On Using the Two-level Model as the Basis of Morphological Analysis and Synthesis of Estonian

Heli Uibo
Institute of Computer Science, University of Tartu, Estonia
heli_u@ut.ee

The paper deals with the problems of describing the Estonian morphological system in the two-level formalism, developed by Kimmo Koskenniemi. The outlines of Estonian morphology are drawn. The basics of the two-level model are given and illustrated with real examples from the experimental Estonian two-level morphology (**EETwoLM**) composed by the author. A detailed example of step-by-step morphological synthesis is given referring to all the relevant lexicons and rules. The compilation and testing processes using the XEROX finite-state software tools are described. Examples of morphological analysis and synthesis are demonstrated. The present stage of the system is characterised and the future perspectives are drawn. Finally, the suitability of the two-level model for the description of Estonian morphology is discussed.

1. Introduction

The module of morphological analysis and/or synthesis is unavoidable in any language engineering tool for Estonian because of its rich morphology. For example, in information retrieval systems, 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, the morphological component should belong to any information retrieval system.

Example 1. To make a query for all occurrences of the word "*rida*" ("row"), without a morphological synthesiser, three different queries are needed (we assume the possibility to add star (*) to the end of the stem instead of inflectional suffices):

Proceedings of NODALIDA 1999, pages 228-242

*rida** (stem in strong grade)

*rea** (stem in weak grade)

ritta (singular additive ("to the row"), quite often used with this word)

The development of EETwoLM is not the first attempt to computerise the morphological analysis and synthesis of Estonian. Ülle Viks (1994) has done important research in the field of morphological classification of Estonian on the basis of pattern recognition theory starting from the 1970s. Viks (1992) has compiled the first morphological dictionary for Estonian as a practical output of the investigations. Further, E. Kuusik and Ü. Viks (1998) have implemented the rule-based morphological analysis and synthesis for Estonian. Heiki-Jaan Kaalep (1996) has developed the speller for Estonian using the results of Viks (1992).

Nevertheless, the growing popularity of the two-level model encourages us to consider its suitability to Estonian morphology. From the practical point of view, the appropriate description of Estonian morphology in the form of lexicons and two-level rules makes the significant move towards the application of Xerox language engineering tools to Estonian language (www.xerox.com/xrce/mltt).

2. The Brief Overview of Estonian Morphology

Estonian language is a member of Finno-Ugric family and is a close relative to Finnish. Estonian morphology is complex – inflected word-forms are built using both agglutination and stem flexion. Nouns have 14-15 cases. Plural forms often have parallel forms.

Table 1. Noun paradigm (word *"käsi"* ("hand")).

Case	Abbrev-iation	Singular		Plural	
		Word-form	Meaning	Form	Meaning
Nominative	N	*käsi*	(the) hand	*käed*	hands
Genitive	G	*käe*	of the hand	*käte*	of the hands

Partitive	P	*kätt*	hand (partial object)	*käsi* etc.
Illative	Ill	*käesse*	into the hand	*kätesse* or *käsisse*
Inessive	In	*käes*	in the hand	*kätes* or *käsis*
Elative	El	*käest*	out of the hand	*kätest* or *käsist*
Allative	All	*käele*	(on)to the hand	*kätele* or *käsile*
Adessive	Ad	*käel*	(up)on, at the hand	*kätel* or *käsil*
Ablative	Abl	*käelt*	from the hand	*kätelt* or *käsilt*
Translative	Trl	*käeks*	for, as the hand	*käteks* or *käsiks*
Terminative	Ter	*käeni*	up to, until the hand	*käteni*
Essive	Es	*käena*	as the hand	*kätena*
Abessive	Ab	*käeta*	without the hand	*käteta*
Comitative	Kom	*käega*	with the hand	*kätega*
Additive	Adt	*kätte*	(in)to the hand	-

Verbs have the following categories in Estonian: person (singular 1st to plural 3rd), voice (personal, impersonal), tense (present, imperfect, perfect, past perfect), mood (indicative, conditional, imperative, quotative).

Table 2. Part of the verb paradigm – infinite verb forms and indicative mood of the finite verb forms (word "*muutma*" ("to change")).

