Baltic J. Modern Computing, Vol. 10 (2022), No. 3, 490-503
https://doi.org/10.22364/bjmc.2022.10.3.21

Latvian Language in the Digital Age: The Main Achievements in the Last Decade

Inguna SKADIŅA[1,2], Baiba SAULĪTE[1], Ilze AUZIŅA[1],
Normunds GRŪZĪTIS[1], Andrejs VASIĻJEVS[2],
Raivis SKADIŅŠ[2], Mārcis PINNIS[2]

[1] Institute of Mathematics and Computer Science, University of Latvia (IMCS UL)
[2] Tilde, Riga, Latvia

[1] firstname.lastname@lumii.lv; [2] firstname.lastname@tilde.lv

Abstract. Ten years ago, when the META-NET Network of Excellence conducted a study on language technology support for European languages, Latvian was included in the category of languages with little or no support. During the last decade, notable progress has been made in the development of language resources and tools for Latvian, particularly regarding the creation of advanced datasets like speech corpora and treebanks, state-of-the-art neural language models, machine translation systems, speech technology, and technologies for natural language understanding and human-computer interaction. This paper provides an overview of the most recent activities in the language technology field in Latvia: national and international initiatives, key language resources and tools, key projects and initiatives. We summarize both the recent activities and the most significant achievements after the publication of the META-NET White Paper on Latvian.

Keywords: Latvian language, language resources and tools, speech recognition and synthesis, machine translation, human-computer interaction

1. Introduction

Ten years ago, the META-NET Network of Excellence conducted an extensive study on 31 European languages on the level of language technology support for these languages. This survey was published in a White Paper book series describing the technology landscape for the European languages (Rehm and Uszkoreit, 2012). Besides general facts about each language, the series describe development in the general language resource and technology areas, as well as in the main application areas of language and speech technology. The series also present a cross-language comparison within four key areas: text analysis, speech and text resources, machine translation, and speech processing. In this report, the Latvian language support in all four key areas was assessed as weak (Skadiņa et al., 2012).

Since the publication of the META-NET White Paper series, the progress and achievements in the key language technology areas for Latvian have been periodically reported through the Baltic HLT conference series and other relevant venues (Skadiņa, 2019; Skadiņa et al., 2016). These reports present notable progress in the rapid

development of language technologies for Latvian, particularly with respect to machine translation, speech recognition and synthesis, natural language understanding and human-computer interaction.

Ten years after the META-NET White Paper series, another pan-European survey was conducted within the European Language Equality (ELE) project[1]. This paper provides an overview of the most recent and significant activities in the language technology area in Latvia, highlighting and elaborating on the findings of the ELE project (Skadiņa et al., 2022). It also provides a broader overview of the key achievements regarding Latvian language resources and technologies since 2012.

2. Language policy and major activities

In general, there is a broad recognition by the research and development community as well as policy makers and government institutions that advancing Latvian language technologies is a critical prerequisite for its survival in the digital age.

Research and development activities in Latvia are supported through different EU and national funding instruments: State research programmes, EU Structural Funds programmes (in particular, through the IT Competence centre projects and the Industry-driven research projects), Latvian Council of Science grants for Fundamental and applied research, EU Horizon 2020, Horizon Europe and CEF programmes.

2.1. National programmes and initiatives

The necessity of language technology support in digital means and importance of language technologies for the long-term survival of the Latvian language has been always recognised in the policy planning documents. However, up to recently, there has been no dedicated language technology research and development program in Latvia. Thus, research and development activities in this area are, in many cases, fragmented and not always sufficiently funded.

Several policy planning documents for 2021–2027 stress the necessity for support of the Latvian language in digital means:

- The State Language Policy Guidelines[2] lists several activities related to the creation and further development of Latvian language resources and tools. Since 2022 the guidelines are being implemented through the three-year State Research Programme “Letonica – Fostering a Latvian and European Society”.
- The Digital Transformation Guidelines for 2021–2027[3] include actions to enable Latvian citizens to access European Digital Space in their native language and to support development of the most important language resources for sustainable and wide use in digital services.

[1] https://european-language-equality.eu
[2] https://likumi.lv/ta/id/325679-par-valsts-valodas-politikas-pamatnostadnem-2021-2027-gadam
[3] https://likumi.lv/ta/id/324715-par-digitalas-transformacijas-pamatnostadnem-20212027-gadam

The information report "On the development of Artificial Intelligence solutions" also lists several directions of action related to the development of AI-based language technologies, such as machine translation, speech technologies, inclusive technologies and terminology databases.

