Parsing Mixed Constructions in a Type Feature Structure Grammar

Jong-Bok Kim[1] and Jaehyung Yang[2]

[1] School of English, Kyung Hee University, Seoul, Korea 130-701
[2] School of Computer Engineering, Kangnam University, Kyunggi, Korea, 449-702

Abstract. Because of the mixed properties of nominal and verbal properties, Korean gerundive phrases (GPs) posit intriguing issues to both theoretical as well as computational analyses. Various theoretical approaches have been proposed to solve this puzzle, but they all have ended up abandoning or modifying fundamental theory-neutral desiderata such as endocentricity (every phrase has a head), lexicalism (no syntactic rule refers to the word-internal structure), and null licensing (abstract entities are avoided if possible) (cf. Pullum 1991, Malouf 1998). This paper shows that it is possible to analyze and efficiently parse the mixed properties of Korean GPs in a way that maintains the desiderata while avoiding abstract entities. This has been achieved through Korean Phrase Structure Grammar, an extension of HPSG that models human languages as systems of constraints on typed feature structures. The feasibility of the grammar is tested by implementing it into the LKB (Linguistics Knowledge Building) system (cf. Copestake 2002).

1 Mixed Properties of Korean Verbal Gerundive Phrases

Like English, Korean gerundive phrases (GP) display verbal properties internally and nominal properties externally (Chung et al. 2001). It is not difficult to find out that they exhibit verbal properties in terms of the internal syntax. One telling piece of evidence comes from the fact that the gerundive verb inherits the arguments from the lexeme from which it is derived. As shown in (1), the gerundive verb takes the same arguments as the lexeme, a nominative subject and an accusative object:[3]

(1) [John-i ku chayk-ul/*uy ilk-ess-um]-i myenghwak-hata
John-NOM that book-ACC/*GEN read-PAST-NMLZ-NOM clear-do
'John's having read the book is clear'

Various other phenomena also show that GPs are internally similar to VPs. The GP can include a sentential adverb as in (2)a; an adverbial element can modify the gerundive verb as in (2)b; the GP can include the sentential negation marker *an* 'not' as in (2)c; it also can contain the full range of auxiliaries as in (2)d:

[3] The paper adopts the following glosses: ACC (accusative), ARG-ST (argument structure), COMP (complementizer), DAT (dative), DECL (declarative), GEN (genitive), HON (honorific), NMLZ (nominalizer), NOM (nominative), NEG (negation), REL (relativizer), SYN (Syntax), SEM (semantics), TOP (Topic), and the like.

(2) a. John-i **papokathi** ku chayk-ul ilk-ess-um
John-NOM foolishly that book-ACC read-PAST-NMLZ
'John's having read the book foolishly'
b. John-i chayk-ul **ppalli/*ppalun** ilk-um
John-NOM book-ACC fast(adv)/*fast(adj) read-NMLZ
'John's reading books fast.'
c. John-i chayk-ul **an** ilk-um
John-NOM book-ACC NEG read-NMLZ
'John's not reading books.'
d. John-i chayk-ul ilk-ko **siph**-um
John-NOM book-ACC read-COMP want-NMLZ
'John's wanting to read books'

Whereas the internal syntax of the GP is much like that of a VP, its external syntax is more like that of an NP. The GP can appear in the canonical NP positions such as subject or object as in (3)a or as a postpositional object in (3)b (cf. Yoon 1989).

(3) a. [ai-ka chayk-ul ilk-um]-i nollapta
child-NOM book-ACC read-NMLZ-NOM surprising
'That child's reading a book is surprising'
b. [John-i enehak-ul kongpwuha-m]-**eytayhay** mollassta
John-NOM linguistics-ACC study-NMLZ-about not.know
'(We) didn't know about John's studying linguistics.'

One thing worth pointing out here is that the GP does not have the full distribution of NPs, either. As demonstrated in (4), the GP cannot serve as the head of a relative clause, implying that the external syntax of the GP is somewhat different from that of a canonical NP.

(4) *John-un [[salam-tul-i __ molulila-ko sayngkakha-n] [Mary-ka ilccik
John-TOP people-PL not.know-COMP think-REL Mary-NOM early
ttenass-um]]-ul alassta.
left-NMLZ knew
'*John knew [Mary's leaving early] that he thought that people wouldn't notice'.

