The Database of Estonian Word Families: 
a Language Technology Resource

Ülle VIKS 1, Silvi VARE and Heete SAHKAI
Institute of the Estonian Language, Tallinn

Abstract. The paper describes a polyfunctional database of Estonian word families which is based on extensive research and contains detailed word formation information about the Estonian vocabulary. It is an XML database integrated into a dictionary management system which offers various possibilities of structure based editing and searching, data reuse etc. The design of the database is based on the word families method, which consists in the organization of words on the basis of common stem morphemes and word formation relations. Until now, the word families method has been used in the compilation of word formation dictionaries. Using the method in the compilation of a database is a novel solution which considerably broadens the access to and the possible uses of word formation data. The database provides material for researchers in computational and general linguistics, language learners and teachers, and lexicographers. The data can also be used in several language technology applications like search engines, text-to-speech synthesis etc. 2

Keywords. word formation, XML database, electronic lexicography, automatic morphology, dictionary management system

Introduction

The paper describes a word formation resource created in the Institute of the Estonian Language: the database of Estonian word families (henceforth DEWF), which is a novel application of the word families method [1]. The database is integrated into a dictionary management system and equipped with a web interface. After completion (in 2012), it will be made available through the Web as a free public resource. The database presents word formation information about Estonian words in an explicit form. The word formation analysis is based on thorough research and has been inserted manually following a fixed schema. The electronic database permits to access the data by a large number of criteria.

There is great need for this type of information. As an agglutinative-fusional language, Estonian is characterized by a rich word formation system in which different word formation kinds, types and means combine in complex ways, and stems are subject to different types of change. Given the complexity and difficult access of word formation data, many important and theoretically interesting phenomena of Estonian grammar have not been properly researched. Comprehensive word formation information is also needed in lexicography; at present, word formation information in

1 Corresponding Author: Senior Computational Linguist, Institute of the Estonian Language, Roosikrantsi 6, Tallinn, 10119 Estonia; E-mail: Ülle.Viks@eki.ee.
2 The study was supported by the National Programme for Estonian Language Technology and by the project SF0050023s09 “Modeling intermodular phenomena in Estonian”.
the dictionaries of Estonian is scarce and often controversial. Another area in which
word formation information is badly needed is language education, since the complex
structure of Estonian words creates comprehension and production difficulties for
language learners. Word formation information has also been insufficiently used in
language technology applications, which is a more widespread problem [2].

1. Word Families and Word Formation

1.1. Theoretical-methodological Background

A word family comprises all the words of a language related by a common stem
morpheme (or its variant). It is headed by the simplex word (the head of the family)
that represents the common stem. Inside the word family, words (family members) are
arranged semasiologically, according to word formation [1]. The words are organized
in an integrated hierarchical network on the basis of a stepwise immediate constituent
analysis [3]. This way of representation permits to visualize the internal structure of
complex words by simultaneously showing their base word and their immediate
constituents (1) ^3.

(1) SPORT ’sport’ > sport=lane ’sport=NOUN SUFFIX’ „sportsman, athlete“ >
sport|las=lik ’sportsman=ADJECTIVAL SUFFIX’ „sportsmanly“ >
sport|las|likk=us ’sportsmanly=NOUN SUFFIX’ „sportsmanship“

As a whole, the word families of a language reflect the structure of the vocabulary
of the language, as well as its entire word formation system, which can thus be grasped
by the user [4, 5, 6, 7]. The word family’s phenomenon embodies the fact that most
words of a language are related to a number of other words by semantic and formal
motivation [8].

The word families method is the method for structuring the vocabulary of a
language and for compiling word formation dictionaries [4, 1, 5, 6]. The compilation of
a word family dictionary consists in segmenting the lexemes into immediate
constituents and organizing them into word families, which requires a large amount of
research and analysis.

Word family dictionaries are a relatively rare type of specific scientific dictionary.
The largest word formation dictionaries exist for German [9, 10] and Russian [7];
smaller dictionaries exist for some other languages. Using the word families method as
the design principle of an electronic database is a novel application of the method and
gives rise to a new type of linguistic resource.

