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

Inga Kaija
Institute of Mathematics and Computer Science,
University of Latvia;
Riga Stradiņš University, Latvia
inga.kaija@rsu.lv

Ilze Auziņa
Institute of Mathematics and Computer Science,
University of Latvia,
Riga, Latvia
ilze.auzina@lumii.lv

Abstract

Copyright and personal data protection are two of the most important legal aspects of collecting data for a learner corpus. The paper explains the challenges in data collection for the learner corpus of Latvian "LaVA" and describes the procedure undertaken to ensure protection of the texts' authors' rights. An agreement / metadata questionnaire form was created to inform the authors of the ways their texts are used and to receive the authors' permission to use them in the stated way. The information, permission, and the metadata questionnaire are printed on one side of an A4 size paper sheet, and the author is supposed to write the text on the other side by hand, thus eliminating the need to identify the author of the text separately. After scanning and adding to the corpus, the text originals are returned to the authors.

1 Introduction

Learner corpora have become increasingly popular, and the demand for such corpora to become available to a wider scope of researchers is growing. However, the creation of publicly available learner corpora includes dealing with personal data protection and copyright issues. A learner corpus of Latvian "LaVA" (Latvian Council of Science Grant Development of Learner corpus of Latvian: methods, tools and applications. No. lzp-2018/1-0527) is being created, and it will be publicly accessible, so these legal issues have to be addressed while still enabling researchers to collect relevant metadata about possible factors impacting language learning outcomes.

The "LaVA" creation is divided into several stages: 1) data collection, 2) data digitization; 3) text correction; 4) automated NLP analysis (morphological analysis); 5) original and corrected text alignment; 6) automatic error annotation and manual review. At least 1000 essays on different topics from students with different language backgrounds are planned to be included in the LaVA corpus.

The initial stage of the project covers development of a methodology for data collection and digitization, development of methodology and guidelines for error annotation, and corpus platform development. Among the most important tasks of this phase were the legal and ethical solutions for the text collection process.

Copyright and personal data protection are two of the most important legal aspects that should be resolved before data collection for the learner corpus is started. Therefore, an agreement and metadata questionnaire form was developed to inform the authors of the inclusion of their works in the corpus and to obtain authors' permission.

There have been efforts to create templates for contracts to help deal with the copyright issues when collecting data for research.[1] While they can be extremely helpful, in the case of creating a learner corpus a more specific compact document is useful where the exact aims and rules of using the texts are

[1] For example, see http://www.meta-net.eu/meta-share/licenses

Inga Kaija and Ilze Auziņa 2020. Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 41–47.

described. The copyright issues might be similar over all kinds of corpora, but the very nature of learner texts makes also anonymity particularly important – not only that of the people mentioned in the texts, but also that of the authors. In many cases, the learners feel self-conscious about their language skills and want their identity to be protected, especially knowing that the data will be available to the public. This, in turn, makes it necessary to specifically agree on the kinds of data the learners provide and the ways they are used.

To protect learners' rights when collecting their texts, an agreement / metadata questionnaire and the procedure of text collection was developed. The present paper lists the main legal and ethical principles considered and describes how the data collection process is carried out.

2 Regulations

The learners, i.e., authors of the texts collected for the corpus, come from various backgrounds and belong to various countries in Europe and outside of it. However, their studies of Latvian (including text writing process) and corpus creation take place in Latvia, so the legislature of Republic of Latvia applies. The regulations regarding personal data protection and copyright issues that concern learner corpus creation in Latvia have been previously described in comparison with the relevant regulations in Lithuania (Znotiņa, 2016). We further list the main legal documents and principles to be observed in each of those areas.

2.1 Copyright

The main document regulating copyright protection in Republic of Latvia is the Copyright Law (AL, 2000), and it 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.

In order to comply with the regulations, the corpus creators have to make it possible for the authors to express their decision explicitly. However, providing the authors with various choices would make the corpus creation process extremely complicated, as all of the different choices would have to be taken into account, especially considering that each of the submitted texts is added a separate consent. Therefore, it was decided that a standardized form for all authors of the texts in corpus must be created. Those authors who would not agree with the common terms could opt out of participating in the project altogether.

