A Computational Implementation of Internally Headed Relative Clause Constructions

Jong-Bok Kim[1], Peter Sells[2], and Jaehyung Yang[3]

[1] School of English, Kyung Hee University, Seoul, Korea 130-701
jongbok@khu.ac.kr
[2] Dept. of Linguistics, Stanford University, USA
sells@stanford.edu
[3] School of Computer Engineering, Kangnam University, Kyunggi, 446-702, Korea
jhyang@kangnam.ac.kr

Abstract. The so-called Internally Headed Relative Clause (IHRC) construction found in the head-final languages Korean and Japanese has received little attention from computational perspectives even though it is frequently found in both text and speech. This is partly because there have been no grammars precise enough to allow deep processing of the construction's syntactic and semantic properties. This paper shows that the typed feature structure grammar HPSG (together with the semantic representations of Minimal Recursion Semantics) offers a computationally feasible and useful way of deep-parsing the construction in question.

1 Introduction

In terms of truth conditions, there is no clear difference between a (Korean) IHRC (internally head relative clause) like (1)a and and EHRC (externally headed relative clause) like (1)b.[1]

(1) a. Tom-un [sakwa-ka cayngpan-wi-ey iss-nun kes]-ul mekessta
Tom-TOP apple-NOM tray-TOP-LOC exist-PNE KES-ACC ate
'Tom ate an apple, which was on the tray.'
b. Tom-un [__ cayngpan-wi-ey iss-nun sakwa]-ul mekessta.
Tom-TOP tray-TOP-LOC exist-PNE apple-ACC ate
'Tom ate an apple that was on the tray.'

Both describe an event in which an apple is on the tray, and Tom's eating it.[2]

Yet, there exist several intriguing differences between the two constructions. One crucial difference between the IHRC and EHRC comes from the fact that

[1] We thank anonymous reviewers for their helpful comments and suggestions. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2005-042-A00056).

[2] The following is the abbreviations used for glosses and feature attributes in this paper: ACC (ACCUSATIVE), COMP (COMPLEMENTIZER), LOC (LOCATIVE), NOM (NOMINATIVE), PNE (PRENOMINAL), TOP (TOPIC), etc.

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 24–31, 2006.

the semantic object of *mekessta* 'ate' in the IHRC example (1)a is the NP *sakwa* 'apple' buried inside the embedded clause. It is thus the subject of the embedded clause that serves as the semantic argument of the main predicate ([1], [2]).

In the analysis of such IHRCs, the central questions thus involve (a) the key syntactic properties, (b) the association of the internal head of the IHRC clause with the matrix predicate so that the head can function as its semantic argument, and (c) the differences between the IHRC and EHRC. This paper provides a constraint-based analysis within the framework of HPSG (Head-driven Phrase Structure Grammar) and implements it in the existing HPSG grammar for Korean using the LKB (Linguistic Building Knowledge) system to check the computational feasibility of the proposed analysis.[3]

2 Implementing an Analysis

2.1 Syntactic Aspects of the IHRC

One main morphological property of the IHRC construction is shown in (2)b: the embedded clausal predicate should be in the adnominal present form of *(n)un*, followed by the so-called bound noun *kes*. This clearly contrasts with the EHRC example (2)a, in which the predicate can have any of the three different markers of tense information:[4]

(2) a. Tom-i __ $_i$ ilk-nun/un/ul chayk$_i$
Tom-NOM read-PRES.PNE/PST.PNE/FUT.PNE book
'the book that Tom reads/read/will read'

b. Tom-un [sakwa-ka cayngpan-wi-ey **iss-nun/*ul** kes]-ul mekessta
Tom-TOP apple-NOM tray-TOP-LOC exist-PNE KES-ACC ate
'Tom ate an apple, which was (lit. 'is') on the tray.'

In traditional Korean grammar, *kes* in the IHRC is called a 'dependent noun', in that it always requires either a modifying determiner or clause, even in a non-IHRC usage:

(3) a.*(i/ku/ce) kes '*(this/that) thing'
b.*(nay-ka mek-un) kes 'the thing (*that I ate)'

This close syntactic relation between the clause and the noun *kes* can also be found in the fact that unlike canonical nouns, it must combine with a preceding adnominal clause:

(4) Na-nun *(kangto-ka unhayng-eyse nao-nun) kes-ul capassta
I-TOP robber-NOM bank-from come-out-PNE KES-ACC caught
'I arrested the robber who was coming out of the bank.'

