The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian

Jelena Kallas, Maria Tuulik, Margit Langemets
Institute of the Estonian Language
jelena.kallas@eki.ee, maria.tuulik@eki.ee, margit.langemets@eki.ee

Abstract

This paper is a report on a lexicographical project completed by the Institute of the Estonian Language. The Basic Estonian Dictionary was published in print in March 2014, and the online version will be available by September 2014. The BED is aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels (A2 to B1) according to the Common European Framework of Reference for Languages.

The dictionary contains about 5,000 headwords, including single items and multi-word lexical items. The BED provides lexicographical information on pronunciation, morphological information, definitions, word formation, government and collocation patterns, multi-word phrases, semantically related words and usage notes. In the online version, sound recordings (mp3 audio files) are provided. Morphological information was generated automatically. The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

The BED contains approx. 400 illustrations, study pages and picture pages (e.g. related to animals). In the appendix, geographical names and grammar tables are included.

In addition, the Dictionary of the Estonian Sign Language (containing approx. 6,700 video recordings), based on the BED database, was published online in March 2014.

Keywords: L2 monolingual lexicography; active dictionary; Estonian

1 Purpose and structure of the dictionary

The Basic Estonian Dictionary (henceforth BED) is a monolingual active dictionary aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels. The dictionary contains about 5,000 headwords, which were chosen on the basis of their frequency in the Estonian Reference Corpus[1], with 250 million tokens as input. In addition, headwords that are necessary in everyday life, but might not be as frequent in corpora, were added, for example *pott* 'pot',

[1] http://www.cl.ut.ee/korpused/segakorpus/ [01/04/2014]

pann 'pan', *jahu* 'flour', and *köhima* 'cough'. To get systemic content, some semantic classes (e.g. animals, plants and professions) were specially analysed.

Headword list includes not only single items, but also multi-word lexical items. Multi-word lexical items presented independently are multi-word verbs – particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down') and expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand') – and multi-word interjections (e.g. *tere õhtust* 'good afternoon'). The headword list of the BED was considered to be the controlled vocabulary list of the whole dictionary, other words were used in the entries. This was intentional so that users can look up unknown words in the dictionary.

Morphological information was generated automatically by using a morphological synthesizer for Estonian[2]. The BED as a learner's dictionary uses a comprehensive form-based presentation of data. For declinable words, grammatical cases in singular and plural, as well as the short form of the Illative, are presented explicitly. For verbs, the *-ma* and *-da* infinitives, and *he/she* forms, the past participle forms are given. However, after automatic generation there was a need for manual control of generated forms. Mostly this was necessary for the identification of homonymy. But forms were also deleted and added according to their frequency in corpora.

Information on pronunciation (palatalization, stress and syllabic quantity (in Estonian, a tripartite correlation of three syllabic quantities of stressed syllables exists)) is presented on the level of basic morphological forms of headwords. This is done by means of special palatalization, quantity and stress marks. Stress is shown only in cases where it is not on the first syllable, the normal stress pattern in Estonian.

Information on word formation is built into the micro-structure. Compounds with the headword as a second element (base word) are presented as references/links to their own entries, without additional information in the entry of the base word. Only transparent compounds, where the meaning of the base word has been preserved, were selected. All referenced compounds are presented as independent headwords as well.

Semantically related words (synonyms, antonyms and paronyms) of headwords are shown using the simplest possible metalanguage, e.g. *sama mis* 'same as' for synonyms and *vastand* 'opposite' for antonyms.

At the end of some entries, there are usage notes. Usage notes show differences between words and help to build vocabulary, e.g. polite phrases related to particular headwords are given and usages prone to error are pointed out.

The XML database of the Basic Estonian Dictionary is organized into several fields: lemma, pronunciation, inflectional information, definition, word formation, government, collocation, multi-word patterns, semantically related words and usage notes.

[2] http://www.eki.ee/keeletehnoloogia/projektid/morfana/ [01/04/2014]

1.1 Government and collocation patterns in the BED

As the BED is an active dictionary, the explicit presentation of syntagmatic relations (government and collocational patterns, also multi-word phrases) are of the utmost importance.