In 2021, Latvia has also approved the Recovery and sustainability plan[4]. Investments are allocated for the development and implementation of high-level skills in three areas: language technology, quantum computing, and HPC. Establishment of Excellence Centre for Language Technology is envisioned to prepare curriculum for language technology teaching, to advance language resources and create platforms and tools for studying and experimentation, and to conduct research involving young researchers. The main research activities planned within this centre are: development of language resources for speech and text processing, creation of large pre-trained language models, advancement of state-of-the-art speech technologies and machine translation, development of software platforms and a shared technical infrastructure for education and research.

2.2. International initiatives

During the last decade, the Latvian language technologies have been part of research, innovation and deployment actions in several FP7, Horizon 2020 and CEF projects on automated translation, speech technologies, human-centred AI, and activities for support digital language equality.

Horizon 2020 projects COMPRISE (Skadiņš and Salimbajevs, 2020), SUMMA[5] and SELMA[6] helped to advance Latvian speech recognition and text analytics technologies, as well as monolingual and cross-lingual intent detection (Kapočiūtė-Dzikienė, 2021).

CEF programme has funded several projects where different language resources and tools were created for Latvian together with several other European languages. Numerous neural machine translation (NMT) engines were developed for translation between Latvian and other EU official languages in the CEF project NTEU (Bié et al., 2020). Domain specific NMT systems were developed in the framework of IADAAPTA project (Castilho, 2019).

Anonymization techniques for Latvian were developed in the CEF programme project MAPA that resulted with an open-source multilingual toolkit for public administrations (Ajausks et al., 2020). New Latvian terminology resources were collected, processed and published in the EuroTermBank database in the CEF projects eTranslation TermBank (Pesliakaitė, 2017) and Federated eTranslation TermBank Network (Lagzdiņš et al., 2022).

[4] https://likumi.lv/ta/id/322858-par-latvijas-atveselosanas-un-noturibas-mehanisma-planu
[5] http://summa-project.eu
[6] https://selma-project.eu

2.3. National initiatives and recent projects

Research and development activities at the national level are mostly supported through three finance instruments: State research programmes, EU Structural Funds programmes, and grants of the Latvian Council of Science.

In 2010, Latvian research institutions and major information technology companies founded the IT Competence Centre (ITCC). The goal of ITCC is to support a long-term cooperation between research organisations and industry to create innovative technologies and prototypes of internationally competitive IT products. Since 2011, more than ten language technology projects have been implemented with support from the ITCC programme (Skadiņa et al, 2016). ITCC supported the creation of the first orthographically and phonetically transcribed Latvian speech corpus, followed by several projects on speech recognition. Several projects addressed machine translation, while others were devoted to intelligent human-computer interaction.

In 2016, the Cabinet of Ministers approved the implementation rules of the Industry-driven research programme of the European Regional Development Fund. In this programme five large projects, mostly implemented through cooperation of research organizations and industry, have been supported. The topics of these projects include: the development of language resources and tools for natural language understanding and generation (Gruzitis et al, 2018), the development of neural network solutions for less resourced languages, multilingual affective human-computer interaction (e.g., Nicmanis and Salimbajevs, 2021), the application of speech technologies for multilingual meeting management, and the transcription of medical speech (Znotiņš et al, 2022).

Funding for the creation, development and maintenance of Latvian language resources and tools has been also received from different national research programmes for Latvian language support. Since 2022, a State Research Programme project "Research on Modern Latvian Language and Development of Language Technology" (LATE) is being implemented aiming to advance research on the grammatical, lexical-semantic, phonetic and phonological systems of the modern Latvian language, and Latvian sign language using data-driven methods, as well as to develop sustainable Latvian language resources and tools.

Since 2018, several projects have been also supported by the Latvian Council of Science, including the development of Latvian Learner Corpus (Dargis et al., 2020a), creation of a pilot Latvian Wordnet and means for neural word-sense disambiguation (Paikens et al, 2022), and research on natural language understanding and generation for human computer interaction (Gosko et al., 2021).

2.4. Infrastructural development

The fundamental support for languages in a digital environment is provided through research infrastructures.

In 2016, Latvia joined European Research Infrastructure Consortium CLARIN (Common Language Resources and Technology Infrastructure).[7] CLARIN-LV is a CLARIN node in Latvia, supporting and collaborating with digital humanities, sharing language resources developed by Latvian academic community, as well as active

[7] https://www.clarin.eu

contributor and participant in international CLARIN activities (Skadiņa et al., 2020). CLARIN-LV mainly focuses on Latvian (and Latgalian) language resources and tools, but not excluding other languages. CLARIN-LV repository of language resources and tools[8] was set up in 2020. The most popular language resources are the open lexical database Tezaurs.lv, Latvian Treebank, Balanced Corpus of Modern Latvian, followed by the NLP pipeline as a service for Latvian – NLP-PIPE. CLARIN-LV is a member of Knowledge Center for Systems and Frameworks for Morphologically Rich Languages SAFMORIL[9] and actively participates in different CLARIN ERIC activities, such as CLARIN Resource Families, Teaching with CLARIN, and CLARIN ParlaMint (Erjavec et al., 2022).

