Finite-State Morphology of Estonian: Two-Levelness Extended

Heli UIBO
Institute of Computer Science
University of Tartu
J. Liivi 2
Tartu 50409, Estonia
Heli.Uibo@ut.ee

Abstract

The paper is concentrated on modeling the Estonian morphology in the framework of two-level morphology model. The result is a consistent description of Estonian morphology, which consists of a network of lexicons (root lexicons cover 2500 most frequent word roots) and two-level rules. The main rule set contains 45 rules, which describe various stem changes. The subset of rules dealing with stem internal changes is applied separately as well. For modeling the derivation process a new solution has been found – to extend the two-levelness into the upper side of the morphological transducer (to the lemmas). It has been shown that finite-state methods are applicable and sufficient for describing Estonian inflectional processes, but word formation rules, especially compounding, require more investigation.

1 Introduction

During the last 25 years the finite-state approach has been the most fruitful one in the field of computational morphology. Although there exist two computerized descriptions of the Estonian morphology (Viks 00; Kaalep 00) it is worth to try to apply finite-state techniques to the Estonian morphology, to make the results comparable to those of other languages.

It is important that a finite-state transducer is bidirectional in its nature, as it describes a regular relation, or a correspondence between two languages. In the simplest case the morphological transducer is a lexical transducer, on the upper side of which are primary forms concatenated with appropriate morphological information and on the lower side – word forms. Each path from the initial state to a final state represents a mapping between a word form and its morphological reading. The morphological analysis can then be understood as the "lookup" operation in the lexical transducer, whereas synthesis – the "lookdown" operation (Beesley & Karttunen 03). The lexical transducer can be composed with rule transducer(s) that convert lexical representation to surface representation, using either two-level (Koskenniemi 83) or replace rules (Karttunen 95).

2 Finite-state morphology of Estonian

2.1 Overview

Estonian is a highly inflected language – grammatical meanings are expressed by grammatical formatives which are affixed to the stem instead of using prepositions. According to more detailed analysis the stem consists of word root and derivational affixes and formative – of features and endings.

The morphological word classes in Estonian:

- nouns (can be declined)
- verbs (can be conjugated)
- indeclinables (remain unchanged)

Nouns have 14-15 cases in singular and plural, there are often parallel forms in plural. Verbs have four moods (indicative, conditional, imperative, quotative), four tenses (present, imperfect, present perfect and past perfect), two modes (personal and impersonal), two voices (affirmative and negative), three persons and two numbers (singular and plural). Derivation is mostly done by affixing:

kiire (Adj) 'quick' *kiire|sti* (Adv) 'quickly'
õppi|ma (V) 'to learn' *õppi|mine* (N) 'learning'

For compounding the concatenation of stems is used. The pre-components of compound nouns can be either in singular nominative, singular genitive and in some cases in plural genitive case. Only the last component is declinable.

Example:
piiri + valve + väe + osa = piirivalveväeosa
border guard power part = 'troup of border guards'
sg gen sg gen sg gen sg nom

There generally exist two different processes in natural language morphology:

1. morphotactics – how to combine word forms from morphemes
   a) concatenative processes (prefixation and suffixation, compounding)
   b) non-concatenative processes (reduplication, infixation, interdigitation)
2. phonological alternations (examples from Estonian)
   a) assimilation (*hi**nd**:hi**nn**a* 'price' sg nom : gen)
   b) insertion (*jooksma:jooks**e**v* 'to run' : 'running')
   c) deletion (*numb**e**r:numbri* 'digit' sg nom : gen)
   d) gemination (*tu**b**a:tu**pp**a* 'room' sg nom : adit)

It has been shown in (Beesley & Karttunen 00) that concatenation, composition and iteration are sufficient means for describing the morphology of languages with concatenative morphological processes. The Estonian morphotactics does not make use of productive non-concatenative processes, thus, theoretically, no problems should occur by the modeling the Estonian morphology by finite-state methods.

The morphological description of Estonian has been built up by the author, lead by the principles of the two-level morphology model (Koskenniemi 83). The two-levelness means that the lexical representations of morphemes are maintained in the lexicons and the task of two-level rules is to "translate" the lexical forms into the surface forms and vice versa. The lexical forms may contain information about the phoneme alternations, about the structure of the word form (morpheme boundaries) etc.

The model is language-independent, but for the different languages the balance between rules and lexicons can be different. The network of lexicons is good for agglutinating languages like Finnish (Koskenniemi 83), Turkish (Oflazer 94) and Swahili (Hurskainen 95), where word forms are built by concatenation of morphemes. Two-level rules are convenient to handle single phoneme alternations. If the stem variants differ more from each other (e.g. *pidu:peo* ('party' sg nom : sg gen) then the stem change can be handled analytically (cf. section 2.4). The Estonian language is both agglutinative and flective. For instance, the word form *hammastega* 'with teeth' is built from the morphemes *hammas+te+ga* and stem flexion rules determine that the stem variant is *hammas* but not *hamba*.

