Experimental Two-Level Morphology of Estonian

Heli Uibo

Institute of Computer Science, University of Tartu, Estonia
J. Liivi 2 - 317, Tartu 50409, Estonia
heli_u@ut.ee

Abstract

The experimental two-level morphology of Estonian is under development at the University of Tartu. The language description, consisting of 45 two-level rules and over 200 lexicons has been implemented and tested using Xerox finite-state tools *twolc* and *lexc*. The root lexicons cover 400 most frequent stems at the present stage of development. The software has been designed to update the lexicon automatically with new stems, including the automatic generation of lexical representations of root lexicon entries. The open problems by describing of word formation processes – derivation and compounding are discussed. The advantages and disadvantages of the two-level model with respect to Estonian morphology are pointed out.

1. Introduction

Morphological component is an important language resource for Estonian language technology. Morphology is needed even in information retrieval systems, where it is usually desirable to make queries using semantic entities, not using special morphological forms of a word. As the word stem often has several shapes in Estonian (often two, but sometimes even four), the lemmatiser would be of significant help.

It is hard to overestimate the complexity of design and implementation of morphological analysers and generators for Estonian. Estonian is a language with a complicated morphology featuring rich inflection and marked and diverse morpheme variation, applying both to stems and formatives (Viks 2000b).

There exist two versions of automatic morphological analyser for the Estonian language – analyser **estmorf** based on so-called unification morphology by Kaalep (2000) and **rule-oriented morphological analyser** by Viks (1995; 2000a). Both analysers can identify the inflectional forms of a word in a text, i.e. establish the stem and case endings, inflected verb forms, etc.

At the same time, in computational morphology, **the Two-Level Morphology** formalism, proposed by Kimmo Koskenniemi (Koskenniemi 1983), has developed nearly to world-wide standard during the past 10-15 years. The language processing tools, based on this formalism, are efficient, due to the small computational complexity of the finite-state automata. Moreover, the language description is represented declaratively in this model and separated from the application programs, instead of hiding it into the program code.

The two-level morphology model has been proved successful for formalising the morphologically very different languages (English, German, Swedish, French, Spanish, Danish, Norwegian, Finnish, Russian, Turkish, Arab, Aymara, Swahili etc.). Thus, we can expect that the model is in fact universal and it may be possible to describe Estonian morphology in this framework as well.

2. Two-Level Description of Estonian Morphology

We are developing the experimental two-level morphology for Estonian (**EETwolM**) – lexicons and two-level rules.

2.1. Lexicons

Evidently, the model is especially well suitable for languages with agglutinative morphology, e.g. Turkish (Oflazer 1994; 2001) and Finnish (Koskenniemi 1985), as the network of lexicons naturally defines the morphotactics of word-forms.

The general rules of morphotactics in Estonian are as follows:

1. The rule for inflected nouns

Noun-form = stem + number + case
e. g. *toolidel = tooli + de + l*
on chairs = chair + plural + adessive

2. The rule for inflected verbs

Verb-form = affirmation/negation + stem + voice (personal/impersonal) + tense (present, past, perfect, pluperfect) + mood (indicative, imperative, conditional, oblique) + number + person

One variant of every category is unmarked, whereas the remaining variants are marked by certain features.

Examples:
singular – unmarked; plural – marked (-d, -de, -te);
indicative mood – unmarked; imperative mood – marked (-gu, -ge), conditional mood – marked (-ks), oblique mood – marked (-vat)

There have been several systems of inflection types for Estonian language. Those, designed for human and printed in the appendices of dictionaries, usually contain ca 100 types. Viks (1994) proposed a new morphological classification for Estonian, based on pattern recognition. It is much more compact and specially designed to formalise the morphological system of Estonian for computers.

The Viks'es type system includes 38 regular inflection types – 26 for nouns and 12 for verbs. 84 words do not belong to any type and are treated as exceptions. Pronouns are handled as exceptions as well.

The network of lexicons in EETwolM has been built up using this type system. Additionally, most of the lexicons of inflection types diverge additionally according to the stem final vowel. During historical processes the stem final vowel has disappeared from the end of lemma (singular nominative) and it is not predictable after the phonological shape of a word stem, which of the four possible stem vowels (**a, e, i, u**) will be used.

In some inflection types, also the third stem variant – short plural stem exists. For stems ending with -e, -i or -u the choice plural stem vowel is regular (Table 1). For words ending with **–a** the plural vowel depends on the phonological shape of the stem and even from those complicated rules there are some exceptions. Thus, we decided to write the plural stem vowel directly to lexicon for words having the –a-stem in singular.

