Language Technology Initiative: Framework for Teaching NLP and Computational Linguistics at the Universities in Latvia

Inguna Skadiņa[1], Jana Kuzmina[1], Marina Platonova[2], Tatjana Smirnova[2], Sergejs Kruks[3]
[1]University of Latvia, [2]Riga Technical University, [3]Riga Stradiņš University
Correspondence: inguna.skadina@lu.lv

Abstract

This short paper provides an overview of language technology related modules and courses developed at three leading universities of Latvia - University of Latvia (UL), Riga Technical University (RTU) and Riga Stradiņš University (RSU).

1 Introduction

Although several language technology (LT) courses have been offered at different universities in Latvia for more than twenty years, with the rapid progress of artificial intelligence (AI) and the continued advancement of LT, the demand for LT modules and courses has increased significantly. What was once a specialized area, primarily within computer science, has now expanded into humanities and social sciences, where LT is increasingly used for text analysis, digital humanities and social media research. Moreover, the scope of humanities has also changed considerably, reshaping how scholars research, interpret, and engage with human culture and communication. This transformation affects both research methods, theoretical perspectives, and practical implementation of digital solutions to explore cultural, linguistic, and historical phenomena (Adolphs and Knight, 2020).

In 2022, the Ministry of Education and Science of Latvia identified the need for designated Language Technology and Computational Linguistics programs, and proposed investing in the development of high-level digital skills to significantly increase the number of LT specialists by 2026. The Language Technology Initiative (LTI) project[1] was initiated to implement this goal. The key objectives of this project are to prepare a curriculum for LT teaching, advance language resources, and create platforms and digital solutions for study and experimentation (Skadiņa et al., 2024).

This short paper introduces the modules, courses, and teaching materials developed within the LTI project through the DigComp 2.2 Competence Framework for resource development for computer science, humanities and social sciences students.

2 Language Technology Modules for Computer Science Students

The Faculty of Science and Technology at UL has developed two LT modules – one for Bachelor's and one for Master's students.

The Master module in Natural Language Processing comprises two courses – "Applications of Language Technology" and "Deep Machine Learning". The "Applications of Language Technology" course introduces state-of-the-art technologies for natural language understanding and generation, speech recognition and synthesis. It aims to provide theoretical knowledge and necessary practical skills to use and integrate existing AI solutions, and to develop innovative ones.

The Bachelor module comprises three courses – "Fundamentals of Natural Language Processing", "Fundamentals of Deep Machine Learning", and "Introduction to the Python Programming Language". The "Fundamentals of Natural Language Processing" course aims to introduce students to LT, covering basic methods and the most important innovations and trends in the field, focusing on data-driven methods and the required language resources. The course is based on selected chapters from Jurafsky and Martin (2025) text-book, adapted and extended for the Latvian context.

The majority of the courses are offered in the Latvian language, with the exception of the "Fundamentals of Natural Language Processing", which is offered in both Latvian and English. Teaching materials from both modules are available to students in the e-Study environment of the UL.[2] Courses

[1] https://www.vti.lu.lv/en/

[2] https://estudijas.lu.lv/?lang=en

Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026), pages 30–33

also make systematic use of Jupyter notebooks,[3] making students active participants in the learning process: they can modify code, run experiments, inspect intermediate results, and explore alternative solutions. A more detailed overview of these modules is presented in Skadiņa et al. (2026).

3 Computational Linguistics and Natural Language Processing for Humanities

The Faculty of Humanities at UL has developed four modules covering all educational levels – Doctoral students in Linguistics, Master students of English Studies and Master students of Latvian Language, Literature and Culture, as well as Bachelor students. The metalanguages of the modules and the courses are both Latvian and English and depend on the language of instruction of the study programmes and their goals.

The module for Doctoral students further develops the academic competence for conducting independent and innovative research in the CL field. It consists of two courses: "Corpus Linguistics in the Context of Digital Humanities" and "Language, Thinking and Language Acquisition".

The module for Master students of English Studies provides in-depth knowledge, skills and competence for LT use in English linguistics and literature studies. It consists of three courses: "Corpus Linguistics", "Programming Languages for Linguists" and "Traditional and Electronic Lexicography".

The module for Master students of Latvian Language, Literature and Culture provides in-depth knowledge, skills and competence for LT use in Latvian linguistics and literature studies. It comprises two courses: "Morphemics, Morphonology and Latvian Language Morpheme Database" (Kalnača et al., 2025) and "Spoken Data Processing and Analysis".

