Processing grammatical information in a dictionary management system

Ülle Viks, Andres Loopmann
Institute of the Estonian Language, Tallinn

www.eki.ee

- Background: Lexicographer's workbench
- Two tools for processing morphological information in dictionaries

Lexicographer's workbench

EELex

National Program for Estonian Language Technology
(2006-2010)

- Aim: to support the dictionary compiler
- EELex consists of
  - software for dictionary writing and management
  - lexical resources including dictionary databases

Lexicographer's workbench: software

The general features of EELex are:

- Unicode support
- XML databases
- XSD schemas
- XSL transformations for generating different views (XML view, Edit view, Layout view)
- click-to-edit
- structural queries and sorting of query results
- export to the MS Word layout format
- team work option (with different levels of user rights)

Lexicographer's workbench: software

- Tools for dictionary processing, e.g.
  - cross-reference checker
  - XML file generator
  - menu compiler
  - bulk corrections interface
  - morphological processing interface
  - layout design interface
  - schema design interface
- Public version of bilingual dictionaries for custom users (http://eksa.eki.ee)
- Estonian language support
  - language (morphological) software

Lexicographer's workbench: software components

- Estonian language morphology software, several DLLs, installed on the local machine
- Internet Explorer uses a digitally signed ActiveX control, which in turn uses morphology software DLLs
- ActiveX control & morphology software installation from web page

Lexicographer's workbench: software components

- IE ActiveX control main tasks:
  - Automatic keyboard change in input text boxes according to the input language (e.g. Estonian -> Russian etc.)
  - Spell check in input text boxes
  - Morphological analysis, synthesis, inflectional type recognition & other morphological functions

Lexicographer's workbench: lexical resources

About 20 dictionaries of different types:

- Bilingual
  - Estonian-Russian dictionary (Est-Ukrainian, Est-Finnish, etc.)
- Monolingual
  - Orthological dictionary
  - Explanatory dictionary
  - Dictionary of foreign words
  - Etymological dictionary
  - Database of word families
- Terminological databases (multilingual)
- Estonian language support
  - Estonian-X dictionary

EELex software and grammatical information

- Tools for processing grammatical information:
  - morphological interface for adding morphological data to word entries
  - bulk corrections interface

Morphological interface: Estonian morphology

- Complicated morphology (agglutinative and inflectional)
  - a great number of inflected forms
  - extensive variation of morphological units
- The interface uses a rule-based system of Estonian morphology
- Morphological data in dictionary entry:
  - basic forms of an inflectional word
  - inflectional type number
  - part of speech

Rule-based morphology of Estonian

Viks, Ülle (2000). Eesti keele avatud morfoloogiamudel. – In Tiit Hennoste (ed.). *Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1*. Tartu: Tartu Ülikooli Kirjastus, 9–36.

Viks, Ülle (2000). Tools for the Generation of Morphological Entries in Dictionaries. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds). *Second International Conference on Language Resources and Evaluation. Proceedings*. Athens: 383–388.

Morphological interface

- Generation of the morphological component for an entry: basic forms and other morphological data
- Generation of the whole paradigm

Morphological interface: problems of input

- Synthesis input = simple word in lemma form:
  - noun: Sg N (*hurmaa* 'persimmon')
  - verb: *ma*-infinitive (*klikkima* 'to click', *auhindama* '*award*')
- Entry word = simple word in lemma form, except
  - compound words: synthesis input = final component (*kõrghoone* 'high-rise building')
  - N Pl: synthesis input = Sg N (*krõpsud* 'crisps', *erarõivad* 'plain clothes')

- Analysis → Synthesis input
  - kõrghoone — kõrg+hoone → hoone
  - *krõpsud* *pl* → krõps
  - erarõivad *pl* era+rõivad → rõivas

Morphological interface: settings

- Recommended:
  - with compound word recognition
  - with dictionary
- Free:
  - word form selection for the entry
  - marking quantity degree and stress (*tõuge, k'auge*)

Morphological interface: dialog

- Necessary for guiding the compilation of morphological entries, enabling the user:
  - to pick the right entry (from homonymous ones) (*vaht* 'foam; guard', *alt* 'from below; alto‘)
  - to delete the overgenerated word forms (*truudus* 'loyalty' – sg; *krõpsud* ‘chips’ – pl)
  - to correct the possible errors

Bulk corrections interface

- In general: correction of a single entry
- Bulk corrections: same correction in many entries throughout the dictionary

Bulk corrections interface: cases of need

- Editing of grammatical data:
  - part-of-speech labelling
  - usage labelling
  - segmentation of compound and derivative words
  - marking of stress, quantity, palatalization
  - etc.
- Postediting after automatic morphological entry generation:
  - homonymous entries
  - overgenerated forms
  - errors

Conclusion

- We hope that our tools for processing grammatical information will facilitate lexicographers' work and enable enhancement of dictionary quality.