Optimizing the finite-state description of Estonian morphology

Heli Uibo
Institute of Computer Science
University of Tartu
J. Liivi 2-339, Tartu 50409
Estonia
heli.uibo@ut.ee

Abstract

The research on modeling the Estonian morphology by finite state devices has been influenced mostly by (Koskenniemi, 1983), (Lauri Karttunen and Zaenen, 1992) and (Beesley and Karttunen, 2000). We have used lexical transducer combined with two-level rules as a general model for describing Estonian morphology. As a novel approach we can emphasize the application of the rules to the both sides of the lexical transducer – both to the lexical representation and to the lemma. In the paper the criteria of optimality of the finite-state description of a natural language morphology and the means of fulfilling these criteria are discussed on the example of Estonian – a language with very rich and complex morphology. Other builders of finite-state morphological transducers may profit from the ideas proposed.

1 Introduction

During the last 25 years the finite-state approach has been the most fruitful one in the field of computational morphology. A morphological finite-state transducer describes the correspondence between word forms and their morphological readings (lemma + morphological features) as 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 and Karttunen, 2003). The lexical transducer can be composed with rule transducer(s) that convert lexical representation to surface representation, using either two-level (Koskenniemi, 1983) or replace rules (Karttunen, 1995).

The finite-state description of Estonian morphology has been built up, lead by the principles of the two-level morphology model (Koskenniemi, 1983). The model consists of a network of lexicons and a set of two-level rules. 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, 1983), Turkish (Oflazer, 1994) and Swahili (Hurskainen, 1995), where word forms are built by concatenation of morphemes. Two-level rules are convenient to handle single phoneme alternations. We will show how we have described the Estonian morphology by the means of finite-state devices and discuss the occurred problems and their solutions.

2 Estonian morphology

Estonian is a highly inflected language – grammatical meanings are expressed by grammatical formatives which are affixed to the stem instead of using prepositions. In some cases the

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 203–209
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

analytical forms – adpositional phrases – can be alternatively used (Table 1, but there is often a style difference. According to a more detailed analysis the stem consists of word root and derivational affixes and formative – of features and endings.

Table 1: Inflected forms of a noun and the corresponding adpositional phrases

Case	Word form	Adpositional phrase	Translation
nominative	raamat	-	book
genitive	raamatu	-	book's
partitive	raamatut	-	book (object)
illative	raamatusse	raamatu sisse	into a book
inessive	raamatus	raamatu sees	in a book
elative	raamatust	raamatu seest	out of a book
allative	raamatule	raamatu peale	onto a book
adessive	raamatul	raamatu peal	on a book
ablative	raamatult	raamatu pealt	from a book
translative	raamatuks	-	(become)a book
terminative	raamatuni	kuni raamatuni	up to a book
essive	raamatuna	raamatu kujul	as a book
abessive	raamatuta	ilma raamatuta	without a book
comitative	raamatuga	koos raamatuga	with a book

Based on the morphological behavior, there are three morphological word classes in Estonian:

- nouns (declinables) – substantives, adjectives, pronouns and numerals;
- verbs (conjugables);
- uninflected words (indeclinables) – adverbs, adpositions, conjunctions and interjections.

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). There is no gender distinction in Estonian.

Derivation is mostly done by suffixing:

kiire (Adj) 'quick' – *kiiresti* (Adv) 'quickly'

õppima (V) 'to learn' – *õppimine* (N) 'learning'

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

The word forms in Estonian are constructed by the following morphological processes:

- Agglutination – concatenation of morphemes, whereas morphemes are clearly distinguishable

  a) declination: *probleemi + de + ta = probleemideta* 'problems' – 'without problems';

  b) conjugation: *ela + ksi + me = elaksime* 'live' – 'we would live';

  c) derivation: *rahu + lik = rahulik* 'peace' – 'peaceful';

  d) compounding: *all + maa + raud + tee = allmaaraudtee* 'subway' ("underground railway").

