

# Treating Multiple-Subject Constructions in a Constraint-Based MT-System

Munpyo Hong

In Korean and Japanese some sentences are to be found which contain more than one nominative-NP. Such constructions are called ‘multiple-subject constructions’ or ‘double-subject constructions’. They do not only raise a question about their syntactic and semantic nature but also cause such problems as structural changes in MT. They must be considered in designing an MT-System between two typologically different languages, for example, between Korean and German or between Japanese and German. In this paper I will show with Korean examples how can this construction be treated for a constraint-based MT-System. Further I will suggest a solution to the translational divergencies caused by the construction of this sort. The solution can be also applied to the Korean-English, Japanese-German, and Japanese-English MT without any modification.

Im Koreanischen und Japanischen gibt es die Sätze, die mehr als eine nominativ-NP enthalten. Solche Konstruktionen werden in der Literatur als ‘Doppel-Subjekt Konstruktionen’ oder als ‘Multiple-Subjekt Konstruktionen’ bezeichnet. Sie stellen nicht nur die Frage nach ihrer syntaktischen und semantischen Struktur, sondern sie verursachen auch Probleme wie strukturelle Veränderungen in der MÜ. Versucht man vor allem ein MÜ-System zwischen sprachtypologisch völlig anderen Sprachen zu konstruieren, so muß diese Konstruktion berücksichtigt werden. In diesem Aufsatz wird gezeigt, wie die Konstruktionen für ein constraint-basiertes MÜ-System analysiert werden können. Darüber hinaus wird eine Lösung zu den durch diese Konstruktionen hervorgerufenen Übersetzungsproblemen vorgeschlagen.

## 1 Introduction<sup>1</sup>

In Korean and Japanese, there are sentences which have more than one nominative-NP in a clause, as seen in (1). In these languages such grammatical functions as subject and object are marked by case markers assigning nominative and accusative case to the noun.

- (1)a. ku yeca-ka ku namca-ka mipta.(Korean)  
Det woman-NOM Det man-NOM hate  
‘The woman hates the man’

---

<sup>1</sup> I wish to give my thanks to Prof. Dr. Johann Haller, Dr. Oliver Streiter and Catherine Pease in IAI for their valuable comments. My warm appreciation goes also to the anonymous reviewers of this paper.

- a' John-ga Mary-ga kirai.(Japanese)  
 John-NOM Mary-NOM hate  
 'John hates Mary'
- b. ku namca-ka ton-i manhta.(Korean)  
 Det man-NOM money-NOM much  
 'The man has much money'
- b' New York-ga kosokenchiku-ga taksan aru.(Japanese)  
 New York-NOM skyscraper-NOM many exist  
 'In New York there are many skyscrapers'
- c. ku yeca-ka son-i yepputa.(Korean)  
 Det woman-NOM hand-NOM beautiful  
 'The woman has beautiful hands'
- c' Nihon-ga dansei-ga tanmei desu.(Japanese)  
 Japan-NOM man-NOM short-life-span be  
 'In Japan, mens are short-life-span'
- d. ku ai-ka mok-i maluta.(Korean)  
 Det child-NOM throat-NOM ?  
 'The child is thirsty'
- d' watashi-wa onaka-ga suitea.(Japanese)  
 I-TOPIC stomach-NOM empty  
 'I am hungry'

Constructions with more than one nominative-NP are called 'Multiple-Subject Constructions(=MSC)'. There have been a few pieces of research about MSCs in Korean and Japanese. Most of them focuss on the syntactic nature of MSCs in the framework of a general linguistic theory such as GB or a grammar formalism such as LFG and HPSG, and some have dealt with the semantic aspect of the construction.<sup>2</sup> However, there have been few attempts to treat MSCs in NLP. MSC does not only raise the question of its syntactic and semantic nature, but also causes such problems as structural changes in machine translation(=MT). This construction must be taken into account in designing an MT-System between two typologically different languages, for example, between Korean and German or Japanese and German. In this paper I would like to suggest that there are two kinds of MSCs with some Korean examples, one being motivated by the lexical features of some predicates, and the other triggered by the semantic characteristics of the subject NP. Both of them will be described in the CAT2 formalism. Then, I will present a solution to the translational divergencies caused by the construction of this sort, which will be implemented in the constraint-based MT-System CAT2<sup>3</sup>. The suggested solution will not be based on the structure-specific transformations but rather on

<sup>2</sup> [Choi94], [Fukui88] and [Kiss81] treated MSCs in Korean and Japanese in the GB framework. [Shin91] tackled the problems in the LFG framework, and [Chang93] in the HPSG framework.