The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

Estonian Sketch Grammar (Kallas 2013) is geared towards the specification of the Estonian Reference Corpus and it contains 85 rules (14 UNARY, four SYMMETRIC, 62 DUAL and five TRINARY grammatical relations). As a result, the system searches for 32 types of lexicogrammatical constructions.

For nouns, the system searches for modifying adjectives, participles, oblique-case substantives, adverbs, pronouns, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

For adjectives, the system searches for modifying adjectives, adverbs, oblique-case substantives, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

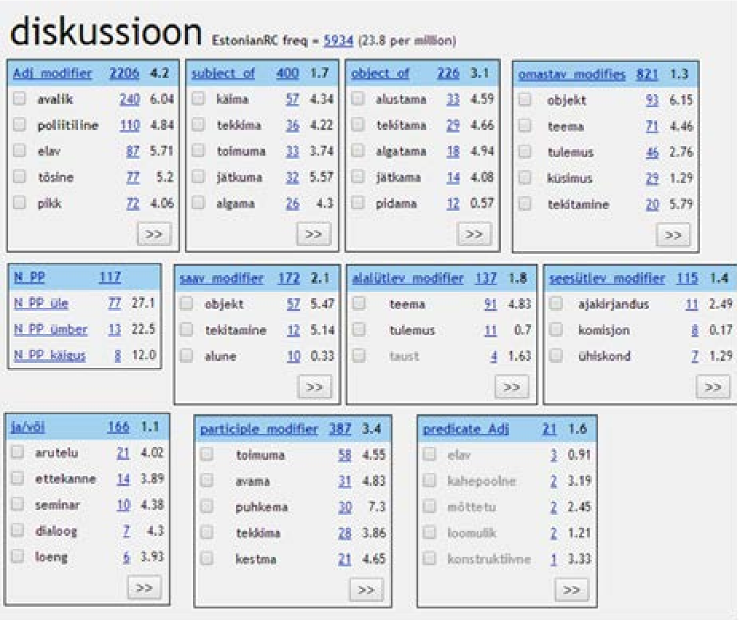
For adverbs, the system searches for modifying adverbs, oblique-case substantives, prepositional phrases and (by identifying conjunctive words) subordinate clauses.

For verbs, the system searches for substantives that function as subjects, objects and adverbials, and also for modifying adjectives, adverbs, prepositional phrases, non-finite verbs, gerundives and (by identifying conjunctive words) subordinate clauses.

Multi-word verbs, i.e. particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down‘), expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand‘), catenative verbs (verb + non-finite verb, e.g. *käima panema* 'start‘, lit. 'make [the engine] work‘), and support verb constructions (e.g. *läbirääkimisi pidama* 'negotiate‘) are considered separately. Since adverbial particles are tagged in the corpus as regular adverbs, a list of adverbial particles was compiled. The system identifies the most frequent adverbial particles used with particular verbs. This feature has great value when lexicographers need to choose what kind of particle verbs should be presented in the dictionary. Secondly, it is possible to see components of expression verbs if the component concerned has the part-of-speech tag X. Other components of multi-word verbs are identified as objects, adverbials or modifying non-finite verbs.

In addition, constructions with the conjunctions *ja/või* ‘and/or’, and *kui/nagu* ‘as’ can be found for all content words. For nouns, the system also searches for predicatives (complements of the copula-like verb *olema* ‘be’).

Figure 1 shows the word sketch for the noun diskussioon ‘discussion’.



diskussioon EstonianRC freq = 5934 (23.8 per million)

	Adj_modifier	2206	4.2
□	avalik	240	6.04
□	poliitiline	110	4.84
□	elav	87	5.71
□	tõsine	77	5.2
□	pikk	72	4.06

	subject_of	400	1.7
□	käima	57	4.34
□	tekkima	36	4.22
□	toimuma	33	3.74
□	jätkuma	32	5.57
□	algama	26	4.3

	object_of	226	3.1
□	alustama	33	4.59
□	tekitama	29	4.66
□	algatama	18	4.94
□	jätkama	14	4.08
□	pidama	12	0.57

	omastav_modifies	821	1.3
□	objekt	93	6.15
□	teema	71	4.46
□	tulemus	46	2.76
□	küsimus	29	1.29
□	tekitamine	20	5.79

