Automatic generation of the Estonian Collocations Dictionary database

Jelena Kallas[1], Adam Kilgarriff[2], Kristina Koppel[1], Elgar Kudritski[1], Margit Langemets[1], Jan Michelfeit[2], Maria Tuulik[1], Ülle Viks[1]

[1] Institute of the Estonian Language, Tallinn, Estonia
[2] Lexical Computing Ltd., Brighton, England
E-mail: jelena.kallas@eki.ee, kristina.koppel@eki.ee, elgar.kudritski@eki.ee, margit.langemets@eki.ee, jan.michelfeit@sketchengine.co.uk, maria.tuulik@eki.ee, ylle.viks@eki.ee

Abstract

This paper reports on the process of the automatic generation of the Estonian Collocations Dictionary (ECD) database. The database has been compiled by the Institute of the Estonian Language in collaboration with Lexical Computing Ltd. The ECD is a monolingual online scholarly dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The dictionary contains about 10,000 headwords, including single and multi-word lexical items. The collocates within each headword are grouped according to the lexico-grammatical structure formed by the collocational phrase, and for collocations example sentences are provided.

For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al., 2004) functions Word List, Word Sketch and Good Dictionary Example (GDEX) were used. The data were automatically extracted in an XML format from the 463-million-word Estonian National Corpus and imported into the XML-based EELex dictionary writing system. To make the importing of automatically extracted data from Sketch Engine into EELex possible, the XML structure for extracted data was matched with the XML structure of ECD in EELex. The ECD project started in 2014 and the dictionary is scheduled to be published in 2018.

Keywords: Corpus Lexicography; Collocations Dictionary; Corpus Query System; Dictionary Writing System; Estonian language

1. Introduction

Due to corpus lexicography development, the automatic generation of lexicographic databases has become a more and more common practise in e-lexicography. Adam Kilgarriff (2013: 78) points out that a corpus can support many aspects of dictionary creation: headword list development; the writing of individual entries, discovering word senses and other lexical units (fixed phrases, compounds, etc.); identifying the salient features of each lexical unit, their syntactic behaviour, the collocations they participate in, and any preferences they have for particular text-types or domains; and providing examples and translations.

As the focus of this article is on collocations, we will discuss the methods that are used for compiling collocations dictionaries and generating collocations databases. Based on the corpus analysis, two main approaches are implemented: automatic and semi-automatic. In the automatic approach, collocational information is automatically extracted from the corpus query system, users get direct access to non-edited collocation patterns and corpora example sentences through web interface, and no editorial work is done in terms of selecting and editing collocations. In the semi-automatic approach, collocational information is automatically extracted from the corpus query system and editorial work is done in order to clean and supplement the database, to reorder the collocates, to edit example sentences, etc.

Examples of the first approach include the projects SkELL (Baisa & Suchomel, 2014) and Wortprofil 2012 (Didakowski & Geyken, 2013). For the SkELL project, the Sketch Engine (Kilgarriff et al., 2004) function Word Sketch was used to discover collocates. By clicking on a collocate, a concordance with highlighted headwords and collocates is shown to users. SkELL uses a large text collection – SkELL corpus – specially gathered for the purpose of English language learning. There are more than 60 million sentences in the SkELL corpus and more than one billion words in total. This amount of textual information provides sufficient coverage of the everyday, standard, formal and professional English language. Wortprofil 2012 provides separated co-occurrence lists for 12 different grammatical relations and links them to their corpus contexts, where the node word and it's collocate co-occur. The co-occurrence lists and their ordering are based on statistical computations over a fully automatic annotated German corpus containing about 1.8 billion tokens.

The second approach was implemented, for example, by Kosem et al. (2013). The corpus data (grammatical relations, collocations, examples and grammatical labels) were automatically extracted from the 1.18-billion-word Gigafida corpus of Slovene. After the data were extracted, they were post-processed by lexicographers. Analytical and editorial tasks were undertaken.

From the user's point of view, both approaches have their advantages. Providing users with edited, proofread material follows the classical conception of academic dictionary publication. The editorial team has full control over the outcome on each level of the dictionary micro-structure (headwords, collocations, example sentences, etc.). Providing users with direct access to the non-edited corpus data also has benefits. New users are often familiar with such software systems as web search engines and they consciously or unconsciously consider the post-processing of outcomes to be a natural task. In addition, direct access to the full set of non-edited corpora examples gives learners a broader overview of a collocation's behaviour in different contexts.

In this paper, we introduce the general concept of the dictionary and describe the approach that we used for the creation of the ECD database (see also Kallas et al., 2015). The data were automatically extracted from the corpus query system Sketch

Engine[1] (Kilgarriff et al., 2004), imported into the dictionary writing system EELex[2] (Langemets et al., 2006; Jürviste et al., 2013) and will be post-processed by lexicographers. We have chosen the semi-automatic method for the following reasons. Firstly, the aim of the project was to compile an academic collocations dictionary with edited content. Secondly, the newest and the biggest Estonian National Corpus (EstonianNC)[3] does not completely fulfil the criteria for a learners' dictionary. The corpus is not balanced; mostly it consists of periodicals, forums and blogs. This means that non-standard language (e.g. slang) is presented and needs to be removed manually. In addition, as the corpus includes field-specific science journals, terminological collocations need to be analysed separately and some removed in order to provide users with general language content only. Also, the output depends on the quality of the lemmatizer, the part-of-speech tagger and the morphological analysis. In terms of the Estonian National Corpus, there are still a lot of mistakes in tagging and as a result of insufficient disambiguation. This influences the quality of the outcome. The previously conducted evaluation of the Estonian Word Sketches revealed that two-thirds or more of the collocations were assessed by lexicographers as relevant and almost one-third were assessed as irrelevant (Kallas, 2013).

2. Estonian Collocations Dictionary

The Estonian Collocations Dictionary is a monolingual online, corpus-driven, scholarly dictionary aimed at learners of Estonian as a foreign language or second language at the upper intermediate and advanced levels (B2 to C1) according to the Common European Framework of Reference for Languages. The dictionary contains about 10,000 headwords, including single lexical items and multi-word lexical items (mostly multi-word verbs).

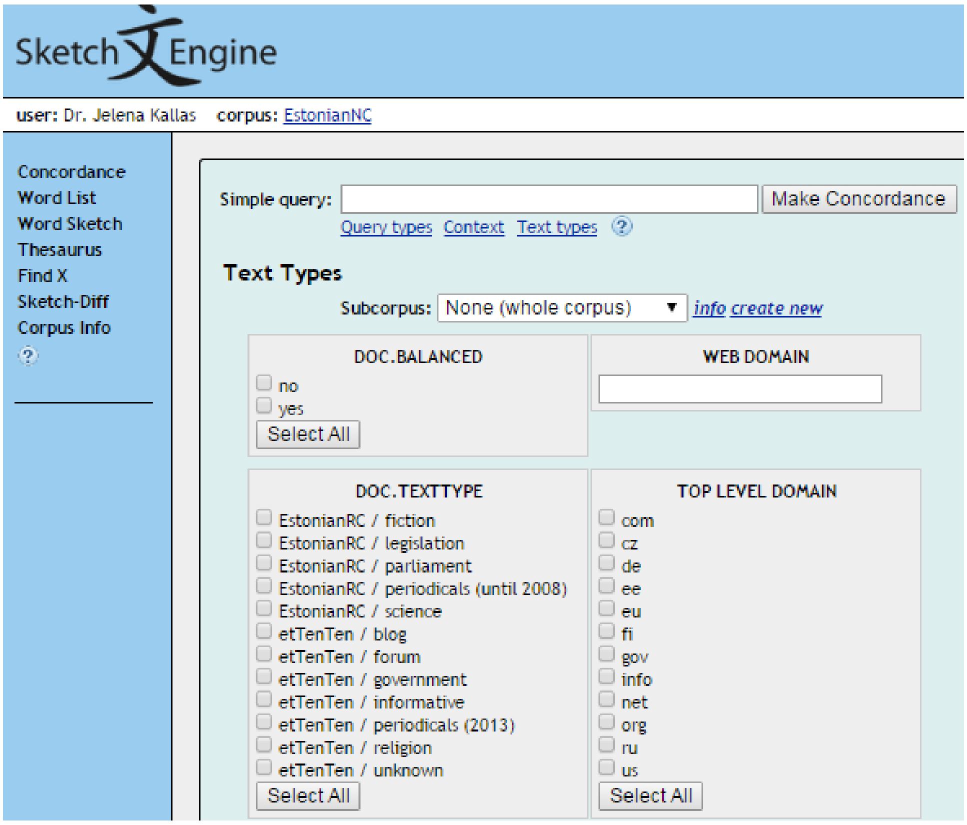
The primary source of the dictionary database is the recently compiled Estonian National Corpus (463 million words). The corpus consists of the Estonian Reference Corpus (contains texts written up to 2008) and the Estonian Web Corpus etTenTen13 (350 million tokens). etTenTen13 was compiled by Lexical Computing Ltd. It was crawled by SpiderLing (Pomikalek & Suchomel, 2012), encoded in UTF-8, cleaned and de-duplicated. The corpus was annotated morphologically, lemmatized, partially disambiguated and annotated by clauses by Filosoft LLC, and installed into Sketch Engine software.

