Latvian Language Resources and Tools: Assessment, Description and Sharing

Andrejs Vasiljevs (andrejs@tilde.com) and Inguna Skadiņa (inguna.skadina@tilde.lv)

THE META-NORD PROJECT AIMS TO ESTABLISH AN OPEN LINGUISTIC INFRASTRUCTURE IN THE BALTIC — ESTONIA, LATVIA AND LITHUANIA, AND NORDIC COUNTRIES — DENMARK, FINLAND, ICELAND, NORWAY AND SWEDEN

State of language technology support for the Latvian language

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
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 Technology: Tools, Technologies and Applications

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
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

Language Resources: Resources, Data and Knowledge Bases

Scores on a scale from 0 (very low) to 6 (very high) were assigned for each criterion.

- The language resources and tools 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.
- A critical **lack** is in many **advanced language technologies**.
- From more advanced technologies, only machine translation (MT) has reached considerable quality and is widely used in public applications
- The more linguistic and semantic knowledge a tool draws on the more gaps there are in the technology.
- Creation of **speech resources and tools** are only in an **initial phase.**
- Number of tools and resources created for Latvian **do not meet industry standards** and cannot be sustained effectively
- There is a **need for** a greater effort to support development of **deep linguistic processing**.

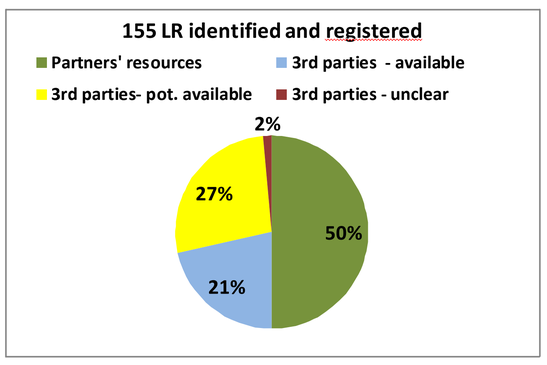
Availability of language resources and tools for languages of Baltic and Nordic countries

	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,
Text Analysis				Danish, Finnish, Norwegian, Swedish	Estonian, Icelandic, **Latvian,** Lithuanian
Resources			Swedish	Danish, Estonian, Finnish, Norwegian	Icelandic, **Latvian**, Lithuanian

Availability of language resources and tools for languages of Baltic and Nordic countries

- An initial comparison across all 30 META-NET languages places three small languages of the Nordic and Baltic region – **Icelandic, Latvian, and Lithuanian – in the bottom cluster**, defined as having major gaps in all of the four key areas.
- 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.
- Targeted **national research and development activities** are urgently needed to fill the identified gaps.

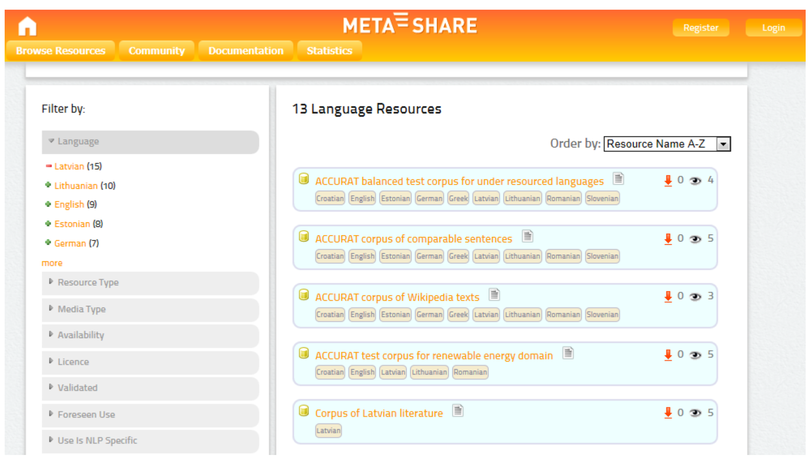
Resource selection, description and population



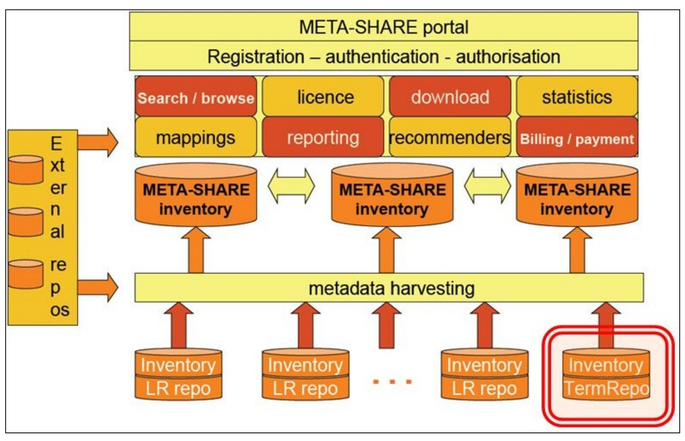
Identified resources are selected according to their quality, relevance, and usability in multilingual services.

Lexical resources	55
Corpora	111
Treebanks	11
Resources for speech	12
WordNets	9
Tools	9
TOTAL	207

The total number of LRs in the second batch (August, 2012).



META-SHARE nodes operational at Tilde, University of Gothenburg, University of Helsinki and University of Tartu.



Terminology repository within META-SHARE network

- Extension of an open linguistic infrastructure with multilingual terminology resources.
- Based on the EuroTermBank terminology databank—the European hub of multilingual terminology that provides an access to 2.3M terms from more than 100 terminology resources and 4 external databases linked in a federated architecture.
- Integration of the EuroTermBank platform into an open linguistic infrastructure by adapting it to relevant data access and sharing specifications.
- Populating EuroTermBank with additional resources of terminology data.

The META-NORD project has received funding from the European Commission through the ICT PSP Programme, grant agreement 270899