Participants from Latvia are among the core members of European Language Grid (ELG) and European Language Equality (ELE) projects. The objective of the Horizon 2020 project European Language Grid is to address fragmentation in the European language technology business and research landscape by establishing the ELG as the primary platform for language technology in Europe and to strengthen European LT business regarding the competition from other continents (Rehm et al, 2021). Various Latvian language processing tools and resources are already available and can be executed on the ELG platform,[10] including machine translation systems, text-to-speech and speech-to-text tools, POS taggers. ELG catalogue also have a comprehensive list of Latvian language resources that is aggregated from META-SHARE, ELRC-SHARE, CLARIN and other repositories.

Latvia sets an example in making language technology services available for public administrations and general public through national language technology platform Hugo.lv. The broad application and high usage of this platform has inspired other countries like Estonia, Croatia, Iceland and Malta to follow this example. Under leadership of Latvian partners Tilde and Culture Information Systems Centre they have joined forces to create and deploy in their countries similar systems in the framework of CEF programme project National Language Technology Platform (Tadić et al., 2022).

3. Language resources

Since the publication of the META-NET White Paper on Latvian, various new language resources have been created, including advanced datasets and models for natural language understanding, speech recognition and synthesis.

3.1. Text and speech corpora

Text corpora have been developed for Latvian already for several decades. In 2012, monolingual text corpora were already rather well represented, while availability of parallel corpora, treebanks and other kind of annotated corpora was weak. Moreover, speech corpora for Latvian were not available yet.

[8] https://repository.clarin.lv
[9] https://www.clarin.eu/content/safmoril-clarin-knowledge-centre-systems-and-frameworks-morphologically-rich-languages
[10] https://live.european-language-grid.eu

Today, many open-access text corpora are accessible through the Korpuss.lv platform, and most of them are forming the Latvian National Corpora Collection (LNCC) – a diverse collection of corpora representing both written and spoken language (Saulīte et al., 2022). The more than 20 corpora of LNCC (its total size currently exceeds 2 billion tokens) represent different types and genres of Latvian written and spoken language. Although the written language is dominant in LNCC, already three relatively large spoken language text corpora are also available: the orthographic transcription of the balanced general-domain 100-hour speech corpus (Pinnis et al., 2014), the orthographic transcription of the balanced 30-hour medical speech corpus (Darģis et al., 2020b), as well as a large subtitle corpus (10M tokens) of public broadcasting.

LNCC is a continuous multi-institutional and multi-project effort, supported by the digital humanities and language technology communities in Latvia. All corpora of LNCC are annotated with a uniform morpho-syntactic annotation scheme which enables federated search and consistent linguistics analysis in all corpora and allows to select and mix various corpora for pre-training large Latvian language models like BERT. Open-access federated search facility is available through the LNCC website, giving an overview about the absolute and relative frequency of a given search term across all the LNCC corpora.

Modern Latvian is primarily represented through the Balanced Corpus of Modern Latvian LVK2018 (Dargis et al., 2020c), which is being extended to 100 million words. For a balanced subset of LVK2018, syntactic and semantic annotation layers have been added to a various extent (12–17 thousand sentences at the time of writing): Universal Dependencies (UD), named entity and co-reference annotations, frame-semantic annotations (Gruzitis et al, 2018). The multilayer corpus is being enhanced and extended through successive projects, aiming at least at 20k annotated sentences. Notably, the latest release of the UD layer contains nearly 17k sentences. It should also be noted that the Latvian UD treebank has been already classified as a relatively big treebank within the CoNLL 2017 and 2018 shared tasks on UD parsing.

Many corpora are also openly accessible from the Opus platform (Tiedemann, 2016) and the ELRC-SHARE repository[11]. Bilingual and multilingual corpora are also stored at Tilde Data Library[12]. Tilde Data Library includes 12.35 billion parallel sentences and 23.85 billion monolingual sentences in 124 languages. Part of this content is publicly available from the ELRC and ELG platforms, while some of them are also browsable through Hugo.lv – the Latvian State Administration Language Technology Platform. However, domain-specific parallel corpora that would allow training and fine-tuning domain-specific MT engines are lacking.