The Estonian National Corpus has 12 subcorpora (see Figure 1).

[1] https://the.sketchengine.co.uk/auth/corpora/ (20.05.15).

[2] http://eelex.eki.ee/ (20.05.15).

[3] ske.li/estonian_national_corpus (20.05.15).



user: Dr. Jelena Kallas **corpus:** EstonianNC

Simple query:

Text Types

Subcorpus: None (whole corpus) ▼ info create new

DOC.BALANCED
□ no
□ yes

WEB DOMAIN

DOC.TEXTTYPE
□ EstonianRC / fiction
□ EstonianRC / legislation
□ EstonianRC / parliament
□ EstonianRC / periodicals (until 2008)
□ EstonianRC / science
□ etTenTen / blog
□ etTenTen / forum
□ etTenTen / government
□ etTenTen / informative
□ etTenTen / periodicals (2013)
□ etTenTen / religion
□ etTenTen / unknown

TOP LEVEL DOMAIN
□ com
□ cz
□ de
□ ee
□ eu
□ fi
□ gov
□ info
□ net
□ org
□ ru
□ us

Figure 1: Subcorpora types of the Estonian National Corpus

Periodicals form 29% of the corpus, forums and blogs form 23%, informative texts 9%, parliament and religion subcorpora 4%, and unknown texts 35%. For text-type identification, Filosoft LCC used 1) domain classification made by the Institute of the Estonian Language (e.g. periodicals and religion), 2) information in web addresses, and 3) the internal structure of the text (e.g. if a text contained a date, time or the word *vasta* 'answer-PRS-2SG', it was classified as a forum)[4]. During the mark-up of the corpus, text-type was added as metadata to the corpus.

In Estonian lexicography, the ECD project is the first dictionary focused exclusively on presenting collocational information in a systematic way. The analysis of Estonian dictionaries (Langemets et al., 2005; Kallas & Tuulik, 2011) determined that traditionally in Estonian dictionaries collocations are presented implicitly on the level of examples. The first attempt to present collocations explicitly was made in the Basic Estonian Dictionary[5] (BED) project (Kallas et al., 2014). The dictionary contains 5,000 headwords, which correspond to B1-level vocabulary. On the first level, collocations were grouped according to the lexico-grammatical structure formed by the collocational phrase, e.g. Adj+N (adjective+noun) or Adv+V (adverb+verb). All together there were 13 types of collocation patterns in BED. On the second level, noun–verb collocations were sub-grouped according to the syntactical function of

[4] http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html (19.05.15).

[5] http://www.eki.ee/dict/psv/ (19.05.15).

nouns (subject, object or adverbial), whereas other collocations were divided into semantically-motivated subgroups.

The ECD methodological conception follows the principles that were elaborated for the Basic Estonian Dictionary. The main difference is that the ECD, as a specialized dictionary, focuses on collocation patterns only; definitions are provided only for polysemous words, and there are no restrictions on vocabulary (in the BED, only words that were given as headwords in the dictionary could be used as parts of collocations). The advantage of the ECD compared to the BED is that we are able to give relevant collocations even if the frequency of one of the collocates is very low, e.g. *konn krooksub* 'frog croaks'. Often these collocations are particularly useful for learners.

For this project we define collocations as semantically transparent, meaningful and statistically significant combinations of content words with other lexical units. The typology of collocation patterns was elaborated for the ECD (see Table 1). Roth (2013: 155) indicates that in collocation lexicography one can distinguish two concepts: *node* and *collocate* (Sinclair, 1966) vs. *base* and *collocator* (Hausmann, 1985). In the ECD, we follow the concept of node and collocate, which means that each component of a collocation can be either a node or collocate, depending on the perspective. We have chosen this approach as we consider it to be more user-friendly. Our aim is for the user to find all frequent collocations connected to the headword in its entry while eliminating the need to navigate between entries. For example, if the user would like to see which nouns in Estonian collocate with the adjective *avar* 'spacious, wide, extensive', as it has a specific range of use, this can be performed within the entry of the adjective.

Noun patterns	
adjective + noun	ilus laul 'beautiful song'
noun (in genitive case) + noun	ekspertide hinnang 'expert opinion' koosoleku otsus 'the decision of the meeting'
noun (in partitive case) + noun	viil leiba 'slice of bread' viil juustu 'slice of cheese'
noun (in adverbial cases) + noun	kullast ehted 'gold jewellery'
noun (as subject) + verb	hobune hirnub 'horse neighs' palavik õtuseb, palavik langeb 'temperature rises, temperature falls'
noun (as object) + verb	arvutit sisse lülitama, arvutit välja lülitama 'turn on a computer / turn off a computer'
noun (as adverbial) + verb	aktsiatesse investeerima 'invest in stocks' arutlusele tulema 'enter into discussion'
noun+adpositional phrase	lepingu kohaselt 'according to a contract'
adverb + noun	raagus puud 'bare trees' omaette tuba 'separate room'
noun + verb in *ma*- or *da*-infinitive	meister valetama 'master to lie' soov laulda 'a wish to sing'
coordinating construction comparison constructions	päike ja tuul 'sun and wind' elu kui kabaree 'life as a cabaret'

Adjective patterns	
adjective + noun	raske otsus ‘hard decision’
noun (in adverbial cases) + adjective	rõõmsates toonides ‘in bright colours’ rõõmsal häälel ‘in a cheerful voice’
adverb + adjective	väga aeglane ’very slow’ silmatorkavalt hea ‘strikingly good’
adjective (in translative case) + verb adjective (in essive case) + verb	rikkaks saama ‘get rich’ rikkana tunduma ‘seem wealthy’
adjective + verb in *ma-* või *da*-infinitive	ilus vaadata ‘nice to look at’ raske mõista ‘hard to understand’
adjective + adjective	igavene suur ‘enormously big’
coordinating constructions comparison constructions	rikas ja ilus ‘rich and beautiful’ valge kui lumi ‘white as snow’ must nagu süsi ‘black as coal’
Adverb patterns	
adverb + adverb	aina rohkem ‘more and more’ väga kiiresti ‘very fast’
adverb + adjective	väga aeglane ‘very slow’
adverb + verb	kiiresti jooksma ‘run fast’
noun + adverb	ideid täis ‘full of ideas’
coordinating construction comparison constructions	hästi ja kiiresti ‘well and fast’ kergelt kui õhk ‘lighter than air’
Verb patterns	
adverb + verb	kiiresti jooksma ‘run fast’
noun (as subject) + verb	hobune hirnub ‘horse neighs’ palavik õtuseb, palavik langeb ‘temperature rises / temperature falls’
noun (as object) + verb	arvutit sisse lülitama, arvutit välja lülitama ‘turn on a computer / turn off a computer’
noun (as adverbial) + verb	aktsiatesse investeerima ‘invest in stocks’
adjective (in translative) + verb adjective (in essive) + verb	täiskasvanuks saama ‘to become an adult’ rikkana tunduma ‘seem wealthy’
infinite verb + finite verb	ajab nutma ‘makes me cry’ jätab maksmata ‘leaves unpaid’
coordinating construction	kirjutama ja lugema ‘to write and read’

