Processing Korean Numeral Classifier Constructions in a Typed Feature Structure Grammar

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

[1] School of English, Kyung Hee University, Seoul, 130-701, Korea
jongbok@khu.ac.kr
[2] School of Computer Engineering, Kangnam University, Kyunggi, 449-702, Korea
jhyang@kangnam.ac.kr

Abstract. The syntactic and semantic complexity of the so-called numeral classifier (Num-Cl) constructions in Korean challenges theoretical as well as computational linguists. We provide a constraint-based analysis of these constructions within the framework of HPSG with the semantic representations of MRS (Minimal Recursion Semantics) and reports its implementation in the LKB (Linguistic Knowledge Building) system.

1 Basic Data and Issues

One of the most salient features of languages like Korean is the complex behavior of numeral classifiers.[3] There exist at least three different environments where the numeral-classifier (Num-CL) expression can appear:[4]

(1) a. Genitive-Case (GC) Type:
sey myeng-uy haksayng-i o-ass-ta
three CL-GEN student-NOM come-PST-DECL
'Three students came.'
b. Noun Initial (NI) Type:
haksayng sey myeng(-i) o-ass-ta
student three CL-NOM come-PST-DECL
c. Noun-Case (NC) Type:
haksayng-i sey myeng-i o-ass-ta
student-NOM three CL-NOM come-PST-DECL

In the GC type, the Num-CL appears with the genitive case marking, preceding the modifying NP. In the NI type, the Num-CL sequence follows a caseless N, whereas in the NC type both the head noun and the following Num-CL are case-marked.

[3] Our thanks go to anonymous reviewers for the comments and suggestions. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2005-042-A00056).

[4] The abbreviations used for glosses and feature attributes in this paper are as follows: 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.

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 3206, pp. 1–8, 2006.

The foremost difficulty in parsing these constructions comes from the NC type in which the Num-CL floats away from its antecedent:

(2) **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.'

Within a system where no movement is allowed, it is not an easy task to correctly link the Num-CL to its remote antecedent.

In order to build a computationally feasible Korean grammar that can yield deep-parsing results, the grammar needs to form these three types of numeral classifier constructions and obtain semantics appropriate for each type. This paper shows that a typed feature structure grammar, HPSG, together with Minimal Recursion Semantics (MRS), is well-suited in providing the proper syntax and semantics of these three types of constructions.[5]

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 and identified the top 8 most frequently-used classifiers:

(3)

CL Type	Frequency	Examples
pen	158	oycwul han pen 'outgoing one CL'
salam	103	swunkem han salam 'policeman one CL'
kaci	70	yuhyung twu kaci 'type two CL'
myeng	56	kkoma han myeng 'child one CL'
kay	50	pang two kay 'room two CL'
mali	27	say han mali 'bird one CL'
cang	25	pyenci han cang 'letter one CL'
tay	20	cenhwa twu tay 'phone two CL'

Of the 694 examples, we identified 86 GC examples, 104 NI examples, and 36 NC examples. The remaining 468 examples consist of 365 anaphoric usages and 103 miscellaneous usages(e.g, ordinal, appositive usages).[6] As expected, the NI type occurs more often than the other two types. The NC patterns are relatively rare partly because the Sejong Corpus we inspected consists mainly of written texts. However, the statistics clearly show these three categories are legitimate constructions and should be taken into consideration if we want to build a robust grammar for Korean numeral classifiers. This research limited its scope to these three main types.

[5] Minimal Recursion Semantics, developed by [1], is a framework of computational semantics designed to enable semantic composition using only the unification of type feature structures. See [1] and [2]. The value of the attribute SEM(ANTICS) in our system represents a simplified MRS.

[6] Examples like *sey myeng-i o-ass-ta* 'three CL-NOM come-PST-DECL' are anaphoric usages in the sense that the antecedent of the Num-CL is within the given context.

3 Implementing an Analysis

3.1 Forming a Numeral-Classifier Sequence and its Semantics

The starting point of the analysis is forming the well-formed Num-CL expressions. Syntactically, numeral classifiers are a subclass of nouns (for Japanese see [3]). However, unlike common nouns, they cannot stand alone and must combine with a numeral or a limited set of determiners:[7] *(twu) kay 'two CL' (Numeral), *(yeleo/myech) kay 'several CL' (Quantifier), and *(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.[8]

(4)

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

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

The feature structure in (4a) 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, (4b) 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. Assuming that only [NUM +] elements can combine with the [CLTYPE +], we can rule out unwanted forms such as **ku myeng* 'the CL'. In addition, unlike quantifier determiners *motun* 'all' as in *ku motun haksayng* 'the all student', nothing can intervene between the NUM

