Yūji Matsumoto
Richard Sproat
Kam-Fai Wong
Min Zhang (Eds.)

Computer Processing of Oriental Languages

Beyond the Orient: The Research Challenges Ahead

LNAI 4285

21st International Conference, ICCPOL 2006
Singapore, December 2006
Proceedings

Springer

Deep Processing of Korean Floating Quantifier Constructions

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. The so-called floating quantifier constructions in languages like Korean display intriguing properties whose successful processing can prove the robustness of a parsing system.[1] This paper shows that a constraint-based analysis, in particular couched upon the framework of HPSG, can offer us an efficient way of parsing these constructions together with proper semantic representations. It also shows how the analysis has been successfully implemented in the LKB (Linguistic Knowledge Building) system.

1 Processing Difficulties

One of the most salient features in languages like Korean is the complex behavior of numeral classifiers (Num-CL) linked to an NP they classify. Among several types of Num-CL constructions, the most complicated type includes the one where the Num-CL floats away from its antecedent:[2]

(1) pemin-i cengmal sey myeng-i/*-ul te iss-ta
criminal-NOM really three CL-NOM/ACC more exist-DECL
'There are three more criminals.'

There also exist constraints on which arguments can 'launch' floating quantifiers (FQ). Literature (cf. [1]) has proposed that the antecedent of the FQ needs to have the identical case marking as in (1). However, issues become more complicated with raising and causative constructions where the two do not agree in the case value:

(2) a. haksayng-tul-**ul** sey myeng-**i/ul** chencay-i-lako mit-ess-ta.
student-PL-ACC three-CL-NOM/*ACC genius-COP-COMP believed
'(We) believed three students to be genius.'

[1] We thank three anonymous reviewers for the constructive 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 are the abbreviations used for the glosses and attributes in this paper: CL (CLASSIFIER), CONJ (CONJUNCTION), COP (COPULA), COMP (COMPLEMENTIZER), DECL (DECLARATIVE), GEN (GENITIVE), LBL (LABEL), LTOP (LOCAL TOP), NOM (NOMINATIVE), PNE (PRENOMINAL ENDING), PST (PAST), RELS (RELATIONS), SEM (SEMANTICS), SPR (SPECIFIER), SYN (SYNTAX), TOP (TOPIC), etc.

Y. Matsumoto et al. (Eds.): ICCPOL 2006, LNAI 4285, pp. 387–394, 2006.
© Springer-Verlag Berlin Heidelberg 2006

b. haksayng-tul-**ul** sey-myeng-**i/ul/*eykey** ttena-key hayessta
student-PL-ACC three-CL-NOM/ACC/*DAT leave-COMP did
'(We) made three students to leave.'

As given in the raising (2a) and causative (2b), the Num-CL *sey myeng* 'three CL' can have a different case marking from its antecedent, functioning as the matrix object. In a sense, it is linked to the original grammatical function of the raised object and the causee, respectively.

Central issues in deep-parsing numeral classifier constructions thus concern how to generate such FQ constructions and link the FQ with its remote antecedent together with appropriate semantics. This paper shows that a typed feature structure grammar, HPSG, together with Minimal Recursion Semantics (MRS), is well-suited in providing the syntax and semantics of these constructions for computational implementations.[3]

2 Data Distribution

We have inspected the Sejong Treebank Corpus to figure out the distributional frequency of Korean numeral classifiers in real texts. From the corpus of total 378,689 words (33,953 sentences), we identified 694 occurrences of numeral classifier expressions. Of these 694 examples, we identified 36 FQ examples, some of which are given in the following:

(3) a. ... salam-i cengmal **han salam-to** epsessta.
person-NOM really one CL-also not.exist
'Nobody was really there.'
b. ... kkoma-lul **han myeng** pwuthcapassta
...little.boy-ACC one CL caught
'(We) grasped one little boy.'

