CLARIN in Latvia: current situation and future perspectives

Inguna Skadiņa
Institute of Mathematics and Computer Science, University of Latvia
Riga, Latvia
Inguna.Skadina@lumii.lv

Abstract

Although Latvia is a CLARIN member supported only by the government of Latvia, it actively participates in CLARIN project activities. The paper presents current situation in Latvia – existing infrastructure (both LRT and technical), activities taken until now and further work and possible co-operation with NEALT countries.

1 Introduction

Language technologies in Latvia have a rather long history starting at the end of the 50s. While work from the earlier period (the 50s – mid-80s) is fixed now only in research papers, resources and tools developed since the mid-80s are collected carefully and many of them are available on the Web.

Currently the Institute of Mathematics and Computer Science (IMCS) of University of Latvia is the main research institution developing language tools for Latvian, while linguistic resources are created, maintained and preserved in many research organizations in Latvia, including IMCS, National Library of Latvia, universities, many research institutes and some enterprises.

Although the CLARIN initiative has been started only recently, the IMCS has been contributing to CLARIN aims already before by collecting, preserving and making public available linguistic resources, by development the Latvian language tools, by co-operating with other research organizations in resource creation and by being Web publisher and maintainer of resources created in other research institutions.

1.1 IMCS: development of resources and tools

Since 1987 the Artificial Intelligence Laboratory (AILab) at the IMCS of the University of Latvia has been concerned with natural language processing. It is one of the major centres dealing with the collection and exploration of Latvian lexical data in the NLP. Within national (funded by National research programmes "Letonika" and "Informatics", Latvian Science foundation, structural funds, Latvian State Language Agency, State Culture Capital Foundation) and international projects, different types of data have been collected, analysed and maintained at the Laboratory (Milčonoka et al., 2004; Grūzītis et al., 2004). Many resources are available on the Web (www.ailab.lv) and are used in humanities research since their creation.

Collecting of Latvian resources at the AILab has been initiated at the end of the 80s, beginning of the 90s when fragments of 'Latvian traditional beliefs' and some chapters of the first Bible translation carried out in the 17th century by Ernest Glück were keyboarded (Spektors and Baltiņa, 1994). Corpus covering the early written Latvian texts (www.ailab.lv/senie.) now contains more than 1 000 000 running words, these are mainly religious texts of the 16th and 17th century (Andronova, 2007).

The collection of modern Latvian texts comprises data from a fiction, a popular science, and some newspapers. At the moment, the number of running words is about 30 millions; about 20 millions words are with HTML mark-up and some 2.5 millions words are with SGML markup. A tiny part of data has been morphologically annotated and disambiguated. Recently the balanced corpus of 1 million words (www.korpuss.lv) has been created by support of State Language Agency.

AILab has collected numerous Latvian dictionaries – mainly explanatory dictionaries and dictionaries of terminology. Main resources are: a Term Bank, covering c. a. 115 000 Latvian terms with their translation equivalents into Russian, English, German and Latin (mainly for terms of medicine and biology) and with term definitions where available; Latvian explanatory dictionary; bilingual Latvian-Russian dictionary, electronic version of Mülenbach-Endzelin's 'Lettisch-deutsches Wörterbuch' (www.ailab.lv/mev), covering c. a. 75 000 headwords and a rich range of examples. The AILab has started an initiative to develop a new electronic dictionary to cover as much Latvian words and their meanings as possible.

Data of a spoken language have been collected and processed at the Laboratory. The speech corpus covers about 20 hours long marked texts.

Apart from lexical data, there are several tools developed for the processing of Latvian: the morphological analyser of Latvian, syntactical analyser, annotation tools etc.

1.2 IMCS: co-operation with other research institutions

National research program *Letonika* aims to facilitate and enhance research activities related to Latvia and the Latvian language, history, culture and other issues. Researchers from sixteen research institutions of Latvia, including the IMCS, participate in this program. Several important resources, such as a database of recently borrowed words, have been created.

Next to national research program *Letonika* many bilateral research projects related to linguistic resources and tools have been realized in co-operation with the University of Latvia and the National Library of Latvia.