These mixed and complicated properties of GP have provided a challenge to syntactic analyses with a strict version of X-bar theory, in particular, with respect to the theory-neutral desiderata such as endocentricity and lexicalism. This paper shows we can provide an effective and systematic way of capturing these mixed and complicated properties without abandoning these desiderata once we adopt the mechanism of multiple classification of category types with systematic inheritance. The grammar we developed as an application of the constraint-based grammar of HPSG to Korean is called Korean Phrase Structure Grammar (KPSG). We have checked the feasibility of the grammar by implementing it into the LKB (Linguistic Knowledge Building) system.

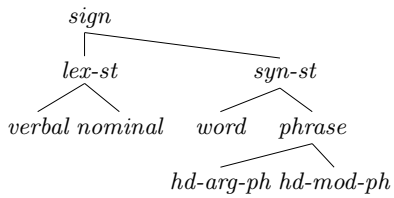
2 Korean Phrase Structure Grammar

2.1 Basic Picture

The Korean Phrase Structure Grammar, aiming to develop an open source grammar of Korean, consists of grammar rules, inflection rules, lexical rules, type

definitions, and lexicon.[4] As in HPSG (Sag et al. 2003), the grammar adopts the mechanism of type hierarchy in which every linguistic sign is typed with appropriate constraints and hierarchically organized. All the linguistic information is thus represented in terms of *sign*. The type *sign* is classified into subtypes as represented in a simplified hierarchy in (5):

(5)



The elements in *lex-st* type, forming the basic components of the lexicon, are built up from lexical processes such as lexical rules and type definitions. Parts of these elements will be realized as *word* to function as syntactic elements. Phrases projected from *word* form basic Korean well-formed phrases such as *hd-arg-ph* (*head-argument-ph*) and *hd-mod-ph* (*head-modifier-ph*). All the syntactic rules in KPSG are either unary or binary. Different from English (and from the Japanese grammar of Siegel and Bender 2002), we assume that Korean adopts the following simplified phrasal well-formed conditions:[5]

(6) Korean X′ Syntax:

```
a. hd-arg-ph:
[ ]->#1,H[ARG-ST <...#1...>]

b. hd-mod-ph:
 [ ] -> [MOD #1], H[#1]

c. hd-filler-ph:
   [ ] -> #1, H[GAP <#1>]

d. hd-word-ph:
 [word] -> [word], H
```

(6)a means that when a head combines with one of its arguments, the resulting phrase is a well-formed phrase. (6)b allows a head to combine with a phrase that modifies it. (6)c is a constraint for a head to form a phrase (with a missing gap) with a filler. (6)d basically generates a word level syntactic element by the combination of a head and a word. This condition in (6)d, not found in languages like English, forms various types of complex predicates found in the language (cf. Kim and Yang 2003).

The type *hd-arg-ph* can easily license basic sentence types such as the following:

[4] The space does not allow us to explicate the morphological and semantic system of the KPSG. As for morphology, we integrated MACH (Morphological Analyzer for Contemporary Hangul) developed by Shim and Yang (2002). This system segments words into sequences of morphemes with POS tags and morphological information. As for semantics, we adopted the Minimal Recursion Semantics developed by Copestake et al. (2001).

[5] Of course, further constraints need to be specified on these phrases. For example, the phrase *hd-word-ph* has additional constraints on the argument structure of the head. See Kim and Yang (2004) for details.

(7) a. [[pi-ka [o-ass-ta]]]. 'It rained.'
rain-NOM come-PST-DECL
b. [John-i [Mary-eykey [chayk-ul [cwu-ess-ta]]]].
John-NOM Mary-DAT book-ACC give-PST-DECL
'John gave Mary a book.'

Since the phrase condition (in particular *hd-arg-ph*) allows a head (lexical or phrasal) to combine with one of its syntactic arguments, KPSG generates only binary structures as represented by the brackets.

One welcoming, desirable consequence of this binary approach concerns the sentence internal scrambling, one of the most complicated facts in the SOV type of language. For example, the sentence in (8) with five syntactic elements can induce 24 (4!) different scrambling possibilities, with the head verb fixed in the final position.

(8) [mayil] [John-i] [haksayng-tul-eykey] [yenge-lul] [kaluchi-ess-ta]
Everyday John-NOM students-PL-DAT English-ACC teach-PST-DECL
'John taught English to the students everyday.