2.2 Personal data protection

Protection of personal data in Republic of Latvia is regulated by the Personal Data Processing Law (FPDAL, 2018) as well as one of the most influential regulations regarding personal data protection in European Union, the European Union's new General Data Protection Regulation (Regulation EU 2016/6791), enforced on May 25 2018 (GDPR 2016). Both of them emphasize the ability to identify a person as a criterion for defining personal data. GDPR states that personal data "means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (GDPR 2016).

The corpus "LaVA" is a beginner learner corpus, and the topics offered for writing to the beginner students often inherently include telling one's own or other people's data (e. g. "Me and my family"). Thus, it is of utmost importance to eliminate the possibility that such personal data would be made publicly available. It can be done by anonymizing the stored data containing personal information, or by avoiding the disclosure of any personal data. In this case, the topics of the texts often require a significant amount of personal data to be included, and anonymization could quickly become a daunting

task. The corpus creators decided to avoid it by requesting the authors of the texts to not include any real personal data and to replace it with imaginary ones instead. This request also has to be included in the form signed by the authors as a part of the conditions for including the texts into the corpus.

Some concerns have been expressed that handwriting can also be seen as personal data as it can be used to recognize the writer. Latvian legislation does not seem to address this issue specifically but it has to be considered when a corpus includes scanned copies of a handwritten text. Here are a few important aspects at play:

- There are many authors of the texts which leads to many examples of similar handwritings. The corpus is expected to contain at least 1000 texts, and each student is offered to submit a text no more than once per semester, for two semesters at most. Therefore, even if all students participated in the project twice (which is not the case), there would still be at least 500 different authors altogether. It makes recognizing someone by handwriting alone highly unlikely.
- The only real information provided about the author is the metadata: gender, other languages spoken, and age at the time of writing (time of writing is between 2018 and 2020, but it is not specified more precisely for any of the texts). In case unusual combinations are found, the corpus creators discuss not including the text into the corpus because unique metadata (such as a rare mother tongue) may give little quantifiable insight into the language learning process in general. This minimizes the possibility of recognizing a person by their handwriting and metadata combination,

During the corpus creation process, the text is seen by the author, the teacher, and no less than three people of the corpus creator team. If any of those people express doubts about the possibility to recognize the author by their handwriting (e. g. the handwriting looks unusual, distinguishable from most), the possibility to not include the text is considered. The amount of texts provided by the participating higher education institutions is large enough that it does not add any pressure on the corpus creators to try including as many texts as possible at the expense of authors' rights protection.

3 Data collection for the learner corpus of Latvian

The main principles of the agreement / questionnaire form are the same ones already used in the learner corpus of the second Baltic language "Esam"[2] (Znotiņa, 2018), but data collection is carried out in a different way. In "Esam", the permissions to use the data were acquired long after the texts were written (in some cases, several years), and all texts were additionally anonymized. In the case of "LaVA", the learners know the texts are going to be included in the corpus when they write them. Besides, the texts in "LaVA" are not further anonymized by the project team, and the data is collected by various people, so the procedure is regulated more strictly in order to maintain uniformity in the received data and information given to the authors.

3.1 Contents of the agreement / questionnaire

An agreement / questionnaire form was created for data collection of the corpus "LaVA". It is written in English because English is used as an intermediary language in studies of Latvian as a foreign language in the higher education institutions of Latvia, so all authors speak this language well. The form is offered to all authors of the texts expected to be included into the corpus, and every text is only included into the corpus after a signed copy of the form is received from the author. The texts are collected from the learners of Latvian in the 1st or 2nd semester of their Latvian language course. If one author submits more than one text (one text during the 1st semester, another one in the 2nd semester), each texts needs to have its own questionnaire filled.

The form is printed on one side of an A4 size paper sheet (for layout, see Picture 1) and includes three parts – an information letter, a permission form, and a metadata collection questionnaire (information about the author).