The FQ type is relatively rare partly because the Sejong Corpus we inspected consists mainly of written texts. However, the statistics clearly show that these FQ constructions are legitimate constructions and should be taken into consideration if we want to build a robust grammar for Korean numeral classifiers.[4]

3 Implementing an Analysis

3.1 Forming a Numeral-Classifier Sequence and Its Semantics

The starting point of our analysis is forming the well-formed Num-CL expressions. Syntactically, numeral classifiers are a subclass of nouns (for Japanese see

[3] Minimal Recursion Semantics, developed by [2], is a framework of computational semantics designed to enable semantic composition using only the unification of type feature structures. See [2] and [3]. The value of the attribute SEM(ANTICS) in our system represents a simplified MRS.

[4] In addition to the FQ type, the Num-Cl can appear with a genitive case marking (GC type) or can combine with a noun without any particle (NI type). For an analysis of the GC and NI types, see [4].

[5]). However, unlike common nouns, they cannot stand alone and must combine with a numeral or a limited set of determiners:[5]

(4) a. *(twu) kay 'two CL' (Numeral)
b. *(myech) kay 'how many' (Interrogative)

Semantically, there are tight sortal constraints between the classifiers and the nouns (or NPs) they modify. For example, *pen* can classify only events, *tay* machinery, and *kwuen* just books. Such sortal constraints block classifiers like *tay* from modifying thin entities like books as in **chayk twu tay* 'book two-CL'. Reflecting these syntactic and semantic properties, we can assign the following lexical information to numerals (*num-det*) and classifiers (*cl-n*) within the feature structure system of HPSG and MRS.[6]

(5)

a.
$$\begin{bmatrix} \textit{num-det} \\ \text{ORTH } \langle \text{sey ''} \rangle \\ \text{SYN} \mid \text{HEAD} \begin{bmatrix} \text{POS } \textit{det} \\ \text{NUM } + \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{HOOK} \begin{bmatrix} \text{INDEX } \textit{i} \\ \text{LTOP } \textit{h2} \end{bmatrix} \\ \text{RELS} \left\langle \begin{bmatrix} \text{PRED } \textit{card\_rel} \\ \text{LBL } \textit{h2} \\ \text{ARG0 } \textit{i} \\ \text{CARG 3} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{cl-n} \\ \text{ORTH } \langle \text{myeng ''} \rangle \\ \text{SYN} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{POS } \textit{noun} \\ \text{CLTYPE } + \end{bmatrix} \\ \text{VAL} \mid \text{SPR } \left\langle \begin{bmatrix} \text{NUM } + \\ \text{INDEX } \textit{i} \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{HOOK} \begin{bmatrix} \text{INDEX } \textit{i} \\ \text{LTOP } \textit{h1} \end{bmatrix} \\ \text{RELS} \left\langle \begin{bmatrix} \text{PRED } \textit{person\_rel} \\ \text{LBL } \textit{h1} \\ \text{ARG0 } \textit{i} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The feature structure in (5a) represents that there exists an individual x whose CARG (constant argument) value is "3". The feature NUM is assigned to the numerals as well as to determiners like *yele* 'several' and *myech* 'some' which combine with classifiers. Meanwhile, (5b) indicates that syntactically a classifier selects a NUM element through the SPR, whereas semantically it belongs to the ontological category *person_rel*. The feature CLTYPE differentiates classifiers from common nouns. An independent grammar rule then ensures that only [NUM +] elements can combine with the [CLTYPE +] expression, ruling out unwanted forms such as **ku myeng* 'the CL'.

3.2 Syntax and Semantics of the Floating Quantifier Constructions

As noted earlier, the Num-CL can float away from the NP it classifies. There exist several supporting phenomena indicating that the FQ modifies the

[5] A limited set of common nouns such as *salam* 'person', *kulus* 'vessel', *can* 'cup', *khep* 'cup', and *thong* 'bucket' can also function as classifiers.