Table 1: Collocation patterns in ECD

Components of collocations are presented as lemmas (e.g. *hea laul* (good-ADJ-SG-NOM song-SG-NOM) ‘good song’, *omaette tuba* (separate-ADV room-SG-NOM) ‘separate room’) or in particular inflectional word forms (e.g. *viil leiba* (‘slice-SG-NOM bread-SG-PART) ‘slice of bread’, *rõõmsates toonides* (bright-ADJ-PL-INE colour-PL-INE) ‘in bright colours’). In this way, learners acquire additional grammatical information, which makes it easier for them to put the collocation into use.

For the grouping of collocations, we use morphosyntactic and syntactic criteria. At the first level, we group collocates according to their word class (with nouns, with adjectives, with adverbs and with verbs). Coordinating and comparison constructions are shown as separate units. At the second level, noun–noun, adjective–noun and

adjective–verb collocates are sub-grouped according to the inflectional word form (case) of the collocate, and noun–verb collocations are sub-grouped according to the syntactical function of the nouns (subject, object or adverbial). For sorting, we rely on raw frequency information and list collocates accordingly.
All collocation patterns are illustrated with example sentences, which were extracted automatically from the EstonianNC and will be post-processed by lexicographers. Where possible, we chose authentic examples, but if needed (e.g. very long sentences, specific vocabulary, slang or rare words) the sentences are shortened and edited.

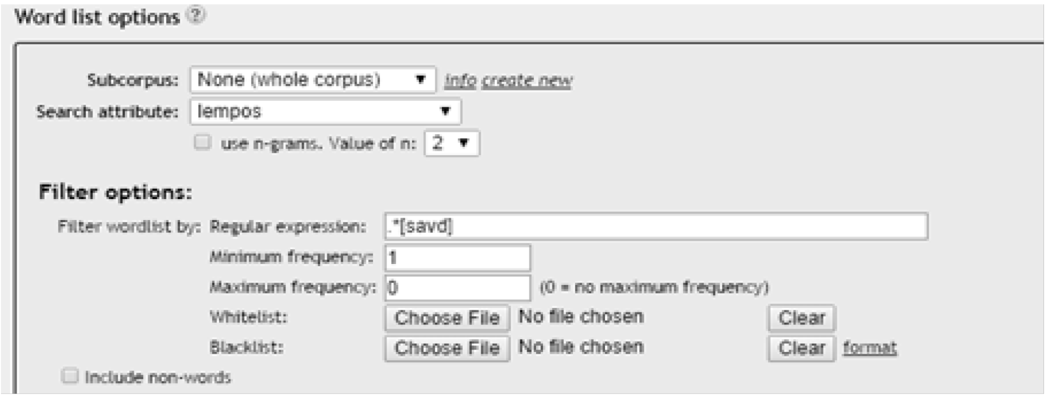
3. Automatic generation of the database

For the automatic generation of the ECD database, we implemented the methodology proposed by Kosem et al. (2013: 35–36). The information was extracted from Sketch Engine (Kilgarriff et al., 2004) in an XML-format and imported into the EELex dictionary writing system (Langemets et al., 2006; Jürviste et al., 2013). The procedure required the following: a selection of lemmas, fine-grained Sketch Grammar, GDEX (Kilgarriff et al., 2008) configuration, settings for extraction and the API script to extract data from Word Sketch.

3.1 Headword list development

The headword list of ECD contains 10,000 headwords. Only content words are presented as headwords: nouns, adjectives, verbs and adverbs. As Kilgarriff et al. (2014: 547) note, collocation dictionaries concern the core of the vocabulary: they are not for very rare words or grammatical words, but for common nouns, verbs and adjectives, which make up 99% of the headword list in a standard dictionary. In the ECD, nouns form 68%, adjectives 14%, verbs 15% and adverbs 3% of the headword list. Only manner adverbs are included in the headword list, e.g. *kergesti* 'easily' and *pehmelt* 'gently'.

For the creation of the headword list, the Sketch Engine function Word List was used.



Word list options

Subcorpus: None (whole corpus) ▾ Info create new
Search attribute: lempos ▾
□ use n-grams. Value of n: 2 ▾

Filter options:

Filter wordlist by: Regular expression: .*[savd]
Minimum frequency: 1
Maximum frequency: 0 (0 = no maximum frequency)
Whitelist: Choose File No file chosen Clear
Blacklist: Choose File No file chosen Clear format
□ Include non-words

Figure 2: Word List function in Sketch Engine

Figure 2 illustrates the general parameters that were used for the headword list generation: the whole corpus is searched; the search attribute is lempos; regular expression is used to identify only words that are tagged as nouns, adjectives, verbs or adverbs; the minimal frequency of the lemma is 1; there is no maximum frequency.

As a basis for the ECD headword list, we took the first 10,500 frequent words, which needed to be checked manually. This was necessary to eliminate "noise" derived from mistakes in tagging and from insufficient disambiguation. Some headwords had to be removed, for example headwords with two kinds of spelling (e.g. *mänedžer* vs. *mänedzher* 'manager', *šokk* vs. *shokk* 'shock', *režiim* vs. *rezhiim* 'regime'), abbreviations (e.g. *eek*, *eur* and *toim*), proper nouns and various terms (e.g. *süsinikdioksiid* 'carbon dioxide').

In parallel with corpus data analysis, we also used already existing lists of multi-word verbs. These lexical units were added manually.

