Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 10-19

Three Kinds of Korean Reflexives: A Corpus Linguistic Investigation on Grammar and Usage

*Beom-mo Kang**
Korea University

This paper discusses the relationship between grammar as linguistic knowledge, as envisaged in Generative Grammar, and usage, the result of performance. In concrete, I analyze the use of Korean reflexives 'caki', 'casin', and 'cakicasin' by examining the occurrences of these reflexives in a 5-million-word Korean corpus, taken from a 10-million-word Korean corpus which is called "KOREA-1 Corpus" compiled at Korea University (H. Kim and B. Kang 1996). This corpus is composed of various genres of Korean texts including 10% of spoken material. From the KWIC concordance of accusative forms of these reflexives, 'cakilul', 'casin-ul', and 'cakicasin-ul', I examined whether a reflexive has a local antecedent or a long-distance antecedent. The result is that 'caki' is almost even in having local and long-distance antecedents, but 'casin' has more and 'cakicasin' has much more local antecedents. I also examined the thematic roles of the local antecedents of reflexives, which shows that 'casin' has relatively more Experiencer antecedents than 'caki' has, although in both cases Agent antecedents dominate. The outcome of these frequency analysis suggests that a tendency (probably not grammaticalized yet), or degree of "naturalness" is real and can be captured in the usage data provided that we have a sizable amount of material manageable in an efficient way as provided by the corpus linguistic method of the present day.

1. LINGUISTIC DATA AND LINGUISTIC INVESTIGATION

The goal of Generative Grammar is to understand the human mind better through the study of linguistic knowledge of native speakers. The linguist's grammar is an explicit representation of such linguistic knowledge, i.e. "language as what a speaker knows". Since linguistic knowledge is very complicated we cannot ask speakers to describe their linguistic knowledge overtly. Instead, generative grammarians have used native speakers' intuition on the grammaticality of sentences as basic linguistic data. This method is in contrast with the earlier structuralists' method of using observable linguistic data, i.e. the product of linguistic performance. The reason why generative grammarians resort to intuition is that language use (performance) is interfered with by nonlinguistic factors such as memory and noise; and that the creativity of language cannot be described on the basis of limited amount of observable linguistic data.

However, if we can study language on the basis of a sizable amount of linguistic data, the amount that has not been available and cannot be handled manually by a single linguist, we are in a different situation. The modern corpus linguistic methods provide such an opportunity. For example, the compilation of COBUILD dictionary was made possible by a 20-million-word English corpus that was processed by efficient computer programs. It not only set a standard for the dictionary compilation but also opened a possibility of using corpora for general linguistic studies (Sinclair 1987).

I do not claim that only the usage data of language should be proper data for any kind of linguistic investigation, but that there are linguistic phenomena that can only be captured on the basis of a corpus and this new method can reveal a new perspective on grammar (cf. Church, et al. 1991). To support such a claim, I take up three kinds of Korean reflexives 'caki', 'casin', and 'cakicasin', which have been extensively investigated by Korean generative grammarians. They used native speakers', usually their own, intuition. In contrast, I used half of a 10-million-word Korean corpus compiled at Korea University, which is called "KOREA-1 Corpus" (Kim and Kang 1996). The KOREA-1 Corpus is composed as follows:

---

* Department of Linguistics, Korea University, Seoul, 136-701 Korea. E-mail: bmkang@kuccnx.korea.ac.kr.

(1) Composition of KOREA-1 Corpus

Spoken	12%
Newspaper	20%
Magazine	10%
Book-Informative	34%
Book-Imaginative	21%
Others	3%
10,000,000 words in total	

In actuality, I used only half of the corpus while keeping the composition kept as above.

2. THE ANTECEDENT CONDITION FOR 'CAKI'

Reflexives are an important topic in Generative Grammar. The binding theory is supposed to explain or describe native speakers' intuition on the grammaticality of the following sentences (I will use the Yale System of Romanization of Hangul in this paper):

(2) a. cengswuka cakilul cohahanta.
Cengswu-Nom self-Acc likes
'Cengswu likes himself.'
b. *cakika cengswulul cohahanta.
self-Nom Cengswu-Acc likes
'Himself likes Cengswu.'

But there are some aspects in the grammar of reflexives that require (observable) linguistic data to be examined, as will be shown below.

