*Human Language Technologies – The Baltic Perspective*
*A. Tavast et al. (Eds.)*


*doi:10.3233/978-1-61499-133-5-265*

# Latvian Language Resources and Tools: Assessment, Description and Sharing

Andrejs VASIĻJEVS and Inguna SKADIŅA[1]
*Tilde*

**Abstract.** This paper describes our work on identification, assessment and cataloguing of Latvian language resources for sharing through an open language resource infrastructure. This work was carried out in the META-NORD project which is the Baltic and Nordic branch of the pan-European network META-NET. Criteria and results of the assessment are provided for the major groups of Latvian language resources and tools. Critical gaps are discussed and general strategy to address them outlined. The on-going work on language resource selection and preparation of metadata is described for their distribution on the pan-European sharing and distribution platform META-SHARE.

**Keywords.** Latvian, language resources, distribution, sharing, META-NORD, META-NET, META-SHARE

## Introduction

In the last decade we can see a strong growth in linguistic resources for many European languages. However, they are located in different places, have been developed in different formats and in many cases are not well documented. High fragmentation and a lack of unified access to language resources are among the key factors that hinder language technology research, development and usage in practical applications. This obstacle is addressed by two complementing pan-European initiatives on language resource infrastructure – META-NET[2] and CLARIN[3].

CLARIN is a long-term pan-European initiative to build a distributed research infrastructure that supports researchers in all fields, especially the Humanities and Social Sciences, dealing with language based material (text, speech, multimodal media) [1]. CLARIN aims to build a federation of trusted centres that will provide language resources and tools through web services with a single sign-on access. The purpose of the infrastructure is to offer persistent services that are secure and provide easy access to language resources [2]. In the summer of 2012, the CLARIN preparatory phase

[1]Corresponding Author: Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, LV 1459, Riga, Latvia; E-mail: Inguna.Skadina@lumii.lv.
[2] Multilingual Europe Technology Alliance, http:// www.meta-net.eu
[3] Common Language Resources and Technology Infrastructure, http://www.clarin.eu

ended and it is now, in its construction phase, CLARIN operates as a part of the European Research Infrastructure Consortium (ERIC).

The META-NET network is dedicated to building the technological foundations of a multilingual European information society by facilitating the creation of an open infrastructure to enable and support large-scale multilingual and cross-lingual services and applications. It is targeted at both industry and researchers who can benefit from centralized online access to variety of multilingual language resources. META-NET encourages development and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide range of subject domains and applications. The META-NET network is supported through several EU projects in the FP7 and ICT-PSP programmes: TE4ME, CESAR, METANET4U and META-NORD. It currently includes 57 research centres from 33 countries.

The goal of the META-NORD project [3] [4] is to establish an open linguistic infrastructure in the Baltic and Nordic countries. The project focuses on the 8 European languages of the Baltic and Nordic region – Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish. This project is the Baltic and Nordic branch of the larger pan-European network META-NET for building a common infrastructure for sharing and distributing language resources.

The following activities in Latvia are being carried out in parallel with similar activities in other European countries united by the META-NET network:

- mapping and describing the national language technology (LT) landscape in terms of language use, language technology and resources, main actors (academy, industry, government and society) that resulted in the Language Whitepaper for Latvian [5] ;
- identifying, selecting and collecting language resources;
- documenting, processing, linking, and upgrading the identified language resources to agreed standards and guidelines;
- setting up an online platform for distribution and sharing of language resources – META-SHARE;
- populating the platform with language resource metadata and actual resources.

In this paper we analyse the status of language resources and tools for the Latvian language from the perspective of a language technology infrastructure, describe the most advanced areas and identify the key gaps. While the situation with the Latvian language resources and tools was described in the previous Baltic HLT conference [6], this paper discusses aspects related to their use in the linguistic infrastructure. We describe results of work on the identification, assessment, documentation of Latvian language resources (LRs) for sharing and distribution through an open online platform. These activities were carried out in the framework of the META-NORD project.

## 1. General Overview of the Latvian Language Resources and Tools

Although work on Latvian language resources and tools can be dated back to the late 1950, it has never been a priority research field in Latvia. Today there are only a few active players in research and industry working on creating language resources and tools for Latvian. Many resources and tools developed in the last two decades are

catalogued in the CLARIN Language resource repository [4] and META-SHARE network of repositories[5].

### *1.1. Latvian Language Resources*

In Latvia work on language resources mainly concentrates around written language, while spoken language resources are in the development stage and not yet available outside the developers' institutions.

The Institute of Mathematics and Computer Science at the University of Latvia (IMCS UL) has collected, created and maintained linguistic resources and tools since the mid 80-ies. Collected and developed resources include different dictionaries [6], corpora and text collections[7] as well as some treebank data. Most of the developed resources are available on the web for browsing, some are downloadable. These resources (except treebanks) have limited or no linguistic annotation (e.g. part of speech tagging or syntactic information), as tools for automatic annotation are mostly in prototype/proof of concept stage.

