UNIVERSITY OF
LATVIA

Summary of
Doctoral Thesis

Roberts Darģis

INFRASTRUCTURE FOR LATVIAN CORPORA DEVELOPMENT AND APPLICATION

Riga 2025

UNIVERSITY OF
LATVIA

FACULTY OF SCIENCE AND TECHNOLOGY

Roberts Darģis

INFRASTRUCTURE FOR LATVIAN CORPORA DEVELOPMENT AND APPLICATION

SUMMARY OF THE DOCTORAL THESIS

Submitted for the degree of Doctor of Science (Ph.D.)
Field of Science – Computer Science and Informatics
Subfield – Computer and Systems Software

Riga 2025

This doctoral thesis was developed at the University of Latvia, Institute of Mathematics and Computer Science, Artificial Intelligence Laboratory over the period of time from 2016 to 2025.

Thesis form: a set of publications in the field of computer science and informatics, the subfield of computer and system software.

Thesis supervisor: Dr. sc. comp. prof. **Guntis Barzdins**, University of Latvia.

Reviewers:

1. Dr. sc. comp. **Karlis Cerans**, University of Latvia;
2. Dr. sc. comp. **Matiss Rikters**, National Institute of Advanced Industrial Science and Technology, Japan;
3. Dr. sc. comp. **Jurgita Kapociute-Dzikiene**, Vytautas Magnus University, Lithuania.

The thesis will be defended at the public session of the Doctoral Committee of Computer science and informatics, and Electrical engineering, electronics, information and communication technology, University of Latvia, at 15.00 on 17th of October, 2025, Raina blvd. 19.

The thesis is available at the Library of the University of Latvia, Raina blvd. 19.

This thesis is accepted for the commencement of the degree of Doctor of Natural Sciences on 17th of October, 2025 by the Doctoral Committee of Computer science and informatics, and Electrical engineering, electronics, information and communication technology, University of Latvia.

Chairman of the Doctoral Committee Dr. sc. comp. **Inguna Skadina**

Secretary of the Doctoral Committee **Sintija Silina**

ISBN 978-9934-36-426-6
ISBN 978-9934-36-427-3 (PDF)

Abstract

This thesis presents research on the development and application of an infrastructure designed for Latvian corpora. It addresses the need for structured and accessible linguistic data to support both natural language processing (NLP) and corpus-based research in digital humanities, political science, and linguistics. The research focuses on developing tools and methodologies for corpus creation, annotation, and accessibility.

The thesis is structured around thematically-related collection of publications, multiple projects and real-world implementations, demonstrating the effectiveness of the proposed infrastructure. The results have significant implications for NLP applications, language modeling, and interdisciplinary academic research.

Contents

1. Introduction

In the rapidly evolving landscape of natural language processing (NLP) and digital humanities, high-quality, language-specific corpora have become a cornerstone for both academic research and technological innovation. Despite significant advances in NLP for well-resourced languages, low-resource languages such as Latvian continue to lag behind, constrained by the limited availability of structured and annotated linguistic data.

1.1 Relevance of the Research Problem

Over the past decade, machine learning has significantly transformed the field of natural language processing. This rapid progress has been driven in part by the emergence of versatile machine learning tools.

The research in this thesis is motivated by the observation that while many machine learning tools are designed for multilingual use, they require extensive and high-quality linguistic data for effective performance.

By focusing on the development of an infrastructure tailored to Latvian corpora, the research fills a significant gap. It not only promotes the integration of the Latvian language into modern NLP tools but also facilitates interdisciplinary studies in computational linguistics, political science, and digital humanities.

1.2 Aim and Objectives of the Research

The primary aim of this research is to foster the seamless integration of the Latvian language into modern NLP tools and corpus-based studies. To reach this main goal, the research focus specific objective.

- **Develop a Tailored Infrastructure:** Design and implement a dedicated infrastructure that streamlines the creation, annotation, and management of Latvian corpora. This includes automated pipelines and intuitive interfaces that minimize manual labor while ensuring high data quality and consistency.
- **Construct Structured and Interoperable Corpora:** Build diverse corpora featuring multi-layered, standardized annotation schemes. This objective ensures that the linguistic data is easily accessible and compatible with international formats, thereby enhancing its utility for both linguistic research and advanced NLP applications.
- **Establish Robust Digital Tools:** Create user-friendly digital platforms that facilitate corpus-based research. These tools are intended to empower researchers in the digital humanities and political sciences by enabling efficient

exploration, analysis, and integration of Latvian language data into broader research and application frameworks.

1.3 Research Hypotheses

- Automated annotation tools and structured data pipelines will reduce the manual effort required for corpus creation while maintaining high accuracy.
- The availability of annotated corpora will improve the development and performance of NLP models tailored for the Latvian language.
- A structured and easily accessible infrastructure for Latvian corpora will significantly enhance corpus-based research methodologies across various disciplines.

1.4 Research Methods

The research methods applied in this thesis were carefully selected to meet the objectives set forth in developing an infrastructure and tools for Latvian corpora creation, annotation, and application. The approach combined both qualitative and quantitative techniques to ensure that the developed tools and procedures were not only technically sound but also practically effective in real-world applications.

- **Literature Review:** comprehensive examination and analysis of existing scientific literature and available technologies were conducted to establish a foundational understanding of corpus development, annotation techniques, and natural language processing frameworks. This review provided a fundamental understanding of the state-of-the-art practices and helped to identify gaps in current methodologies and to inform the design of novel methods tailored to the linguistic characteristics of Latvian.
- **Quantitative and Qualitative Evaluation:** the performance of the infrastructure was rigorously assessed using established metrics. Key performance indicators, such as annotation accuracy, processing efficiency, and error reduction, were measured to determine the effectiveness of the automated processes compared to manual methods. This quantitative analysis provided a clear basis for evaluating improvements and benchmarking against existing solutions. Complementarily, qualitative assessments were conducted, including error analyses and manual verification, to ensure linguistic accuracy and data consistency.
- **Controlled Experiments:** several controlled experiments were designed to assess the impact of individual components within the infrastructure. By comparing variants of the automated annotation and data transformation algorithms in a controlled setting, the research identified the most effective approaches for handling linguistic nuances and minimizing errors.
- **Case Studies and Real-world Implementations:** real-world case studies were implemented across multiple projects such as parliamentary corpora,

learner corpora, and speech corpora. These practical applications validated the robustness and general applicability of the infrastructure across various domains, including linguistics, digital humanities, and NLP applications.
- **Iterative Development and Adaptation:** the design and implementation of the corpus infrastructure were carried out through an iterative process. Prototypes were developed, tested, and refined continuously based on pilot evaluations and expert feedback. This cycle of development allowed for the rapid incorporation of practical insights and ensured that the system evolved in response to both technical challenges and user requirements.

Together, these research methods provided a robust framework for the development, evaluation, and continuous improvement of an integrated infrastructure that supports the creation, annotation, and application of Latvian corpora.

1.5 Main Results

This thesis presents a comprehensive set of research findings that significantly enhance the development, management, and application of Latvian corpora.

