RESEARCH AND INNOVATION IN LANGUAGE TECHNOLOGY AT THE ARTIFICIAL INTELLIGENCE LABORATORY

Ilze Auziņa, *Dr. philol.*, leading researcher
Normunds Grūzītis, *Dr. sc. comp.*, leading researcher
Guntis Bārzdiņš, *Dr. sc. comp.*, Professor, leading researcher, Corresponding Member of the Latvian Academy of Sciences
Institute of Mathematics and Computer Science, University of Latvia

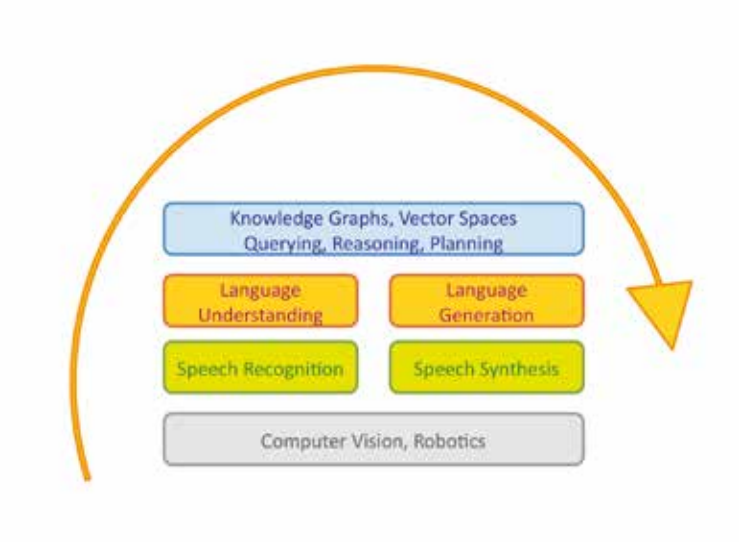
The Artificial Intelligence Laboratory (AI Lab) at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL), founded in 1992, conducts research in natural language processing (NLP) and machine learning (ML). Both research directions are closely related and have gained significant boost through the implementation of numerous innovation projects together with industry partners and through the international cooperation.

AI Lab particularly focuses on cross-lingual natural language understanding (NLU) and generation (NLG) by combining knowledge-based and machine learning approaches. Our work on NLU includes speech recognition, information extraction and knowledge graph construction from unstructured texts and audio recordings, as well as image and video data. The work on NLG includes text generation from data and abstract meaning representations, as well text-to-speech synthesis. We conduct research in NLU and NLG in several directions and aspects:

- Speech-to-text recognition and text-to-speech synthesis for Latvian.
- Syntactic and semantic parsing for cross-lingual information extraction, question answering, human–computer interaction.
- Multilingual text generation from abstract meaning representations and from data.
- Controlled natural language for knowledge representation.
- Creation of annotated language resources for NLP and ML: machine-readable dictionaries, text and speech corpora (training and evaluation datasets).
- Machine learning for NLU and NLG, also in connection to robotics, computer vision and grounded language learning.

The core research group consists of twenty staff members: leading researchers, researchers and research assistants. Since the foundation of AI Lab, there have been computer scientists and linguists working side by side; in fact, both fields are represented in rather equal proportions. Most of the leading researchers are also involved in teaching at the University of Latvia, which facilitates knowledge transfer and student involvement.

AI Lab has recently implemented multiple research and innovation projects on large-scale information extraction and speech recognition. Most notably, we have established a long-term partnership with the largest Latvian news agency LETA, developing scalable software platforms for fully automated and human-in-the-loop media monitoring. This relies on state-of-the-art speech recognition, semantic parsing and knowledge representation technologies.



Schematic representation of natural language understanding (NLU) and natural language generation (NLG)

Together with LETA, we have implemented more than ten national and EU-level projects. Through this partnership, our researchers are collaborating also with the global broadcasters, namely Deutsche Welle (DW) and BBC, helping to solve language processing tasks for the big-data media monitoring and multilingual news production.