One of the reasons why the Korean reflexive 'caki' has been treated so importantly in the Korean grammar is that it differs from English counterparts in that it can take a long-distance antecedent as well as a local antecedent. For example,

(3) pyengswunun cengswuka cakilul silhehantako sayngkakhanta.
Pyengswu-Top Cengswu-Nom self-Acc dislikes-Comp thinks
'Pyengswu thinks that Cengswu dislikes himself/him.'

In this sentence, the antecedent of 'caki' can be either 'Pyengswu', the subject of the matrix sentence, or 'Cengswu', the subject of the embedded sentence. 'Caki' is a kind of long-distance reflexive, which is found in many other languages such as Chinese, Japanese, Swedish and Icelandic (Kang 1988). Many linguists have discussed the antecedent condition of 'caki' with respect to syntactic structures, thematic roles, and discourse factors. Among others, there have been discussions regarding whether 'caki' tends to take a long-distance antecedent or a local antecedent. For example, which do you prefer in the above sentence as the antecedent of 'caki', 'Pyengswu' or 'Cengswu'?

There have been opposing claims: a claim that 'caki' tends to take a local antecedent versus a claim that 'caki' tends to take a long-distance antecedent. Among others, let us examine Chang's (1986) and Moon's (1996) claims.

Chang (1986) claims that 'caki' tends to take antecedents in the following order, according to the general "Local Antecedent First" principle:

(4) [+NCL] sentence > [-NCL] sentence, i.e.
simplex S > complex S > paragraph

According to him, 'caki' in the following sentence tends to take 'Mia' rather than 'Chelswu' as the antecedent.

(5) chelswunun miaka caki chaykul ilhepelyesstako (naeykey) allye cwuessta.
Chelswu-Top Mia-Nom self book-Acc lost (I-Dat) informed
'Chelswu informed me that Mia lost her/his book.' (preference: mia > chelswu)

But it seems to me that 'Chelswu' equally tends to be the antecedent for 'caki'.

In contrast, Moon (1996) suggested the following reflexive interpretation constraints in the framework of Optimality Theory.

(6) OT Constraints on Reflexive Interpretation

THC thematic hierarchy constraint (Agt > Exp > Goal,Pat,Th,Sour > Loc)
LDC local domain preference constrain
LPC larger domain preference constrain
SOC subject-orientedness constraint
CCC c-command constraint
DBC discourse binding constraint

These constraints are assumed to be universal and an individual language is assumed to select some of them and impose an order of preference on the selected constraints. English and Korean differ in the preference ordering as follows, he claims.

(7) English reflexives THC > CCC > SOC > LDC > DBC
Korean 'caki' THC > LPC > SOC > CCC > DBC

According to Moon (1996), the main difference between English and Korean is that English reflexives follow LDC (local domain preference) but Korean 'caki' observes LPC (larger domain preference), so that 'caki' tends to take a long-distance antecedent. For example,

(8) Johni Billi Tomi cakilul piphanhantako mitnuntako malhayssta.
John-Nom Bill-Nom Tom-Nom self-Acc criticized believed said
'John said that Bill believed that Tom criticized him/himself.'
(John > Bill > Tom)

But again, I cannot decide which antecedent is better for 'caki'. (Moreover, such a sentence with a threefold embedding with 'caki' never occurs in the corpus.)

Then, does the discussion on local/long-distance antecedents of reflexives have no meaning at all? I believe it has some significance, but only when the question is taken up on the basis of the use of language since our intuition about the preference of antecedents without context is not clear. I will follow the following research strategy:

(9) We examine the use of 'caki-lul', the accusative form of 'caki,
i.e. we count the local and long-distance antecedents of 'caki-lul'.

Note that I will not examine all of the word forms of 'caki' but only the accusative 'caki-lul'. The reason is that other frequent forms of reflexives, i.e. nominative and genitive forms, are not suitable for the present study for the following reasons. Nominative forms are used as subjects and subjects usually do not take an antecedent in the same sentence due to the c-command constraints, so that there cannot be a local antecedent in a strict sense. Genitive forms are part of NPs, which are treated as local domains in some theoretical persuasions. In such a theory, a genitive NP is a kind of

subject and this cannot take a local antecedent at all. To avoid such problems I take only 'caki-lul' as the object of study.