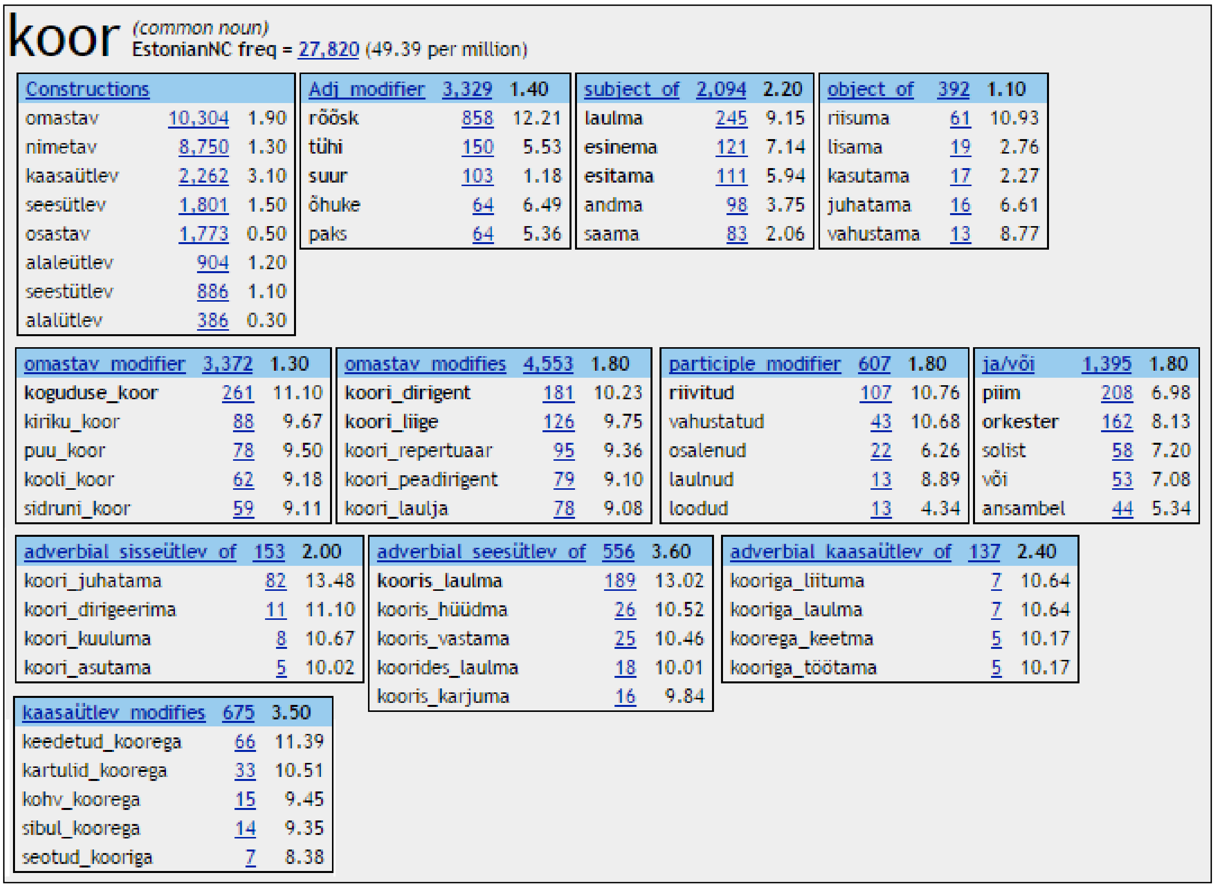
After the headword list was developed, it was divided into two frequency classes: for Class I the most frequent 5,000 words, with a minimum frequency in EstonianNC of 5057; and for Class II the 5,000 mid-frequency words, with a minimum frequency in EstonianNC of 1057. Different settings for extraction were elaborated for different frequency classes (see section 3.4).

3.2 Sketch Grammar

For the detection of collocations, the Sketch Engine function Word Sketch was used. A word sketch is a summary of a word's grammatical and collocational behaviour (Kilgarriff et al., 2004).

Estonian Word Sketch Grammar is geared towards the specification of the Estonian National Corpus and relies on lists of syntagmatic relations of Estonian nouns, adjectives, adverbs and verbs, formed on the basis of traditional and formal grammar descriptions (Kallas, 2013). Word Sketch Grammar version 1.5 for Estonian was completed in 2013 and contained 85 rules. In 2014 the new version of Sketch Grammar was elaborated. Version 1.6 has 109 rules, including 16 *unary*-type rules (which make it possible to analyse the usage of inflectional forms of nouns and adjectives), four *symmetric*-type rules (which detect coordinating and comparison constructions, for example *päike ja tuul* 'sun and wind', *ilus ja noor* 'beautiful and young', and *hoolima ja hoolitsem* 'to care and to take care'); 16 *dual*-type rules (which make it possible to search for co-occurrences of two lemmas, for example *päike* + *paistma* 'sun + shine'), and 73 *colloc*-type rules (which make it possible to detect three-word collocations, for example *hoolitsema laste eest* 'to take care of the kids', and make it possible to present two-word collocations in a way that one component is presented as a lemma and the other in the particular inflectional form, for example *kari lambaid* (flock-SG-NOM sheep-PL-PART) 'flock of sheep', *rääkima aktsendita* (talk-INF accent-SG-ABE) 'talk without an accent', and *suhtuma lugupidamisega* (treat-INF respect-SG-COM) 'to treat with respect+'[6].

Colloc-type rules proved to be very efficient for Estonian Sketch Grammar. Estonian has a rich morphological system: the nouns decline in 14 cases both in singular and plural; and verbs are inflected for tense, person, mood and voice (Liin et al., 2012). For that reason, presenting collocates as lemmas makes the whole collocation very opaque. *Colloc*-rules are particularly useful in the case of homonyms. Figure 3 displays a selection of grammatical relations for the homonyms *koor_1* (choir-SG-NOM): *koori* (choir-SG-GEN) vs. *koor_2* (peel-SG-NOM; cream -SG-NOM): *koore* (peel-SG-GEN; cream-SG-GEN), i.e. 'choir' vs. 'peel; cream': *kooris laulma* (choir-SG-INE sing-INF) 'sing in a choir', *kooriga liituma* (choir-SG-COM join-INF) 'join a choir', but *koorega kartulid* (peel-SG-NOM potato-PL-NOM) 'potatoes with peels', *koorega kohv* (cream-SG-COM coffee-SG-NOM) 'coffee with cream', etc.



koor (common noun)
EstonianNC freq = 27,820 (49.39 per million)

Constructions		
omastav	10,304	1.90
nimetav	8,750	1.30
kaasaütlev	2,262	3.10
seesütlev	1,801	1.50
osastav	1,773	0.50
alaleütlev	904	1.20
seestütlev	886	1.10
alalütlev	386	0.30

Adj modifier	3,329	1.40
rõõsk	858	12.21
tühi	150	5.53
suur	103	1.18
õhuke	64	6.49
paks	64	5.36

subject of	2,094	2.20
laulma	245	9.15
esinema	121	7.14
esitama	111	5.94
andma	98	3.75
saama	83	2.06

object of	392	1.10
riisuma	61	10.93
lisama	19	2.76
kasutama	17	2.27
juhatama	16	6.61
vahustama	13	8.77

omastav modifier	3,372	1.30
koguduse_koor	261	11.10
kiriku_koor	88	9.67
puu_koor	78	9.50
kooli_koor	62	9.18
sidruni_koor	59	9.11

omastav modifies	4,553	1.80
koori_dirigent	181	10.23
koori_liige	126	9.75
koori_repertuaar	95	9.36
koori_peadirigent	79	9.10
koori_laulja	78	9.08

participle modifier	607	1.80
riivitud	107	10.76
vahustatud	43	10.68
osalenud	22	6.26
laulnud	13	8.89
loodud	13	4.34

ja/või	1,395	1.80
piim	208	6.98
orkester	162	8.13
solist	58	7.20
või	53	7.08
ansambel	44	5.34

adverbial sisseütlev of	153	2.00
koori_juhatama	82	13.48
koori_dirigeerima	11	11.10
koori_kuuluma	8	10.67
koori_asutama	5	10.02

adverbial seesütlev of	556	3.60
kooris_laulma	189	13.02
kooris_hüüdma	26	10.52
kooris_vastama	25	10.46
koorides_laulma	18	10.01
kooris_karjuma	16	9.84

adverbial kaasaütlev of	137	2.40
kooriga_liituma	7	10.64
kooriga_laulma	7	10.64
koorega_keetma	5	10.17
kooriga_töötama	5	10.17

kaasaütlev modifies	675	3.50
keedetud_koorega	66	11.39
kartulid_koorega	33	10.51
kohv_koorega	15	9.45
sibul_koorega	14	9.35
seotud_kooriga	7	8.38

Figure 3: Word Sketch for the noun *koor* 'choir; peel; cream' (from etTenTen13)

[6] For more on directives used in the Sketch Grammar, see https://www.sketchengine.co.uk/documentation/ wiki/SkE/GrammarWriting (20.05.15).

The new Sketch Grammar version 1.6 includes all of the lexico-grammatical structures that will be presented in the collocations dictionary (see Table 1). After the new version of Estonian Sketch Grammar was elaborated, settings for extraction were developed for nouns, adjectives, adverbs and verbs; we decided on such parameters as the frequency of the grammatical relation, the frequency of the co-occurrence of the collocates and the score of collocation (see section 3.4).

3.3 GDEX configurations

GDEX (Kilgarriff et al., 2008) is a tool that rates the quality of sentences and helps the lexicographer to select the best. GDEX works as a filter: it evaluates syntactic and lexical features of sentences and sorts concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of sentences: the better candidates are at the top of the list and the not-so-good ones at the bottom. The theoretical framework for GDEX development is proposed in Kilgarriff et al. (2008) and Kosem et al. (2011) and Kosem et al. (2013).

