A Constrained Finite-State Morphotactics for Korean

Eunsok Ju
Department of Language and Information
Yonsei University
ju@lex.yonsei.ac.kr

Chongwon Park
Department of Composition/English
University of Minnesota Duluth
cpark2@d.umn.edu

Minhaeng Lee
Department of Language and Information
Yonsei University
leemh@yonsei.ac.kr

Kiyong Lee
Department of Linguistics
Korea University
klee@korea.ac.kr

Abstract

In this paper, we propose a constrained finite-state model, named **cfsm**, for Korean morphotactics and attempt to show how it can successfully treat some major morphological problems in Korean. As a preliminary descriptive framework, this model adopts the Korean morphological system **Komor** by Lee (1999) to lay out some basic problems in Korean morphotactics and describe linear approaches to their possible solutions. This descriptive step is then followed by various testing steps executed by using Xerox's finite-state development tools, namely **xfst** for creating finite-state networks and **lexc** specifying natural language lexicons. With **Komor**'s constraints represented in feature structures and appropriately implmented into **xfst** and by making **Komor** run on **xfst**, the proposed **cfsm** is expected to fully benefit from the descriptive groundwork of **Komor** and the finite-state processing power of **xfst**.

1. Aim and Approach

This work aims at developing a constraint-based finite-state model named **cfsm** for Korean morphology. For this purpose, it re-implements Lee's (1999) Korean morphology system **Komor,** which had been implemented with a C-augmented grammar tools named MALAGA[1], by using Xerox's finite-state development tools, namely **xfst** for creating finite-state networks and **lexc** specifying natural language lexicons. With **Komor**'s constraints represented in feature structures and appropriately implmented into **xfst** and by making **Komor** run on **xfst**, the proposed **cfsm** is expected to fully benefit from the descriptive groundwork of **Komor** and the finite-state processing power of **xfst**.

In designing and implementing our proposed model **cfsm**, we strictly adhere to the principle of possible continuation that has been advocated by Hausser (1989) and Beesley and Karttunen (2003) in recent years. The main operation in executing **cfsm** is concatenation (without backtracking), but strictly constrained by some requirement conditions on feature-value matching. In Korean, for instance, noun and verbal stems concatenate with a sequence of their suffixal particles or endings to form well-formed word forms. But this concatenation is often constrained by their particular syllable structure or degree of regularity. For example, a verbal stem '먹 mek' may simply concatenate with a tense marker to form another stem, but its syllable structure may restrict the choice of its ending, thus allowing the stem '먹었 mek.ess' only. This linear approach with necessary constraints is, however, considered adequate especially in treating the morphology of agglutinative languages like Korean.

[1] MALAGA is an acronym for "Malaga accepts left-associative grammars with attributes", developed by Bjoern Beutel and others at Department of Computational Linguistics, Erlangen-Nuernberg University.

2. Design and Representation

The proposed model **cfsm** consists of three main modules: (1) Lexicon, (2) Grammar, and (3) Output.

2.1. Lexicon

The Lexicon again consists of three sub-lexicons: (1) Basic Lexicon, (2) Extended Lexicon with allomorphic variations, and (3) Enlarged Lexicon augmented with syllable-complete surface forms. These sub-lexicons are constructed semi-automatically by a preprocessing engine implemented preferably by Java that provides a list of multi-character symbols, a set of lexical definitions and concatenation rules for the execution of Xerox's morphotactic engines[2].