[6] The value of LBL is a token to a given EP (elementary predicate). The feature HOOK includes externally visible attributes of the atomic predications in RELS. The value of LTOP is the local top handle, the handle of the relations with the widest scope within the constituent. See [2] for the exact functions of each attribute.

following verbal expression. One phenomenon is the substitution by the proverb *kule-* 'do so'. As noted in (6), unlike the NI type, only in the NC type, an FQ and the following main verb can be together substituted by the proverb *kulay-ss-ta*:

(6) a. namca-ka [sey myeng o-ass-ko], yeca-to kulay-ss-ta
man-NOM three CL come-PST-CONJ woman-also do-PST-DECL.
'As for man, three came, and as for woman, the same number came.'
b. *[namca sey myeng-i] o-ass-ko, yeca-to [kulay-ss-ta]

This means that the FQ in the NC type is a VP modifier, though it is linked to a preceding NP.

Coordination data also support a VP modifier analysis:

(7) [namhaksayng-kwa] kuliko [yehaksayng-i] [sey myeng-i] oassta
boy student-and and girl student-NOM three CL-NOM came
'The total 3 of boys and girls came.'

The FQ 'three-CL' cannot refer to only the second conjunct 'girl students': its antecedent must be the total number of boys and girls together. This means the FQ refers to the whole NP constituent as its reference. This implies that an analysis in which the FQ forms a constituent with the preceding NP then cannot ensure the reading such that the number of boys and girls is in total three.

Given this VP-modifier treatment, the following question then is how to link an FQ with its appropriate antecedent. There exist several constraints in identifying the antecedents. When the floating quantifier is case-marked, it seems to be linked to an argument with the same case marking. However, further complication arises from examples in which either the antecedent NP or the FQ are marked not with a case marker, but a marker like a TOP:

(8) a. haksayng-tul-i/un sakwa-lul sey kay-lul mekessta
student-PL-NOM/TOP apple-ACC three CL-ACC eat
'As for the students, they ate three apples.'
b. sakwa-lul haksayng-tul-i/un sey kay-lul mekessta

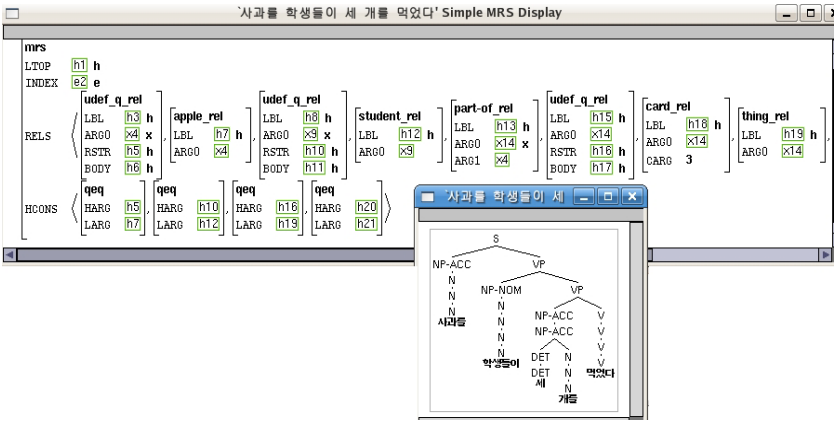
The data suggest that a surface case marking cannot be a sole indicator for the linking relation, and that we need to refer to grammatical functions. What we can observe is that, regardless of the location, the NOM-marked FQ is linked to the subject whereas the ACC-marked FQ is linked to the object. This observation is reflected in the following lexical information:[7]

[7] When the FQ has a delimiter marker (rather than a case marker) or no marker at all, it will refer to one of the elements in the ARG-ST (argument structure). Its antecedent will be determined in context.

