Aggregating dictionaries into the language portal Sõnaveeb: issues with and without solutions

Kristina Koppel, Arvi Tavast, Margit Langemets, Jelena Kallas
quick overview of Ekilex and Sõnaveeb

issues with and without solutions:

• consistency of information and avoiding duplicates when unifying the dictionaries
• turning dictionary-specific information into customizations of the central service
• deciding on deliberate ambiguities
• parsing data fields containing more than one data element, including textual condensation
• moving from annotating form (e.g. italics) to annotating content (e.g. a citation)
• moving from (near) duplicates to sensible information fragments
• deciding between an app and a responsive web page
• possible legal problems regarding the authorship of the new central resource
• NEW: concerns about the quantity and quality of information
why?

assumption: users look for information about the language, not about dictionaries

avoid duplication and conflicts

make dictionary compilation realistic

reduce confusion for users and sponsors
aggregated search vs one dictionary
aggregated search vs one dictionary

present differences to the user

search engine

dictionary 1
dictionary 2
dictionary 3

resolve differences

dictionary 1
dictionary 2
dictionary 3

one dictionary

search engine
Ekilex data model

lexeme:
this word
in this meaning
as described in this dictionary

<table>
<thead>
<tr>
<th>Word</th>
<th>Lexeme</th>
<th>Meaning</th>
</tr>
</thead>
</table>

- spreadsheet: thing with rows and columns
- table: thing with four legs
Previously: many dictionaries
initial state: 2 words
                2 meanings
                2 dictionaries

word 1: cat
  - lexeme 1: Beware of the cat!
    - ex.: Our cat had kittens.
    - Every cat had a special task.
  - meaning 1: A small domesticated animal with soft hair, originated from a wild cat [DicEst]

word 2: cat
  - lexeme 2: Our cat had kittens.
    - The cat and the dog got along well.
  - meaning 2: A small hairy animal, widely kept as a pet [BED]
one dictionary

word 1

lexeme 1
ex.: Beware of the cat!
    Our cat had kittens.
    Every cat had a special task.

meaning 1
def.: A small domesticated animal with soft hair,
      originated from a wild cat [DicEst]

lexeme 2
ex.: Our cat had kittens.
    The cat and the dog got along well.

meaning 2
def.: A small hairy animal, widely kept as a pet [BED]

words combined:  1 word
                 2 meanings
                 2 dictionaries
words and meanings combined: 1 word 1 meaning 2 dictionaries
one dictionary

words and meanings combined:

1 word
1 meaning
2 dictionaries
one dictionary

words and meanings combined:
- 1 word
- 1 meaning
- 2 dictionaries
one dictionary

words and meanings combined:

1 meaning
2 dictionaries

why is the meaning duplicated?

and why does it have two nearly identical definitions?
all combined:  
1 word  
1 meaning  
1 dictionary
all combined: 1 word
1 meaning
1 dictionary

resolve these manually
initial situation
headwords aggregated
headwords aggregated
examples aggregated

GDEX:
- examples: suitable for learners all
- definitions:
  - long
  - short
  - simple

Explanatory:
- morphology:
  - basic
  - frequent
  - all

MAB:
- headwords:
  - frequent
  - suitable for A2 level
  - those that have collocations all

Backbone:
- dict 2
- dict 3
- dict 4
- dict 6
### Collocations

<table>
<thead>
<tr>
<th>equivalents</th>
<th>collocations:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>suitable for learners</td>
</tr>
<tr>
<td></td>
<td>frequent</td>
</tr>
<tr>
<td></td>
<td>salient</td>
</tr>
<tr>
<td></td>
<td>all</td>
</tr>
</tbody>
</table>

### GDEX

<table>
<thead>
<tr>
<th>equivalents</th>
<th>examples:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>suitable for learners</td>
</tr>
<tr>
<td></td>
<td>all</td>
</tr>
</tbody>
</table>

### Explanatory

<table>
<thead>
<tr>
<th>equivalents</th>
<th>definitions:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>long</td>
</tr>
<tr>
<td></td>
<td>short</td>
</tr>
<tr>
<td></td>
<td>simple</td>
</tr>
</tbody>
</table>

### MAB

<table>
<thead>
<tr>
<th>equivalents</th>
<th>morphology:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>basic</td>
</tr>
<tr>
<td></td>
<td>frequent</td>
</tr>
<tr>
<td></td>
<td>all</td>
</tr>
</tbody>
</table>

### Backbone

<table>
<thead>
<tr>
<th>equivalents</th>
<th>headwords:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>frequent</td>
</tr>
<tr>
<td></td>
<td>suitable for A2 level</td>
</tr>
<tr>
<td></td>
<td>those that have collocations</td>
</tr>
<tr>
<td></td>
<td>all</td>
</tr>
<tr>
<td>Category</td>
<td>Descriptions</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bilingual</td>
<td>frequent, good for reversing the dictionary, all</td>
</tr>
<tr>
<td>Collocations</td>
<td>suitable for learners, frequent, salient, all</td>
</tr>
<tr>
<td>GDEX</td>
<td>suitable for learners, all</td>
</tr>
<tr>
<td>Explanatory</td>
<td>long, short, simple</td>
</tr>
<tr>
<td>MAB</td>
<td>basic, frequent, all</td>
</tr>
<tr>
<td>Backbone</td>
<td>frequent, suitable for A2 level, those that have collocations, all</td>
</tr>
</tbody>
</table>

one dictionary
number of headwords?

60000 as in the popular dictionary
150000 as in the largest dictionary
1 million most frequent lemmas
6 million as in the current corpus
x million as in the next corpus
minimum amount of information?

corpus example and frequency

definition

equivalents

manually picked examples

translations of examples
minimum quality of information?

errors of tokenising, tagging etc
ungrammatical
politically incorrect
not approved by language planning
noise of distributional semantics
thank you

Kristina Koppel
Arvi Tavast
Margit Langemets
Jelena Kallas

This work has been partially supported by receiving funding from the European Regional Development Fund.