<sup>3</sup> CAT2 is both a grammar formalism and an MT-System. About the CAT2, cf.[Sharp94]

the simple linguistic knowledges such as subcategorization, linear precedence and semantic constraints. The solution can also be applied to the Japanese-to-German or Japanese-to-English MT without any modification.

## 2 Lexically motivated MSC

### 2.1 Idiomatic expressions

MSCs are found in some idiomatic expressions. Here are some examples.

- (2)a. ku        namca-ka    pay-ka        kophuta  
       Det        man-NOM    stomach-NOM    ?  
       ‘The man is hungry’
- b. Ku        ai-ka        mok-i        maluta  
       Det        child-NOM    throat-NOM    ?  
       ‘The child is thirsty’

In these examples ‘pay-ka kophuta’ and ‘mok-i maluta’, in (2a,b) respectively, are sentences as themselves and also predicates of the whole sentences taking as their subject the nominative-NP ‘ku namca-ka’ and ‘ku ai-ka’. The idiomatic expressions which allow for MSCs have only two nominative-NPs, i.e., if one more NP with nominative case is attached to the sentence, the sentence will be ungrammatical.

The reason we have to handle these sentences as idiomatic expressions is that they do not show compositionality in building their meanings. In (2.a) the verb ‘kophuta’ has no meaning of its own and is only used when combined with the NP ‘pay-ka’. The translation pairs in (2) also show the lexical gaps between Korean and English. For the analysis of the idiomatic expressions, I suggest an ‘extended support verb construction’. Support verb construction(=SVC) is actually composed of a predicative noun which has its own argument structure and a verb which has almost no meaning except temporal and aspectual information. [Mesli91] and [Streiter96] dealt with SVCs in the CAT2 formalism, and [Krenn94] presents the syntactic and semantic analysis of SVCs in HPSG framework. Therefore I will not go into the details of SVCs in this paper. The nouns ‘pay’ and ‘mok’ in (2) are not predicative nouns in the traditional sense. But these sentences are similar to SVCs, in that the verbs ‘kophuta’ and ‘maluta’ do not have their own meanings and the corresponding translations.<sup>4</sup> Thus I will suggest that the SVC be applied to these expressions,

---

<sup>4</sup> It was pointed out by an anonymous reviewer that in Japanese, in contrast to Korean, it is difficult to say that the predicates have no meaning of their own in such constructions, as seen in the following examples.

(1) sono hito-ga onaka-ga suita(that person is hungry)

(2) sono hito-ga onaka-ga itai(that person has stomach-ache)

too. The idea of the SVC analysis in CAT2 was that the predicative noun has an argument structure of its own, and the verb shares the argument structure of the predicative noun by the so-called ‘argument transfer’.<sup>5</sup>

The noun ‘mok(=throat)’ takes one argument which bears the ‘theme’ role and has the nominative case. It also takes a support verb(=*vsup*) ‘maluta’ as a complement. The verb ‘maluta’ has no semantic information except temporal and aspectual information {speech=simul, aspect=dur} and inherits the semantic information of the noun by variable binding(SEM). The argument structure of the verb will be unified with that of the predicative noun by the argument transfer(ARG1). The lexical entries for the predicative noun ‘mok’ and the verb ‘maluta’ are illustrated in Fig.1