N_PP	117	
N_PP_üle	77	27.1
N_PP_ümber	13	22.5
N_PP_käigus	8	12.0

	saav_modifier	172	2.1
□	objekt	57	5.47
□	tekitamine	12	5.14
□	alune	10	0.33

	alaütlev_modifier	137	1.8
□	teema	91	4.83
□	tulemus	11	0.7
□	taust	4	1.63

	seesütlev_modifier	115	1.4
□	ajakirjandus	11	2.49
□	komisjon	8	0.17
□	ühiskond	7	1.29

	ja/või	166	1.1
□	arutelu	21	4.02
□	ettekanne	14	3.89
□	seminar	10	4.38
□	dialoog	7	4.3
□	loeng	6	3.93

	participle_modifier	387	3.4
□	toimuma	58	4.55
□	avama	31	4.83
□	puhkema	30	7.3
□	tekkima	28	3.86
□	kestma	21	4.65

	predicate_Adj	21	1.6
□	elav	3	0.91
□	kahepoolne	2	3.19
□	mõttetu	2	2.45
□	loomulik	2	1.21
□	konstruktiivne	1	3.33

Figure 1: Word sketch of the noun *diskussioon* ‘discussion’.

The word sketch offers the lexicographer the most frequent collocates that occur as adjectival modifiers (e.g. *avalik* 'public‘, *poliitiline* 'political‘, *elav* 'lively‘, *tõsine* 'serious‘, *pikk* 'long‘ and *avatud* 'open‘), various oblique-case substantive modifiers (e.g. *diskussiooni objekt/teema/tulemus* 'object/topic/result of discussion‘) and in the 'and/or‘ relation to the node word (e.g. *diskussioon ja arutelu* 'discussion and debate‘).

Also identified are relations where the node word functions as subject (e.g. *diskussioon käib/tekib/jätkub* 'discussion takes place/starts/continues‘) and object (e.g. *diskussiooni alustama/algatama/jätkama/avama* 'start/initiate/continue/open a discussion‘).

The most frequent extracted patterns are mostly included in the entry of particular words and registered in the dictionary database.

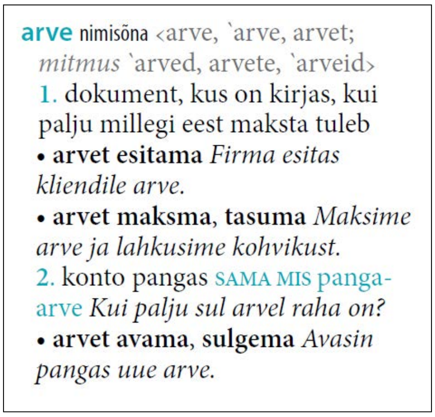
In the BED database, the government pattern field contains data about the government pattern, together with attributes for the type of government (object, case, adposition, infinitive and conjunctive word government), as well as the position of the complements and complementation variability.

The collocation pattern field contains data about the collocation pattern, together with attributes for the type of collocation. Collocation patterns are described by means of categorical and functional-relational labels. There are 13 types of collocation in the BED database:

- N(S)+V   noun (as grammatical subject) + verb: *päike paistab/tõuseb/loojub* ‘sun shines/rises/sets’;
- N(O)+V   noun (as grammatical object) + verb : *arvutit sisse lülitama* ‘switch on the computer’;
- N(A)+V   noun (as adverbial modifier) + verb: *kinnisvarasse investeerima* ‘invest in property’;
- Adj+V   adjective + verb: *määravaks osutuma* ‘prove decisive’;

- Adv+V adverb + verb: *kiiresti jooksma* 'run fast';
- N+N noun + noun: *ekspertide hinnang/arvamus* 'assessment/opinion of experts';
- Adj+N adjective + noun: *hea/halb eeskuju* 'good/bad example';
- Num+N numeral + noun: *sada meetrit/kilo* 'hundred meters/kilograms';
- Adv+N adverb + noun: *kergesti süttiv* 'easily flammable';
- Adv+Adv adverb + adverb: *väga aeglaselt* 'very slowly';
- Prep+N preposition + noun: *enne/pärast jõule* 'before/after Christmas';
- N+Post noun + postposition: *interneti/raadio kaudu* 'on television/ radio'.