The morphological phenomena occurring in the Estonian language have been divided between rules and lexicons as follows:

- The rules of phonotactics, different stem flexion types and morphological distribution have been formalized as two-level rules.
- The rules of morphotactics have been described in the network of lexicons.
- The stem final alternations have been divided between lexicons and rules. Most of the alternations concerning only one grapheme have been formalized as rules. Handling the change of a whole segment by two-level rules requires several rules to be coordinated (Trosterud & Uibo 05) and therefore, the stem final changes like *hobune : hobuse : hobust* are handled by continuation lexicons.

2.2 The network of lexicons

The network of lexicons was designed after the morphological classification by Ülle Viks (Viks 92), which is based on pattern recognition. It is compact and oriented for automatic morphological analysis. It contains 38 inflection types – 26 for nouns and 12 for verbs. 84 words (including most of the pronouns) are handled as exceptions. We have additionally splitted some noun types according to the stem final vowel.

Each inflection type has been modeled as a number of linked lexicons. The first group generates stem variants (lexicon 28 in Figure 2), the second group locates the stem variants in paradigm (lexicons TP_28at and TP_28an) and the third builds the base forms and their analogy forms (lexicons An_ma … An_takse). This kind of structure has been inspired by (Viks 92).

The paradigms of all the noun and verb inflection types have been described in the network of lexicons. Comparison of adjectives, productive derivation and compounding have also been implemented, using continuation lexicons. The word formation rules are too general yet. Nevertheless, the problem is application-dependent. For information retrieval, the problem of overgeneration is of less importance than for spelling check (Uibo 02).

2.3 Problems with lexicons

The network of lexicons seems to be a powerful tool: following the links between different lexicons word roots, derivation suffixes, inflectional features and endings can be combined into grammatical word forms. However, a number of problems occured in practice:

- As there are many inflection types in Estonian, the number of continuation lexicons is also

high (164) and the network of lexicons becomes difficult to manage. But the number could be even bigger if we did not use two-level rules for handling stem internal changes (Trosterud & Uibo 05).

- Using word lists does not fit into the model, however it is needed to constrain the overgenerating derivation and compounding.
- The principle that the rules of morphotactics and the distribution of stem variants are described by lexicons and the phonological relations of stem variants are formalized as two-level rules cannot always be followed. Stem final alternations have often become individual properties of a word and are not predictable by phonological rules.
- The network of lexicons would be best readable if for each morpheme there is exactly one lexicon. In the existing network of lexicons the morphemes are often splitted that cuts the readability down.

2.4 Rules

The majority of two-level rules handle stem flexion and phonotactics. The most interesting inflection type from the point of view of phonological changes is characterized by weakening consonant gradation – the deletion of *b*, *d*, *g* or *s* – and also changes in the immediate neighbourhood of the disappeared consonant – the lowering of the surrounding vowels.

Example list of words belonging to the type:

*mad**u** : mao*	*si**g**a : sea*	*u**b**a : oa*
*lu**g**u : loo*	*k**ä**si : käe*	*s**ü**si : söe*

There should be a rule for handling the deletion (`$` is the weak grade marker):

```
SC:0 <=> Vok: _ Vok: %$:;
```

And another rule for vowel lowering:

```
HVow:LVow<=>Bgn_LV:  StemVow:  %$:  ;
                Bgn Vow: LV: _ %$: ;
where     HVow in (u ü i)
          LVow in (o ö e)
matched ;
```

The last rule has two contexts: the lowering can occur both in the right (*mad**u** : mao*) and in the left context (*s**i**ga : s**e**a*). In (Uibo 00) the stem flexion types and the discovery process of rules have been discussed in details. Figure 1 gives an overview of the whole rule set.

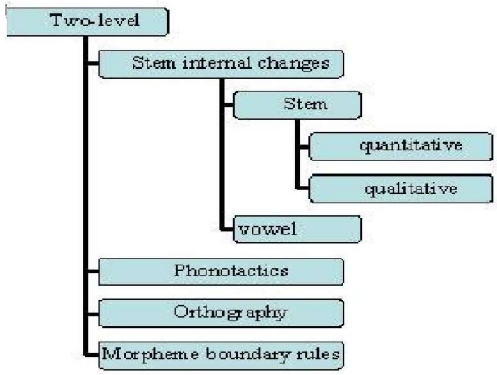

Figure 1: Two-level rules for Estonian

3 A new approach to word formation modeling – two-levelness extended

All Estonian verbs are subject to productive derivation processes resulting in the word forms exemplified in Table 1.