Singular stem vowel	Plural stem vowel
-e	**-i**
-I	**-e**
-u	**-e**

Table 1. Dependence of the plural stem vowel from the singular stem vowel.

Nearly 86 % of the nouns are subject to some kind of changes of stem end. Next, a short overview of possible alterations of stem end will be given.

(1) addition of stem final vowel (*kass:kass**i***)

(2) deletion of stem final **s** (*sipelga**s**: sipelga*)

(3) addition of stem final vowel combined with deletion of the vowel (usually **e**) before the last consonant **(l, m, n, r)** (*vank**e**r:vankr**i**, piib**e**l:piibl**i***)

(4) ne-se-alternation (*hobu**ne**:hobu**se***)

(5) addition of syllable **–da** or **–me** (*tore:tore**da** habe:habe**me***)

Two first alternation types can be co-occur with stem internal changes (gradation). Only second and third alternation type are handled by two-level rules in EETwolM. Still, the model gives the opportunity to handle such “unnatural” stem changes as (4) and (5) conveniently by small lexicons.

The network of lexicons includes ca 200 additional lexicons as the result of divergence within inflection types that describe the inflection processes according to Viks (1994). Some very general rules of derivation and compounding have been also implemented. The root lexicons of substantives, verbs and adjectives contain 400 most frequent Estonian stems in total. The frequency lexicon has been extracted from the corpus of written Estonian (Hennoste et al 1998).

The network of lexicons is quite hard to administrate, if the morphotactis of the language is not the simplest one (the same problem pointed out by Oflazer (2001)) due to the large number of small lexicons. To avoid overgeneration in word formation and to handle exceptions, Oflazer (2001) also used finite-state constraints. It seems to be right choice for handling Estonian derivation and compounding as well.

2.2. Rules

Two-level rules are convenient tool to handle various regular stem alternations evolving only one phoneme. As an opposite to lexicons, the number of rules never grows over 50 and usually the rules are quite simple.

One of the major phenomena to be described by rules is certainly Estonian **consonant gradation**.

One can differentiate between the following kinds of gradation in Estonian (Viks 1979):

(1) alternations of III and II duration, not expressed by written form: s*oovida - soovin, vang - vangi.* (This kind of gradation should be handled, if the III duration has been coded in the lexical representation: s‘*oovida - soovin, v‘ang - vangi.*

Designing EETwolM we assumed the written language as the surface level. But the suitable markers can be introduced into root lexicons in principle.

(2) alternation of long and short geminate (*ku**kk**uda - ku**k**un, avan**ss**i - avan**s**i*);

(3) alternation of strong and weak stops (*lam**p**i - lam**b**i, tõr**k**uda - tõr**g**un*);

(4) assimilation (ka**nd**ma - ka**nn**an, va**rs** - va**rr**e);

(5) replacement of a weak stop by rules b:v, d:j (*kae**b**ama - kae**v**ata*, *ra**d**a - ra**j**a*);

(6) deletion of a stop (k,p,t,g,b,d) or s (nu**g**a - noa, käs**k**ida - käsin, ve**s**i - vee).

Hint stresses in his dissertation (1997) that the gradation is not any more a phonological, but morphophonological phenomenon in present day Estonian. The reasons for gradation have disappeared from the word-forms, the gradation itself is a kind of relict. Some kinds of gradation (especially qualitative changes – assimilation, replacement and deletion) are not productive any more. Young people prefer to decline some of the traditionally gradated words so that the stem does not change internally. New words do not join to the inflection types having the feature of qualitative stem internal changes.

As the stem gradation is not phonologically caused any more, the words that undergo those changes should be marked in the lexicon. We have used capital letters (B,D,G,K,P,T,S) to denote the gradation in the stem. The rules are mostly sensitive to these special capital letters only. Moreover, there are weak grade markers placed in the right positions in the lexicons of inflection types. These are indicators for two-level rules, which handle stem flexion.

Default correspondences are: B:b, D:d, G:g, S:s, K:k, P:p, T:t and S:s.

For EETwolM 45 two-level rules have been written that deal with stem flexion, phonotactics, orthography and morphophonological distribution.

3. Implementation

The rules and lexicons have been developed and tested using Xerox finite-state software tools **twolc** (Karttunen, Beesley 1992) and **lexc** (Karttunen 1993).

The current work in progress includes the enlargement of the coverage of root lexicons. “The development of the morphological analyser is an iterative process, whereby the human informant will revise and/or refine the information previously elicited based on the feedback from test runs of the nascent analyses. The iterative process will converge to a wide-coverageanalyser coming slowly at the beginning (where morphological phenomena and lexicon abstractions are being defined and tested), but significantly speeding up once wholesale root form acquisition starts.” (Oflazer et al. 2001) We hope to reach the stage of the iterative process soon where it becomes faster. We have designed a software for updating the lexicon automatically. It uses the type recognition rules by Viks (1995; 2000a) and an algorithm constructed by ourselves. The algorithm derives the lexical representation for any new word which should be included into root lexicon, based on the inflection type given by type detection module and the phonemes of the first syllable of the given word.