Collocations of the same type are divided into semantic sets and presented explicitly as separate bundles. Figure 2 shows the entry for *arve* 'invoice, account' in the printed version of the BED.



arve nimisõna ‹arve, \`arve, arvet; *mitmus* \`arved, arvete, \`arveid›
1. dokument, kus on kirjas, kui palju millegi eest maksta tuleb
• **arvet esitama** *Firma esitas kliendile arve.*
• **arvet maksma, tasuma** *Maksime arve ja lahkusime kohvikust.*
2. konto pangas SAMA MIS panga-arve *Kui palju sul arvel raha on?*
• **arvet avama, sulgema** *Avasin pangas uue arve.*

Figure 2: The BED entry for the noun *arve* 'invoice, account' in the printed version.

1.2 Extra materials

The BED also contains approx. 400 illustrations. These are single illustrations with legends, structural illustrations (particular objects are highlighted by means of arrows), functional illustrations (mostly for adpositions), scenic illustrations (mostly for phrasal verbs) and nomenclatory illustrations (see figure 3).

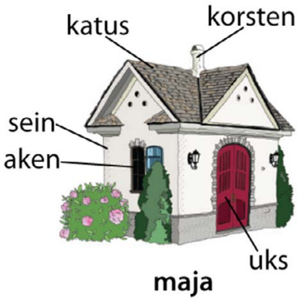
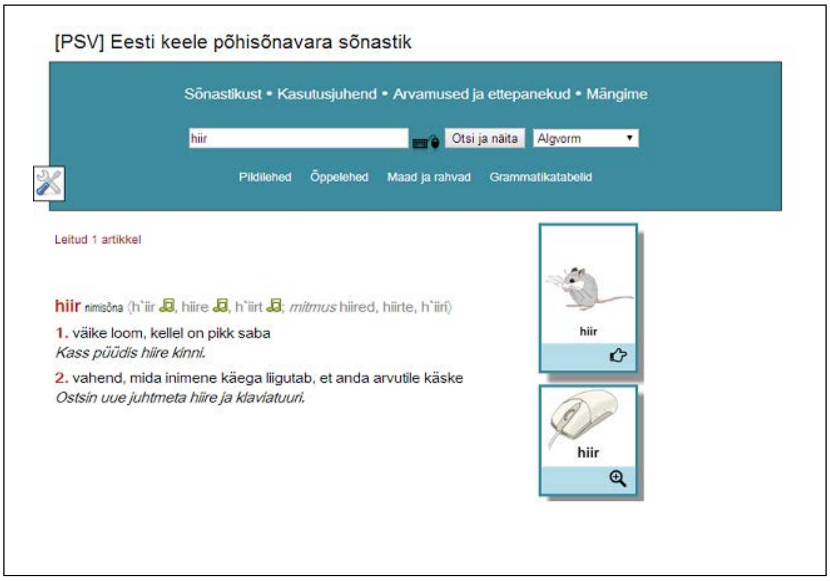


Figure 3: Nomenclatory illustration for the entry *maja* 'house'.

Besides illustrations, the dictionary has a centre section of 16 study pages (including instructions for producing numbers, time and dates, writing letters and emails, punctuation marks, common abbreviations, useful phrases) and 17 picture pages (e.g. on insects, animals, flowers and transportation). In the appendix, a list of countries, people and languages is given, as well as grammatical tables. Grammatical tables show how to decline and conjugate words, also they give guidance for producing all other word forms when moving on from the basic forms given in the dictionary.

2 The BED as an online-dictionary

The online version of the BED has some innovative features, which are implemented in the Estonian lexicography for the first time. Figure 4 illustrates the interface of the online version of BED.