[2] Available online: http://www.esamkorpuss.lv

The former consists of:

- basic information about the project, the institutions that are carrying it out, and contact information;
- brief instructions for the learner;
- information about the security of data on the server used for the corpus and privacy;
- 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 used in the corpus).

The permission includes seven statements the author agrees to comply with by signing the form:

- The author agrees 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.
- The author confirms that none of the data in this text can lead to identification of any existing people. This condition is also particularly stressed when instructing the students; not all of them have a clear understanding what personal data is, so teachers who participate in the project sometimes explain the concept and help deciding what kind of data must be replaced.
- The author agrees that the text is anonymous and their name is not mentioned anywhere on the corpus website or its public documentation. While copyright issues are often solved by crediting the author, the standard solution was decided to be anonymity. There may be some authors who would not mind their names to be associated with their texts, but, the more authors are known, the easier it is to recognize the others who do not want it. The questionnaire also states that 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. However, it was later decided that associating several texts with the same author would not give enough research possibilities. Moreover, it would potentially enable one to recognize an author based on the combined contents of the texts, thus undermining the anonymity and personal data protection factors in play.
- 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 an 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.).
- 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.

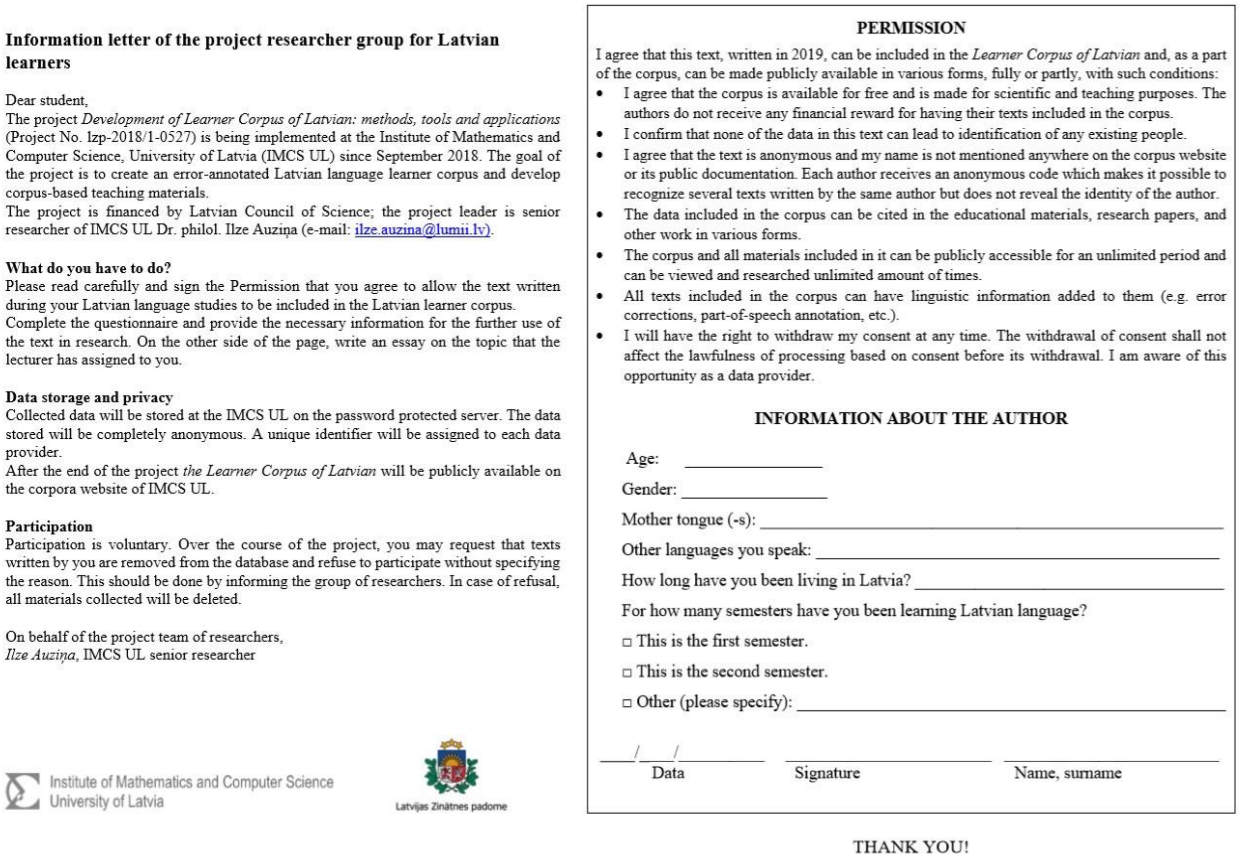
Finally, the metadata collection questionnaire requests the author to provide some information about factors that may influence their target language production: age, gender, mother tongue(-s), other spoken languages, the length of residence in Latvia, and the number of semesters studying Latvian language in a higher education institution.

