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

eLex 2015, UK
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

eLex 2015, UK
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) ‘good song’ and in word form, e.g. viil leiba (slice-SG-NOM bread-SG-PART) ‘slice of bread’
- 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

eLex 2015, UK
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

eLex 2015, UK
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
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

eLex 2015, UK
Sketch Grammar (version 1.6)

109 rules

- 16 unary-type rules

- Four symmetric-type rules (päike ja tuul ‘sun and wind’, ilus ja noor ‘beautiful and young’)

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

- 73 colloc-type rules, to search for
  - Three-word collocations, e.g.
    V_PP: hoolitsema X eest ‘to 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) ‘flock of sheep’
    rääkima aktsendita (talk-INF accent-SG-ABE) ‘talk without an accent’

eLex 2015, UK
Word Sketch for the homonyms

*koor₁* (choir-SG-NOM) : *koori* (choir-SG-GEN) vs.

*koor₂* (peel-SG-GEN, cream -SG-NOM) : *koore* (peel-SG-GEN; cream-SG-GEN)

eLex 2015, UK
GDEX configurations: parameters of example sentences in the learner's dictionary, academic dictionary and WebCorpus

<table>
<thead>
<tr>
<th></th>
<th>Number of words</th>
<th>Average sentence length (words)</th>
<th>Average word length (characters)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Substantives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary (BED)</td>
<td>3–9</td>
<td>5.08</td>
<td>5.6</td>
</tr>
<tr>
<td>academic dictionary (ED)</td>
<td>3–12</td>
<td>6.42</td>
<td>6.7</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>4–40</td>
<td>15.8</td>
<td>5.2</td>
</tr>
<tr>
<td><strong>Adjectives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>3–10</td>
<td>5.08</td>
<td>5.3</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>5–11</td>
<td>6.44</td>
<td>6.7</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>3–37</td>
<td>15</td>
<td>5.23</td>
</tr>
<tr>
<td><strong>Verbs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>3–7</td>
<td>4.36</td>
<td>6.21</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>2–10</td>
<td>4.72</td>
<td>5.66</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>6–56</td>
<td>16.9</td>
<td>6</td>
</tr>
<tr>
<td><strong>Adverbs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>3–11</td>
<td>5.44</td>
<td>4.96</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>3–13</td>
<td>5.74</td>
<td>6.1</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>7–42</td>
<td>16.8</td>
<td>5.64</td>
</tr>
</tbody>
</table>
Subordinate clauses in the learner's dictionary, academic dictionary and WebCorpus

<table>
<thead>
<tr>
<th></th>
<th>Percentage of subordinate clauses (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Substantives</strong></td>
<td></td>
</tr>
<tr>
<td>learner's dictionary (BED)</td>
<td>0%</td>
</tr>
<tr>
<td>academic dictionary (ED)</td>
<td>12%</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>18%</td>
</tr>
<tr>
<td><strong>Adjectives</strong></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>0%</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>14%</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>58%</td>
</tr>
<tr>
<td><strong>Verbs</strong></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>8%</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>10%</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>76%</td>
</tr>
<tr>
<td><strong>Adverbs</strong></td>
<td></td>
</tr>
<tr>
<td>learner’s dictionary</td>
<td>20%</td>
</tr>
<tr>
<td>academic dictionary</td>
<td>16%</td>
</tr>
<tr>
<td>etTenTen13</td>
<td>76%</td>
</tr>
</tbody>
</table>

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

eLex 2015, UK
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|siiia|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 sample of generated database

XML sample of generated database
Presentation of the data in DWS

eLex 2015, UK
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, jääts ja šokolaad* ‘cakes, biscuits, **ice cream** and chocolate’

*eLex 2015, UK*
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 Kollokatsioonituuvastaja

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)
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

References


eLex 2015, UK
Dictionaries