(9)
$$\left[\begin{array}{l} \textit{num-cl-mw} \\ \text{ORTH}\ \langle \text{sey myeng-i} \rangle \\ \text{HEAD}\left[\begin{array}{l} \text{POS}\ \textit{noun} \\ \text{CASE}\,|\,\text{GCASE}\ \textit{nom} \\ \text{MOD}\left\langle \left[\begin{array}{l} \text{POS}\ \textit{verb} \\ \text{SUBJ}\left\langle \text{NP}_i \right\rangle \end{array}\right] \right\rangle \end{array}\right] \\ \text{SEM}\,|\,\text{HOOK}\,|\,\text{INDEX}\ i \end{array}\right] \left[\begin{array}{l} \textit{num-cl-mw} \\ \text{ORTH}\ \langle \text{sey myeng-ul} \rangle \\ \text{HEAD}\left[\begin{array}{l} \text{POS}\ \textit{noun} \\ \text{CASE}\,|\,\text{GCASE}\ \textit{acc} \\ \text{MOD}\left\langle \left[\begin{array}{l} \text{POS}\ \textit{verb} \\ \text{COMPS}\left\langle \text{NP}_i, \ldots \right\rangle \end{array}\right] \right\rangle \end{array}\right] \\ \text{SEM}\,|\,\text{HOOK}\,|\,\text{INDEX}\ i \end{array}\right]$$

As given in (9), the NOM-marked *num-cl-mw* modifies a verbal element whose SUBJ has the same index value, whereas the ACC-marked *num-cl-mw* modifies a verbal element which has at least one unsaturated COMPS element whose INDEX value is identical with its own INDEX value. What this means is that the NOM or ACC marked *num-cl-mw* is semantically linked to the SUBJ or COMPS element through the INDEX value. Our system yields the following parsing results for (8b):[8]

(10)



As seen from the parsed syntactic structure, the FQ *sey kay-lul* 'three CL-ACC' (NP-ACC) modifies the verbal expression *mek-ess-ta* 'eat-PST-DECL'. However, as noted from the output MRS, this modifying FQ is linked with its antecedent *sakwa-lul* 'apple-ACC' through the relation *part-of_rel*. Leaving aside the irrelevant semantic relations, let's see *card_rel* and *apple_rel*. As noted, the ARG0 value (x14) of *part-of_rel* is identified with that of *card_rel* whereas its ARG1 value (x4) is identified with the ARG0 value of the *apple_rel*. We thus can have the interpretation that there are three individuals x14s which belongs to the set x4.

4 Case Mismatches

Further complication in parsing FQ constructions comes from raising, causatives, and topicalization where the FQ and its antecedent have different case values.

[8] The attribute HCONS is to represent quantificational information. See [3].

In such examples, the two need not have an identical case value. For example, as given in (11b), the ACC-marked raised object can function as the antecedent of either the NOM-marked or ACC-marked FQ:

(11) a. namcatul-i [yecatul-**i** sey myeng-**i/*ul** chakhata-ko] mitessta.
men-NOM women-NOM three-CL-NOM/*ACC honest-COMP thought
'Men thought that three women are honest.'
b. namcatul-i yecatul-**ul** sey myeng-**i** chakhata-ko mitessta.

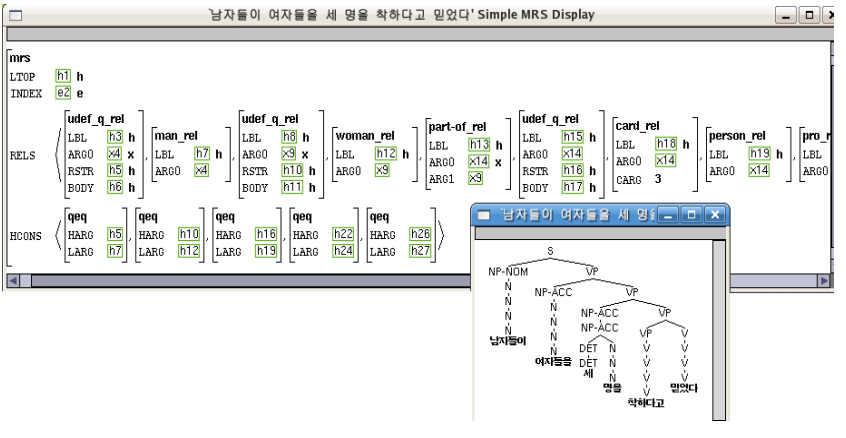
In the present analysis in which the case-marked FQ is linked to either the SUBJ or a COMPS element as given in (12), we can expect these variations. Let us consider the lexical entry for the raising verb *mitessta* 'believed':