The language technology company Tilde has been developing language tools and resources since the early 90-ties. For the Latvian language Tilde has created and maintains numerous general, specialized and terminological dictionaries, thesaurus, monolingual and bilingual text corpora, encyclopaedias, and some multimedia content. Among the tools created by Tilde for Latvian are a morphology analyser, spelling and grammar checker, hyphenator, named entity recognizer, term identifier, text to speech synthesizer and others. These resources are available online for browsing through the portal *letonika.lv* [8] and they are included in Tilde's applications for PCs and smartphones (e.g. Tildes Birojs, Tildes Tulkotājs).

Many Latvian language resources have been developed using proprietary annotation schemas that do not comply with common standards. This situation is improving as IMCS is participating in the ISOcat initiative and works on Multex-East morpho-syntactic specification for Latvian. Tilde works on upgrading several resources for compliance to standards to include them in the META-SHARE infrastructure.

### *1.2. Latvian Language Tools*

For Latvian the basic language tools, such as spelling checker, grammar checker and hyphenation tools, are rather well developed and widely used. These tools have reached the quality of commercial software and are included *Tilde* and *Microsoft* products. However, more advanced resources and tools are mostly in an early prototype stage or have not been developed yet.

From more advanced technologies, only machine translation (MT) has reached considerable quality and is widely used in public applications (e.g. Tilde Translator, Google Translator, Bing Translator). However, the quality of machine translation depends on the availability of language resources, which are limited for such a small language as Latvian. As a result current machine translation systems provide

[4] http://www.clarin.eu/view_resources, http://www.clarin.eu/view_tools
[5] http://www.meta-share.eu
[6] http://www.tezaurs.lv/
[7] http://www.korpuss.lv/
[8] http://www.letonika.lv/

reasonably good translation quality in some domains, while in other domains MT output is mostly unusable.

Although research on semantic analysis and controlled languages is progressing at IMCS, these tools are not yet catalogued. Moreover, more sophisticated tools, including text generation, dialogue systems or tools for robust parsing are missing or in the initial prototype stage.

The situation is even more dramatic with speech tools: while text-to-speech synthesizers for Latvian were developed several years ago and are in public use (e.g. Tilde TTS Visvaris), work on the speech recognition is very rudimentary and has not yet produced any usable tools.

### *1.3. Assessment of Latvian Language Tools and Resources*

One of the main objectives of the META-NORD project is to provide a description of the national landscape in terms of the language use, language-savvy products and services, language technologies and resources, their current level of development, main actors in the language technology field, public policies and programmes, prevailing standards and practices, main drivers and roadblocks. These descriptions form a set of language white papers for each of the network's languages, including Latvian [5].

To assess language technologies and resources within the framework of the META-NET, resources and tools for European languages, including the Latvian language, were identified and analysed using common criteria: quantity, availability, quality, coverage, maturity, sustainability and adaptability. Scores on a scale from 0 (very low) to 6 (very high) were assigned for each criterion. Results of this exercise are presented in Table 1.

**Table 1.**State of language technology support for the Latvian language.

| | Quantity | Availability | Quality | Coverage | Maturity | Sustainability | Adaptability |
|---|---|---|---|---|---|---|---|
| **Language Technology: Tools, Technologies and Applications** | | | | | | | |
| Speech Recognition | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Speech Synthesis | 2 | 3 | 4 | 3 | 4 | 3 | 4 |
| Grammatical analysis | 2,5 | 2 | 3 | 3,5 | 4 | 3 | 4 |
| Semantic analysis | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Text generation | 1 | 2 | 1 | 2 | 2 | 1 | 2 |
| Machine translation | 3 | 4 | 3 | 3 | 4 | 3 | 4 |
| **Language Resources: Resources, Data and Knowledge Bases** | | | | | | | |
| Text corpora | 2 | 4 | 4 | 3 | 3 | 3 | 4,5 |
| Speech corpora | 1 | 0 | 1 | 1 | 1 | 1 | 3 |
| Parallel corpora | 1 | 3 | 2 | 2 | 3 | 4 | 4 |
| Lexical resources | 3 | 3,5 | 4 | 3 | 4,5 | 4,5 | 4,5 |
| Grammars | 2 | 1 | 3 | 2 | 3 | 4 | 3 |

This assessment indicates that in a number Latvian language research areas there are tools available, although their quality or functionality could be significantly improved. Currently, text analysis components and language resources for Latvian cover the linguistic phenomena to a certain extent and form part of applications involving mostly shallow natural language processing, e.g., spelling and grammar correction. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas.

