Why Korean Is Not a Regular Language: A Proof

Yongkyoon No*
Chungnam National University

Yongkyoon No. 2001. Why Korean Is Not a Regular Language: A Proof. *Language and Information 5.2*, 1–8. Natural language string sets are known to require a grammar with a generative capacity slightly beyond that of Context Free Grammars. Proofs regarding complexity of natural language have involved particular properties of languages like English, Swiss German and Bambara. While it is not very difficult to prove that Korean is more complex than the simplest of the many infinite sets, no proof has been given of this in the literature.
I identify two types of center embedding in Korean and use them in proving that Korean is not a regular set, i.e. that no FSA's can recognize its string set. The regular language **i salam i (i salam ul)^j michi (key ha)^k essta** is intersected with Korean, to give {**i salam i (i salam ul)^j michi (key ha)^k essta** | $j, k \geq 0$ **and** $j \leq k$}. This latter language is proved to be nonregular. As the class of regular sets is closed under intersection, Korean cannot be regular. **(Chungnam National University)**

1. Introduction

A natural language is, at a level of abstraction, simply a set of all grammatical sentences in that language. As this set is an infinite one, its specification cannot be completed by enumeration. It has to be done with rules and a fundamental question in linguistics has for some time been what type of rules is needed for natural language.

There has been copious work done on this issue of "complexity" in the last few decades of the last century. Most linguists seem to agree that natural language string sets require a grammar slightly more powerful than a Context Free Grammar. That Context Free Grammars are not sufficient for the specification of the string set of natural languages comes from languages such as Swiss German (Shieber, 1985) and Bambara (Culy, 1985). Evidence, albeit not as solid, also comes from English and Dutch.

In light of the current understanding of the issue of complexity, it would be expected that Korean would require a grammar of a comparable generative capacity. It would be surprising if Korean turned out a substantially less complex set. What I set out to do is determine the position of Korean in the complexity hierarchy. My first step to this end is answering the question: why is Korean beyond the weak generative capacity of regular grammars?

Chomsky and Miller (1958) ask exactly the same question about English. They show that English is not regular (i.e., not describable by a regular grammar) with a construction they call "center embedding". The strings in (1), though not all of them are totally acceptable, must be regarded as grammatical in any reasonable description of the English

* Department of Linguistics, Chungnam National University, 220 Koong dong, Yuseong gu, Taejeon 305-764. E-mail: yno@linguist.cnu.ac.kr

language.

(1) a. The cheese that the rat ate stank.

b. ?The cheese that the rat that the cat saw ate stank.

c. ??The cheese that the rat that the cat that the dog chased saw ate stank.

The existence of this very construction, i.e., the relative clause with a gap in a non-subject position, coupled with the fact that a relative clause occurs after its head noun, provides the evidence that English is not a regular set. The reader is referred to their original work for a detailed proof.

My strategy is the same. What has been missing in the literature is the part of Korean syntax which is associated with this property. Which constructions reveal center embedding in Korean? In section 2, I will describe two constructions that are center embedding constructions in Korean. A rigorous proof will be given in section 3 that the string set of Korean is not a regular set. Some suggestions will be made for future research in the last section.

2. Two center embedding constructions in Korean

Korean certainly has relative clauses. It allows relativization on non-subjects as well as one on subjects. However, relative clauses do not give rise to center embedding constructions in Korean. This is due to the opposite headedness of Korean relative to English in the placement of a relative clause with respect to its head noun: a relative clause comes immediately before its head nominal in Korean.

2.1 Thinking backwards from the grammar

We have to look elsewhere for a possible source of center embedding in the language. Despite the fact that what this article is concerned with is the mathematical properties of the language as a string set and it is not assuming any particular grammatical rules, I see it fit to show how one can arrive at the crucial constructions by thinking deductively, i.e. thinking backwards from known aspects of the grammar.[1] The first type of center-embedding constructions that comes to mind must involve a PS rule which has a recursive category X in its left hand side and has the same node X some place other than the two edges in its right hand side. This rule must have at least three nodes in its right hand side. As there is no PS rule in Korean that is widely accepted as involving more than three nodes, the rule may well have no more than three nodes in its right hand side. In that case, it must look like (2).