[3] The LKB, freely available with open source (`http://lingo.stanford.edu`), is a grammar and lexicon development environment for use with constraint-based linguistic formalisms such as HPSG. cf. [3].

[4] These three prenominal markers in the EHRC extend their meanings to denote aspects when combined with (preceding) tense suffixes.

These examples show that the pronoun *kes* selects an adnominal clause as its complement, and that the IHRC requires a specific inflected form of its predicate.

Then, what is the relationship between the whole IHRC clause including *kes* and the matrix verb? To relate the matrix verb with this construction with an 'internal semantic head', it was assumed in transformational grammar that it was necessary to introduce an empty category such as *pro* to the right of the adnominal clause, on the assumption that the IHRC is an adjunct clause (Jhang 1991). However, there is ample evidence showing that the clause is a direct syntactic nominal complement of the matrix predicate. One strong argument against an adjunct treatment centers on the passivization of the IHRC clause. As shown in (5), an object IHRC clause can be promoted to the subject of the sentence.

(5) [Tom-i talli-nun kes]-i Mary-eyeuyhayse caphiessta
Tom-NOM run-PNE KES-NOM Mary-by be.caught
'Tom, who was running, was caught by Mary.'

Another fact concerning the status of the IHRC comes from stacking: whereas more than one EHRC clause can be stacked, only one IHRC clause is possible:

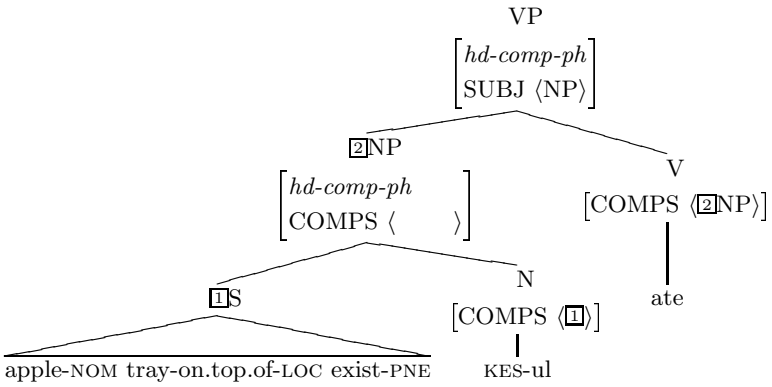
(6) a.*kyongchal-i [**kangto-ka unhayng-eyse nao-nun**]
police-NOM [robber-NOM bank-from come.out-PNE]
[ton-ul hwumchi-n] **kes**-ul chephohayssta
money-ACC steal-PNE KES-ACC arrested
'(int.) The police arrested a thief coming out of the bank, stealing money.'

b. kyongchal-i [__ **unhayng-eyse nao-nun**]
police-NOM [bank-from come.out-PNE]
[ton-ul hwumchi-n] **kangto**-lul chephohayssta
money-ACC steal-PNE robber-ACC arrested
'(int.) The police arrested a thief coming out of the bank, stealing money.'

This contrast implies that the adnominal clause which is the IHRC has the canonical properties of a complement clause.

Based on these observations, we assume the structure (7) for the internal and external structure of the IHRC in (1)a:

(7)



As represented in the tree, *kes* combines with its complement clause, forming a *hd-comp-ph* (*head-complement-ph*). This resulting NP also functions as the complement of the matrix verb *ate*.

2.2 Semantic Aspects of the IHRC and Related Constructions

One thing to note is that IHRCs are syntactically very similar to DPCs (direct perception constructions). IHRCs and DPCs both function as the syntactic argument of a matrix predicate. However, in the IHRC (8)a, the internal argument *John* within the embedded clause functions as the semantic argument of 'caught'. Meanwhile, in (8)b it is the whole embedded clausal complement that functions as its semantic argument:

(8) a. Mary-nun [John-i talli-nun kes]-ul **capassta**.
Mary-TOP John-NOM run-PNE KES-ACC caught
'Mary caught John who was running.'

b. Mary-nun [John-i talli-nun kes]-ul **poassta**.
Mary-TOP John-NOM run-PNE KES-ACC saw
'Mary saw John running.'

