The database of Estonian Word Families

Ülle Viks, Silvi Vare, Heete Sahkai
Institute of the Estonian Language
Outline

1. Background
   – What is a word family
   – The word families method
2. Data
3. Design: editing, query, Web interface
4. Applications
A word family (WF) is the set of all the words in the vocabulary of a language that contain a common stem morpheme:

- **aed** 'garden n.'
- **aednik** 'gardener'
- **aedmaasikas** 'garden strawberry'
- **aeda pidama** 'garden v.'
- **aiapidaja** 'gardener'
Word family

The WF is introduced by the simplex word that represents the common stem – the head of the family:

- AED
- aednik
- aedmaasikas
- aeda pidama
- aiapidaja
Word family

The words in the WF – the family members – are analyzed into immediate constituents and are assigned a word formation type:

- **aed=nik** ‘garden=noun suffix’
- **aed+maasikas** ‘garden+strawberry’
- **aeda pidama** ‘garden.partitive keep’
- **aia+pida=ja** ‘garden.genitive+keep+noun suffix’
Word family

The words in the WF – the family members – are organized hierarchically according to mutual word formational relations: each word is preceded by its base word and followed in turn by the derivations and compounds that are based on it:

AED ‘garden’
lasteaed ‘child.gen.pl+garden’ “kindergarten”
lasteaednik ‘kindergarten=noun suffix’ “kindergarten teacher”
Word family

- **ELA#MA**
  - ela=mu
  - kahe+pere+ela|mu
  - ela=nik
  - ela|nik=kond
  - el=u
  - el|u=s
  - el|u=tu
  - abi+el|u
  - abi|ell|u astu#ma
    - abi|ell|u|£#ma
    - abi|ell|u=mine
    - abi|ell|u|mis+ette|pane|k
The word families method

- Consists in organizing the entire vocabulary of a language into word families
- A method for structuring the vocabulary of a language
- A way of representing the word formation of a language
- The method used in the compilation of word formation dictionaries
- Consists in the word formation analysis of all the words of a language
- Presupposes a detailed description of word formation in the language
Principal word formation dictionaries


The WF method as the design principle of an electronic database

- A new type of linguistic resource
- Greatly improved access to word formation data and description
- A wide range of potential applications
Data

- The inventory of words is based on the latest large general dictionaries of Estonian
- The word formation analysis is based on the descriptive grammar of Estonian and subsequent research into Estonian word formation
- 8880 word families
- 192 000 items in total
Units of the macrostructure of the database: the word family

**aed subst.**

- **[P_TUL]**
  - aed=ik subst. *väike aed*
  - aed=nik subst.
    - maa|stiku+aed|nik subst. *(tegeleb maastiku kujundamisega)*

- **[P_LS1]**
  - botaanika+aed subst.
  - ema+aed subst. *aiand. (kust võetakse seemneid, pook- ja pistoksi)*
  - las#te+aed subst.
    - las#te+aed=nik subst. *lasteaiakasvataja*
    - las#te+aia+kasva|tal|ja subst.
    - las#te+aia+laps subst.

- **[P_LS2]**
  - aed+maasikas subst.
    - aed=maasika+kee|d|is subst.
  - aia+maja subst.

- **[P_YH2]**
  - aeda pida#ma
    - aia+pida=ja subst.
    - aia+pida=mine subst.
Units of the macrostructure of the database: the word family

• The word family is introduced by the head of the word family (a simplex word) and constituted of the family members.
• The family members are organized hierarchically by step of formation.
• The maximal number of steps found in the database is seven.
Units of the macrostructure of the database: the word family

• On the first level of the hierarchy, the head is followed by all the words based on it – the first-step formations. For clarity of presentation, the first-step formations are divided into separate blocks according to their word formation kind: derivatives (P_TUL), compounds by the second constituent (P_LS1), compounds by the first constituent (P_LS2), verbal expressions by the second constituent, and verbal expressions by the first constituent (P_YH2).

• Each first step formation is again followed by the eventual second-step formations, that is the words that are in turn based on it, and so forth.
Units of the macrostructure: family members

- **AED** ‘garden’
- **aed=nik** ‘garden=noun suffix’
  - maastiku+aed|nik
    ‘landscape.gen+garden|noun suffix’
- **aed+maasikas** ‘garden+strawberry’
- **aeda pidama** ‘garden.partitive keep’
  - aia+pida=ja
    ‘garden.genitive+keep+noun suffix’
The units of the microstructure

Each head of family and each family member has its own microstructure, separate fields for representing grammatical and lexical information about them:

• Homonym number
• Part-of-speech
• Definition
• Subject label
• Usage label
• Context
• …
Design

- Embedded in the dictionary management system EELex
- Based on a specially designed XML schema that follows the hierarchical structure of word families
- Provided with a Web interface
The dictionary management system EELex

- A web-based toolset for dictionary writing and management
- Stores universal reusable databases encoded in a standard XML format
- Provides tools for editing, query, layout design

www.eki.ee
EELex editing window

• The editing window is divided into the editing pane and the layout pane, which are mutually connected by click.
• In the editing pane, data can be edited both in table form and in the XML code.
EELex editing window: table view
EELex editing window: XML view
Editing

• For the hierarchical DEWF, important editing functions are the adding, deleting and moving of whole structural groups (blocks and family members).

• Another important editing function is bulk corrections because a large number of words occur in two or more word families and this function permits to modify all these occurrences at once.
Editing: block moving
Editing: bulk corrections
Query

• The EELex software permits to conduct structure based queries by every labelled group, element and attribute.

• The search results can be sorted in different ways: each column can be sorted in increasing, decreasing and reverse order (i.e. by the final letters of words).
Query: words with a usage label
Query: part-of-speech = adv.  
(reverse order by final letters)
Query: the block zone is empty
Web interface

• The resources completed in EELex are made available through the Web as free public resources.

• The Web interface supports structure-based querying.
Web interface: structure-based query
Web interface: search results display

Word families are often extremely large and the queried item may thus be difficult to find in the whole entry. Therefore we use a match-based display: in the initial search result, only family members containing the element that matches the search criteria are displayed, together with the family member[s] immediately preceding it in the hierarchy. The remaining part of the entry is hidden behind green plus-buttons.

In order to display the other family members on the same level of the hierarchy the user has to click on the green button.
Web interface: search results display
Web interface: search results display
Applications

Estonian is typologically an agglutinative-fusional language characterized by extensive stem variation and the abundance of formatives. The majority of Estonian vocabulary consists of derivations and compounds with possibly quite complex structure.

The needs generated by the Estonian word formation system:

• language description
• language education
• lexicography
• language technology
Applications: research

- Data for research into word formation and related areas
- The process of the compilation of the database has already given rise to studies into several problematic and less researched phenomena of Estonian word formation
Applications: language education

• A tool for learning Estonian word formation and the vocabulary of Estonian
• Permits to generate different types of learner’s dictionaries of word formation
• A tool for teachers for compiling custom teaching materials
Applications: lexicography

• Helps to compile the lists of headwords of dictionaries
• Provides the word formation segmentation of the complex headwords of dictionaries
• Provides the lists of selected derivatives and compounds to be included in the entries of dictionaries
Applications: language technology

- Word formation module of automatic morphology
- Information retrieval
- Speech synthesis
- Integrated lexicon and grammar system
Thank you!