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

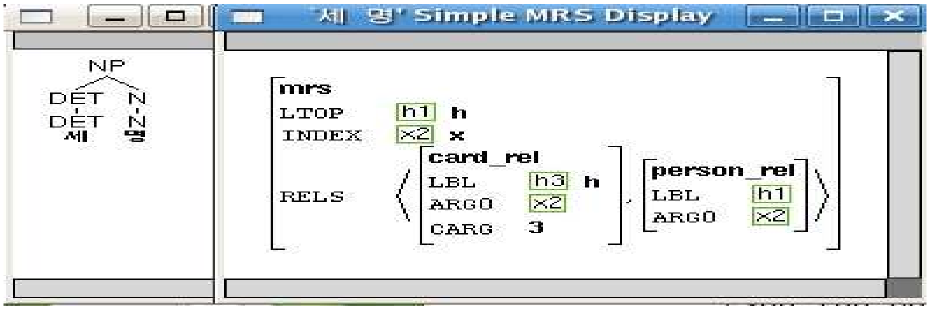
[8] The value of LBL is a token to a given EP (elementary predicate). The feature HOOK includes externally visible attributes of the atomic predications in RELS. The value of LTOP is the local top handle, the handle of the relations with the widest scope within the constituent. See [1] for the exact functions of each attribute.

and CL. Our grammar captures these multi-word like properties by treating the Num-CL sequence as a multiword (*mw*) expression formed by the following rule:[9]

(5) Num-CL Rule:

$$\left[\textit{num-cl-mw}\right] \rightarrow \begin{bmatrix}\textit{num-det}\\ \text{NUM} +\end{bmatrix}, [\text{CLTYPE} +]$$

When this rule is incorporated in our existing grammar and implemented in the LKB system[10], we then generate a right syntactic structure with the following MRS representation:



```
mrs
LTOP    h1 h
INDEX   x2 x
RELS  < [card_rel           ]   [person_rel ] >
        [LBL    h3 h        ] , [LBL   h1   ]
        [ARG0   x2          ]   [ARG0  x2   ]
        [CARG   3           ]
```

As represented here *sey myeng* ‘three CL’ forms a simple NP with the meaning that there are three individuals ‘x2’ which ontologically belongs to *person_rel*.

3.2 Genitive Case Type

Following [5], we assume the attachment of GEN case particle *-uy* to a nominal will add the information on GCASE (grammatical case) as well as the specification on the MOD feature:

(6)

$$\begin{bmatrix}\textit{num-cl-gen}\\ \text{ORTH}\ \langle \text{sey myeng-uy}\ \rangle\\ \text{SYN}\begin{bmatrix}\text{HEAD}\begin{bmatrix}\text{POS}\ \textit{noun}\\ \text{CASE}\,|\,\text{GCASE}\ \textit{gen}\\ \text{MOD}\ \langle \text{NP}_j\rangle\end{bmatrix}\end{bmatrix}\\ \text{SEM}\,|\,\text{RELS}\left\langle\begin{bmatrix}\text{PRED}\ \textit{card\_rel}\\ \text{LBL}\ \textit{h2}\\ \text{ARG0}\ \textit{i}\\ \text{ARG1}\ \textit{3}\end{bmatrix}, \begin{bmatrix}\text{PRED}\ \textit{person\_rel}\\ \text{LBL}\ \textit{h1}\\ \text{ARG0}\ \textit{i}\end{bmatrix}, \begin{bmatrix}\text{PRED}\ \textit{part-of\_rel}\\ \text{ARG0}\ \textit{i}\\ \text{ARG1}\ \textit{j}\end{bmatrix}\right\rangle\end{bmatrix}$$

[9] The type *num-cl-mw* is a subtype of *hd-spr-ph* formed by the combination of a head and its specifier.

[10] 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. 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. [4].

Unlike the simple expression *sey myeng*, the GEN marked expression *sey myeng-uy* adds an additional constraint: the MOD value indicates that the expression that the *num-cl-gen* modifies must be a nominal expression whose index value is associated with it through *part-of_rel*. Unlike the determiners, the GEN-marked NP functions as a modifier to a completely saturated NP as in *John-uy ku chinkwu* 'John-GEN the friend' or *ku John-uy chinkwu* 'the John-GEN friend'. In capturing such an NP property, our grammar introduces the Head-MOD rule (generating *hd-mod-ph*) that allows the combination of an adnominal element and its head, generating an appropriate syntactic structure and semantic representations.

3.3 NI (Noun-Initial) Type

The cleft sentences in (7) indicate that unlike in the NC type, in the NI type the head noun forms a strong syntactic unit with a following Num-CL:

(7) a. ku sensayngnim-ul mos ka-key ha-n kes-un
that teacher-ACC not go-COMP do thing-PNE
[haksayng sey myeng]-i-essta.
three-CL-GEN student-COP-PAST
'What made the teacher not leave were five students.'
b. *ku sensayng-nim-ul moskakey han kes-un [haksayng-i sey myeng-i]-ess-ta.

