LaVA – Latvian Language Learner corpus

Roberts Darģis[1], Ilze Auziņa[1], Inga Kaija[1,2], Kristīne Levāne-Petrova[1], Kristīne Pokratniece[1]

Institute of Mathematics and Computer Science, University of Latvia[1],
Rīga Stradiņš University[2]

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

Corpus Creation Pipeline

Data digitalization
↓
Text correction
↓
Morphological annotation
↓
Automated error annotation

Agreement and questionnaire form

Information letter of the project researcher group for Latvian learners

Dear student,
The project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The goal of the project is to create an error-annotated Latvian language learner corpus and develop corpus-based teaching materials.
The project is financed by Latvian Council of Science; the project leader is senior researcher of IMCS UL Dr. philol. Ilze Auziņa (e-mail: ilze.auzina@lumii.lv).

What do you have to do?
Please read carefully and sign the Permission that you agree to allow the text written during your Latvian language studies to be included in the Latvian learner corpus.
Complete the questionnaire and provide the necessary information for the further use of the text in research. On the other side of the page, write an essay on the topic that the lecturer has assigned to you.

Data storage and privacy
Collected data will be stored at the IMCS UL on the password protected server. The data stored will be completely anonymous. A unique identifier will be assigned to each data provider.
After the end of the project *the Learner Corpus of Latvian* will be publicly available on the corpora website of IMCS UL.

Participation
Participation is voluntary. Over the course of the project, you may request that texts written by you are removed from the database and refuse to participate without specifying the reason. This should be done by informing the group of researchers. In case of refusal, all materials collected will be deleted.

On behalf of the project team of researchers,
Ilze Auziņa, IMCS UL senior researcher

FLPP FUNDAMENTAL AND APPLIED RESEARCH PROJECTS
Institute of Mathematics and Computer Science University of Latvia

Information letter

PERMISSION

I agree that this text, written in 2020, can be included in the *Learner Corpus of Latvian* and, as a part of the corpus, can be made publicly available in various forms, fully or partly, with such conditions:

- I agree that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.
- I confirm that none of the data in this text can lead to identification of any existing people.
- I agree that the text is anonymous and my name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which ma… possible to recognize several texts written by the same author but does not reveal the … author.
- The data included in the corpus can be cited in the educational ma… papers, and other work in various forms.
- The corpus and all materials included in it can be publicl… unlimited period and can be viewed and researched unlimited amount of …
- All texts included in the corpus can have … ation added to them (e.g. error corrections, part-of-speech annotation, etc.).
- I will have the right to withdraw my consent a… time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. I am aware of this opportunity as a data provider.

Permission

INFORMATION ABOUT THE AUTHOR

Age: _____
Gender: _____
Mother tongue (-s): _____
Other languages you speak: _____
How long have you been living in Latvia? _____
For how many semesters have you been learning Latvian langu…
□ This is the first semester.
□ This is the second semester.
□ Other (please specify): _____

Questionnaire

___/___/_____ Date _____ Signature _____ Name, surname

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.

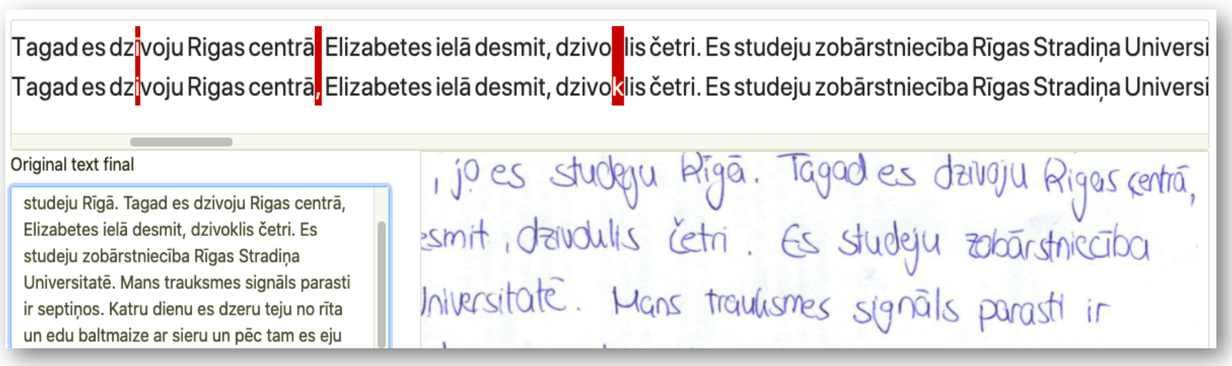
Length of the texts:
- Preferred text length – at least 100 words

