CLARIN
Common Language Resources and Technology Infrastructure

FLPP
FUNDAMENTAL AND APPLIED RESEARCH PROJECTS

# Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection

**Inga Kaija**[1,2], Ilze Auziņa[1]

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

[2] Riga Stradiņš University, Latvia

CLARIN annual conference 2019

Leipzig, Germany
30.09.2019.–02.10.2019.

# Overview

- Information about the project
- The agreement / metadata collection form
- Data collection procedure

# About the Project

FLPP
FUNDAMENTAL AND APPLIED RESEARCH PROJECTS

Latvian Council of Science Grant *Development of Learner corpus of Latvian: methods, tools and applications* (No. lzp-2018/1-0527)

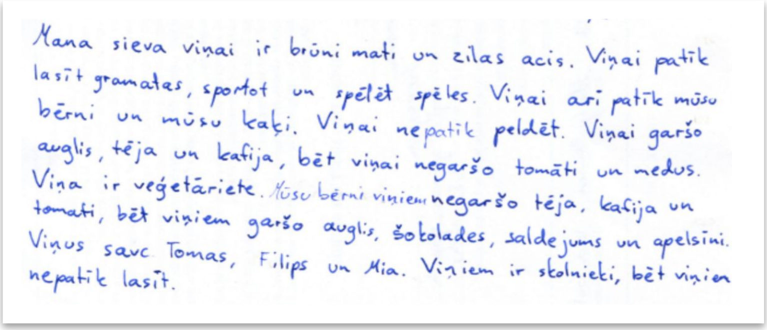
Duration of the projects: 2018–2021

The project has several interrelated goals:

- creation of infrastructure for corpus collection and development of data collection and annotation methodology (both error annotation and morphological annotation);
- development of an error-tagged *Learner Corpus of Latvian* (LaVA);
- development of corpus-based learning materials and self-assessment web platform.

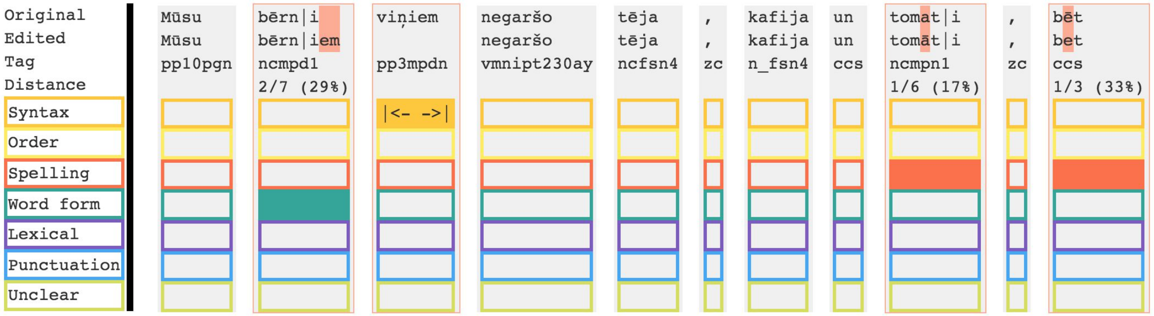
# Stages of LaVA corpus development

- Methodology and tools for data collection and annotation process (including agreement / questionnaire form)
- Integrated multifunctional platform for data uploading, annotating and search
- Data collection and digitization
- Correction of the learners’ texts (Creation of the target hypothesis)
- Morphological (plus lemma and POS) and error annotation