Another way of co-operation is storing, maintenance and providing Web access to linguistic resources from the IMCS servers. Currently this type of co-operation has been established with the Latvian Language Institute, the Institute of Literature, Folklore and Arts and the University of Latvia.

2 CLARIN activities in Latvia

In 2006 IMCS and Tilde company were invited to join CLARIN initiative. Since then we actively participate in CLARIN project activities and coordinate CLARIN related activities in Latvia. Latvia is not a CLARIN consortium member yet, however we plan to join CLARIN consortium soon. Until now activities of the IMCS are financed by the Ministry of Education and Science of the Republic of Latvia.

The first year of the CLARIN project was very important for activities in Latvia. This year, two significant activities have been initiated, i. e., the CLARIN Latvia project and the National Corpus of Latvia. Since these initiatives are closely related and the target audience is very similar, we have joined our efforts in dissemination activities and in tasks related to assessing the current state of the art in language resources and tools.

2.1 CLARIN presented to the Latvian State Language Commission

On April 2, 2008 the CLARIN initiative was presented at the workshop organized by the Latvian State Language Commission. It was the first time the CLARIN initiative was presented to the researcher community in Latvia and it received a positive feedback from the participants.

The current situation in language technologies and resources was presented and discussed in a workshop. This workshop gathered ca. 30 participants – representatives of research institutes, universities, publishing houses, libraries and companies dealing with the Latvian language resources.

2.2 CLARIN National Contact Point

One of the first activities in Latvia was establishment of the National Contact Point and development of the CLARIN Latvia Web page (www.clarin.lv). The Web page is used very actively to promote CLARIN related activities in Latvia and Europe. Not only information related to project activities is published, but also different materials which could be useful for users of the infrastructure are published.

Potential contributors and users of CLARIN infrastructure are regularly informed about the CLARIN activities in Latvia by e-mails.

2.3 National seminar

On November 3, a seminar *CLARIN project and the National Corpus* was organized by the IMCS, the Latvian State Language Commission and the National Library of Latvia in order to bring together the potential CLARIN community of Latvia – owners and developers of resources, language technology developers and users of linguistic resources and tools.

The morning session was devoted to the CLARIN project. The CLARIN project coordinator Steven Krauwer presented the mission and role of the CLARIN initiative, emphasizing that all languages (widely and less widely used) are equally important in CLARIN. Participants of the seminar were introduced with CLARIN aims and tasks in Latvia, they were asked to participate actively in the creation of the CLARIN network of expertise in Latvia.

The afternoon session was devoted to the Latvian National Corpus initiative. The current state of Latvian corpus, aims of the National Corpus initiative group, experience of the Czech National Corpus and on-going work on Latvian National Digital Library (being the biggest repository of Latvian culture) were presented in the session.

The meeting was closed by a very interesting discussion on issues related to corpus, copyright issues and access to language tools.

2.4 CLARIN National Advisory Board

The CLARIN National Advisory Board was established during the National seminar with the aim to prioritize goals and tasks of the CLARIN project in Latvia and to facilitate the development of the CLARIN infrastructure.

The CLARIN National Advisory Board includes 17 members from the fields of academia, industry and government. The board members are experts in different CLARIN related issues, such as creation, maintenance and preservation of language resources, development of language technologies, language policy related issues and usage of LRT in social sciences and humanities (SSH). Tasks of the Advisory board include setting priorities and providing recommendations related to goals of CLARIN project in Latvia.

2.5 Workshop on corpus resources

On 4–5 February, a workshop *Corpus of Modern Latvian and its usage* was run at the IMCS. The aim of the workshop was to introduce SSH researchers with possibilities of corpus and corpus exploration tools in their research work. Initially the workshop was planned as a one-day event, but because of great interest, it turned into a two-day session.

The workshop revealed two important issues:

- It is very important to organize practical workshops, where researchers are introduced with possibilities of LRT infrastructure
- There is a big gap between technology and resource providers and users of language resources and tools

This was the first practical workshop; we plan to continue this work by organizing more workshops of this kind.

