

Klex: A Finite-State Transducer Lexicon of Korean

Na-Rae Han

Department of Linguistics, University of Pennsylvania, Philadelphia, PA 19104, USA,
nrh@ling.upenn.edu,
WWW home page: <http://www.cis.upenn.edu/~nrh/klex.html>

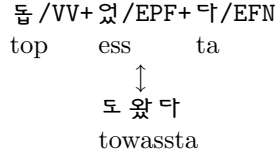
Abstract. This paper describes the implementation and system details of Klex, a finite-state transducer lexicon for the Korean language, developed using XRCE's Xerox Finite State Tool (XFST). Klex is essentially a transducer network representing the lexicon of the Korean language with the lexical string on the upper side and the inflected surface string on the lower side. Two major applications for Klex are morphological analysis and generation: given a well-formed inflected lower string, a language-independent algorithm derives the upper lexical string from the network and vice versa. Klex was written to conform to the part-of-speech tagging standards of the Korean Treebank Project, and is currently operating as the morphological analysis engine for the project.

1 Introduction

Korean is a highly agglutinative language, with productive use of post-position markers following nouns, verbal ending suffixes for verbs and adjectives, as well as frequent use of derivational morphology. Breaking up Korean words into smaller morphological components is the necessary first step in any natural language processing system of Korean, as well as an essential task for computerized language-learning tools. Klex is a fully operational Korean lexicon system whose underlying mathematical representation is a form of data structure known as the finite-state transducer or FST.¹ The transducer's upper level consists of strings representing a sequence of lexical forms of morphemes, each followed by their part-of-speech tag; the lower level consists of the fully inflected surface forms produced by concatenation of the morphemes, with relevant phonological and morphotactic processes applied:

¹ The name Klex is used somewhat ambiguously: it refers to the binary FST network that constitutes the lexicon; it also loosely refers to the entire morphological analysis and generation system of Korean lexicon built around the binary FST, with a few auxiliary networks, XML database and other helper scripts and utilities.

- (1) Mapping for $\text{돕} + \text{았} + \text{다}$ /top+ess+ta/ ‘help+Past+Declarative’:



A transducer network as a whole consists of all such possible morpheme-sequence/word pairs in the language. Given the lower inflected form, a language-independent algorithm can produce the analyzed morpheme sequence (the process of “looking-up”). Conversely, the transducer can be used in producing the fully inflected surface form of grammatical sequence of morphemes (the opposite of “looking-up”, hence Xerox’s terminology of “looking-down”). These two operations, namely morphological analysis and generation, are the most typical applications of lexical transducers.

In the remainder of the paper, we will present implementation details of the system, and then discuss the main characteristics of Korean morphology and some of the design aspects of Klex which are aimed at providing optimal solutions for Korean.

2 Implementation

The system is implemented using XRCE(Xerox Research Centre Europe)’s XFST software platform. It consists of the main binary FST network, some auxiliary networks, and scripts which perform the morphological analysis and generation operations. The source codes for the networks are built from an XML dictionary and some helper scripts.²

2.1 Overall Design

The main FST binary of Klex is built with three modules which are FSTs themselves: the lexicon FST, the rule FST, and the encoding-converter FST. These transducers are combined together in an operation known as composition to form the single lexical transducer which erases intermediate levels of representation to directly encode the relation between analysis strings and surface orthographical strings (Figure 1).

The lexicon FST is the backbone of the architecture: it is a network of legitimate Korean words, which maps lexical forms with POS in the upper level to their abstract representation at the bottom. The rule FST is composed at the bottom of the lexicon FST to apply morpho-phonological and orthographical processes to abstract symbols to produce the romanized surface strings.

² The main binary FST network of Klex, along with its source code, was released by Linguistic Data Consortium (LDC) catalog number LDC2004L01 and ISBN 1-58563-283-x.