- **Automated Annotation and Data Pipelines:** A comprehensive infrastructure has been designed and implemented to streamline the processes of corpus creation, annotation, and management. The development of tailored tools and methods reduces the need for manual labor and ensures high data quality.
- **Development of Diverse Latvian Corpora:** Several specialized corpora have been developed as part of this research, including parliamentary corpora, error-annotated corpora, and various speech corpora. These resources fill existing gaps in Latvian language resources and open up new opportunities for research in the digital humanities.
- **Integration of Diverse Corpora:** Multiple corpora from various domains have been standardized and integrated into a unified infrastructure – Korpuss.lv. This integration facilitates advanced natural language processing applications and supports extensive corpus-based research in digital humanities and linguistics.

These results represent a substantial research conducted at the Artificial Intelligence Laboratory of the Institute of Mathematics and Computer Science (IMCS) at the University of Latvia (UL). The research is the outcome of collaborative efforts, with the author playing a major role in achieving these results.

1.6 Practical Significance and Approbation

Beyond the theoretical contributions, the research has a strong practical impact.

- **Validation Through High-Impact Projects:** The methodologies and tools developed in this thesis have been successfully implemented in multiple national and international projects, confirming the practical relevance of the

research results. The research has been supported by highly competitive initiatives funded by the European Regional Development Fund, the Latvian Council of Science, and various state research programs.

- **Academic Recognition:** The infrastructure's approval is reflected in its widespread adoption across interdisciplinary projects and its positive reception within the academic community. Frequent citations in scholarly publications further underscore its impact, significance, and long-term sustainability.
- **Advancement in Language Technology:** By providing a standardized and scalable platform, this research significantly enhances access to Latvian linguistic data. It not only supports the development of language technologies, but also advances digital humanities research by making corpora more discoverable and easier to analyze.

The results of this thesis have been validated in adaption and training of multiple models.

- **Ilvars - Latvian Male VITS Text-to-Speech Model**[1] - A neural model for text-to-speech (TTS) synthesis in Latvian. Trained using VITS on a 25-hour speech corpus of audiobooks read in a male voice. Available for academic and non-commercial purposes via an API.
- **Latvian BERT base model**[2] - A BERT model pretrained on Latvian language data using the masked language modeling and next sentence prediction objectives. (Znotins and Barzdins, 2020)
- **General-purpose Latvian ASR model**[3] - This is a fine-tuned whisper-large-v3 model for Latvian, trained by AiLab.lv using two general-purpose speech datasets: the Latvian part of Common Voice 19.0, and the latest version of the Latvian broadcast dataset LATE-mediji.
- **General-purpose Latgalian ASR model**[4] - This is a fine-tuned whisper-large-v3 model for Latgalian, trained by AiLab.lv using Latgalian part of Common Voice 19.0.

1.7 Research Projects

The research presented in this thesis has been implemented and validated across a diverse range of high-impact projects, demonstrating both its versatility and real-world relevance. Each project has contributed unique insights and validated different aspects of the proposed infrastructure, reinforcing its practical significance across multiple domains.

[1] Ilvars - Latvian Male VITS Text-to-Speech Model - https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/89

[2] Latvian BERT base model - https://huggingface.co/AiLab-IMCS-UL/lvbert

[3] General-purpose Latvian ASR model - https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-lv-late-cv19

[4] General-purpose Latgalian ASR model - https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-latgalian-2503

European Regional Development Fund *University of Latvia and its institutes in the European research space - excellence, activity, mobility and capacity* (2018-2022) (1.1.1.5/18/I/016)

The project "University of Latvia and Institutes in the European Research Area – Excellence, Activity, Mobility, Capacity" aimed to strengthen Latvia's research institutions by enhancing their competitiveness, mobility, and capacity within the European Research Area. Key activities included preparing high-level project proposals for international programs, organizing and participating in conferences, and fostering collaborations in quantum computing and European research infrastructures.

Initial development of Latvian repository CLARIN-LV, NoSketch Engine deployment and corpus preparation in necessary formats.

Latvian Council of Science Fundamental and Applied Research Projects *Development of Learner Corpus of Latvian: Methods, Tools and Applications (LaVA)* (2018-2021) (lzp-2018/1-0527)

The project aimed to establish a research foundation for studying the peculiarities of Latvian language acquisition by creating a Latvian language learner corpus. Its specific objectives were to develop a methodology for annotating errors in the corpus, build the corpus and its infrastructure, analyze learners' mistakes both quantitatively and qualitatively, and create corpus-based learning materials and a self-assessment web platform.

Development of error-annotation shcema, methodology, pipeline and infrastructure for learner corpus LaVA.

State Research Programme Latvian Language *Latvian Language* (2018-2021) (VPP-IZM-2018/2-0002)

The overarching goal of the program was to strengthen the sustainability, linguistic quality, and competitiveness of the Latvian language as a fundamental element of Latvia's identity and national values, both within the country and in the global linguistic context.

One of the research objectives was to develop and analyze Latvian language corpora needed for studying Latvian as a native, second, foreign, and heritage language. Development of error annotation methodology and infrastructure, development of VVPP corpus

European Regional Development Fund Industry-Driven Research *Latvian Speech Recognition and Synthesis for Medical Applications (RUTA:MED)* (2019-2022) (1.1.1.1/18/A/153)

The project aimed to develop Latvian language resources necessary for adapting speech recognition and synthesis technologies for radiology and other medical fields. It also sought to develop the required software components and evaluate

their suitability for radiology and medical applications.

Methodology and infrastructure for MediSpeech corpus development was resereched in context with this thesis.

State Research Programme Digital Resources of the Humanities *Digital Resources for Humanities: Integration and Development* (2020-2022) (VPP-IZM-DH-2020/1-0001)

The aim of this project was to support the development and use of digital humanities resources and tools in Latvia. It prevented fragmentation by fostering collaboration between institutions, ensured wider visibility and accessibility, analyzed user needs to improve usability, and advanced digital humanities research and education.

The first veriions of Korpuss.lv platform was developed in this project.

European Union Horizon 2020 *SELMA – Stream Learning for Multilingual Knowledge Transfer* (2021-2023) (Grant agreement No 957017)

The aim of this international project was to build a multilingual open-source platform that could process very large volumes of content and featured a continuous learning AI system. It was designed to help media monitors and journalists make sense of huge content streams (big data analysis) and enabled them to enrich audiovisual output through transcription, translation, voice-over, and subtitling, thus making it more accessible.

The datasets developed using infrastructure and annotation methodology presented in this thesis were validated in international colloboration to develop an multilingual platform and NLP models.

State Research Programme Letonika – Fostering a Latvian and European Society *Research on Modern Latvian Language and Development of Language Technology (LATE)* (2022-2024) (VPP-LETONIKA-2021/1-0006)

The aim of the project was to advance research on the grammatical, lexical-semantic, phonetic, and phonological system of the modern Latvian language and Latvian sign language using data-driven methods, as well as to develop sustainable Latvian language resources and tools.

The speech corpus infrastructure and transcription methodology was developed and approbated by developing LATE-mediji and LATE-sarunas.

EU Recovery and Resilience Facility *Language Technology Initiative* (2023-2026) (2.3.1.1.i.0/1/22/I/CFLA/002)

The goal of the project was to develop and refine large-scale language models (LLM), grammars, and lexicons, to create technologies for monolingual and multilingual audiovisual data processing, and to build resources and tools to sup-

port the learning of language technologies for both developers and users.