- Flexion – morpheme having the same meaning changes its shape in different grammatical forms, e.g. *tuba : toa* 'room' sg nominative : sg genitive;

- Suppletivity – the forms in the paradigm come from absolutely different stems that historically have been words with similar meanings, e.g. *minema : lähen 'to go : I go', hea : parem 'good : better', üks : esimene 'one : the first', kaks : teine 'two : the second'*;

- Analyticality – multi-word forms

  a) verb forms with auxiliaries, e.g. *oli tehtud 'had been done', on söönud 'has eaten'*;

  b) chain verbs, e.g. *hakkab olema 'will be', paneb põlema* 'switches on';

  c) phrasal verbs, e.g. *alla kirjutama* 'to sign';

  d) idiomatic expressions, e.g. *jalga laskma* 'to escape';

  e) adpositional phrases, e.g. *laua peal* 'on the table', *metsa sees* 'in the forest', *minu järel* 'after me'.

- Reduplication – the repetition of the stem (sometimes in a slightly varied shape).

  This phenomenon occurs in some descriptive adverbs and adjectives only, e.g. *kilin-kõlin 'jingle-jangle', kimpsud-kompsud 'bundles', siiruviiruline 'striate', pilla-palla 'higgledy-piggledy', sahker-mahker 'hugger-mugger'*.

3 Finite-state morphology of Estonian

The morphological phenomena 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 (An example is given on Figure 1.
- The rules of morphotactics have been described in the network of lexicons.
- The stem final alternations have been divided between lexicons and rules.

Figure 1: "b,d,g,s deletion"

```
"b,d,g,s deletion"
  GC:0 <=> V: _ V: $:;
```

(The symbol $ marks the weak grade and GC is the set of gradating consonants.)

The network of lexicons was designed after the morphological classification of Ülle Viks (Ülle Viks, 1992) which is based on pattern recognition. This classification 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. There is a branching inside some noun types according to the stem final vowel in our lexicon. The lexicons of inflection types (noun types 01-26 and verb types 27-38) contain a number of linked lexicons. The first group generates the stem variants, the second group locates the stem variants in paradigm and the third builds the base forms and their analogy forms.

Noun declination, verb conjugation, comparison of adjectives, productive derivation and compounding have been implemented, using the continuation lexicons. Agglutinative processes occuring by declination of nouns and conjugation of verbs have been described by three layers of lexicons (cf. Figure 2):

1. continuation lexicon for each inflection type (lexicon 10_NE-SE-S). There are references to these lexicons from the root lexicons of word classes (Substantive, Verb, Adjective, etc.).
2. allocation of stem variants in the paradigm (lexicons An_SgN, An_SgG ... An_PlP_id)
3. adding of grammatical features and endings (lexicons Cases_1 and Cases_2)

The finite-state description of Estonian is a little unbalanced – the network of lexicons plays the major role, but Viks's type system allows to reuse the automatic inflection type detection module developed for this particular system.

The Estonian finite-state morphology has been implemented using the XEROX tools LEXC, TWOLC and XFST. There are 45 two-level rules. The network of lexicons covers all the inflection types. The stem lexicon contains ca 2500 most frequent word roots, based on the frequency dictionary of Estonian (Kaalep and Muischnek, 2002). Additionally, the network of lexicons include ca 200 continuation lexicons, which describe the stem final changes, noun declination, verb conjugation, derivation and compounding.

4 Optimizing the finite-state description of Estonian

The question could arise, in which sense should the finite-state description be optimal. We can consider it from the point of view of efficiency (computer) and maintainability (human).

- Time- and space-complexity of the resulting transducer should be minimized. It is important, although nowadays the processor speed and amount of operative memory are not any more so critical resource than in 1980s.
- The system of lexicons and rules should be not only machine-readable, but also human-readable and easy to update. Everybody who has tried to build a system consisting from both lexicons and rules knows how complicated it might get to update the rules and lexicons, if the system is not reasonably structured.

