LaVA – Latvian Language Learner corpus

Roberts Dargiš1, Ilze Aužina1, Inga Kaija1,2, Kristīne Levāne-Petrova1, Kristīne Pokratniece1
Institute of Mathematics and Computer Science, University of Latvia1, Rīga Stradiņš University2

Overview
This paper presents the Latvian Language Learner Corpus (LaVA) developed at the Institute of Mathematics and Computer Science, University of Latvia. LaVA corpus contains 1015 essays (190k tokens and 790k characters excluding whitespaces) from foreigners studying at Latvian higher education institutions and who are learning Latvian as a foreign language in the first or second semester, reaching the A1 (possibly A2) Latvian language proficiency level. The corpus has morphological and error annotations. The corpus is publicly available at: http://www.korpuss.lv/id/LaVA

Agreement and questionnaire form

Text digitalization and correction

Agreement and questionnaire form

Published formats and interfaces

The corpus is published in the corpus homepage for easy browsing. The homepage also provides concordancer for simple queries. More advanced queries can be constructed in the noSketchEngine instance. The files necessary for a researcher to upload the LaVA corpus in the SketchEngine are available in the Download section of the corpus homepage.

Data source and data characteristics

Texts are written by hand on the other side of the form.

Authors of the texts:
- Higher education students from 5 universities of Latvia
- Living in Latvia for a relatively short time
- Learning Latvian at the beginner level for the 1st or 2nd semester

Topics of the texts:
- Teachers choose the desired topic based on pedagogical needs
- My friends, My family, My day, My studies, etc.

Length of the texts:
- Preferred text length – at least 100 words

Gender of the authors of the texts: women – 63%, male – 37%

Age of text authors:
- 88% of authors are between the ages of 17 and 25
- 12% are between the ages of 26 and 46

Mother tongue:
- Language learners have indicated 35 different mother tongues
- The most common are German (37%), Swedish (11%) and Finnish (9%)

Error Analysis

Spelling, inflectional and word formation, punctuation, and lexical errors are marked automatically when the original and corrected texts are typed, and then manually checked. Syntax and combined errors are marked manually.

Error Type | Relative frequency
--- | ---
Inflectional & word formation | 40%
Spelling | 33%
Lexical | 19%
Punctuation | 6%
Syntactic | 1%
Complex | 1%

Table 1. Error frequency by error type relative to tokens with errors

| Error Type | Relative frequency
--- | ---
Diacritic marks | 78.1%
Capital letters | 7.8%
Missing letters | 4.1%
Redundant letters | 3.6%
Complex | 6.4%

Table 2. Frequency of spelling errors by error type