The date, signature, name, and surname of the author is needed to ensure the author's full agreement with the aforementioned statements in the permission, but is not included in the metadata of the corpus. In case any of the authors later decided to revoke their consent, their name could also be used to find the text that should be deleted.

It is important to note that the information, permission agreement, and the metadata questionnaire are integrated into one document which is then printed on one side of an A4 size paper sheet. The other side of the form is blank, and authors are requested to hand-write an essay there. This eliminates the need to have any identifying information in or around the text. Since the texts are given back to their authors to ensure educational feedback, it is important for teachers to know who should receive which one of the texts. If the text were written on a separate piece of paper, it would therefore require some kind of identification that would complicate the process of avoiding personal data inclusion. The length of texts normally does not exceed the amount that is easily fitted on an A4 size paper sheet because the 1st and 2nd semester (level A1 or A2) students are usually not writing extended essays yet. This approach may not be suitable if longer texts (probably in higher language skill level) are collected. Any alternative that does not complicate matters is possible; in "LaVA", some students who prefer to write on different paper

or who needed more than one sheet of paper, stapled the text (on one piece of paper) and form (on another one) together before submitting.

When the questionnaire is completed and the essay is written, both sides of the page are scanned, and the data is further used in building the corpus. The scanned copy of the written text becomes an integral part of the corpus.



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

Institute of Mathematics and Computer Science
University of Latvia

Latvijas Zinātnes padome

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!

Picture 1: The layout of the agreement / questionnaire form

3.2 Data collection procedure

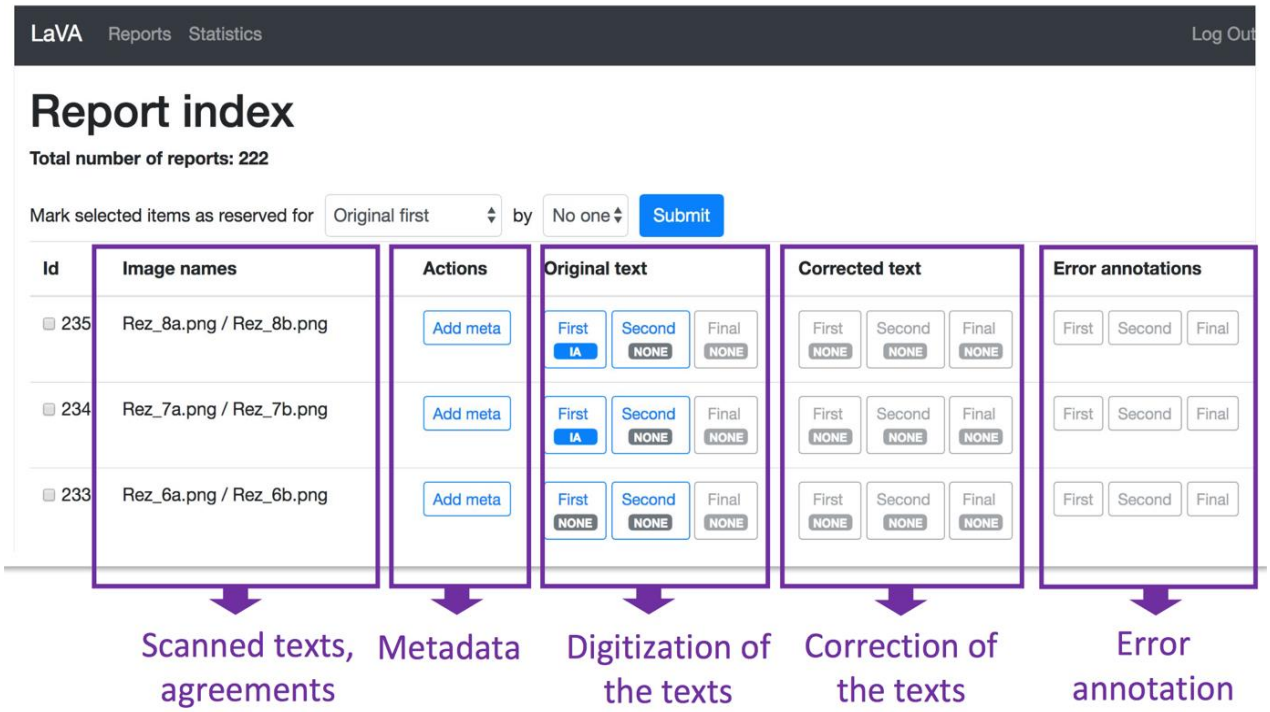
The authors of the texts are all higher education students who have been living in Latvia for a relatively short time and are learning Latvian language at the beginner level for the first or second semester. Teachers are allowed to choose the desired topic and length of the text, and study materials may be used when writing. The teachers who collect the texts instruct the students about the copyright and personal data protection system used in the project, and remind them particularly that regardless of the topic no real personal information should be included in the text. 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.