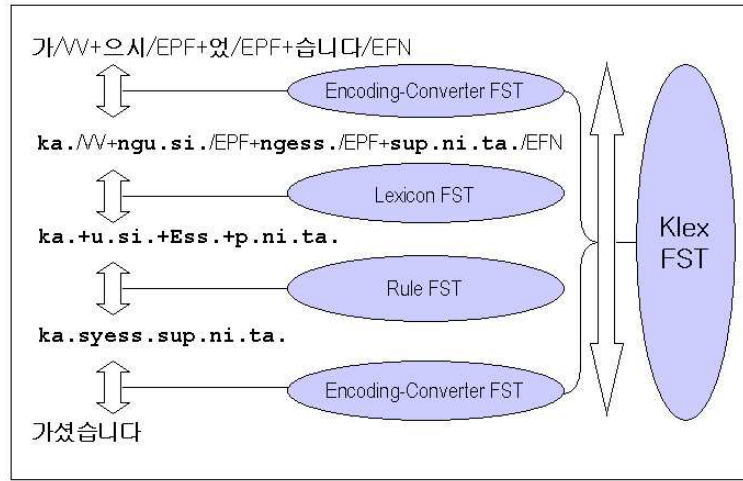


Fig. 1. Klex and component FSTs

Composing the lexicon FST and the rule FST results in a FST lexicon of Korean with roman transliteration of Korean strings on both levels; composing an encoding-converter FST on top and bottom produces the final product which handles Korean-encoded input and output. Currently Klex is configured for KSC-5601 (EUC-KR) encoding; however, it can be recompiled to use other encoding schemes such as UTF-8.

While *hangul*, the Korean script, is an alphabetic system with symbols for consonants and vowels in the language, modern computerized encoding schemes of Korean take the syllabic unit as the encoding block, thereby rendering the internal structure of syllables, e.g. individual alphabetic characters, opaque. This makes the task of writing rules, which must address vowels and consonants in the phonological inventory of the language, impossible with Korean encodings. For this reason, roman transliteration was used for the core part of the system. The romanization scheme adopted in Klex is a variant of the “Yale romanization system”, a popular choice among linguists. A few modifications were made: (1) end-of-syllable is marked with “.” to ensure a one-to-one mapping between the romanized and Korean alphabets; (2) the Korean consonant alphabet character “ㅇ” is always mapped to “ng” even in the onset position where it actually lacks phonological value, a measure taken in order to facilitate the process of writing rules.

2.2 The Dictionary

Klex relies on a dictionary in XML format, which serves as a database of lexical entries and their lexical properties, including morphotactic information and their underlying forms. Figure (2) shows examples of typical entries.

```

<entry>
  <rom>ngu.lo.</rom>
  <POS>PAD</POS>
  <form>u.lo.</form>
  <lemma>으 톨</lemma>
  <morpho>MiddleCosa:MiddleCosa;</morpho>
</entry>

<entry>
  <rom>top.ta.</rom>
  <POS>VVt</POS>
  <form>toP.</form>
  <irr>ㅂ</irr>
  <lemma>돕 탁</lemma>
  <morpho>vv</morpho>
</entry>

```

Fig. 2. Dictionary entries for 으로 /ulo/ ‘with’ and 돕 /top/ ‘to help’

Each entry contains the following fields: the romanized form, the part-of-speech, the underlying form and the verbal conjugation class where relevant, the lemma form, and the morpho-syntactic information. There is a set of scripts written in perl which scan this dictionary and then create a source code for the lexicon FST based on the information in the fields.

The vocabulary included in the lexicon was obtained from many sources, including *Minjung Eutteum Korean Dictionary for Elementary School Students* (민중 초등학교 으뜸 국어사전: Minjungseorim, 1998), the Korean English Treebank Annotation (Palmer et al., 2002) and other various texts. The lexicon was expanded and fine-tuned by testing against these various corpora, the process of which included fixing undesirable outputs and adding missing lexical entries. The XML dictionary contains about 148,000 entries: the vast majority of them are nouns (about 133,000), and a small fraction of them belongs to the affix category (total of 622).

2.3 Morphotactics

The morphotactic grammar is formulated as a set of continuation classes in LEXC grammar which are then compiled into the lexicon FST. Much attention was paid to the morphotactics of postposition markers, verbal ending suffixes, derivational morphology and finally, noun compounding.