1. Infinite (declined) forms		
Morphological meaning	Abbreviation	Example
Supine (illative)	Sup	*muutma*
Supine inessive	Sup In	*muutmas*
Supine elative	Sup El	*muutmast*
Supine translative	Sup Tr	*muutmaks*
Supine abessive	Sup Ab	*muutmata*
Supine impersonal	Sup Ips	*muudetama*
Infinitive	Inf	*muuta*
Gerund	Ger	*muutes*

Participles (Pts):			
Present (Pr) participle	Personal (Ps)	Pts Pr Ps	*muutev*
	Impersonal (Ips)	Pts Pr Ips	*muudetav*
Past (Pt) participle	Personal	Pts Pt Ps	*muutnud*
	Impersonal	Pts Pt Ips	*muudetud*

2. Finite (conjugated) forms					
2.1. Indicative mood (Ind)					
Personal voice (Ps)					
Tense		Number	Person		
Present	Affirm-ation	Sg	1.	Ind Pr Ps Sg1	*muudan*
			2.	Ind Pr Ps Sg2	*muudad*
			3.	Ind Pr Ps Sg3	*muudab*
		Pl	1.	Ind Pr Ps Pl1	*muudame*
			2.	Ind Pr Ps Pl2	*muudate*
			3.	Ind Pr Ps Pl3	*muudavad*
	Negation (Neg)			Ind Pr Ps Neg	*ei muuda*

Imperfect	Affirm-ation	Sg	1.	Ind Ipt Ps Sg1	*muutsin*
			2.	Ind Ipt Ps Sg2	*muutsid*
			3.	Ind Ipt Ps Sg3	*muutis*
		Pl	1.	Ind Ipt Ps Pl1	*muutsime*
			2.	Ind Ipt Ps Pl2	*muutsite*
			3.	Ind Ipt Ps Pl3	*muutsid*
	Negation			Ind Ipt Ps Neg	*ei muutnud*
Present perfect	Affirmation			Ind Pf Ps	*on muutnud*
	Negation			Ind Pf Ps Neg	*ei ole muutnud*
Past perfect	Affirmation			Ind Ppf Ps	*oli muutnud*
	Negation			Ind Ppf Ps Neg	*ei olnud muutnud*
Impersonal voice (Ips)					
Present	Affirmation			Ind Pr Ips	*muudetakse*

	Negation	Ind Pr Ips Neg	*ei muudeta*
Imperfect	Affirmation	Ind Ipt Ips	*muudeti*
	Negation	Ind Ipt Ips Neg	*ei muudetud*
Present perfect	Affirmation	Ind Pf Ips	*on muudetud*
	Negation	Ind Pf Ips Neg	*ei ole muudetud*
Past perfect	Affirmation	Ind Ppf Ips	*oli muudetud*
	Negation	Ind Ppf Ips Neg	*ei olnud muudetud*

3. The Outlines of the Two-level Model, Illustrated with Examples in Estonian

The two-level morphology model was proposed by Kimmo Koskenniemi in his dissertation (1983). By now, the model has been used for morphological parsing of English, German, Swedish, Norwegian, Danish, Finnish, French, Turkish, Swahili etc. The main features of the two-level model are the following:

- The language description, consisting of rules and lexicons, is **separated** from the application programs.
- The model is **bidirectional** – it is oriented to morphological analysis as well as to morphological synthesis.
- **The two-levelness** of the model means that the deep representations of morphemes rather than morphemes themselves are maintained in lexicons. From those all the real word-forms can be produced with the help of two-level rules and links between lexicons.

Example 2. The lexical and surface representation of the word-form "*kaob*" ("disappears").

Lexical representation: k a D u $ + b

Surface representation: k a 0 o 0 0 b

The surface representation of a word-form is theoretically a sequence of phonemes. Practically, it tends to be the written form because of the better availability of written texts, as mentioned by Koskenniemi (1997:101).

The lexical representation can contain

a) surface phonemes ("k", "a", "u", "d", "b");
b) lexical phonemes ("D" corresponds to d in the strong grade and either disappears or assimilates in the weak grade);
c) special symbols for morpheme boundaries and morphophonological features ("+" indicates the boundary between stem and inflectional ending, "$" is the weak grade marker).

The representations are aligned with zero-characters.

- **Rules** and **lexicons** are two major parts of the model.
- **The set of rules** is like a filter, through which the lexical representation can be seen as surface representation and vice versa.
- The rules are **not ordered** and all of them have to be satisfied at the same time.
- Rules are implemented as **finite-state automata**.
- A finite-state automaton can be represented as a regular expression, thus the rules are coded as **regular expressions**.