(12) a. $\begin{bmatrix} \text{HEAD} \mid \text{POS } \textit{verb} \\ \text{SUBJ } \langle \boxed{1}\text{NP} \rangle \\ \text{COMPS } \langle \boxed{2}\text{S} \rangle \\ \text{ARG-ST } \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix}$ b. $\begin{bmatrix} \text{HEAD} \mid \text{POS } \textit{verb} \\ \text{SUBJ } \langle \boxed{1}\text{NP} \rangle \\ \text{COMPS } \langle \boxed{2}\text{NP}_i, \boxed{3}\text{VP[SUBJ } \langle \text{NP}_i \rangle] \rangle \\ \text{ARG-ST } \langle \boxed{1}, \boxed{2}, \boxed{3} \rangle \end{bmatrix}$

(12a) represents the lexical entry for *mitessta* 'believed' in (11a) selecting a sentential complement. Meanwhile, (12b) represents the raising verb 'thought' in (11b) in which the subject of the embedded clause is raised as the object. That is, *yecatul-ul* 'women-ACC' functions as its object even though it originally (semantically) functions as the subject of the embedded clause.

Equipped with these, our grammar generates the following parsing results for (11a):

(13)



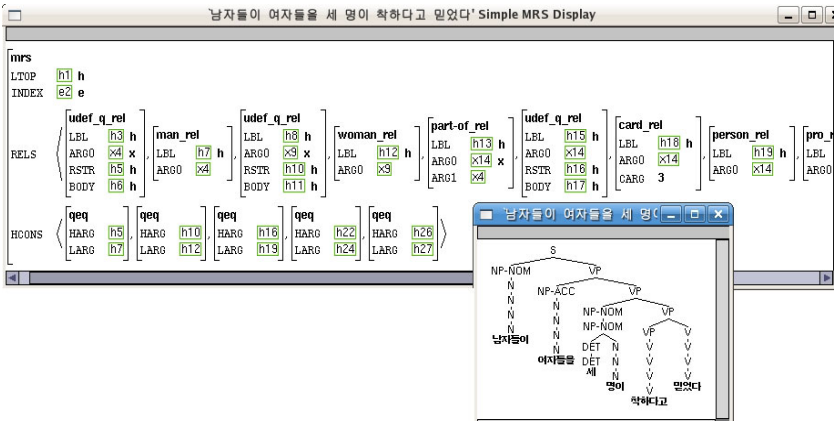
Syntactically, as noted from the parsed structure, the ACC-marked FQ *sey myeng-ul* 'three CL-ACC' (NP-ACC) modifies the VP *chakhata-ko mitessta* 'honest-COMP believed'.[9] Meanwhile, semantically, the ACC-marked FQ is linked to the ACC-marked object *yecatul-ul* 'woman-ACC'. This is because in

[9] Our grammar allows only binary structures for the language. One strong advantage of assuming binary structures comes from scrambling facts. See [7].