I used 10-million-word KOREA-1 corpus but out of the KWIC examples obtained from it, I used only half of them (by taking alternate examples), so that the figures reported in this paper is as good as based on a 5-million-word corpus. Since 'caki' has different meanings of 'porcelain' and 'magnetism' (i.e. there are homonyms) , I had to examine every instance and exclude irrelevant examples. The final result is 316 instances of the use of reflexive 'caki-lul'.

Some authentic examples of local antecedents and long-distance antecedents follow. First, local antecedents:

(10) ... kyewunay elum sokeyse chengcenghakey [cakilul] kakkwuessten
kolccakiuy malkun mwulimye...
'... transparent water in the gorges that preserved itself intact under ice ...'

(11) thum issnun taylo ilkese kamsanghamye tewuk [cakilul] katatumnun
ilto kwuicwunghan ilita.
'It is important for you to read books often and prepare yourself.'

The first example is a case of relative clause.

Second, long-distance antecedents:

(12) kuttay kunun ssikssikhamye [cakilul] kkyeana cwuten tekhoka tteolunta.
'At that time, she remembered Tekho, who embraced her while breathing hard.'

(13) phaymi [cakilul] way ttalawassnunyako cimeykey mwutnunta.
'Pam asks Jim why he followed her.'

In the first example, the antecedent of 'caki' is not 'Tekho', which is the head of the relative clause; in the second example, the antecedent is 'Paym', which is not a local subject.

The manual examination of each instance of 'caki-lul' gives the following figures:

(14) Antecedents of 'caki-lul'
local: 151 long-distance: 165 total: 316

These figures show that 'caki' is used almost equally with local and long-distance antecedents, although there is a slight difference (52% : 48%). In a sense, such a neutral characteristic of 'caki' caused opposing views on its antecedents.

When we compare the use of 'caki' with that of other Korean reflexives 'caki' and 'cakicasin', a clearer picture of the use of 'caki' emerges.

3. THE USE OF REFLEXIVES: 'CAKI', 'CASIN', AND 'CAKICASIN'

Up until now, 'casin' and 'cakicasin' have attracted less attention than 'caki' from Korean grammarians. Many grammarians assume that descriptions of 'caki' can be extended to these other forms of reflexives in Korean. Sung (1981) and Im (1987) are among the exceptions.

Sung (1981) compares 'caki' and 'casin' and states the differences as follows:

(15) a. 'casin' tends to take a local antecedent.
b. 'caki' does not take a first person antecedent.
c. For local antecedents, 'casin' is used more often.
d. 'casin' can be used for emphatic purposes, as in 'John casin'.

The second (b) and the fourth (d) points are not controversial. But the first and third points (a,c) can cause some controversy. When we consider sentences with reflexives in isolation, without any context, it seems hard to verify these claims intuitively. For the following sentence, I cannot tell which antecedent is better for 'casin', 'Pyengswu' or 'Cengswu' , although Sung (1981) claims that 'casin' is better than 'caki'.

(16) cengswunun cakilul/casinul miwehanta.
Cengswu-Top self-Acc dislikes
'Cengswu dislikes himself.'

It seems that even though there is no difference in grammaticality judgements, the latter is more natural than the former. But I do not think that such a judgement is applicable to every case. Rather, I will show later that the difference in naturalness comes from the meaning of the verb 'miwehata'("to hate"). As Barlow (1996) shows, (English) reflexives also tend to occur more often with some specific verbs than with other verbs.

Im (1987) claims that one of the semantic differences between 'caki' and 'casin' is that 'caki' is [+conscious] but 'casin' is [-conscious]. It is why the difference in grammaticality of the following sentences results, he argues. The situation where the sentence is used is that Chelswu kicks the statue of him. Notice that kicking is a conscious act.

(17) chelswuka cakilul/*casinul pallo chassta.
Chelswu-Nom self-Acc with-foot kicked
'Chelswu kicked himself (with his foot).'

Grammaticality judgements are his, but I think that the difference is not a matter of grammaticality but a matter of preference or naturalness. This point becomes clearer in a situation where Chelswu exploded the statue of him. In the following examples, the first sentence is perfect and the third sentence is totally unacceptable, but the second is clearly different from either of the two.