The only difference between (8)a and (8)b is the matrix predicate, which correlates with the meaning difference. When the matrix predicate is an action verb such as *capta* 'catch', *chepohata* 'arrest', or *mekta* 'eat' as in (8)a, we obtain an entity reading for the clausal complement. But as in (8)b we will have only an event reading when the matrix predicate is a type of perception verb such as *po-ta* 'see', *al-ta* 'know', and *kiekhata* 'remember'.

The key point in our analysis is thus that the interpretation of *kes* is dependent upon the type of matrix predicate. Hence the lexical entries in our grammar involve not only syntax but also semantics. For example, the verb *cap-ta* 'catch' in (9) lexically requires its object to refer to a *ref-ind* (referential-index) whereas the verb *po-ta* 'see' in (10) selects an object complement whose index is *indiv-ind* (individual index) whose subtypes include *ref-ind* and *event-ind*, indicating that its object can be either a referential individual or an event.[5]

[5] The meaning representations adopted here involve Minimal Recursion Semantics (MRS), developed by [4]. This is a framework of computational semantics designed to enable semantic composition using only the unification of type feature structures. The value of the attribute SEM(ANTICS) we used here represents simplified MRS, though it originally includes HOOK, RELS, and HCONS. The feature HOOK represents externally visible attributes of the atomic predications in RELS (RELATIONS). The value of LTOP is the local top handle, the handle of the relations with the widest scope within the constituent. The value of XARG is linked to the external argument of the predicate. See [4] and [5] for the exact function(s) of each attribute. We suppress irrelevant features.

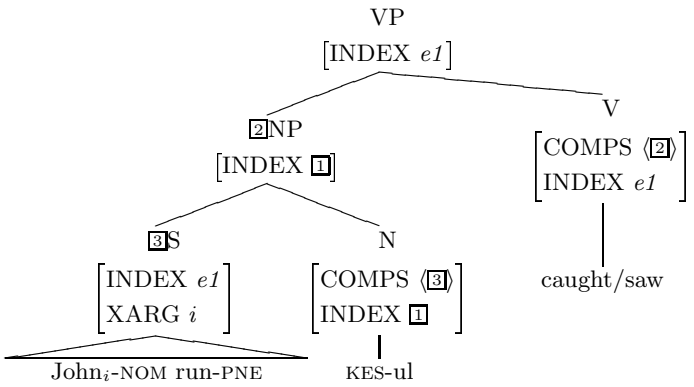
(9)

a.
$$\begin{bmatrix} \langle \text{cap-ta 'catch'} \rangle \\ \text{SYN} \mid \text{VAL} \begin{bmatrix} \text{SUBJ } \langle \text{NP}_i \rangle \\ \text{COMPS } \langle \text{NP}_j \rangle \end{bmatrix} \\ \text{SEM} \mid \text{RELS} \left\langle \begin{bmatrix} \text{PRED } \textit{catch\_v\_rel} \\ \text{ARG0 } \textit{e1} \\ \text{ARG1 } _i[\textit{ref-ind}] \\ \text{ARG2 } _j[\textit{ref-ind}] \end{bmatrix} \right\rangle \end{bmatrix}$$

b.
$$\begin{bmatrix} \langle \text{po-ta 'see'} \rangle \\ \text{SYN} \mid \text{VAL} \begin{bmatrix} \text{SUBJ } \langle \text{NP}_i \rangle \\ \text{COMPS } \langle \text{NP}_j \rangle \end{bmatrix} \\ \text{SEM} \mid \text{RELS} \left\langle \begin{bmatrix} \text{PRED } \textit{see\_v\_rel} \\ \text{ARG0 } \textit{e1} \\ \text{ARG1 } _i[\textit{ref-ind}] \\ \text{ARG2 } _j[\textit{ind-ind}] \end{bmatrix} \right\rangle \end{bmatrix}$$