Further development of korpuss.lv platform and related services.

Latvian Council of Science Fundamental and Applied Research Projects
***Common Writing Errors in Latvian: Corpus-Driven Error Analysis and Text Correction* (Norma) (2024–2026) (lzp-2023/1-0481)**

The project aims to create a semi-automatically annotated corpus of errors made by native speakers of Latvian, documenting, correcting, and explaining the most common mistakes in the language. The error corpus will be used to develop a more advanced grammar checker, one that not only detects technical inaccuracies and basic spelling or punctuation errors, but also identifies issues in sentence structure and construction.

This project expands and validates the experience and methodology developed in previous projects for building error corpora. While earlier efforts focused on learner corpora for non-native speakers of Latvian, this initiative is based on a native speaker error corpus, which contains more complex syntactic and stylistic mistakes.

1.8 Publication of the Research Results

The research results presented in this thesis are organized as a compilation of publications. In total, the author has contributed as a co-author to 37 publications, the majority of which are indexed in Scopus, with three of these achieving the highest recognition as Q1 journal articles. From this substantial collection, a selected subset of 14 core publications has been included in the thesis. The author of this thesis is the main author in 10 of these core publications, and 13 of them are indexed in Scopus.

The core publications detail the design, implementation, and application of the Latvian corpora infrastructure, showcasing its development from initial conception through to real-world deployment. These publications form the central pillars of the thesis and provide detailed insights into the methodologies, experiments, and results obtained.

1. **R. Dargis**, G. Rabante-Busa, I. Auzina and S. Kruks, *ParliSearch – A system for large text corpus discourse analysis*, Human Language Technologies – The Baltic Perspective, Vol. 289, IOS Press. (2016). Scopus, WoS

2. **R. Dargis**, I. Auzina, U. Bojars, P. Paikens and A. Znotins, *Annotation of the Corpus of the Saeima with Multilingual Standards*, Proceedings of the 2018 ParlaCLARIN Workshop. (2018).

3. **R. Dargis**, I. Auzina and K. Levane-Petrova, *The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners*, Proceedings of the 11th International

Conference on Language Resources and Evaluation (LREC), pp. 4111–4115. (2018). Scopus, WoS

4. U. Bojars, **R. Dargis**, U. Lavrinovics and P. Paikens, *LinkedSaeima: a Linked Open Dataset of Latvia's Parliamentary Debates*, Proceedings of the 15th SEMANTiCS Conference, Vol. 11702, Springer, pp. 50–56. (2019). Scopus

5. **R. Dargis**, N. Gruzitis, I. Auzina and K. Stepanovs, *Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian*, Human Language Technologies – The Baltic Perspective, Vol. 328, IOS Press, pp. 135–141. (2020). Scopus, WoS

6. **R. Dargis**, I. Auzina, K. Levane-Petrova and I. Kaija, *Detailed Error Annotation for Morphologically Rich Languages: Latvian Use Case*, Human Language Technologies – The Baltic Perspective, Vol. 328, IOS Press, pp. 241–244. (2020). Scopus, WoS

7. **R. Dargis**, I. Auzina, K. Levane-Petrova and I. Kaija, *Quality Focused Approach to a Learner Corpus Development*, Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 392–396. (2020). Scopus, WoS

8. N. Gruzitis, **R. Dargis**, V. Lasmanis, G. Garkaje and D. Gosko, *Adapting Automatic Speech Recognition to the Radiology Domain for a Less-Resourced Language: The Case of Latvian*, Intelligent Sustainable Systems, Vol. 333, Springer, pp. 267–276. (2022). Scopus

9. **R. Dargis**, I. Auzina, I. Kaija, K. Levane-Petrova and K. Pokratniece, *Corpus Based Self-Assessment Platform for Latvian Language Learners*, Baltic Journal of Modern Computing, Vol. 10(3), pp. 392–401. (2022). Scopus, WoS

10. **R. Dargis**, I. Auzina, I. Kaija, K. Levane-Petrova and K. Pokratniece, *LaVA – Latvian Language Learner corpus*, 13th Language Resources and Evaluation Conference (LREC), pp. 727–731. (2022). Scopus, WoS

11. B. Saulite, **R. Dargis**, N. Gruzitis, I. Auzina, K. Levane-Petrova, L. Pretkalnina, L. Rituma, P. Paikens, A. Znotins, L. Strankale, K. Pokratniece, I. Poikans, G. Barzdins, I. Skadina, A. Baklane, V. Saulespurens and J. Ziedins, *Latvian National Corpora Collection – Korpuss.lv*, 13th Language Resources and Evaluation Conference (LREC), pp. 5123–5129. (2022). Scopus, WoS

12. **R. Dargis**, A. Znotins, I. Auzina, B. Saulite, S. Reinsone, R. Dejus, A. Klavinska and N. Gruzitis, *BalsuTalka.lv – Boosting the Common Voice Corpus*

for Low-Resource Languages, Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 2080–2085. (2024). Scopus

13. **R. Dargis** and B. Saulite, *Korpuss.lv – a Versatile Platform for Digital Humanities*, Baltic Journal of Modern Computing, Vol. 12(4), University of Latvia, pp. 636–645. (2024). Scopus, WoS

14. I. Auzina, N. Gruzitis, **R. Dargis**, G. Rabante-Busa, D. Gosko, J. Vempers, R. Kivkucans and A. Znotins, *Recent Latvian Speech Corpora for Linguistic Research and Technology Development*, Baltic Journal of Modern Computing, Vol. 12(4), University of Latvia, pp. 646–658. (2024). Scopus, WoS

In addition to these primary contributions, the author has also participated in several international, peer-reviewed publications that, while indirectly related to the thesis topic, further demonstrate the broader impact of the research on the fields of computational linguistics and digital humanities.

1. **R. Dargis** and A. Znotins, *Baseline for keyword spotting in Latvian broadcast speech*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

2. I. Auzina, M. Pinnis and **R. Dargis**, *Comparison of rule-based and statistical methods for grapheme to phoneme modelling*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

3. G. Garkaje, E. Zilgalve and **R. Dargis**, *Normalization and automatized sentiment analysis of contemporary online Latvian Language*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

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

5. I. Auzina, K. Levane-Petrova, G. Rabante-Busa, **R. Dargis** and A. Fabregas, *Designing an annotated longitudinal Latvian children's speech corpus*, Human Language Technologies – The Baltic Perspective, Vol. 289, IOS Press. (2016). Scopus, WoS

6. A. Spektors, I. Auzina, **R. Dargis**, N. Gruzitis, P. Paikens, L. Pretkalnina, L. Rituma and B. Saulite, *Tezaurs.lv: the largest open lexical database for Latvian*, Proceedings of the 10th International Conference on Language Re-

sources and Evaluation (LREC). (2016). Scopus, WoS

7. **R. Dargis** and I. Auzina, *Towards a Modern Text-to-Speech System for Latvian*, Human Language Technologies – The Baltic Perspective, Vol. 307, IOS Press, pp. 26–29. (2018). Scopus, WoS

8. O. Urek, A. Vulane, **R. Dargis**, A. Taurina, T. Zirina and H. G. Simonsen, *Latvian CDI: methodology, developmental trends and cross-linguistic comparison*, Journal of Baltic Studies, Vol. 50(3), Routledge, pp. 285–305. (2019). Scopus, WoS

9. I. Auzina, **R. Dargis** and K. Levane-Petrova, *Latviešu valodas apguvēju kļūdu analīze: pareizrakstības kļūdas*, Vārds un tā pētīšanas aspekti, LiePA, pp. 220–227. (2019).

