Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary

Madis Jürviste, Jelena Kallas, Margit Langemets, Maria Tuulik, Ülle Viks

Institute of the Estonian Language
Tallinn, Estonia
E-mail: madis.jyrviste@eki.ee, jelena.kallas@eki.ee, margit.langemets@eki.ee, maria.tuulik@eki.ee, ylle.viks@eki.ee

Abstract

This paper introduces new functions of the EELex dictionary writing system, developed at the Institute of the Estonian Language. Recently the EELex system has gained many new functions and query possibilities: scheme editor; article preview generator; bulk corrections; testing cross-references; generating Estonian morphological information; displaying the complete data or only chosen types of data; adding image, video and audio files; exporting dictionary data to Word format. By the example of the corpus-based active Basic Estonian Dictionary we describe the implementation of new EELex features for compiling, editing and presenting dictionary data. The dictionary is being compiled for Estonian language learners at the beginner and lower-intermediate levels and will be published in 2013 in both paper and electronic versions. In addition, the Estonian module of the corpus query system Sketch Engine (Kilgarriff et al., 2004) for the extraction and presentation of government and collocation patterns will be illustrated.

Keywords: dictionary writing system; corpus query system; learner lexicography

1. Introduction

EELex (Langemets at al., 2010, henceforth EELex)[1] is a web-based dictionary writing system for compiling, editing and presenting dictionary data, allowing simple and advanced structure-based queries and the sorting of query results; in addition, EELex offers possibilities for group working and has Estonian language support. Altogether nearly 30 dictionaries of various types and different structures have been compiled using the EELex system: monolingual and bilingual dictionaries, terminological and grammatical databases, etc. All dictionaries compiled in the EELex system have a standard XML-markup, which makes EELex a multi-purpose language resource that can be used by lexicographers, terminographers and ordinary end-users.

With the EELex system we intend to create a more efficient bridge between the IT domain and lexicography. The problem of modern lexicography seems to be that only a very small number of the huge variety of possibilities offered by the recent developments in information technology are actually put into practice: the great majority of new e-dictionaries are in substance the same old traditional dictionaries that are simply displayed on an electronical support platform. A veritably modern dictionary should be able to implement in practice the various solutions offered by technology, from corpus-based compilation methods up to real comfort of use (of the final product). Taking into account the possibilites of hypertext functionality one should avoid strictly linear text restrictions, e.g. preferring semantic criteria in entry compilation to use of strict alphabetic order, but also and foremost offering the end-user a user interface that would surpass these restrictions (cf. Atkins, 2002: 11). As electronic dictionaries long ago passed the stage of physical data storage capacity limits, the only real 'capacity' limit to consider is the amount of data shown to the dictionary user on the screen: the user should not be overloaded with too much data (for example a very long and complex article containing various types of data entirely displayed at once in its full form). Rather, the user has to be given the possibility of easily finding and exploring these large amounts of data.

The Basic Estonian Dictionary is taking steps towards a solution of this kind. The conditions within which to achieve such a result have been created: the dictionary is being compiled using an advanced dictionary writing system (EELex), and we use comprehensive Estonian language corpora via the Sketch Engine corpus query system (Kilgarrif et al., 2004)[2]. To the end-user of the e-dictionary we plan to offer a solution thoroughly different from a traditional print dictionary format. In the following chapters we would like to introduce these two tools, EELex and Sketch Engine for Estonian, in more detail, by using the example of the Basic Estonian Dictionary.

2. Specialised Features of the EELex Dictionary Writing System

Even though the main functions of EELex were ready by 2010 (Langemets et al., 2010; Viks et al., 2010), the system is being continuously developed. There are several new functions in the areas of dictionary compilation and customisation, editing and publishing; it is also possible to make global queries from different EELex-based dictionaries.

2.1 Dictionary compilation and customisation

The possibilities for dictionary compilation have been

[1] http://eelex.eki.ee/ (20.09.2011).