2.6 Workshop at Rēzekne Higher Education Institution

Rēzekne Higher Education Institution was established only in 1993. Most of teachers are young and open to new technologies and new methods in their research and education. The workshop in Rēzekne was inspired by the corpus workshop in Riga. Participants of the workshop were teachers and students of this institution. Similarly to corpus workshop in Riga this workshop revealed a great necessity for hands-on sessions and discussions on practical issues.

3 CLARIN infrastructure: the state-of-the-art

The IMCS actively participates in the following CLARIN activities: WP2 Technical Infrastructure, WP3 Humanities projects, WP5 LRT Overview, WP8 Construction and Exploitation Agreement. In Latvia work is organized around these activities. Regular internal meetings are held to exchange information between participants from different work packages.

3.1 Overview of LRT in Latvia

The IMCS actively participates in WP5 by collecting and analyzing information about tools and resources developed in Latvia which could serve as a basis for the CLARIN research infrastructure. During the National seminar two questionnaires have been distributed – one concerning resources and tools created in institutions (WP5) and the other about research projects related to the usage of language resources and technologies (WP3).

The results obtained from the questionnaire showed that there are only two institutions, namely, the IMCS and Tilde, who are developers of the Latvian language tools, technologies and applications.

The situation is much better with resources – almost all institutions whose work is related to resources – either they are linguists or computer scientists – have developed or collected some resources. There are many resources in electronic form available, however many of them are available only internally as text files. At the same time most of the resource owners are interested to share their resources and to include them in the CLARIN infrastructure.

The questionnaire as well as a workshop on corpus usage revealed one problem – even if the resources are publicly available, many potential users don't know about their existence or don't know how to explore or apply them to their own research.

3.2 Technical infrastructure

The IMCS has long term experience in telecommunications and Internet technologies. In 1992 IMCS UL has founded the Academic Network LATNET, in 2007 it was renamed to SigmaNet, the National Research and Education Network (NREN), which provides access to the GEANT2 infrastructure and offers various services.

The main goal of the research is to provide Latvian academic institutions with high quality network services according to the position of the European Union. Research focuses on practical aspects such as design and development of optical networks and deployment of high-performance gigabit network connectivity; data privacy and network security issues; technical and legal aspects of creating and keeping e-documents; legal aspects of networks usage; Grid solutions, methods and software; establishment of Grid infrastructure and the National Grid Centre, etc.

The IMCS has initiated the consolidation of academic Grid resources into the National grid network of Latvia. The IMCS currently has two operational Grid clusters of 12 and 20 CPUs. These clusters are accredited in the EGEE.

On 1 May 2008, the BalticGrid Second Phase (BalticGrid-II) project has started. It is designed to increase the impact, adoption and reach, and to further improve the support of services and users of the recently created e-Infrastructure in the Baltic States.

The IMCS is also participating in the GEANT project and is both a regional NREN (National Research and Education Network) and a CA (Certification Authority) accredited by the EUGridPMA, who coordinates the trust fabric for e-Science grid authentication in Europe.

The institute has experience in execution of large scale NLP tasks in the BalticGRID infrastructure, namely dependency chunking and morphological tagging of the whole Latvian web corpus.

Being member of the CLARIN project is a stimulus to make our language resources and tools widely accessible and compliant with established standards. Our long term intention is to become a CLARIN-conformant national-level service and metadata providing centre. For the preparatory phase, however, we have selected some existing resources and tools that can be rather rapidly integrated in the emerging CLARIN infrastructure.

4 Institutions and co-operation

4.1 The Latvian State Language Commission

The Latvian State Language Commission was established in 2002 by the President of Latvia with the aim to analyze the situation of the state language and to design recommendations for strengthening the status of Latvian as the official language.

The State Language Commission activated necessity to develop language technologies and resources; in order to achieve this, a sub-commission on the Latvian language in New Technologies has been established. The tasks defined by the sub-commission, can be grouped into two major categories. First, tasks related to the creation of a scientific basis for the introduction of the Latvian language use in new technologies. Second, a practical work aimed at the introduction and use of Latvian in new technologies, as well as the use of the new technologies in language development. Tasks set by the sub-commission are included in the State Language Policy Basic Guidelines for 2005–2014 (Vasiļjevs, 2008).