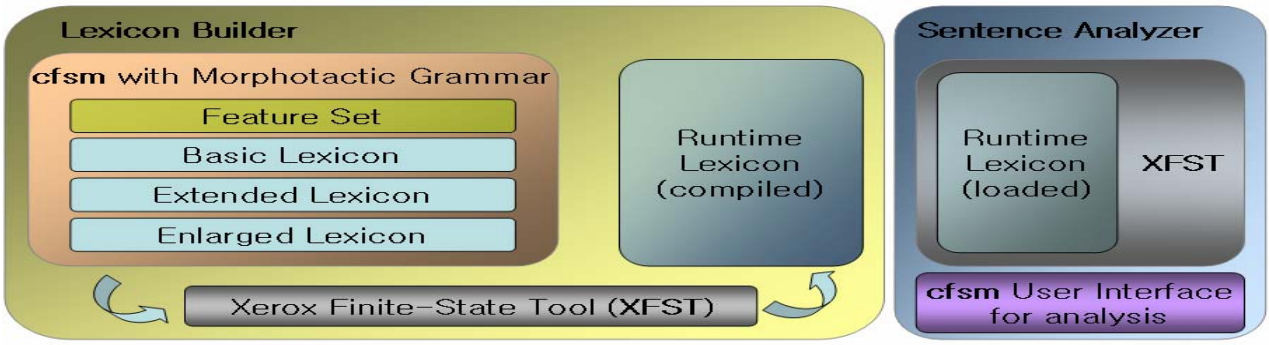

Figure 1: cfsm for Korean

Each lexical item listed in these sub-lexicons is associated with an appropriate feature structure which is represented in a feature-value matrix or an equivalent XML format. For illustration, consider a list of well-formed word forms in Korean like[3]:

(1) a. 춥다 chwup.ta, (be cold)
b. 추웠다 chwu.wess.ta, (was/were cold)

The feature structure associated with each of these word forms can be represented in a matrix form as follows:

(2) a.
$$\begin{bmatrix} \text{SURF} & \text{'춥다 chwup.ta'} \\ \text{SEM} & \begin{bmatrix} \text{CONTENT} & \langle \text{state}, \text{'being\_cold'} \rangle \end{bmatrix} \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{SURF} & \text{'추웠다 chwu.wess.ta'} \\ \text{SEM} & \begin{bmatrix} \text{CONTENT} & \langle \text{state}, \text{'being\_cold'} \rangle \\ \text{TENSE} & \text{past} \end{bmatrix} \end{bmatrix}$$

2.1.1. Basic Lexicon

The Basic Lexicon lists basic lexical units, or morphemes, for instance like verbal stems and their suffixal endings with minimally necessary information.

(3) a.
$$\begin{bmatrix} \text{SURF} & \text{'춥 chwup'} \\ \text{POS} & \text{adjective} \\ \text{FORM} & \text{p-class} \\ \text{SEM} & \begin{bmatrix} \text{CONTENT} & \langle \text{state}, \text{'being\_cold'} \rangle \end{bmatrix} \end{bmatrix}$$

[2] We use the non-commercial version of the Xerox engines that are provided with Beesley and Karttunen (2003).
[3] The Yale romanization of Hangul is adopted as is done by most Korean linguists. See Lee (1994).

b.
```
[SURF 'ㅆ ss'
 POS verbalEnding
 TYPE prefinal
 REQUIRES [BRIDGE null]
 SEM [TENSE past]]
```

c.
```
[SURF 'ㄼ keyss'
 POS verbalEnding
 TYPE prefinal
 SYLLABLE closed
 SEM [MODALITY conjectured]]
```

d.
```
[SURF 'ㄴ n'
 POS verbalEnding
 TYPE wordFinal
 RESULTS [POS adnoun]]
```

e.
```
[SURF '다 ta'
 POS verbEnding
 TYPE sentenceFinal
 SEM [SENTENCETYPE declarative]]
```

2.1.1. Extended Lexicon for Allomorphic Variations and Constraints

By a set of allomorphic rules, the Basic Lexicon is then automatically extended to the Extended Lexicon that is designed to contain all of the possible allomorphic forms. Since the adjectival stem '춥 chwup' belongs to the p-variation class of semi-irregularity, the Extended Lexicon should contain the following two items:

1. Allomorphic Extension: Adjectival Stems

(4) a.
```
[SURF '춥 chwup'
 POS adjective
 BASEFORM '춥 chwup'
 SYLLABLE closed&dark
 SEM [CONTENT <state, 'being_cold'>]]
```

b.
```
[SURF '추우 chwu.wu'
 POS verb
 BASEFORM '춥 chwup'
 BRIDGE e
 SYLLABLE open&dark
 SEM [CONTENT <state, 'being_cold'>]]
```

Note above that each allomorph contains information about its syllable features: the syllable analysis of the stem '춥 chwup' shows that its (final) syllable is closed, thus ending in a consonant, and is classed as dark because its vowel is neither 'a' or 'o'. On the other hand, the final syllable of the stem '추우 chwu.wu' is analyzed as open and dark.