In 2016, the successful collaboration allowed LETA to join the EU Horizon 2020 big-data project on Scalable Understanding of Multilingual MediA (SUMMA; 688139), together with leading European universities, BBC, DW and others; a team of AI Lab researchers was also fully involved through this partnership. In 2021, we have joined a new H2020 project on Stream Learning for Multilingual Knowledge Transfer (SELMA; 957017), coordinated by DW. Our team is responsible for the integration of the whole SELMA platform.

AI Lab has also established a close partnership with several research groups and public service providers working in the humanities and social sciences areas in Latvia. We have previously developed a text processing system for the National Library of Latvia, dealing with part-of-speech tagging and named-entity recognition in a billion word corpus of OCR-scanned texts. This joint work is now continued on a new level in a state research programme on Digital Resources for Humanities (VPP-IZM-DH-2020/1-0001). In partnership with the Latvian Language Institute, AI Lab is continuing a constant development of extensive machine-readable dictionaries of Latvian and of an integrated on-line dictionary platform Tezaurs.lv (VPP-IZM-2018/2-0002), which is one of the most widely and frequently used Latvian language resources. A relatively new but rapidly growing research direction in digital humanities is also the creation of resources and tools for corpus-based analysis of language learners and for the consequent development of improved learning aids (lzp-2018/1-0527). This has strengthened our collaboration with regional universities and with the Latvian Language Agency.

For many years, AI Lab is also collaborating with the Faculty of Communication at Rīga Stradiņš University, developing tools and datasets for monitoring aggressiveness and hate speech in user-generated content of news portal communities and for annotating and analysing transcripts of parliamentary debates.

In 2013, AI Lab together with LETA and the language technology company Tilde created the first general-purpose machine learning dataset for Latvian speech recognition. Since then, speech recognition of commercial quality has been rapidly developed by the leading language technology companies in Latvia. Recently, we have established a partnership with Rīga East University Hospital, the largest hospital in Latvia. Within an industry-driven research programme of the European Regional Development Fund, we are creating and adapting language resources and models for Latvian speech recognition in radiology (1.1.1.1/18/A/153). These components are being further integrated in a prototype of an automated dictation platform for medical reporting.

Although AI Lab primarily focuses on the language resource and technology development for the less-resourced Latvian language, our teams have successfully participated in international NLU and NLG evaluation campaigns on well-resourced languages as well. In 2016 and 2017, we scored the best results in the SemEval shared tasks on Abstract Meaning Representation parsing and generation from and to English.

Although the language resources that we develop are Latvian-specific, the formal annotations that we add are compatible with state-of-the-art representations. AI Lab is among the early data contributors to the international Universal Dependencies (UD)

Staff of the Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, in 2020: from right to left: the 2nd Normunds Grūzītis, the 5th Ilze Auziņa, the 6th Guntis Bārzdiņš

framework for syntactic parsing, coordinated by the Uppsala University. The Latvian UD dataset (a so called treebank) is already qualified as a comparatively big one and is being used in multilingual settings by many research teams world-wide. Similarly, we are active participants in the global FrameNet initiative for cross-lingual frame-semantic parsing. Our experience in creating FrameNet annotations on top of UD treebanks has attracted interest from researchers working on other languages, and it has also inspired us in the development of novel tools for frame-semantic media monitoring.

IMCS UL is the national coordinator of CLARIN ERIC, the European research infrastructure of language resources and technology for the humanities and social sciences (1.1.1.5/18/I/016). IMCS UL is also the national competence centre of the European Language Grid platform (ELG), as well as a national technology anchor point of the European Language Resource Coordination initiative (ELRC) and an observer of the European Lexicographic Infrastructure (ELEXIS). Being a part of these infrastructures allows us to make the language resources and tools developed by AI Lab available and, first of all, discoverable to other research teams and technology companies in Europe and world-wide, which is essential for a relatively small language like Latvian not only to survive but flourish in the digital era.