[2] http://www.sketchengine.co.uk/ (20.09.2011).

improved. EELex also has new functions such as scheme editor, user management interface and article preview generator.

The scheme editor allows the user to customise the XML-structure of the dictionary. It is possible to add and delete elements or attributes, as well as to change the labels and properties of existing elements and attributes.

With the user management interface, the editor can assign different rights to different working group members.

With the article preview generator, the user can modify the dictionary entry preview. The user can set the font, size, colour and background of each element; set a character, text or line break between, in front of or after a specific element or group of elements; show or hide specific elements in the article editing preview and print preview; assign conditions for element display (according to the value of the attribute or neighbouring elements); and assign a hyperlink to an element.

2.2. Dictionary editing

Dictionary editing is made easier by new functions such as bulk corrections, automatic generation of morphological information, use of multimedia data and list data manager. Bulk corrections (see Figure 1) allow simultaneous corrections in all entries of a certain type (entries that do not correspond to the exact criteria can be excluded separately).

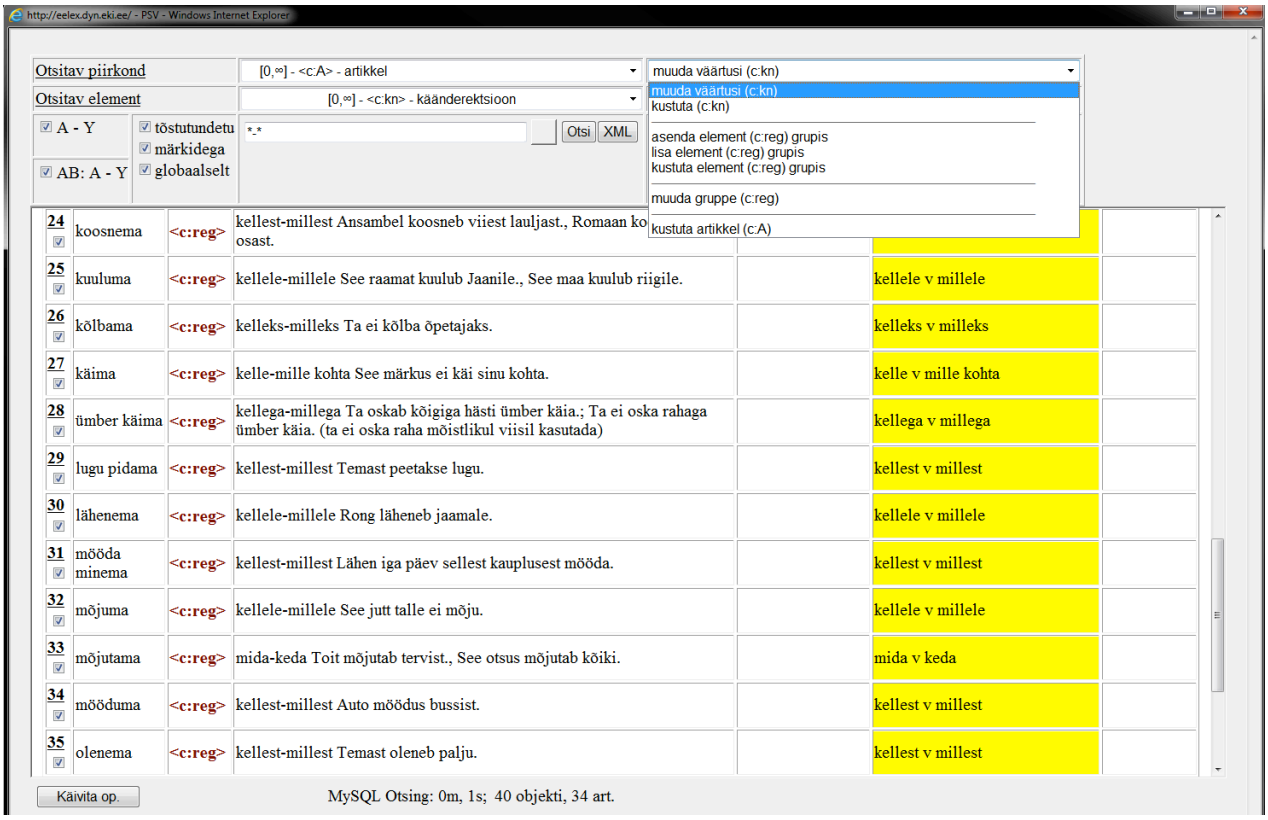

Figure 1: EELex editing: bulk corrections

Automatic generation of morphological information for the entries is possible in two modes: global and local. In global mode, the morphological data is added simultaneously to all entries; homonyms get several inputs, which requires obligatory postediting. In local mode the data is added through a dialogue and the editor can solve problems directly during the working process.

As for multimedia data, audio, video and image files can be inserted into an entry and opened directly.

The list data manager (see Figure 2) allows the user to add and delete list elements like labels, attribute values, semantic types, and so forth.

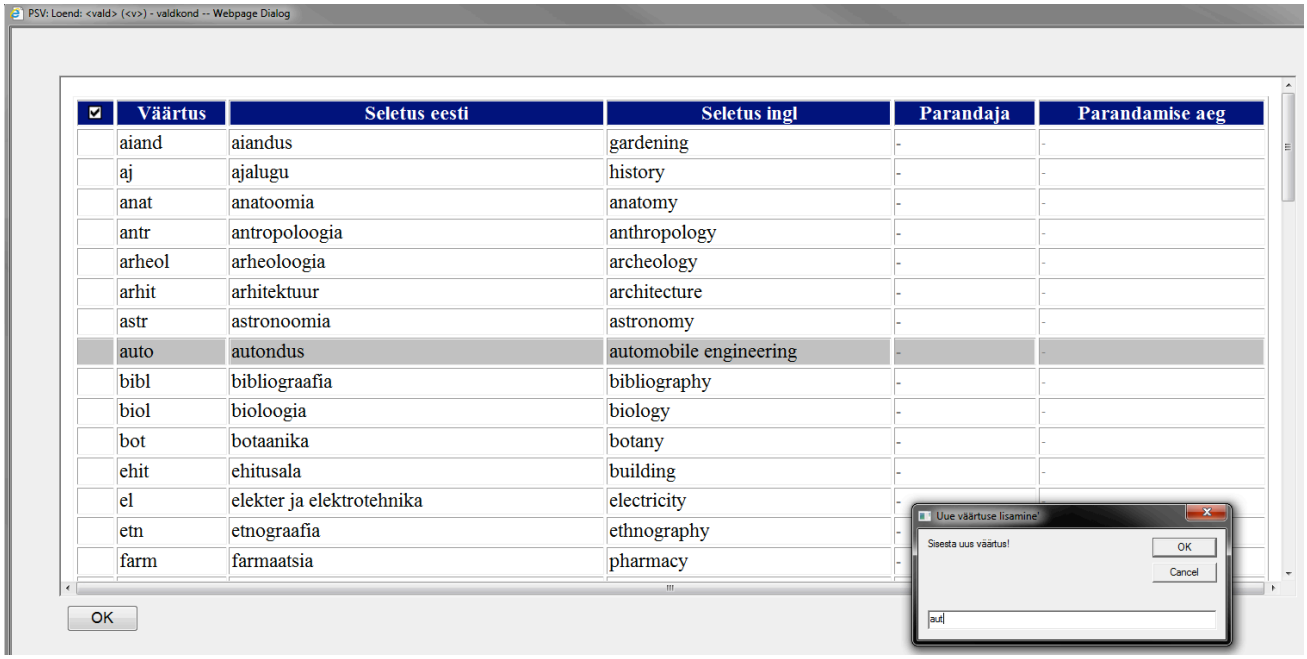

Figure 2: EELex editing: list data manager

2.3. Publishing

EELex allows dictionaries to be published in print as well as electronic versions.