(2) $X \longrightarrow A\ X\ B$

As X is a recursive category, it could be one of $\overline{N}$, S, and VP. The recursive rule involving $\overline{N}$ introduces a single node as its modifier. Hence, $\overline{N}$ does not fit X's slot in (2). S, however, may fit X's slot in (2). In such a case, B would have to be a verb and A, a phrase for the subject of the sentence. As PP's headed by the nominative postposition

1. I am grateful to a referee of this journal for pointing this out to me. The previous version of this article was unclear about the role of the PS rules in this subsection with respect to the proof it purports to give.

may mark the subject in this language, the PS rule in (3) would underlie one type of center embedding constructions.[2]

(3) S ⟶ PP S V

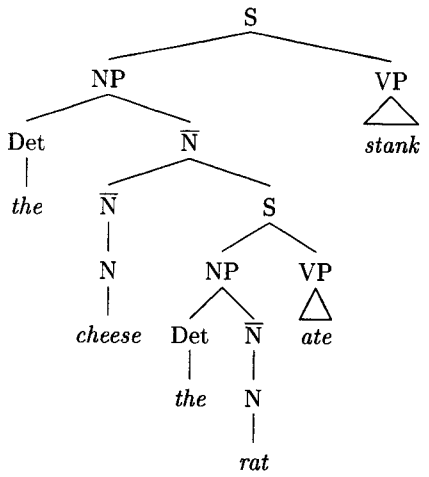
In addition, a VP may also fit X's slot in (2), in which case B would have to be a verb and A, an accusative PP. The rule in (4) underlies another type of center embedding constructions.

(4) VP ⟶ PP VP V

The two rules above, i.e., those in (3) and (4), have a lot in common. Each is a recursive rule in its own right: the node to the left of the arrow appears again in the right hand side of the arrow. In addition, the recursive node appears at a place other than an edge in the local tree. Each recursive node in (3) and (4) is flanked by a PP on its left and by a V on its right. I will call this type of center-embedding "nonedge mono-self-embedding".

The other type of center embedding constructions arises as we concentrate on formal properties of the relative clause construction with a nonsubject gap. The configuration behind the English center embedding construction(s) can be characterized: a phrasal node X can occur at one edge of another node Y which occurs at the other edge of Y's ancestor node Z. Confirm this with the following tree diagram.

(5)



The embedded S in (5) appears at the right edge of its ancestor node NP, which appears at the left edge of its ancestor node, the root S. The rule in (3) might be taken as embodying an incorrect analysis of a construction that is better analyzed with the following set of rules.

(6) S ⟶ PP VP
VP ⟶ S V

2. Readers who take the nominative and accusative markers of the language as affixes rather than as (clitic) words may susbstitute "noun phrase with the nominative marker" for my "postpositional phrase headed by the nominative postposition". The decision does not affect anything in this article.

Here, the node S appears at the left edge of its ancestor node VP, which in turn appears at the right edge of its ancestor node S. The relevant configurational property of the Korean construction described with the rules in (6) is exactly the same as that of the center-embedding construction in English. I will call that type of center-embedding involving this property an "Edge-flip type of self-embedding."

A terminological clarification is in order here. What I call "nonedge mono-self-embedding" is a special case of direct recursion. Note that there are cases of direct recursion that are not of "nonedge mono-self-embedding". A rule system which has the property of recursion which consists of a single rule may of course fail to reveal nonedge mono-self-embedding. Such rule systems would typically consist of a left recursive or right recursive rule. Similarly, what I call "Edge-flip type of self-embedding" is a special case of indirect recursion. Many rule systems that are indirectly recursive exhibit self-embedding of a nonedge type.[3]

2.2 Nonedge mono-self-embedding

So-called Periphrastic Causative construction falls under this group. The verbs *mantul* and *ha* enter into it. With the verb *ha*, we get sentences in (8).

(7) i salam i michi -essta.
this person NOM become insane PAST DECL
'This person became insane.'

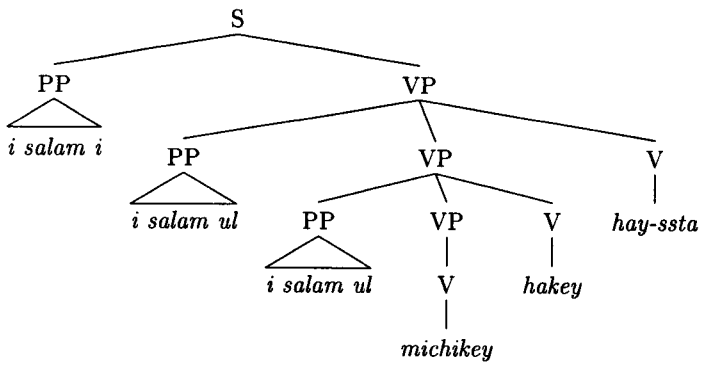
(8) a. i salam i i salam ul michi-key hay-ssta.
'This person made this person become insane.'

b. ?i salam i i salam ul i salam ul michi-key ha-key hay-ssta.
'This person made this person make this person insane.'

c. ??i salam i i salam ul i salam ul i salam ul michi-key ha-key ha-key hay-ssta.
'This person made this person make this person make this person insane.'