In addition, there exist various examples indicating that the NI type behaves like a synthetic compound or multiword expression. For example, the N and the following Num-CL sequence cannot be separated at all:[11]

(8) haksayng (*ku) sey myeng-i o-ass-ta
student the three CL-NOM come-PST-DECL
'Three students came.

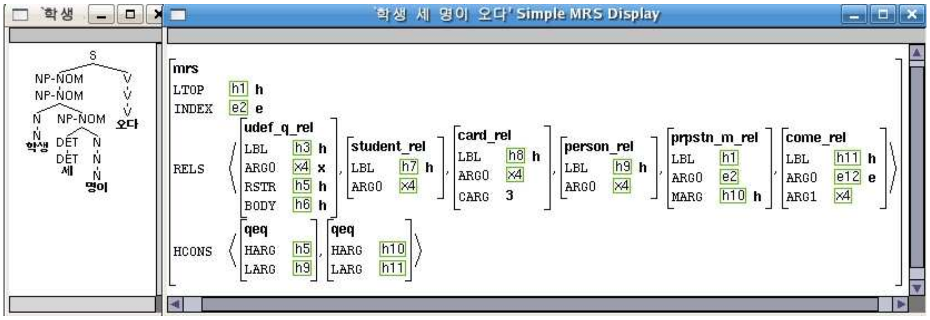
Such a tight syntactic cohesion supports the idea that the NI sequence is another multiword expression formed by a rule like the following:

(9) NI Compound Formation Rule:

$$\begin{bmatrix} \textit{cn-num-cl-mw} \\ \text{HEAD} \mid \text{MOD} \ \langle\ \rangle \end{bmatrix} \rightarrow \begin{bmatrix} \textit{cn} \\ \text{INDEX } i \end{bmatrix}, H \begin{bmatrix} \textit{num-cl-mw} \\ \text{CLTYPE} + \\ \text{INDEX } i \end{bmatrix}$$

The resulting type *cn-num-cl-mw*, unlike *num-cl-mw*, has an empty MOD value, indicating that it can function not as a modifier but as an argument. This formation rule will eventually license the combination of *haksayng* 'student' with *sey myong* as a multiword expression, generating the following structure and MRS for (8):

[11] A long pause between the two improves the example, but such an example can be taken to be an NC type.



```
mrs
LTOP   h1 h
INDEX  e2 e
RELS   < [udef_q_rel  LBL h3 h  ARG0 x4 x  RSTR h5 h  BODY h6 h],
         [student_rel  LBL h7 h  ARG0 x4],
         [card_rel  LBL h8 h  ARG0 x4  CARG 3],
         [person_rel  LBL h9 h  ARG0 x4],
         [prpstn_m_rel  LBL h1  ARG0 e2  MARG h10 h],
         [come_rel  LBL h11 h  ARG0 e12 e  ARG1 x4] >
HCONS  < [qeq  HARG h5  LARG h9], [qeq  HARG h10  LARG h11] >
```

As represented in the structure, the common noun *haksayng* 'student' combines with the *num-cl-mw* expression *sey myeng* in accordance with the formation rule in (9). They both have the same index value with their own semantic contributions as given in the RELS values. The NP then functions as the ARG1 of the *come_rel* relation which projects a propositional message (*prpstn_m_rel*).

3.4 Noun-Case Type

The NC type allows the NOM or ACC-marked NP to be followed by the identical case-marked NUM-CL (called FQ here) which even can float away from the NP as noted in (2). There exist several supporting phenomena indicating that the FQ modifies the following verbal expression. One phenomenon is the substitution by the proverb *kule-* 'do so'. As noted in (10), 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*:

(10) 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.

The 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, a 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:

(11) a. haksayng-tul-i/un sakwa-lul sey kay-lul mekta
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 mekta

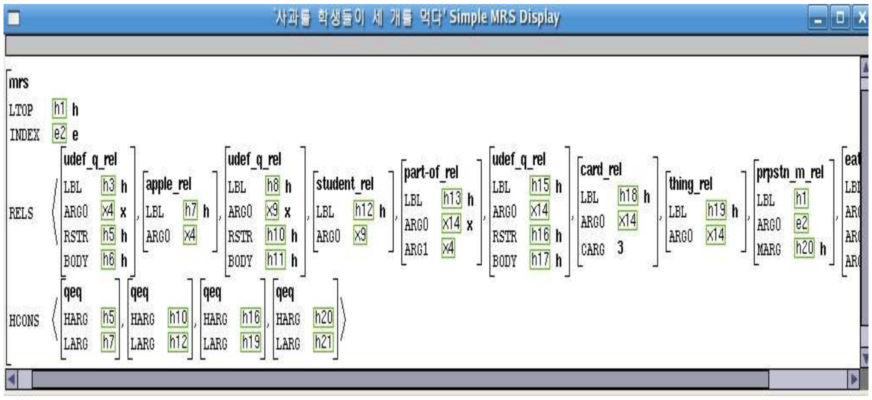
This implies that a surface case marking cannot be a sole indicator for the linking relation, and that we need to refer to grammatical functions. Regardless of its location, however, we can observe that the NOM-marked FQ is linked to the subject whereas the ACC-marked FQ is linked to the object. This observation is reflected in the following lexical information:

(12)

$$\text{a.}\begin{bmatrix} \textit{num-cl-mw} & \\ \text{ORTH } \langle \text{sey myeng-i} \rangle & \\ \text{HEAD} & \begin{bmatrix} \text{POS } \textit{noun} \\ \text{CASE} \mid \text{GCASE } \textit{nom} \\ \text{MOD} \left\langle \begin{bmatrix} \text{POS } \textit{verb} \\ \text{SUBJ} \langle \text{NP}_i \rangle \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM } \mid \text{HOOK} \mid \text{INDEX } \textit{i} & \end{bmatrix} \quad \text{b.}\begin{bmatrix} \textit{num-cl-mw} & \\ \text{ORTH } \langle \text{sey myeng-ul} \rangle & \\ \text{HEAD} & \begin{bmatrix} \text{POS } \textit{noun} \\ \text{CASE} \mid \text{GCASE } \textit{acc} \\ \text{MOD} \left\langle \begin{bmatrix} \text{POS } \textit{verb} \\ \text{COMPS} \langle \text{NP}_i , ... \rangle \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM } \mid \text{HOOK} \mid \text{INDEX i} & \end{bmatrix}$$

As given in the lexical information, the case-marked *num-cl-mw* functions as a specifier to a verbal expression, but quantifies over an argument with the same case value.

(13)



'사과를 학생들이 세 개를 먹다' Simple MRS Display

```
mrs
LTOP   h1 h
INDEX  e2 e
       udef_q_rel       apple_rel      udef_q_rel       student_rel     part-of_rel      udef_q_rel       card_rel        thing_rel       prpstn_m_rel    eat
RELS < LBL  h3 h    ,   LBL  h7 h  ,   LBL  h8 h    ,   LBL  h12 h  ,   LBL  h13 h   ,   LBL  h15 h   ,   LBL  h18 h  ,   LBL  h19 h  ,   LBL  h1     ,  LBI
       ARG0 x4 x        ARG0 x4        ARG0 x9 x        ARG0 x9         ARG0 x14 x       ARG0 x14         ARG0 x14        ARG0 x14        ARG0 e2          ARC
       RSTR h5 h                       RSTR h10 h                       ARG1 x4          RSTR h16 h       CARG 3                          MARG h20 h       ARC
       BODY h6 h                       BODY h11 h                                        BODY h17 h                                                        ARC
        qeq          qeq            qeq            qeq
HCONS < HARG h5  ,   HARG h10  ,    HARG h16  ,    HARG h20 >
        LARG h7      LARG h12       LARG h19       LARG h21
```

As given in (12), the NOM-marked *num-cl-mw* thus 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. As given in (13), this system provides a right MRS for (11b). The output MRS links the ARG0 value of *apple_rel* with the ARG0 value of the CL *thing_rel*.

4 Future Work and Conclusion

Our grammar has been implemented in the HPSG for Korean. In testing its performance and feasibility for parsing numeral classifier constructions, we used 100 sentences from the identified 226 sentences (GC, NI, and NC type) extracted from the Sejong corpus as noted in section 2, and 100 grammatical and 50 ungrammatical sentences extracted from the literature. As noted before, the grammar successfully constructed three main types of Num-CL constructions together with appropriate semantic representations. One strong merit of this analysis, as we have seen, is that it can capture the syntactic and semantic aspects of the NC type in which the NP and the FQ are not adjacent but in remote positions.

Our approach still needs to cover other types of numeral classifier constructions and then expand its coverage for authentic data. However, the test results provide a promising indication that the grammar, built upon the typed feature structure system, is efficient enough to build proper syntactic as well as semantic representations for the complex numeral classifiers.

References

1. Copestake, A., Flickenger, D., Sag, I., Pollard, C.: Minimal recursion semantics: An introduction. Manuscript (2003).
2. 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.
3. Bender, E. M., Siegel, M.: Implementing the syntax of Japanese numeral classifiers. In: Proceedings of IJCNLP-04. (2004).
4. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI Lecture Notes. Center for the Study of Language and Information, Stanford (2001).
5. 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.