Some non-local dependencies are observed in the morphotactic grammar of Korean, most of which involve semantic constraints between suffixes encoding aspectual senses and verbal and adjectival stems. For example, final verbal ending 는다 /nunta/ is incompatible with adjective stems, since it encodes the “habitual” aspect, therefore: 먹는다 /meknunta/ “eats (habitual)” vs. *작는다 /*caknunta/ “*is small (habitual)”. Other suffixes such as the honorific marker 으시

/usi/ can intervene, which makes the dependency a non-local one. Verb and adjective stems in Korean share the same morphotactics for the most part, which makes the option of writing separate continuation classes for the two categories redundant and impractical. The standard method of Flag Diacritics, available in the XFST platform, provides an economical solution: a bit of Flag Diacritic indicating its POS is set for adjective roots, which is then checked against a matching bit of Flag Diacritic associated with the habitual suffix ending at runtime, ruling out any path that traverses the two incompatible morphemes.

2.4 Rewrite Rules for Morpho-Phonological Alternations

Korean is known to display a wide variety of morpho-phonological processes, which the rule FST was designed to handle. The rule FST itself is a single transducer which maps the abstract morphophonemic strings to the surface orthographical strings, built by applying a composition operation on an ordered sequence of 100 individual replace rules.

The initial input to the set of rules, e.g. the lower strings of the lexicon FST, contains rich abstract information which allows the rules to set their targets precisely. Morpheme boundaries, where all alternations occur, are marked with “+”. Morphemes that go through morpho-phonological processes are given abstract representations which deviate from the romanization scheme, such as inclusion of upper-case letters and missing onset consonant. The rules are formulated to target these symbols and environments.

The replace rules, comparable to the rewrite rules used traditionally in phonological derivations (Chomsky and Halle, 1968), handle three major groups of alternations: transformations at the right edge of irregular verbal roots; phonological processes occurring on morpheme boundaries, mostly involving verbal ending suffixes; and allomorphy in post-position markers. The first group of alternations involve six classes of irregular verbs of Korean, whose roots are subject to different sets of phonological processes when followed by transformation-triggering suffixes. The second group of phonological processes mostly involves verbal ending suffixes, such as vowel harmony, epenthesis, glide formation and u deletion³.

Figure 3 illustrates derivations of three irregular verbal roots 듣 /tut/ “to hear”, 자르 /calu/ “to cut”, and 젓 /ces/ “to stir”⁴ and a regular verbal root 가 /ka/ “to go”, followed by any verbal ending with an initial abstract vowel E. The key morpho-phonological processes involved here are: “T” irregular stem operation, “L” irregular stem operation, “S” irregular stem operation, vowel harmony, vowel merge and u (which is the epenthetic vowel in Korean) deletion. The rule [. .] -> ng || . + _ VOW; addresses an orthographic issue: Korean orthography inserts a phonologically empty consonant “ㅇ” into an empty onset position. At the very end, the cleanup rule removes remaining traces of the morpheme boundary.

³ /u/ functions as the epenthetic vowel in Korean phonology.

⁴ Called “ㄷ”, “ㄹ”, and “ㅅ” irregular verbs respectively in the school grammar of Korean.

듣+어 ⇒ 들어 “to hear”	자르+어 ⇒ 잘라 “to cut”	젓+어 ⇒ 저어 “to stir”	가+어 ⇒ 가 “to go”	rewrite rule
tuT.+E.	ca.Lu.+E.	ceS.+E.	ka.+E.	vowel harmony E → a [o a] (CON . CON u) . + -; E → e;
tuT.+e.	ca.Lu.+a.	ceS.+e.	ka.+a.	u deletion u . + → [..] CON _ [e a];
n/a	ca.La.	n/a	n/a	vowel merge a . + → [..], e . + → [..] \w _ VOW
n/a	n/a	n/a	ka.	“T” irregular T → ɫ _ . + E;
tul.+e.	n/a	n/a	n/a	“L” irregular . L → ɫ . ɫ _ VOW;
n/a	cal.la.	n/a	n/a	“S” irregular S → [..] _ . + VOW;
n/a	n/a	ce.+e.	n/a	insert empty onset consonant [..] → ng . + _ VOW;
tul.+nge.	n/a	ce.+nge.	n/a	remove morpheme boundary + → [..];
tul.nge.	n/a	ce.nge.	n/a	

