Corpus-Based Study of Word Order Variations in Korean

Seoin Shin[1]

1. Introduction

This paper aims to show the results of an analysis on word order variations in Korean using a syntactically annotated Korean corpus. Korean has long been said to have free word order, because noun phrases usually have postpositions as case markers. However, the case markers do not guarantee all the scrambled orders. In some cases, one order is more acceptable than the others in real language use; in some other cases, certain order is not acceptable in certain reason.

(1) a. 리나가 민희에게 선물을 주었다.
Rina-ga Minhee-ege seonmul-eul ju-ess-da.
Rina-SBJ Minhee-I_OBJ present-D_OBJ give-PAST-FINAL
'Rina gave Minhee a present.'

b. 리나가 선물을 민희에게 주었다.
Rina-ga seonmul-eul Minhee-ege ju-ess-da.
Rina-SBJ present-D_OBJ Minhee-I_OBJ give-PAST-FINAL
'Rina gave a present to Minhee.'

(2) a. 리나가 민희를 친구로 알았다.
Rina-ga Minhee-reul chingu-ro al-ass-da.
Rina-SBJ Minhee-OBJ friend-ADJ regard-PAST-FINAL
'Rina regareded Minhee as a friend.'

b. *리나가 친구로 민희를 알았다.
Rina-ga chingu-ro Minhee-reul al-ass-da.
Rina-SBJ friend-ADJ Minhee-OBJ regard-PAST-FINAL
'*Rina regareded Minhee as a friend.'

Example 1: Word order variations in Korean

[1] Department of Korean Language and Literature, Seoul National University
e-mail: seoin.shin@gmail.com

In Example 1, sentence (1a) is more natural than sentence (1b), but both of them are acceptable. Even though two sentences have same grammatical meaning, they have different meaning in the sense of information structure. If one says the sentence in order of (1b), then indirect object '*Minhee-ege*' is been focused. On the other hand, sentence (2b) has a scrambled order of sentence (2a), but the order of (2b) is not acceptable. Between (1a) and (1b) word order is a matter of preference; but between (2a) and (2b) word order is a matter of correctness. In this paper the factors that cause word order variation will be analyzed.

If one word order is preferred to other possible orders in constructing a sentence, this particular word order can be expected to occur more frequently in a corpus. The current study uses a corpus to find out the extent of and the determining factors in word order variations in Korean.

2. Corpora

In this research, Korean parsed corpora of *21c Sejong project* are used. *21c Sejong project* is a government-funded project to build corpora and produce an electronic dictionary for natural language processing. The parsed corpus of *21c Sejong project* has been built since 2002. The human annotators use tools that aid to build the corpora, but most of the processes are being done manually. All of the parsed corpora are written text, and they are composed of news, novels, essays, etc. The size of corpus which is used in this paper is 570,064 words; 52,292 sentences.

```
; 어머니가 힐끔 아버지의 얼굴을 바라다봤다.
(S-     (S      (NP_SBJ 어머니/NNG + 가/JKS)
        (VP     (AP 힐끔/MAG)
                (VP     (NP_OBJ(NP_MOD 아버지/NNG + 의/JKG)
                                (NP_OBJ 얼굴/NNG + 을/JKO))
                (VP 바라다보/VV + 았/EP + 다/EF))))
        (S- + ./SF))
```

Figure 1: *21C Sejong project* Korean parsed corpora

The syntactic annotations have been added on POS tagged corpora. Each sentence has two parts; raw sentence and annotated sentence. The raw sentence starts with semi-colon (;) and the annotated sentence has both POS tags and syntactic tags. The POS tags are presented with slash (/) after each word: NNG for noun, VV for

verb, JKS for subject marker, EF for final ending, etc. The syntactic tags are presented with parentheses on phrases and clauses: SBJ for subject, OBJ for object, CMP for complement, AJT for adjunct, etc. Syntactic annotations are presented on the opening parentheses.

For the research on word order variation, the predicate-argument patterns of the each verb were extracted from this corpus. The process of pattern extraction is illustrated in Shin (2005). Here is the example of the extracted patterns.

~e/JKB^NP_AJT	~i/JKS^NP_SBJ	~eul/JKO^NP_OBJ	nae/VV	‘submit’
~eun/JX^NP_SBJ	~eul/JKO^NP_OBJ	~euro/JKB^NP_AJT	ori/VV	‘cut’
		~eun/JX^NP_SBJ	us/VV	‘laugh’

Figure 2: Representation of arguments: postpositions and arguments

Each line of Figure 2 presents the realized argument structure of each verb in sentences. The list shows the order of the arguments in authentic data. In this list, some of the verbs have same pattern, so that verbs can be classified into same group according to the patterns. The pattern has relevance to meaning of the verb.