Example 3. Example of a rule: "The disappearance of **D** in the weak grade"

D:0	⇔	SylBg Vow Vow: _ (StemFinVow:) %$:;	!*laud-laua*
		Vow Vow Liq _ StemFinVow %$:;	!*siirdama-siirata*
		[e | i: | u: | ü:] _ StemFinVow: %$: ;	!*vedama-vean, rida-rea*
		õ _ e %$: ;	!*õde-õe, põdeda-põen*
		[Cons - [r|s]] a _ u: %$: ;	!*kaduda-kaon*

The rule should be read as "The lexical symbol D corresponds to zero-character (i.e. disappears on the surface) in one of the following contexts and only there." In a context description the underline character "_" denotes the place of the pair **D:0** between the left and right contexts.

There is a possibility to define sets of characters to make the rules shorter and more readable, e.g. **Vow** stands for vowels, **StemFinVow** for possible stem final vowels, **Cons** for consonants. Sometimes it is also convenient to give names to frequent word segments, e.g. **SylBg** means the beginning of a syllable. Note that there is "$" (the weak grade marker) at the end of each right context, thus the disappearance takes place only in the weak grade.

The exclamation mark indicates the beginning of comments – we have provided every context with 1-2 example-words that help to understand the context.

- **The network of lexicons** consists of a stem lexicon and a number of small lexicons describing stem end alternations, inflectional and derivational processes.
- The network of lexicons is implemented as a **finite-state transducer**.
- A lexical entry includes **morphological information**, **lexical representation** and the name of the **next lexicon**.

Example 4. The structure and co-operation of lexicons.

```
LEXICON V11 !Conjugation class 1, endings which can be added to infinitive stem
+Inf:+da #;
+Ger:+des #;
+Kvt+Pr:+vat #;
+Ind+Pf:+nud #;
+Ind+Ipt: Ipt2;
+Ind+Imp+Pr: Imp;
LEXICON Ipt2 !The imperfect tense endings for the word types "elama" and "õppima"
 +Sg1:+sin #;
 +Sg2:+sid #;
 +Sg3:+s #;
 +Pl1:+sime #;
 +Pl2:+site #;
 +Pl3:+sid #;
LEXICON Imp  !The endings for imperative mood, except Sg2.
 +Sg1:+gu #;
 +Sg3:+gu #;
 +Pl1:+gem #;
 +Pl2:+ge #;
 +Pl3:+gu #;
```

Let us illustrate the details of information represented in the lexicons with two records:

+Inf	**:**	**+da**	**#**
+Ind+Imp+Pr	**:**		**Imp**
morphological information	separator	lexical representation	link to the next lexicon ('#' – the end of word-form)

If we take the word "*õppima*" ("to learn"), the LEXICON V11 builds the following forms:

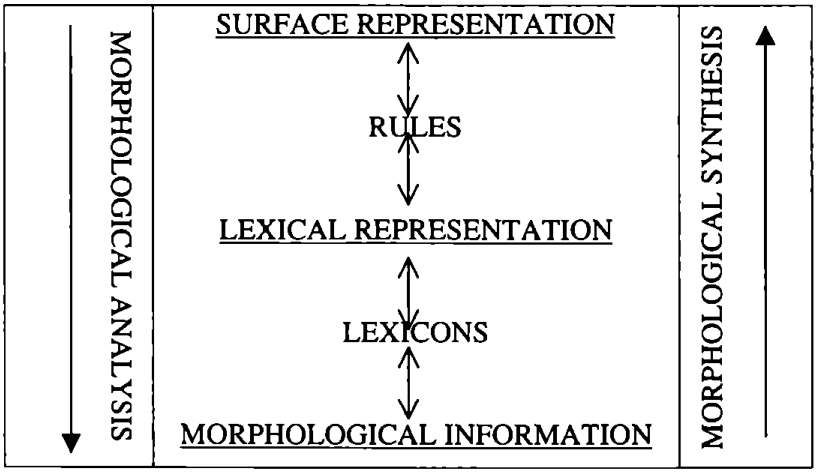
da-infinitive	- *õppida*
gerund	- *õppides*
quotative mood, present tense	- *õppivat*
indicative mood, present and past perfect tense	- *õppinud*

Further word-forms can be built going along the links to the lexicons Ipt2 and Imp:

indicative mood, imperfect tense	- *õppisin, õppisid, õppis, õppisime, õppisite*
imperative mood, present tense	- *õppigu, õppigem, õppige*

As have been said previously, the model can be the basis for morphological analysis as well as for synthesis. Both analysis and synthesis mean the sequential application of the rule automata and the lexical transducer, but in different order, as seen on figure 1.