[PSV] Eesti keele põhisõnavara sõnastik

Leitud 1 artikkel

hiir nimisõna (h`iir, hiire, h`iirt; *mitmus* hiired, hiirte, h`iiri)

1. väike loom, kellel on pikk saba
Kass püüdis hiire kinni.

2. vahend, mida inimene käega liigutab, et anda arvutile käske
Ostsin uue juhtmeta hiire ja klaviatuuri.

Figure 4: Online BED entry for the noun *hiir* 'mouse'.

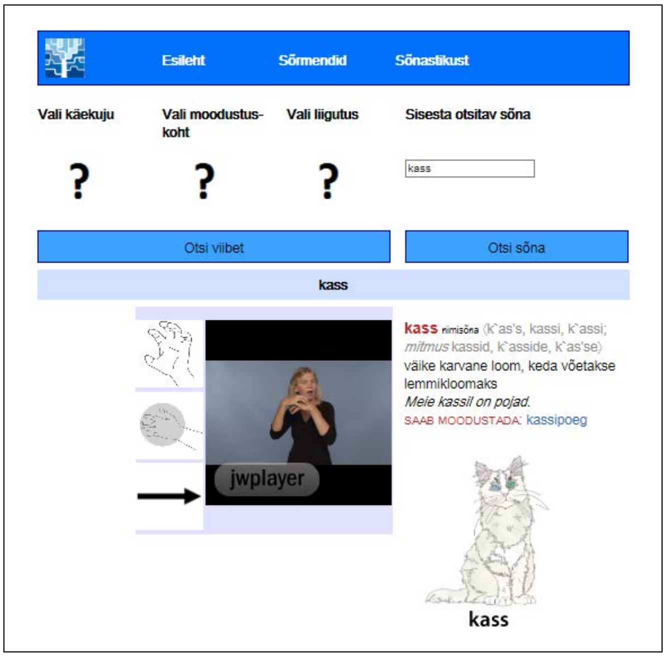
Pictures are aligned with particular word meanings. If the picture is topically related to one of the special picture pages provided in the dictionary as extra material (e.g. *hiir* 1. 'mouse' as related to *animals*), then these pictures are linked together. Otherwise it is possible to enlarge the picture.

Green musical note symbols indicate that there are sound recordings (mp3 audio files) linked to particular morphological forms. The audio files were pre-recorded.

The contents of the entire dictionary have been morphologically analysed. As a result, users can click on any word in a definition or example to find the entry for that word. And, vice versa, it is possible to type into the search box a word in any form (previous dictionaries allowed for searching only on the basis of lemma), and the lemma entry will be provided.

3 The BED as a basis for the Dictionary of the Estonian Sign Language

The BED database was used to compile the online Estonian Sign Language – Estonian Language dictionary[3]. Figure 5 illustrates the interface of the dictionary. There are approx. 6,700 video files. For every sign, the BED database contains information on the initial hand form, the location where the sign is articulated (face, lips, cheek, chest, neutral space etc.) and the movement with which the sign is formed. Based on these three parameters, it is possible – for the first time in Estonian lexicography – to search for a certain sign by choosing the hand form, the location, or the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign. The interface allows for searching in the opposite direction as well, making it possible for the non-deaf to learn Estonian Sign Language.



kass nimisõna ⟨k`as's, kassi, k`assi; *mitmus* kassid, k`asside, k`as'se⟩
väike karvane loom, keda võetakse lemmikloomaks
Meie kassil on pojad.
SAAB MOODUSTADA: kassipoeg

Figure 5: Online entry for the noun *kass* 'cat' in the Estonian Sign Language – Estonian Language online-dictionary.