To clarify the GDEX parameters for Estonian, we used the example sentences of the Basic Estonian Dictionary (BED) and the Dictionary of Estonian (ED) (Langemets et al., 2010, to be published in 2018), and compared them to etTenTen13 web corpora sentences. The BED and ED dictionaries were used as the gold standard for dictionary example sentences. BED example sentences are compiled by lexicographers. They are didactic units and the aim is to show how words are used in context. The target audience of the ED is not language learners but well-educated native speakers. For that reason, the level of lexicographic adaptation of example sentences is much lower. etTenTen13 corpus sentences are fully authentic.

We analysed such parameters as the minimum and maxumum number of words in a sentence, sentence length, word length and the number of subordinate clauses. Only sentences with substantives, adjectives, adverbs and verbs were taken into account. For each part of speech we analysed 150 sentences from three sources: 50 sentences from the BED, 50 sentences from the ED, and 50 sentences from etTenTen13. Tables 2 and 3 summarize the results of the analysis.

Quantitative analysis of the parameters clearly showed the peculiarity of sentences. Example sentences in BED, which has teaching purposes, are usually very short (the maximum number of words is 11, the average number of words in a sentence is 4.36–6.44). Sentences in ED are also rather short: the maximum number of words is 13 and the average number of words in a sentence is 4.72–6.42. Authentic sentences in corpora have very different characteristics. The difference is extremely large: the number of words in a sentence extends to 56 and the average number of words in a sentence is 15–16.9.

	Number of words	**Average sentence length (words)**	**Average word length (characters)**
Substantives			
BED	3–9	5.08	5.6
ED	3–12	6.42	6.7
etTenTen13	**4–40**	**15.8**	5.2
Adjectives			
BED	3–10	5.08	5.3
ED	5–11	6.44	6.7
etTenTen13	**3–37**	**15**	5.23
Verbs			
BED	3–7	4.36	6.21
ED	2–10	4.72	5.66
etTenTen13	**6–56**	**16.9**	6
Adverbs			
BED	3–11	5.44	4.96
ED	3–13	5.74	6.1
etTenTen13	**7–42**	**16.8**	5.64

Table 2: Parameters for BED and ED example sentences and etTenTen13 corpora sentences

Average word length varies only between 4.96 and 6.21 characters. At the same time, words in Estonian can be quite long, e.g. *kiiruisutamismeistrivõistlused* 'speed skating championships' (30 characters); so it is reasonable to also set maximum word lengths.

	Percentage of subordinate clauses (%)
Substantives	
BED	0%
ED	12%
etTenTen13	18%
Adjectives	
BED	0%
ED	14%
etTenTen13	**58%**
Verbs	
BED	8%
ED	10%
etTenTen13	**76%**
Adverbs	
BED	20%
ED	16%
etTenTen13	**76%**

Table 3: Percentage of subordinate clauses in BED, ED and etTenTen13 corpora sentences

The analysis of subordinate clauses showed that the number of subordinate clauses was rather small in the BED and ED example sentences, while authentic sentences in etTenTen13 web corpora included more subordinate clauses (18% in the case of substantives, 58% in the case of adjectives, and 76% in the case of verbs and adverbs) (see Table 3).

The reason for this might be that the lexicographer thinks of the example sentence as an addition to the definition and chooses not to add information that does not really illustrate a word’s use. Sentences in web corpora reflect the desire and the need to provide readers with more context.

It also appeared that all the sentences in BED and ED included a predicate. In corpus sentences, there were a lot of elliptic sentences. Corpus sentences are also characterized by a large number of proper nouns and numbers.

Based on the empirical analysis of the sentences and also on the theoretical framework proposed by Kilgarriff et al. (2008), Kosem et al. (2011) and Kosem et al. (2013), we developed the following classifiers for GDEX for Estonian:

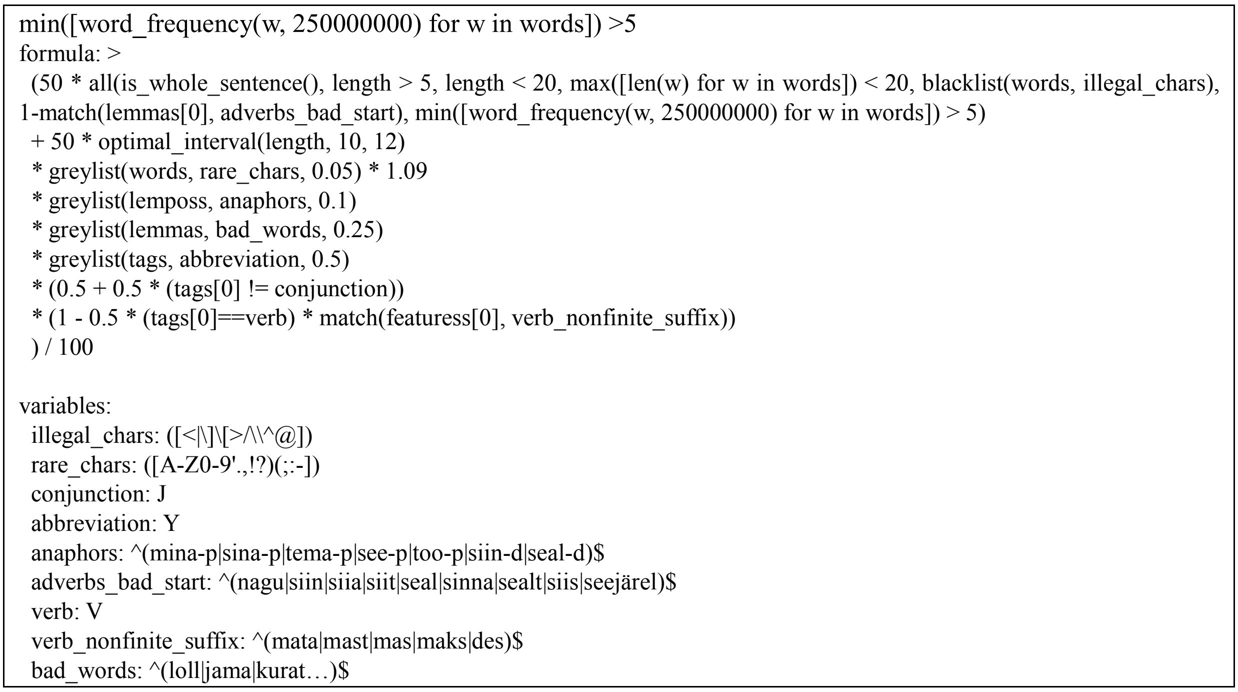
- whole sentences starting with capital letter and ending with (.), (!) or (?);
- sentences longer than five words;
- sentences shorter than 20 words;
- penalize sentences which contain words with a frequency of less than five words;
- penalize sentences with words longer than 20 characters;
- penalize sentences with more than two commas, or with brackets, colons, semicolons, hyphens, quotation marks and dashes;
- penalize sentences with words starting with capital letters. Penalize sentences with H (=Proper noun) and Y (=abbreviation) POS-tags;
- penalize sentences with “bad words”;
- penalize sentences with the pronouns *mina* 'I', *sina* 'you', *tema* 'he/she', *see* 'it' and *too* 'that', and the adverbs *siin* 'here' *seal* 'there';
- sentences should not start with the pronouns *mina* 'I', *sina* 'you' or *tema* 'he/she', or the local adverbs e.g. *siin* 'here' and *seal* 'there';
- penalize sentences which start with punctuation marks (typical informal texts) and with J (=conjunction) POS-tags;
- penalize sentences where lemmas are repeated;

- penalize sentences with tokens containing mixed symbols (e.g. letters and numbers), URLs and email addresses.