4. Discussion

4.1. Word Formation Problems

The present system performs normally on simple inflected nouns and verbs but the word formation processes have not been described in the sufficient extent yet. Both derivation and compounding are partly very productive in Estonian, but too general rules cause overgeneration.

There is a need to study derivation and compounding processes in present day Estonian using the corpus to formalise these processes more specifically in lexicons and rules. The theoretical studies give very few material in this aspect of morphology.

Viks (2000a) has given the following general rules for derivation:

(1) The suffix -*mine* can be added unconditionally to any verb stem (*hakka|ma* → *hakka|mine, leid|ma* → *leid|mine*). The result is a substantive, denoting the process expressed by the verb.

(2) The suffix -*mus* can be added to the verb stem, if it is ended by -*ne*, -*i* or -*u* (*paljune[ma* → *paljune|mus*, *leppi[ma* → *leppi|mus*, *h`arju[ma* → *h`arju|mus*). The result is an abstract substantive.

(3) The adverbial suffix -*lt* can be added to the stem variant of the adjective that is used in singular genitive. (ablas:apla → apla|lt, punane:punase → punase|lt, kurb: kurva → kurva|lt);

(4) The substantive suffix *–us* can be added to the inflected stem (in strong grade or unchanged stem) (ablas : apla → apl|us, napp : nappi → n`app|us, tore : toreda → tored|us, kasulik : kasul`ikku → kasul`ikk|us).

It is quite difficult to find any rules for compounding. One that is sure is that they should take into account lexical semantics.

Finally, let us consider the problem of word formation in the point of view of spelling checkers, as it is the most important application of morphological analysis and synthesis.

Generally, the word formation, including derivation and compounding, is free in Estonian. Everybody can invent new words, never used before, just adding several stems one after another. It would be annoying if spell checker would underline those words with red colour. Thus, the list of possible compounds and derivatives cannot be finite. On the other hand, if the morphological description allows very productive word formation, other kind of problems occur. One half of the words in Estonian texts is homonymous in average. Sometimes also derivatives and compounds happen to have the same spelling as a non-compound inflected word-form.

If we allow unrestricted word formation, depending e.g. only from the word class, the result is, that lots of the spelling errors remain out of sight of the speller, as it counts the word-form for grammatical derivative or compound.

Let us consider the following examples.

1. *naljakass* = *nalja+kass* S Sg Nom (*cat of fun*) – possible, but weird compound
The probability of occurrence of the word is nearly 0. What actually was meant, is the word *naljakas (funny)*, but occasionally the writer held his finger too long on the key '**s**' at the end of the word.

2. **kaustatud* instead of *kasutatud*
Typing quickly, it is very easy to make typos like this. If we apply the principle of analogy, this kind of derivation could be possible from any substantive:

õnn (*happiness*)-> *õnnetu* (*unhappy*)
kodu (*home*) -> *kodutu* (*homeless*)
kaust (*folder*) -> **kaustatu* (*one having no folder*)

But again, nobody would use the derived form **kaustatud*, at the same time the wordform *kasutatud* (*used*) is very frequent in everyday language usage.

It also depends on the user of word processor – some people prefer to be informed about every possible spelling mistake they do, others get angry, if their style of building new words is considered ungrammatical.

How to find the proper extent of productivity of word formation rules, is still an open problem.

4.2. Suitability of the Two-Level Model for Describing Estonian Morphology

Estimating the suitability of the two-level morphology model to Estonian language one can point out the following positive features (Uibo 2000):

1. Using the deep representation is an advantage because the lexical entries can include other information additionally to the pure phonemic consistence of a word or its part:

1.1. There is a possibility to use special denotations for phonemes having more than one surface variant. This is a great advantage, as the type of stem flexion generally does not depend on the phonemic shape of the stem in the present-day Estonian (Some kinds of stem flexion are not productive any more.)

1.2. The lexical information can contain morphophonological features and morpheme boundaries, which are often used by rules.

2. The rule set is not ordered. The compilation of ordered rule set would be complicated because it is difficult to count the influence of all the previous rules in the sequence to the left and right context.

3. A rule can point to the arbitrarily far context. E.g. there can be a rule which should check the stem final character, without knowing the number of syllables.