The first Latvian speech corpus was created in 2012–2013 (Pinnis et al, 2014). The corpus contains 100 hours of transcribed speech, which was a key starting point for the rapid development of speech recognition solutions for Latvian. However, access to this speech corpus is limited, and currently the only open-access Latvian speech corpora are LaRKo and Common Voice Latvian, each of them contain about 8 hours of annotated speech data. In addition, several domain-specific speech corpora have been created (e.g., a medical speech corpus for the radiology domain (Dargis et al., 2020b)).

The development of a general and open-access Latvian language speech corpus has recently started in the National Research Programme's "Letonica – Fostering a Latvian

[11] https://elrc-share.eu/repository/search/

[12] https://www.tilde.com/products-and-services/data-library

and European Society" project "Research on Modern Latvian Language and Development of Language Technology"[13] (LATE). In this project, a balanced open-access speech corpus of at least 100 hours will also be created, as well as a quality speech corpus for text-to-speech synthesis.

Multimodal corpora are still not available for Latvian, although the development of a pilot sign language corpus is also planned in LATE project.

3.2. Lexical resources

Latvian digital lexical resources are being developed already for a long time. Tezaurs.lv is the largest open lexical dataset and on-line dictionary for Latvian (Spektors at al., 2016). The dictionary is popular not only among researchers, but also widely used by the general public: translators, journalists, students and many others, receiving 30–40M requests per year (generated by more than 100k unique users per month). It is regularly updated, and currently contains more than 380,000 lexical entries that are compiled from more than 300 sources.

The development of another important lexical resource, Latvian WordNet, is currently underway. The chosen methodology for word sense splitting and linking is based on corpus evidence and the data from Tezaurs.lv, ensuring a theoretical foundation that has been fine-tuned for both the actual use of Latvian and the linguistic tradition. Furthermore, the links between synonym sets of Latvian WordNet and Princeton WordNet are also being added (Paikens et al., 2022).

Different lexicons (mostly bilingual) are available from the Letonika.lv portal. It contains electronic dictionaries for widely used language pairs (Latvian and English, French, German and Russian), as well as dictionaries of the languages of the Baltic countries: Latvian and Lithuanian, Latvian and Estonian.

Latvian terminology is consolidated in the European Terminology Bank (Rirdance and Vasiļjevs, 2006)[14] and the Latvian national terminology portal.[15] Today, EuroTermBank contains about 3.5 million entries (14.5 million terms) from 463 collections in 44 languages. As part of the CEF project FedTerm, Latvian national terminology portal is integrated with EuroTermBank in a federated network that interlinks terminology portals from different European countries.

4. Tools, technologies and applications

4.1. Text analysis tools

Various basic text processing tools, such as tokenizers and sentence splitters, morphological analysers and taggers, spelling and grammar checkers have been available for Latvian for several decades. Spelling and grammar checking tools are available for users through Microsoft and Tilde products, as well as some open-source text processors. Various open-source Latvian NLP tools are integrated into NLP--PIPE:

[13] http://www.digitalhumanities.lv/projects/vpp-late/
[14] https://www.eurotermbank.com
[15] https://termini.gov.lv

a modular pipeline for text tokenisation and sentence splitting, morphological tagging, named entity recognition, syntactic dependency parsing, etc. (Znotins and Cirule, 2018).

With introduction of large pre-trained language models that can be fine-tuned for different NLP tasks several part-of-speech taggers, dependency parsers (Znotiņš and Bārzdiņš, 2020) and named entity recognizers (Vīksna and Skadiņa, 2020) have been developed. Finally, several sentiment analysis tools have been created as well.

4.2. Natural language understanding (NLU) and generation (NLG)

Regarding NLU and NLG, apart from neural transformer encoder, transformer decoder and transformer encoder-decoder models for Latvian, experiments with the interlingual knowledge-based representations, namely FrameNet, Abstract Meaning Representation (AMR) and Grammatical Framework, have also been conducted for Latvian, English and other languages. This demonstrates the expertise and potential of combining machine learning and knowledge-based approaches for state-of-the-art NLG for both high-resourced and less-resourced languages for practical use cases when predictability and precision is as important as fluency and scalability (Ranta et al., 2020).

4.3. Machine translation

With respect to machine translation (MT), the situation for the Latvian language has changed considerably comparing to 2012. Today, most global companies, which offer machine translation services, support also Latvian (e.g., Google Translate[16], Bing Microsoft Translator[17], DeepL[18], Amazon Translate[19], Watson Language Translator[20], and others). In 2017, Latvian was included as a competition language in the shared task of news translation of the Conference on Machine Translation[21]. Neural machine translation (NMT) systems that were developed by Tilde were recognised among the best systems (Bojar et al., 2017). Based on these results Tilde together with partners who provided language resources developed the EU Council Presidency Translator, which has already been used in 8 countries (Pinnis et al., 2020, 2021). However, not having enough data for various narrower domains still limits development of MT engines for specific domains and lesser resourced languages like Latvian.