A most effective grammar would no doubt be the one that can capture all such scrambling possibilities within minimal processing load. In KPSG, the conditions on *hd-arg-ph* and *hd-mod-ph* allow us to generate all the word ordering possibilities for cases like (8). The following is one of the three encoded rules for the *hd-arg-ph* in the LKB system:[6]

```
head-arg-rule-1 := hd-arg-ph &
[SYN.ARG-ST #2,
ARGS<#1, syn-str&[SYN.ARG-ST [FIRST #1,
                              REST #2]]>].
```

Such a rule basically licenses a head to combine with only one of its complement(s), resulting in a binary structure. This kind of simple X′ syntax is enough to capture the intriguing sentence internal scrambling without positing various movement operations.

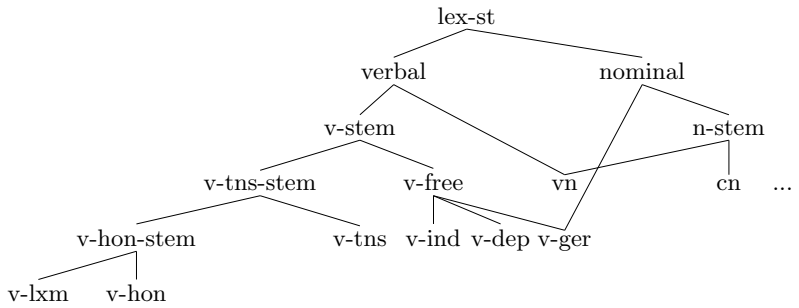
2.2 The Structure of Lexicon and Forming Gerundive Verbs

The starting point of structuring the lexicon in the KPSG is parts of speech in the language. Like the traditional literature, the KPSG assumes *verbal*, *nominal*, *adverbial*, and *adnominal* as the language's basic categories. These are further subclassified into subtypes. For example, the type *verbal* is taken to have the rather simplified hierarchy given in (9):[7]

[6] Since the LKB does not allow a set operation, the LKB implementation requires to write three *head-arg-rule*s, depending on which argument in the ARG-ST combines with the head:

[7] The lexicon only provides elements for *v-lxm* and all the other type elements are generated from the morpho-syntactic, semantic constraints. See Kim and Yang 2003.

(9)



The key point of capturing the mixed properties of GPs comes form the multiple inheritance mechanism in which the type of gerundive verbs (*v-ger*) is declared to be the subtype of both *v-free* and *nominal* as represented here (cf. Kim 1998 and Kim and Yang 2003). One main difference from traditional grammar is the assignment of the HEAD feature POS: It assigns [POS *verb*] not to *verbal* but to the type *v-stem*, a subtype of *verbal*, and [POS *noun*] not to *nominal* but to the type *n-stem*, a subtype of *nominal*: there arises thus no conflict in the inheritance of the POS value on *v-ger*.[8]

The KPSG thus differs from standard approaches in that it introduces three related features POS, VERBAL, and NOMINAL. This makes the grammar flexible enough to refer to each of these features when necessary. For example, all [VERBAL +] objects will have non-empty ARG-ST values, only the [POS *noun*] element will serve as the head of a relative clause, all [NOMINAL +] elements could serve as the host of the genitive marker, etc. This system then makes it unnecessary to introduce an additional part of speech such as *gerundive* as Malouf (1998) did for English.

In forming a gerundive verb, the KPSG thus starts from a transitive verb lexeme *ilk-* 'read' given in (10)a and and then forms a *v-tns-stem* with the attachment of the past tense suffix *-ess*. This *v-tns-stem* then combines with the nominalizer suffix *um*, forming the type *v-ger*. In this process, verbal properties (e.g. POS and VERBAL value) are inherited from *v-lxm* and *v-tns-stem*, whereas their nominal properties (e.g. NOMINAL) are incurred from its supertype *nominal*. The gerundive verb *ilk-ess-um* 'read-PST-NMLZ' will thus have at least the lexical information given in (10)b:

(10)

a.
$$\begin{bmatrix} \textit{v-lxm} \\ \text{ORTH ilk-} \\ \text{SYN} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{POS } \textit{verb} \\ \text{VERBAL } + \\ \text{NOMINAL } - \end{bmatrix} \\ \text{ARG-ST } \langle \boxed{1}\text{NP}, \boxed{2}\text{NP} \rangle \end{bmatrix} \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{v-ger} \\ \text{ORTH [ilk + ess] + um} \\ \text{SYN} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{POS } \textit{verb} \\ \text{VERBAL } + \\ \text{NOMINAL } + \end{bmatrix} \\ \text{ARG-ST } \langle \boxed{1}\text{NP}, \boxed{2}\text{NP} \rangle \end{bmatrix} \end{bmatrix}$$

[8] One great advantage of this system is that it can also successfully capture the mixed properties of verbal noun. Notice here that the verbal noun type *vn* is slightly different: In the present system, this type will have [NOMINAL +, VERBAL +], but [POS *noun*]. These feature specifications will predict the facts that verbal nouns are in part nominal and in part verbal though morphologically they are more like nouns unlike gerundive verbs.