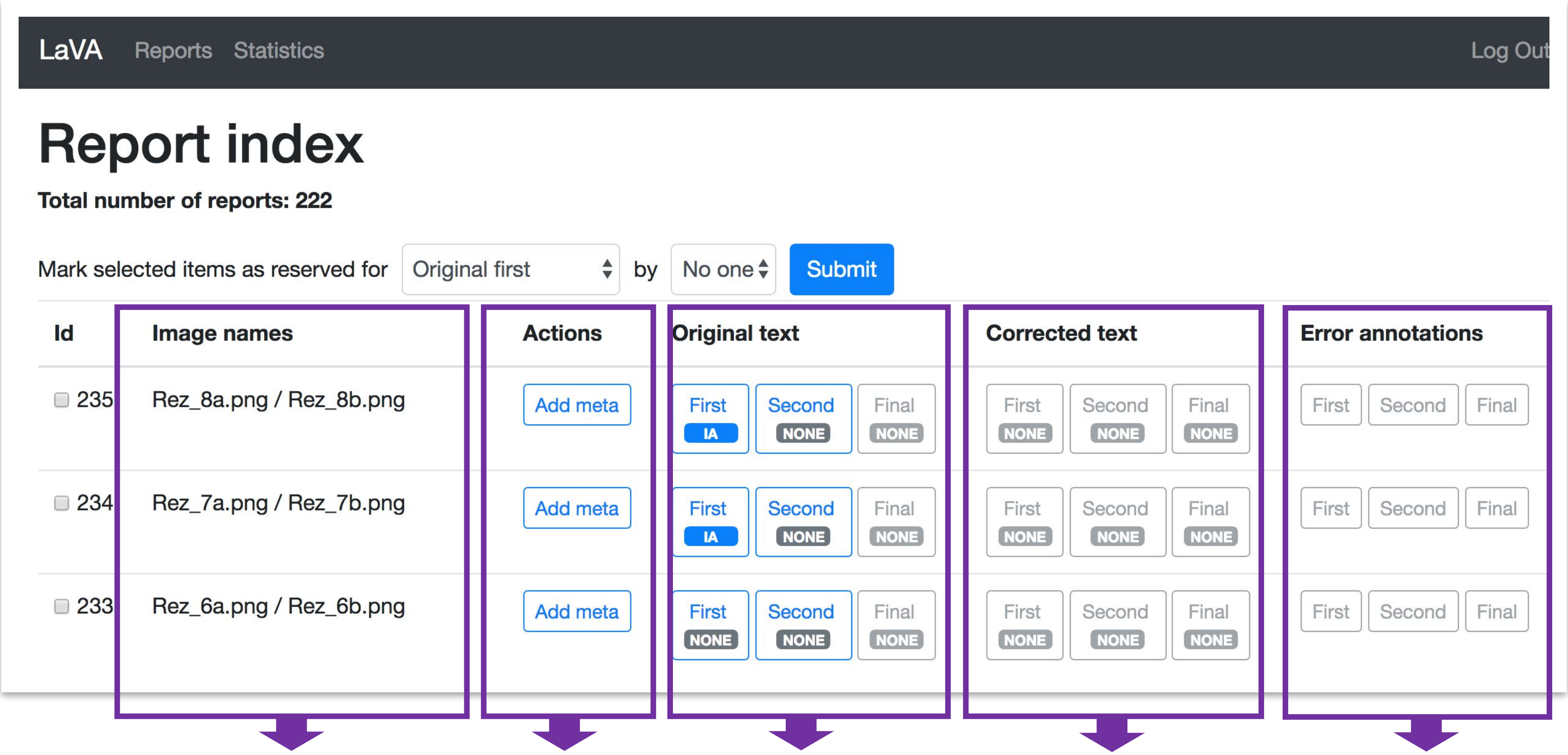
Mana sieva viņai ir brūni mati un zilas acis. Viņai patīk
lasīt gramatas, sportot un spēlēt spēles. Viņai ari patīk mūsu
bērni un mūsu kaķi. Viņai nepatīk peldēt. Viņai garšo
auglis, tēja un kafija, bēt viņai negaršo tomāti un medus.
Viņa ir veģetāriete. Mūsu bērni viņiem negaršo tēja, kafija un
tomati, bēt viņiem garšo auglis, šokolades, saldejums un apelsīni.
Viņus sauc Tomas, Filips un Mia. Viņiem ir skolnieki, bēt viņien
nepatīk lasīt.

***Manai sievai*** *ir brūni mati un zilas acis. Viņai patīk lasīt* ***grāmatas****, sportot un spēlēt spēles. Viņai arī patīk mūsu bērni un mūsu kaķi. Viņai nepatīk peldēt. Viņai garšo* ***augļi,*** *tēja un kafija,* ***bet*** *viņai negaršo tomāti un medus. Viņa ir veģetāriete. Mūsu* ***bērniem*** *negaršo tēja, kafija un* ***tomāti, bet*** *viņiem garšo* ***augļi, šokolāde****, saldējums un apelsīni. Viņus sauc Tomass, Filips un Mia.* ***Viņi*** *ir skolnieki,* ***bet*** *viņiem nepatīk lasīt.*