In 2012, the dominant machine translation paradigm was phrase-based statistical machine translation (SMT). However, since late 2016, the state-of-the-art paradigm is neural machine translation (Bojar et al., 2016). This paradigm shift has impacted machine translation research also for Latvian. Prior to 2016, work focused on improving MT quality for Latvian with the use of external morphological taggers or parsers (e.g., Skadiņš et al., 2010), domain-specific terminology (e.g., Pinnis, 2015), domain adaptation (e.g., Pinnis et al., 2013; Pinnis and Skadiņš, 2012), and other methods. Although such methods are also relevant nowadays, the paradigm shift required

[16] https://translate.google.com
[17] https://bing.com/translator
[18] https://www.deepl.com/translator
[19] https://aws.amazon.com/translate
[20] https://www.ibm.com/cloud/watson-language-translator
[21] https://statmt.org/wmt17/translation-task.html

complete rework of these methods since NMT differs substantially from SMT. The shift to NMT has also spurred research in areas that are specifically important for NMT, such as input representations for neural networks that would improve word splitting consistency for morphologically rich languages (such as Latvian) (Pinnis et al., 2017), NMT system robustness (e.g., Bergmanis et al., 2020), and others.

Recent NMT research in Latvia and for Latvian has been focused on the following topics: terminology integration in NMT (Bergmanis and Pinnis, 2021a, 2021b), analysis of biases in NMT systems (Stafanovičs et al., 2020), robustness of NMT systems (Bergmanis et al., 2020), speech translation (Alves et al., 2020), and others.

4.4. Speech technology

For many years, speech technology support for Latvian was almost non-existent due to the lack of data for training speech recognition and synthesis models. Shortly after the transcribed 100-hour corpus of spoken Latvian was created, several automatic speech recognition (ASR) systems were developed (Salimbajevs and Strigins, 2015; Znotins et al., 2015). Today, the accuracy of these systems is comparable to the state of the art.

General-purpose Latvian speech synthesisers and recognisers developed by Tilde[22] and IMCS UL[23] are publicly available and are constantly advanced with new features (Nicmanis and Salimbajevs, 2021). Domain-specific speech transcription systems are also being developed, most notably for the medical domain: IMCS UL together with Riga East University Hospital have developed an ASR system with the focus on radiological and histopathological examination reports, as well as medical case histories (Dargis et al, 2020b; Znotiņš et al., 2022). Tilde has researched methods for adaptation of ASR to medical domain with untranscribed audio (Salimbajevs and Kapočiūte-Dzikiene, 2022), and together with Children's Clinical University Hospital developed an ASR system focusing on psychiatry, paediatrics and radiology.

There is an ongoing work on different online solutions encompassing Latvian speech recognition. That includes live event transcription for people with hearing impairments, captioning of video recordings, live transcription of online video meetings[24].

4.5. Human-computer interaction

With the renaissance of AI and availability of computational resources that have made deep learning techniques applicable to natural language processing tasks, the human-computer interaction with help of virtual assistants has become actual topics again.

Today several task-oriented virtual assistants, which help users finding answers to their questions, can communicate in Latvian. Virtual assistants are also used by public services. For example, at the time of writing, Hugo.lv[25] lists 15 virtual assistants for different public administration services, including the Latvian State Radio and Television Centre, the Courts Administration, the Bank of Latvia, the Rural Support Service, and many others. These have been developed using the capabilities of the

[22] https://www.tilde.lv/tildes-balss
[23] https://selma-project.github.io
[24] https://speech.tilde.ai
[25] https://hugo.lv

conversational AI platform tilde.ai. It allows users to create their own virtual conversational agents for specific tasks. These agents support both text and voice input, can recognise intents expressed in the input, and deliver response using text, visual media, or voice modalities.

However, natural language understanding is still not solved problem and thus a lot of work needs to be done to create technologies for deeper language understanding and human-computer interaction. Several steps to this direction have been already made by researching intent detection (Balodis and Deksne, 2019; Kapočiūtė-Dzikienė et al., 2021), slot filling (Gosko et al, 2021) and next dialogue action prediction techniques (Deksne and Skadiņš, 2020, 2021; Skadiņa and Goško, 2020).

5. Conclusion

We have provided an overview of the current state of the Latvian language in the digital environment. Since the publication of the META-NET White Papers, notable progress has been made in the development of various language resources and tools for Latvian.

