



Eesti Keele Instituut



Example sentences in Estonian learners' dictionaries

Kristina Koppel

Institute of the Estonian Language /
University of Tartu

General outline

- Short overview of functions of the IEL
 - Dictionary projects
- Estonian Collocations Dictionary
 - Automatic extraction of the dictionary database
- Example sentences in learners' dictionaries
 - Good Dictionary EXample (GDEX)
- Future plans

Institute of the Estonian Language

- National Research and Development institution
- Researches modern Estonian, the history of the Estonian language, Estonian dialects and Finno-Ugric cognate languages
- Guarantees the norms of the Estonian standard language
- Has the status of the Office of Onomastic Expertise



Public functions of the IEL

- Compilation and upgrading dictionaries and databases
- Free public language advice
- Language care and language planning
- Coordination of nationwide terminological work
- The collection and development of language archives
- Services for people with special needs: speech technology (speech synthesis), dictionaries of sign language
- Online services for language learners

Dictionary projects (1)

- „Estonian Dictionary of Word Families“ (2012)
 - 9000 word families
 - For linguists, lexicographers, teachers of Estonian, language enthusiasts
- „Basic Estonian Dictionary“ (BED) (2014)
 - For learners of Estonian at levels A2-B1
 - 5000 headwords
 - Auxiliary materials (picture pages, study pages, table of countries, peoples, languages, grammar tables)
 - Audio files, speech synthesis

Dictionary projects (2)

- „Estonian Collocations Dictionary“ (ECD) (2018)
 - Corpus-based monolingual online scholarly dictionary
 - For learners of Estonian at levels B2-C1
- Team:



PhD Jelena Kallas
project manager, scientific
secretary-computational
lexicographer



Maria Tuulik
lexicographer-junior
researcher, PhD student



Kristina Koppel
lexicographer-junior
researcher, PhD student



PhD Geda Paulsen
Senior-lexicographer

Collocations

- Semantically transparent, meaningful and statistically significant combinations of content words with other lexical units, e.g. *ilus ilm* ‘beautiful weather’, *tugev vihm* ‘heavy rain’, *koer haugub* ‘the dog barks’

Automatic extraction of the ECD database

- Generated semi-automatically
- CQS Sketch Engine
- Estonian National Corpus (463 million words)
- XML-based in-house DWS EELEX

- a selection of lemmas
- Sketch Grammar
- GDEX configuration
- API script

- 10,939 headwords
 - 82,678 gramrels
- 493,971 collocates
- 2,469,855 example sentences



saabas 'boot' in EELex

The screenshot displays the XML structure of the word 'saabas' in the EELex database. The XML is organized into several levels: <x:cmg>, <x:relg>, <x:reln>, <x:rfr>, <x:rsc>, <x:colg>, <x:colloc>, <x:col>, <x:msj>, <x:cfr>, <x:csc>, and <x:cng>. The <x:cng> elements contain example sentences illustrating the word's usage in different contexts. The right-hand panel shows the word 'saabas' with its frequency data for various grammatical categories and parts of speech.

XML Structure:

- <x:cmg x:csi="A">
 - <x:relg>
 - <x:reln>Adj_modifier
 - <x:rfr>1304
 - <x:rsc>1.926536
 - <x:colg>
 - <x:colloc>uued saapad
 - <x:col>uued
 - <x:msj>saapad
 - <x:cfr>148
 - <x:csc>1.642161
 - <x:cng>
 - <x:cn>Uueks aastaks oleks vaja veel uusi &ba;saapaid&bl; ja seelikut.
 - <x:cn x:as="ab">Talv saab läbi ja ei taha uute &ba;saabaste&bl; ostmisele raha kulutada.
 - <x:cn x:as="ab">Lapsel on vaja uued &ba;saapad&bl; osta ja breketid panna.
 - <x:cn x:as="ab">Kullo Kender oli oma uued &ba;saapad&bl; täiega välja teeninud.
 - <x:cn x:as="ab">Eile õhtul otsisid uisuliidu esindajad Berni poodidest juba uut &ba;saabast&bl;.
 - <x:colg>
 - <x:colloc>vanad saapad
 - <x:col>vanad
 - <x:msj>saapad
 - <x:cfr>78
 - <x:csc>2.836141
 - <x:cng>
 - <x:cn x:as="ab">Vanad &ba;saapad&bl; polnud küll veel päris läbi, aga ikkagi.
 - <x:cn x:as="ab">Väga ebameeldiv oli vanade &ba;saabastega&bl; vana-aasta õhtul välja minna.
 - <x:cn x:as="ab">Minu praegusi alles olevad vanad &ba;saapad&bl; on metallist lukuga.

