Automatic generation of the Estonian Collocations Dictionary database

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

Outline

- approaches to collocations dictionary compilation
- tools
  - Corpus Query System
  - Dictionary Writing System
- Estonian Collocations Dictionary
- automatic generation of the dictionary database
  - headword list development
  - collocations
  - example sentences
  - settings for extraction
- lexicographic analysis and editing of the database
- future development

Approaches to the compilation of collocations dictionaries

- manual

- semi-automatic
    - MacMillan Collocations Dictionary

- automatic

Automatic Collocations Dictionaries - corpus-derived electronic databases of recurrent word combinations

- automatically generated listings
    - positional extraction methods
    - relational extraction methods

    e.g. SkELL, HASK collocation dictionaries, Wortprofil 2012

Adam Kilgarriff (29.09.2014): *.. the best approach may be for LCL to provide a 'dump' of the data (headwords + collocations + examples) from which you then choose collocations and examples in different software, maybe a spreadsheet. I know you were not so keen on that approach, but it could be a lot faster - do reconsider*

Estonian Collocations Dictionary: general conception

- monolingual online scholarly dictionary
- for learners of Estonian as a foreign language or second language at the upper intermediate and advanced levels
- 10,000 headwords, incl. single lexical items and multiword verbs
- only content words as headwords: nouns (68%), adjectives (14%), verbs (15%) and adverbs (3%)
- corpus-driven
- the primary source: Estonian National Corpus (463 million words)
- collocations are listed on the basis of frequency
- collocates are presented in lemma form, e.g. *hea laul* (good-ADJ-SG-NOM song-SG-NOM) ˈgood songˈ and in word form , e.g. *viil leiba* (slice-SG-NOM bread-SG-PART) ˈslice of breadˈ
- two levels:
    - 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)
    - noun+verb collocations are sub-grouped according to the syntactical function of nouns (subject, object or adverbial)
    - example sentences

Estonian Collocations Dictionary: compilation stages

Stage I: automatic compilation of database (2014)

Stage II: lexicographic analysis and editing of the database

- sense division
- writing definitions
- adding style and domain labels

- adding collocations
- deleting all irrelevant information
- editing of example sentences
- identifying multiword lexical items and phraseology

corpus consultation is needed

Tools

- the 463-million-word Estonian National Corpus
  https://the.sketchengine.co.uk/auth/corpora/

- corpus query system Sketch Engine (Kilgarriff et al., 2004)
  - Word List
  - Word Sketch
  - GDEX

- EELex dictionary writing system (Jürviste et al., 2013)

Automatic generation of the database

- extraction of information from Sketch Engine in an XML-format

- import of information into the EELex dictionary writing system

What do we need:

- a selection of lemmas
- fine-grained Sketch Grammar
- GDEX (Kilgarriff et al., 2008) configuration
- settings for extraction

Adam Kilgarriff (30.09.2014):
LCL to prepare a db for 50-100 hwds
to specify
*   - gramrel (or a list of gramrels in a file)*
*   - GDEX configuration*
*   - number of examples per collocate*
*   - number of collocates per grammatical relation*
*   - minimum frequency of a collocate*
*   - minimum frequency of a grammatical relation*

We shall
sort by frequency
Min freq = 10
+ve Dice
We'll come back to TBL approaches if this does not give everything that is wanted

eLex 2015, UK

Headword list development

Source: Estonian National Corpus

Tool: Word list option of Sketch Engine

Two frequency classes:

Class I: the most frequent 5,000 words, with a minimum frequency in EstonianNC of 5057

Class II: the 5,000 mid-frequency words, with a minimum frequency in EstonianNC of 1057

Problems:

▪lemmatization and tagging mistakes

▪multiword units

▪proper nouns

▪terms

▪homonyms

Sketch Grammar (version 1.6)

109 rules

- 16 unary-type rules

- four symmetric-type rules (päike ja tuul ʻsun and wind', ilus ja noor ʻbeautiful and young')

- 16 dual-type rules, to search for co-occurrences of two lemmas, e.g.
  päike + paistma ʻsun + shine'

- 73 colloc-type rules, to search for

  - three-word collocations, e.g.
  V_PP: hoolitsema X eest ʻto take care of X'

  - two-word collocations in a way that one component is presented as a lemma and the other one in the particular inflectional form, e.g.
  kari lambaid (flock-SG-NOM sheep-PL-PART) ʻflock of sheep'
  rääkima aktsendita (talk-INF accent-SG-ABE) ʻtalk without an accent'

Word Sketch for the homonyms
koor_1 (choir-SG-NOM) : *koori* (choir-SG-GEN) vs.
koor_2 (peel-SG-GEN,cream -SG-NOM) : *koore* (peel-SG-GEN; cream-SG-GEN)