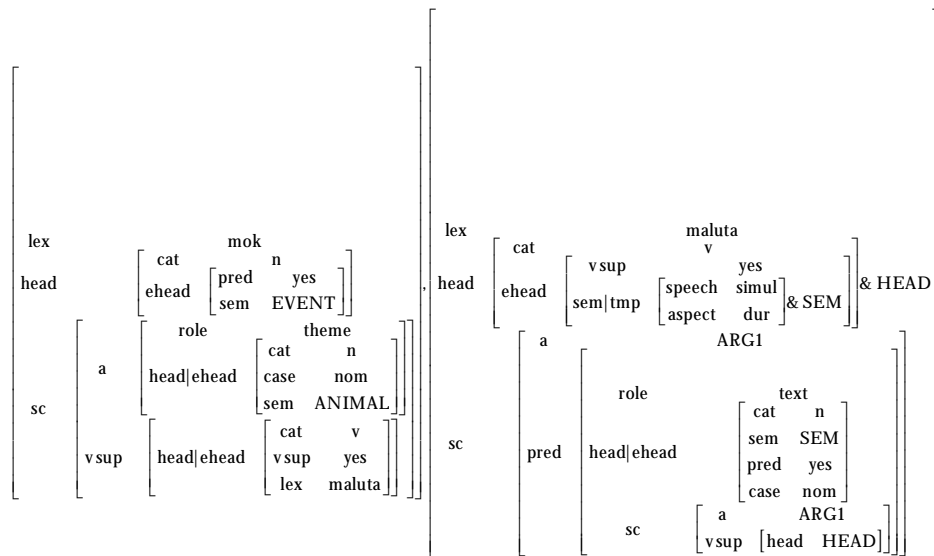


Fig.1: The lexical entries for the predicative noun ‘mok’ and the verb ‘maluta’

For the transfer of SVCs, we only need the predicative noun, discarding the support verb. The semantic information of the verb will not go lost but is stored, as its ‘sem’ value will be unified with that of the predicative noun. The lexical transfer rule for the word ‘mok’ is illustrated in (3).

(3)  $t\_mok\_durst = \{lex = mok, head = \{ehead = \{pred = yes\}\}\}.[] \Leftrightarrow \{lex = durst\}.[]$

The German word ‘durst(=thirst)’ takes also a support verb ‘haben(=have)’, and it will be generated in the intervening stages between the interface structure and the surface structure for the German sentence. The system will then generate the correct German translation ‘Das Kind hat Durst’ for the Korean sentence ‘ku aika moki maluta(=The child is thirsty)’.

(3) sono hito-ga onaka-ga ippai(that person has had enough)

(4) kono resutoran-wa itsumo suite-imasu(this restaurant is always full)

<sup>5</sup> The idea of argument transfer is originally brought out by [Grimshaw88].

## 2.2 ‘psyche’-adjectives and ‘exist’-verbs

The so-called ‘psyche’-adjectives and the ‘exist’-verbs also allow for MSCs, as seen in (1.a) and (1.b). The ‘psyche’-adjectives express the emotional state of the subject referent. The ‘exist’-verbs are the verbs which can be paraphrased in English and German with ‘there are (many) sth.’ and ‘es gibt (viel) etw.’.

As in the case of the idiomatic expressions, these sentences only allow for two nominative-NPs. These predicates have only two arguments in their argument slot. Therefore if one more argument is attached to, it cannot be subcategorized for. However, the problem is how to relate the thematic roles with the appropriate NPs. Because the case information is the only method for it, it is problematic in this case.

- (4) silhta(=hate)      <goal, theme>  
                               haksayng(student)    swuhak(mathematics)  
                               \*swuhak                    haksayng  
       issta(=have)      <goal, theme>  
                               namca(man)                ton(money)  
                               \*ton                            namca

Therefore we need more information to link the thematic role to the correct NP. For this purpose, let us observe word order in Korean.

Korean displays relatively free word order. Except for the fixed word order in a phrase, there are no strict restrictions among the NPs in a sentence. The NPs can be freely scrambled in a sentence without changing the propositional meaning of the sentence.

