Unified Data Modelling for Presenting Lexical Data: the Case of EKILEX

Arvi Tavast, Margit Langemets, Jelena Kallas, Kristina Koppel
Institute of the Estonian Language, Tallinn
EURALEX 2018
Recurring problem: disconnected dictionaries
Disconnected dictionaries

- 140 dictionaries
- 3 separate DWSs
- different data models and formats
- inconsistencies
- duplication
EKILEX project: new DWS, import legacy data
EKILEX project

- Institute of the Estonian Language (EKI)
- Developer OÜ Tripledev
- End user products are separate projects
  - WordWeb, Dec 2018
  - TermWeb, applying for funding
EKILEX design choices in brief

• Merge dictionaries into one

• Centralise common data elements

• Relational database

• Graph structure with n:m between word and meaning
Merge dictionaries (instead of linking)

- information about language, not about dictionaries
- remove duplication of work
- resolve existing conflicts

How is this possible?
- as a single publisher, we can afford it
- before, during or after import
- tolerate duplicates while not yet merged
Centralise common data  (instead of linking or copying)

• morphology
• collocations
• etymology

• example sentences
• quantitative data
• etc

How is this possible?
• most data elements are independent of dictionary
• new process guidelines in the institute
  – do definitions, not the Explanatory Dictionary
• data elements may still be specialised for use cases
  – do definitions for learners, not the Learners' Dictionary
Relational database (instead of rdf or json)

• developers available
• tried and tested, robust
• standards are needed for data exchange, not storage

How is this possible?
• export to any existing or future standard format
  – rdf
  – json
  – NB: as needed
Graph structure

• n:m between word and meaning
• semasiology or onomasiology
• nothing is repeated

How is this possible?
• show as a tree if the user so prefers
  – lexicographers: senses of a word
  – terminologists: terms denoting a concept

(instead of tree)
n:m between word and meaning
n:m between word and meaning

• common word list
• common meaning list

The link table we call the **lexeme**:  
• this word in this meaning in this dataset  
• a dictionary is a mapping between words and meanings  
  – (while we still have dictionaries)
Expected challenges

• Data quality issues
  – duplicates, inconsistencies
• Merging words
  – 1300 homonyms in Estonian
  – 2 person/days per dictionary
• Merging meanings
  – much more labour-intensive
  – differences intentional
Unexpected challenges

- Differences between datasets even bigger than expected
- User resistance even bigger than expected
- Many differences intentional
- Unification tends to flatten valid distinctions
- Machine-readable data looks redundant to human readers
- Conflicts are explicit in the new system
  - not (yet) resolved
  - painfully evident to the user
Conclusions, lessons and plans

• More courage to depart from status quo
  – dictionaries
  – data formats and standards
  – data elements
• Import lexical data, not dictionaries
• Empirical quantification
  – from corpora
  – from users: crowdsourcing, experiments, games
• Export for ELEXIS etc as needed