One parameter was that a sentence should contain a verb as a predicate; otherwise, the sentence was elliptical. But this parameter would only be possible to implement if the corpus was semantically annotated.

The blacklist is based on a list of words (compiled by Filosoft LCC[7]) that the Estonian speller should not offer as replacements for unknown words. To supplement the list, we analysed words in the EDE dictionary that were marked as vulgar, pejorative, colloquial or slang. We added such words as *türa* 'dick', *narkots* 'dope', etc. We also added internet acronyms (*omg*, *wtf*, *lol*, *irw*) and curse words in English and Russian (*fuck*, *pohui*) and their adapted variants (*fakk*, *pohh*). The final list contained 446 words.

Figure 4 illustrates the API script written by Jan Michelfeit for the Estonian GDEX configuration.



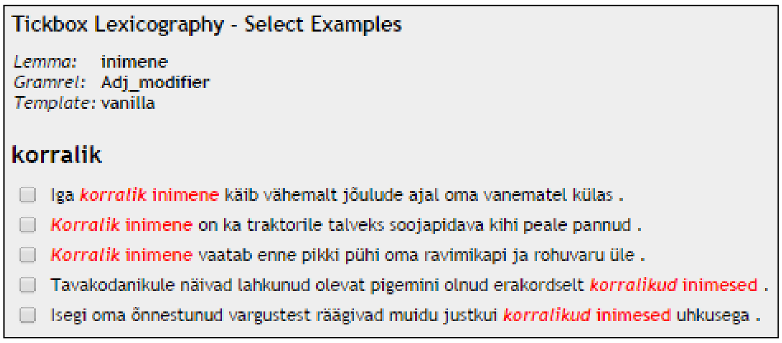
```
min([word_frequency(w, 250000000) for w in words]) >5
formula: >
  (50 * all(is_whole_sentence(), length > 5, length < 20, max([len(w) for w in words]) < 20, blacklist(words, illegal_chars),
1-match(lemmas[0], adverbs_bad_start), min([word_frequency(w, 250000000) for w in words]) > 5)
  + 50 * optimal_interval(length, 10, 12)
  * greylist(words, rare_chars, 0.05) * 1.09
  * greylist(lemposs, anaphors, 0.1)
  * greylist(lemmas, bad_words, 0.25)
  * greylist(tags, abbreviation, 0.5)
  * (0.5 + 0.5 * (tags[0] != conjunction))
  * (1 - 0.5 * (tags[0]==verb) * match(featuress[0], verb_nonfinite_suffix))
  ) / 100

variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:-])
  conjunction: J
  abbreviation: Y
  anaphors: ^(mina-p|sina-p|tema-p|see-p|too-p|siin-d|seal-d)$
  adverbs_bad_start: ^(nagu|siin|siia|siit|seal|sinna|sealt|siis|seejärel)$
  verb: V
  verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
  bad_words: ^(loll|jama|kurat…)$
```

Figure 4: GDEX configuration file[8]

As a result, the output of GDEX improved substantially. Figure 5 illustrates that after the GDEX parameters were applied, there were considerably fewer subordinate clauses in the output and sentences were generally shorter.

[7] The authors thank Heiki-Jaan Kaalep (Filosoft LCC) for the list.

[8] The list of 'bad words' is skipped.



Tickbox Lexicography - Select Examples

Lemma: inimene
Gramrel: Adj_modifier
Template: vanilla

korralik

□ Iga *korralik* inimene käib vähemalt jõulude ajal oma vanematel külas .
□ *Korralik* inimene on ka traktorile talveks soojapidava kihi peale pannud .
□ *Korralik* inimene vaatab enne pikki pühi oma ravimikapi ja rohuvaru üle .
□ Tavakodanikule näivad lahkunud olevat pigemini olnud erakordselt *korralikud* inimesed .
□ Isegi oma õnnestunud vargustest räägivad muidu justkui *korralikud* inimesed uhkusega .

Figure 5: Automatically generated sentences for the collocation *korralik inimene* 'decent person'

For each collocation, we extracted five sentences, but for less frequent collocations there could be fewer than five examples in total. In this case, the program gave all examples without applying the parameters.

For future research testing additional GDEX classifiers proposed by Kosem et al. (2013) could be considered. For example, position of lemma, second collocate (collocate of collocate), or Levenshtein distance could be applied. We could test also different GDEX configurations for each word class.

3.4 Settings for extraction

The parameters used for the extraction of data were the following:

- a list of grammatical relations for nouns, adjectives, verbs and adverbs was elaborated. For nouns, we extracted 23 grammatical relations, for adjectives nine grammatical relations, for verbs 27 grammatical relations and for adverbs five grammatical relations;
- the minimal frequency of a collocate: 10 (for the frequency I class) and five (for the frequency II class);
- the minimal salience of a collocate: positive Dice, except for three grammatical relations (N_PP, Adj_PP and V_PP) we added that the Dice should be at least 2.00 (if less than 2.00 it is mostly noise);
- the minimum frequency of the grammatical relation: 10;
- the minimum salience of the grammatical relation: positive Dice;
- the number of examples sentences for a collocate: five.

We extracted collocates in a fixed order according to grammatical relations, e.g. for nouns first come adjectives, then verbs, then other nouns, then and/or-grammatical relations. For some grammatical relations we also used stop-lists (e.g. modal verbs as collocates of nouns). Extracted collocates were ranked by frequency.

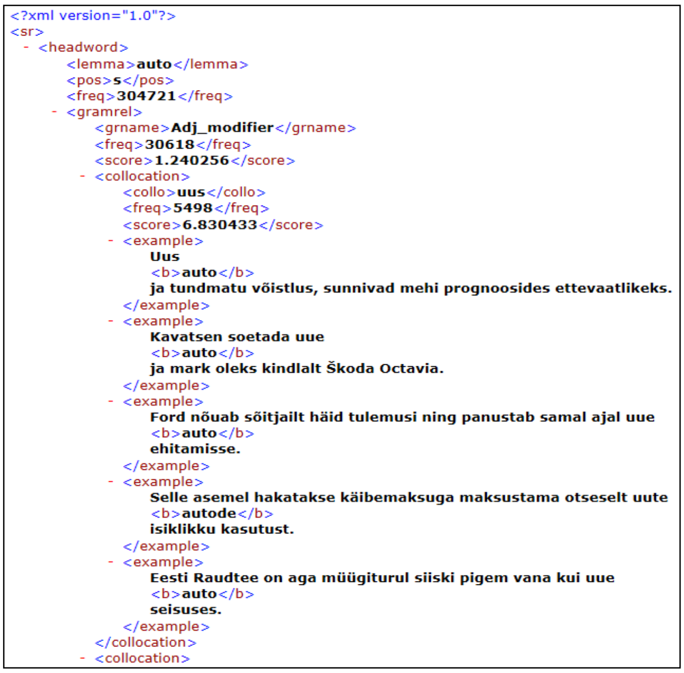
We also extracted all possible information about the frequency of collocates and grammatical relations:

- general frequency of lemmas;
- overall frequency of grammatical relations;
- overall score of grammatical relations;
- frequency of each collocate;
- score of each collocate.

Also GDEX-score could be extracted to show lexicographers how well the particular sentence corresponds to the parameters.

In perspective, it is possible to use frequency numbers for adding frequency labels (‘star rating’) to identify high-frequency, mid-frequency and low-frequency words. Also, statistical data can be used for different kinds of visualization of lexical data in the dictionary interface.

The data were extracted from Sketch Engine in XML-format (see Figure 6) and imported into the dictionary writing system EELex (Langemets et al., 2006; Jürviste et al., 2011) (see Figure 7). To make the importing of automatically extracted data from Sketch Engine into EELex possible, the XML structure for extracted data was matched with the XML structure of the ECD in EELex.



