Chapter 24
Language Report Latvian

Inguna Skadiņa, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis

Abstract Ten years ago, when META-NET conducted a study on Language Technology support for Europe's languages, Latvian was assessed as a language with little or no support (Skadiņa et al. 2012). During the last decade, progress has been made in the development of language resources and tools for Latvian, particularly with respect to advanced datasets and language models, machine translation solutions, speech technologies, and technologies for natural language understanding and human-computer interaction. This chapter provides a summary of the current state of the Latvian language, the only official language of Latvia, in the digital environment and highlights the most important activities in the language technology field.

1 The Latvian Language

Latvian is the official language of the Republic of Latvia. There are about 1.5 million native speakers, 1.38 million of which live in Latvia. By the end of 2017, Latvian was the mother tongue of 60.8% of the country's resident population. Latvian is spoken as a second language by around 0.5 million people of other ethnicities. Latvian has three dialects: the Central, Livonic, and High Latvian dialect.

The Latvian language uses the phono-morphological basis of orthography. Latvian punctuation is based on the grammatical punctuation principle. Latvian orthography almost fully corresponds to the pronunciation. The present-day Latvian orthography basis is the Latin script. The Latvian standard alphabet consists of 33 letters, including letters with diacritical marks.

Standard Latvian has 26 consonant phonemes, 12 vowels (six short and six long), and 10 diphthongs. Vowel length is phonemic and plays an important role in distinguishing the lexical and grammatical meaning of words. Most Latvian words are stressed on the first syllable. Syllables with long vowels, diphthongs, and diphthongical combinations of vowel and sonorant in the centre are subject to certain intonation

Inguna Skadiņa · Ilze Auziņa · Baiba Valkovska · Normunds Grūzītis
University of Latvia, Latvia, inguna.skadina@tilde.com, ilze.auzina@lumii.lv, baiba.valkovska@lumii.lv, normunds.gruzitis@lumii.lv

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_24

patterns. In a few areas, three patterns of tone or intonation are distinguished: level (also drawling, even) tone, falling tone, and broken tone.

From a language typology perspective, Latvian has a classic Indo-European (Baltic) system. However, for regional and historical reasons, Latvian grammar also displays some features more similar to those found in Finno-Ugric languages (Kalnaca and Lokmane 2021). Latvian is a fusional, mainly suffixing language with a rich system of forms and word formation. A distinction is made between inflected and non-inflected word classes. Nouns inflect for number and case, adjectives inflect for case, number, gender and definiteness, and verbs may inflect for tense, mood, voice and person (Nau 1998). Word order is relatively free, i. e., pragmatically governed, but the basic word order is subject verb object (SVO).

2 Technologies and Resources for Latvian

Research and development activities in Latvia are being supported through different EU and national finance instruments and are usually organised around short-term projects. The lack of a dedicated LT programme, however, leads to fragmentation of research and development activities and complicates the development of larger resources and long-term cooperation between institutions. Progress and key achievements are regularly reported through the Baltic HLT conferences and other events (Skadiņa 2019; Skadiņa et al. 2022, provide recent overviews).

Most open-access monolingual text corpora for Latvian are listed on Korpuss.lv (Saulīte et al. 2022). Modern Latvian is primarily represented through the Balanced Corpus of Modern Latvian (LVK2018, Dargis et al. 2020). A balanced subset of LVK2018 includes several annotation layers: named entities, co-references, Universal Dependencies (UD), FrameNet and PropBank annotations, as well as Abstract Meaning Representation (AMR) (Gruzitis et al. 2018). Many parallel corpora are openly accessible from OPUS, ELG and ELRC-SHARE. Bilingual and multilingual corpora are also stored on Korpuss.lv and the Tilde Data Library.[1] Domain-specific parallel corpora for the development of domain-specific MT engines are lacking.

The first Latvian speech corpus was created in 2012/2013. It contains 100 hours of transcribed speech. However, access is limited, and currently the only open-access Latvian speech corpora are very small. Multimodal corpora are still not available for Latvian, although the development of a sign language corpus is ongoing in the State Research Programme “Letonika”.

Tezaurs.lv is the largest open lexical dataset and online dictionary for Latvian (Spektors et al. 2016). It is regularly updated, and currently contains more than 380k single- and multi-word entries, compiled from 300+ sources. A Latvian WordNet is being created as an extension to Tezaurs.lv. Different lexicons (mostly bilingual) are available from the Letonika.lv portal, including dictionaries for widely used language pairs, as well as dictionaries of the languages of the Baltic countries.