(18) a. chelswuka cakilul phokphahayssta.
Chelswu-Nom self-Acc exploded
'Chleswu exploded himself.'
b. ?chelswuka casinul phokphahayssta.
Chelswu-Nom self-Acc exploded
c. *chelswuka caphokhayssta.
Chelswu-Nom self-exploded

In a situation where he really exploded himself by exploding a bomb around his waist, we can use any of the three sentences. Usually strict reflexivity is required of morphological reflexives such as 'caphokhata' (cf. Lidz 1997, Reinhart and Reuland 1993).

In the remainder of this paper, by means of examining corpus data, I will show that most of the differences between 'caki' and 'casin' are not a matter of grammaticality but a matter of naturalness or preference and that some of the differences are due to the kinds of verbs that take reflexives as objects.

Not only 'caki' and 'casin', I also examined 'cakicasin', a reflexives which is less often used than the other two. The result is as follows:

(19) Antecedents of Reflexives

	caki-lul	casin-ul	cakicasin-ul
local	151	311	66
long-distance	165	123	5
total	316	434	71

(cakilul-casinul: X2 = 44.062, p<0.001)

What this table shows is as follows:

(20) a. The Korean reflexives (accusative) are used in the following order of frequency.
'casin-ul' > 'caki-lul' > 'cakicasin-ul'.
b. 'cakicasin-ul' is much less used than the other two.
c. Both 'caki-lul' and 'casin-ul' can take local and long-distance antecedents freely.
d. 'caki-lul' takes local and long-distance antecedents equally well.
e. 'casin-ul' takes local antecedents much more than long-distance ones.
f. 'cakicasin-ul' takes local antecedents almost exclusively.

The most remarkable fact is that 'casin' takes both local and long-distance antecedents but much more local antecedents. Figure 1 shows the difference of the three kinds of reflexives more clearly.

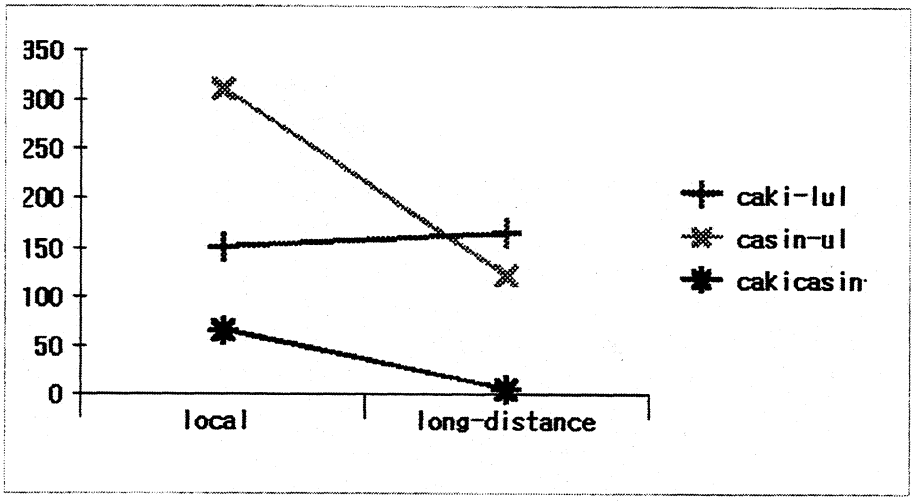

Figure 1 Local and long-distance antecedents of reflexives

In a sense, this finding supports Sung's (1981) and Im's (1987) claims that 'casin' tends to be used with local antecedents. But notice that a quarter of the occurrences of 'casin' are with long-distance antecedents. The figures obtained from corpus data show this tendency in a quantitative way and on a more firm basis.

Putting aside the question of local/long-distance antecedents, let us consider semantic differences among reflexives. This problem can be taken up by examining what kinds of verbs are used with reflexives as objects. Here is the research strategy:

(21) When (accusative forms of) reflexives are used with local antecedents, we examine the thematic roles of the antecedent, which is in fact the subject of the (transitive) verbs used.

In concrete, I classified the verbs into two big categories: verbs with Agent subjects and verbs with Experiencer subjects. Here, Experiencer verbs include not only verbs of passive experience such as 'alta' ("to know") and 'insikhata' ("to recognize") but also verbs of more active experience such as 'pota' ("to see") and 'palkyenhata' ("to discover"). Generally, this class includes any verbs of perception and verbs of (change of) mental states.