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

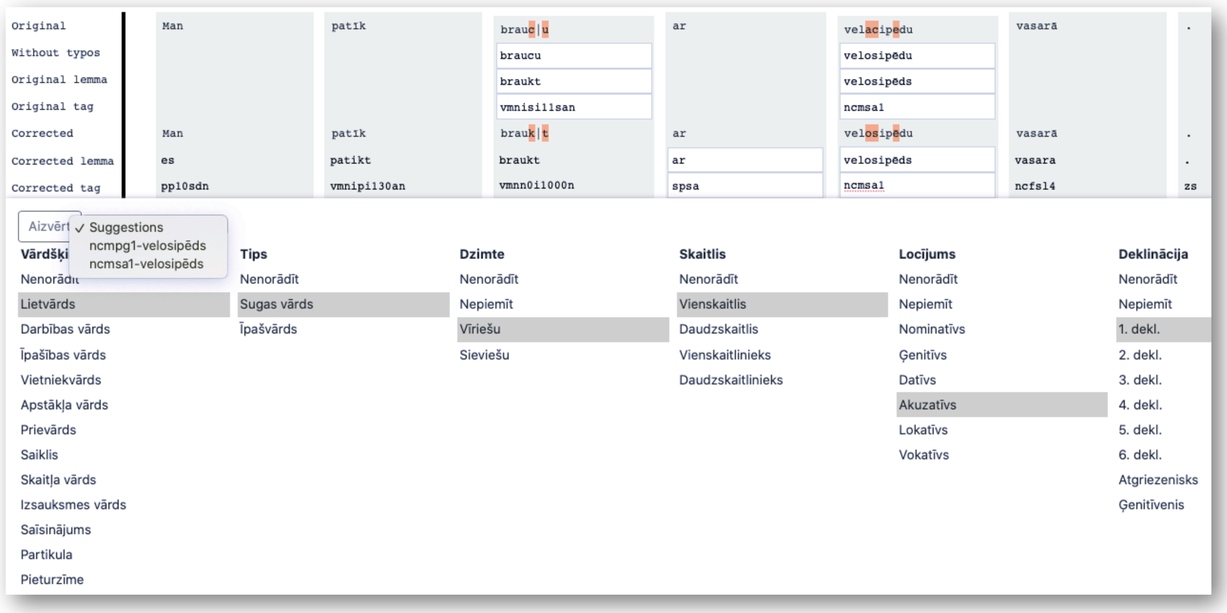
Age of text authors:
- 88% of authors are between the ages of 17 and 25
- 12% are between the ages of 26 and 46

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

Text digitalization and correction



Morphological and error annotation



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.

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

http://lava.korpuss.lv

FLPP FUNDAMENTAL AND APPLIED RESEARCH PROJECTS
SRP State Research Programme

The work reported in this paper is a part of the project *Development of Learner Corpus of Latvian: methods, tools and applications* (No. lzp-2018/1-0527), financed by Latvian Council of Science, and the Latvian State Research Programme project *Research on Modern Latvian Language and Development of Language Technology* (No. VPP-LETONIKA-2021/1-0006). This work has also received support from the Latvian Language Agency under grant agreement No. 4.6/2019-029