10. N. Gruzitis, **R. Dargis**, L. Rituma, G. Nespore-Berzkalne and B. Saulite, *Deriving a PropBank Corpus from Parallel FrameNet and UD Corpora*, Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet, pp. 63–69. (2020).

11. **R. Dargis**, P. Paikens, N. Gruzitis, I. Auzina and A. Akmane, *Development and Evaluation of Speech Synthesis Corpora for Latvian*, Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6633–6637. (2020). Scopus, WoS

12. **R. Dargis**, K. Levane-Petrova and I. Poikans, *Lessons Learned from Creating a Balanced Corpus from Online Data*, Human Language Technologies - The Baltic Perspective, Vol. 328, IOS Press, pp. 127–134. (2020). Scopus, WoS

13. I. Auzina, I. Kaija, K. Levane-Petrova, K. Pokratniece and **R. Dargis**, *Latviešu valodas apguvēju korpusa (LaVA) izmantošana pētniecībā un mācību uzdevumu izstrādē*, Latviešu valodas apguve. XIII Starptautiskais baltistu kongress, LiePA, pp. 142–161. (2021).

14. L. Skestere and **R. Dargis**, *Agenda-Setting Dynamics during COVID-19: Who Leads and Who Follows?*, Social Sciences, Vol. 11(12), pp. 556. (2022). Q1, Scopus, WoS

15. I. Skadina, I. Auzina, **R. Dargis** and A. Voitkans, *CLARIN valodas resursu un rīku pētniecības infrastruktūra humanitārajām un sociālajām zinātnēm*, Letonica, Vol. 47, pp. 312–327. (2022). Scopus

16. I. Skadina, I. Auzina, **R. Dargis**, E. Lasmanis and A. Voitkans, *CLARIN-LV:*

Many Steps till Operation, CLARIN Annual Conference, pp. 9–13. (2022).

17. A. Znotins, **R. Dargis**, N. Gruzitis, G. Barzdins and D. Gosko, *RUTA:MED – Dual Workflow Medical Speech Transcription Pipeline and Editor*, Natural Language Processing and Information Systems, Vol. 13286, Springer, pp. 209–214. (2022). Scopus, WoS

18. I. Auzina, **R. Dargis**, B. Saulite, N. Gruzitis, M. Grasmanis, A. Spektors and K. Stepanovs, *Specializēta latviešu valodas runas korpusa un izrunas vārdnīcas izveide vizuālās diagnostikas izmeklējumu lingvistiskai analīzei un sistemātiskai transkribēšanai*, Letonica, Vol. 47, pp. 244–262. (2022). Scopus

19. I. Auzina, **R. Dargis**, I. Kaija, K. Levane-Petrova and K. Pokratniece, *Valodas korpusu izmantošana latviešu valodas uzdevumu automātiskā ģenerēšanā*, Letonica, Vol. 47, pp. 264–282. (2022). Scopus

20. B. Saulite, I. Auzina and **R. Dargis**, *Nacionālā korpusu kolekcija Korpuss.lv*, Linguistica Lettica, Vol. 31(1), LU Latviešu valodas institūts, pp. 202–223. (2023). Scopus

21. T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubesic, K. Simov, A. Pancur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrimsson, C. Coltekin, J. de Does, K. Depuydt, T. Agnoloni, G. Venturi, M. C. Perez, L. de Macedo, C. Navarretta, G. Luxardo, M. Coole, P. Rayson, V. Morkevicius, T. Krilavicius, **R. Dargis**, O. Ring, R. van Heusden, M. Marx and D. Fiser, *The ParlaMint corpora of parliamentary proceedings*, Language Resources and Evaluation, Vol. 57, Springer, pp. 415–448. (2023). Q1, Scopus, WoS

22. **R. Dargis**, G. Barzdins, I. Skadina, N. Gruzitis and B. Saulite, *Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams*, Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, pp. 289–293. (2024). Scopus

23. T. Erjavec, M. Kopp, N. Ljubesic, T. Kuzman, P. Rayson, P. Osenova, M. Ogrodniczuk, C. Coltekin, D. Korzinek, K. Meden, J. Skubic, P. Rupnik, T. Agnoloni, J. Aires, S. Barkarson, R. Bartolini, N. Bel, M. C. Perez, **R. Dargis** and e. al., *ParlaMint II: advancing comparable parliamentary corpora across Europe*, Language Resources and Evaluation. (2024). Q1, Scopus, WoS

These publications, collectively, not only validate the research methodology but also underscore the long-term practical and academic significance of the work.

2. Corpora

This chapter presents an overview of the corpora that have been developed using the infrastructure and methodology described in Chapter 3. These corpora serve as a key component in the approbation of the tools and methodologies. By presenting these corpora first, this chapter sets the stage for a deeper exploration of the infrastructure and methodological considerations that form the central research problem of this thesis.

2.1 Latvian National Corpora Collection (LNCC)

Latvian corpora are increasingly being used for pre-training large language models, such as LVBERT (Znotins and Barzdins, 2020), LitLatBERT (Ulcar et al., 2021), and GPT2-LV (Plenert, 2021). The zero-shot learning capability of large language models depends not only on model size and data volume but also on the quality and breadth of encyclopedic knowledge within the training corpora. This has led to the concept of ”GoodData” (Press, 2021). The groundbreaking GPT-3 (Brown et al., 2020) language model was trained on 750GB of primarily English GoodData, while GPT-SW3, designed for the comparatively ”small” Swedish language, was trained on a 100GB text collection (Ekgren et al., 2022). The Latvian National Corpora Collection (LNCC), with its current aggregated size of nearly 10GB and broad coverage, represents a step toward establishing Latvian GoodData – essential for training high-quality Latvian large language models essential for various downstream tasks, particularly in zero-shot natural language understanding (NLU) and natural language generation (NLG), which are shaping the future of NLP.

Since Latvian is a relatively low-resource language and even smaller than Swedish, it is unlikely that a single corpus will ever match the size or quality needed for pre-training language models as large as GPT-3 for English. Furthermore, no single balanced Latvian text corpus can fully meet the needs of modern lexicography and grammar research. New text types and sources, such as user-generated content and spoken language, as well as specialized domains, remain insufficiently covered.

LNCC aims to unify multiple corpora into a single collection with a standardized format. The goal is to encompass diverse Latvian language use cases and include all major text types and genres – such as news, social media, blogs, books, scientific texts, debates, and essays – while balancing both quality and size. Achieving this objective requires a continuous, multi-institutional, and multi-project effort, supported by the Digital Humanities and Language Technology communities in Latvia. Currently, LNCC includes 39 corpora developed by 13 institutions.