Vaade Standard

1 **saabas** nimisõna {S} 8522

(tavaliselt mitmuses)

Omadussõnaga

Adj_modifier **1304** (1.926536)
uued saapad **148** (1.642161)
vanad saapad **78** (2.836141)
mustad saapad **59** (3.879268)
porised saapad **51** (8.323421)
head saapad **37** (0.046835)
kõrged saapad **31** (2.376672)
valged saapad **28** (2.804342)
madalad saapad **28** (3.156037)
korralikud saabas **28** (2.856138)
rasked saapad **28** (1.449554)
pikad saapad **27** (1.443609)
soojad saapad **17** (2.342852)
tavalised saapad **16** (1.622823)
märjad saapad **15** (4.173928)
pruunid saapad **15** (4.712365)
katkised saapad **14** (5.351257)
mugavad saapad **14** (3.451556)
karvased saapad **12** (5.185818)

Tegusõnaga

subject_of **643** (2.281389)

rahvalaul 'folk song' in online version of ECD

rahvalaul nimisõna 2841 2015-12-18T13:57:45

Omadussõnaga

Näited

•

Adj_modifier 427

vaimulik rahvalaul

• Vaimulikud **rahvalaulud** on midagi enam kui lihtsalt vaimuliku sisuga laulud.

regivärsiline rahvalaul

vana rahvalaul

riimiline rahvalaul

lõppriimiline rahvalaul

traditsiooniline rahvalaul

mitmehäälneline rahvalaul

ehe rahvalaul

lüüriline rahvalaul

müütiline rahvalaul

iidne rahvalaul

•

participle_modifier 65

tuntud rahvalaul

Tegusõnaga

Näited

•

subject_of

rahvalaul **kõlab**

• Kontserdil kõlavad lüürilised rahvalaulud mõisahärradest, lossipreilidest ja armastusest .

•

object_of 72

rahvalaule **laulma**

rahvalaule **esitama**

rahvalaule **koguma**

• Ta kogub **rahvalaule**.

rahvalaule **teadma**

rahvalaule **õppima**

Nimisõnaga

Näited

•

omastav_modifies 468

rahvalaulude **töötlus**

• Lauupeo programm koosnes **rahvalaulude** töötlustest ja kaasaegsete heliloojate loomingust.

rahvalaulude **kogumik**

rahvalaulude **kogumine**

rahvalaulude **laulmine**

rahvalaulude **esitamine**

rahvalaule **viis**

ja/või

Näited

•

ja/või 145

muusika ja rahvalaul

• Kavas oli põhiliselt eesti vaimulik muusika ja **rahvalaul**.

rahvaluule ja rahvalaul



Good Dictionary Example (GDEX)

- software part of Sketch Engine
- evaluates syntactic and lexical features
- scores and sorts the sentences

```
formula: >
(50 * is_whole_sentence() * blacklist(words, illegal_chars) * blacklist(lemmas, parsnips)
+ 50 * optimal_interval(length, 10, 14)
* greylist(words, rare_chars, 0.1)
* greylist(tags, pronouns, 0.1)
) / 100
variables:
illegal_chars: ([<|\|\\[>/\\^@])
rare_chars: ([A-Z0-9'.,!?) (::-])
pronouns: PRON.*
parsnips: ^(tory, whisky, jesus, cowgirl, meth, commie, bacon)$
```

Syntax of GDEX configuration files

GDEX for Estonian (1)

- **2014:** Pilot study on sentences in BED, Dictionary of Estonian, web corpus etTenTen13
 - Dictionary examples – gold standard
 - Web corpus sentences – fully authentic
- ?sentence length, word length, the number of subordinate clauses, keyword position etc.
- First configuration → extraction of example sentences for ECD

GDEX for Estonian (2)

- **PhD thesis:** „Parameters of example sentences for Estonian learners' dictionaries“ (supervisors: Jelena Kallas, Raili Pool)
- **2015:** second configuration
 - Corpus containing only sentences picked out by GDEX
- **2016:** STSM in Ljubljana University/Trojina Institute (host: Iztok Kosem) → third configuration