This is always to some extent a subjective matter which phenomena to describe by the means of rules and which by using lexicons. As an objective matter, productivity is the key issue here. We have to be aware, how productive are the rules that participate in the word inflection, derivation and compounding processes. If the rule is absolutely productive then it is easy

Figure 2: Linked lexicons describing the agglutinative processes in noun morphology

```
LEXICON Substantive
! Root lexicon of substantives
hobu 10\_NE-SE-S; ! horse
LEXICON 10\_NE-SE-S
! Inflection type for 3-syllable words
! ending in -ne.
:ne An\_SgN; ! Stem variant -ne
             ! (nominative stem)
! continue from the lexicon An\_SgN.
:se AnŽ\_SgG; ! Stem variant -se
              !(genitive stem)
! continue from the lexicon An\_SgG.
:s An\_SgP\_t; ! Stem variant -s
               ! (consonant stem)
! continue from the lexicon An\_SgP\_t...
:s An\_PlG\_te; ! ... or from the lexicon
                ! An\_SgN.
:se An\_PlP\_id;
LEXICON An\_SgN
+Sg+N:0 GI; ! sg nominative +
            ! optional stress particle -gi.
Compound; ! precomponent of a compound
LEXICON An\_SgG
+Sg+G:0 GI;
+Sg:0 Cases\_1;
! Singular cases are built from the
! genitive stem.
+Pl+N:+d GI;
Compound;
LEXICON An\_SgP\_t
+Sg+P:+t GI;
LEXICON An\_PlG\_te
+Pl+G:+te GI;
+Pl:+te Cases\_1; ! Plural cases are built
                  ! from the consonant stem
                  ! + plural feature -te.
LEXICON An\_PlP\_id
+Pl+P:+id GI;
+Pl:+i Cases\_2; ! short plural
LEXICON Cases\_1
! Case endings from illative to comitative
+Ill:+sse GI;
+In:+s GI;
+El:+st GI;
+All:+le GI;
+Ad:+l GI;
+Abl:+lt GI;;
+Trl:+ks GI;
+Ter:+ni GI;
+Es:+na GI;
+Ab:+ta GI;
+Kom:+ga GI;
LEXICON Cases\_2  ! Cases illative
                  ! ...translative
+Ill:+sse GI;
+In:+s GI;
...
+Trl:+ks GI;
```

Figure 3: Lexical representation of stem variants of the word *jõgi* 'river'.

```
          jõgi
jõG=i     jõe
          jõge
          jõkke
```

to formalize either as a rule or a part of the network of lexicons. For example, for regular stem changes we have applied the "many in one"-solution – all the possible stem variants are encoded as a single lexical entry in the root lexicon, using lexical symbols (morphophonemes) that correspond to different phonemes on the surface (Figure 3). Stem internal and phonologically caused stem final changes are handled by lexical symbols. Two-level rules state the legal correspondences between lexical and surface phonemes, depending on the current morphophonological context.

As an example of regular non-phonologically caused stem changes once again consider Figure 2 to see how non-phonologically caused stem final changes e.g. hobune : hobuse : hobust (horse sg nom, gen, part)have been described using continuation lexicons.

Exceptions always cause problems and often lead to inelegant solutions in the language description. Extreme exceptions should be certainly listed in the lexicon. But how to handle the semi-productive morphotactic and phonological rules? We have to choose between introducing lots of new lexical features which trigger the rules and are hard to memorize and writing lots of small lexicons which are linked between themselves and change the network of lexicons to a "spider's net". Beesley and Karttunen (2003) propose a more convenient solution from the human viewpoint – flag diacritics, but it has its own drawback – we can lose in system's efficiency by using it.

Building the finite-state description of Estonian we tackled the fact that it was especially the network of lexicons which got too complicated to maintain. The problems with the rules had more to do with efficiency and they have been solved in rather early stages of the system's development. Therefore, we concentrated on the optimization of the network of lexicons.