These lexical entries will then project an identical syntactic structure for (8)a and (8)b, represented together here in (10):

(10)



As represented in the structure, in both constructions *kes* selects an adnominal S as its complement and forms a *hd-comp-ph* with it. The resulting NP serves the complement of the main verb *caught* or *saw*. However, semantically, due to the lexical entries in (9), the object of *caught* is linked to the external argument (XARG) *robber* whereas that of *saw* in (9)b is linked to the event denoted by the S.[6] The type of predicate thus determines whether the INDEX value of *kes* will be identified with that of the S or that of its XARG, as presented in the lexical entries:

(11)

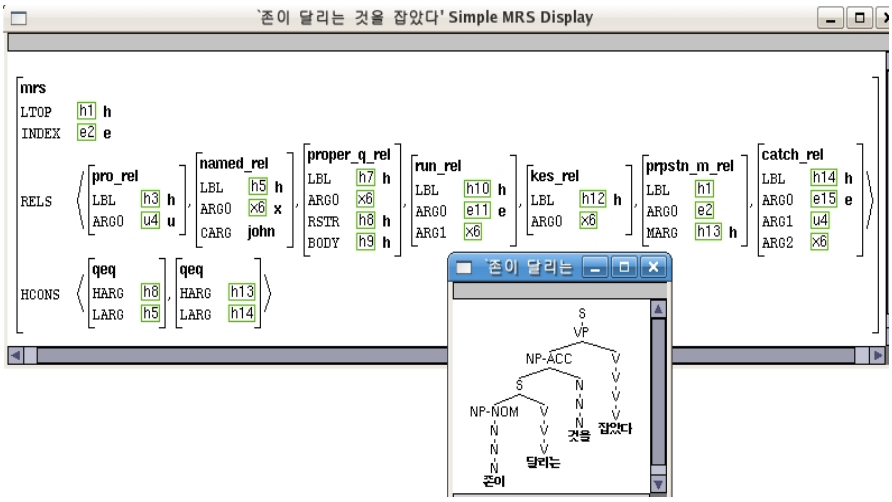
a.
$$\begin{bmatrix} \langle \text{kes} \rangle \\ \text{SYN} \begin{bmatrix} \text{HEAD} \mid \text{POS } \textit{noun} \\ \text{VAL} \mid \text{COMPS } \langle \text{S}[\text{INDEX } \textit{e1}] \rangle \end{bmatrix} \\ \text{SEM} \mid \text{HOOK} \mid \text{INDEX } \textit{e1} \end{bmatrix}$$

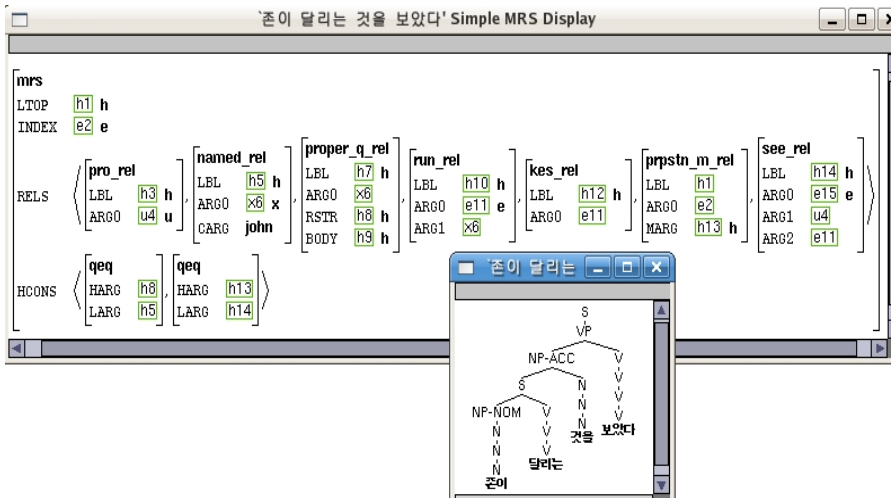
b.
$$\begin{bmatrix} \langle \text{kes} \rangle \\ \text{SYN} \begin{bmatrix} \text{HEAD} \mid \text{POS } \textit{noun} \\ \text{VAL} \mid \text{COMPS } \langle \text{S} \begin{bmatrix} \text{XARG } i \end{bmatrix} \rangle \end{bmatrix} \\ \text{SEM} \mid \text{HOOK} \mid \text{INDEX } i \end{bmatrix}$$

[6] The feature XARG refers to the external argument in control constructions like *John tries to run*. The XARG of *run* is thus identified the matrix subject *John*. See [5] for details.

