Learner corpus *Esam*:
A new corpus for researching Baltic interlanguage

Inga Znotiņa
Liepāja University • Ventspils University College • Rīga Stradiņš University
Inga.S.Znotina@gmail.com

The corpus is being built as a part of the author's ongoing PhD research.

Baltic languages: Latvian and Lithuanian

Baltic interlanguage: the interlanguage that forms when a person with the background of one Baltic language learns the second Baltic language

Design of the corpus

Name of the corpus: *esam* means 'we are' in Latvian as well as in colloquial Lithuanian

Materials included in the corpus:
- Written texts
- By university students who are learning the second Baltic language
- Written independently as a homework
- Allowed to use any materials
- 2007–2014 (but newer materials will be added later)
- On various topics:
  «My family and friends»
  «The place where I would like to return»
  «A strange day in my life» etc.
- Length of the texts: 45–500 words

Data collection

- **Universities in Latvia and Lithuania**
- **Texts written for learning purposes**
- **Collected without a specific goal**

- Later given to the creator of the corpus
- Authors tracked down, asked for permission
- A standardized permit signed by each author who agrees
  - Texts can be made publicly accessible
  - Identity of the author not revealed (except author list, if agreed)
  - Other terms
- Texts anonymized
- Texts included in the corpus

Markup and annotation

Markup:
So far only anonymization
- Information that can reveal author's identity is:
  - Removed – tag <izlaid>
  - Replaced with grammatically similar – tags <anon></anon>

Other metadata collected that could later be turned into markup

Annotation:
- So far none
- In the future:
  - Error annotation
  - Part of speech annotation
  - Syntactic sentence type annotation

The sample corpus

- **Publicly available since 14th June 2015**
- **On esamcorpus.wordpress.com**

- About 15 000 words, 68 texts
- Anonymized
- Non-annotated
- Only texts in Lithuanian, written by Latvian students
- A metadata file included:
  - Code of the text
  - Code of author
  - Topic
  - Number of words
  - Language of the text
  - Semester
  - Language of instruction

Access to the corpus

**.txt* files available for downloading

- Stored on a free file exchange server *files.fm* (link on corpus's website)
- Each file can be downloaded separately or in a **.zip* archive
- **NO software provided!**
- Markup / annotation in *XML* tags
- Possibly – other versions in future
- Encoding – *UTF-8*
- Tested to work with Anthony Lawrence's *AntConc*
- **Terms and conditions** (based on permits signed by authors):
  - Using allowed for education and research
  - Can be cited in educational and/or research papers
  - Corpus use for commercial aims is forbidden
  - Can be researched with any software of the researcher's choice
  - The creator of the corpus is not responsible for texts' content
  - The files are only accessible from the link given on this website

Future plans

Already collected (with permits)
- More than 40 000 words
- More than 200 texts
- Only in Lithuanian

Working on permits for other already collected texts in Latvian as well as Lithuanian

Possibly – longer texts too

Teachers are still collecting texts from current students

So far – only:
- Lithuanian taught in Latvia
- Latvian taught in Lithuania

In the future maybe – texts written by learners in target language setting?

Special thanks to:
- The authors for allowing to use their texts in research
- Teachers of second Baltic language for helping to gather materials for the corpus

Find it here:

esamcorpus.wordpress.com