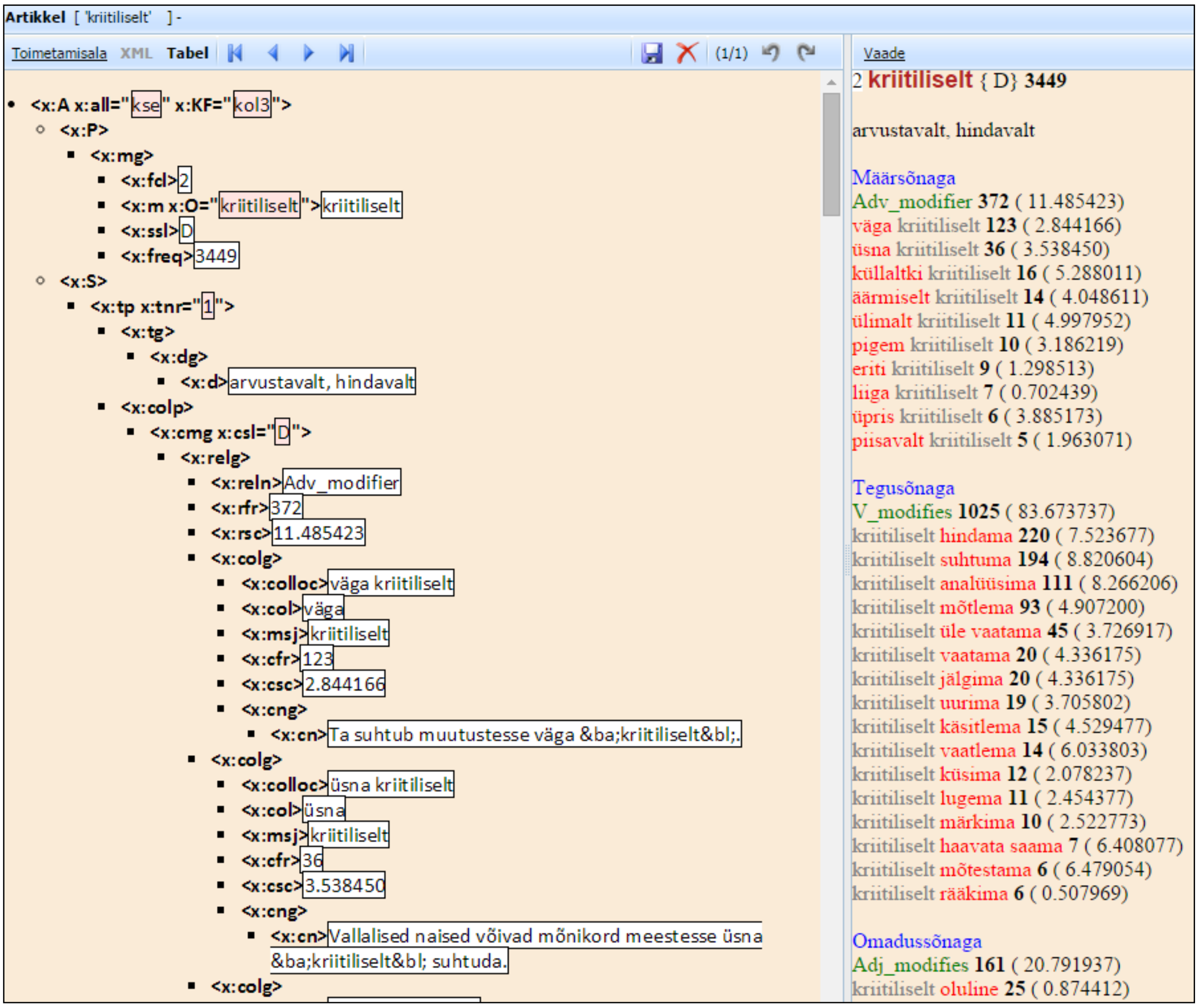
```
<?xml version="1.0"?>
<sr>
 - <headword>
      <lemma>auto</lemma>
      <pos>s</pos>
      <freq>304721</freq>
    - <gramrel>
         <grname>Adj_modifier</grname>
         <freq>30618</freq>
         <score>1.240256</score>
       - <collocation>
            <collo>uus</collo>
            <freq>5498</freq>
            <score>6.830433</score>
          - <example>
               Uus
               <b>auto</b>
               ja tundmatu võistlus, sunnivad mehi prognoosides ettevaatlikeks.
            </example>
          - <example>
               Kavatsen soetada uue
               <b>auto</b>
               ja mark oleks kindlalt Škoda Octavia.
            </example>
          - <example>
               Ford nõuab sõitjailt häid tulemusi ning panustab samal ajal uue
               <b>auto</b>
               ehitamisse.
            </example>
          - <example>
               Selle asemel hakatakse käibemaksuga maksustama otseselt uute
               <b>autode</b>
               isiklikku kasutust.
            </example>
          - <example>
               Eesti Raudtee on aga müügiturul siiski pigem vana kui uue
               <b>auto</b>
               seisuses.
            </example>
         </collocation>
       - <collocation>
```

Figure 6: XML sample of generated database

As a result, we generated a database of ECD which contains 10,939 headwords, 82,678 grammatical relations, 493,971 collocates and 2,469,855 example sentences (five example sentences for each collocate). Additionally, the database includes the part-of-speech and overall frequency number of each headword, the overall frequency of each gramrel and collocate, and the score of each gramrel and collocation.

Currently, the database is being examined, edited and supplemented by lexicographers. The manual inspection and analysis of the collocates that were disregarded in the automatic extraction process are being carried out by lexicographers.

Preliminary observations regarding editing collocations are that deleting is necessary mainly in the case of mistakes in tagging and from insufficient disambiguation; in the case of specific terms that are not part of general Estonian (*analüütiline filosoofia* 'analytical philosophy'); and in the case of very frequent words that do not combine salient collocations with headwords: *mees* 'man', *naine* 'women', *tegema* 'to do', *ajama* 'to make; to drive', etc.



```
• <x:A x:all="kse" x:KF="kol3">
   ○ <x:P>
      ▪ <x:mg>
         ▪ <x:fcl>2
         ▪ <x:m x:O="kriitiliselt">kriitiliselt
         ▪ <x:ssl>D
         ▪ <x:freq>3449
   ○ <x:S>
      ▪ <x:tp x:tnr="1">
         ▪ <x:tg>
            ▪ <x:dg>
               ▪ <x:d>arvustavalt, hindavalt
         ▪ <x:colp>
            ▪ <x:cmg x:csl="D">
               ▪ <x:relg>
                  ▪ <x:reln>Adv_modifier
                  ▪ <x:rfr>372
                  ▪ <x:rsc>11.485423
                  ▪ <x:colg>
                     ▪ <x:colloc>väga kriitiliselt
                     ▪ <x:col>väga
                     ▪ <x:msj>kriitiliselt
                     ▪ <x:cfr>123
                     ▪ <x:csc>2.844166
                     ▪ <x:cng>
                        ▪ <x:cn>Ta suhtub muutustesse väga &ba;kriitiliselt&bl;.
                  ▪ <x:colg>
                     ▪ <x:colloc>üsna kriitiliselt
                     ▪ <x:col>üsna
                     ▪ <x:msj>kriitiliselt
                     ▪ <x:cfr>36
                     ▪ <x:csc>3.538450
                     ▪ <x:cng>
                        ▪ <x:cn>Vallalised naised võivad mõnikord meestesse üsna
                        &ba;kriitiliselt&bl; suhtuda.
                  ▪ <x:colg>
```

2 **kriitiliselt** { D} **3449**

arvustavalt, hindavalt

Määrsõnaga
Adv_modifier **372** (11.485423)
väga kriitiliselt **123** (2.844166)
üsna kriitiliselt **36** (3.538450)
küllaltki kriitiliselt **16** (5.288011)
äärmiselt kriitiliselt **14** (4.048611)
ülimalt kriitiliselt **11** (4.997952)
pigem kriitiliselt **10** (3.186219)
eriti kriitiliselt **9** (1.298513)
liiga kriitiliselt **7** (0.702439)
üpris kriitiliselt **6** (3.885173)
piisavalt kriitiliselt **5** (1.963071)