Based on our experience we could give to the builders of finite-state lexical transducers the following recommendations:

1. Avoid redundancy in lexicons.

   For example, substantives and adjectives decline similarly, but behave differently as regards to derivation, compounding and comparation processes. However, this can be a compromise between the size and readability.

2. Split up the lexicon into the system of linked lexicons logically.

   One should follow the rules of morphotactics; it is better if the morphemes are not split into smaller parts. This principle can be followed quite naturally for completely agglutinative languages. In the Estonian language, however, the boundary between the morphemes is often blurred, e.g. in the word form *anti'*was given' (the lemma is *and|ma*)the phoneme *t*is "shared" between the stem and formative .

3. Use meaningful lexical forms.

   A good example here is the description of Estonian consonant gradation (Trosterud and Uibo, 2005)where we denote the gradating consonants with corresponding uppercase letters. However, there is a problem with deletion of the phoneme s which can occur as a result of stem internal consonant gradation but it can also be dropped from the stem end in some other inflection types. We have chosen to denote the disappearing s in both cases by uppercase S, but as a consequence the contexts in the S:0 rule come from substantially different sources (Figure 4).

4. Keep the number of lexicons reasonable.

   The general principle is to try to use rules instead of lexicons as much as possible.

5. Minimize the size of root lexicons.

   This is done by avoiding stem doubling in inflection, derivation and compounding and also by unknown word guessing using the phoneme patterns as regular expressions for productive inflection types where the inflection types is unambiguously determined by the phonological shape of the stem.

Figure 4: Rule: Deletion of stem internal and stem final s.

```
S:0 <=> Bgn V [C+] (GC:GCstr) V:  _ StemEnd;
! küngas-künka
        Bgn V GC:GCstr C V  _  StemEnd;
! kobras-kopra
        Bgn V  _ (V) \%$:; ! käsi-käe
        Bgn V V s  _ V \%$:;  ! kauss-kausi
                where GC in  (G B D K P T )
! set of gradating consonants, except for S
                     GCstr  in  (g b d k p t )
! corresponding strong grade phonemes
           matched;
```

5 A step towards better readability and reduction of lexicon size – two-levelness extended

One step towards better readability, satisfying the requirements 3 and 5 from the previous section is our novel idea to use the two-level representation also for **lemma stems** in the root lexicons. We will take a closer look at this approach.

The majority of Estonian verbs are subject to productive derivation processes. The problem arised with the verbs with weakening stem flexion, for which the base form (supine) is in the strong grade (*lugema'*to read'). The morphological information for the derived word form, outputted during morphological analysis, should contain the derived base form, which is sometimes in the weak grade (*loetav 'one that is being read', loetud 'read (finished)', loetu 'one that has been read'*).

The lexical transducer picks up the strong-grade stem and the word class V (verb), but finding a derivational suffix from the word form, it might turn out, that it is a substantive or an adjective with weak lemma. The initial solution was to include the verbs with weakening stems into root lexicons three times – once into the root lexicon of verbs and in both strong and weak grade into the root lexicon of verbal derivatives (Figure 5).

We have found a helpful solution to the weak grade verb derivatives problem: to extend the two-levelness to the left side of the lexical transducer (to the lemma). The approach has been applied for verbs with stem flexion (Figure 6).

Figure 5: Derivation from verb roots: storage consuming solution

```
LEXICON Verb
lugema+V:luGe V2;

LEXICON Verb-Deriv
loe VD0;
luge VD1;
```

Figure 6: Derivation from verb roots: a better solution

```
LEXICON Verb
luGe V2;
```

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 surface form, stem flexion rules have to be applied onto the left side of the lexical transducer. The resulting morphological transducer of Estonian can be formulated as follows:

$((LexiconFST)^{-1} \circ RulesFST_1)^{-1} \circ RulesFST$,

where LexiconFST is the lexical transducer, RulesFST is the rule transducer (the intersection of all the two-level rules) and

$RulesFST_1 \subset RulesFST$

is the intersection of stem flexion rules. The operations used are composition and inversion. More about the extension of two-levelness in Uibo (2005).