Deriv. suffix	Example *(lugema)*	Translation (*to read*)	Word class
-ja	**luge**ja	reader (person)	Subst
-mine	**luge**mine	reading (process)	Subst
-v	**luge**v	reading	Adj
-tav	**loe**tav	being read	Adj
-nud	**luge**nud	having read	Adj
-tud	**loe**tud	read (finished)	Adj
-nu	**luge**nu	one who has read	Subst
-tu	**loe**tu	one that has been read	Subst

Table 1: Productive derivation from verbs

Modeling the productive derivation from verbs with **weakening consonant gradation**, i.e. verbs for which the primary form (supine) is in the strong grade but some inflected forms in the weak grade, we have run into a serious problem. Namely, the information for the derived word form, outputted during the analysis, should contain the derived primary form, which can be in the weak grade (*loetav, loetud, loetu*).

The lexical transducer picks up the strong-grade stem and the word class V (verb), but it may be a derived word with a weak lemma. The initial solution was to include the weakening verbs into root lexicons three times – into the root lexicon of verbs and into the root lexicon of verbal derivatives in both strong and weak grade (Figure 2).

```
LEXICON Verb  ! Root lexicon
  lugema+V : luGe 28;

LEXICON 28  ! Building of stem variants
        TP_28at;
: $     TP_28an;

! Distribution of stem variants in the paradigm
LEXICON TP_28at      ! luge+…
 An_ma;
 Am_mata;
 An_v;
 An_sin;
 An_sime;
 An_da;
 An_ge;
 Ja_mine;

LEXICON  TP_28an ! loe+…
 An_b;
 An_me;
 An_tud;
 An_takse;

! Base forms and their analogy forms.

LEXICON An_ma
 ma+V+sup+ill :                  ma        GI;
 ma+V+quot+pres+ps :             vat       GI;

LEXICON An_v
 ma+V+partic+pres+ps :           v         GI;
 v+A+pos+sg+nom+partic :         v         02_A;
 …

LEXICON An_takse
 ma+V+indic+pres+imps+af :       takse   GI;

! Productive derivation
LEXICON Verb-Deriv
  loe Partic/N-N;
  luge Partic/N-T;

LEXICON Partic/N-N
  tav+A :                 tav       A_02_A;
  tav+S :                 tav       Axx;
  tud+A+Sg+N :            +tud      #;
  tu+S :                  tu        01;

LEXICON Partic/N-T
  v+A  :        v         A_02_A;
  nud+A  :      nud       #;
  nu+S  :       nu        01;
                          Ja_mine;
LEXICON Ja_mine
 ja+S :         +ja       01;
 mine :         +m        12_nE-SE-S;
 mata+A :       +mata     #;
```

Figure 2: Derivation from verbs: a storage consuming solution

Finally we have found a helpful solution to the weak grade verb derivatives problem: **to extend the two-levelness to the upper side of the lexical transducer** (to the lemma). The solution has been implemented as sketched on Figure 3.

```
LEXICON Verb
  luGe 28;

! inflection like in figure 2; skipped
! productive derivation
LEXICON 28_deriv
 ja+S          :         +ja       Szz;
mine+S         :         +mine     Sqq;
v+A            :         +v        Aww;
$tav+A         :         @+tav     Aww;
+nud+G         :         +nud      #;
$tud+G         :         $+tud     #;
nu+S           :         +nu       Scc;
$tu+S          :         $+tu      Sdd;

LEXICON Substantiiv
 Scc;
  ...
LEXICON Adjektiiv
Aww;
```

Figure 3: Derivation modeling: a better solution

As a result, the productive verb derivatives do not require three, but only one record in the root lexicon. To get the lemma in the correct surface form, stem flexion rules have to be applied onto the upper side of the lexical transducer. The resulting morphological transducer of Estonian can be formulated as follows:

$$((\text{LexiconFST})^{-1} \circ \text{RulesFST}_1)^{-1} \circ \text{RulesFST}$$

Here LexiconFST is the lexical transducer, RulesFST is the rule transducer (the intersection of all two-level rules) and RulesFST_1 is the intersection of consonant gradation rules. The operations used are composition and inversion.

The percentage of verbs is about 15 % among the 10 000 most frequent words of written Estonian (Kaalep & Muischnek 02). Thus, after the extension of two-levelness the number of records in root lexicons will decrease ca 23 %.

4 Implementation

The rules and lexicons are compiled into finite-state transducers using the Xerox finite-state tools *twolc* (Karttunen & Beesley 92) and *lexc* (Karttunen 93).

In the course of the project some additional tools have been developed:

- A tool for automatic updating of root lexicons (generates the lexical representation and detects the inflection type);
- A tool for testing the morphological analyzer on correctly tagged corpus. The program lists the words tagged correctly and incorrectly as well as unknown words.

The testing and lexicon extending cycle will go on, as the present coverage of the lexicon is about 30 % only.