Such a cross-classification of the type *v-ger*, allowing multiple inheritance, is also reflected in the feature descriptions in the LKB. The following represents a sample source code:

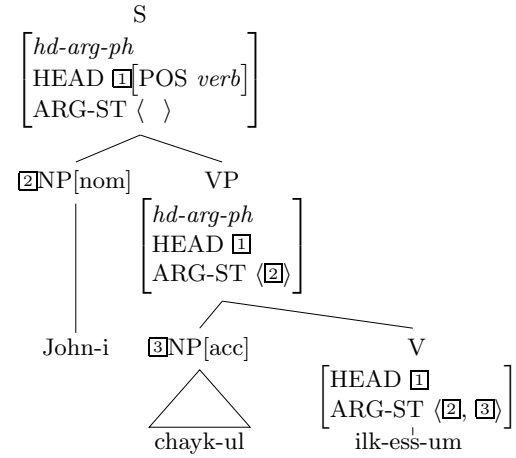
```
v-ger := v-free & n-stem &
 [ SYN #syn & [HEAD.MOD <>],
   SEM #sem,
   ARGS <v-tns-stem & [SYN #syn, SEM #sem]>].
```

As observed here, as a subtype of *v-free* and *n-stem*, the type *v-ger* thus inherits the constraints from its supertypes. Being a subtype of *v-free*, it inherits verbal properties from the type *v-free*, selecting arguments and assigning case values to them. Since it is a subtype of *nominal*, *v-ger* would undergo nominal suffixation processes such as case attachment. In addition, the grammar introduces the binary-valued features VERBAL on the type *verbal* and NOMINAL on the type *nominal*, which plays crucial roles in capturing mixed properties as well as various generalizations in the Korean grammar.

2.3 Projecting Gerundive Verbs into Syntax

Once we build up a gerundive verb with rich information that could be relevant in syntax, we then now need to look at how it is projected in syntax. Within the KPSG, the lexical entry in (10)b will be projected into structure like (11):

(11)



As noted, the gerundive verb *ilk-ess-um* 'read-PST-NMLZ' inherits all the other properties such as argument structure value from the verb lexeme *ilk–* 'read'. This explains why the gerund selects a nominative subject, can be modified by an adverb, allows sentential adverbials within the clause, combines with the sentential negative marker, occurs with an auxiliary verb, and the like. Because the gerundive verb selects the same argument(s) as the verb lexeme it is derived from, the phrase formed by the gerundive and one of its complements will be a well-formed *hd-arg-ph*. This resulting VP combines with the remaining nominative argument, the subject. This is what the top node S in (11) represents, reflecting the internal properties of GPs.

Prevalent are morphological and syntactic phenomena supporting this line of approach. Support for the treatment of the gerundive predicate induces a projection of [POS verb] comes from (a) the presence of a tense and an agreement suffix and (b) the possibility of heading an independent sentence as in (12):

(12) sensayngnim-i chayk-ul **ilk-usi-ess-um**.
teacher-NOM book-ACC read-HON-PAST-NMLZ
'The teacher has read the book.'

The analysis also provides a simple way of capturing relativization and extraction phenomena. Though GPs externally act like noun phrases (because of the NOMINAL feature), they do not serve as the head of a restrictive relative clause as repeated here in (13).

(13) *John-un [[salam-tul-i __ molulila-ko sayngkakha-n] [Mary-ka ilccik
John-TOP people-PL not.know-COMP think-REL Mary-NOM early
ttenass-um]]-ul alassta.
left-NMLZ knew
'*John knew [Mary's leaving early] that he thought that people wouldn't notice'.

