The principles and structure of the Estonian Etymological Dictionary

Sven-Erik Soosaar
Institute of the Estonian Language, Tallinn

The Estonian Etymological Dictionary (EED) has been a project of the Institute of the Estonian Language (IEL) since 2003. Due to the urgent necessity for an etymological dictionary it was decided to start from a short and not too detailed version tailored for the general public with no philological background and to broaden this version later in order to compile a scientific dictionary. The next step involved concrete decisions about the material to be included into the first version of the dictionary.

1. Principles of selection of stems for the dictionary

Estonian is a Uralic language spoken mainly in Estonia. It has enjoyed conscious language planning at least since the beginning of 20th century. Numerous loanwords have entered since into literary language mainly from Russian, German, English and Finnish. Most of these are international words, which contain phonemes not found in Estonian inherited vocabulary (rendered by graphemes *f, š, z, ž*) or phonemes which phonotactical distribution is different compared to inherited vocabulary (*g, b, d* in the beginning of words, long vowels in successive syllables etc). These words are traditionally called in Estonian foreign words (*võõrsõnad*) in line with German practice, where *Fremdwörter* are traditionally considered a subsection of loanwords with special properties. In Estonia special dictionaries are published, where the origin and meaning of Estonian foreign words are explained (Võõrsõnade leksikon 1961, 2006 (7th ed.), Võõrsõnastik). But since the exact borderline between foreign words and loan words is disputed, the selection of words for such dictionaries is always problematic.

The basis for the selection of main entries for EED was the latest normative dictionary of Estonian (Eesti keele sõnaraamat, ÕS 1999). All the entries were checked and compounds and derivations were excluded. Our team of lexicographers and etymologists decided not to take into account recent loanwords, accepting only stems, which are found in a signpost of Estonian lexicography, the Estonian-German dictionary by Ferdinand Johann Wiedemann and its 2nd edition, printed in 1893. Not all stems found in Wiedemanns dictionary are included. Excluded are stems found in words beginning with letters *b, d, g, f* and words marked with asterisk (*), which are recent loans (*Fremdwörter*) and neologisms or as expained in preface: ‘Eigenthum neuerer Schriftsteller, Neubildungen, nicht der Volkssprache angehörend’. Examples of such words are *jāguar* (in contemporary spelling *jaaguar* ‘jaguar’), *kupel* (*kuppel* ‘dome’), *pank* (*pank* ‘bank’), *pendel* (*pendel* ‘pendulum’), *sekertārius* (*sekretär* ‘secretary’), *wilosowērima* (*filosofeerima* ‘to philosophize’).

2. Etymological resources

As a result of the selection we got a wordlist of 6308 stems to be etymologically analyzed. For this research were used following main sources:

1. Estonian etymological card catalogue of the IEL containing over 50 000 cards with articles on etymologically treated Estonian words and notes about possible etymologies;
2. Estnisches Etymologisches Wörterbuch, an unfinished manuscript of Julius Mägiste (1900-1978) of 4100 pages printed in Finland in 1982 and 2000;

3. Etymological dictionaries of Finnish (‘Suomen kielen etymologine sanakirja’ (6 volumes 1955-1978), ‘Suomen sanojen alkuperä’ (3 volumes 1992-2000), ‘Nykysuomen sanakirja’ (2004) and other languages.

We have also synchronized our word list with the manuscript of the new edition of ‘Võõrsõnade leksikon’ (VL2011), a dictionary of foreign words, which is due to be published in 2011. All stems not included in our dictionary should be included in VL2011 and vice versa.

3. Technical solutions and timetable

The preparatory activities of the dictionary began already in the 1970s, but the work of compiling etymological entries for the dictionary started in 2003. We used simple text files with provisional XML markup in hope that it would facilitate a future transfer of the material to a dictionary compiling software. The appropriate software was developed in the IEL and was ready for use in 2006, but the team of EED could not use it initially, since other dictionaries of the IEL were added to the system. The most important features of EELex are:

- web-based operating environment, which enables working virtually everywhere;
- Unicode-support;
- possibility of collective work;
- several search forms;
- two views: text with XML-markup (see Table 1) and with layout (see Figure 1).