5 Conclusion and perspectives

The finite-state approach has been resulted in a consistent description of the Estonian morphology, consisting of a network of lexicons and two rule sets. 45 rules that handle stem flexion, phonotactics, orthography and morphophonological distribution. A subset of stem flexion rules is used separately as well. The root lexicons contain 2500 most frequent words, based on the frequency dictionary of Estonian (Kaalep & Muischnek 02). There are 164 continuation lexicons which describe stem final changes, noun declination, verb conjugation, derivation and compounding.

A new solution has been proposed for modeling derivation: two-levelness has been partly extended to the upper side of the lexical transducer – to the lexical representations of the lemmas of forms productively derivable from the verb stems. The proposed approach may be applied for other languages where the word stems change in the course of derivation.

It has been shown that two-level representation is useful for the description of the stem internal changes, especially because the stem flexion does not depend on the phonological shape of a stem in the contemporary Estonian any more. The network of lexicons, combined with rules, having effect on morpheme boundaries, naturally describe the morphotactic processes. Lexicons are also good for describing non-phonological stem end alternations.

However, some open problems remain to be solved for the Estonian finite-state morphology:

- To increase the coverage of root lexicons.
- To guess the analysis of unknown words. The idea is to include a regular expression (e.g. $CVVC^{+}V$) in the root lexicon for each productive inflection type.
- To constrain the overgeneration of compound words by semantic constraints.
- To include the finite-state morphological component into practical applications. The most interesting idea in this perspective is to work on fuzzy information retrieval that is tolerant to misspellings and typos.

6 Acknowledgements

The research on finite-state morphology of Estonian has been supported by the Estonian Science Foundation grant No. 4605. Our thanks also go to Kimmo Koskenniemi, Lauri Karttunen, Kenneth Beesley and Trond Trosterud for encouraging discussions.

References

(Beesley & Karttunen 00) K. Beesley, L. Karttunen. *Finite-State Non-Concatenative Morphotactics*. In "Proceedings of SIGPHON-2000" 5th Workshop of the ACL Special Interest Group in Computational Phonology, Centre Universitaire, Luxembourg. 1-12.

(Beesley & Karttunen 03) K. Beesley, L. Karttunen. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications. Stanford, USA 2003.

(Kaalep & Muischnek 02) H.-J. Kaalep, K. Muischnek. *Eesti keele sagedussõnastik*. (*The frequency dictionary of written Estonian*). University of Tartu Press, Tartu 2002.

(Hurskainen 95) A. Hurskainen. *Information Retrieval and Two-Directional Word Formation.* Nordic Journal of African Studies 4 (2): 81-92 (1995).

(Karttunen 93) L. Karttunen. *Finite-State Lexicon Compiler.* Technical Report. ISTL-NLTT-1993-04-02. April 1993. Xerox Palo Alto Research Centre. Palo Alto, California.

(Karttunen 95) L. Karttunen. *The Replace Operator*. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. ACL-95, pp. 16-23, Boston, Massachusetts.

(Karttunen & Beesley 92) L. Karttunen, K. Beesley. *Two-Level Rule Compiler*. Technical Report. ISTL-92-2. Xerox Palo Alto Research Centre. Palo Alto, California.

(Koskenniemi 83) K. Koskenniemi. *Two-level Morphology*: *A General Computational Model for Word-Form Recognition and Production*. University of Helsinki, Dept of General Linguistics. Publications No. 11. Helsinki 1983.

(Oflazer 94) K. Oflazer. *Two-level Description of Turkish Morphology,* Literary and Linguistic Computing, Vol. 9, No:2 (1994).

(Trosterud & Uibo 05) T. Trosterud, H. Uibo. *Consonant gradation in Estonian and Sámi: two-level solutions.* Festschrift in honor of Professor Kimmo Koskenniemi's 60th anniversary. CSLI Publications 2005.

(Uibo 00) H. Uibo. *Kahetasemeline morfoloogiamudel eesti keele arvutimorfoloogia alusena*. (*Two-level morphology model as a basis for computational morphology of Estonian*) In "Arvutuslingvistikalt inimesele." Publications of the Department of General Linguistics, University of Tartu No. 1, pp. 37-72. Tartu 2000.

(Uibo 02) H. Uibo. *Experimental Two-Level Morphology of Estonian.* In "LREC 2002. Third International Conference on Language Resources and Evaluation." Las Palmas de Gran Canaria, Spain. Proceedings. Vol. III. pp. 1012 – 1015.

(Viks 92) Ü. Viks. *A Concise Morphological Dictionary of Estonian I:* Introduction & Grammar. Tallinn 1992.

(Viks 00) Ü. Viks. *Eesti keele avatud morfoloogiamudel*. (*Open morphology model of Estonian*). In "Arvutuslingvistikalt inimesele." Publications of the Department of General Linguistics, University of Tartu No. 1, pp. 9-36. Tartu 2000.