4 The BED as a lexical resource in the Dictionary Writing System EELex

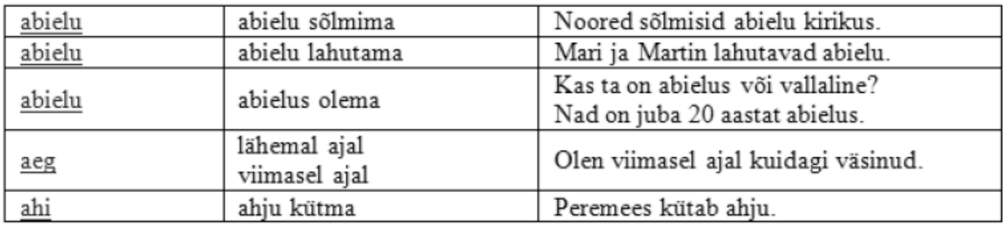
The BED was compiled in the web-based dictionary writing system EELex[4] (Jürviste et al. 2011). Nearly 50 dictionaries of different types (monolingual and bilingual, general and learner's dictionaries, etc.)

[3] http://www.eki.ee/dict/viipekeel/ [01/04/2014]
[4] http://eelex.eki.ee/ [01/04/2014]

with a standard XML mark-up make EELex a multi-purpose lexicographical database. XML-based compilation allows for the generation of different outputs: for example, specialised dictionaries based on partial database output (Kallas, Langemets 2012). There are two options for the automatic generation of specialised dictionaries: reorganising the preview (and layout) of the existing dictionary articles, or generating a new dictionary database (i.e. to clone only a part of the source database).

The function of the article preview generator makes it possible to modify the preview, i.e. to set a character, text or line break between, in front of or after a specific element or group of elements, to show or hide specific elements in the article editing preview, to assign a condition for displaying a specific element (according to the value of the attribute or neighbouring elements) or to assign a hyperlink to an element. So, by specifying elements in the print preview, it is possible to get output consisting of only those elements that are specified by the user.

The same result may be achieved by the customization of the regular XML query. It is possible to select particular elements to be displayed instead of the whole content of the dictionary article. Table 1 shows a dictionary-like extract from the BED database consisting only of the following elements: lemma, collocational patterns and usage example.



abielu	abielu sõlmima	Noored sõlmisid abielu kirikus.
abielu	abielu lahutama	Mari ja Martin lahutavad abielu.
abielu	abielus olema	Kas ta on abielus või vallaline? Nad on juba 20 aastat abielus.
aeg	lähemal ajal viimasel ajal	Olen viimasel ajal kuidagi väsinud.
ahi	ahju kütma	Peremees kütab ahju.

Table 1: An example of the collocations extracted from the BED database.

In this way, it is possible to reuse the BED database in order to generate specialised dictionaries (e.g. a dictionary of government and collocations).

5 Conclusion

The XML-based compilation makes the BED database a useful lexical resource, which can be used in different ways for development materials meant for the teaching and learning of the Estonian language as a second or a foreign language. The dictionary is special in many ways. It is the first monolingual dictionary meant for learners of Estonian at the elementary and intermediate levels (previously there were bilingual dictionaries). Government and collocation patterns were analysed and selected using the Sketch Engine corpus query system. The online version allows learners to listen to the pronunciation of words. In addition, the morphological analysis implemented in the BED make it a very innovative and user-friendly dictionary.

The first online Estonian Sign Language – Estonian Language dictionary has also been compiled. This dictionary is unique in that it enables the deaf dictionary user to find the Estonian equivalent for a sign, and not only for a word.

In future, it may be possible to convert the dictionary web page into a language-learning portal by combining the dictionary with other resources (corpora, different specialised dictionaries etc.).

6 References

Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., Viks, Ü. (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In I. Kozem, K. Kozem (eds.) eLexicography in the 21st Century: New Applications for New Users, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovenian Studies, pp. 106-112.

Kallas, J., Langemets, M. (2012). Automatic Generation of Specialized Dictionaries Using the Dictionary Writing System EELex. In A. Tavast, K. Muischnek, M. Koit (eds.) Human Language Technologies – The Baltic Perspective, Proceedings of the Fifth International Conference Baltic HLT 2012. IOS Press, (Frontiers in Artificial Intelligence and Applications), pp. 103-110.

Kallas, J. (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. PhD thesis. Tallinn: Tallinna Ülikool.

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.

Acknowledgements

The project was supported by the National Programme for Estonian Language and Cultural Memory (2009–2013), as well as the National Programme for Estonian Language Technology (2011–2017).