Inga Znotiņa, inga.s.znotina@gmail.com, Latvia

Learner corpus of the second Baltic language: annotation and data comparability

Göteborg

06.12.2017.

Contents

The corpus «Esam»

- Not a learner corpus of Baltic languages in general
- Second Baltic language:
  - for Lithuanians – Latvian
  - for Latvians – Lithuanian
  - taught in some higher education institutions in Latvia and Lithuania, mostly as a part of philology studies
- Bi-directional

Description of the corpus

- Beginner learner corpus
  - 1st and 2nd semester of language studies at university (adults?)
  - roughly A level

- Only written texts
  - various topics (title added as metadata)
  - by hand or computer, 2007-2014
  - original texts rather than translations (or are they?)

Description of the corpus

- Full texts
- No specialization
- Small, unbalanced: ~52 000 tokens
  - ~45 000 Latvian → Lithuanian
  - ~7 000 Lithuanian → Latvian

- Publicly accessible, anonymized; no registration needed
- Runs on *TEITOK*, hosted on a private server

Annotation of the corpus

- Lemmatization
- POS annotation – standard tags

- Both done at the same time, using *TEITOK* interface
- Annotated manually
- Some things might be done differently in other corpora (e.g. diminutives as separate lemmas or not)

Annotation of the corpus

- Syntactic annotation:
  - Only sentence types
  - Done manually in the XML files
  - Classification (see next slide)
    - not used (?) in other corpora
    - not very different
    - language specific ?

Annotation of the corpus

- Utterances by model of sentence; quasi-sentences separately
  - Utterances
    - Simple sentence
      - Unextended
      - Extended
    - Complex sentence
    - Compound sentence
    - Mixed-compound sentence
  - Quasi-sentences
  - Sentence of unclear structure

Annotation of the corpus

- Error annotation
  - Texts are corrected (target hypothesis)
  - Each token where the target hypothesis differs from the original is given an error tag
  - Error taxonomy – (heavily) adapted S. Granger's taxonomy for French (2003)
    - Language specific
    - Level specific

Annotation of the corpus

- Error annotation issues:
  - sometimes it is difficult to guess what the (beginner) author meant
  - errors are not always token-level
    - various syntactic errors
    - more than one error in a token
  - an error can have more than one cause
    - mans mājas (my (*m.sg.*) home (*f.pl.*): gender or number?)
  - errors create new errors
    - Viņas uzbūve (structure, *f.*) → augums (complexion, *m.*) ir smalka (*f.*→*m.*), spēcīga (*f.*→*m.*)

Thoughts on comparability

- Inner comparability
  - Bi-directional data → comparing the directions
  - Balancing issues (6x more Lithuanian texts)

- The concept of Baltic interlanguage
  - Do LT-LV and LV-LT interlanguages share common traits as opposed to XX-LT and XX-LV interlanguages?
  - Specifics of learning a language that is closely related to one's own
  - Only comparison can show

Thoughts on comparability

- Technologically – yes (?)
- Data – yes, especially:
  - beginner data of other learners of Baltic languages
  - more advanced learners of the second Baltic language
  - not native speakers!

Thoughts on comparability

- Annotation:
  - Lemmatization – yes, now
  - POS annotation – yes, now
  - Syntactic annotation – yes, probably
  - Error annotation – not before resolving the issues

- *Concept → data* or *data → concept*
  - Fragmentation in the field

Thoughts on comparability

- Public access:
  - great accessibility, wider range of corpus users
  - various legal issues: copyright, personal data protection…
  - less data (?)
- Technologies, skills required
  - creating the corpus
  - using the corpus
- Communication

Paldies! Ačiū! Tack! Thank you!

inga.s.znotina@gmail.com

www.esamkorpuss.lv