The number of kinds of Agent verbs found in the corpus is 180 in total and that of Experiencer verbs is 58 in total. The numbers of the occurrences of reflexives with each kind of verbs are shown below.

(22) Theta Roles of Local Antecedents

	caki-lul	casin-ul	cakicasin-ul
Agent antecedent	107	196	33
Experiencer ant.	36	108	29

(cakilul - casinul: $X2=4.472$, $p < 0.05$)

What these figures show is as follows:

(23) a. Both 'caki' and 'casin' have more Agent antecedents than Experiencer antecedents.
b. Contrasted with 'caki', 'casin' takes relatively more Experiencer antecedents. ($X^2=4.472$, $p < 0.05$)

The graph in Figure 2 shows the differences more clearly.

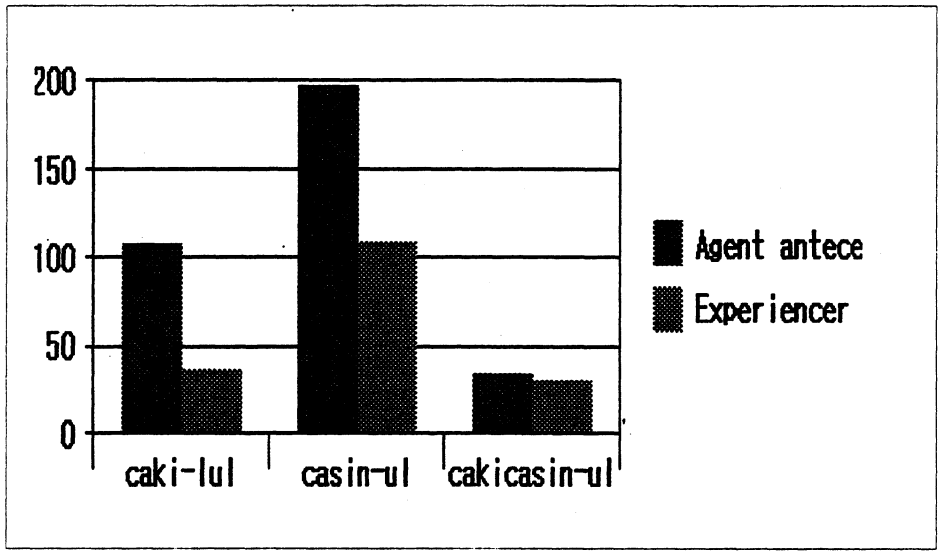

Figure 2 Thematic roles of antecedents of reflexives

The findings show that 'caki' and 'casin' have some semantic differences but they are not absolute but only relative. Compared with 'caki', 'casin' takes relatively more Experiencer antecedents. However, in absolute numbers of occurrence, 'casin' takes more Agent antecedents than Experiencer antecedents, and 'caki' not only takes Agent antecedents but also takes Experiencer antecedents. For example, although we can say that 'caki-lul' is more natural than 'casin-ul' as an object of the verb 'chata' ("to kick"), and 'casin-ul' is more natural than 'caki-lul' as the object of the verb 'miwehata' ("to hate"), we cannot say that the opposite cases result in an absolute ungrammaticality.

Another controversy about different kinds of reflexives is the one regarding their status as nouns. Im (1987) classified 'caki' as a pronoun and 'casin' as a (common) noun for the following reason. In Korean it is possible to have a construction of "pronoun + noun", but "pronoun + pronoun" construction is not possible, usually when the pronoun is understood as a possessor.

(24) a. wuli aki
we baby 'our baby'
b. *wuli ku / *wuli kunye
we he we she

'casin' patterns with a noun but 'caki' patterns with a pronoun.

(25) a. wuli casin
we self
b. *wuli caki
we self

So it seems that 'caki' is more like a pronoun and 'casin' is more like a noun. However, when we consider the construction of "caki/casin + noun", the situation becomes reversed: 'caki + noun' is possible but 'casin + noun' is not.

(26) cakipiphan 'self-criticism', cakipwuceng 'self-denial'
cakipocon 'self-preservation', ...
(27) *casinpiphan, *casinpwuceng, *casinpocon, ...

Since "pronoun + noun" is not possible as follows, we may claim that 'casin' is more like a noun than 'caki' in this respect.