Other lexical items also contain information about their syllable structure through syllable analysis, as shown in the following:

2. Allomorphic Extension: Past Tense Markers

(5) a.
```
[SURF '쓰 ss'
 POS verbalEnding
 SYLLABLE closed&incomplete
 REQUIRES [BRIDGE null]
 SEM [TENSE past]]
```

b.
```
[SURF '았 ass'
 POS verbalEnding
 SYLLABLE closed
 REQUIRES [SYLLABLE closed&clear]
 SEM [TENSE past]]
```

c.
```
[SURF '었 ess'
 POS verbalEnding
 SYLLABLE closed
 REQUIRES [SYLLABLE closed&dark]
 SEM [TENSE past]]
```

d.
```
[SURF '였 yess'
 POS verbalEnding
 SYLLABLE closed
 REQUIRES [BRIDGE ye]
 SEM [TENSE past]]
```

e.
```
[SURF 'ㅕ쓰 ess'
 POS verbalEnding
 SYLLABLE closed&incomplete
 REQUIRES [BRIDGE e]
 SEM [TENSE past]]
```

f.
```
[SURF 'ㅣ쓰 iss'
 POS verbalEnding
 SYLLABLE closed&incomplete
 REQUIRES [BRIDGE i]
 SEM [TENSE past]]
```

Six variants are listed as past tense markers in the Extended Lexicon. Each has its own SYLLABLE and REQUIRES conditions for combining with an appropriate stem: the last ‘ㅣ쓰 iss’, for instance, combines with ‘하 ha’ to produce ‘했 hayss’(do+past).

4. Allomorphic Extension: Past Adnominal Endings

(7) a.
```
[SURF 'ㄴ n'
 POS verbalEnding
 FORM wordFinal
 SYLLABLE incomplete
 REQUIRES [SYLLABLE open]
 RESULTS [POS adnoun]]
```

b.
```
[SURF '은 n'
 POS verbalEnding
 FORM wordFinal
 REQUIRES [SYLLABLE closed]
 RESULTS [POS adnoun]]
```

2.1.3. Enlarged Lexicon

This Extended Lexicon is then further augmented to the Enlarged Lexicon in two ways: (1) syllable completion and (2) minimal ordering.[4] Since Hangul characters, say '사랑 sa.rang', can be analyzed into phonemic units, 'ㅅ s', 'ㅏ a', 'ㄹ r', 'ㅏ a', and 'ㅇ ng', or into syllables, '사 sa' and '랑 rang', a choice is necessary between phonemic and syllabic analysis for morphological processing.[5] The **komor** system by Lee (1999) chose the phonemic analysis, thus analyzing '추웠다 cwu.wess.ta' into '추우/ㅓ/ㅆ/다 chwu.wu/e/ss/ta'. As in Shim and Yang (2002, 2004), on the other hand, the present work has taken the alternative way of analyzing characters into complete syllables, thus resulting in a morphological analysis as in: '추웠/다 chwu.wess/ta' instead of '추우/ㅓ/ㅆ/다 chwu.wu/e/ss/ta'.

The second approach reduces the number of the steps of morphological processing. In the given example, that number reduces from three to one. This process, however, requires the pre-processing of surface forms like '추웠 chwu.wess' that combines any syllable-incomplete forms like 'ㅓ ㅆ ess' into syllable-complete forms like '웠 wess'. In our case, the Intermediate Lexicon must thus contain both the forms '추웠 chwu.wess' and '추운 chwu.wun', since both the verbal endings 'ㅓ ㅆ ess' and 'ㄴ n' are syllable-incomplete.