3. Word order variations and obligatoriness

In order to find tendency of word order variation, we are going to look at the sentences which have ‘NP-*i* NP-*eul* NP–*ro* V’ pattern in surface structure.

word order	frequency
NP-i NP-ro NP-eul V	88
NP-i NP-eul NP-ro V	46
NP-ro NP-i NP-eul V	15
NP-ro NP-eul NP-i V	0
NP-eul NP-i NP-ro V	2
NP-eul NP-ro NP-i V	0
total	151

Table 1: word order variations of 'NP-*i*, NP-*ro*, and NP-*eul*'

Basic information about word order can be gathered from table 1. First of all, even though case markers represent the functions of noun phrases in a sentence, the word order is not totally free. There exists preferred word order. The subject NP-*i* prefers to

be at the beginning of a sentence, and it doesn't like to be in the very front of a verb.

Secondly, in many cases subject comes before object, but object rarely comes before subject. We can make sure that Korean is an SOV language. The order of subject, object and verb has well been known, but it is hard to find a description on the position of the other elements. It is necessary to pay attention to the rest element: NP-*ro*.

(3) 현우는 눈물어린 눈으로 마을을 내려다봤다.
Hyunwoo-neun nunmuleori-n nun-euro maeul-eul naeryeodabo-ass-da.
Hyunwoo-SBJ tearful eye-ADV village-OBJ look down-PAST-FINAL.
Hyunwoo looked down the village with tearful eyes.

(4) 경찰은 이씨를 범인으로 단정했다.
Gyengchal-eun issi-reul beomin-euro danjeonghassda.
police-SBJ Mr. Lee-OBJ culplit-O.C. conclude-PAST-FINAL.
The police concluded that Mr. Lee is the culprit.

In the sentence (3) NP-*(eu)ro* is an adjunct, and in the sentence (4) NP-*(eu)ro* is an complement. In surface structure both of them have *(eu)ro* as a case marker, but one represents adjunct and the other represent complement. They show difference in word order.

word order	NP-ro: complement	NP-ro: adjunct	total frequency
NP-i NP-ro NP-eul V	18	70	88
NP-i NP-eul NP-ro V	33	13	46
NP-ro NP-i NP-eul V	0	15	15
NP-ro NP-eul NP-i V	0	0	0
NP-eul NP-i NP-ro V	1	1	2
NP-eul NP-ro NP-i V	0	0	0
total	52	99	151

Table 2: word order variations according to obligatoriness of NP-*ro*

From table 2, we can find that the obligatoriness of NP affects word order variation. First of all, when NP-*ro* is a complement, it prefers following object to preceding object: the frequency of 'NP-*i* NP-*eul* NP-*ro* V' were almost two times of the frequency of 'NP-*i* NP-*ro* NP-*eul* V'. Meanwhile, when NP-*ro* is an adjunct, it prefers following subject to preceding subject or following object: the frequency of 'NP-*i* NP-*ro* NP-*eul* V' was overwhelming.

Secondly, when NP-*ro* is a complement, the distribution of it is more restricted: NP-*ro* cannot occur before subject. On the contrary, when NP-*ro* is an adjunct, it can move freely: it can precede or follow subject and it can follow object as well.

Thirdly, when NP-*ro* is an adjunct, that verb is considered to have the pattern 'NP-*i* NP-*eul* V'. In that case, the order of subject and object is rather fixed though wherever adjunct NP-*ro* is inserted. Among three possible orders, adjunct NP-*ro* prefers to occur following subject closely.

(5)	a.	NP-i	(NP-ro)	NP-eul		V
	b. (NP-ro)	NP-i		NP-eul		V
	c.	NP-i		NP-eul	(NP-ro)	V

As seen above, word order is affected by the obligatoriness of noun phrases. The location of required element is rather fixed: on the contrary the position of optional element is relatively flexible.

In Somers (1987:189-190), the relative centrality of the participants in a sentence was represented as a series of concentric orbits around the central predicator. In the hierarchy of complement-middle-adjunct-peripheral, complement orbits inner circle and extra-peripheral orbits outer circle. It is true in English: the more required the inner, the more optional the outer. The statement can be revised for the order of noun phrases in the language where the case marker is developed such as Korean: the more required the more fixed, the more optional the more flexible.

4. Word order variations and thematic role

In this chapter, four groups of verbs that have 'NP-*i* NP-*eul* NP-*ro* V' as a basic pattern will be examined. The verbs in first group are *ormgida*, *idongsikida*, etc, which are meaning 'move'. This kind of verbs needs three arguments: agent, theme, and goal.

