Tools for the Generation of Morphological Entries in Dictionaries

Ülle Viks

Institute of the Estonian Language Roosikrantsi 6, EE 10119 Tallinn, Estonia ylle@eki.ee

Abstract

The lexicographer's tool introduced in the report represents a semiautomatic system to generate the section of morphological information for Estonian words in dictionary entries. Estonian is a language with a complicated morphology featuring (1) rich inflection and (2) marked and diverse morpheme variation, applying both to stems and formatives.

The kernel of the system is a rule-based automatic morphology with separate program modules for every linguistic subsystem such as syllabification, recognition of part of speech and type of inflection, stem variation, morpheme and allomorph combinatorics. The modules function as rule interpreters applying formal grammars in an editable text format.

The system enables generation of the following: (1) part of speech, (2) type of inflection, (3) inflected forms, (4) morphonological marking: degree of quantity, morpheme boundaries (stem+formative, component boundaries in compounds), (5) morphological references for inflected forms considerably different from the headword.

The system permits of set-up, so that the inflected forms to be generated, the style of morphonological marking and the criteria for reference selection are all up to the user to choose. Full automation of the system application is restricted mainly by morphological homonymy.

1. Occasion

Estonian lexicographers do need a tool for entering morphological information in dictionary entries. As a rule, dictionary is not meant to provide an elaborate description of grammar. Depending on the type of the dictionary its author's prime concern is usually an explanation of the meaning of words, or providing words with translation equivalents, etc. etc., not grammar. Yet, grammatical information cannot be left totally aside, in particular if the language in question has a rich and complex morphology. It is rather inevitable in translation dictionaries, particularly if the source and target language differ typologically. Our tool is meant to help lexicographers shake off most of the grammar problems and concentrate on their main task, which is, of course, presentation of lexical information.

Inflectional morphology of Estonian is very complicated. First, there is a considerable number of inflected forms: nouns change in case (14-15 cases) and number (singular, plural) and have numerous parallel forms in the plural, while verbs change in mood, tense, voice, person and aspect, having nearly 60 different simple forms, in addition to the analytical forms of tense and negation.

Another essential peculiarity of the Estonian language is the extensive variation of its morphological units. Variation may characterise both morphological formatives, cf.

```
Part Sg: aasta/\underline{t} - p\ddot{a}eva/ - kuu/\underline{d}

'year' - 'day' - 'month'

Inf: ela/\underline{da} - haka/\underline{ta} - vii/\underline{a}

'to live' - 'to start' - 'to take'
```

and stems, cf.

<u>tuba</u> – <u>t`oa</u> – <u>t`uppa</u> – <u>tub/e</u> 'room' Nom Sg – Gen Sg – Adt Sg – Part Pl <u>k`and/ma</u> – <u>kanna/b</u> – <u>k`an/tud</u> 'to carry' Supin – Sg3 Pr Ind – Ips Past Participle

A stem may be subject to different kinds of variation. Although a stem may also remain unchanged (h'aige – h'aige 'ill'), in most cases it is patient of stem-end changes (p'eegel – p'eegli 'mirror', v'aene – v'aese 'poor') or of

stem-grade changes (*läige* – *l`äike* 'shine'), or of both (*liige* – *l`iikme* 'member'). Moreover, the location of stem variants in paradigms can follow different alternation patterns, cf. *läige* 'shine', *liige* 'member', *k`äik* 'walk' (strong grade variants are underlined):

```
Nom Sg Gen Sg Part Sg Gen Pl Part Pl läige <u>l'äike</u> läige/t läige/te <u>l'äike</u>/id liige <u>l'iikme</u> liige/t <u>l'iikme</u>/te <u>l'iikme</u>/id <u>k'äiku</u> k'äiku k'äiku/de k'äiku/sid & k'äik/e
```

2. Grammatical Information in Dictionaries

The quantum and choice of grammatical information as well as the manner of its representation depend on the size and type of the dictionary in hand, as certain principles do not equally fit different dictionaries (cf. a large many-volume vs. pocket-size dictionary, a monolingual vs. bi-/multilingual dictionary, a specialised vs. general dictionary, dictionaries for learners vs. tourists, dictionaries for active vs. passive language acquisition, etc.).

Grammatical information is usually placed after the headword. In some active dictionaries we can also find grammatical information after translation equivalents.