The courses for bachelor students provide "Introduction to Applied Linguistics and Language Technologies", explore "Digital Opportunities for Language Learning and Research" and apply automated text analysis tools to investigate the EU and project management discourse.

The learning outcomes of the modules comprise various digital areas, i.e. information and data literacy, digital content creation and problem-solving. In particular, the activities aiming at information and data literacy comprise several competences at advanced and at highly specialised level.

4 Language Technology Studies through MOOC

The Faculty of Computer Science, Information Technology and Energy of RTU offers two study modules comprising three MOOCs (massive open online courses) each, which are are available at the RTU online course learning platform.[4] The interdisciplinary curricula of the study courses have been designed to meet the needs of the students irrespective of their background. Each study course provides students with the opportunity for distant self-paced study using a variety of learning materials, including at least 50 interactive videos, engaging case studies aimed at the development of advanced competence in using language technologies, and a wide range of H5P activities, or those created using HTML5 Package.

The module "Language Technologies for Multimodal Information Processing" comprises three study courses, namely, "Digital Semantics and Pragmatics", "Multimodal Digital Semiotics", and "Digital Sentiment Analysis". The courses have been designed to meet the educational needs of post-graduate students who seek to expand the range of their digital competences and skills.

The curriculum of the course "Digital Semantics and Pragmatics" is arguably the most saturated and demanding in terms of time and effort students are expected to invest. Students learn the principles of compositional and distributional semantics, generative grammar and develop skills in using word vectors, word clouds and word nets as tools in organizing and representing semantic knowledge. Students also develop skills in using transformers in text processing, summarization and generation.

The course "Multimodal Digital Semiotics" not only guides students through the state-of-the-art and new ideas concerning the role of semiotic analysis in contemporary media studies, it also provides a platform for the students to advance their digital creativity, digital storytelling and digital semiosphere development skills. Through a series of engaging H5P activities and technology-intense tasks, students use and customize LT tools for solving a range of tasks in such areas as digital news analysis, semiotics of games and advertising, inter-semiotic translation and big multimodal data analysis.

[3] The Notebooks for different courses are available from GitHub repositories: Language Technology courses: https://github.com/LUMII-AILab/NLP_Course; Python course: https://github.com/CaptSolo/LU_Python_course

[4] https://moocinfo.rtu.lv/en/

The "Digital Sentiment Analysis" course is perhaps the most practice-driven study course in the module. It provides a comprehensive view of sentiment and emotion analysis in various media and allows students to develop high-level skills for a wide range of applications in lexicography, marketing and culture. Students apply NLP solutions for data and text mining and processing using pre-trained language models, develop skills in using and customization of sentiment libraries, and learn how to use such frameworks and libraries as TensorFlow, PyTorch, and Hugging Face Transformers.

The study module "Language Technologies for Translation Theory and Practice" offers undergraduate students to enroll in one of the three courses: "Digital Edutainment Elements in Translation", "Machine Learning for Textual Data Processing", and "Machine Translation Skillset". It is interesting to note that the course "Digital Edutainment Elements in Translation" appeared useful not only for the students but also for academic staff of RTU, specifically keeping in mind that edutainment as an educational paradigm uniting education with entertainment is currently gaining momentum.

5 Natural Language Processing and Social Sciences

The Faculty of Social Sciences at the RSU offers postgraduate students a course on Discourse Analysis (DA). Lectures dedicated to qualitative methods provide an overview of theories of meaning and teach methodologies of Critical DA and Cognitive DA. Lectures dedicated to quantitative methods introduce the Latvian National Corpus Collection[5] and teach the extraction of data from the corpora using the NoSketch Engine program (Kilgarriff et al., 2004). Besides the publicly available corpora, students work with texts of their research interests in Linux. Students learn the basics of Bash programming language that enables creation of ad hoc mini-corpora and extraction of data from them. During practical assignments, students extract words that designate essentially contested concepts in Latvian language and interpret their contextual meaning. In current political, legal, and media discourse, words like 'welfare state', 'politics', and 'social integration' are applied inconsistently. The problem derives from "a straightforward failure to specify the relationship between 'term' and 'meaning', involving confusion about concepts" (Collier et al., 2006). To raise language users' awareness of the meaning of concepts, students are encouraged to submit their findings to the online Modern Latvian Language Dictionary (Zuicena et al., 2025).[6] Therefore, the students of the Social Work program suggested a new definition of the word *invaliditāte* (disability) that complies with modern social policy.