This grammar in which lexical information interacts with the other syntactic components ensures that the perception verb *saw* combines with an NP projected from (11)a whereas the action verb *caught* with an NP projected from (11)b. Otherwise, the resulting structure will not satisfy the selectional restrictions of the predicates.

Incorporating this into our Korean grammar,[7] we implemented this analysis in the LKB and obtained the following two parsed trees and MRSs for the two examples:





[7] The current Korean Resource Grammar has 394 type definitions, 36 grammar rules, 77 inflectional rules, 1100 lexical entries, and 2100 test-suite sentences, and aims to expand its coverage on real-life data.

Leaving aside the irrelevant parts, we can see that the two have the identical syntactic structures but different semantics. In the former, the ARG0 value of *kes* is identified with the *named_rel* (for 'John') but in the latter it is identified with *run_rel*.

The analysis thus provides a clean account of the complementary distribution of the IHRC and the DPC. That is, according to our analysis, we obtain an entity reading when the index value of *kes* is identified with that of the external argument. Meanwhile, we have an event reading when the index value is structure-shared with that of the adnominal S. This analysis thus correctly predicts that there exist no cases where the two readings are available simultaneously.

One of the welcome predictions that this analysis brings is that the canonical antecedent of the pronoun *kes* is the external argument:

(12) [haksayng-i aktang-ul cha-nun kes-ul] capassta
student-NOM rascal-ACC kick-PNE KES-ACC caught
'(I) caught a student, who was then kicking a rascal.'

Even though one can catch either a student or a rascal, the semantic object of the verb 'catch' is not the object but the external argument *haksayng* (attested by our implementation but not included here because of limits on space).

3 Discussion and Conclusion

The analysis we have presented so far, part of the typed-feature structure grammar HPSG for Korean aiming at working with real-world data, has been implemented into LKB (Linguistic Knowledge Building System) to test its performance and feasibility.

We first inspected the Sejong Treebank Corpus (33,953 sentences) and identified 4,610 sentences with [S[FORM *nun*] + *kes*]. Of these, we inspected the 518 ACC marked examples, but found only 3 IHRC examples. Another 154 examples used *kes* in a cleft construction, and 361 as direct perception examples. Among these, we selected canonical types of the IHRC constructions to check if the grammar can parse them both in terms of syntax and semantics. As we have shown in section 2.2, the grammar is quite successful in picking up the appropriate semantic head from the IHRC. Of course, issues remain of extending the coverage of our grammar to parse more real-life data and further identifying other constructional types of *kes*, such as cleft usages.

Any grammar, aiming for real world application, needs to provide a correct syntax from which we can build semantic representations in compositional ways. In addition, these semantic representations must be rich enough to capture compositional as well as constructional meanings. In this respect, the analysis we have sketched here seems to be promising in the sense that it provides appropriate semantic representations for the IHRC and DPC in a compositional way, suitable for applications requiring deep natural language understanding.

References

1. Kim, Y.B.: Relevance in internally headed relative clauses in korean. Lingua **112** (2002) 541–559
2. Chung, C., Kim, J.B.: Differences between externally and internally headed relative clause constructions. In: Proceedings of HPSG 2002, CSLI Publications (2003) 3–25
3. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford (2002)
4. Copestake, A., Flickenger, D., Sag, I., Pollard, C.: Minimal recursion semantics: An introduction. Manuscript (2003)
5. Bender, E.M., Flickinger, D.P., Oepen, S.: The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In Carroll, J., Oostdijk, N., Sutcliffe, R., eds.: Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics, Taipei, Taiwan (2002) 8–14