- (5) a. ku    yeca-ka    ku    chayk-ul    ku    ai-eykey    cwuessta.  
       b. ku    chayk-ul    ku    yeca-ka    ku    ai-eykey    cwuessta.  
       c. ku    ai-eykey    ku    chayk-ul    ku    yeca-ka    cwuessta.  
       ‘ The woman gave the book to the boy’

In the above examples (5) the 3 NPs(ku yeca-ka, ku chayk-ul, ku ai-eykey) can be scrambled without changing the propositional meaning of the sentence(‘The woman gave the book to the boy’)<sup>6</sup>, as long as the main verb(=cwuessta) is at the end of the sentence. However, we can observe that MSCs with the ‘psyche’-adjectives and the ‘exist’-verbs do not allow for this.

- (6)a.\*    swuhak-i                    ku    haksayng-i                    silhta  
                               mathematics-NOM Det    student-NOM                hate  
       b. \*    ku            chayk-i                    ku    namca-ka                    issta.  
                               Det            book-NOM                Det man-NOM                have

<sup>6</sup> The propositional meaning of a sentence is the meaning only built by its subparts, i.e., without the discourse information such as topics and comments.

The scrambling of the NPs in these sentences is not allowed unlike in other Korean sentences. This word order constraint can be formulated with the following f-rules<sup>7</sup>.

I introduced a feature ‘psy’ and ‘exist’ in order to differentiate the ‘psyche’-adjectives and the ‘exist’-verbs from the other predicative adjectives and the verbs. The ‘psyche’-adjectives have the AV-pair {head={ehead={psy=yes}}}} and the ‘exist’-verbs {head={ehead={exist=yes}}}} in their lexical entries. The other verbs are assigned {head={ehead={psy=no}}}} and {head={ehead={exist=no}}}} as a default value by the f-rules (7).

- (7) f\_head\_ehead\_psy\_no=={head={cat=v,ehead={psy=no}}}.[].  
 f\_head\_ehead\_exist\_no=={head={cat=v,ehead={exist=no}}}.[].

These f-rules are applied to the lexical items, whilst the default values are assigned to them. The lexical entry for the word ‘silhta(=hate)’ is illustrated in Fig.2<sup>8</sup>.

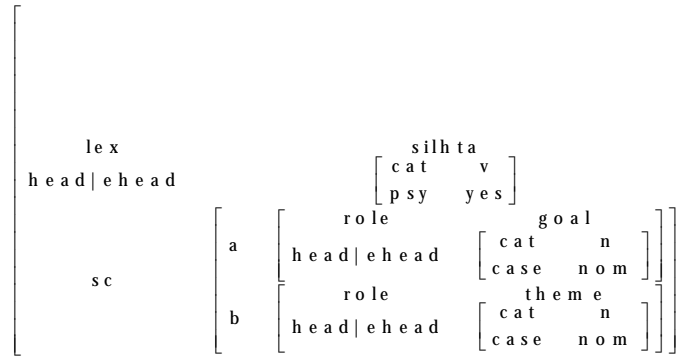


Fig. 2: The lexical entry of ‘silhta(=hate)’

The lexical entries of the ‘exist’-verbs look similar to those of the ‘psyche’-adjectives, except that they have ‘exist=yes’ instead of ‘psy=yes’. To restrict the mapping between the second nominative-NP and the ‘goal’ role, the following f-rule was designed.

- (8) f\_psy\_exist\_goal\_theme={}.[{{role=goal,head={ehead={cat=n,case=nom}}}}>>{head=nil},  
 {head={ehead={cat=v,({psy=yes});{exist=yes}}}}].[]].

This ‘f\_psy\_exist\_goal\_theme’ rule says if a nominative-NP with the role ‘goal’ is followed by a ‘psyche’-adjective or an ‘exist’-verb, it will inevitably fail. It will filter out ungrammatical sentences like in (6).

The ‘psyche’-adjectives do not pose difficult problems in translation. However, some problems arise in translating some ‘exist’-verbs.

<sup>7</sup> The f-rules in CAT2 formalism perform two functions; they can assign default values to the lexical entries, and they can also constrain the objects built by the b-rules which are the CAT2 counterparts of the ID-rules in HPSG.

<sup>8</sup> For the ease of implementation, I classified the ‘psyche’-adjectives as stative verbs.