The sentences in (8) would all be given an analysis, in a standard PSG description of the language, involving the self-embedding rule in (4). Under such a grammar, sentence (8b) would be analyzed:

(9)



3. A referee wrongly suggested that my distinction of the two types is the same as the already available distinction between direct and indirect recursions. I hope this paragraph helps him/her clarify the difference.

2.3 Edge flip type

The second type is illustrated with a verb of saying or a propositional attitude verb: *ha* 'say' and *mit* 'believe'. A special verb group consisting of the processual noun *sayngkak* 'thought process' and the light verb *ha* also illustrates it.

(10) i salam i ttena -ssta.
this person NOM leave PAST DECL
'This person left.'

(11) a. i salam i i salam i ttena-sstako sayngkak hay-ssta.
'This person thought this person left.'

b. ?i salam i i salam i i salam i ttena-sstako sayngkak hay-sstako sayngkak hay-ssta.
'This person thought this person thought this person left.'

c. ??i salam i i salam i i salam i i salam i ttena-sstako sayngkak hay-sstako sayngkak hay-sstako sayngkak hay-ssta.
'This person thought this person thought this person thought this person left.'

3. The proof

Let L_1 be the language described by the regular expression **i salam i (i salam ul)^j michi (key ha)^k essta**. Intersect L_1 with Korean to get the language L[4].

(12) $L = L_1 \cap$ Korean

What do we get when L_1 is intersected with Korean? The answer depends on the grammaticality judgment of some crucial strings of Korean words. Consider the following.

(13) a. i salam i michi-key hay-ssta.
'This person made x become insane.'

b. ?i salam i i salam ul michi-key ha-key hay-ssta.
'This person made this person make x insane.' or
'This person made x make this person insane.'

c. ??i salam i i salam ul i salam ul michi-key ha-key ha-key hay-ssta.
'This person made this person make this person make x insane.'
'This person made this person make x make this person insane.'
'This person made x make this person make this person insane.'

4. It has been suggested by a referee that (11) had better be replaced with a simpler-sounding "Korean contains L". This cannot be so for the following reason. (11) means THE CONJUNCTION of the following two statements:
(i) Korean contains L
(ii) Korean does not contain any string in $L_1 - L$.
Without showing (ii) in addition to showing (i), the proof would become inadequate. Note that the well-known context-free language $a^n b^n$ is a proper subset of the regular language $a^m b^n$ and hence showing that some language contains the former in itself has nothing to do with showing that it is non-regular. This is exactly one of the formal slips Postal (1962) makes in his unsuccessful attempt to show that Mahawk is non-context-free. See Pullum and Gazdar (1982).

Virtually any noun phrase, postpositional phrase with the nominative postposition, or one with the accusative postposition can be suppressed in Korean. A zero in Korean can play the role an intersentential pronoun plays in other languages. This is why all the strings in (13) are to be taken as grammatical. In fact, (13a) is as acceptable as any other Korean sentences and (13b), with either reading, is quite acceptable. Any degree of clumsiness that could be attributed to (13b) and (13c) would be subject to a systematic explanation of a psycholinguistic nature. For this subject, the reader is referred to Johnson (1998) and Morrill (2000).

The next set of strings to be considered before we establish L, namely the intersection of L_1 and Korean, involves throwing in more occurrences of *i salam ul* in the grammatical sentences of (7) and (8).

(14) a. *i salam i i salam ul michi-essta.

b. *i salam i i salam ul i salam ul michi-key ha-essta.

c. *i salam i i salam ul i salam ul i salam ul michi-key ha-key hay-ssta.

A principle similar to what Chomsky (1981) calls "θ-criterion" is at work in Korean and a grammatical sentence cannot contain more accusative PP's than there are transitive verbs in it. (14a) has no transitive verbs and there being an accusative PP in it renders it ungrammatical; (14b) has one transitive verb and there being two accusative PP's makes it ungrammatical.

We can now turn to our original question: what do we get when L_1 is intersected with Korean? The answer seems to be as in (15).

(15) $L = \{\text{i salam i (i salam ul)}^j \text{ michi (key ha)}^k \text{ essta} \mid j, k \geq 0 \text{ and } j \leq k\}$

In order to prove that Korean is not a regular set, I need to prove that L is not regular. In proving that L is not regular, I first prove that a slightly simpler looking language, $\{a^j b^k \mid j, k \geq 1 \text{ and } j < k\}$ is not regular, namely, (16). Then, I will show that this language is a homomorphic image of our language L.

(16) $\{a^j b^k \mid j, k \geq 1 \text{ and } j < k\}$ is not regular.

As is ususal in proving the nonregularness of an infinite language like the one in (16), I will rely on the pumping lemma.

Lemma 1

(Hopcroft and Ullman, 1979, page 56) Let L be a regular set. Then there is a constant n such that if z is any word in L, and $|z| \geq n$, we may write $z = uvw$ in such a way that $|uv| \leq n, |v| \geq 1$, and for all $i \geq 0, uv^i w$ is in L. Furthermore, n is no greater than the number of states of the smallest FA accepting L.