Almost all LNCC datasets are automatically tokenized and morphologically tagged. Morpho-syntactic annotation is performed using the open-source IMCS UL

morphological tagger, which achieves 92.7% full morphological tag accuracy and 97.6% lemmatization/part-of-speech (POS) accuracy (Paikens et al., 2013; Paikens, 2016). The common Latvian tagset, developed and refined over the years at IMCS UL, is a positional tagset generally aligned with the MULTEXT-EAST standard (Erjavec, 2012), with adaptations for the specific characteristics of Latvian.

2.2 Latvian Parliamentary Corpora

This chapter outlines efforts to enhance the usability of Latvian parliamentary data for researchers in digital humanities and political science.

This research laid the foundation for incorporating the Latvian parliamentary corpus into the ParlaMint corpus – a comparable and interoperable collection of parliamentary debates from 29 European countries and autonomous regions (Erjavec et al., 2023, 2024).

2.2.1 Corpus of the Saeima

The Corpus of the Saeima (Parliament of Latvia) (Auziņa et al., 2019) was first published in 2016. Initially, the corpus was released in plain text format with speaker annotations and other metadata.

The source data for this corpus was extracted from the Saeima's website ootnotehttps://www.saeima.lv/lv/transcripts/category/21, where verbatim reports of all Saeima sessions are published in HTML format. The texts were processed using a semi-automatic pipeline to identify speech boundaries and speakers. The text was then segmented into utterances, ensuring each utterance contained only a single speaker's speech. The corpus includes data spanning from 1993 to 2022.

2.2.2 Automatic Annotation with Multilingual Standards

With the increasing availability of corpora in different languages, it became evident that unannotated corpora were insufficient for facilitating comparative research. Enhancing corpora with additional annotation layers based on widely used multilingual standards enables cross-linguistic research without requiring proficiency in all the languages involved.

The Corpus of the Saeima has been enriched with multiple annotation layers:

- Morphosyntactic information for linguistic analysis, including lemmas, morphological tags, and syntactic dependencies.
- Automated translations into English.
- Named entity mentions linked to the Wikidata knowledge base.

The morphosyntactic annotation layer includes lemmas, part-of-speech tags, morphological features, and syntactic dependencies following the Universal Dependencies standard format. Texts are automatically tokenized, lemmatized, and morphologically analyzed using a CMM-based tagger (Paikens, 2016). Syntactic

dependencies are inferred by a neural transition-based dependency parser (Znotins, 2016) trained on the Latvian Universal Dependencies corpus version 2.1 (Pretkalnina et al., 2016).

Latvian speeches were translated into English using a neural machine translation system (Barone et al., 2017). The unreviewed machine-generated translations facilitate quantitative analysis and improve searchability for international researchers. However, since automated translations may lack accuracy, professional translation is recommended for qualitative analysis.

A named entity linking system was developed based on prior research in news corpora analysis (Paikens, 2016). This approach leverages structured Wikidata extracts (Ismayilov et al., 2018) as the entity knowledge base. The entity alias information in Wikidata was extended with Latvian morphological inflections and automatically generated name variants for people and organizations, ensuring accurate linking of corpus mentions to Wikidata identifiers.

In the Corpus of the Saeima, 3392,530 mentions of 2,998 unique entities were identified. Additionally, 165,000 out of 497,000 utterances contained entity mentions.

Automatic tokenization, morphological, and syntactic annotations are published in CoNLL-U format, a simple plain-text encoding. The dataset includes columns specifying the word index, word form, lemma, part-of-speech tag, full morphological tag, morphological features, head of the current word, universal dependency relation to the head, and spacing information.

2.2.3 LinkedSaeima

LinkedSaeima is a Linked Data representation of the Corpus of the Saeima. It contains structured information about parliamentary proceedings and the entities mentioned therein, represented using Wikidata identifiers. Linked Data facilitates structured representation of parliamentary debates by defining the properties of parliamentary meeting objects and their relationships.

The LinkedSaeima dataset was created by converting the Latvian parliamentary debate corpus into Linked Data, following the Europarl framework. It includes structured metadata and links to Wikidata. The dataset comprises approximately 4.9 million RDF triples, enabling advanced querying and integration with global Linked Data infrastructures.

The dataset includes 497,221 speeches (utterances) from 1,293 parliamentary meetings. These speeches were delivered by 690 speakers holding 162 different roles and contain 392,530 mentions of 2,998 unique Wikidata entities. The dataset provides structured information about the following object classes:

- **Meeting** (*lpv_eu:SessionDay*) – Represents a parliamentary plenary session, typically consisting of multiple speeches.
- **Speech** (*lpv_eu:Speech*) – An individual utterance delivered by a single speaker during a meeting.
- **Speaker** (*lpv:Speaker*) – The person delivering a speech.

- **Role** (*lpv:PoliticalFunction*) – The political function held by the speaker at the time of the speech (e.g., Prime Minister). A speaker may appear in multiple roles over time.

The LinkedSaeima data model follows the LinkedEP project's structure and adopts the LinkedPolitics vocabulary, referenced in this paper using the prefixes *lpv* and *lpv_eu* (Van Aggelen et al., 2017). The primary innovation of this dataset, compared to LinkedEP, is the inclusion of named entity information. This is represented using the *schema:mentions* property, which links entities to their corresponding Wikidata identifiers. Additionally, speaker roles are "materialized" by assigning URI identifiers, enabling dataset querying (e.g., filtering speeches by Ministers of Foreign Affairs) and linking roles to external datasets. Speaker roles (*lpv:PoliticalFunction*) may also contain links to corresponding Wikidata entities.

2.3 Speech Corpora

Speech corpora play a essential role not only in the development of language technology – particularly automatic speech recognition (ASR) and text-to-speech synthesis (TTS) – but also in enhancing our understanding of phonetics and prosody, morphology and syntax, semantics, and pragmatics.

As demand grows for speech and language technology to support the rapid development of various open-source and commercial applications, as well as modern studies in linguistics and digital humanities (DH), diverse speech corpora have become invaluable resources. Corpora contribute to improving and evaluating ASR and TTS models, end-user applications, complex workflows, DH research, and study aids.

This section introduces several conceptually distinct Latvian speech corpora, to whose development the author has made a significant contribution.

2.3.1 LATE Conversational Corpus (LATE-sarunas)

The LATE conversational corpus (Auziņa et al., 2024b) includes recordings and orthographic transcriptions of private conversations, interviews, and public speeches.

Each audio recording in the corpus is accompanied by metadata, including the speaker's gender and age group (12–15, 16–25, 26–50, 51–75, 76+), as well as information about the speech type: dialogue or monologue, spontaneous or prepared speech, etc.

The LATE-sarunas corpus comprises 35 hours of recordings from more than 300 speakers. The dataset is distributed via the CLARIN-LV repository with a CLARIN Restricted license.

2.3.2 LATE Media Speech Corpus (LATE-mediji)

The LATE media speech corpus (Auziņa et al., 2024a) includes recordings of broadcasts from Latvian public media, encompassing both spontaneous and prepared speech.

The speech data is transcribed according to standard Latvian orthography, adhering to punctuation and other grammatical rules.

The LATE-mediji corpus contains 70 hours of recordings from more than 250 speakers. The dataset is distributed via the CLARIN-LV repository with a CLARIN Academic license.