Fig. 3. Cascade of rules applied to verb forms (CON: consonant; VOW: vowel)

The top strings are the abstract representations that are given as the input to the rule FST; the sequence of rules are applied and the bottom strings are resulting surface strings. Note that the precise ordering among the rules is crucial: the vowel merge rule must apply before the “S” irregular rule, otherwise an illegitimate surface string `*ce.` would be derived instead of the correct string `ce.nge.`. This dissected view of the rule module easily gives the illusion of a procedural model of rule application; ultimately, the cascade of rules are compiled into a single rule FST which maps the abstract strings on the top directly to the surface strings at the bottom. This “composition of sequential rules” architecture achieves the elegance of the mathematically equivalent model of KIMMO-style two-level morphology (Kokenniemi, 1983, 1984; Karttunen, 1983) while granting the linguist a higher level of flexibility and ease in formulating rules.

3 Language-Specific Issues

3.1 Lexical Representation of Affixes

The analyzed upper strings of Klex take the following form: `morph1/POS1+morph2/POS2+...+morphn/POSn`, where morphemes are separated by “+” and each morpheme is followed by “/” and its part-of-speech tag. Korean affixes, therefore, are represented as a lexical item, rather than combinations of grammatical features such as `+plural`, `+honorific` and `+past-tense`, the popular approach taken mostly by finite-state lexicons of European languages. This is a rather inevitable design decision given the sheer number of Korean affixes; currently there are 622 of them listed in the system. Translating each one of them into a combination of binary features is not only an infeasible task but also an undesirable one; many affixes in Korean convey their own semantic and pragmatic senses that are not easily decomposed into matrices of binary features.

The feature representation of affixes can be more useful depending on the nature of applications, especially those that involve generation. The current setup requires the user to supply all component morphemes in their correct representative forms along with their correct part-of-speeches in order to obtain the inflected surface forms, which mandates a sophisticated level of knowledge on user’s part. For a system that generates Korean words, however, a single verbal root marked with some grammatical features such as `+polite`, `+past-tense`, `+interrogative` might be considered a more reasonable input, which suits the feature representation scheme well. Although the system is not presently set up to handle such an alternative representation, its flexibility allows it to be modified as such with relative ease. First a subset of affixes would have to be selected excluding those that are inessential from the generation perspective. Their entries in the XML dictionary must then be augmented with appropriate feature representations. Finally a set of modified scripts can then produce a LEXC source script with the feature representations of suffixes instead of their lexical form and the POS in the upper level.

The set of part-of-Speech tags used is fully compliant with the specification of the Korean Treebank Project Phase 2. The set, which includes a total of 33 tags, is based on the one employed by the Korean Treebank Project Phase 1 (Han & Han, 2001) with some newly introduced modifications.⁵

3.2 Allomorphy in Klex

A large number of inflectional suffixes and post-position markers in Korean have allomorphs, whose distributions are conditioned by the phonological environment in which they appear. For example, the “topic” proposition marker takes three different forms 은 /un/, 는 /nun/, and ㄴ /n/; the past-tense pre-verbal-ending suffix 았 /ess/, 았 /ass/, and ㅆ /ss/.

The predominant position taken by past and present systems of Korean morphological analysis has been not to posit a single lexical representation for such sets of allomorphs, opting instead to output appropriate allomorphic forms within context.