A critical lack is in many advanced language technologies. The more linguistic and semantic knowledge a tool draws on the more gaps there are in the technology. There is a need for a greater effort to support development of deep linguistic processing. To build sophisticated applications such as quality machine translation there is a need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text.

Creation of speech resources and tools are only in an initial phase. Obviously, further efforts are required to develop the most used language resources, such as speech resources and parallel corpora for machine translation. Widely used resources as transcribed speech corpora, Wordnet, FrameNet and treebanks or are not available for the Latvian language.

Moreover, a number of tools, resources, and data formats created for Latvian do not meet industry standards and cannot be sustained effectively. A concerted programme is required to standardise data formats and API's.

Finally there is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. As a result the Latvian language is less equipped with language resources and tools than most of other official EU languages. In critical language technology areas like speech processing and language resources, Latvian does not reach the quality and coverage not only of English, but also of several under-resourced languages of the Baltic and Nordic region (Table 2).

Our findings show that targeted national research and development activities are urgently needed to fill the identified gaps.

**Table 2.** Availability of language resources and tools for languages of Baltic and Nordic countries.

| | **Excellent** | **Good** | **Moderate** | **Fragmentary** | **Week/No** |
|---|---|---|---|---|---|
| **Speech Processing** | | | Finnish | Danish, Estonian, Norwegian, Swedish | Icelandic, **Latvian,** Lithuanian |
| **Machine Translation** | | | | | Danish, Estonian, Finnish, Icelandic, **Latvian,** Lithuanian, Norwegian, Swedish |
| **Text Analysis** | | | | Danish, Finnish, Norwegian, Swedish | Estonian, Icelandic, **Latvian,** Lithuanian |
| **Resources** | | | Swedish | Danish, Estonian, Finnish, Norwegian | Icelandic, **Latvian,** Lithuanian |

## 2. Resource Selection and Description

Another goal of the META-NORD project is to contribute to a pan-European digital resource exchange facility by identifying and collecting resources in the Baltic and Nordic countries.

At the initial stage of the project we identified more than 40 Latvian language resources and tools. The identified language resources were assessed using a number of criteria: *availability* (openly available LR versus restricted proprietary), *suitability* for LT development, *multilinguality* (in terms of support multilingualism or linking between languages), *longevity* (maintenance of LR), *quality*, *extensibility* (level of available documentation and metadata description) for inclusion into a LR sharing infrastructure.

The assessment procedure revealed several serious problems. First is an unclear legal situation which restricts making use of digital texts. Many resources have some limitations for use or they have been created by so many institutions and people that clarification of intellectual property rights becomes practically impossible. Together with politicians or policy makers, researchers should try to develop laws and regulations that enable researchers to use publicly available texts for language-related R&D activities.

Secondly, many resources are in an early prototype stage thus not usable for application or even for filling gaps in the basic language technology tools. Finally, many resource owners are not ready to share their resources. As the result we were able to select 10 resources to document, process and upgrade to the agreed standards for inclusion in the META-SHARE repository (see next section).

These selected resources are described using the common metadata model developed within the META-NET network [7]. The metadata model is inspired by the *component-based mechanism* (Component MetaData Infrastructure, CMDI), according to which semantically coherent elements are grouped together to form components [8]. *Components* are the core building blocks of the metadata model. They consist of *elements* (categories) that are used to encode specific descriptive features. In order to accommodate flexibility, the elements belong to two basic levels of description:

- an initial level providing the basic elements for the description of a resource (*minimal schema*), and
- a second level with a higher degree of granularity (*maximal schema*), providing detailed information on a resource and covering all stages of LR production and use [7].

The minimal schema contains those elements considered indispensable for LR description (from the provider's perspective) and identification (from the consumer's perspective) [9].

## 3. Setting up and Populating an Online Sharing Platform

For distribution and sharing of language resources META-NET is establishing a distributed online platform META-SHARE [10]. It consists of independent META-SHARE nodes set up in different countries and interlinked into a federated repository. This online infrastructure provides facilities for describing, storing, preserving, and making publicly available LR repository in an open, user-friendly and trustworthy way. Among different language data that can be considered useful for different purposes, META-SHARE places a strong focus on language data that are important in language

technology development for building applications useful to the EU citizen, primarily in his everyday communication and information search needs.

Currently the META-NORD project has set up four META-SHARE nodes: at Tilde[9], University of Gothenburg[10], University of Helsinki[11] and University of Tartu[12]. The Tilde based node is the master node, where information about catalogued language resources is collected. In the next version of META-SHARE software synchronization of the metadata repository will be implemented, making a full catalog browsable from any META-SHARE node. At the time of submitting this paper (July, 2012) 207 language resources and tools have been catalogued by the META-NORD consortium: 55 lexical resources, 111 corpora, 11 treebanks, 12 speech resources, 5 monolingual wordnets, 4 pilot bilingual wordnets, and 9 tools.