Although the Latvian language is used by a rather small number of speakers, and it is often categorised as less resourced, it is represented rather well not only by different language resources (digital libraries, text and speech corpora, lexicons, etc.) but also by core language technologies. Concerning more advanced technologies, Latvian has a reasonably good support for machine translation, speech recognition and synthesis, while solutions that involve deep state-of-the-art natural language understanding, like virtual assistants and text summarisers, are less developed.

There are still significant gaps with respect to availability, size and technology readiness level (TRL) of language resources. With respect to language resources, significant gaps are identified for both monolingual and multilingual data of all forms: written, spoken and multimodal. For example, datasets that represent conversational data, question answering, knowledge bases, informal language or specific domain are small or non-existent. There are almost no spoken and multimodal open-data available.

Domain-specific parallel and multilingual data that would allow training and fine-tuning MT engines are insufficient, while the current open-access monolingual text corpora are too small for training massive language models like GPT-3. Consequently, there is a lack of large pre-trained language models (both general and domain-specific) and lack of benchmarks for specific NLP tasks, e.g., Latvian GLUE or SQUAD.

Creation of such models is limited not only by availability of necessary data but also by insufficient hardware infrastructure, which could be solved through significant long-term support for research infrastructures.

Another important aspect is IPR and GDPR regulations that need to be more flexible, allowing wider use of IPR protected data for the development of language resources and technologies in a way that does not harm the interests of the authors.

Overall, similarly to many other languages of Europe, there is insufficient amounts of quality corpora, including monolingual corpora, currently available for Latvian, as well as insufficient computational resources, for training large-scale SOTA language models like GPT-3. However, there are resources and competence available for pre-training relatively smaller language models like BERT and GPT-2, and for fine-tuning large pre-trained multilingual models like mT5 and XLS-R for various downstream tasks.

Limited availability of human resources leads to gaps and limitations also in language technology development. Although the Latvian LT industry and research

groups have demonstrated excellent results in LT adaption for morphologically rich languages (which is not a trivial task), they are less present among leaders in the development of world-class novel language technology solutions.

Finally, gaps and fragmentation in research and development activities related to LT is a result of short, project-based (mostly 2–3 years, sometimes even 1 year, rarely 5 years) research and development funding and disproportion between funding for research (TRL 1–4) and industrial activities (TRL 5–8).

With respect to policies and instruments, strong national and international support is necessary for further Latvian language research and development activities, including dedicated long-term LT programs that provide equal support for both research and industrial activities. Close synchronisation between national and international activities is necessary, especially, with respect to research infrastructures and research priorities.

Acknowledgements

This study has been supported by the European Language Equality project funded from the EU under the grant agreement № LC-01641480 – 101018166, the National Research Programme project "Digital Resources for Humanities: Integration and Development" (VPP-IZM-DH-2020/1-0001), and the European Regional Development Fund projects "University of Latvia and institutes in the European Research Area – Excellency, activity, mobility, capacity" (1.1.1.5/18/I/016), "AI Assistant for Multilingual Meeting Management" (1.1.1.1/19/A/082), and "Latvian Speech Recognition and Synthesis for Medical Applications" (1.1.1.1/18/A/153).

References

Ajausks, Ē., Arranz, V., Bié, L., Cerdà-i-Cucó, A., Choukri, K., Cuadros, M., ..., Zweigenbaum, P. (2020). The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 471–472.

Alves, D., Salimbajevs, A., Pinnis, M. (2020). Data augmentation for pipeline-based speech translation. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 73–79.

Balodis, K., Deksne, D. (2019). FastText-Based Intent Detection for Inflected Languages. *Information*. 10 (5), 161, pp. 1–16.

Bergmanis, T., Stafanovičs, A., Pinnis, M. (2020). Robust Neural Machine Translation: Modeling Orthographic and Interpunctual Variation. *Human Language Technologies–The Baltic Perspective*, pp. 80–86.

Bergmanis, T., Pinnis, M. (2021a). Facilitating Terminology Translation with Target Lemma Annotations. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3105–3111.

Bergmanis, T., Pinnis, M. (2021b). Dynamic Terminology Integration for COVID-19 and Other Emerging Domains. *Proceedings of the Sixth Conference on Machine Translation*, pp. 821–827.

Bié, L., Cerdà-i-Cucó, A., Degroote, H., Estela, A., García-Martínez, M., Herranz, M., ... Vasiļevskis, A. (2020). Neural Translation for the European Union (NTEU) Project. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 477–478.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., … Zampieri, M. (2016). Findings of the

2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 169–214.