| Original | Mūsu | bērn\|i | viņiem | negaršo | tēja | , | kafija | un | tomat\|i | , | bēt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Edited | Mūsu | bērn\|iem | | negaršo | tēja | , | kafija | un | tomāt\|i | , | bet |
| Tag | pp10pgn | ncmpd1 | pp3mpdn | vmnipt230ay | ncfsn4 | zc | n_fsn4 | ccs | ncmpn1 | zc | ccs |
| Distance | | 2/7 (29%) | | | | | | | 1/6 (17%) | | 1/3 (33%) |
| Syntax | | | \|<- ->\| | | | | | | | | |
| Order | | | | | | | | | | | |
| Spelling | | | | | | | | | 🟥 | | 🟥 |
| Word form | | 🟩 | | | | | | | | | |
| Lexical | | | | | | | | | | | |
| Punctuation | | | | | | | | | | | |
| Unclear | | | | | | | | | | | |

# Integrated platform for data uploading, markup, annotating and search



## Report index

**Total number of reports: 222**

Mark selected items as reserved for Original first by No one Submit

| Id | Image names | Actions | Original text | | | Corrected text | | | Error annotations | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| □ 235 | Rez_8a.png / Rez_8b.png | Add meta | First IA | Second NONE | Final NONE | First NONE | Second NONE | Final NONE | First | Second | Final |
| □ 234 | Rez_7a.png / Rez_7b.png | Add meta | First IA | Second NONE | Final NONE | First NONE | Second NONE | Final NONE | First | Second | Final |
| □ 233 | Rez_6a.png / Rez_6b.png | Add meta | First NONE | Second NONE | Final NONE | First NONE | Second NONE | Final NONE | First | Second | Final |

Scanned texts, agreements

Metadata

Digitization of the texts

Correction of the texts

Error annotation

# Main legal documents regarding copyright issues in Latvia

**The Copyright Law**

- the main document regulating copyright protection in Republic of Latvia
- states that:
  - texts written as a part of study process are protected by copyright, unless otherwise stated in the study agreement between the author and the study institution;
  - in order to make the text (or part of it) available to the public, a written permission must be received from the author;
  - the author has the right to decide to be recognized as an author and to decide when, how many times etc. the work can be accessed.

# Main legal documents regarding personal data protection in Latvia (1/2)

**The Personal Data Processing Law**

- the main document regulating protection of personal data in Republic of Latvia
- GDPR compliant (regarding our needs)
- regulates who and how manages personal data processing (in institutions) and complaints of violations (in the country)
- regulates the cases in which personal data can be processed without data subject's agreement (not ours)

# Main legal documents regarding personal data protection in Latvia (2/2)

**GDPR**

- states that:
  - it is personal data if we can identify the person
  - data subject is aware who processes data and why, knows who receives it further
  - data subject can revoke permission
  - it is legal to store the data necessary for agreements
  - information provided to the data subject must be clear

# The layout of the agreement / questionnaire form

The form is printed on one side of an A4 size paper sheet and includes **three parts**:

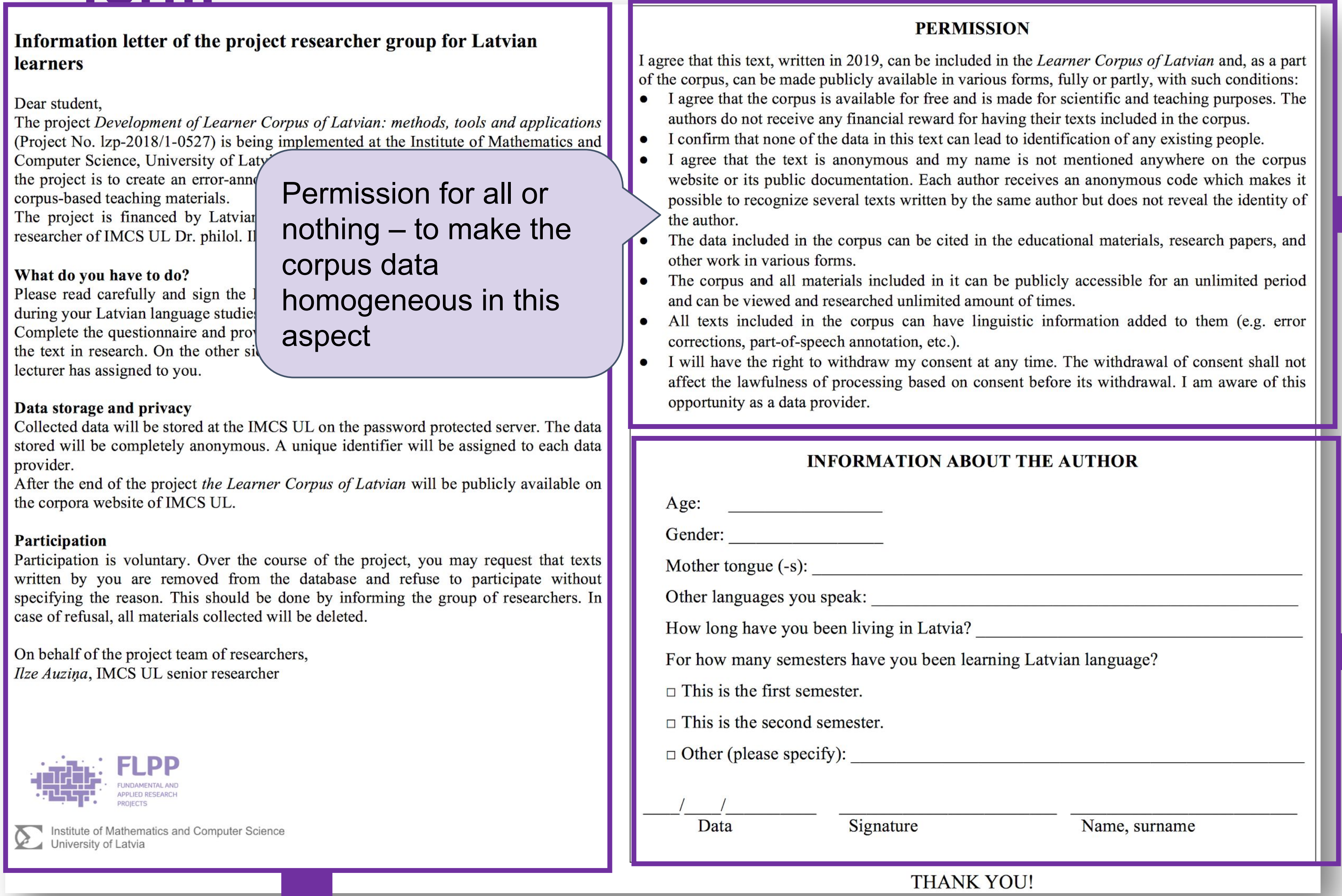
- an information letter,
- a permission,
- a metadata collection questionnaire (information about the author).

At the end of the form there is a space for

- date,
- signature,
- name and surname of the author.

Making the document legally binding; not disclosed to anyone outside the project

# The layout of the agreement / 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 Latv
the project is to create an error-ann
corpus-based teaching materials.
The project is financed by Latvian
researcher of IMCS UL Dr. philol. Il

**What do you have to do?**
Please read carefully and sign the l
during your Latvian language studie
Complete the questionnaire and pro
the text in research. On the other si
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

**PERMISSION**

I agree that this text, written in 2019, 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 makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched unlimited amount of times.
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).
- I will have the right to withdraw my consent at any 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.

**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 language?

□ This is the first semester.

□ This is the second semester.

□ Other (please specify): _________________________________________________

____/____/__________ Data

__________________________ Signature

__________________________ Name, surname

THANK YOU!

Permission for all or nothing – to make the corpus data homogeneous in this aspect

Permission

Questionnaire

Information letter

# Information letter

The letter of the project researcher group for learners consists of:

- basic information about the project, the institutions that are carrying it out, and contact information; **[PERSONAL DATA, COPYRIGHT]**
- brief task instructions for learner; **[CORPUS DEVELOPMENT]**
- information about the security of data on the server used for the corpus and privacy; **[PERSONAL DATA]**
- explanation on expressing one's will regarding participation in the project (i.e. what to do if the author decides they no longer want their texts to be a part of the corpus). **[PERSONAL DATA, COPYRIGHT]**

# Permission (1/2)

The permission includes seven statements the author agrees to comply with by signing the form (1-3):

- The author agrees that the corpus is available **for free** and is made for scientific and teaching purposes; and they do not receive any financial reward for having their texts included in the corpus. **[FUNDING]**
- The author confirms that none of the data in this text can lead to identification of any existing people. **[PERSONAL DATA; NO ADDITIONAL DATA SUBJECTS]**
- The author agrees that the text is anonymous and their name is not mentioned anywhere on the corpus website or its public documentation. **[COPYRIGHT / PERSONAL DATA]** *Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.* **[POSSIBLE PERSONAL DATA ISSUE?]**