The user would probably have it the most convenient, if all of the necessary grammatical information were given right in the entry. (A good example is the Collins COBUILD, in which each headword is followed by a full list of inflections.) We cannot afford it, however, in the case of languages with a rich morphology, like Estonian, in which a word may appear in dozens of different forms. Here, evidently, at least part of the grammatical information has to be left out of the dictionary entry and moved to a congruous grammar survey.

As grammar and dictionary can be regarded as two different genres of language description, their object may coincide to a considerable extent. It is just more expedient to describe some phenomena in a grammar, whereas some others rather belong to a dictionary. A dictionary associates data with each individual word, pointing out its inflections, usage information, morphemic division, class, etc. A grammar employs a more general approach characterising the ways to generate inflected forms, rules of form usage, a general view and principles of classification, class

descriptions, etc. So, if such a dictionary and grammar are integrated by means of cross-references both ways, the result is an optimum economy in information presentation.

For over a century now Estonian lexicographers have been using the following (economical) way of material presentation: the word entry contains a couple of inflected forms and a type number, while the cocrresponding type descriptions can be found either in the grammar part of the same dictionary or sometimes in a separate volume. The same principle, only formalised, underlies the first morphological dictionary specially written for the Estonian language and called *Väike vormisõnastik* 'Concise Morphological Dictionary' (CMD) (Viks, 1992). This dictionary is based on a systematically elaborated model of classificatory morphology (Viks, 1977; 1994a) that has served as basis for several computerised systems of Estonian morphology (Hein, 1994; Kuusik, 1994), including an Estonian spelling checker (Kaalep, 1996).

In the CMD every word is provided with part of speech and every inflected word is provided with (1) some inflected forms (to display all variations of stems and formatives); (2) inflectional type number (referring to a type description available in the grammar part). The grammar volume contains a survey of the Estonian morphology and the type descriptions. For the sake of clarity as well as economy the type descriptions consist of nothing but basic forms as those are the forms to reveal all morphological differences distinguishing between the types. In addition there are rules of analogy enabling the synthesis of a whole paradigm out of the basic forms given.

Beside the main entries the CMD carries grammatical reference entries to create a link between different forms of one and the same word, e.g. t ' $oa \rightarrow tuba$, t ' $uppa \rightarrow tuba$ 'room', pr ' $ae/b \rightarrow pr$ 'aadi/ma 'to fry'. Reference entries are irrelevant in computer dictionaries enabling access by inflected forms as well, but they are very helpful for a human user of a traditional translation dictionary, particularly in the case of languages with extensive stem variation. If, for example, a beginner reading an Estonian text happens to encounter such forms as toas 'in the room' or tuppa 'into the room; inside', he/she would probably have a hard time finding the lemma tuba 'room', but for references (cf. the English forms $took \rightarrow take$).

Our morphological entry generator has been designed to generate all grammatical data necessary for the formation (synthesis) of Estonian inflected forms, as well as for understandig them (analysis). If a concrete dictionary happens not to need all those data, one may feel free to tailor his/her own selection in the phase of preliminary set-up.

3. Tools

3.1. Rule-based Morphology

Generation of morphological entries in dictionaries is one of the applications of the rule-based morphological system developed at the Institute of the Estonian Language (Viks, 1994b). The rule-based system means that all regular morphological processes are realized by means of program modules (rule interpreters) based on formalized language descriptions (formal grammars). Only irregular cases are included in the inner lexicons of the system.

At the present moment the Estonian language finds itself in a situation of unstability and change. The current social changes have brought along rapid innovative processes in the lexicon, presenting a lexicographic challenge. Close contacts with other languages reflect on the grammar as well, creating a necessity to introduce some changes in the fixed grammar rules.

Our rule-based system is open in two aspects. On the one hand, it is open to handle new words, making it possible to analyse and synthesise the forms of such words that fail in system lexicons but are morphologically regular (new words, as a rule, have a regular morphology). On the other hand, the system is open for changing the rules. Rule change can be effected without even changing the software, as input to the rule interpreting modules comes from ordinary editable text files.

As a rule, each program module first checks the list of exceptions. If the input word is detected there, the module gets its output data out of the same list. If not, rules are applied for a regular output.