Castilho, S., Resende, N., Gaspari, F., Way, A., O'Dowd, T., Mazur, M., ... Šics, V. (2019). Large-scale machine translation evaluation of the iADAATPA Project. *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pp. 179–185.

Darģis, R., Auziņa, I., Levāne-Petrova, K., Kaija I. (2020a). Quality Focused Approach to a Learner Corpus Development. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 392–396.

Darģis, R., Grūzitis, N., Auziņa, I., Stepanovs, K. (2020b). Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 135–141.

Darģis, R., Levāne-Petrova, K., Poikāns, I. (2020c). Lessons Learned from Creating a Balanced Corpus from Online Data. *Human Language Technologies – The Baltic Perspective*, IOS Press, pp.127–134.

Deksne, D., Skadiņš, R. (2020). Interactive Learning of Dialog Scenarios from Examples. *Frontiers in Artificial Intelligence and Applications: Human Language Technologies – The Baltic Perspective. Proceedings of the Ninth International Conference Baltic HLT 2020*, IOS Press, pp. 87–94.

Deksne, D. and Skadiņš, R. (2021). Predicting Next Dialogue Action in Emotionally Loaded Conversation. *Proceedings of the Future Technologies Conference*, LNNS, 358, pp. 264–274.

Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, pp. 1–34.

Goško, D., Znotiņš, A., Skadiņa, I., Grūzītis, N., Nešpore-Bērzkalne G. (2021). Domain Expert Platform for Goal-Oriented Dialog Collection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pp. 295–301.

Grūzītis, N. Pretkalniņa, L., Saulīte, B., Rituma, L., Nešpore-Bērzkalne, G., Znotiņš, A., Paikens, P. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 4506–4513.

Kapočiūtė-Dzikienė, J., Salimbajevs, A., Skadiņš, R. (2021). Monolingual and cross-lingual intent detection without training data in target languages. *Electronics*, **10**(12), 1412, pp. 1–24.

Krišlauks, R., Pinnis, M. (2020). Tilde at WMT 2020: News Task Systems. *Proceedings of the Fifth Conference on Machine Translation*, pp. 175–180.

Lagzdiņš, A., Siliņš, U., Pinnis, M., Bergmanis, T., Vasiļevskis, A., Vasiļjevs, A. (2022). Open Terminology Management and Sharing Toolkit for Federation of Terminology Databases. *Proceedings of the 13th International Conference on Language Resources and Evaluation* (LREC 2022), pp. 6310-6316.

Nicmanis, D., Salimbajevs. A. (2021). Expressive Latvian Speech Synthesis for Dialog Systems. *Proceedings of Interspeech Show & Tell Contribution*, pp. 3321–3322.

Paikens, P., Grasmanis, M., Klints, A., Lokmane, I., Pretkalniņa, L., Rituma, L., Stāde, M. and Strankale, L. (2022). Towards Latvian WordNet. *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pp. 2808–2815.

Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation. *Proceedings of the Eighteenth Annual Conference of the European Association for Machine Translation* (EAMT 2015), pp. 89–96.

Pinnis, M., Auziņa, I., Goba, K. (2014). Designing the Latvian speech recognition corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pp. 1547–1553.

Pinnis, M., Bergmanis, T., Metuzāle, K., Šics, V., Vasiļevskis, A., Vasiļjevs, A. (2020). A tale of eight countries or the EU council presidency translator in retrospect. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, Volume 2: User Track*, pp. 525–546.

Pinnis, M., Busemann, S., Vasiļevskis, A., van Genabith, J. (2021). The German EU Council Presidency Translator. *KI-Künstliche Intelligenz*, (36), pp. 99–104.

Pinnis, M., Krišlauks, R., Deksne, D., Miks, T. (2017). Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. *Proceedings of the 20th International Conference of Text, Speech and Dialogue* (TSD), 10415 LNAI, pp. 237–245.

Pinnis, M., Skadiņa, I., Vasiļjevs, A. (2013). Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English Latvian IT Localisation. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING), pp. 224–235.

Pinnis, M., Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. *Human Language Technologies–The Baltic Perspective. Proceedings of the 5th International Conference Baltic HLT*, Vol. 247, pp. 176–184.

Pesliakaitė, E. (2017). eTranslation TermBank–naujas tarptautinis projektas Europos Komisijos automatinio vertimo kokybei pagerinti. *Terminologija*, (24), pp. 205–209.

Ranta, A., Angelov, K., Gruzitis, N., Kolachina, P. (2020). Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2), pp. 425–486

Rehm, G., Piperidis, S., Bontcheva, K., Hajic, J., Arranz,V., Vasiļjevs, A., Backfried, G. et al. (2021). European Language Grid: A Joint Platform for the European Language Technology Community. *Proceedings of the 16th Conference of the European Chapter of ACL: System Demonstrations*, pp. 221–230.