For the Latvian language nine resources are currently available on the META-SHARE node representing different types of LRs. These resources were selected according to the criteria described in the previous section. The special focus was on resources that are important for language technology development, especially resources that can be used to build applications that help to overcome language barriers. These resources include bi/multi-lingual lexicons (e.g. Latvian-Lithuanian Dictionary), monolingual corpora (e.g. Corpus of Latvian Literature), parallel corpora (e.g. Latvian-English N-gram Corpus of Legislation of Republic of Latvia) and terminology resources (e.g. EuroTermBank). Special attention has been paid to resources and tools that have been created in the EU co-funded projects where Tilde is the coordinator or partner responsible for development of particular resource (e.g. outcomes of ACCURAT and EASTIN-CL projects). Before resources have been catalogued, IPR issues are clarified and licensing conditions fixed in the metadata.

Tilde is also working on integration of EuroTermBank [11] terminology databank making it a dedicated META-SHARE node for terminology resources.

## Conclusion

While several basic language resources and tools are rather well represented for the Latvian language, more advanced resources and tools are missing. Since LT has never been a priority research field in Latvia there are rather large gaps in language resources and in the tools needed for a sustainable development of the Latvian language. These gaps exist not only in comparison with the widely spoken languages, but also in comparison with the lesser spoken languages that have benefited from a dedicated language technology programmes, e.g., Estonian.

A focused long-term endeavour, such as a Language Technology Programme, would address this situation providing resources necessary for creation of Latvian LT resources to foster research, innovation, and development. The need for large amount of data and high complexity of language technology systems make it vital for Latvia to participate and contribute to the META-NET and CLARIN infrastructures. Recent advancements in such sophisticated areas as machine translation lead us to believe that with targeted efforts and cooperation of research institutions, funding organizations,

[9] http://metashare21.tilde.lv/
[10] http://spraakbanken.gu.se/metashare/
[11] http://metashare.csc.fi/
[12] http://metashare.ut.ee/

universities and industry Latvian can become one of the most technologically advanced languages of multilingual Europe.

## Acknowledgements

The work described in this paper was carried out in the META-NORD project which has received funding from the European Commission through the ICT PSP Programme, grant agreement no 270899.

## References

[1] T. Váradi, P. Wittenburg, S. Krauwer, M. Wynne, K. Koskenniemi. CLARIN: Common Language Resources and Technology Infrastructure. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 1244-1248, 2008.

[2] P. Wittenburg, N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, L. Lemnitzer, B. Maegaard, M. Piasecki, J. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R. Veenendaal, T. Váradi., M. Wynne. Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure. *Proceedings of the Seventh conference on International Language Resources and Evaluation. (LREC'10),* 2010, May 19-21, Valletta, Malta, 60-63, 2010.

[3] Vasiļjevs, B.S. Pedersen, K. De Smedt, L. Borin, Lars and I. Skadiņa, META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure, *NODALIDA 2011 workshop Visibility and Availability of LT Resources, NEALT Proceedings Series,* Vol.13, 18-22, 2011.

[4] I. Skadiņa, A. Vasiljevs, L. Borin, K. de Smedt, K. Linden, E. Rognvaldsson. META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries. *Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, 107-114, Chiang Mai, Thailand, 2011.

[5] I. Skadiņa, A. Veisbergs, A. Vasiļjevs, T. Gornostay, I. Keiša, A. Rudzīte. *The Latvian Language in the Digital Age*. Springer, 2012 (in print).

[6] I. Skadiņa, I. Auziņa, N. Grūzītis, K. Levāne-Petrova, G. Nešpore, R. Skadiņš, A. Vasiļjevs. Language Resources and Technology for the Humanities in Latvia (2004–2010). *Proceedings of the Fourth International Conference Baltic HLT 2010*, IOS Press, Frontiers in Artificial Intelligence and Applications 219 (2010), 15-22.

[7] M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz and V. Mapelli. The META-SHARE Metadata Schema for the Description of Language Resources. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 1090-1097, 2012.

[8] D. Broeder, T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari and P. Wittenburg, Foundation of a Component-based Flexible Registry for Language Resources and Technology, *Proceedings of the 6th International Conference of Language Resources and Evaluation*, 2008.

[9] C. Federmann, B. Georgantopoulos, R. del Gratta, O. Hamon, B. Magnini, D. Mavroeidis, S. Piperidis, M. Schroeder, M. Speranza. *META-NET Deliverable D7.1.1 – META-SHARE Functional and Technical Specification*, 2011.

[10] S. Piperidis. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 36-42, Beijing, 2012.

[11] A. Vasiļjevs, K.D. Schmitz, Collection, harmonization and dissemination of dispersed multilingual terminology resources in an online terminology databank, *Proceedings of TSTT 2006, Third International Conference on Terminology, Standardization and Technology Transfer*, 265-272, 2006.