4.2 Latvian National Corpus initiative

For a number of years a development of the Latvian corpus has been among key priorities of the language policy of Latvia. Still practical implementation was hindered by a lack of funding, coordination and a limited awareness in the humanities community. To coordinate the activities of different institutions and to raise general awareness the Latvian National Corpus initiative was finally launched and a working group established in 2008.

The Latvian National Corpus (LNC) initiative has linked efforts of the Latvian State Language Commission, the National Library of Latvia, the biggest resource holders and the universities.

The Latvian National Corpus has been envisioned as a Latvian building block in CLARIN's common language resource infrastructure. It will be an open on-line resource providing access to federated resources from different research institutions and content providers.

4.3 Co-operation with universities and research institutes

The National seminar and corpus workshop revealed that research community of Latvia has a great interest in implementation CLARIN infrastructure and they are interested and ready to contribute to it.

The following institution expressed their interest in the CLARIN infrastructure: the Academy of Science of Latvia, Daugavpils University, the Institute of Latvian language, the Institute of Literature, Folklore and Arts, the Latvian State Language Agency, the Latvian State Language Commission, Liepāja University, the Ministry of Education and Sciences, the National Library of Latvia, Rēzekne Higher Education Institution, Riga Teacher Training and Educational Management Academy, Tilde, the Translation and Terminology Centre, the University of Latvia, Ventspils University College.

5 Future perspectives

Now, when the potential contributors and users of the CLARIN infrastructure have been introduced to the project, the IMCS continues work on fulfilling the aims of the CLARIN preparation phase.

There are several activities where we see our role in next years of the preparation phase. First, we will continue to contribute to the EU CLARIN project and will work to prepare Latvia for the construction phase of the project.

Second, we will continue activities related to knowledge transfer to SSH research community. Already after corpus seminar several institutions showed interest to contribute their resources. We will continue to provide technical support to them.

Third, we plan to implement the CLARIN infrastructure prototype at least for the IMCS, thus becoming a real CLARIN centre.

Fourth, we will actively co-operate with other institutions in Latvia to create nationally important resources, such as the National Corpus.

Until now co-operation with other CLARIN countries was based on informal discussions of the CLARIN implementation scenarios in other countries, however we are open for closer co-operation in future, especially with Baltic region and NEALT countries.

Acknowledgements

The CLARIN project activities are supported by the Ministry of Education and Science of the Republic of Latvia within project 'Participation of Latvia in the first period of preparation phase of the CLARIN project' (September, 2008 – April, 2009). Author would like to thank my colleagues Everita Andronova, Normunds Grūzītis, Ināra Opmane and Andrejs Spektors from IMCS and Andrejs Vasiļjevs from Tilde who contributed to this paper.

References

Andronova, Everita. 2007. The Corpus of Early Written Latvian: current state and future tasks. *Proceedings of Corpus Linguistics 2007,* Birmingham, UK. Electronic publication: (http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf).

Grūzītis, Normunds., Ilze Auziņa, Sanita Bērziņa-Reinsone, Kristīne Levāne-Petrova, Everita Milčonoka, Gunta Nešpore, Andrejs Spektors. 2004. Demonstration of resources and applications at the Artificial Intelligence Laboratory, IMCS, UL. *Proceedings of the first Baltic conference 'Human Language Technologies – the Baltic Perspective'*, Riga, pp. 38–42.

Milčonoka, Everita, Normunds Grūzītis, Andrejs Spektors. 2004. Natural language processing at the Institute of mathematics and computer science: 10 years later. *Proceedings of the first Baltic conference "Human Language Technologies - the Baltic Perspective"*, Riga, pp. 6–11.

Spektors, Andrejs and Maija Baltiņa. 1994. Latviešu valodas vēsturisko tekstu datu bāzes izveide. *Valoda un tehnika Eiropā 2000,* Riga, p. 30.

Vasiļjevs Andrejs. 2008. The influence of new technologies upon the Latvian language. *Break-out of Latvian.* Zinātne, pp. 345–355.