To prepare a print dictionary, a file in MS Word format with a .doc extension is created, based on a specific query and for a defined set of entries. The entries in the .doc file have the same layout as in the editing window display preview. The query system allows the user to use the database of one particular dictionary to compile different versions of print dictionaries by selecting different structural elements to be shown in the printed version.

To allow electronic publishing of dictionaries, we host a web interface based on the XML format, allowing public access to published dictionaries[3]; lexicographers can also access dictionaries that are being compiled, but are not yet published, via the Institute's intranet. Users can search for information based on entry structure using the web interface; in addition to a simple query or a query from the whole entry, the user can restrict the search to a specific structure element, such as headwords, domain labels, examples, grammatical information, etc.

It is often difficult to easily find the relevant information in long entries. To display the entries in a concise manner, we have conceived a step-by-step display system: initially, only those parts of the entry that contain the queried information are displayed, although the user can also display other parts of the entry (e.g. the Dictionary of Estonian Word Families[4], Viks et al., 2011). We also plan to add the possibility to use structure filters, to allow multiple views, i.e. the display of only predefined structural elements of the entry, such as the headwords and definitions without examples.

2.4. Global queries

As the number of dictionaries in the EELex system is constantly increasing, a global query interface has been created for lexicographers, allowing a better overview of all dictionaries. The query allows the user to search for a headword in other EELex dictionaries or common keywords in different dictionaries.

3. The Basic Estonian Dictionary Database

The Basic Estonian Dictionary (henceforth BED) is an active dictionary, designed for learners of Estonian (Kallas, 2010; Kallas, Tuulik, 2011). It uses an XML entry scheme containing a detailed entry structure, which the editor-in-chief can complete and restructure if necessary, using the EELex scheme editor. Altogether the scheme has eleven blocks (pronunciation, inflectional formation, definition, word formation, government and collocation patterns, etc.) (see Figure 3).

[3] http://portaal.eki.ee/ (20.09.2011).

[4] http://www.eki.ee/dict/sp/ (20.09.2011).