2.3.3 Radiology Speech Corpus (LVMED)

Despite the widespread use of modern medical technology in Latvia, medical reports are still produced manually. The largest healthcare institutions in Latvia either maintain in-house transcription centers or outsource transcription services. However, as the number of diagnostic examinations continues to grow, both doctors and patients frequently experience delays of several days before receiving their reports. Moreover, transcription services are expensive, making them inaccessible for regional healthcare institutions.

General-purpose ASR systems trained on broad speech and text corpora are not suitable for medical reports due to their specialized language. The word error rate (WER) is too high for practical use, making domain-specific ASR adaptation essential.

Inspired by the successful implementation of Estonian ASR in radiology (Paats et al., 2018; Alumäe et al., 2017), ASR adaptation resources were developed for Latvian, including the 30-hour LVMED corpus (Auziņa et al., 2022). This corpus consists of manually transcribed speech recordings specifically selected to support ASR training and evaluation in the medical domain.

The corpus was created using real dictation recordings from hospital transcription center archives. The selected recordings represent approximately 10% of the total 300-hour archive, ensuring diversity in both speakers and acoustic conditions. Audio recordings were filtered based on quality, excluding low-quality recordings (e.g., those with excessive compression, low bitrate, or low sample rate). High-quality recordings can be artificially degraded if necessary, but the reverse is not possible. As a result, the corpus includes approximately 70 unique speakers.

To avoid overfitting the acoustic model and to improve generalization, an approximately equal number of recordings were selected for each speaker. Since recording and acoustic conditions often vary even for the same speaker, no more than one recording per day was chosen per speaker, and the time intervals between selected recordings were evenly distributed.

Annotation guidelines were adapted for medical dictation transcription, drawing on experience gained from developing a general-purpose Latvian speech corpus (Pinnis et al., 2014, 2016).

2.3.4 *BalsuTalka.lv and BolsuTalka.lv*

Between mid-2023 and mid-2024, the Latvian section of the multilingual Common Voice corpus[1] expanded significantly in both quantity and diversity. This was achieved through the national crowdsourcing initiative BalsuTalka.lv[2], in which a carefully curated text corpus was read by thousands of individuals of different ages and nationalities, both from Latvia and the diaspora. In late 2023, this campaign was successfully extended to Latgalian, which had not previously been represented in Common Voice.

The first step in creating or expanding a Common Voice (CV) speech corpus is to submit and validate a text corpus comprising well-formatted sentences for reading. Before the campaign, the Latvian CV corpus contained around 7,000 sentences, mostly sourced from movie subtitles. To enhance diversity and increase the number of recorded speech samples, the corpus was expanded to nearly 30,000 sentences. This effort enriched the corpus with a broader range of text genres, functional styles, and vocabulary. Readability and conversational style were prioritized, incorporating expressive elements such as questions, exclamations, dialogues, and conversational fragments. The CV 18.0 release[3] includes 293 recorded hours for Latvian, with 244 validated hours contributed by 6,086 speakers.

To establish a CV corpus for Latgalian, the Mozilla CV user interface was localized, and an initial set of 5,000 Latgalian sentences was curated and submitted to the platform. Selection criteria included adherence to standard Latgalian orthography, phonetic and intonational diversity (narrative, interrogative, exclamatory), and content variety. Text snippets from dictionaries, short dialogues, and phraseology from fiction and non-fiction were manually added.

The Latgalian ”Bolsu tolka” campaign was organized as an extension of the Latvian ”Balsu talka” initiative, benefiting both efforts. The current Latgalian CV text corpus contains almost 10,000 sentences, while the CV 18.0 Latgalian release includes 27 recorded hours (from 321 speakers) and 25 validated hours. All Common Voice datasets are openly available for both research and commercial use.

2.4 Error-Annotated Corpora

Error-annotated corpora are collections of texts, typically produced by language learners, that have been systematically analyzed to identify, correct, and categorize linguistic errors.

Learner corpora have been collected and analyzed for more than 25 years. While there are numerous learner corpora for English (Granger et al., 2009; Gilquin et al., 2010), similar resources for other languages are also gaining popularity. These include French (Granfeldt et al., 2006), Swedish (Volodina et al., 2019), Nor-

[1] https://commonvoice.mozilla.org

[2] Approximate translation of ”balsu talka”: voice harvesting; however, ”talka” conveys the idea of voluntary communal work towards a common goal.

[3] https://commonvoice.mozilla.org/lv/datasets

wegian (Tenfjord et al., 2006), Dutch (Lemmens and Perrez, 2010), Japanese (Gries and Adelman, 2014), Arabic (Alfaifi et al., 2014), Chinese (Wang et al., 2015), Portuguese (Mendes et al., 2016), and others.

The demand for learning Latvian as a foreign language is increasing. Latvian as a foreign language is taught not only at higher educational institutions in Latvia but also in more than 20 universities worldwide (Šalme, 2011; Laizane et al., 2018). Therefore, corpus-based and corpus-driven teaching materials are essential for international students acquiring Latvian both in Latvia and abroad. Learner corpora are not only essential for studying the language of learners but also play a significant role in the development of educational applications.

An error-annotated corpus development schema and platform were designed and validated through the creation of two corpora. Details about the schema and annotation process are provided in section 3.2.

2.4.1 Corpus of the Tests of the State Language Proficiency Testing (VVPP)

The initial version of the error annotation pipeline was developed and validated during the creation of the VVPP corpus (Auziņa et al., 2018). This corpus was compiled as part of the Latvian Language Agency's research project *Quality of the Latvian Language: Results of the State Language Proficiency Test*.

The corpus consists of successfully passed tests from the State Language Proficiency Testing (Certification), which evaluates an applicant's state language proficiency level. A total of 900 tests were included, with 150 tests for each proficiency level (A1, A2, B1, B2, C1, C2). Upon passing the State Language Proficiency Examination, an Applicant receives a state language proficiency certificate, which is required for employment and obtaining a permanent residence permit.

Unfortunately, this dataset is not publicly available due to privacy concerns.

2.4.2 Latvian Language Learner Corpus (LaVA)

The infrastructure was further refined during the development of the LaVA corpus (Auziņa et al., 2021). This corpus was created as part of the Latvian Council of Science Fundamental and Applied Research Project *Development of a Learner Corpus of Latvian: Methods, Tools, and Applications (LaVA)*.

The LaVA corpus contains 1,015 essays (190,000 tokens and 790,000 characters, excluding whitespaces) written by foreign students at Latvian higher education institutions. These students are learning Latvian as a foreign language in their first or second semester, typically reaching the A1 (or possibly A2) Latvian language proficiency level.

Copyright and personal data protection are among the most important legal aspects that must be addressed before initiating data collection for a learner corpus. The purpose of the agreement form is to inform authors about how their texts will be used and to obtain their explicit consent for such use (Kaija and Auzina, 2020).

A structured agreement/questionnaire form was developed for corpus data collection. This form is printed on one side of an A4-sized paper and consists of three parts: an information letter, a permission form, and a metadata collection questionnaire (containing details about the author). The metadata includes gender, age, mother tongue, knowledge of other languages, and the duration of Latvian language study. The reverse side of the form is left blank, where authors are requested to handwrite their essays.

The most common essay topics include: *Me and My Family*, *My Routine*, and *My Studies*. To protect personal data, students are encouraged to use fictional information instead of real-life details.

