP_SBJ**아 버 지** /**NNG** + 가 /**JKS**))
(VP_MOD (NP_AJT**생 전** /**NNG** + 에 /**JKB**)
(VP_MOD**쓰** /**VV**+어 /**EC**+**두** /**VX**+ ㄴ /**ETM**)))
(VP (VP_MOD묵 /VV +은 /ETM)
(VP일 기 책 /NNG +이 /VCP +었 /EP +다 /EF))))
(S- +./SF))

As shown in (10), the missing object in a relative clause is marked as ***T*** and the invisible *wh*-operator is assumed to be at the top position of the relative clause and is marked as ***op*** in the PKT. However, the missing object of a relative clause is not marked at all in the ST as shown in (11).

Although the PKT marks trace information in Korean relative clauses and topicalized constructions, the semantic binding between a trace and its coreferential antecedent (or filler) is not marked. For example, the missing object in (10) refers to the same object as the head noun *mwucenki* 'radio'. However, there is no mark-up that captures this semantic relation. In contrast, the position of an invisible *wh*-operator has been marked even though Korean does not have relative pronouns.

Korean does not have long-distance dependency constructions with *wh*-phrases as in English. However, previous research, including Kang (1996), Lee et al. (2004), etc., showed that long-distance dependency constructions

that license traces exist and include relative clauses, topic constructions, and *tough* predicate constructions. In order to retrieve the meaning of long-distance dependency constructions, the trace information needs to be syntactically represented, and furthermore, semantic binding between the trace and its filler needs to be specified. By adding syntactic and semantic annotations of traces, treebanks can provide correct structural representations and semantic parsing for long-distance dependency constructions.

With respect to topic constructions, the PKT analysis is inconsistent; the topicalized subject is treated like the subject without licensing a trace in the subject position as in (12). However, the topicalized object or complement licenses a trace in its original position, appearing in the sentence initial position as in (13). While the topicalized subject *taytay hochwul tayho* appears in the subject position without licensing a trace in (12), the topicalized object *kwuenhan* in (13) licenses a trace in its original position as object of the predicate *kaciko*.

(12) 대대 호출 대호 -는 "정산 "이지요 .
taytay hochwul tayho-nun Jengsan-iciyo
battalion calling code-Top Jengsan-Cop
'The calling code of the battalion is "Jengsan."'
(S (NP-SBJ대 대 /NNC
호 출 /NNC
대 호 /NNC+는 /PAU)
(VP (NP "/SLQ
(NP정 산 /NPR)
"/SRQ
이 /CO+지 요 /EFN)) ./SFN)

(13) 주파수를 바 꾸 는 권한 은 대 대 참모장 이
cuphaswulul pakkwu-nun kwuenhan-un taytay chammocangi
frequency change-Rel right-Top battalion chief of staff
가지고 있읍니다 .
kaciko isssupnita.
have be
'As for the right to change the frequency, the chief of staff has it.'
(S (NP-OBJ-1 (S (NP-SBJ *pro*)
(VP (NP-OBJ주 파 수 /NNC+를 /PCA)
바 꾸 /VV+는 /EAN))
(NP권 한 /NNC+은 /PAU))
(S (NP-SBJ대 대 /NNC

참 모장 /NNC+이 /PCA)
(VP (VP **(NP-OBJ *T*-1)**
가 지 /VV+고 /EAU)
있 /VX+습 니다 /EFN)) ./SFN)

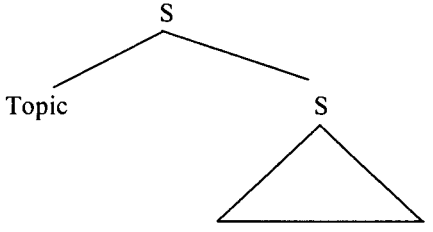
The only difference between subject topicalization and object topicalization is that the topicalization of the subject does not change the linear order of the given sentence. However, the PKT approach is also inconsistent for object topicalization. Consider the following example where the topicalized object appears with the missing subject.