The preparations for transfer of the material of EED into EELex dictionary management system were started in 2008. For this, usual steps for a new dictionary to be added to the EELex were taken: 1. creation of the description of structure; 2. defining parameters for views; 3. defining presentation models for adding new entries and groups; 4. composing menu lists (Viks, Loopmann 2009: 47). These tasks were completed and the transfer took place in January 2010. Since then all editing is done in EELex system.

All entries were covered by the end of 2008 and in 2009 we started editing entries. The manuscript should be edited by the end of 2010 and proof-read by June 2011.

4. Data fields

The entry contains following main sections: header, body and additional data.

5. The header

The header contains following fields: (1) headword, (2) two conjugational forms, (3) optional compound group for stems, which appear only in compounds, (4) meaning(s), 5. type of stem (inherited or loan and if loan, then source language).

6. The body

The body contains of (1) main dialectal variants, (2) same-stem words (only irregular derivatives), (3) loan sources and their equivalents in daughter languages (if the loan is from a protolanguage), (4) etymological cognates in other Finno-Ugric languages, and (5) etymological commentaries about problems in possible other etymological theories.

7. Additional data

Additional data contains bibliographical references and editorial notes. Bibliographical references are included in the database but excluded from the dictionary, since it would considerably inflate the volume and it is presumed that general public is not so much interested in the sources of the etymology as the etymology itself. Another field is for comments and notes for co-editors and co-authors.

As an example of an entry in EED we have chosen headword *kägar* 'a crumpled piece of material' (a possible Balto-Finnic-Lapp stem) in table and layout (Figure 1) and XML markup (Table 1).