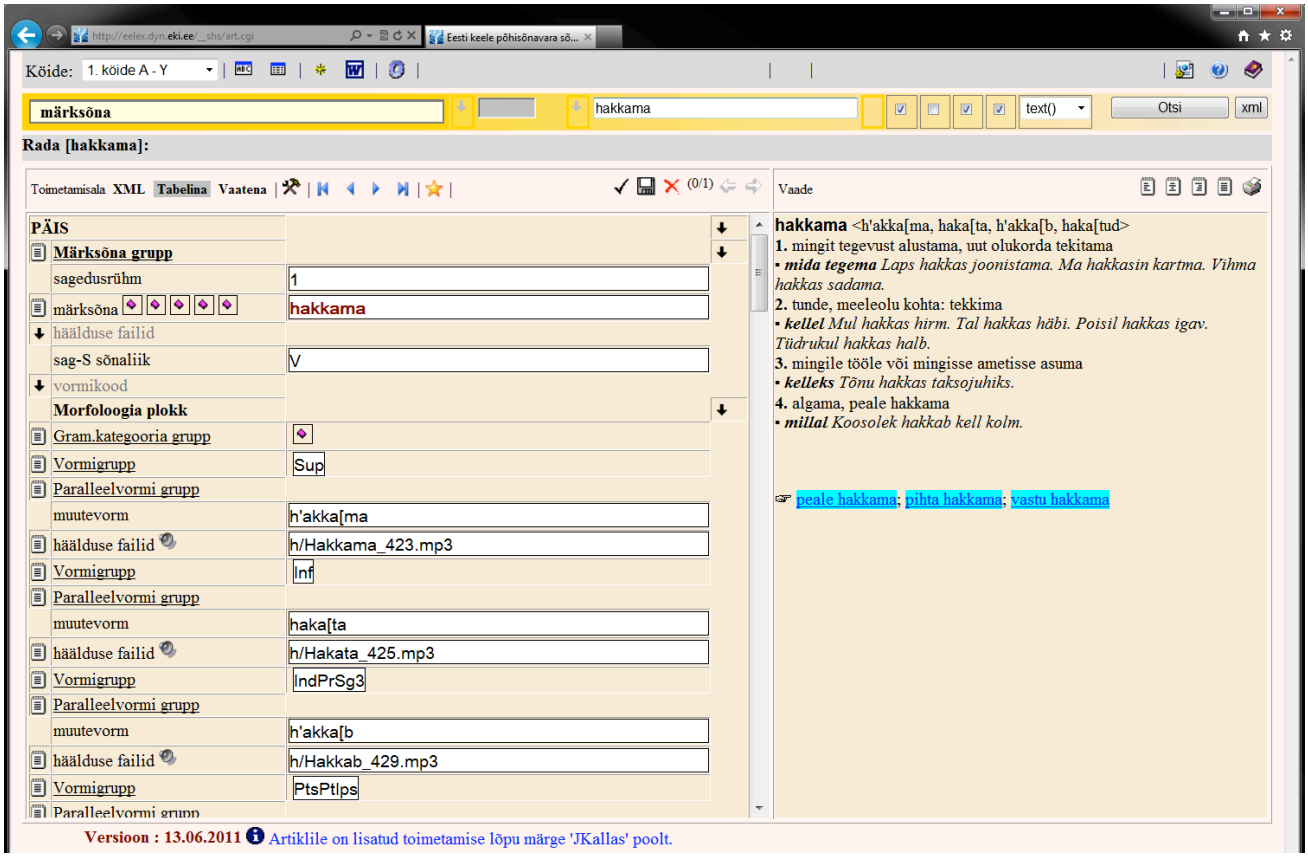

Figure 3: EELex editing window: the table view

For the purpose of coherence the metadata of all entries contains information about the semantic type(s) of the word sense(s): the dictionary is being compiled and edited according to semantic types (for the semantic classification of nouns, cf. Langemets, 2010).

In the BED compilation process we use all the new editing functions offered by EELex: bulk corrections, automatic generating of morphological data, use of multimedia data, the list data manager. In the following paragraphs we would like to describe the elements that clearly show the differences between the print and the electronic version.

In the pronunciation block we present a sound recording (mp3 audio file) for the most important forms in addition to the graphical presentation of the headword pronunciation. Morphological information is generated automatically. The BED as a learner's dictionary uses a comprehensive form-based presentation of data. In the printed version we give only the minimal inflectional paradigm (i.e. only the main forms: for nominal words the three first cases in singular and plural, for verbs the primary main forms, with secundary main forms if necessary). On the other hand, in the e-version the user will have the option to open the word's full inflectional paradigm: for nominal words all 14 cases in singular and plural, and for verbs all the finite and non-finite forms of the personal voice as well as of the impersonal voice. In the BED, several cross-references are used to show links between words: in the word formation block we use links to show word formation relationships (words that can be formed on the basis of the headword); in the lexico-semantic relationship block links are used to show synonyms, antonyms and paronyms. All the cross-references can be checked with a specific tool in EELex. However, this is not an automatic calculation yet, like in TLex[5] where the updating of sense and homonym numbers is fully automated.

In the sign language block the BED has a video recording of the relevant sign. For every sign the BED contains information about the initial hand form (the handshape with which the sign is articulated; should the handshape change during the formation of the sign, only the initial hand form is shown), the location where the sign is articulated (i.e. face, lips, cheek, chest, neutral space, etc.) and the movement with which the sign is formed. Based on these three parameters it is possible to search for a certain sign choosing the hand form, the location, and the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign.

It is possible to add images to the entries. Images are presented both in the paper version as well as in the electronic version of the BED.

[5] http://tshwanedje.com/tshwanelex/ (20.09.2011).