(14) 신호는 언제나 사용합니다.
sinho-nun enceyna sayonghapnita
signal-Top always use
(S (NP-SBJ *pro*)
(VP (NP-OBJ 신호 /**NNC**+ 는 /**PAU**)
(VP (ADVP언 제나 /ADV)
(VP (VV사 용 /NNC+하 /XSV+ㅂ 니다 /EFN)))) ./SFN)

The linear order has not been changed in the topicalized sentence in (14). Thus, the topicalized element has been analyzed as appearing in the original object position without licensing a trace, which contrasts with (13).

In general, a topicalized element appears in sentence initial position with the topic marker *nun/u.* Thus, the topicalized sentences have been analyzed as having the following structure.

(15)



Given the general properties of Korean topic constructions, a consistent analysis is one in which the topicalized element appears in the topic position, which is higher than the position of the subject, while its trace is licensed in the original position, as in (13).

3.4 Zero Anaphor Marking

The final point relates to the syntactic marking of zero elements, which are different from traces. In topic prominent languages like Korean and Japanese, a repeated nominal element has a null surface realization, called a *zero pronoun*, in contexts where an explicit pronoun would be used in English. 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?

According to the PKT guidelines, only missing obligatory arguments are marked as *pros*, and this excludes missing optional arguments and adjuncts. Issues regarding zero elements in the PKT have been already discussed in Lee et al. (2004). They argue that zero pronoun mark-ups of the PKT are inconsistent because they fail to clarify the concept of obligatory vs. optional argument and because they pose unnecessary zero positions for subjectless predicates or idiosyncratic expressions.

In contrast to the PKT, which overgenerates zero pronouns, the ST analysis does not include any empty categories in a sentence by arguing that only the surface structure of a given sentence is considered. However, this approach loses all the information required for the retrieval of semantic interpretations as well as for correct syntactic representations. At the discourse processing level, unrealized elements are important for tracking the attentional state of a discourse or the center of a given dialog in topic-oriented languages like Korean and Japanese. This has been shown in Walker et al. (1994), Iida (1998), Hong (2000), etc.

In addition, from a practical point of view, conducting language processing tasks in treebanks without zero element mark-ups makes it difficult to extract exact argument realization patterns of predicates. Treebanks are a useful resource for identifying the argument realization patterns of a certain predicate. The patterns can be correctly captured when missing argument information is specified in the treebank and compared to the subcategorization frame of the given predicate.

By considering the importance of zero anaphors, we claim that treebank annotations need to be developed that put additional focus on how information of zero anaphors can be identified and marked and what kind of linguistic features are necessary for applying anaphor resolution algorithms. A detailed discussion of anaphor annotations can be found in Lee et al. (ms.).

4 Additional Features

In addition to suggesting four issues relating to syntactic and semantic

annotations for high-level processing and correct linguistic analyses, we argue for adding more sophisticated morphosyntactic classifications. For example, speech act information is a necessary morphosyntactic feature for treebank annotation. In Korean, five different types of verbal suffixes are associated with different speech acts, such as declaratives, interrogatives, imperatives, propositives, and exclamatives. Some examples are given in (16).

(16) a. Ø 자 -니 ?
Ø *ca-**ni***? (Question)
sleep-Q
'Are (you) sleeping?'
b. Ø 잘 -래
Ø *ca-**llay***. (Declaration)
sleep-will
'(I) will sleep.'
c. Ø 자 -자
Ø *ca-**ca***. (Request)
sleep-let's
'Let's sleep.'

The morphological information associated with speech acts can be useful for identifying sentence types and analyzing discourse structures, as well as developing question-answering systems.