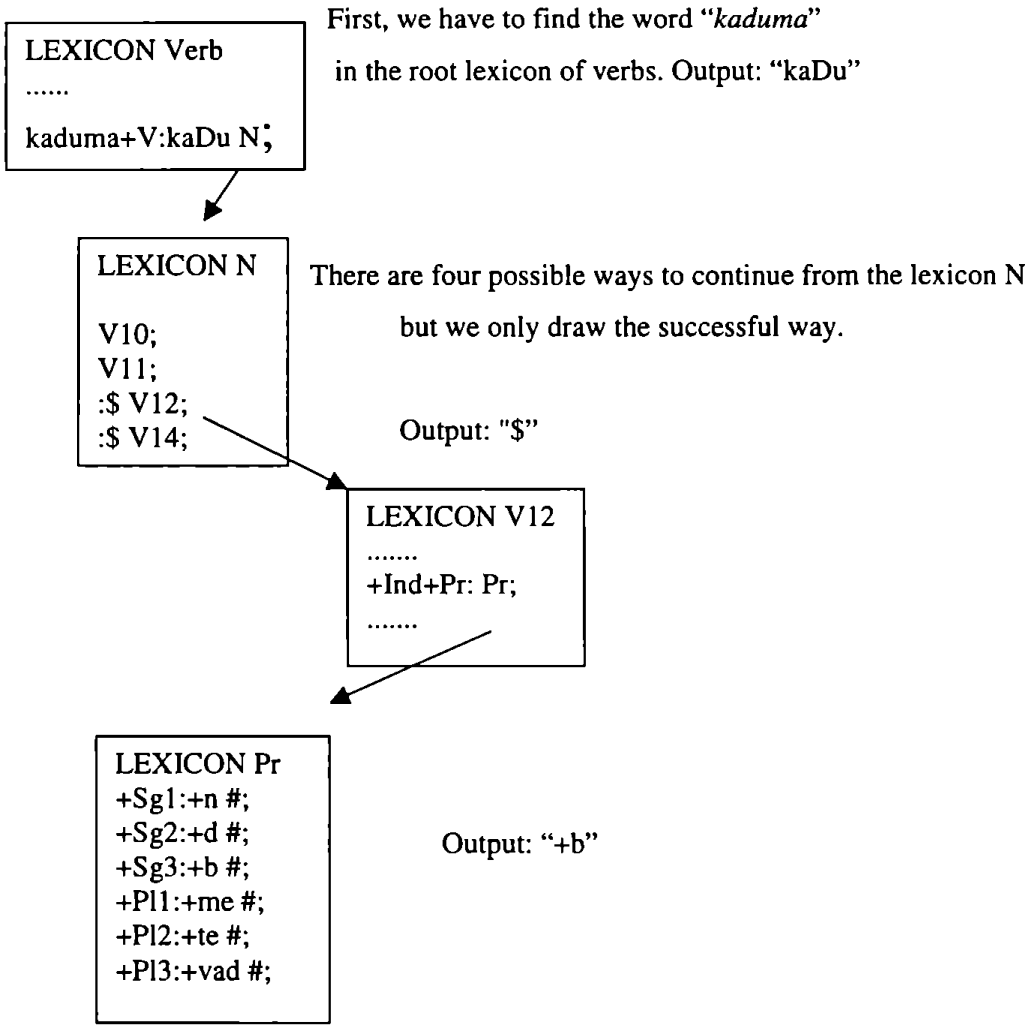
Figure 1. Morphological analysis and synthesis in the two-level model.



4. A Detailed Example of Morphological Synthesis

Example 5. The synthesis of the word-form "*kaob*" (verb "*kaduma*", indicative mood, present tense, singular, 3rd person).