As the BED is an active dictionary, the explicit presentation of syntagmatic relations is of the utmost importance. Government and collocation patterns are presented in separate blocks and grouped according to codes presented, in the form of a drop-down menu (all menus are created with the list data manager). Figure 4 shows the codes for collocational patterns in the BED.

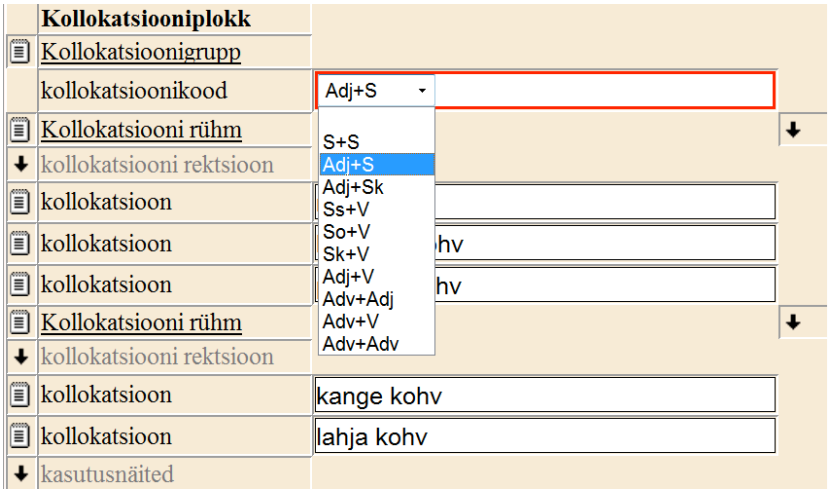

Figure 4: Codes for collocational patterns in the BED

The most frequent government and collocation patterns are analysed and selected using the Sketch Engine corpus query system.

4. Sketch Engine for Estonian: extraction and presentation of government patterns and collocations

Sketch Engine (Kilgarriff et al., 2004) is a web-based corpus query system for several languages (including, among others, French, Spanish, Japanese, etc.). Since autumn 2010 it has been used at the Institute of the Estonian Language to compile two Estonian language dictionaries: the Basic Estonian Dictionary and the explanatory one-volume Dictionary of Estonian (to be published in 2015).

Sketch Engine for Estonian uses the Estonian Reference Corpus[6] of 250 million tokens as input. The corpus had previously been annotated morphologically, lemmatised, partially disambiguated, and annotated by clause by Filosoft LLC[7]. Due to the agglutinative structure of the Estonian language the annotation involves not only lemmata and morphological features (POS-tag), but also inflections, e.g. *majas* /S/maja/s/sg_in 'in the house'.

In order to identify the grammatical relations between words, the syntagmatic (syntactic and collocational) properties of proper nouns, verbs, multi-word verbs (phrasal, prepositional and phrasal-prepositional), adjectives, ordinal numerals, and adverbs were investigated. As a result, 38 grammatical relations for Estonian were defined, using regular expressions and query language IMS Corpus Workbench. The system searches for grammatical relations which correspond to POS-tags and morphological inflections (e.g. such categories as *subject, object, oblique objects, adverbials, modifiers, adverb*); constructions with conjunctions *ja/või* 'and/or', *kui/nagu* 'as'; predicative (complements of the copula-like verb *olema* 'be'); various combinations of finite verbs with non-finite verbs; oblique objects and adverbials of particle verbs, prepositional verbs and noun prepositional phrases. All possible syntactic government patterns are brought forth: the system shows explicitly all possible case, adposition and infinitive government patterns for substantives and adjectives; object, case, adposition, and infinitive government for verbs; case government for particle verbs; adposition government for prepositional verbs and case government for adverbs. Figure 5 shows the word sketch for noun *usk*, 'faith'.