(28) *wulipiphan / *kupiphan / *nepiphan
we-criticism he-criticism you-criticism

Therefore, the cases of compounding do not help us to judge which of 'caki' and 'casin' is more noun-like. Examining the use of reflexives sheds some light on this issue. Generally, nouns tend to be modified freely ('a beautiful flower', 'the boy who I met yesterday', etc), the modification of pronouns is not usual since pronouns are used to refer to discourse referents directly. Of course, nonrestrictive modification is possible for pronouns as well as for proper names: 'clever I', 'beautiful Mary'. Therefore, although modification is not impossible with pronouns, we can say that common nouns are modified much more often than pronouns. Regarding 'caki' and 'casin', if one of them is used with more modification than the other, we can say that the former is more noun-like. Without extra constraints, modification is possible for both:

(29) sihemey silphayhan caki/casin
in-the-exam failed-Rel self
'self who failed in the exam'

The examination of the corpus data gives the following figures. The result is based on the examination of both local and long-distance reflexives and it shows that 'casin' is more modified than 'caki'. This in turn suggests that 'casin' is more noun-like than 'caki' in this respect.

(30) Modification of reflexives

	caki-lul	casin-ul	cakicasin-ul
modified	13	50	0
unmodified	303	384	71

(cakilul-casinul $X2=13.038$, $p < 0.001$)

4. CONCLUSIONS

In this paper, we examined the aspects of the use of Korean reflexives 'caki', 'casin', and 'cakicasin' on the basis of a Korean corpus and discussed how the findings contribute to the grammar of Korean reflexives. The three kinds of Korean reflexives are not so different as to show differences in grammaticality. It is just a matter of preference and naturalness, and the corpus data show that this difference in naturalness is real.

Let us return to the question raised in the first part of the paper, namely the relationship between grammar and the use of language. In other words, what does the use of language tell us about grammar as the knowledge about language?

Linguistic knowledge is implicit and some kind of such knowledge can be sought after by soliciting intuition about grammaticality from native speakers. But some knowledge cannot be sought after in such a manner. Intuition about degrees of naturalness is not clear and it can only be justified on the basis of observable linguistic data, i.e. a corpus. Frequency data provide the ground for quantification of such degrees of naturalness. In a sense, I claim that a balanced corpus reflects speakers' unconscious intuition collectively and the examination of the use of three kinds of Korean reflexives is a step toward proving such a claim. Or, at the least, the result of this investigation will provide a solid base on which further theorizing may proceed.

ACKNOWLEDGEMENTS

The research reported in this paper is supported in part by the faculty research fund of Korea University.

REFERENCES

Barlow, M. (1996) "Corpora for Theory and Practice," in International Journal of Corpus Linguistics 1-1, 1-37.

Chang, S.J. (1986) "Discourse Function of Anaphora: Reflexives," in Hangul 194, 121-155.

Church, K., W. Gale,, P. Hanks, and D. Hindle (1991) "Using Statistics in Lexical Analysis", in U. Zernik (ed.) Lexical Acquisition, Hillside: LEA, 115-164.

Im, H.B. (1987) A Study on Korean Reflexives, Seoul: Shingumunwhasa. [written in Korean]

Kang, B. (1988) "Unbounded Reflexives", in Linguistics and Philosophy 11-4, 415-456.

Kim, H.G. and B.M. Kang (1996) "KOREA-1 Corpus: Design and Composition", Korean Linguistics 3, 233-258. [written in Korean]

Kučera, H. and W.N. Francis (1967) Computational Analysis of Present-Day American English, Providence: Brown University Press.

Lidz, J. (1997) "Anti-antilocality", presented at Long-Distance Reflexives Workshop, Cornell University.

Moon, S.C. (1996) "'Optimality in Anaphoric Binding", handout for a presentation at Korean Society for Language and Information, Nov. 1996.

Reinhart, T. and E. Reuland (1993) "Reflexivity", in Linguistic Inquiry 24, 657-720.

Sinclair, J. (ed.) (1987) Looking Up, London: Collins.

Suh, J.S. (1996) Korean Grammar, Seoul: Hanyang Univ. Press. [written in Korean]

Sung, K.S. (1981) "Another Discussion on Korean Reflexives - 'caki' and 'casin' -", in Hangul 172, 29-55 [written in Korean].