As hinted, the only thing the grammar needs to specify is the constraint that a relative clause modifies a phrase projected from the feature [POS *noun*] not the one with [NOMINAL +] as represented in part of the relative clause modifying rule in the LKB description:

```
head-rel-mod-rule := binary &
 [ SYN ...
   ARGS < phrase & [ SYN [ HEAD.MOD < #1 & [ SYN.HEAD.POS noun,
                                             SEM.INDEX #2 ] >,
                           ... ] ],
          syn-st & #1 & [ SYN.VAL [ ARG-ST #argst,
                                    GAP <! !> ] ] > ].
```

As indicated in the first element of the ARGS value, the relative clause can modify an element whose head bears [POS *noun*] value here. This will then correctly block any relative clause from modifying a gerundive verb head though it behaves like a nominal element.

It is possible to extract an element from GPs, indicating that the GP behaves more like Ss and less like NPs in terms of the external syntax:

(14) **ku chayk-un** na-nun [John-i __ ilkess-um]-ul mitnunta
that book-TOP I-TOP John-NOM __ read-NMLZ-ACC believe
'That book, I believe John read.'

This is unexpected when considering the external status of the GP to be a pure NP. The KPSG, allowing extraction from a sentence level element ([POS *verb*]), takes the GP to be just like a sentence with the positive NOMINAL feature. Nothing thus blocks this extraction.

3 Some Further Consequences

Unlike a clitic or a phrasal approach that treats the nominalizer as a phrasal element or a clitic (cf. Yoon 1989), the present lexical analysis takes it to be a pure suffix, reflecting its morphological properties (cf. Kim 1998). However, examples like (15), in which the nominalizers seem to scope over the two coordinate sentences, seem to devalue such a lexical approach.[9]

(15) [[A-ka sakwa-lul mek-ess]-ko [B-ka maykcwu-lul masi-ess]-m]
A-NOM apple-ACC eat-PST-CONJ B-NOM beer-ACC drink-PST-NMLZ
'A ate apples and B drank beer.'

In our analysis, this is also predictable. Since the second GP is also a type of sentence, cases like (15) are coordination of two sentences. There is no category mismatch in our analysis: the second conjunct is different from the first one only in its FORM value.

The present analysis also provides a simple way of dealing with cases where the subject is realized as genitive:

(16) ?[John-uy chayk-ul ilk-um]
John-GEN that book-ACC read-NMLZ-NOM
'John's reading books '

As we can observe, the case value on the subject of the gerundive verb is different from that on the subject of the English gerund. Korean allows only nominative or genitive. The example in (16) differs from the nominative subject GP only in the way that the VP combines with a genitive specifier to form a possessive noun phrase. This case analysis is predicted within a rule-based case theory (cf. Kim 2004) in which the subject of a verbal element with the feature [VERBAL +] gets NOM whereas the specifier of [NOMINAL +] gets GEN. This system then would allow the following four possibilities:

(17) a. [John-i [chayk-ul [ilk-um]]] (ilk-um: [VERBAL +])
b. [John-uy [chayk-ul ilk-um]] (ilk-um: [VERBAL +, NOMINAL +])
c. [John-uy [chayk-uy ilk-um]] (ilk-um: [NOMINAL +])
d. ?[John-i [chayk-uy ilk-um]] (ilk-um: [NOMINAL +, VERBAL +])

As noted in the bracket with what feature is relevant for the case assignment in a sense, we could observe that the enriched information on the gerundive verb *ilk-um* can license GEN to its arguments due to the feature NOMINAL.

[9] Treating *um* as an independent lexical or phrasal element brings a serious drawback since this means an element like *v-tns-stem* to appear in syntax. That is, this would allow an element like *mek-ess* 'eat-PST' to freely appear in syntax contrary to the fact that only *v-free* elements can appear in Korean syntax.

4 Testing the Feasibility of the Analysis

The grammar we have built within the typed-feature structure system here, eventually aiming at working with real-world data, has been first implemented into the LKB.

In testing its performance and feasibility, we used the SERI Test Suites '97 after the successful parsing of the self-designed 250 sentences. The SERI Test Suites (Sung and Jang 1997), carefully designed to evaluate the performance of Korean syntactic parsers, consists of total 472 sentences (292 test sentences representing the core phenomena of the language and 180 sentences representing different types of predicate). In terms of lexical entries, it has total 440 lexemes (269 nouns, 125 predicates, 35 adverbs, and 11 determiners) and total 1937 word occurrences.