- (9)a. ku haksayng-i chinkwu-ka manhta.  
 Det Student-NOM friend-NOM have many  
 ‘Der Student hat viele Freunde(=The student has many friends)’
- b. ku kyoswu-ka chayk-i manhta.  
 Det Professor-NOM book-NOM have many  
 ‘Der Professor hat viele Bücher(=The professor has many books)’

As we can see in the example sentences in (9), the Korean verb ‘manhta’ is expressed by a verb(=haben) and a modifier of the argument of the verb(=viel) in German. Such translational divergencies in MT are treated in most transfer-based approaches by the complex transfer rules. But the drawback of such an approach is well-known<sup>9</sup>, i.e., if some complex structural changes are interwoven, the rules often fail to be applied to. To avoid these problems, we try to solve it on the lexical level, maintaining the interface structure of the source language. We assume that the second NP-complement of the verb ‘manhta(=have many)’ is plural, i.e., if it is a countable noun, its number value will be plural. Thus we encoded the information that the second complement-NP has the AV pair {head={ehead={sem={bound=many}}}} in the lexical entry of the predicate.

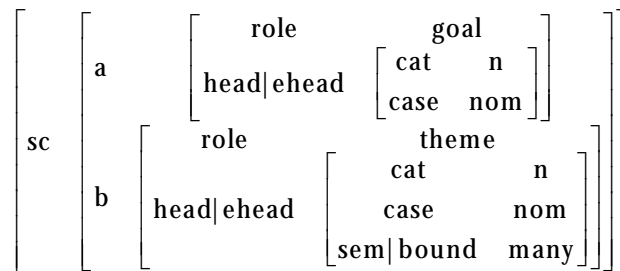
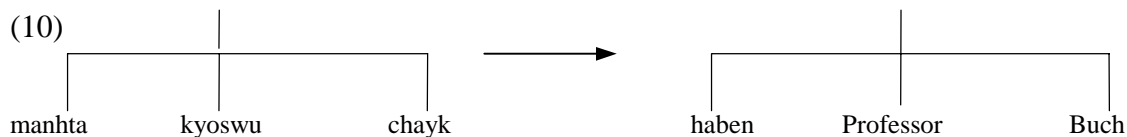


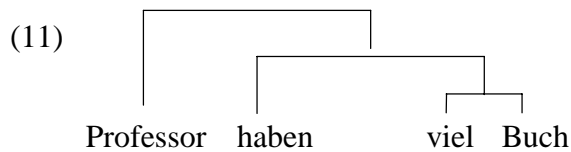
Fig. 3: The subcategorization frame of ‘manhta’

The interface structure of the Korean sentence in (9.b) will then be transferred to the German interface structure without any structural changes as illustrated in (10).



The AV pair {head={ehead={sem={bound=many}}}} of the second argument, in this case ‘Buch(=book)’, will trigger the generation of the adjective ‘viel(=many)’ before the noun in the generation phase like (11).

<sup>9</sup> cf [Dorr93] pp.25



### 3 Semantically motivated MSCs

Until now we have observed MSCs triggered by the subcategorization of the predicates. These predicates were either idiomatic predicates or the predicates with the special features, [+psy] or [+exist]. They allow for only two nominative-NPs and no more. But there are other MSCs which do not belong to either group.

- (12)a. peter-ka emeni-ka celmta.  
 Peter-NOM mother-NOM young.  
 ‘Peter’s mother is young’
- b. hankwuk-i namca-ka swumyeng-i ccalpta  
 Korea-NOM men-NOM life-span-NOM short.  
 ‘It is in the case of Korea that men are short-life-span’
- c. ku tayhakkyo-ka tokmwunkwa-ka 2 haknyen-i yehaksayng-i  
 i yepputa.  
 DET university-NOM department of German-NOM sophomore-NOM female  
 students-NOM beautiful.  
 ‘It is in the German department of the university that the female sophomore students are beautiful’

These constructions deviate from the lexically driven constructions, in that they do allow for more than two nominative-NPs. Furthermore there do not seem to be common semantic features among the predicates, except that they are all adjectives. [Chang93] classified above examples as a ‘whole-and-part’ relation. But this classification is not appropriate, because we cannot classify the NPs in the sentence (12.a), ‘Peter’ and ‘his mother’, as the relation of ‘whole-and-part’.