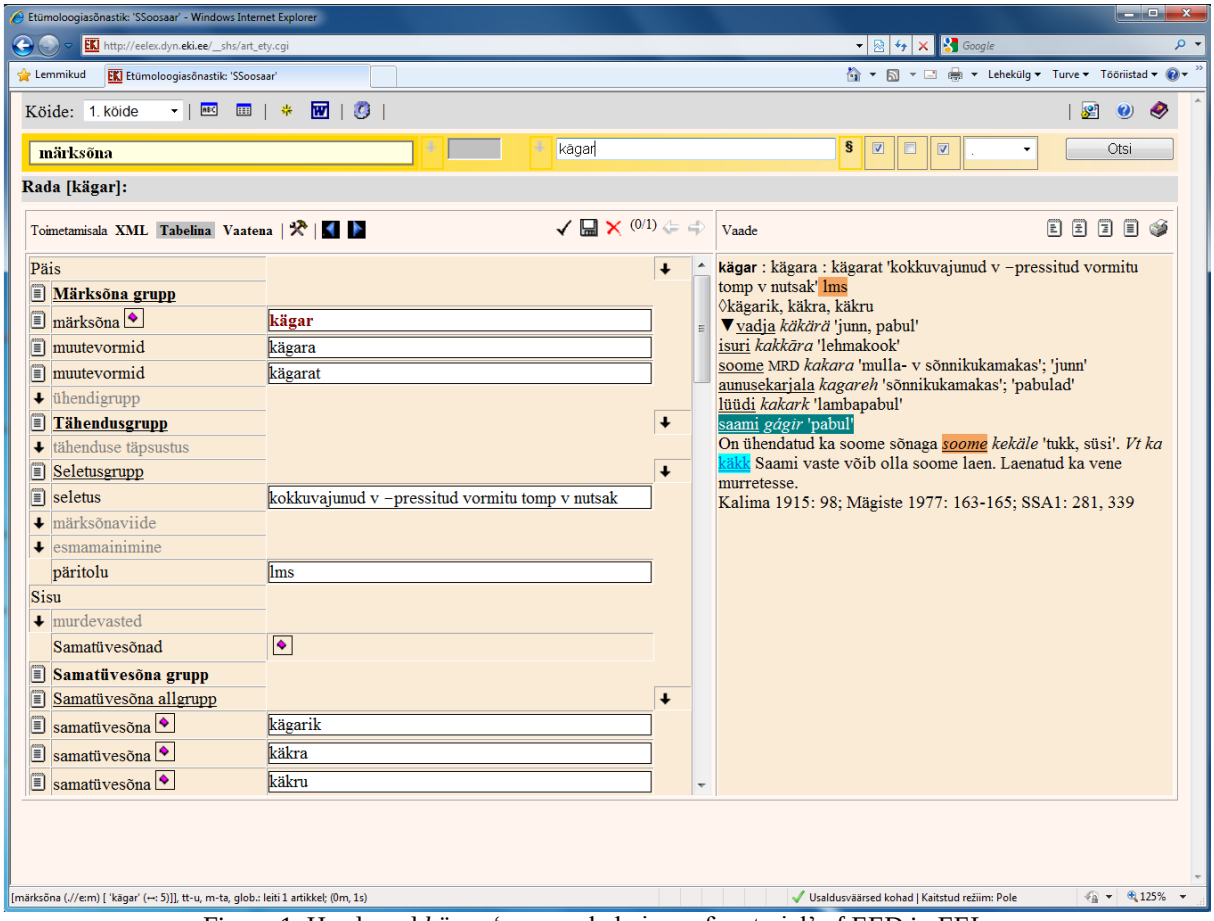

Figure 1. Headword *kägar* 'a crumpled piece of material' of EED in EELex

- <e:A>
  - <e:P>
    - <e:mg>
      - <e:m **e:O**="kägar">kägar
      - <e:mv>kägara
      - <e:mv>kägarat
    - <e:dg>
      - <e:glg>
        - <e:gl>kokkuvajunud v -pressitud vormitu tomp v nutsak
    - <e:po>lms
  - <e:S>

```
<e:tul>
    <e:dxg>
        <e:dxgg>
            <e:dx>kägarik
            <e:dx>käkra
            <e:dx>käkru
<e:va>
    <e:su>
        <e:kp e:k="vadja">
            <e:xg>
                <e:xgg>
                    <e:x>käkärä
                    <e:glg>
                        <e:gl>junn, pabul
        <e:kp e:k="isuri">
            <e:xg>
                <e:xgg>
                    <e:x>kakkāra
                    <e:glg>
                        <e:gl>lehmakook
        <e:kp e:k="soome">
            <e:xg>
                <e:xgg>
                    <e:s>mrd
                    <e:x>kakara
                    <e:glg>
                        <e:gl>mulla- v
                        sõnnikukamakas
                    <e:glg>
                        <e:gl>junn
        <e:kp e:k="aunusekarjala">
            <e:xg>
                <e:xgg>
                    <e:x>kagareh
                    <e:glg>
                        <e:gl>sõnnikukamakas
                    <e:glg>
                        <e:gl>pabulad
        <e:kp e:k="lüüdi">
            <e:xg>
                <e:xgg>
                    <e:x>kakark
                    <e:glg>
                        <e:gl>lambapabul
        <e:kp e:k="saami" e:ky="1">
            <e:xg>
                <e:xgg>
                    <e:x>gágir
                    <e:glg>
                        <e:gl>pabul
```

- <e:co>
  - On ühendatud ka soome sõnaga
    - <e:cx **e:k="**soome">kekäle
    - <e:gl>tukk, süsi

  .

    - <e:cvt **e:cvtl="**Vt ka">käkk

  Saami vaste võib olla soome laen. Laenatud ka vene murretesse.

- <e:L>
  - <e:bibl>
    - <e:aut>Kalima
    - <e:ia>1915
    - <e:lhk>98
  - <e:bibl>
    - <e:aut>Mägiste
    - <e:ia>1977
    - <e:lhk>163-165
  - <e:bibl>
    - <e:aut>SSA1
    - <e:lhk>281
    - <e:lhk>339
- <e:data>ety.pdf#kägar
- <e:G>c113d3fa-21c8-4b87-8ded-df67b0193881
- <e:K>SSoosaar
- <e:KA>2010-01-27T15:56:10
- <e:T>SSoosaar
- <e:TA>2010-04-19T13:52:41

Table 1. Headword *kägar* ‘a crumpled piece of material’ of EED in XML-markup

References

Erelt, T. (ed.). (1999). *Eesti keele sõnaraamat: ÕS 1999*. Tallinn: Eesti Keele Sihtasutus.
Mägi, R. (ed.). (2005). *Võõrsõnastik*. Tallinn: TEA Kirjastus.
Mägiste, J. (1982). *Estnisches Etymologisches Wörterbuch*. Helsinki.
Raun, A. (1982). *Etümoloogiline teatmik*, Rome-Toronto: Maarjamaa.
Vääri, E. (2006). *Võõrsõnade leksikon*. Tallinn: Valgus.
Wiedemann, F. J. (1893). *Estnisch-Deutsches Wörterbuch*. St. Petersburg.