# Permission (2/2)

The permission includes seven statements the author agrees to comply with by signing the form (4-7):

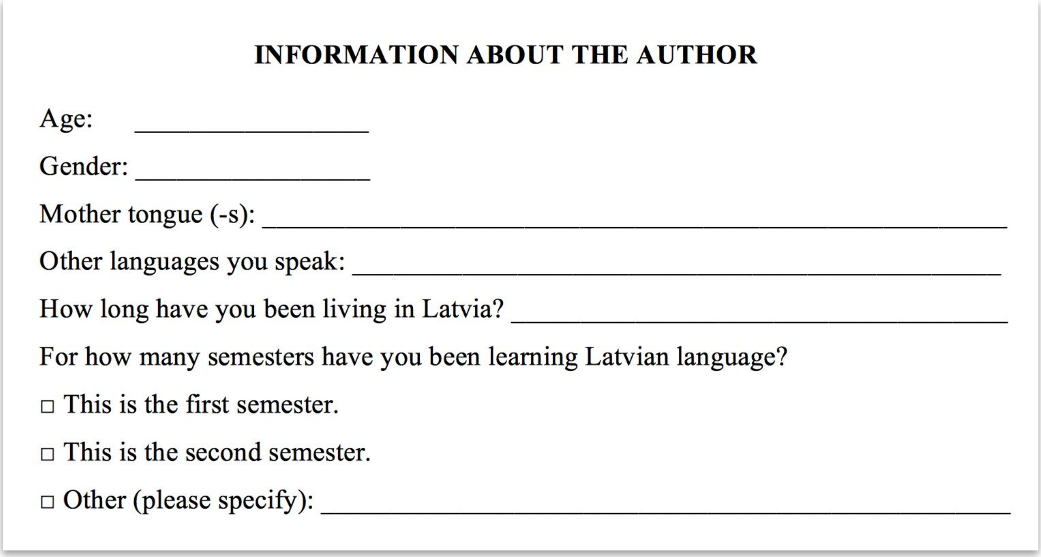
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms. **[COPYRIGHT]**
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched an unlimited amount of times. **[COPYRIGHT. METADATA =/= PERSONAL DATA]**
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.). **[COPYRIGHT]**
- The author will have the right to withdraw their consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. The author is aware of this opportunity as a data provider. **[REVOKING PERMISSION]**

# Metadata collection questionnaire

- Age,
- Gender,
- Mother tongue(-s),
- Other spoken languages,
- The length of residence in Latvia,
- The number of semesters studying Latvian language in a higher education institution.

**Shown in corpus!**

Not personal data, unless a unique combination



**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 language?

□ This is the first semester.

□ This is the second semester.

□ Other (please specify): ______________________________________________

# **Data collection procedure** (1/2)

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

**Authors** of texts

- Higher education students
- Living in Latvia for a relatively short time (mostly – 1 year or less)
- Learning Latvian at the beginner level for the first or second semester

**Topics** of texts

- Teachers choose the desired topic based on pedagogical needs
- For example: *My friends, My family, My day*

**Length** of texts

- Teachers choose the length of the text
- Preferred text length – at least 100 words

# Data collection procedure (2/2)

**Instructions for learners**

- The teachers instruct the students about the copyright and personal data protection system used in the project.
- If the topic contradicts this idea (e.g. “My friends and my family”), students are instructed to write about imaginary people or replace the real information with false one.
- Study materials can be used when writing.
- If the writing task is mandatory, students not wishing to participate are allowed to not fill the form.

# Conclusions

- The permission / metadata collection form is relatively simple and seeks to minimise unnecessary personal data processing.
- If any text is suspected to include any real personal data, the author is contacted once more by the teacher / data collector.
- The form can be used as a basis for agreements in data collection for other learner corpora in countries which have similar personal data and copyright protection regulations.

CLARIN
Common Language Resources and
Technology Infrastructure

FLPP
FUNDAMENTAL AND
APPLIED RESEARCH
PROJECTS

# Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection

**Inga Kaija**[1,2], Ilze Auziņa[1]

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

[2] Riga Stradiņš University, Latvia

CLARIN annual conference 2019

Leipzig, Germany
30.09.2019.–02.10.2019.