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

Constructions			Adj modifier	3,329	1.40	subject of	2,094	2.20	object of	392	1.10
omastav	10,304	1.90	rõõsk	858	12.21	laulma	245	9.15	riisuma	61	10.93
nimetav	8,750	1.30	tühi	150	5.53	esinema	121	7.14	lisama	19	2.76
kaasaütlev	2,262	3.10	suur	103	1.18	esitama	111	5.94	kasutama	17	2.27
seesütlev	1,801	1.50	õhuke	64	6.49	andma	98	3.75	juhatama	16	6.61
osastav	1,773	0.50	paks	64	5.36	saama	83	2.06	vahustama	13	8.77
alaleütlev	904	1.20									
seestütlev	886	1.10									
alalütlev	386	0.30									

omastav modifier	3,372	1.30	omastav modifies	4,553	1.80	participle modifier	607	1.80	ja/või	1,395	1.80
koguduse_koor	261	11.10	koori_dirigent	181	10.23	riivitud	107	10.76	piim	208	6.98
kiriku_koor	88	9.67	koori_liige	126	9.75	vahustatud	43	10.68	orkester	162	8.13
puu_koor	78	9.50	koori_repertuaar	95	9.36	osalenud	22	6.26	solist	58	7.20
kooli_koor	62	9.18	koori_peadirigent	79	9.10	laulnud	13	8.89	või	53	7.08
sidruni_koor	59	9.11	koori_laulja	78	9.08	loodud	13	4.34	ansambel	44	5.34

adverbial sisseütlev of	153	2.00	adverbial seesütlev of	556	3.60	adverbial kaasaütlev of	137	2.40
koori_juhatama	82	13.48	kooris_laulma	189	13.02	kooriga_liituma	7	10.64
koori_dirigeerima	11	11.10	kooris_hüüdma	26	10.52	kooriga_laulma	7	10.64
koori_kuuluma	8	10.67	kooris_vastama	25	10.46	koorega_keetma	5	10.17
koori_asutama	5	10.02	koorides_laulma	18	10.01	kooriga_töötama	5	10.17
			kooris_karjuma	16	9.84			

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

GDEX configurations: parameters of example sentences in the learner's dictionary, academic dictionary and WebCorpus

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

eLex 2015, UK

Subordinate clauses in the learner's dictionary, academic dictionary and WebCorpus

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

eLex 2015, UK

GDEX configurations (version 1.0)

- 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 shouldn't 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

Configuration file for GDEX

```
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)$
```

eLex 2015, UK

Settings for extraction

- for nouns 23 grammatical relations, for adjectives 9 gramrels, for verbs 27 gramrels, and for adverbs 5 gramrels
- the minimal frequency of the grammatical relation: **10**
- the minimal salience of the grammatical relation: **positive Dice**
- the minimal frequency of a collocate:
  - **10** (for the frequency I class)
  - **5** (for the frequency II class)
- the minimal salience of a collocate: **positive Dice**
- the number of collocates
  - 5-20 (depending on gramrel)
- the number of example sentences for a collocate: **5**

Generation of the database

- data were extracted from Sketch Engine in XML-format and
- imported into the dictionary writing system EELex

Database contains:

- 10,939 headwords
- 82,678 grammatical relations
- 493,971 collocates
- 2,469,855 example sentences
- the part-of-speech and overall frequency number of each headword
- the overall frequency of each gramrel and collocate
- the score of each gramrel and collocation

```xml
<?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>
```

XML sample of generated database

eLex 2015, UK

Presentation of the data in DWS

transformation of extacted collocations in DWS

Lexicographic analysis and editing of the database: selecting collocates (1)

Problems:

- quality of lemmatization, POS tagging and disambiguation
- **outcome depends on corpus content** (especially the influence of WebCorpus): internet language, slang and terminological collocations
- collocates extracted in lemma form need to be changed manually and presented in word form
- multiword collocations are extracted as duals
- **statistics**: raw frequency can be used as the basis (more suitable for A1 and A2 levels)
- salient collocates and statistical collocates in the range -5 to 5 should to be checked and added
- collocations of very frequent words often dominate and might not be relevant (e.g. *mees* 'man', *naine* 'women', support verbs, and modal verbs)
- multiword expressions (phrasal verbs, idioms) are not identified
- ja/või 'and/or' – gramrel, often not a the dual relationship, but part of a list, e.g.

    koogid, küpsised, **jäätis** *ja šokolaad* 'cakes, biscuits, **ice cream** and chocolate'

Lexicographic analysis and editing of the database: selecting collocates (2)

Solutions:

- bigger corpora
- the longest-commonest match
- possibility of updating the database
- Sketch Grammar development (NB! collocational span)

Lexicographic analysis and editing of the database: selecting examples

Problems:

- **outcome depends on corpus content** (especially the influence of WebCorpus)
- no "good sentences" for low-frequency words
- other resources should be used, e.g. WebCorp, Google, Keeleveeb and Kollokatsioonituvastaja
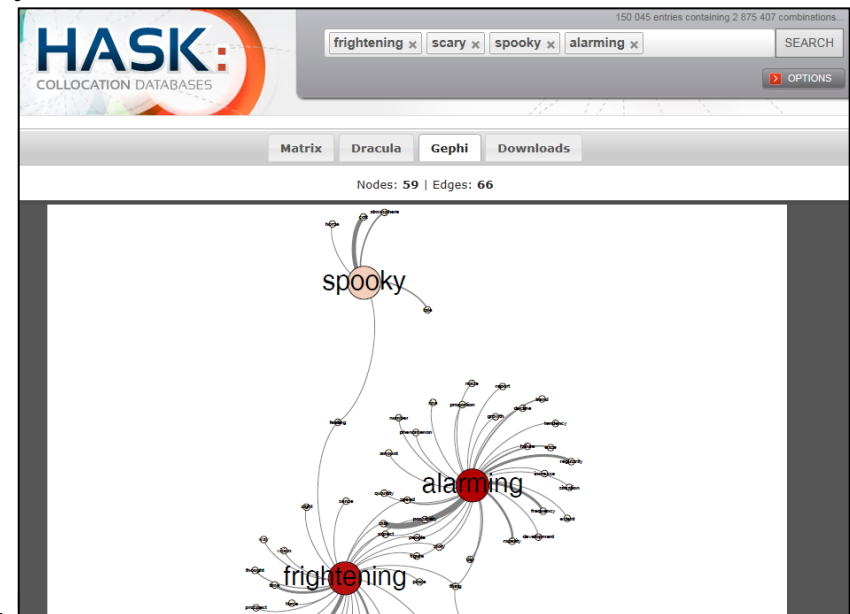
Solutions:

- bigger corpus
- GDEX development:
- testing of additional parameters, e.g. second collocate (collocate of collocate)
- position of lemma
- sorting according to GDEX score
  (Kosem et al. 2013)

kollok grupp (collocation)	
eelnev märksõna	
kollokaat (collo)	kaunis
järgnev märksõna ◆	küünlajalg
vald	
stiil	
kollok sagedus (freq)	5
kollok skoor (score)	1.294864
näidete grupp	
kol-näide (example) ab	Ma ei mäleta, milleks – ju pidi plekitükist viimaks saama kaunis &ba;küünlajalg&bl;.
kol-näide (example) ab	Lisaks vanadele piltpostkaartidele saab kambris näha kauneid &ba;küünlajalgu&bl; ning õlest ja kullatud käbidest valmistatud ehteid.
kol-näide (example) ab	Traditsioonilised jõululaua ilustused on kaunitesse kõrgetesse &ba;küünlajalgadesse&bl; asetatud tavalised antiikküünlad, jõulutähed vm. lilleseaded punaste lilledega.
kol-näide (example) ab	Õnneks ongi Tallinnas olemas jõulumeeleolule täiesti sobilik helesinine lavakujundus – Nukuteatri õuesaalis Andres Dvinjaninovi lavastatud "Bullerby jõuludel", millele Mart Haberi-Tarvo Pilleri disainibüroo Shishi vaid detailid lisas: kaunid &ba;küünlajalad&bl;, laudlinad, kuused, "ütleb moešõu korraldaja Janne Klooren.
kol-näide (example) ab	Kõigepealt tänusõnad Küllikile, kes juba kümme aastat on korraldanud KÕKU suvepäevi, kingiks sai ta nimelise särgi, kauni &ba;küünlajala&bl; ja toompihlaka istiku.

Future development

- direct access to corpus
  - access to statistical collocates
  - more example sentences (pre-compiled corpus is needed)
- release of the dictionary: layout design
  - options:
    - print-like layout (horizontal or vertical listings)
    - folded view (some units are folded away until clicked on)
    - panel view (screen is split into parts)
    - graph-based visualization
    - cloud visualization
    - combination of options, e.g.
    HASK automatic collocation dictionaries
- new techniques for editing process:
  - ? crowdsourcing
- user research

eLex 2015,

References

- Evert, S. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations.*
  Phil. Diss. Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.
- 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.
- KILGARIFF, A., RYCHLY, P. 2010. Defining the Definiendum. De Schryver, Gilles-Maurice (ed.): A way with words: recent advances in lexical theory and analysis: a festschrift for Patrick Hanks.Kampala: Menha Publishers.pp.299–312.
- 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.

References

- 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.; 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.

- 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.

- 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.; 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.

Dictionaries

BED = *Eesti keele põhisõnavara sõnastik*. 2014. J. Kallas, M. Tiits, M. Tuulik, M. Jürviste, K. Koppel (eds.). Tallinn: Eesti Keele Sihtasutus.

ED = *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.