(6) a. 경찰은 이씨를 유치장으로 옮겼다.
Kyungchal-eun issi-reul yuchijang-euro ormgi-ess-da.
police-SBJ Mr. Lee-OBJ the house of custody-ADJ ormgi-PAST-FINAL.
The police took Mr. Lee into custody.

b. 경찰은 유치장으로 이씨를 옮겼다.
Kyungchal-eun yuchijang-euro issi-reul ormgi-ess-da.
police-SBJ the house of custody-ADJ Mr. Lee-OBJ move-PAST-FINAL.
The police took Mr. Lee into custody.

In the case of this kind of verb, both of (6a) and (6b) are possible. However, (6a) is more frequent than (6b). In other words, when NP-*ro* represents a goal of a movement, the order of agent-theme-goal is more natural.

The verbs in second group are *ganjuhada, yeogida*, *chwigeuphada*, etc. They all mean 'regard'. This kind of verbs needs three arguments: agent, theme, and goal as well. In this case, however, the characteristic of goal is different from that of (6). This kind of goal is 'identificational goal' of Jackendoff (1990).

(7) 남들은 그들을 하나의 단위로 취급한다.
Namdeul-eun geudeul-eul hana-eui danwi-ro chwigeubha-n-da.
other people-SBJ them-OBJ one group-O.C. treat-PRESENT-FINAL
Other people treat them as a gang.

(8) 유학자들은 이러한 것을 불교의 약점으로 간주한다.
Yuhakjadeul-eun ireoha-n geot-eul bulgyo-eui yakjeom-euro ganjuha-n-da.
confucianists-SBJ these thing-OBJ week points of Buddhism-O.C. regard-PRESENT-FINAL
Confucianist regards these things as week points of Buddhism.

The object and object complement (O.C.) of these sentences have a 'is a' relation: 'they are a gang' in example (7) and 'these things are week points of Buddhism' in example (8). For that reason, in this case, theme always precedes goal(identificational goal). The arguments of the verb in this group have relative fixed position. In the corpus only the order of 'NP-*i* NP-*eul* NP-*ro* V' was found and the scrambled order of it was not found.

The verbs in third group are *kkumida*, *jangsikhada*, etc, which are meaning 'decorate'. This kind of verbs needs three arguments: agent, theme, and instrument.

(9) 그들은 아기자기한 소품으로 의자 주변을 꾸몄다.
Geu-deul-eun agijagiha-n sopum-euro uija jubyeon-eul kkumi-eoss-da.
they-SBJ cute accessories-ADJ uija jubyeon-OBJ kkumi-PAST-FINAL
They decorated around the chair with cute accessories.

(10) 그들은 촛대 주변을 솔가지로 꾸몄다.
Geu-deul-eun chosdae jubyeon-eul solgaji-ro kkumi-eoss-da.
they-SBJ chosdae jubyeon-OBJ solgaji-ADJ kkumi-PAST-FINAL
They decorated candlestick with pine twigs.

The position of NP-*ro* representing instrument is relatively free. The order of agent-instrument-theme is slightly more frequent than agent-theme-instrument.

The verbs in fourth group are *samda* and *mandeulda*. They mean 'make'. In this case, if the position of 'NP-*ro*' is changed, the meaning of the sentence would be changed.

(11) 그는 불량학생들을 양자양녀로 삼았다.
Geu-neun bulyanghaksaengdeul-eul yangjayangnyeo-ro sam-ass-da.
he-SBJ disorderly students-OBJ adopted son and daughter-O.C. make-PAST-FINAL
He adopted disorderly students.

(12) 그는 맹물을 포도주로 만들었다.
Geu-neun maengmul-eul podoju-ro mandeul-eoss-da.
he-SBJ water-OBJ wine-O.C. make-PAST-FINAL
He turned water into wine.

(13) 이들은 그 수익금으로 장학금을 만들었다.
Ideul-eun geu suikgeum-euro janghakgeum-eul mandeul-eoss-da.
they-SBJ the proceeds-ADJ scholarship-OBJ make-PAST-FINAL
They found a scholarship with the preceeds.

(14) 그들은 전시회장을 쑥대밭으로 만들었다.
Geudeul-eun jeonsihoijang-eul ssudaebat-euro mandeul-eoss-da.
They-SBJ exhibition hall-OBJ wormwood-O.C. make-PAST-FINAL
They turned the exhibition hall into ruins.

In the example (11), if the positions of NP-*eul* and NP-*ro* are switched, then the meaning of the sentence would be totally changed: 'He made his adopted son and daughter into disorderly students'. In this switched order, NP-*ro* would be regarded as an instrument. It is same in example (12): if the positions of NP-*eul* and NP-*ro* are switched, then the sentence would have opposite meaning: 'He turned wine into water'.