GDEX for Estonian (3)

- Two training datasets:
 - selected examples
 - rejected/non-selected examples
- A script for extracting features of datasets:
 - sentence length
 - sentence initial word (determined by word class)
 - keyword position
 - number of pronouns, adverbs, numerals, commas etc. in the sentence

GDEX editor

Old GDEX configuration

```
formula: >
(50 * all(is_whole_sentence(), blacklist(words, illegal_chars), min([word_occurrences(w) for w in
* 50 * optimal_interval(length, 8, 20)
* sum([1/length for w in words if word_frequency(w, 1000000) > 1])
) / 100
variables:
illegal_chars: ([<|\\|>|\\@])
```

Corpus

estonianNC

Metadata

- info.id
 info.author
 info.newspaperNumber
 info.heading
 info.article
 info.exercise
 info.subheading
 info.bottom
 info.chapter
 info.title
 info.unk
 doc.t2id
 doc.tid
 doc.urldomain
 doc.id
 doc.length
 doc.url
 doc.web_domain
 doc.crawl_date
 doc.langdiff
 doc.texttype
 doc.filename
 doc.balanced
 doc.wordcount
 p.heading

CQL query

[lemma="kits"]

Concordance size: 7792

GDEX configuration

```
formula: >
(50 * all(is_whole_sentence(), blacklist(words, illegal_chars), min([word_occurrences(w) for w in
* 50 * optimal_interval(length, 8, 20)
* sum([1/length for w in words if word_frequency(w, 1000000) > 1])
* _('RARE', greylist(words, rare_chars, 0.05))
) / 100
variables:
illegal_chars: ([<|\\|>|\\@])
rare_chars: ([A-Z0-9'.!?:;-])
```

Sample size

100

Minimum distance: 0.3

0.3

Test

Old rank	Rank	Sentence	Old score	Score	RARE
1	1	Kibe töö vähese arvu töölistega ei anna laulmiseks mahti .	0.99	0.98	0.85
2	10	Alates aprillist on traataiaga piiratud platsil iga päev käinud kibe töö .	0.99	0.94	0.50
3	13	Kibe töö käib 13. korrusel tervisekeskuses , kus seab kümmekond ehitajat .	0.94	0.93	0.75
4	6	Kohe algas ka kibe töö staari imago kujundamisel .	0.89	0.98	0.80
5	18	Päike on juba ammu looja läinud , aga kunstiakadeemia õmblustoas käib kibe töö .	0.89	0.91	0.85
6	16	Balti jaamas käib kibe töö , sest perroonide uuendamine peab lõppema enne jõule .	0.89	0.92	0.90
7	44	Samal ajal käib Kautla baaslaagris kibe töö , et kogu retk edukalt kulgeks .	0.86	0.37	0.55
8	17	Ma saan ainult ette kujutada kui kibe töö käib mängude , veebilehtede etc .	0.85	0.91	0.95
9	24	Praegu käib tõllakuuris kibe töö ja tundub , et restaureerida on veel üsna palju .	0.85	0.88	0.85
10	23	Keset kibedat tööd varises ta California põietavale rannalivalle .	0.83	0.89	0.80



What are good example sentences

- Ideally short
- Syntactically and grammatically not complex
- No low-frequency words
- Headword/collocation in its typical context
- Help to understand the meaning of an unknown word


Parameters of example sentences for Estonian (1)

- full sentence
- sentence length 4-20 tokens
- optimal length 6-12 tokens
- contains a verb
- word length <20 characters
- some characters are prohibited (e.g. <|\\|>^@•.*#_~), some penalized (e.g. ;:“,”«»””×...§-)
- certain words (e.g. *pigem* ‘rather’, *seetõttu* ‘for that reason’), word pairs (e.g. *seda enam* ‘even more’, *teiste sõnadega* ‘in other words’), sentence initial tags (e.g. conjunction, abbreviation, interjection) are prohibited in the beginning of the sentence

Parameters of example sentences for Estonian (2)

- words with a frequency of <5 are prohibited
- lemmas with a frequency of <1000 are penalized
- keyword repetition is prohibited
- some pronouns, words from graylist (inc. sensitive words, profanities), abbreviations, proper names, certain non-finite constructions are penalized
- sentences containing more than 2 verbs, more than 1 adverb, more than 1 pronoun, more than 1 conjunction, more than 1 proper name, more than 1 number and more than 1 comma are penalized

GDEX output in SkE

Query raamat 204,301 > GDEX 204,301 (362.74 per million) 

Page of 10,216 [Next](#) | [Last](#)

517862 Raamatuid on lugenud tänaseks terves maailmas juba üle 10 miljoni inimese .

676700 Hiljuti sai valmis kaheksas raamat , mille ise olen kirjutanud .

Tegelikult pole alt mindud ühegi raamatuga .