3.2. Program Modules for Inflection

The main five program modules used in the generation of morphological entries are as follows:

- (1) inflectional type recognition
- (2) generation of stem variants
- (3) morphological synthesis
- (4) lemmatisation (morphological analysis)
- (5) boundary recognition in compound words

All these program modules are realised as dynamic linked libraries (DLL) that can be used in other applications as well (s. http://www.eki.ee/tarkvara/).

The morphology modules work in two modes: with and without mark of quantity. Although the Estonian writing systematically overlooks certain phonetic differences, some of them have morphonological status, and should thus be reflected in more respectable dictionaries. Many Estonian inflected forms differ only by degree of quantity, e.g.

kutse (II degree: Nom Sg) – *k`utse* (III degree: Gen Sg) 'invitation',

kooli (II degree: Gen Sg) – k`ooli (III degree: Part Sg) 'school'.

M1: inflectional type recognition

The recognition rules in Module 1 work on two phonological features: number of syllables and stem-final sounds. The input of the module is a lemma (Nom Sg or Supin, which are the normal entry forms of a headword in Estonian dictionaries), the output informs the user of its type of inflection (TOI) and part of speech (POS).

M2: generation of stem variants

Stem variants are generated by means of re-writing rules included in a number of subgrammars. The generation algorithm is navigated by a stemguide containing typewise information about the derivative relationships between different stem variants and the set of subgrammars to be applied in order to produce the necessary stem variants. The input of Module 2 consists of a lemma stem, type of inflection and part of speech. The module generates all possible stem variants of this lemma, together with the corresponding stem codes.

M3: morphological synthesis

Morphological synthesis is based on rules of morpheme and allomorph combinatorics:

- (a) morphotactic rules that determine depending on the part of speech – the choice of the paradigm (set of inflected forms), and
- (b) allotactic (combinatorial) rules that determine depending on the type of inflection the choice of the stem code and formative variant for any inflected form in a paradigm.

As input, Module 3 takes type of inflection and part of speech, plus stem variants with stem codes. The output consists of inflected forms together with their form codes.

M4: lemmatisation (morphological analysis)

Most of the rules and exceptions used by Module 4 of morphological analysis are the same as those employed by the above block of synthesis, but the program modules are run by a different master algorithm. Here the input is an inflected form, the output consisting of its lemma, type of inflection and part of speech, and the form code.

Lemmatisation needs to be applied to plural headwords in order to find their initial form (Nom Sg), as this is the form that serves as input for the rest of the modules. Even if a headword is in the plural, the singular forms should be provided in the dictionary entry, as they may be used, for example, in compound words, like in *püksid* 'trousers', but *püksi+rihm* 'trouser belt'.

M5: boundary recognition in compound words

The rules for finding component boundaries in compound words (Module 5) function in close co-operation with the above Module 4 of morphological analysis. Those rules proceed from the general phonotactic rules of the language, the possible combinations and frequency of occurrence of components, etc. For morphological purposes it is essential to find the final component, as this is the only part patient of inflection, e.g.

laste+tuba – laste+toa – laste+tuppa, etc. 'nursery' *tuba+teater – tuba+teatri – tuba+teatreid*, etc. 'chamber theatre'

toa+nurk - toa+nurga - toa+nurka, etc. 'corner of the room'

3.3. Entry Generation

As can be concluded from the above survey of modules our system enables the generation of the following data: (1) part of speech, (2) type of inflection, (3) inflected forms, (4) morphonological transcription marks in the inflected forms: degree of quantity, morpheme boundaries (stem+formative, boundaries in compound words). These data enable the formation of the necessary entries by means of two special modules:

(6) generation of reference entries

(7) entry design

M6: generation of reference entries

Morphological references provide a link from an inflected form to the headword, which is important, if the spelling of the inflection differs considerably from that of the headword, e.g.

toa→tuba, tuppa→tuba, prae/b→praadi/ma

Here it is necessary to state the threshold difference, as it is not expedient to give a reference entry to every stem variant. Simpler and regular cases can be described by rules placed in a special grammar survey so that the user can easily form his/her own headword.

M7: entry design

As for entry design, it is a process beginning with system set-up and continuing after the necessary data have been generated. The final stage of entry design consists of several subtasks such as locating different kinds of information, technical reduction of data, adding layout markup, etc. As here the solutions depend on the concrete dictionary, the choices are left up to the lexicographer to make.