Another useful piece of morphosyntactic information is to specify verbal nouns that require arguments. As shown in (17), the verbal noun *myenglyeng* 'command' licenses an embedded clause as its argument. During the last decade, verbal noun constructions have been one of the most controversial topics in Korean linguistics. The current treebank annotations do not have verbal noun information, although they do provide mark-ups of some light verbs that combine with verbal nouns. Systematic mark-ups of verbal nouns will make it possible to extract the exact patterns of their argument realization in corpora. Therefore, the annotated treebanks can be used as a tool for identifying linguistic hypotheses and drawing generalizations. In addition, marking verbal nouns will present information that is crucial for event nominal tagging that is part of TimeML (http://www.timeml.org). TimeML is a markup language for temporal and event expressions, and it pursues temporal and event recognition for question and answering systems. Annotations of verbal nouns will increase the usability of treebanks by facilitating event taggings using TimeML.

(17) 통제 소대 -에 지휘 관측소 -를 점령하라 -고
thongcey sotay-e cihwi kwanchukso-lul cemlyenghala-ko
controlling platoon-to command observatory-Acc occupy-to
명령을 내림 -으로써
myenglyeng-ul naylim-ulosse
command-Acc put-ing-by
'By putting a command to occupy a commanding observatory to the controlling platoon.'
(VP (NP-ADV통 제 /NNC
소 대 /NNC+에 /PAD)
(VP (S-COMP (NP-SBJ *pro*)
(VP **(NP-OBJ 지휘 /NNC**
관 측 소 /NNC+를 /PCA)
(VV점 령 /NNC+하 /XSV+라 고 /ECS)))
(NP-OBJ명 령 /NNC+을 /PCA*Verbal Noun*)
내 리 /VV+ㅁ /ENM+으 로 써 /PAD))))

5 Conclusion

Treebank annotations are a significant and useful source for applied language processing in addition to theoretical linguistic research. Thus, it is crucial to represent linguistic phenomena correctly and to provide more enriched information with broad applicability. In this paper, we examined four issues with respect to Korean treebank annotation and suggested how they can be properly handled. In addition, we argue that adding more sophisticated morphosyntactic and discourse features will improve the quality of annotated corpora and increase their usability.

References

Choi, Hyeon-bae. 1971. *Wulimalpon.* Seoul: Jeongumsa.

Han, Chung-hye, Na-Rae Han, Eon-suk Ko, Heejong Yi, and Martha Palmer. 2002. Development and Evaluation of a Korean Treebank and its Application to NLP. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC, 2002).

Han, Na-Rae. 2004. Korean null pronouns: Classification and annotation. *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, 33–50.

Hong, Minpyo. 2000. Centering theory and argument deletion in spoken Korean. *The Korean Journal of Cognitive Science* 11(1):9–24.

Iida, Masayo. 1998. Discourse coherence and shifting centers in Japanese texts. In *Centering Theory in Discourse*, ed. Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince. Oxford University Press.

Kameyama. 1985. Zero Anaphora: The Case of Japanese. Doctoral dissertation, Stanford University.

Kang, Hyun-hwa. 1996. Tongsa Yenkyel Kwusenguy Tatankyesengey Kwanhan Yenkwu. Doctoral dissertation, Yonsei University.

Lee, Sun-Hee, Donna Byron, and Whitney Gegg-Harrison. 2004. Annotations of Zero Pronoun Resolution in Korean Using the Penn Korean Treebank. In *The 3rd Workshop on Treebanks and Linguistic Theories (TLT 2004)*, 75–88. Tübingen, Germany.

Lee, Sun-Hee and Seok Bae Jang (ms.) *Why Is Zero Marking Important?*

Nam, Ki-shim and Young-Guen Ko. 2002. *Standard Korean Grammar* (*Phyocwun Kwuke Mwunpeplon*). Top Publishing Co.

Walker, Marilyn A., Masayo Iida, and Sharon Cotes. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2):193–232.

Yonsei Korean Dictionary. 1999. Doosan Dong-A Publishing Co.

Guidelines of the Sejong Treebank. Korea University.

Department of Computer Science & Engineering
Ohio State University
Columbus, OH 43210
shlee@ling.ohio-state.edu

Department of Linguistics
Georgetown University
Box 571051
37th and O Streets, NW
Washington, D.C. 20057-1051
sbj3@georgetown.edu