[6] http://www.cl.ut.ee/korpused/segakorpus/ (20.09.2011).
[7] http://www.filosoft.ee/index_en.html (20.09.2011)

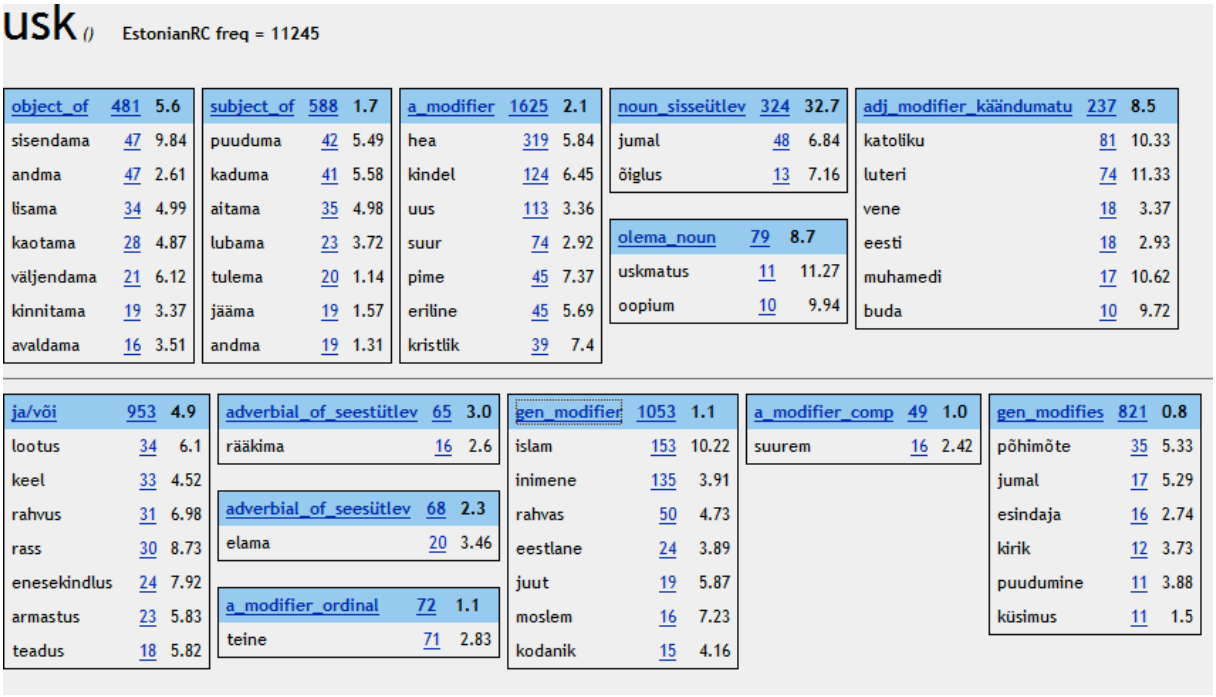

Figure 5: Word sketch for *usk*, 'faith'

The word sketch of the noun *usk*, 'faith', reveals the following patterns: noun (as subject) + verb (e.g. *usk puudub* 'to have no faith'), noun (as object) + verb (e.g. *usku sisendama* 'to inspire faith'), adjective + noun (e.g. *hea usk* 'good faith'; *katoliku usk* 'catholic faith'), coordinated nouns (e.g. *usk ja lootus* 'faith and hope'), and case government pattern (*usk* [kellesse-millesse] 'faith [in] someone or something'), etc. The most frequent extracted patterns are included in the entry of this particular noun (see Figure 6).

usk
1. kindel sisemine veendumus, milles ei kahelda *Aimult ravimid ei aita, ka usku peab olema.*
▪ ***kellesse-millesse*** *Mul on tema võimetesse usku. Ta leidis uuesti usu jumalasse.*
● **usku kaotama** *Ma olen kaotanud usu inimestesse.*
● **kindel**, **suur**; **pime usk**
2. religioon *Mis usku sa oled?*
● **katoliku, luteri usk**

Figure 6: BED entry for noun *usk*, 'faith'

The dictionary entry for *usk* presents information about the most frequent government patterns (*usk* [kellesse-millesse]) and collocations (*usku kaotama*; *kindel, suur, pime usk*; *katoliku, luteri usk*).