The verb *samda* and *mandeulda* has two kinds of patterns: 'NP-*i* NP-*ro* NP-

eul V' and 'NP-*i* NP-*eul* NP-*ro* V'. In the former pattern, the first NP is agent, the second NP is instrument, and the third NP is theme; and in the latter pattern, the first NP is agent, the second NP is theme, and the third NP is goal. In the latter case, the relationship of NP-*eul* and NP-*ro* is similar to that of example (7-8).

In example (13), as only one way of change is possible, two noun phrase cannot be switched. If you want to change the sentence while maintaining the meaning of the sentence, you must change the pattern into this: 'NP-*i* NP-*eul* NP-*ro* V'.

Example (14) seems to be similar to (11-13) at a glance, but either other kind of order nor other kind of pattern is allowed for this meaning. Because noun phrase of NP-*ro* is a result of the change, it cannot be an instrument of in the other pattern. In this case the word order is fixed.

For the verb *mandeulda*, according to the patterns three senses of (12-14) must be discriminated in a dictionary.

.

(15) *mandeulda*

Sense 1: [NP1-*i* NP2-*eul* NP3-*ro* V]

NP1 makes NP2 with NP3.

Sense 2: [NP1-*i* NP2-*ro* NP3-*eul* V]

NP1 makes NP3 with NP2.

Sense 3: [NP1-*i* NP2-*eul* NP3-*ro* V]

NP1 turns NP2 into NP3.

For the reason we have seen until now, when we develop automatic Korean-English translator, we need to take care of the pattern and the word order. If we translate noun phrase by noun phrase, it would be hard to capture the similarity between sense 1 and sense 2 and the difference between sense1 and sense 3. Here we can find the significance of the study of sentence pattern and word order variation.

5. Conclusion

In this paper, we have observed word order variation of 'NP-*i* NP-*eul* NP-*ro* V' pattern. According to the corpus, we reconfirmed SOV is most natural order in Korean. However, the position of the adverb NP-*ro* and the tendency of the variation of it were varied according to the meaning of the verb. The corpus showed not only which word order is most frequently used out of many theoretically possible orders but also what factors play a role in restricting these variations.

One of the factors, which decideed the extent of the word order variations, was obligatoriness. It was found that ‘NP-*ro*’ occurred in front of a predicate when it was a complement in an argument structure of the predicate, while its position was not fixed when ‘NP-*ro*’ was an adjunct. Another factor was a thematic role. In some cases one word order was preferred: a theme preceded an identificational goal in the sentences of the ‘regard’ group; in other cases word order was flexible: a locative goal preceded or followed theme. On the other hand, the existence of other kinds of pattern affected the flexibility of word order. An instrument did not affect the position of the preceding theme or the following theme in the sentences of the ‘decorate’ group, while instrument was always precede theme in the sentences of ‘make’.

A study of sentence pattern and word order variation is necessary to increase the performance of automatic parsing and automatic translation. Syntactically annotated corpora provide useful data for this kind of study to us. In order to analysis the tendency of word order variation, many kinds of linguistic concepts in various dimensions are necessary. The corpus-based study can enrich the discussion on word order variation.

References

Abeille, A. (ed.) (2003) *Treebanks: Building and Using Syntactically Annotated Corpora*, Boston: Kluwer.

Allerton, D. J. (1982) *Valency and the English Verb*, Academic Press.

Brill, E. and P. Resnik (1994) A Rule-based Approach to Prepositional Phrase Attachment Disambiguation. In *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING, 1994)*, Kyoto, Japan.

Grimshaw, J. (1990) *Argument Structure*, The MIT Press.

Heylen, K. and D. Speelman (2003) A Corpus-based Analysis of Word Order Variation: The Order of Verb Arguments in the German middle field. In *Proceedings of Corpus Linguistics 2003*. University of Lancaster. pp. 320–29.

Hunston, S. and G. Francis (2000) *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam and Philadelphia: John Benjamins.

Jackendoff, R. (1990) *Semantic Structures*, The MIT Press.

Nelson, G., S. Wallis, and B. Aarts (2002), *Exploring natural language*, John Benjamins.

Shin, S. (2005) Automatic Pattern Extraction for Korean Sentence Parsing, in *Proceedings from The Corpus Linguistics Conference Series*, Vol. 1, no. 1.

University of Birmingham. Available on-line from http://www.corpus.bham.ac.uk/PCLC.

Somers, H. L. (1987) *Valency and Case in Computational Linguistics*, Edinburgh University Press.

Williams, E. (1994) *Thematic Structure in Syntax*, The MIT Press.