Suppose $\{a^j b^k \mid j, k \geq 1 \text{ and } j < k\}$ in (16) were regular, and let n be the integer in the pumping lemma and consider $a^n b^{n+i} (i \geq 0)$, whose length is obviously greater than n. We may thus write $a^n b^{n+i} = uvw$ for some strings u, v, and w with $|v| \geq 1$ and $|uv| \leq n$. Let us concentrate on v, especially on the fact that $|uv| \leq n$ forces v to consist of only a's. Suppose that $v = a^s$ for $s \geq 1$. Then, if $u = a^r$, we have $w = a^{n-(r+s)} b^{n+i}$. It follows that $uv^{n+i+2}w = a^r a^{(n+i+2)s} a^{n-(r+s)} b^{n+i} = a^{(n+i+1)s+n} b^{n+i}$, which is not in $\{a^j b^k \mid j, k \geq 1 \text{ and } j < k\}$. Thus $\{a^j b^k \mid j, k \geq 1 \text{ and } j < k\}$ cannot be regular since it fails to satisfy the pumping lemma.

Now that (16) has been proved, we are almost through in the immediate task of proving that the language L in (15) is not regular. We need to show that a very special sort of an onto function, called homomorphism, exists from L to $\{a^j b^k | \ j,k \geq 1 \text{ and } j < k\}$. Such a function does indeed exist and is defined as in (17).

(17) h(i salam i) = a h(key ha) = b
h(i salam ul) = a h(essta) = b
h(michi) = b

The class of regular sets is closed under homomorphisms. See Hopcroft and Ullman (1979, page 61). If L were regular, its homomorphic image, $h(L)$ would also be regular. As $\{a^j b^k | \ j,k \geq 1 \text{ and } j < k\}$, which is $h(L)$, is not regular, L could not be either.

Now we can return to the main theme of this paper, namely, proving that Korean is not regular. Recall that L is the intersection of Korean and the regular language **i salam i (i salam ul)^j michi (key ha)^k essta.**

The class of regular sets is closed under intersection (Hopcroft and Ullman, 1979, page 59). L would be regular if Korean were regular, since L_1 is regular. L is not regular, as shown in the preceding paragraphs. Hence, Korean is not regular. This ends the proof.

4. Conclusion

The string set of Korean is not regular. I proved this by identifying center embedding constructions in the language. The regular language **{i salam i (i salam ul)^j michi (key ha)^k essta** $| \ j,k \geq 0\}$ is intersected with Korean, giving L_1 (=**{i salam i (i salam ul)^j michi (key ha)^k essta** $| \ j,k \geq 0$ **and** $j \leq k\}$). That the latter is not a regular language is shown by: (i) the fact that the class of regular languages is closed under homomorphisms, (ii) showing that there is a homomorphism from L_1 onto $\{a^j b^k | \ j,k \geq 1 \text{ and } j < k\}$, and (iii) proving, relying on the pumping lemma, that $\{a^j b^k | \ j,k \geq 1 \text{ and } j < k\}$ is not regular. By *reductio ad absurdum*, Korean is not regular.

Along the way, two types of center embedding were identified. The monadic non-edge self-embedding type is the one related to the periphrastic causatives of multiplely indirect nature; the edge-flip type is the one for English relative clauses with nonsubject gap and for Korean sentences denoting "nested belief" or multiplely indirect quotations. Korean is a language in which major argument phrases may fail to appear: it provides a bit of complexity to the proof.

While not surprising at all, the findings in this paper justify the common practice of adopting Context Free grammars in describing the syntax of the Korean language. It is yet to be considered whether the language as a set of strings, has properties that cannot be captured with a device as powerful as CFG's.

References

Chomsky, Noam. 1981. *Lectures on Government and Binding.* Foris Publication, Dordrecht. Now published by Mouton de Gruyter.

Chomsky, Noam and George A. Miller. 1958. Finite state languages. *Information and control,* 1(2).

Culy, Christopher. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy,* 8:345–351.

Hopcroft, John E. and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley Publishing Company, Reading, Massachusetts.

Johnson, Mark. 1998. Proof nets and the complexity of processing center embedded constructions. *Journal of Logic, Language and Information,* 7(4):433–447.

Morrill, Glyn. 2000. Incremental processing and acceptability. *Computational Linguistics*, 26(3):319–338.

Postal, Paul. 1962. *Some syntactic rules in Mahawk*. Ph.D. thesis, Yale University. Published by Garland, New York, 1979.

Pullum, Geoffrey K. and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343.

Submitted on: April 23, 2001
Accepted on: August 2, 2001