The Enlarged Lexicon can further be enlarged by containing complex nominal particles like '에서부터라도 ey.se.pwu.te.ra.to' (even from) or complex verbal endings like '으시었겠 u.si.ess.keyss) (Bridge+Honorific+Past+Conjectured). These complex particles or endings are found to occur with very low frequency in a very large corpus and to be not worth of being generated by a set of rules. Hence, they can be preprocessed and included in the Enlarged Lexicon. Such a decision reduces the load of ordering rules to a great extent.[6]

2.2. Grammar

By Grammar here is meant a set of morphotactic rules based on the Enlarged Lexicon. The basic, perhaps sole operation in **cfsm** for Korean is concatenation and this operation applies linearly in a left-associative manner to a pair of input surface forms to generate a well-formed new string and eventually a well-formed word form through repeated applications. On the basis of the Enlarged Lexicon devised here, our **cfsm** should be able to analyze or generate the following word forms:

(8) a. 춥다 chwup.ta, (state of being cold)
b. 춥겠다 chwup.keyss.ta (conjectured state of being cold)
c. 추웠다 chwu.wess.ta (past state of being cold)
d. 추웠겠다 chwu.wess.keyss.ta (past conjectured state of being cold)
e. 추운 (cold)chwu.wun (Adnoun, past state of being cold)

[4] Shim and Yang (2002) and Shim and Yang (2004) proposed (1) and Kang (2002) (2), also through personal communication.
[5] See Lee (1994).
[6] This observation was communicated to us by Seung-Shik Kang through personal communication and also in Kang (2002).

2.2.1. Preprocessing Complex Verbal Endings

In **komor** the word form '추웠겠다 chwu.wess.keyss.ta' is analyzed as a series of concatenation as in:

(9) 추우 stem + ㅕ bridge + ㅆ past + 겠 conjecture + 다 sfinal

But **cfsm** analyzes it as:

(10) a. Either 추웠 stem,past + 겠 conjecture + 다 sfinal
b. or 추웠겠 stem,past,conjecture + 다 sfinal

Here we have an option to choose either (a) or (b). Consider the following:

(11) a. 먹었다 mek.ess.ta (eat/Verb+Past)
b. 먹겠다 mek.keyss.ta (eat/Verb+Conjectured)
c. 먹었겠다 mek.keyss.ta (eat/Verb+Past+Conjectured)
d. *먹겠었다 mek.keyss.ess.ta

But this list contains an ill-formed word-form, namely '먹겠었다 mek.keyss.ess.ta'. To eliminate it, **komor** uses an ordering constraint so that the past tense verbal ending '었 ess' should precede the conjectural ending '겠 keyss'.

Another way to solve this problem is to treat a sequence of verbal endings like '었겠' as a complex verbal ending and list them in the Enlarged Lexicon. Such a decision is statistically motivated, as Kang (2002) claims that the preprocessing of such complex verbal endings must increase the efficiency of morphological processing.

Our proposed Enlarged Lexicon will then contain the following complex verbal endings:

Enlarged Lexicon: Complex Verbal Endings

(12) a.
```
SURF '었겠 ess.keyss'
POS verbalEnding
SYLLABLE closed
REQUIRES [SYLLABLE closed&dark]
SEM [TENSE past
     MODALITY conjectured]
```

Since there is no longer a process of concatenating of '었 ess' and '겠 keyss' in our system, it is necessary to pre-generate verbal stems like '추웠겠 chwu.wess.keyss' in the Enlarged Lexicon to obtain well-formed word-forms like '추웠겠다 chwu.wess.keyss.ta'.