The quality of the word sketches depends on the quality of the morphological disambiguation. Errors in the output are mainly caused by errors or shortcomings in the morphological annotation. Secondly, the content of the corpus input should be balanced. At the moment (September 2011) 75 per cent of the texts in the corpus represent the media, i.e. journals and newspapers, fiction being significally under-represented in the corpus. We plan to upload the new version of the Estonian reference corpus into Sketch Engine in January 2012. This new version will contain more texts and additional information, including syntactic tags and data about text types.

When choosing example sentences for the BED we use two Sketch Engine functions: Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) and Tickbox Lexicography Template. In the Estonian module we use the language-independent *vanilla*-version in which the selection of examples depends on the length of the sentence (usually 5–10 words), initial capital letter and punctuation marks. At present, the data has to be manually selected in the Word Sketch and then manually copied into the dictionary. In the future we are planning to create a link between EELex and Sketch Engine to provide the possibility of direct transfer of corpus example sentences into dictionary entries.

5. Conclusion and perspectives

The dictionary writing system EELex has several new functions at different levels: set-up and customisation of a dictionary project (scheme editor, user management

interface, article preview generator), editing a dictionary (bulk corrections, automatic generating of morphological data, multimedia data input option, list data manager), electronic publishing (structure-based queries, step-by-step entry data display) and global query interface.

XML-based compilation will allow us to display content for the end-user in the web interface in layers, so that users themselves can choose what information will be displayed on the screen. For example, one might choose to display full entries, entries without the morphological information and/or pronunciation, and whether or not to display government patterns, usage examples, etc.

Moreover, using the EELex global query function, the user will have the option to search for a word in other dictionaries compiled using this system (for example explanatory dictionaries, bilingual dictionaries, etc.).

Being compiled in the EELex system, the BED project is a single database that contains all dictionary related data. This database allows the generation of different outputs: for example a (static) print dictionary or a (dynamic) e-dictionary, as well as specialised dictionaries based on partial database output.

6. Acknowledgements

This project has been supported by research grant SF0050023s09 ('Modelling Intermodular Phenomena in Estonian'), National Programme 'Estonian Language and Cultural Memory (2009–2013)' ('The Basic Estonian Dictionary', 2010–2013) as well as the National Programme for Estonian Language Technology (2011–2017) ('The Modifying of the Lexicographer's Workbench', 2011–2012).

7. References

Atkins, B.T.S. (2002). Bilingual Dictionaries – Past, Present and Future. In M. Corréard (ed.) *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins.* EURALEX, pp. 1–29.

Kallas, J. (2010). The development of scholary lexicography of the Estonian language as a Second Language in a historical and a theoretical perspective. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress.* Leeuwarden: Fryske Academy, pp. 648–651.

Kallas, J., Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. In *Eesti Rakenduslingvistika Ühingu aastaraamat 7*. Tallinn: Eesti Rakenduslingvistika Ühing, pp. 59–75.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress.* Lorient: Université de Bretagne Sud, pp. 105–116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress.* Barcelona: Universitat Pompeu Fabra, pp. 425–432.

Langemets, M. (2010). Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Tallinn: Eesti Keele Sihtasutus.

Langemets, M., Loopmann, A. & Viks, Ü. (2010). Dictionary management system for bilingual dictionaries. In S. Granger, M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications*. *Proceedings of eLex 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, pp. 425–430.

Viks, Ü., Vare, S. & Sahkai, H. (2010). The Database of Estonian Word Families: a Language Technology Resource. In I. Skadina, A. Vasiljevs (eds.) *Human Language Technologies. The Baltic Perspective. Proceedings of the Fourth International Conference, Baltic HLT 2010*. Amsterdam: IOS Press, pp. 169–176.