Special attention is paid to the use of ambiguous grammatical forms (AGF), which have become widespread in the media during the last decade (Kruks, 2026). Ambiguity occurs when a word has more than one meaning (Gillon (1990); Kennedy (2011)). Combining the characteristics of different parts of speech, simultaneously an AGF can designate a subject, object, or predicate; a process, or result. Using AGFs, the sender conceals the agent and some aspects of the action and assumes no responsibility for the content of the proposition. In this capacity, language becomes an instrument of power relations (Kruks, 2024). Students quantitively compare the incidence of AGFs across corpora and assess their manipulative potential using the CDA methodology.

6 Conclusion

This paper provides an overview of eight modules developed for students in computer science, humanities and social sciences at Latvian universities. The information about all developed modules and courses is available on the LTI project website.[7] The study courses were launched in Spring 2024, and over the course of three semesters, more than 750 students in Latvia have completed one or several courses.

The framework of module design and integration into the curriculum has been validated by external stakeholders (Latvian Information and Communication Technology Association experts) and has been highly evaluated by students in their post-course quensionnaires and proven to be successful.

Although the LTI project ends in June 2026, the modules and courses described in this paper have been integrated into the curricula and will continue to be offered within the respective faculties beyond the project duration. Several courses developed at UL are currently being adapted for distance learning, expanding their accessibility and long-term sustainability.

[5] https://korpuss.lv/

[6] https://mlvv.tezaurs.lv

[7] Developed modules and courses: https://www.vti.lu.lv/en/education/study-modules/

Acknowledgments

The "Language Technology Initiative" project (No. 2.3.1.1.i.0/1/22/I/CFLA/002) is funded by the European Union Recovery and Resilience Mechanism Investment and the National Development Plan.

References

Svenja Adolphs and Dawn Knight. 2020. *The Routledge Handbook of English Language and Digital Humanities*. Routledge handbooks in English language studies. Routledge.

David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. 2006. Essentially contested concepts: Debates and applications. *Journal of Political Ideologies*, 11(3):211–246.

Brendan S. Gillon. 1990. *Ambiguity, generality, and indeterminacy: Tests and definitions*, volume 85, page 391–416.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Andra Kalnača, Tatjana Pakalne, and Kristīne Levāne-Petrova. 2025. Database of Latvian morphemes and derivational models: ideas and expected results. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 279–286, Tallinn, Estonia. University of Tartu Library.

Christopher Kennedy. 2011. *23. Ambiguity and vagueness: An overview*, pages 507–535. De Gruyter Mouton, Berlin, Boston.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The sketch engine. *Proceedings of the 11th EURALEX International Congress*, pages 105–116.

Sergejs Kruks. 2024. Ambiguous grammaratical forms in Latvian corpora. *Letonica*, 56:160–175.

Sergejs Kruks. 2026. Ambiguous grammar of news in Latvian online media.forthcoming. *Letonica*.

Inguna Skadiņa, Guntis Bārzdiņš, Uldis Bojārs, Normunds Grūzītis, and Pēteris Paikens. 2026. Teaching NLP in the AI era: Experiences from the university of latvia. In *Proceedings of the Seventh Workshop on Teaching NLP*. Association for Computational Linguistics.

Inguna Skadiņa, Jana Kuzmina, Sergejs Kruks, Marina Platonova, Tatjana Smirnova, and Ilze Auzina. 2024. Language technology initiative - bridging the gap between research and education. In *CLARIN Annual Conference Proceedings*, pages 26–30.

Ieva Zuicena, Ieva Auziņa, Santa Briede, Irēna Ilga Jansone, Ieva Kuplā, Gunta Lejniece, Ilga Migla, Laimdota Oldere, Ārija Ozola, Vija Požarnova, Sanda Rapa, Anitra Roze, Imants Šmidebergs, Dorisa Šnē, Māra Šnē, Agris Timuška, Mikus Grasmanis, Lauma Pretkalniņa, and Artūrs Znotiņš. 2025. Dictionary of contemporary latvian language (MLVV) (2025-12-21). CLARIN-LV digital library at IMCS, University of Latvia.