2.2.2. Conforming to the Xerox engines

Our engine **cfsm** automatically converts the Enlarged Lexicon to the rule format suitable for executing Xerox's **xfst** and **lexc**. Besides a list of multi-character symbols that are used in defining the rules, it consists of a set of definitions and a sequence of rules called LEXICON. Here is a portion of it obtained from our sample Enlarged Lexicon:

```
Definitions
 adjective =
 [
  춥
  "@P.SURF.'춥 chwup'@"
  "@P.POS.adjective@"
  "@P.TYPE.stem@"
  "@P.BASEFORM.'춥 chwup'@"
  "@P.SYLLABLE.closed@"
  "@P.SEM.[CONTENT <state,'cold'>]@"
 |
  추웠
  "@P.SURF.'추웠 chwu wess'@"
  "@P.POS.adjective@"
  "@P.BASEFORM.'춥 chwup'@"
  "@P.SYLLABLE.closed&dark@"
  "@P.SEM.[CONTENT   <state,'cold'>   TENSE
past]@"
 ] ; ! the end of adjective entries

 verbalEnding =

 [
  았
  ! Check constraint
  "@R.SYLLABLE.closed&clear@"
"@R.TYPE.stem@"

  "@P.SURF.'았 ass'@"
  "@P.POS.verbalEnding@"
  "@P.SYLLABLE.closed@"
  "@P.SEM.[TENSE past]@"
 |
  았겠
  ! Check constraint
  "@R.SYLLABLE.closed&clear@"
"@R.TYPE.stem@"

  "@P.SURF.'았겠 ass keyss'@"
  "@P.POS.verbalEnding@"
  "@P.SYLLABLE.closed@"
  "@P.SEM.[TENSE       past       MODALITY
conjectured]@"
 ] ; ! the end of verbalEnding entries

 adnoun =
 [
  추운
  "@P.SURF.'추운 chwu wun'@"
  "@P.POS.adnoun@"
  "@P.FORM.wordFinal@"
  "@P.BASEFORM.'춥 chwup'@"
  "@P.SYLLABLE.closed@"
  "@P.SEM.[CONTENT <state,'cold'>]@"
 ] ; ! the end of adnoun entries


 verbEnding =
 "@D.FORM.wordFinal@" ! D = Disallow

 [
  다
  "@P.SURF.'다 ta'@"
  "@P.POS.verbEnding@"
  "@P.FORM.sentenceFinal@"
  "@P.SEM.[SENTENCETYPE declarative]@"
 ] ; ! the end of verbEnding entries

!! the end of Definitions

LEXICON Root
 adjective;
 verb;
 adnoun;

LEXICON adjective
 <adjective> verbalEnding;
 <adjective> verbEnding;

LEXICON verb
 <verb> verbalEnding;
 <verb> verbEnding;

LEXICON adnoun
 <adnoun> #;

LEXICON verbalEnding
 <verbalEnding> verbEnding;

LEXICON verbEnding
 <verbEnding> #;
```

Table 1: Sample Enlarged Lexicon

2.2.3. Compilation

By giving a command apply up or simply up on **xfst** for '추웠다', we can get the following analysis:

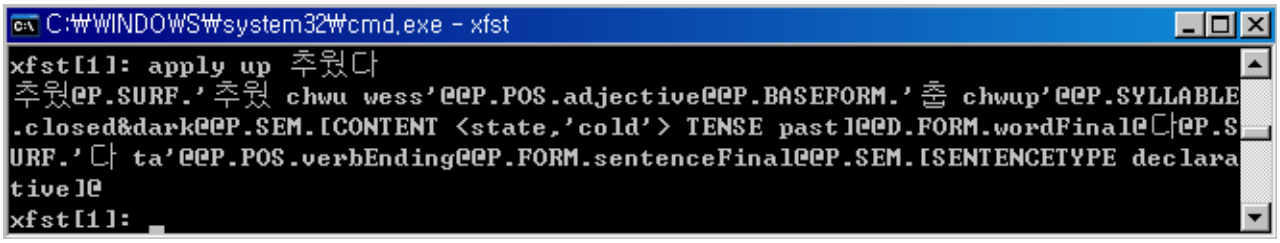

Figure 2: Analyzing for '추웠다' in Xerox Finite-State Tool (**xfst**) / Command Mode

In order to display the results of analysis or generation, **cfsm** runs an interface that represents each lexical item in a matrix form, showing how they are processed and concatenated step by step. Here is an example.