3.4. System Set-up

As a rule, the generative abilities of the system tend to exceed the needs of a particular user. Therefore the output should better be tailored for a concrete dictionary, and this takes place during system set-up. The options are several:

- the list of inflected forms to be generated
- the style of morphonological mark-up
- the kinds of information to be added
- criteria for reference selection

What inflections are to be represented usually depends on the target group of dictionary users:

- All basic forms are necessary for the generation of a whole paradigm. This is important for an active learner's dictionary.
- Forms representing only the stem variants are sufficient for the reader to recognise and bring together different-looking forms of one and the same word.
- A few essential forms help either to recognise more important inflections or to differentiate between homonyms. This suffices for a user of a relatively advanced language proficiency.

The user can inform the system of what information is necessary, i.e. whether

- the entry should consist just of inflections or whether the type of inflection and/or part of speech should be added,
- inflection names should be supplied,
- morpheme boundaries (stem+formative, component boundaries in compound words) should be provided,
- degree of quantity should be marked.

Additional information concerning the use of concrete inflections (e.g. *pl.*, *usu. in interior local cases, indeclinable in pre-attributive position*, etc.) has to be supplied by the lexicographer, at the present state of the art.

The figure provides a sketchy outline of the process of morphological entry generation.

3.5. Problems

The main obstacle in the way of a full automation of the system application is morphological homonymy. Most of the headwords with homonymous spelling belong to different paradigms. E.g. the headword *ehe* can be inflected as: *ehe eheda* A 2 'pure' or *ehe `ehte* S 6 'adornment' or *ehe `ehtme* S 5 'waist'. The headword *meene* can be inflected as: *meene m`eene* S 6 'souvenir' or *m`eene m`eese* A 2 'with honey'.

In this case the system produces several sets of morphological data as the information necessary for disambiguation (selecting the correct paradigm) is not available from the mere headword. Neither syntactical nor statistical methods can help perform morphological disambiguition in headwords. We can only go through the full text of dic-

tionary entries trying to hit upon some useful knowledge. Yet even this strategy need not have the desired outcome, as the illustrations need not include the sufficiently transparent forms. So the problem has to be solved in dialogue mode.

4. Applications

Hitherto the system has been used to generate the necessary data on the Estonian morphology for three translation dictionaries (an earlier version of the system has been described in Kuusik & Lind & Viks (1995)). The following sample entries contain the same headwords handled in different dictionaries: tuba 'room', lastetuba 'nursery', andma 'to give'.

4.1. Estonian-Russian Dictionary

Romet & Liiv (1997-2000) is a bilingual dictionary for translators and learners (60,000 entries in 4 volumes).

Data:

- after the headword: all basic forms, POS, TOI; compound division, morpheme division; quantity degree marks
- in the references: morpheme division
- in the grammar part: paradigm surveys, type descriptions and rules of analogy

Example:

main entries:

```
tuba <tuba t'oa t'uppa, tuba[de tuba[sid &
          tub/e S 18> ...
      lastetuba <+tuba t'oa t'uppa, tuba[de
          tuba[sid & tub/e S 18> ...
      andma <`and[ma `and[a anna[b `an[tud,
           `and[is `and[ke V34*>...
reference entries:
      toa → tuba
       tuppa \rightarrow tuba
       tube → tuba
      anna[b \rightarrow andma
      an[tud \rightarrow andma
```

4.2. Norwegian-Estonian and Estonian-Norwegian Dictionary

Farbregd & Kangur & Viks (1998) is a general bilingual dictionary (20,000 Estonian entries).

- on the headword: compound division, morpheme divi-
- after the headword: inflected forms representing all different stem variants, TOI; morpheme division
- after the translation equivalent: inflected form representing the main stem variant, TOI
- in the references: morpheme division
- in the grammar part: paradigm surveys, type descriptions and rules of analogy

Example:

```
main entries (Estonian-Norwegian):
       tuba [toa, adt. tuppa 18e] ...
      laste·tuba [toa, adt. tuppa 18e] ...
      andma [annab, ips. antud 34] ...
main entries (Norwegian-Estonian):
      tuba [toa 18e]
      an|dma [-nab 34]
```

```
reference entries (Estonian-Norwegian):
```

```
toa → tuba
tuppa \rightarrow tuba
annab \rightarrow andma
antud \rightarrow andma
```

4.3. Estonian-Norwegian and Norwegian-Estonian **Pocket Dictionary**

Farbregd & Lepp & Viks (1999) is a bilingual dictionary for tourists (15,000 Estonian entries).