The preferred text length of each individual text is at least 100 words, as this was decided to be long enough for the learners to be able to use various phrases and constructions which demonstrates their skills of using vocabulary and grammar, as well as other aspects of language use. The maximum length of a text has not been set but rarely exceeds ~270 words.

After the texts are digitized for inclusion in the corpus, the originals are given back to the teacher who corrects them according to the needs of the pedagogical process, and then hands the texts back to the students, once more reminding them about the possibility to revoke the permission if need be (such as accidental inclusion of real personal data etc.).

The corpus is built on an integrated multifunctional platform (Figure 2) that provides a single interface for uploading, digitizing, annotating and search. At the same time, the web platform can also be used for storing scanned copies of essays, comparing texts entered and corrected by two independent digitizers, editing automatically morphological annotated and error-annotated texts, and making inter-annotator agreement.

Collected essays with metadata are handwritten; therefore, they need to be digitized for further data processing steps. The digitization is being carried out in three steps: (1) scanning of the assignments and essays; (2) metadata input; (3) text rewriting in digital format. Scanned images of the assignments help to validate data correctness if any concerns arise. Metadata is entered manually, and the authors' names are not included to retain anonymity.



Picture 2: An integrated multifunctional platform for data uploading, mark-up, annotating and search.

4 Conclusions

The agreement / metadata collection questionnaire form used in the learner corpus "LaVA" is relatively simple and it helps minimise the amount of additional paperwork involved in the creation of the corpus and gives learners a chance to exercise their rights. 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.

Acknowledgements

The work reported in this paper is part of the project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) that is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The project is financed by Latvian Council of Science.

This work is also a part of the Latvian State Research Programme “Latvian Language” (No. VPP-IZM-2018/2-0002) subproject “Acquisition of Latvian Language” and the European Structural Funds project No. 1.1.1.5/18/I/016 that are being implemented at IMCS UL.

References

[AL 2000] Autortiesību likums, 48/150 (2059/2061), 27.04.2000. [Viewed on April 29, 2019]. Available online: https://likumi.lv/doc.php?id=5138

[FPDAL 2018] Fizisko personu datu apstrādes likums, 132 (6218), 04.07.2018. [Viewed on April 29, 2019]. Available online: https://likumi.lv/ta/id/300099-fizisko-personu-datu-apstrades-likums

[GDPR 2016] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Inga Znotiņa. 2016. Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība. *Vārds un tā pētīšanas aspekti* 20 (2): 219–227.

Inga Znotiņa. 2018. Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas. Doctoral dissertation. Liepaja : Liepaja University.