[Fukui88] argued using Japanese examples that in such constructions the NP is adjuncted to the so-called X-bar node and thus a theoretically unlimited number of NPs can be attached to the sentence. However, this argument did not take into account of the semantic aspect of the construction. Because in (12.a) and (12.b), for example, if one more nominative-NP is attached to, the sentence will be ungrammatical like (13).

- (13) a. \* hans-ka peter-ka emeni-ka celmta.  
 Hans-NOM peter-NOM mother-NOM young
- b. \* ilpon-i hankwuk-i namca-ka swumyeng-i ccalpta  
 Japan-NOM Korea-NOM men-NOM life-span-NOM short.

[Shin91] pointed out an interesting aspect of this construction in the framework of Montague semantics. According to him, the reason these predicates allow for MSCs is that the type of subject is  $\langle e, t \rangle$ , i.e., that of a common noun in the type-theoretical terminology. So it cannot be combined with a predicative adjective, the type of which is also  $\langle e, t \rangle$ , so that such sentences cannot have a truth value. To avoid this, the language needs some other devices to restrict the denotation of the NP, the subject of the predicate. Thus if the sentence (12.a) did not have the NP ‘peter-ka’, the sentence would be ungrammatical, because the subject NP ‘emeni’ denotes the set of entities which are mothers of someone in a given model.

(14) \* emeni-ka celmta.<sup>10</sup>  
 Mother-NOM young

The first nominative-NP in (12.a) picks up a certain entity in the set of ‘emeny(mother)’, i.e., ‘Peter’s mother’ to assign the property of ‘being young’. The sentence (12.a) then is a semantically saturated sentence. Thus no NP can be attached to it like (13.a).

At this point two questions must be answered. Firstly, when can nominative-NPs be attached to? Secondly, what is the limit to the number of nominative-NPs which can be attached to? My answer to the first question is that the nominative-NP can be attached to when the referentiality of the subject NP is not definite or deictically not fixed. As is well known, in Korean and Japanese the existence of a determiner in an NP is not obligatory from a syntactic point of view. In many cases the NPs in these languages are just bare nouns with case markers. The number and the referentiality of these NPs are often deduced from a context and the world knowledge. The answer to the second question follows automatically, i.e., if the referentiality of the subject is definite or deictically fixed, no more nominative-NPs can be attached to. In this construction the right most nominative-NP, i.e., the NP which is at the very left of the predicate will be the subject of the sentence, and the others will be only the specifiers of this subject, semantically restricting the denotation of the subject NP.

The algorithm I adopt for the analysis of the semantically driven MSC is (15).

(15) In  $S_n = NP_{Nom1}, NP_{Nom2}, \dots, NP_{Nomi-1}, NP_{Nomi}, AP$   
 ( $NP_{Nomi}$  abbreviates a nominative-NP in ‘i’th position in a sentence)  
 step 1: if  $NP_{Nomi}$  is definite or deictically fixed, then go to step 3  
           otherwise go to step 2  
 step 2: set  $i=i-1$ , and if  $i \geq 1$ , then go to step 1  
           otherwise fail<sup>11</sup>

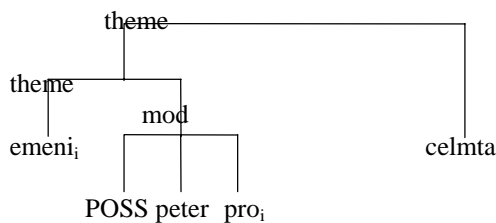
<sup>10</sup> In this paper I do not treat the generic reading of the sentences. Therefore the possible generic reading of (14), ‘every mother in this world is young without exception’, is excluded.

step 3: if  $i=1$ , then succeed  
 otherwise fail

To determine the referentiality and the number of an NP in Korean and Japanese automatically is also a very difficult problem, which can not be handled here.<sup>12</sup> In this paper I simply assume that the referentiality of an NP in Korean is definite in case i) the NP is composed of a noun and a definite article or ii) the NP is a proper name like ‘John’ or ‘ilpon(=Japan)’.