Klex diverges from other systems by treating allomorphs as having a single representative form. All allomorphs of a given lexical item therefore show up as a single form in the upper (analyzed) string. For example, the topic markers in 학교-는 /hakkyo-nun/ ‘school-Top’, 학생-은 /haksayng-un/ ‘student-Top’ and 너-ㄴ /ne-n/ ‘you-Top’ are equally assigned 은 /PAU in the analyzed strings:

- (2) 은 /un/ as the representative form for the Korean topic postposition:

학교 /NNC+은 /PAU	학생 /NNC+은 /PAU	너 /NPN+은 /PAU
hakkyo un	haksayng un	ne un
↓	↓	↓
학교 는	학생 은	넌
hakkyonun	haksayngun	nen

The criteria used in determining the representative form among allomorphs are as follows:

- (3) Criteria for determining the representative form
- a. The representative form should be fully syllabic, i.e. 은 /un/ is chosen over ㄴ /n/.
 - b. The form for the post-consonantal environment is chosen, i.e. 이 /i/ instead of 가 /ka/.
 - c. Epenthetic vowels are included, i.e. 으로 /ulo/ and not 로 /lo/.⁶
 - d. For vowel harmony, 어 /e/ is chosen over 아 /a/, i.e. 어서 /ese/ and not 아서 /ase/.

⁵ The POS tagging guideline for the Korean Treebank Phase 1 can be found at: <ftp://ftp.cis.upenn.edu/pub/ircs/tr/01-09/>.

⁶ This clause is in fact redundant, as epenthetic vowels are used in post-consonantal environments only which is covered by criterion (b).

Note that these representative forms are to be distinguished from the “abstract underlying representation”: the representative forms are those that function as the dictionary entry; the abstract underlying representations are their romanized counterparts with abstract symbols that are used system-internally, and which eventually undergo morpho-phonological transformations through applications of rewrite rules. For example, ㄹ /ess/ is the representative form for the the past-tense verbal ending suffix with its underlying abstract representation Ess..

3.3 The Guesser Modules

A successful morphological analysis requires that the root of the encountered word be listed in the lexicon database. Even with Klex’s extensive dictionary, it certainly cannot provide full coverage for the continuously evolving vocabulary of the language. To handle cases of novel words, two “guesser” modules are implemented: one dealing with novel roots belonging to open classes of part-of-speech and the other dealing with Korean person names.

Some part-of-speech categories, such as affixes, verbs and adjectives of Korean are considered a closed class: they consist of a closed set of vocabulary items, and addition of new vocabulary items is rare. On the other hand, common nouns (NNC), proper nouns (NPR), adverbs (ADV) and interjections (IJ) are open classes which allow novel vocabulary items, including newly formed words or borrowed words, to be added more freely. Phonologically possible roots are defined for each of the four part-of-speeches, and a guesser lexicon is compiled with the guessed roots in place of real roots. The guessed roots are subject to the same morphotactic grammar and morpho-phonological alternations as real roots. Figure 4 shows how 핸섬하다 /haynsemhata/ “is handsome” is handled by the module. $\text{핸섬}^{\sim}\text{Guess/NNC+하/XSJ+다/EFN}$ (haynsem[~]Guess/NNC+ha/XSJ+ta/EFN) is the correct guess, with the borrowed word 핸섬 (/haynsem/) is a guessed common noun root (~Guess/NNC) followed by an adjectivization suffix ha/XSV .⁷

```

xfst[1]: apply up 핸섬하다
핸섬하~Guess/NPR+이/CO+다/EFN
핸섬하~Guess/NNC+이/CO+다/EFN
핸섬~Guess/NNC+하/XSV+다/EFN
핸섬~Guess/NNC+하/XSJ+다/EFN ← CORRECT GUESS!
핸섬하다~Guess/ADV
핸섬하다~Guess/NPR
핸섬하다~Guess/IJ
핸섬하다~Guess/NNC

```

Fig. 4. Guesser module analyzes novel 핸섬하다 /haynsemhata/ “is handsome”

⁷ Foreign vocabularies lose their original part-of-speech and are uniformly treated like nouns in Korean.