Input: "kaduma+V+Ind+Pr+Sg3"



Thus, moving along the links between lexicons the **lexical representation "kaDu$+b"** has been composed.

Next, the rules will be applied. The rule "Disappearance of D" is satisfied with the pair D:0 in the context k a _ u: $: (see the last context). Thus we get "ka0u$+b".

```
Rule "Disappearance of D"
D:0 ⇔ SylBg Vow Vow: _ (StemFinVow:) %$:;
          Vow Vow Liq _ StemFinVow %$:;
          [ e | i: | u: | ü: ] _ StemFinVow: %$: ;
                        õ _ e %$:
             [Cons- [rls] ] a _ u: %$: ;      !kaduda-kaob
```

```
Rule "Lowering of Vowels"
HighVow:LowVow ⇔ SylBg _ DCons: [aleli:lu:](l) %
        SylBg Vow DCons: _ %$: ;
            where HighVow in (u ü i)
                  LowVow in (o ö e)
            matched ;
```

The second rule accepts the pair u:o in
the context k a D: _ $: . The result is "ka0o$+b".

After applying the whole rule set the default pairs "$:0" and "+:0" have their turn. The result is "ka0o00b". After the deletion of zero-characters **the surface representation "kaob"** is ready.

5. The Implementation of the Two-level Model Using the XRCE Software

The rules and lexicons were developed and tested using the XEROX finite-state tools *lexc* (finite-state **lex**icon **c**ompiler developed by L. Karttunen (1993)) and *twolc* (**two**-level rule **c**ompiler developed by L. Karttunen & K. Beesley (1992)).

The process of testing the correctness and consistency of the lexicons and rules usually proceeds as follows:

- The rule and lexicon files have to be composed in a word processor in the described formats.
- Rules coded as regular expressions are compiled to automata with the two-level rule compiler "*twolc*". Lexicons are compiled into a lexical transducer with the lexicon compiler "*lexc*" .
- Next, the rules and lexicons can be composed with the help of the program "*lexc*".
- There are some possibilities to test the correctness of the language description in the program "*lexc*". One can analyse single word-forms using the directive "*lookup* <word-form>" and produce word-forms using the directive "*lookdown* <primary form+morphological information>"). We can use the directive "*random-surf*" for generating word-forms randomly using the existing lexicons and rules.

Example 6. Test of morphological analysis and synthesis using the program *lexc*.

lexc> lookup pead

pea+S+Sg+P ("head", substantive, singular partitive)

pea+S+Pl+N ("head", substantive, plural nominative)

pidama+V+Ind+Pr+Sg2 ("must", verb, indicative mood, present tense, singular, 2nd person)

As we can see, words can be morphologically ambiguous in Estonian. By the way, in Estonian texts about 50 % of the word-forms are morphologically ambiguous.

lexc> lookup ajalehepoisina

aeg+leht+poiss+S+Sg+Ess

In Estonian compound words the first components usually remain unchanged, while the last component is subject to inflection. At the same time, pre-components can be either in the nominative or in the genitive case. There is no general rule for choosing the right case for the pre-components.

lexc> lookdown kallis+A+Spr+Pl+El

kalleimaist

kalleimatest

The example demonstrates the existence of parallel forms, i.e. word-forms having the same grammatical meaning but different forms.

Example 7. Random surface string generation, using the program *lexc*.

lexc> random-surf

Use (s)ource or (r)esult? [r]:

NOTE: Using RESULT.

käed	+	"hands"
pessa	+	"into the nest"
öeldavaid	+	"said by somebody" (Adj), plural partitive
kohaehe	?	"the decoration of a place" – strange compound
vanalt	+	"from the old" (Adj)

pimeduse	+	"of the darkness"
ülejätnuiks	?	strange compound, hard to translate
eemaltõppijad	?	"ones who learn from a distance" – slightly strange compound
nähtuta	+	"without the seen thing"
läksin	+	(I) "went"

To the output of the program approximate English translation as well as signs "+" or "?" are added. Every normal word-form is followed by "+". If the word-form is not used, it is marked by "?". The mistakes are caused by the overgeneration of compounds and derivatives.