1.2. Word Formation in Estonian

Estonian is typologically an agglutinative-fusional language and its morphology is
characterized by extensive stem variation and the abundance of formatives (inflectional

---

3 Symbols used in examples: # demarcates stems and inflections; = demarcates stems and affixes if
derivation is the last step of formation; | demarcates stems and affixes if derivation is not the last step
of formation; + demarcates constituents of compounds if compounding is the last step of formation;
¤ demarcates constituents of compounds if compounding is not the last step; , demarcates interfixes.
and derivational affixes) [11, 12, 13]. The majority of Estonian vocabulary consists of derivations and compounds and can thus be organized into relatively large families.

The two main kinds of Estonian word formation are derivation (including conversion) and compounding. A repeated application of the same word formation kind, or an alternation of different kinds, may give rise to lexemes with quite complex structure, displaying concatenations of several affixes or stems, or combinations of stems and affixes. The hierarchical structure of word families permits to visualize these consecutive word formation steps in an explicit manner.

A frequent phenomenon that complicates the Estonian word formation are different types of stem changes, which may give rise to very different stem forms. Stem changes may lead to accidental formal similarities, which in turn give rise to interpretation difficulties. The word families representation clearly disambiguates these coinciding forms as they occur in different parts of the hierarchy and have different internal structures.

A phenomenon that has given rise to theoretical discussion (cf. [14, 15, 16]) and to misinterpretations in Estonian lexicography and language teaching are the so-called synthetic compounds. In dictionaries they are treated as compounds whereas theoretically they are regarded in Estonian as a subtype of derivation involving simultaneously the attachment of a suffix and the compounding of stems [17, 18, 19], e.g. keel#t kast#ma 'tongue#PARTITIVE water#INFINITIVE’ „to drink“ > keele+kast=e 'tongue.GENITIVE+water=NOUN SUFFIX’ „a drink“. In DEWF they are explicitly represented as derivations based on syntactic phrases.

2. The Database of Estonian Word Families

2.1. The Data

The Estonian word families are compiled on the basis of the latest general dictionaries of Estonian, which cover the majority of the vocabulary of contemporary written Estonian. The database contains at present 8880 lexical entries, i.e. word families, with a total of about 186 000 items, and 940 simplex words with no attested derivations or compounds. The analysis of the data is based on the descriptive grammar of Estonian [11] and on the subsequent research into Estonian word formation, e.g. [20–24, 19].

The basic unit of the macrostructure of the DEWF is 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, strictly following the motivational relations between words. On the first level, the head is followed by all the words based on it – the first-step formations, each of which is again followed by the eventual second-step formations, and so forth. For clarity of presentation, the first-step formations are divided into separate blocks according to their word formation kind: derivatives, compounds, verbal expressions.

To illustrate, Figure 1 presents the word family aed ‘garden, yard; fence’ in a strongly abbreviated form. The head of the word family is followed by the first-step formations (in separate blocks), e.g. aed=nik “gardener”, las#te+aed “kindergarten”, aeda pida=ma “to garden”. The second-step formations are e.g. maa|stiku+aed|nik “landscape architect”, las#te+aed=nik “kindergarten teacher”, aia+pida=ja “gardener”, and so forth. The maximal number of steps found in the database is seven.
On the level of the microstructure of the DEWF, the principal units of description are the head of the word family and all the family members. Inside each family member, special symbols (cf. footnote 3) are used to code its internal word formation structure. Separate fields represent grammatical and lexical information characteristic to family members and the head, e.g. part-of-speech (provided for all words), definition, subject or usage label, homonym number, context etc.

2.2. Technical Solutions

The word families database is integrated into the dictionary management system EELex (henceforth EELEX), a web-based toolset for dictionary writing and management (lexicographer’s workbench), which has been developed in the Institute of the Estonian Language [25]. At present, EELEX contains the databases of about 20 dictionaries of various types. The databases stored in EELEX are universal reusable language resources encoded in a standard XML format. This permits to exchange data both internally and externally to the system. Each dictionary has a specially designed XML schema. In the case of the DEWF, the schema follows the hierarchical structure of the word families. The schema serves as the basis for editing, searching and the layout design.

**Editing.** The EELEX software permits to use various structure based editing functions. The data is displayed simultaneously in two formats: the editing window is divided into the editing pane and the layout pane, which are mutually connected by click (Figure 2). Data can be edited in the editing pane both in the table form and in the XML code.