4. The net of lexicons is convenient to handle
non-phonologically caused stem end alternations;
rules of morphotactics;
productive derivation and compounding (partly).

The difficulties occurred while compiling the Estonian experimental two-level morphology:

Converting a word segment as a whole (e.g. *joo|ma-juu|a*), is not possible and causes two or more rules to be co-ordinated.

The introducing of word lists independent of rules of morphotactics is very difficult in the lexicon system - it increases the amount of different lexicons very quickly (e.g. for compound-words generation semantics-based lists are needed).

There is a serious problem in derivation process: as the derivation often changes the word class there should be the possibility to "turn back" and delete the original morphological information belonging to the derivation base. The implemented solution to get the correct result is not elegant: the verb stems have been doubled in the lexicon of verb derivatives and the assignment of morphological information has been deferred in the case of deriving adjectives from nouns.

5. Conclusion

Concluding, the experimental two-level morphology of Estonian has shown that the model is quite usable for Estonian simple word recognition and production.

However, the derivation and compounding processes need to be studied more thoroughly to avoid overgeneration.

The efficiency of the implementation of the rules and lexicons as finite-state transducers is undoubtedly an advantage in the program's performance.

Unfortunately, the objective comparison with the other lemmatisers and spelling checkers is not possible, as the Estonian two-level lexicon's coverage is not comparable with the lexicons of the working systems yet.

6. Acknowledgements

The research on the experimental two-level morphology of Estonian is supported by Estonian Science Foundation under Grant No. 4605.

7. References

Hennoste, T., Koit, M., Roosmaa, T. & Saluveer, M. 1998. Structure and Usage of the Tartu University Corpus of Written Estonian. *International Journal of Corpus Linguistics*. Vol 3(2), 1998, 279--304.

Hint, M. 1997. Typological problems of Estonian grade alternation and prosodical system. Dissertation at the University of Helsinki, Dept of Baltic-Finnic Languages.

Kaalep, H.-J. & Vaino, T. 2000. Full Morphological Analysis in the Toolbox of a Linguist. Abstracts. *Congressus Nonus Internationalis Fenno-Ugristarum. Pars II* Tartu 2000, 342--343.

Karttunen, L. & Beesley, K. R. 1992. Two-level Rule Compiler. Technical Report. ISTL-92-2. October 1992. Xerox Palo Alto Research Centre. Palo Alto, California.

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

Koskenniemi, K. 1983. 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.

Koskenniemi, K. 1985. An application of the two-level model to Finnish. In Fred Karlsson, editor, Computational morphosyntax: a report on research 1981 – 1984. University of Helsinki, Dept of General Linguistics.

Oflazer, K. 1994. Two-Level Description of Turkish Morphology. *Literary and Linguistic Computing* Vol. 9 (2), 175--198

Oflazer, K. 2001. Two-Level Morphology and Finite-State Methods: a Consumer's View. Panel presentation at the event "Twenty years of Finite-State Methods", Aug 22, 2001, University of Helsinki.

Oflazer, K., Nirenburg, S. & McShane, M. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics* Vol. 27 (1)

Uibo, H. 2000. On Using the Two-level Model as the Basis of Morphological Analysis and Synthesis of Estonian. In Proceedings from the 12th "Nordiske datalingvistikkdager" (pp. 228--242). Ed. by Torbjørn Nordgård. Trondheim: Department of Linguistics, NTNU.

Viks, Ü. 1979. Kuidas formaliseerida tüvemuutusi. (How to formalize the stem alternations). *Keel ja Kirjandus* No. 11, 671-677. (in Estonian)

Viks, Ü. 1994. Eesti keele klassifikatoorne morfoloogia. (Classificatory morphology of Estonian) *Dissertationes philologiae Estonicae Universitatis Tartuensis* ; 1

Viks, Ü. 1995. About rule-oriented morphology of Estonian. In Abstracts of Posters Presented at the 10th Nordic Conference of Computational Linguistics NODALIDA-95 (pp. 28--30). Helsinki.

Viks, Ü. 1995. Rules for Recognition of Inflection Types. In Automatic Morphology of Estonian 2. (Research Reports) Ed. by Ü. Viks. (pp. 23--45). Tallinn, Insitute of Estonian Language.

Viks, Ü. 2000a. Eesti keele avatud morfoloogiamudel (Open Morphology Model of Estonian) – In *Arvutuslingvistikalt inimesele*. Dept of General Linguistics, University of Tartu. Publications No. 1. (pp. 9--36). Tartu, University of Tartu. (in Estonian)

Viks, Ü. 2000b. Tools for the Generation of Morphological Entries in Dictionaries. In Proceedings of the 2nd International Conference on Language Resources and Evaluation LREC2000. Athens