The present system correctly generated all the lexemes in the test suites and inflected words. In terms of parsing sentences, the grammar (syntactically and semantically) parsed 423 sentences out of total 472.[10] Unfortunately, the test suites do not include any gerundive phrases, which seems to be a serious drawback when considering their frequent usages in real life; we included 50 test sentences for gerundive constructions adopted from the literature on Korean gerundive phrases (e.g. Kim 1998, Lapointe 1993, Yoon 1989) and the Sejong Project Basic Corpus. The present system successfully parsed all these sentences. Another promising indication of the test is that its mean parse (average number of parsed trees) for the total 458 (423 plus 35) sentences is 1.97, controlling spurious ambiguity at a desired level.

As noted here, the test results provide clear evidence that the KPSG, built upon typed feature structure system, offers high performance and can be extended to large scale of data. Since the test suites here include most of the main issues arising in analyzing major Korean constructions, we believe that further tests for designated corpora will surely achieve nearly the same result of high performance too.

5 Conclusion

This paper has shown that it is possible to analyze English and Korean GPs in a way that maintains the lexical integrity principle, captures endocentricity, and avoids empty categories. This has been achieved through the development of KPSG, an extension of HPSG, that could reflect the language particular properties. HPSG is a sign-based grammar in which the basic unit of linguistic object *sign* is a structured complexes of linguistic information, represented by *typed feature structure*. The feasibility of the grammar developed has been checked with its implementation into the LKB system. The result of an existing test suite and self-constructed experimental data is quite promising though there still remains an issue of testing it with a large scale of corpora.

[10] Failed 49 sentences are related to the grammar that the current system has not yet written. For example, the SERI Test Suites include examples representing phenomena such as left dislocations of the subject, gapping, and non-subject *pro* drops. It is believed that once we have a finer-grained grammar for these phenomena, the KPSG will be able to parse these remaining sentences.

Acknowledgements

We are grateful to the three anonymous reviewers for comments and suggestions. We also thank Sae-Youn Cho, Jae-Woong Choe, Chan Chung, and Yongkyoon No for helpful comments. The first author also acknowledges the financial support by the Brian Korea 21 project in the year of 2004.

References

Chung, Chan, Jong-Bok Kim, Byung-Soo Park, and Peter Sells. 2001. Mixed Categories and Multiple Inheritance Hierarchies in English and Korean Gerundive Phrases. *Language Research* 37.4: 763–797.

Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars.* CSLI Publications.

Kim, Jong-Bok. 1998. Interface between Morphology and Syntax: A Constraint-Based and Lexicalist Approach. *Language and Information* 2: 177-233.

Kim, Jong-Bok and Jaehyung Yang. 2003. Korean Phrase Structure Grammar and Implementing it into the LKB System (In Korean). *Korean Linguistics* 21: 1–41.

Kim, Jong-Bok and Jaehyung Yang. 2004. Projections from Morphology to Syntax in the Korean Resource Grammar: Implementing Typed Feature Structures. In *Lecture Notes in Computer Science* Vol 2945, pp 13-24, Springer-Verlag.

Malouf, Robert. 1998. *Mixed Categories in the Hierarchical Lexicon.* Stanford: CSLI Publications.

Sag, Ivan, Tom Wasow, and Emily Bender. 2003. *Syntactic Theory: A Formal Approach.* Stanford: CSLI Publications.

Shim, Kwangseob and Yang Jaehyung. 2002. MACH: A Supersonic Korean Morphological Analyzer. In *Proceedings of Coling-2002 International Conference*, pp.939-45, Taipei.

Siegel, Melanie and Emily M. Bender. 2002. Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization.* Coling 2002 Post-Conference Workshop. Taipei, Taiwan.

Sung, Won-Kyung and Myung-Gil Jang. 1997. SERI Test Suites '95. In *Proceedings of the Conference on Hanguel and Korean Language Information Processing.*

Lapointe, Steven. 1993. Dual Lexical Categories and the Syntax of Mixed Category Phrases. In A. Kathol & M. Bernstein (eds.), *Proceedings of ESCOL*, 199-210.

Pullum, Geoffrey. 1991. English Nominal Gerund Phrases as Noun Phrases with Verb-Phrase Heads. *Linguistics* 29: 763–799.

Yoon, James. 1989. *A Restrictive Theory of Morphosyntactic Interaction and Its Consequences.* Ph.D. Dissertation. University of Illiniois, Urbana-Champaign.