Because we do not allow an NP to function as a modifier in CAT2 MT-System, we need some other lexical units to express the relationship between the NPs. The semantic relation between these two NPs in (12a), ‘Peter’ and ‘emeni(=mother)’, is that of ‘possession’ in a broad sense. Therefore the underlying semantic structure of the sentence could be expressed like this.

(16)



‘POSS’ is an abstract lexical unit which expresses the ‘possessor-possessed’ relationship. The idea behind this strategy is that in every language there are some words expressing the ‘possession’ relationship. Therefore it is unnecessary to say explicitly that a word ‘A’ expressing ‘possession’ relationship in one language is translated to the word ‘B’ in another language. What must be said is only that the ‘possession’-relationship is transferred to other languages without changes. The Korean interface structure (16) will be transferred to the German interface structure without any structural changes.

Then the generation module of the target language can generate the corresponding word for the concept. In our system, the German morpheme ‘-s’ or ‘von jm’ will be generated for the ‘POSS’. Using this strategy, we could keep the interface structure of the both languages unchanged.

## 4 Conclusion

In this paper I suggested that there are two kinds of MSCs in Korean. One type is lexically motivated. It can be analyzed with the subcategorization of the predicates and word order constraints. Further I showed that the proposed

<sup>11</sup> If the rule fails to be applied to, the rescue rules in CAT2 system are activated for the robustic processing. As a result, the user gets the rough default translation of the source sentence.

<sup>12</sup> About the countability and referentiality of an NP in Japanese, cf. [Bond94]

methods do not require any complex transfer for the translation. The other type is semantically motivated. The referentiality of the subject plays the most important role in describing the construction. Also here we could avoid the transformation by adopting a lexical function ‘POSS’.

Determining the countability and referentiality of an NP in Korean and Japanese automatically is a difficult problem, which is very important for the correct analysis of MSCs. It should be done along with the researches about MSCs in future works.

## References

- [Bond94] Francis Bond, Kentaro Ogura and Satoru Ikehara(1994): Countability and Number in Japanese to English Machine Translation, *Proceedings Vol.1, COLING 94*, pp.32-38
- [Chang93] Chang, Seok-Jin(1993): *Information-based Korean Grammar*, Language & Information Research Assn. Seoul
- [Choi94] Choi, Myung-Won(1994): Kasustheorie und Mehrfachnominativekonstruktion im Koreanischen, Heidelberg, SFB 340: *Sprachtheoretische Grundlagen für die CL/Bericht Nr.57*
- [Dorr93] Dorr, Bonnie Jean(1993): *Machine Translation: A View from the Lexicon*, The MIT Press
- [Fukui88] Fukui, N(1988): Deriving the differences between English and Japanese: A case study in parametric Syntax, *English Linguistics 5*.
- [Grimshaw88] J. Grimshaw, A. Mester(1988): Light Verbs and  $\theta$ -Marking, *Linguistic Inquiry, Vol.19, Nr.2*, pp.205-232
- [Kiss81] Kiss, Katalin E.(1981): On the Japanese ‘double subject’ construction, *The linguistic review vol.1. NO.2*, pp.155-170
- [Krenn94] Brigitte Krenn, Gregor Erbach(1994): Idioms and Support Verb Construction: *German in Head-Driven Phrase Structure Grammar*, pp.365-395, CSLI
- [Mesli91] Mesli, Nadia(1991): Funktionsverbgefüge in der maschinellen Analyse und Übersetzung: Linguistische Beschreibung und Implementierung in CAT2 Formalismus. *Eurotra-D Working Papers 19*. IAI, Saarbrücken, Germany
- [Sharp94] Sharp, Randall(1994): CAT2 Reference Manual Version 3.6, *IAI Working Papers Nr.27*, Saarbrücken, Germany
- [Shin91] Shin, Soo-Song(1991): *Das Verstehen der Unifikationsgrammatik - Lexikalische Funktionale Grammatik*, Hanshin Verlag, Seoul

[Streiter96] Streiter, Oliver(1996): *Linguistic Modelling for Multilingual Machine Translation*, Shaker Verlag