Linnaametnike piltidest on koostajal plaanis välja anda raamat .

Eestis ilmub septembri algul veel üks raamat Estonia katastroofist .

263862 Ühel inimesel võib raamatuid olla käes tuhandete kroonide eest .

38427 Raamat on heas korras (raamatus on pühendus) .

Rohkete fotodega illustreeritud raamatu " Presidendi lapsed " kujundas Silver Vahtre .

211160 Raamat " Viiskümmend halli varjundit " ilmub eesti keeles neljapäeval , 22. novembril .

251678 Raamatud ja igasugu käsikirjad ja paberid on mul mitmel pool laiali , tugitoolis , põrandal .

Samasuguseid imelikke juhtumisi on ka teistes Pratchetti raamatutes .

Sa tegid mulle au , võttes seda raamatut nii tõsiselt ja nii suure armastusega .

227326 Guido Knoppi nime all ilmunud raamatud pole suurt üldse tema enda kirjutatud .

Niguliste muuseumis esitleti eile raamatut " Höbedakamber " .

Kes teosega tutvunud , arvavad , et see on väärt raamat , sobiv kink jõuluvana kotist .

Mletšini raamatus tehakse juttu neist kõigist , kuid erilist tähelepanu pälvivad muidugi välisministrid .

Võtke lahti nn kollane raamat ja te näete , kui palju on ette nähtud raha investeringuteks Kunstimuuseumile .

278216 Kord kuus laenutab Õisu raamatukogu rahvamajas raamatuid ja see toob eeskätt just vanemad inimesed kohale .

Korraldajad üritavad süüdata maailma kõrgeima jaanitule , mis saaks ära märgitud ka Guinnessi rekordite raamatus .

Pärnu maantee 10 asuva Rahva Raamatu poe juhataja sõnul võrdub poolele pinnale tõmbumine kaupluse aeglase väljasuretamisega .

Page of 10,216 [Next](#) | [Last](#)

What the future holds for ECD and GDEX

- Integrate corpora with ECD
 - User clicks on a collocate → concordances picked out by GDEX will be shown
- Estonian-Finnish online dictionary (2018)
 - Joint project of Institute for the Languages of Finland and Institute of the Estonian Language
 - Using automated lexicography
- Linking the ECD database to any bilingual dictionary designed for producing Estonian as L2

.. more in the future

- Course book corpus
 - Digitization of course books
 - Annotation of full texts and vocabulary lists
 - Due to time limitation no evaluation process of books/chapters
- Sentence selection
 - Combining rule-based approach with machine learning methods

References

- Gantar, P. & Kosem, I. & Krek, S. (2016). Discovering automated lexicography: the case of the Slovene database. In *International Journal of Lexicography*, 29 (2), 200–225. doi: 10.1093/ijl/ecw014
- Kallas, J. & Koppel, K. & Tuulik, M. (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. [New possibilities in corpus lexicography based on the examples of the Estonian Collocation Dictionary.] In *Eesti Rakenduslingvistika Ühingu aastaraamat*, 11, 75–94. doi: 10.5128/ERYa11.05
- Kallas, J.; Kilgarriff, A.; Koppel, K.; Kudritski, E.; Langemets, M.; Michelfeit, J.; Tuulik, M.; Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, 1–20.
- 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 Linguística Aplicada, Universitat Pompeu Fabra, pp. 425–432.
- Koppel, K. (2017). Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. [Automatic detection of good dictionary examples in Estonian learner’s dictionaries.] In *Eesti Rakenduslingvistika aastaraamat*, pp. 53–71.
- Kosem, I.; Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In Proceedings of eLex 2011, pp. 151–159.
- 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, 17–19. http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf
- 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 20 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.