The data was collected from five universities: Riga Stradiņš University (87%), Rezekne Academy of Technologies (4%), University of Latvia (3%), Liepaja University (3%), and the Latvian Academy of Culture (3%).

3. Infrastructure and Methodology

This chapter outlines the technical and methodological foundations that support the development of the Latvian corpus presented in the previous chapter. The following sections provide an in-depth exploration of pipelines, annotation schemas, and tools that facilitate corpus creation and processing. They also demonstrate how these components address the challenges of developing corpora for a morphologically rich and relatively low-resource language like Latvian.

3.1 Speech Corpora

Creating large, curated speech corpora is a labor-intensive task. Modern ASR development requires less data to achieve competitive results – primarily through fine-tuning large, pre-trained multilingual models such as XLS-R (Babu et al., 2022), Whisper (Radford et al., 2023), and MMS (Pratap et al., 2024).

This section presents various approaches used to develop several conceptually distinct Latvian speech corpora. These corpora serve both data-driven research in digital humanities (DH) and the development of ASR models for general-purpose and domain-specific applications.

3.1.1 Speech Corpora Transcription

Transcription format is a vital aspect of speech corpora, as it directly influences their usability, interpretability, and application across various research and technology development areas. The choice of transcription format can significantly affect the depth of analysis and the precision of computational applications.

Defining a comprehensive transcription format requires multiple decisions, as it must include various layers of information. A well-defined transcription format determines how pronunciation variations, text formatting, and speech phenomena (such as self-repair, false starts, and mispronunciations) are represented. The orthographic transcription style significantly impacts data usability. For example, writing numbers in words (”twenty-five”) versus digits (”25”) affects both linguistic readability and automatic speech recognition (ASR) performance. Similarly, transcription rules must specify whether to explicitly mark words spoken in a different language, which is particularly important in multilingual corpora. Abbreviations present another challenge, as their pronunciation may differ from their written form. Deciding whether to transcribe ”USA” as spoken (”U-S-A”) or as a single word affects ASR training. Dictation corpora introduce additional complexity, as they may include commands such as ”full stop”, ”new line”, or ”comma”, which must be clearly distinguished from naturally spoken punctuation to avoid confusion.

All speech corpora introduced in Section 2 adhere to the same general transcription principles. In orthographic transcription, sentences are segmented based on syntactic and prosodic cues. Non-verbal elements, unclear speech, and physiological noise are annotated in the transcriptions. Transcription is performed in a single layer, with all additional information noted using specific markup syntax, allowing different transcription layers to be extracted based on specific use cases.

Annotations in square brackets indicate deviations from standard pronunciation norms (e.g., ”lasām [lasam]”; ”interesanti [intresanti]”), as well as the pronunciation of abbreviations and foreign words (e.g., ”SIA [si ā]”; ”ZZS [zē zē es]”; ”Rail [reil] Baltica [boltik]”). They also capture word truncated words (e.g., ”četrdesmit [čēesnt]”), and the reading of numbers, which requires contextual syntactic agreement of word forms (e.g., ”7.8 [septiņi komats astoņi] grami” – nominative; ”līdz 1940. [tūkstoš deviņsimt četrdesmitajam] gadam” – dative).

In the LVMED corpus, which also contains dictation commands, commands that are meant to be executed are annotated in curly brackets (e.g., ”full stop”, ”new line”).

For corpora where a written text was available beforehand, a separate transcription layer is preserved, enabling further research. The LVMED corpus also includes the text of medical reports given to patients, allowing analysis and training of models on how transcriptions can be edited into finalized texts for patient documentation. Additionally, the LATE-Mediji corpus contains subtitles for specific segments, enabling analysis of how spoken text is shortened and what information is omitted to conform to subtitle guidelines.

Segments containing unclear or ambiguous pronunciation can negatively affect ASR training. However, these segments are particularly valuable for specific linguistic analyses. Detailed orthographic transcriptions enable deeper linguistic analysis and allow filtering of high-quality data for ASR training.

3.1.2 Crowdsourcing Speech Corpus

The availability of large-scale open speech corpora for lesser-spoken languages has been limited. This was the case for Latvian, a language spoken by approximately 1.5 million people. While several closed Latvian speech corpora exceeding 100 hours existed and were utilized for training automatic speech recognition (ASR) models, only a few small open datasets were accessible at the beginning of 2023. Among these, the Latvian Common Voice 13.0 dataset, containing 18 hours of recorded speech, was the largest.

Through a national crowdsourcing initiative, collaboratively organized by multiple institutions, the Latvian Common Voice corpus grew tenfold in both size and speaker diversity within a year. A subsequent initiative was successfully launched for Latgalian, a recognized endangered historic variant of Latvian with around 150,000 speakers. The goal of both initiatives was not only to increase the volume of data but also to enhance its diversity – capturing a broad range of speakers, accents, text genres, styles, intonations, grammatical structures, and vocabulary.

During the planning phase, careful consideration was given to the structure that would yield the highest participation and the greatest impact. Two main options were explored: building a custom platform or using an existing one hosted by a global initiative. A global platform could ensure broader visibility and greater long-term impact but might struggle to engage the local community. A custom platform would allow for a more tailored message and stronger cultural connection but would require significant development effort and might limit global reach.

Ultimately, a hybrid approach was adopted to combine the best of both worlds. A culturally resonant landing page "BalsuTalka.lv" was created to direct users to the Mozilla Common Voice (CV) platform[1].

This targeted, culturally aware campaign led to widespread participation both within Latvia and among the diaspora. "BalsuTalka," in both its name and design, reflected a strong Latvian identity. By leveraging local platforms and culturally relevant storytelling, the campaign integrated national symbols and values to resonate with the audience and drive engagement. The simple, focused landing page offered a clear call to action, avoiding the confusion that might arise when promoting a large global platform with diverse objectives.

Mozilla's Common Voice platform provided the technical backbone for the project – a reliable, scalable infrastructure with strict privacy and security measures. Its multilingual, standardized data repository is well-recognized within the international NLP community, making Latvian data both accessible and compatible with broader multilingual research and language models. This ensured that the collected data was not only preserved but also readily usable, maximizing its long-term value.

The initiative was officially launched on May 4, 2023, to coincide with Latvia's Restoration of Independence Day. Early promotional efforts included widespread coverage on national television, radio, and online platforms. One of the most successful campaigns was hosted on the popular Latvian online dictionary Tezaurs.lv, which generated over 16,000 unique conversions within three months.

Before the initiative began, the Latvian Common Voice corpus (version 13.0, March 2023) included 18 recorded and 14 validated hours from 321 speakers. Within six months, version 15.0 (September 2023) had grown to 165 recorded and 88 validated hours, contributed by 2,773 speakers. By March 2024, version 17.0 reached 277 recorded and 223 validated hours from 5,712 speakers, placing Latvian among the top languages in terms of contributor participation relative to the total number of native speakers.

In parallel, a Latgalian Common Voice corpus was launched. The first batch of 5,000 sentences was carefully selected and submitted, adhering to orthographic standards and ensuring phonetic and intonational variety. By the release of Common Voice 17.0, nearly 10,000 Latgalian sentences had been recorded, resulting in 24 validated hours of speech contributed by 250 speakers.