6. Results

The experimental two-level morphology for Estonian (EETwoLM) has been composed:

- There are 45 two-level rules in the rule set that deal with stem flexion, phonotactics, orthography and morphophonological distribution.
- The net of lexicons consists of root lexicons for all word classes containing a total of ≈350 different word roots and of over 200 small lexicons describing the stem end alternations, conjugation of verbs and declination of nouns.
- The lexicons and rules express most of the phenomena occurring in Estonian morphology.
- The system is consistent in its present stage: we can get correct results to both morphological analysis and synthesis in the range of word stems occurring in the root lexicons.

7. Future Perspectives

The coverage of stem lexicons can be enlarged semi-automatically, using the electronic version of Viks (1992) and the type-detection module developed in the Institute of Estonian Language (see the webpage *www.eki.ee/tarkvara*). To adapt EETwoLM exactly to the morphological classification after Viks (1994), some changes have to be introduced into the network of lexicons.

A consistent and lexically satisfactory description of Estonian morphology in the two-level formalism can be the basis of automatic morphological analysis and synthesis. Simultaneously, two-level-morphology-based language engineering software in XRCE (spelling checker, information retriever a. o) would be applicable to Estonian language.

8. The Estimation of the Suitability of the Two-level Morphology Model to Estonian Language

During the composition of EETwoLM some features of the two-level model proved very useful. We have given the overview of them in Uibo (1999:55):

1. Using the lexical representation is an advantage because the lexical entries can include other information additionally to the pure sequence of letters:
   - 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.
   - 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 an 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 checking the number of syllables.
4. If a pair occurs in several contexts having nothing common neither in content nor in form the corresponding contexts can be listed on the right side of one and the same rule. It prevents from introducing new and meaningless lexical characters. E.g. the pair "S:0" is possible both in the weak grade of the words with s:0-alternation within the stem and at the end of a class of words ending with "s". In the first case the lexical phoneme "S" is situated in between vowels, in the second case it is found at the end of the stem.
5. The net of lexicons is convenient to handle
   - non-phonologically caused stem end alternations (org. "ne/se", "0/me");

- rules of morphotactics;
- productive derivation and compounding (partly).

However, we have also pointed to some difficulties in Uibo (1999:56) that have occurred in the course of the description of Estonian morphology in the two-level formalism:

1. The word class is subject to change during the derivation processes, but the morphological information is composed moving along the pointers between lexicons in one direction. Return to the previous steps, thus the deletion and replacement of the word class is not possible. Now the problem has been solved artificially: the verb derivatives are in a separate lexicon and for the productively derivable adjectives the determination of word class has been deferred.
2. It is inconvenient to introduce word lists into the lexicon system that do not coincide with the inflection types. The lists are needed e.g. for words with exceptional forms, for words having additive case and short plural, and especially for compound word production.

The hypothetical solution of the listed problems could be the combination of the two-level model with another model that would help to overcome the above-listed limitations.

9. Conclusion

The experiments on EETwoLM have shown that the two-level model is quite usable for Estonian simple word recognition and production. However, the net of lexicons is not very well suitable for modelling the derivation and compounding processes. The efficiency of the implementation of the rules and lexicons as finite-state transducers is definitely an advantage. Unfortunately, the objective evaluation of EETwoLM is not possible yet, as the coverage of the lexicons is insufficient for real text processing.

References

Kaalep, H.-J. 1996. ESTMORF, a Morphological Analyser for Estonian. *Estonian in the Changing World* / edited by H. Õim, 43-97. Tartu: University of Tartu, Dept of General Linguistics.

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

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

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

Koskenniemi, K. 1997. Representations and Finite-State Components in Natural Language. *Finite-state Language Processing* / edited by E. Roche and Y. Schabes, 99-116. Language, Speech, and Communication Series. Cambridge, Massachusetts, London, England: The MIT Press.

Kuusik, E. & Viks, Ü. 1998. The Rule-based Morphological Synthesis. *Arvutimaailm* (*The World of Computers*) 1/1998, 43-45, 63, 2/1998, 19-21 (in Estonian).

Uibo, H. 1999. The Estonian Word-Form Analysis and Generation, Using Two-Level Morphology Model. M.Sc. thesis. Tartu: University of Tartu, Institute of Computer Science (in Estonian).

Viks, Ü. 1992. *A Concise Morphological Dictionary of Estonian.* Tallinn: Institute of Estonian Language and Literature.

Viks, Ü. 1994. Classificatory Morphology. Ph.D. thesis. Tartu: University of Tartu. (in Estonian).