The percentage of verbs is about 15 % among the 10000 most frequent words of written Estonian (Kaalep and Muischnek, 2002). Consequently, after the extension of two-levelness the number of records in root lexicons would decrease ca 23 %. The testing and lexicon extending cycle will go on, as the present coverage of the lexicon is about 30 % only.

6 Conclusion and future work

We have given an overview of the finite-state description of Estonian morphology and pointed out the criteria of optimality of the description from the point of view of both efficiency and maintainability of the system. These criteria have arised from the practice. The present finite-state description of Estonian is far from being perfect yet, but we have the ideas how to reorganize the system to improve first of all its human-readability.

We can bring forth the following strengths of the finite-state approach:

- The two-level representation is useful for the description of the Estonian stem internal changes, especially because the stem flexion type 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.
- The lexicons are useful for describing the non-phonologically caused stem end alternations.
- Due to the possibility to compose finite-state transducers we can use an economic solution for modelling productive verbal derivation: we have extended the two-levelness partly to the upper side of the lexical transducer – to the lexical representations of the lemmas of forms productively derivable from the verb roots. The proposed approach may be applied in describing the morphology of languages, where the word stems are subject to change during productive derivational processes.

There are some open problems for which we know in the best case a theoretical solution:

- How to guess the analysis of unknown words? The idea that has so far tested only on a very limited lexicon is to have a root as a regular expression (e.g. CVVCV) in the root lexicon for each productive inflection type.
- The balance between productivity and lexicalization – how complex is it to describe partially productive derivation types by minilexicons and continuation links (instead of including the derivatives into stem lexicons as independent stems)? Which derivation types to consider productive enough? Which are the formal features that could be used to handle some processes in derivation by rules?
- How to constrain the overgeneration of compound words? The idea is to apply

the semantic features. An alternative idea is to use weighted finite-state transducers trained on the corpus data.

- To include the finite-state 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.

7 Acknowledgments

The research on finite-state morphology of Estonian has been supported by the Estonian Science Foundation grant No. 4605 (2001-2003). Our thanks also go to Kimmo Koskenniemi, Lauri Karttunen, Kenneth Beesley and Trond Trosterud for encouraging discussions.

References

Kenneth R. Beesley and Lauri Karttunen. 2000. Finite-state non-concatenative morphotactics. In *Proceedings of SIGPHON-2000. 5th Workshop of the ACL Special Interest Group in Computational Phonology*, pages 1–12, Centre Universitaire, Luxembourg.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, USA.

Arvi Hurskainen. 1995. Information retrieval and two-directional word formation. *Nordic Journal of African Studies*, 4(2):81–92.

Heiki-Jaan Kaalep and Kadri Muischnek. 2002. *Eesti keele sagedussõnastik (The frequency dictionary of written Estonian)*. University of Tartu Press, Tartu.

Lauri Karttunen. 1995. The replace operator. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. ACL-95*, pages 16–23, Boston, Massachusetts.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-form Production and Generation*. Publications of the Department of General Linguistics, University of Helsinki. University of Helsinki, Helsinki.

Ronald M. Kaplan Lauri Karttunen and Annie Zaenen. 1992. Two-level morphology with composition. In *Proceedings of the 14th conference on Computational linguistics*, volume 1, pages 141–148, Nantes, France.

Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Trond Trosterud and Heli Uibo. 2005. Consonant gradation in estonian and s?mi: two-level solution. In *Inquiries into words, constraints and contexts: Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*, CSLI Studies in Computational Linguistics ONLINE, pages 136–150. CSLI Publications.

Heli Uibo. 2005. Finite-state morphology of estonian: two-levelness extended. In *Proceedings of Iternational Conference Recent Advances in Natural Language Processing. RANLP 2005*, pages 580–584, Borovets, Bulgaria.

Ülle Viks. 1992. *A Concise Morphological Dictionary of Estonian*, volume 1. Eesti Teaduste Akadeemia, Keele ja Kirjanduse Instituut, Tallinn.