Tegusõnaga
V_modifies **1025** (83.673737)
kriitiliselt hindama **220** (7.523677)
kriitiliselt suhtuma **194** (8.820604)
kriitiliselt analüüsima **111** (8.266206)
kriitiliselt mõtlema **93** (4.907200)
kriitiliselt üle vaatama **45** (3.726917)
kriitiliselt vaatama **20** (4.336175)
kriitiliselt jälgima **20** (4.336175)
kriitiliselt uurima **19** (3.705802)
kriitiliselt käsitlema **15** (4.529477)
kriitiliselt vaatlema **14** (6.033803)
kriitiliselt küsima **12** (2.078237)
kriitiliselt lugema **11** (2.454377)
kriitiliselt märkima **10** (2.522773)
kriitiliselt haavata saama **7** (6.408077)
kriitiliselt mõtestama **6** (6.479054)
kriitiliselt rääkima **6** (0.507969)

Omadussõnaga
Adj_modifies **161** (20.791937)
kriitiliselt oluline **25** (0.874412)

Figure 7: The presentation of the extracted data in EELex: editing window in XML view (left) and dictionary entry preview (right).

Regarding example sentences, although the initial idea was to present edited example sentences for each collocation, this proved to be too time-consuming. For one group, this can amount to 20 collocations and for one headword there are several collocational groups, thus leading to more than 200 sentences per entry. Therefore, we decided to give separate example sentences only for each collocation containing a verb and provide at least one example per group for other grammatical relations: adjective–noun, noun–noun, adverb–adjective, etc.

Figure 7 demonstrates the presentation of an outcome in the dictionary writing system EELex.

4. Conclusions

For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al., 2004) Word List, Word Sketch and Good Dictionary Example (GDEX) functions were used. The data were automatically extracted in an XML format from the 463-million-word Estonian National Corpus and imported into the XML-based EELex dictionary writing system (Langemets et al., 2006; Jürviste et al., 2011). To make the importing of automatically extracted data from Sketch Engine into EELex possible, the XML structure for extracted data was matched with the XML structure of the ECD in EELex.

We implemented the methodology proposed by Kosem et al. (2013). The procedure required the following: a selection of lemmas, fine-grained Sketch Grammar, GDEX (Kilgarriff et al., 2008) configuration, the API script to extract data from Word Sketch and settings for extraction. The list of lemmas was compiled using the Word List function. The latest Sketch Grammar version 1.6 was developed and improved; it includes all of the lexico-grammatical structures that will be presented in the ECD. The Grammar contains 116 rules in total. For the extraction of dictionary examples, the first version of GDEX for Estonian was developed. Classifiers connected with sentence optimum length, word optimum length, number of punctuation marks, word frequency, lemma repetition, anaphors, tokens with capital letters and symbols, abbreviations and a list of 'bad words' were proposed and implemented. The use of classifiers brought significant improvements to the output.

For automatic extraction, the following parameters were specified: a list of grammatical relations, minimum frequency and salience of grammatical relations, the number of collocates per grammatical relation, the minimum frequency and salience of a collocate, and the number of examples per collocate.

As a result, the database contains 10,939 headwords, 82,678 grammatical relations, 493,971 collocates and 2,469,855 example sentences (five example sentences for each collocate). Additionally, the database includes the part of speech and overall frequency number of each headword, the overall frequency of each gramrel and collocate, and the score of each gramrel and collocation. Currently, the database is being examined, edited and supplemented by lexicographers.

5. Acknowledgements

The Estonian collocations dictionary project is supported by the National Programme for Estonian Language and Cultural Memory II (2014–2018) and the National Programme for Estonian Language Technology (2011–2017). The authors would also like to thank COST Action "European Network of e-Lexicography" for the opportunity to share knowledge and participate in network activities.

6. References

Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Tribun EU, 2014. pp. 63–70.

Didakowski, J. & Geyken, A. (2013). From DWDS corpora to a German Word Profile – methodological problems and solutions. In Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network "Internet Lexicography". Mannheim: Institut für Deutsche Sprache. (OPAL - Online publizierte Arbeiten zur Linguistik X/2012), pp. 43–52. Available at: http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf.

EDE: *Eesti keele seletav sõnaraamat I–VI* [*The Explanatory Dictionary of Estonian*]. (2009). 2nd edition. M. Langemets, M. Tiits, T. Valdre, L. Veskis, Ü. Viks (eds.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan (eds.) *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.06.1984. Tübingen: Niemeyer*, pp. 118–129.

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. Kosem & K. Kosem (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. Available at: http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-13.pdf.

Kallas, J. & Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics], 7, pp. 59–75.

Kallas, J. (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus - ja õppeleksikograafias. [Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography.] Tallinn: Tallinn University. Dissertations on Humanities Sciences.

Kallas, J.; Tuulik, M. & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian. In A. Abel, C. Vettori & N. Ralli (eds.) Proceedings of the XVI EURALEX International Congress: The User in Focus, 15–19 July 2014, Bolzano/Bozen. Bolzano/Bozen: European Academy, pp. 1109–1119. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX_Part_3.pdf.

Kallas, J.; Koppel, K. & Tuulik, M. (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. [New possibilities in corpus lexicography based on the example oft he Estonian Collocations Dictionary.] Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics], 11, pp. 75–94.

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 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra*, pp. 425–432.

Kilgarriff, A. (2013). Using Corpora and the Web as Data Sources for Dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography.* Bloomsbury, London. Chapter 4.1, pp. 77–96.

Kilgarriff, A.; Rychlý, P.; Jakubicek, M.; Kovář, V.; Baisa, V. & Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. In N. Calzolari, N. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (eds.) *LREC (Language Resources and Evaluation Conference), Reykjavik, Iceland*, pp. 454–552. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/52_Paper.pdf.

Kosem, I.; Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, pp. 151–159. Available at: http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-19.pdf.

Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia*, pp. 17–19. Available at: http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf.

Langemets, M.; Mägedi, M. & Viks, Ü. (2005). Süntaktiline info sõnastikus: probleeme ja väljavaateid. *Eesti Rakenduslingvistika Ühingu aastaraamat* [*Estonian Papers in Applied Linguistics*], 1, pp. 71–98.

Langemets, M.; Loopmann, A. & Viks, Ü. (2006). The IEL dictionary management system of Estonian. In G.-M. De Schryver (ed.) DWS 2006: Proceedings of the

Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System. Turin, 5th September 2006. Turin: University of Turin, pp. 11–16. Available at: http://nlp.fi.muni.cz/dws06/dws2006.pdf.

Langemets, M.; Tiits, M; Valdre, T. & Voll, P. (2010). In spe: ühekőiteline eesti keele sõnaraamat. *Keel ja Kirjandus*, 11, pp. 793–810.

Liin, K.; Muischnek, K.; Müürisep, K. & Vider, K. (2012). *Eesti keel digiajastul* [*The Estonian Language in the Digital Age*]. Valge raamatu sari [White Paper Series]. G. Rahm ja H. Uszkoreit (eds.). Heidelberg [etc.]: Springer.

Pomikalek, J. & Suchomel, V. (2012). Efficient web crawling for large text corpora. In A. Kilgarriff & S. Sharoff (eds.) *Proceedings of the 7th Web-as-Corpus workshop, Lyon, France*, pp. 39-43.

Roth, T. (2013). Going Online with a German Collocations Dictionary. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia, pp. 152–163. Available at: http://eki.ee/elex2013/proceedings/eLex2013_11_Roth.pdf.

Sinclair, J. (1966). Beginning the Study of Lexis. In C. E. Bazell et al. (eds.) In Memory of J. R. Firth. London: Longman, pp. 410–430.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/