Korean person names pose another challenge. Klex implements an auxiliary FST module tailored to recognize them. Korean person names in most common cases consist of three-syllables, one for the surname and the the other two for given names⁸. The list of 137 known Korean surnames is obtained from census data. Given names are usually built by combining two syllables from a pool of chinese characters – while the Korean syllabic structure permits 11,172 possible syllabic combinations of which 2,350 are in wide use, the range of syllables commonly used in names is relatively small: we hand-picked 262 of them. A model of Korean person names is then described as a regular expression: a syllable from the surname set, followed by two from the pool of syllables for given names. With the regular expression in place for a proper-noun root, the FST for Korean names can recognize legitimate Korean person names followed in some cases by titles or postposition markers. This is by no means a robust model of Korean names, but provides a good enough measure for the task of guessing.

4 Conclusions

Klex is a full-scale FST-based lexicon model for morphological analysis and generation of Korean words. The finite-state technology, which XRCE’s XFST suite implements, provides an elegant yet powerful mathematical framework for designing such a system for morphologically complex languages such as Korean. The XFST suite by XRCE provides particularly powerful and flexible tools for a linguist seeking to develop such a system, since their composition-based modular architecture lets the developer model distinct aspects of morphology such as morphotactics and morpho-phonemic alternations into separate modules of transducers, which are then combined into a single transducer network that is both structurally simple and efficient. Klex was developed to fit the specification of the Korean Treebank Project, and is currently employed as the morphological analyzer for the project. It is also available by licensing through the Linguistic Data Consortium (LDC).

Acknowledgements

This research has been partially supported by various sources, including the Korean Treebank Project at the University of Pennsylvania, ARO grant DAAD 19-03-2-0028, a 5-year grant (BCS-998009, KDI, SBE) from the National Science Foundation via TalkBank, and the Linguistic Data Consortium. We would like to thank Xerox (XRCE) for making their tools available to the public. Also special thanks go to: Ken Beesley and Lauri Karttunen who provided valuable insights and guidance through the initial stages of the project; Mike Maxwell for his help in putting together the final product; and finally Martha Palmer for her continuous support throughout the course of the project.

⁸ Only 9 Korean surnames with two syllables are documented. Monosyllabic Korean given names are not rare; given names longer than 2 syllables are a rarity.

References

1. Back, D.H., Lee, H., Rim, H.C.: A structure of korean electronic dictionary using the finite state transducer. In: Proceedings of the 7th Symposium for Information Processing of Hangul and Korean (한글 및 한국어 정보처리 학술대회). (1995) in Korean.
2. Beesley, K.R., Karttunen, L.: Finite-State Morphology: Xerox Tools and Techniques. CSLI Publications, Stanford, California (2003)
3. Han, C.H., Han, N.R.: Part of speech tagging guidelines for penn korean treebank. Technical report, IRCS, University of Pennsylvania (2001)
4. Han, N.R.: Klex: Finite-state lexical transducer for korean. Linguistic Data Consortium (LDC) catalog number LDC2004L01 and ISBN 1-58563-283-x (2004)
5. Han, N.R.: Morphologically annotated korean text. Linguistic Data Consortium (LDC) catalog number LDC2004T03 and ISBN 1-58563-284-8 (2004)
6. Kim, S.: Korean Morphology (우리말 형태론). Tap Publishing, Seoul, Korea (1992) in Korean.
7. Ko, Y.: A Study of Korean Morphology (국어형태론연구). Seoul National University Press, Seoul, Korea (1989) in Korean.
8. Koskenniemi, K.: Two-level morphology: A general computational model for word form recognition and production. Publication No: 11, Department of General Linguistics, University of Helsinki (1983)
9. Minjungseorim, ed.: Minjung Eutteum Korean Dictionary for Elementary School Students (민중 초등학교 으뜸 국어사전). Minjungseorim, Seoul, Korea (1998) in Korean.
10. Palmer, M., Han, C.H., Han, N.R., Ko, E.S., Yi, H.J., Lee, A., Walker, C., Duda, J., Xue, N.: Korean english treebank annotations. Linguistic Data Consortium (LDC) catalog number LDC2002T26 and ISBN 1-58563-236-8 (2002)