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

Various text analysis tools such as tokenisers and sentence splitters, morphological analysers and taggers, spelling and grammar checkers, syntactic and semantic parsers, named entity recognisers, and text classifiers are available for Latvian. Open-source components are integrated into a Latvian NLP pipeline as a service.[2]

Regarding natural language understanding and generation, experiments with the interlingual UD, FrameNet, AMR, BERT, GPT, etc. models for Latvian demonstrate the potential of combining machine learning and knowledge-based approaches.

With respect to machine translation (MT), the situation has changed a lot since 2012. Besides MT solutions provided by global companies, the company Tilde provides customised MT solutions for complex, highly inflected languages, particularly smaller European languages. MT systems developed by Tilde have been recognised among the best systems for four consecutive years (2017-2020) at WMT international news translation shared tasks (Pinnis et al. 2019). These results allowed Tilde together with partners to develop the EU Council Presidency Translator which has been used already in eight countries (Pinnis et al. 2020).

Several speech recognition and synthesis systems have been developed for Latvian by Tilde, the national news agency LETA, and the University of Latvia. Several virtual assistants can communicate in Latvian, e. g., Hugo.lv (Skadins et al. 2020) lists more than 10 virtual assistants for different public services.

Latvia is a member of CLARIN (Skadiņa et al. 2020) and focuses on Latvian and Latgalian resources and tools. CLARIN-LV participates in the CLARIN Knowledge Center for Systems and Frameworks for Morphologically Rich Languages.

3 Recommendations and Next Steps

Today, Latvian has a rather stable position in the digital world. However, the situation could change dramatically, if efforts and investments in LT are not increased in R&D and language policy. Strong national and European support is necessary for further Latvian research and development activities, including dedicated long-term LT programmes, that provide equal support for both research and industrial activities. To narrow the digital divide, there is pressing urgency for novel techniques that would bring less-resourced languages to a level comparable to the state-of-the-art results for resource-rich languages. Moreover, close synchronisation between national and international activities is necessary.

References

Dargis, Roberts, Kristine Levane-Petrova, and Ilmars Poikans (2020). “Lessons Learned from Creating a Balanced Corpus from Online Data”. In: *Human Language Technologies – The Baltic Perspective*. Vol. 328. IOS Press, pp. 127–134. DOI: 10.3233/FAIA200614.

[2] http://nlp.ailab.lv

Gruzitis, Normunds, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens (2018). "Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 4506–4513.

Kalnaca, Andra and Ilze Lokmane (2021). *Latvian Grammar*. University of Latvia.

Nau, Nicole (1998). *Latvian*. Vol. 217. Lincom Europa.

Pinnis, Mārcis, Toms Bergmanis, Kristīne Metuzāle, Valters Šics, Artūrs Vasiļevskis, and Andrejs Vasiļjevs (2020). "A Tale of Eight Countries or the EU Council Presidency Translator in Retrospect". In: *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2)*. Association for Machine Translation in the Americas, pp. 525–537.

Pinnis, Mārcis, Rihards Krislauks, and Matiss Rikters (2019). "Tilde's Machine Translation Systems for WMT 2019". In: *Proceedings of the 4th Conference on Machine Translation (Volume 2)*. Florence, Italy: ACL, pp. 327–334. DOI: 10.18653/v1/W19-5335.

Saulīte, Baiba, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, and Valdis Saulespurēns (2022). "Latvian National Corpora Collection – Korpuss.lv". In: *Proceedings of the 13th LREC Conference*.

Skadiņa, Inguna (2019). "Some Highlights of Human Language Technology in Baltic Countries". In: *Databases and Information Systems*. Vol. 315. IOS Press, pp. 18–30. DOI: 10.3233/978-1-61499-941-6-18.

Skadiņa, Inguna, Ilze Auziņa, Normunds Grūzītis, and Arturs Znotiņš (2020). "Clarin in Latvia: From the preparatory phase to the construction phase and operation". In: *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries*, pp. 342–350.

Skadiņa, Inguna, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis (2022). *Deliverable D1.22 Report on the Latvian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-latvian.pdf.

Skadiņa, Inguna, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte (2012). *Latviešu valoda digitālajā laikmetā – The Latvian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/latvian.

Skadins, Raivis, Marcis Pinnis, Arturs Vasilevskis, Andrejs Vasiljevs, Valters Sics, Roberts Rozis, and Andis Lagzdins (2020). "Language Technology Platform for Public Administration". In: *Human Language Technologies – The Baltic Perspective*. Ed. by Utka Andrius, Vaicenoniene Jurgita, Kovalevskaite Jolantai, and Kalinauskaite Danguole. Vol. 328. FAIA. IOS Press, pp. 182–190.

Spektors, Andrejs, Ilze Auzina, Roberts Dargis, Normunds Gruzitis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, and Baiba Saulite (2016). "Tezaurs.lv: the largest open lexical database for Latvian". In: *Proceedings of LREC 2016*.