---

4 Symbols and conventions: [P_TUL]: derivatives; [P_LS1]: compounds by the right constituent; [P_LS2]: compounds by the left constituent; [P_YH2]: verbal expressions by the left constituent; PART-OF-SPEECH; SUBJECT LABEL; definition.
For the hierarchical DEWF, important editing functions are the adding, deleting and moving of whole structural groups (blocks and family members). As the same word may occur in several word families (e.g. a compound containing two or more stems), a useful function is bulk corrections, which permit to make the same change simultaneously in all the entries corresponding to the defined criteria.

**Query.** The EELex software permits to conduct structure based queries by every labelled group, element and attribute. Words can be searched e.g. by structural elements of a word: different affixes, inflections, constituents of compounds, coding symbols, etc., as well as by additional information: part-of-speech, definition, subject label, word formation type, etc. The search system permits to use regular expressions, logical operators, and symbol classes. 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). The entries returned as the result of the search can be exported to a MS Word file in layout format.

**Web Interface.** DEWF, like all the resources completed in EELex, will be made available through the Web as a free public resource (http://portaal.eki.ee/). The public versions of the dictionaries are primarily addressed to ordinary users, but the more specific needs of researchers, students, lexicographers and teachers are taken into account as well. More specific material can be searched using the structure-based query, which permits to define more precise search criteria. Another function currently being developed is the complex query, which permits to combine the values of several attributes in the same query.
A special solution was needed for the display of the search results, as the DEWF entries are often extremely large and the queried item thus difficult to find, even with the aid of colouring of text or structure menus which have been used for large dictionaries [26]. Our solution is the following: in the initial search result, the minimum of information is displayed: the head of the family and the family member matching the query criteria, together with the item immediately preceding it in the hierarchy. The remaining parts of the entry can by visualized by clicking on the „+“ icons (Figure 3).

3. The Applications of the Database

DEWF has a whole range of possible applications. First of all, DEWF considerably broadens the possibilities of the study of Estonian word formation and related areas like lexical semantics. The process of the compilation of the database has already given rise to studies into several problematic and less researched phenomena like the back-formation of verbs [22], conversion [23, 24], or reanalysis [19].

Secondly, language learners can use the DEWF as a tool for learning Estonian word formation and the vocabulary of Estonian. The export function of EELex permits to compile exercises and other teaching material, and to generate various types of learner’s dictionaries of word formation. DEWF can also be used in the creation of interactive electronic systems for language learning.

Thirdly, DEWF will be used in language technology to create the independent word formation module as part of the rule-based morphology of Estonian, which already covers the fully regular word formation [27]. DEWF provides the necessary
data for writing the word formation rules and in part permits to generate the rules automatically. In addition to morphology, the word formation module is necessary in text-to-speech synthesis [28] since the pronunciation of a word may depend on its word formation structure. Another possible area of application of the word formation module is information retrieval, e.g. in search engines: if the queried word is not found, another word from the same word family may provide useful information.

And finally, DEWF has already been used in lexicography, e.g. in the compilation of the lists of headwords of dictionaries. Thanks to the data reuse and data export and import functions of EELex, the data of DEWF will be used in the other dictionaries of the system: it will provide the word formation segmentation of the complex headwords of new dictionaries, and the lists of selected derivatives and compounds to be included in the entries. The first application of this type will be the learner’s dictionary of the basic vocabulary of Estonian (ab. 4000 entries), which will contain all the relevant lexicographic information: definitions, inflectional and derivational morphology, syntax, semantic classes etc.

DEWF is also a good starting point for the creation of electronic lexicon and grammar systems like Word Manager, DeKo [2] or canoonet German Dictionaries and Grammar (http://www.canoo.net).

4. Conclusion

DEWF is in many ways a novel resource. First of all, large lexicographic word formation resources exist so far only for a small number of languages, since they require a large amount of research and contain relatively specific information, which however is indispensable in the research, acquisition and language technology applications of languages with rich and complex morphology. Secondly, DEWF is a novel application of the word families method so far used in the compilation of word formation dictionaries: using the method as the design principle of an electronic database considerably broadens the access to word formation data. Thirdly, DEWF is a new type of application of a dictionary management system. As a polyfunctional resource, DEWF will have many applications in language technology (automatic morphology, text-to-speech synthesis, search engines etc.), lexicography (word formation data in general dictionaries), language education (learner’s dictionaries of word formation, teaching materials etc.) and research.

References