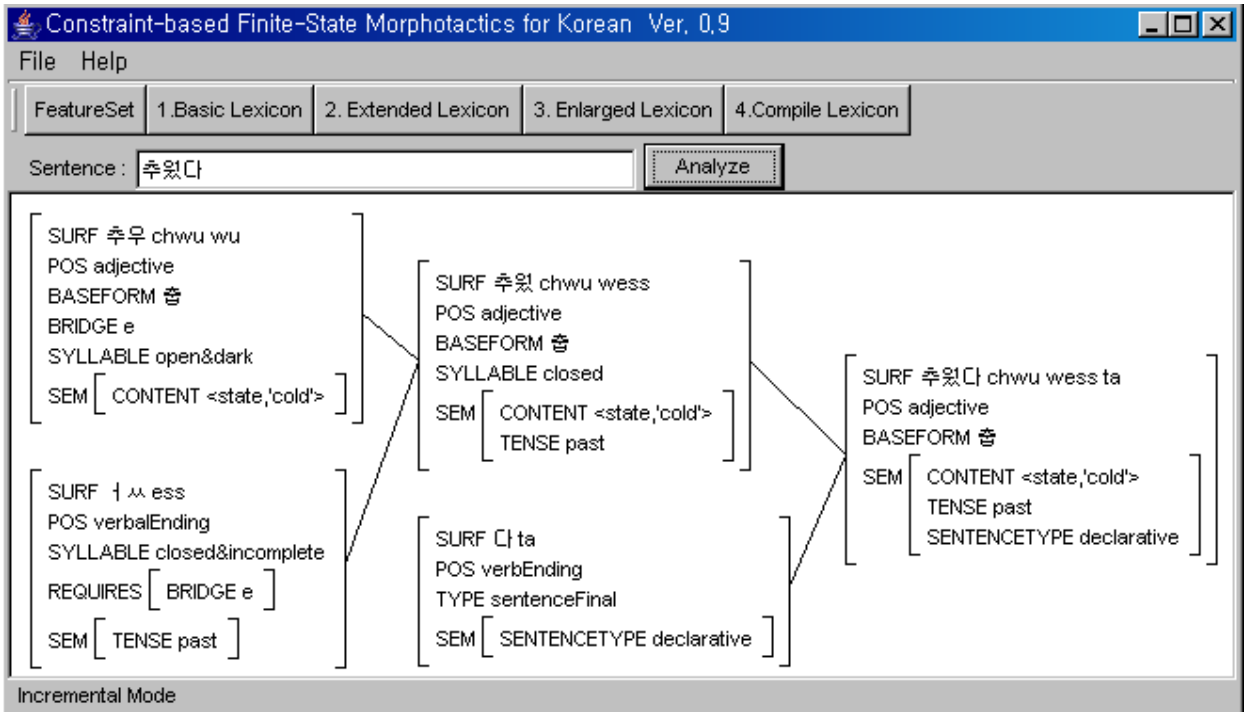

Figure 3: Analyzing for '추웠다' in Constraint-based Finite-State Morphotactics (**cfsm**) UI

3. Concluding Remarks

Our proposed **cfsm** is a newborn baby, although its theoretical conception dates back to the beginning of **Komor**. In this paper we have attempted to find automatic ways of converting **Komor** into **cfsm**, thus allowing the use of finite-state tools for morphotactics like Xerox's **xfst** and **lexc**. The primary benefit of this conversion is that the comparatively complete description of the Korean morphology in the rule-based **Komor** can directly be imported into the proposed **cfsm**.

4. References

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford.

Hausser, Roland. 1989. *Computation of Language: An Essay on Syntax, Semantics and Pragmatics in Natural Man-Machine Communication*. Springer, Berlin.

Kang, Seung-Shik. 2002. *Korean Morphological Analysis and Information Retrieval*. Hong-Rung Science Publications, Seoul. [Written in Korean.]

Lee, Kiyong. 1994. Hangul, the Korean writing system, and its computational treatment. *LDV-Forum*, 11.2:26.43.

Lee, Kiyong. 1999. *Computational Morphology*. Korea University Press, Seoul. [Written in Korean].

Shim, Kwangseob and Jaehyung Yang. 2002. Mach: A supersonic Korean morphological analyzer. *Proceedings of the 19th International Conference on Computational Linguistics* (COLING-2002), pages 939.945.

Shim, Kwangseob and Jaehyung Yang. 2004. Very high speed korean morphological analysis based on adjancy conditions check. Journal of the Korea Information Science Society: Software and Applications, 31.1:89.99. [Written in Korean.]