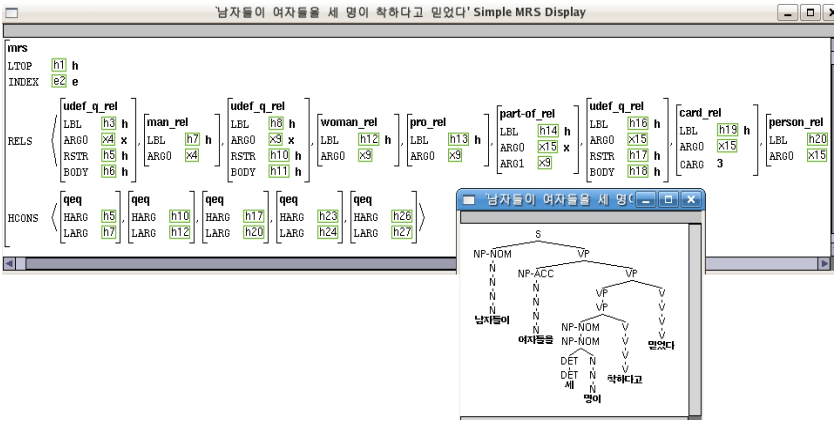
our grammar the antecedent of the ACC-marked FQ must be an unsaturated complement of the VP it modifies. As noted from the semantic relations *part-of_rel*, *card_rel* and *woman_rel* in the parsed MRS, this linking relation is attested. That is, the ARG0 value (x9) of *woman_rel* is identified with the ARG1 value of *part-of_rel* whereas the ARG0 value of *card_rel* is identical with the ARG0 value of *part-of_rel*. Thus, the semantic output correctly indicates that the individuals denoted by the FQ is a subset of the individuals denoted by the antecedent.

For the raising example (11b), our grammar correctly produces two structures. Let's see (14) first. As seen from the parsed syntactic structure here, the FQ *sey myeng-i* 'three CL-NOM' (NP-NOM) modifies the complex VP *chakhata-ko mitessta* 'honest-COMP believed'. However, in terms of semantics, the FQ is linked to the subject of the VP that it modifies.[10]This linking relation is once again attested by the MRS structure here. As noted here, the two semantic arguments of *part-of_rel*, ARG0 and ARG1, have identical values with the ARG0 value of *card_rel* (x14) and *man_rel* (x4), respectively.

(14)



(15)



[10] As another semantic constraint, the FQ can be linked only to a sentential internal argument.

Meanwhile, as given in the second parsing result (15), the FQ *sey myeng-i* 'three CL-NOM' modifies the simple VP *chakhata-ko* 'honest-COMP' only. Since the VP that the FQ modifies has only its SUBJ unsaturated, the SUBJ is the only possible antecedent. The output MRS reflects this raising property: The ARG0 value of *part-of_rel* identified with that of *card_rel* whereas its ARG1 value is identified with the ARG0 value of *woman_rel*. Our system thus correctly links the NOM-marked FQ with the ACC-marked antecedent even though they have different case values.

5 Future Work and Conclusion

One of the complicated issues in building a robust parsing system is whether to cover empirical as well as psychological (intuition-based) data. Even though examples like the case mismatches in FQ occur not often in the corpus data we inquired, we need to deal with such legitimate constructions if we want to develop a system aiming for reflecting the fundamental properties of the language in question.

The grammar we have built within the typed-feature structure system and well-defined constraints, eventually aiming at working with real-world data, has been implemented in the HPSG for Korean. We have shown that the grammar can parse the appropriate syntactic and semantic aspects of the FQ constructions. The test results provide a promising indication that the grammar, built upon the typed feature structure system, is efficient enough to build semantic representations for the simple as well as complex FQ constructions.

References

1. Kang, B.M.: Categories and meanings of Korean floating quantifiers-with some reference to Japanese. Journal of East Asian Linguistics **11** (2002) 375–398
2. Copestake, A., Flickenger, D., Sag, I., Pollard, C.: Minimal recursion semantics: An introduction. Manuscript (2003)
3. 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
4. Kim, J.B., Yang, J.: Processing Korean numeral classifier constructions in a typed feature structure grammar. In: Lecture Notes in Artificial Intelligence. Volume 2945. Springer-Verlag (2006) To appear
5. Bender, E.M., Siegel, M.: Implementing the syntax of Japanese numeral classifiers. In: Proceedings of IJCNLP-04. (2004)
6. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI Lecture Notes. Center for the Study of Language and Information, Stanford (2001)
7. Kim, J.B., Yang, J.: Projections from morphology to syntax in the Korean resource grammar: implementing typed feature structures. In: Lecture Notes in Computer Science. Volume 2945. Springer-Verlag (2004) 13–24