- on the headword: compound division, morpheme divi-
- after the headword: inflected forms representing the main stem variants (for uncompounded words only), morpheme division
- in the references: morpheme division
- in the grammar part: paradigm surveys

Example:

```
main entries:
```

```
tuba <toa, adt tuppa>
       laste-tuba`
       an|dma <-nab -tud>
reference entries:
       toa → tuba
       tuppa → tuba
       annab \rightarrow andma
       antud \rightarrow andma
```

Symbols and abbreviations used in examples:

```
the third degree of quantity
[/
        morpheme boundaries
        compound boundary
        abridgment mark
&
        or (between the parallel forms)
S
        substantive noun
A
        adjective
        verb
Nom
        nominative
Gen
        genitive
Part
        partitive
Adt
        aditive (= short illative)
        singular
Sg
Ρl
        plural
        present
Pr
```

indicative Ind impersonal Ips TOI type of inflection **POS** part of speech

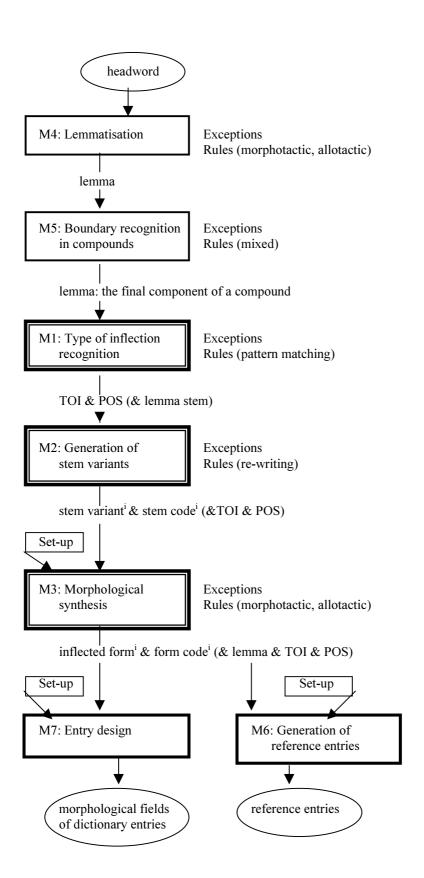


Figure: Morphological entry generation

5. References

- Farbregd, T. & Kangur, S. & Viks, Ü. (1998). Norsk-estisk estisk-norsk ordbok. Norra-eesti eesti-norra sõnaraamat. Tallinn.
- Farbregd, T. & Lepp, H. & Viks, Ü. (1999). *Estisk Lommeordbok*. Oslo: Kunnskapsforlaget.
- Hein, I. (1994). Practical realisation of the morphological analysis. In *Automatic Morphology of Estonian 1*. Tallinn
- Kaalep, H.-J. (1996). ESTMORF, a Morphological Analyzer for Estonian. In H. Õim (ed.), *Estonian in the Changing World*. Tartu.
- Kuusik, E. (1994). Morphological synthesis of Estonian based on the agglutination strategy. In *Automatic Morphology of Estonian 1*. Tallinn.
- Kuusik, E. & Lind, P. & Viks, Ü. (1995). *An Estonian Morpho-generator for Dictionaries*. Preprint, Tallinn.
- Romet, A. & Liiv, M. (eds.) (1997-2000). *Eesti-vene sõnaraamat 1-2*. Tallinn.
- Viks, Ü. (1977). Klassifikatoorse morfoloogia põhimõtted. Tallinn.
- Viks, Ü. (1992). A Concise Morphological Dictionary of Estonian. I: Introduction & Grammar. II: Dictionary & Appendices. Tallinn.
- Viks, Ü. (1994a). *Eesti keele klassifikatoorne morfoloogia*. Dissertationes philologiae Estonicae Universitatis Tartuensis 1 (DrPhil thesis). Tartu.
- Viks, Ü. (1994b). A morphological analyzer for the Estonian language: the possibilities and impossibilities of automatic analysis. In *Automatic Morphology of Estonian 1*. Tallinn.