Rehm, G., Uszkoreit, H. (Eds.) (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Springer.

Rirdance, S., Vasiļjevs A. (Eds.) (2006). *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project*. (ISBN 9984-9133-4-1) EuroTermBank Consortium.

Salimbajevs, A., Strigins, J. (2015). Latvian Speech-To-Text Transcription Service. *Proceedings of Interspeech 2015*, pp. 722–723.

Salimbajevs, A., Kapociute-Dzikiene, J. (2022). Automatic Speech Recognition Model Adaptation to Medical Domain Using Untranscribed Audio. *Digital Business and Intelligent Systems*, Springer, Cham, pp. 65-79.

Saulīte, B., Darģis, R., Grūzītis, N., Auziņa, I., Levāne-Petrova, K., Pretkalniņa, L., Rituma, L., Paikens, P., Znotiņš, A., Strankale, L., Pokratniece, K., Poikāns, I., Bārzdiņš, G., Skadiņa, I., Baklāne, A., Saulespurēns, V., Ziediņš, J. (2022). Latvian National Corpora Collection – Korpuss.lv. *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pp. 5123–5129.

Skadiņa, I., Auziņa, I., Valkovska, B., Grūzītis, N. (2022). *D1.22*. Report on Latvian. *Deliverable of the project European Language Equality*. ELE Consortium.

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I., Rudzīte, A. (2012). *The Latvian Language in the Digital Age*. Springer.

Skadiņa, I. (2019) Some Highlights of Human Language Technology in Baltic Countries. *Databases and Information Systems X*, IOS Press, pp. 18–30.

Skadiņa, I., Auziņa, I., Deksne, D., Skadiņš, R., Vasiļjevs, A., Gailuna, M., Portnaja I. (2016). Filling the gaps in Latvian BLARK: Case of the Latvian IT Competence Centre. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 3–11.

Skadiņa, I., Auziņa, I., Grūzītis, N., Znotiņš, A. (2020). Clarin in Latvia: From the preparatory phase to the construction phase and operation. *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pp. 342–350.

Skadiņa, I. Goško, D. (2020). Towards Hybrid Model for Human-Computer Interaction in Latvian. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 103–110.

Skadiņš, R., Goba, K., Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 125–132.

Skadiņš, R., Salimbajevs, A. (2020). The COMPRISE Cloud Platform. *Proceedings of 1st International Workshop on Language Technology Platforms*, pp. 108–111.

Spektors, A., Auziņa, I., Darģis, R., Grūzītis, N., Paikens, P., Pretkalniņa, L., Rituma, L., Saulīte, B. (2016). Tezaurs.lv: the largest open lexical database for Latvian. *Proceedings of the 10th International LREC Conference*, pp. 2568–2571.

Stafanovičs, A., Bergmanis, T., Pinnis, M. (2020). Mitigating Gender Bias in Machine Translation with Target Gender Annotations. *Proceedings of the Fifth Conference on Machine Translation*, pp. 629–638.

Tadić, M., Farkaš, D., Filko, M., Vasiļevskis, A., Vasiļjevs, A., Ziediņš, J., Motika, Ž., Fishel, M., Loftsson, H., Guðnason, J., Borg, C. (2022). National Language Technology Platform for Public Administration. *Proceedings of the LREC 2022 Workshop Language Resources and Evaluation Conference*, pp. 46-51.

Tiedemann, J. (2016). OPUS – Parallel Corpora for Everyone. *Baltic Journal of Modern Computing (BJMC), 4(2), Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT)*, pp. 384.

Vīksna, R., Skadiņa, I. (2020). Large Language Models for Latvian Named Entity Recognition. Frontiers in Artificial Intelligence and Applications, volume 328: Human Language Technologies–The Baltic Perspective, pp. 62-69.

Znotins, A., Polis, K., Dargis, R. (2015). Media monitoring system for Latvian radio and TV broadcasts. Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH).

Znotiņš, A., Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 183–189.

Znotiņš, A., Darģis, R., Grūzītis, N., Bārzdiņš, G., Goško, D. (2022). RUTA:MED – Dual Workflow Medical Speech Transcription Pipeline and Editor. *Natural Language Processing and Information Systems*, LNCS vol. 13286, Springer, pp. 209–214.

Znotiņš, A., Bārzdiņš, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. *Human Language Technologies–The Baltic Perspective*, IOS Press, pp. 111–115.

Received August 21, 2022, accepted September 5, 2022