This initiative clearly demonstrates that localized, culturally relevant approaches to crowdsourcing speech data can successfully mobilize community par-

[1] https://commonvoice.mozilla.org

ticipation. The newly developed corpora have already proven valuable for enhancing ASR capabilities and advancing linguistic research.

3.2 Infrastructure for Error Annotation

Error annotation plays a vital role in understanding and improving learner language production. The infrastructure developed for this purpose is designed to support the systematic annotation of linguistic errors in texts written by Latvian language learners.

The annotation schema is particularly suited for languages with relatively free word order and rich morphology, such as Latvian. As a morphologically rich language with a high degree of inflection, Latvian presents unique challenges in error annotation, including distinctions in vowel length, diphthongs, and flexible word order. Moreover, as a phonetic language with a relatively straightforward orthography-to-phonology relationship, spelling errors can be systematically categorized and analyzed.

To support this annotation process, a robust infrastructure was developed to facilitate the digitization, correction, and morphological annotation of learner texts. This infrastructure was validated through the creation of two key corpora: the *Corpus of the Tests of the State Language Proficiency Testing (VVPP)* and the *Latvian Language Learner Corpus (LaVA)*.

The following sections detail the components of this infrastructure, outlining the corpus creation pipeline and the error annotation schema that underpin the systematic categorization of learner errors in the Latvian language.

3.2.1 Corpus Creation Pipeline

The corpus creation pipeline consists of four steps:

1. Data digitization
2. Text correction
3. Morphological annotation
4. Error annotation

Each step is performed independently by two annotators, with inconsistencies resolved by a third independent annotator. Error types are automatically determined based on morphological annotations and the alignment between the original and corrected text.

Most of the essays are handwritten and require digitization. Only a few recent essays are typed by students due to distance learning during COVID-19 restrictions. Character-level agreement for text digitization between the two annotators is 97.4%.

In the text correction step, the original text is edited to a grammatically correct version based on the assumed target hypothesis (Auzina et al., 2020). Character-level agreement for text correction between the two annotators is 96.8%.

Both the original and corrected texts are morphosyntactically annotated. The initial annotation is generated by the IMCS morphological tagger (Paikens, 2016) and subsequently manually verified by two annotators. Morphosyntactic annotations include part-of-speech tags, lemmas, and other Latvian-specific morphological and syntactic information. Additionally, a token with corrected typos is added to the original token. The alignment between the original token and the corrected token is used in the error analysis step to determine the exact types of errors made by learners. Without this alignment, distinguishing between typos and inflectional or word formation errors would not be possible. Annotator agreement for all layers is 92.5%, with individual layer agreement ranging from 95.5% for the original tag to 99.3% for the corrected lemma.

The final step of corpus creation is automatic error annotation. Currently, errors are classified into six categories, but since this step is automated, modifications can be made as needed. The six categories are: spelling errors, inflectional and word formation errors, lexical errors, punctuation errors, syntactic errors, and complex errors.

Time for each activity is automatically measured in the corpus platform. The timer stops if the annotator is inactive for more than 15 seconds and resumes upon detecting activity (mouse movement or keyboard input). The time required for each step is shown in Table 3.1. On average, processing a single essay takes 45 minutes, as the initial steps must be completed twice.

Step	Rate (characters per minute)	Average time per step (minutes)
Initial digitization	128.9	6.1
Final digitization	370.6	2.1
Initial text correction	177.8	4.4
Final text correction	274.1	2.8
Initial annotation	95.1	8.4
Final annotation	117.1	6.6

Table 3.1
Time consumption per corpus creation step

3.2.2 Error Annotation Schema

In this error annotation schema, errors are not annotated directly. Instead, detailed error codes are extrapolated from five easily identifiable properties: the original token without typos, the original lemma, the original tag, the corrected lemma, and the corrected tag (Figure 3.1).

Spelling errors are determined automatically by comparing the original token with the version corrected for typos. Character-level alignment, combined with

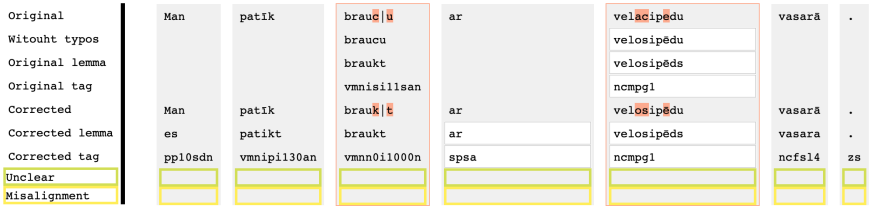

Figure 3.1: **Error annotation interface**

a rule-based system, enables the precise identification of incorrectly used character pairs. This information facilitates both qualitative research (by providing fine-grained search capabilities) and quantitative analysis (by grouping extracted character errors).

Morphological tags contain extensive information, including part-of-speech tags. Since punctuation marks have distinct tags, punctuation errors are easily identified – if the corrected token differs from the original token and both are punctuation marks, a punctuation error is recorded.

Lexical errors occur when the lemma of the corrected token differs from the lemma of the original token. While subtypes cannot be determined automatically, they can be assigned manually for unique token pairs, as lexical errors are not context-dependent.

The remaining errors fall under grammatical errors, which can be analyzed in detail using morphological tags.

Two additional properties can be annotated – *unclear* and *misalignment*. *Misalignment* is used when alignment errors occur. *Unclear* is used for cases where annotation is ambiguous or when a broader context affects the error but cannot be captured within the current scheme. Such cases are reviewed to refine the annotation scheme and incorporate syntax-level error annotations if needed.

3.3 Infrastructure for Digital Humanities

Language research seeks to understand the structure, function, and use of language through systematic study. Corpora, which are large, structured collections of textual or spoken language data, are invaluable in this endeavor (McEnery and Hardie, 2012). They provide a robust empirical basis for both qualitative and quantitative analyses, enhancing the accuracy and depth of linguistic research. By using corpora, researchers can achieve greater objectivity and reproducibility in their studies. As language continues to evolve, the role of corpora in documenting and analyzing these changes grows, making them an indispensable resource in linguistic research.

Despite the numerous advantages, several challenges limit the effective use of corpora. Many corpora are not readily accessible to researchers due to complex formats or the need for specialized software to prepare the data for analysis. This often requires advanced technical skills, creating barriers for researchers without

such expertise. Additionally, running annotation tools can be computationally demanding.

This section describes efforts made to promote corpus-based research in the digital humanities and linguistics by improving the accessibility of corpora.

3.3.1 Korpuss.lv

Korpuss.lv was specifically developed to serve as the primary access point for the Latvian National Corpora Collection (LNCCC).

The website's user interface (UI) and user experience (UX) are designed to facilitate the discovery of various corpora. The site is available in both Latvian and English. The homepage features a list of corpus information cards equipped with filtering and sorting functionalities, enabling researchers to identify the most relevant corpora for their studies.

Each corpus card provides key details, including the corpus code, full name, data period, size, developers, and a search button (if an online search is available). Clicking on a card leads to a detailed corpus information page.

The corpus information page offers comprehensive details and related resources, including links to the corpus homepage, search functionalities, word frequency lists, metadata tags associated with the corpus, and the CLARIN repository. These resources simplify access for researchers, facilitating corpus exploration in their scholarly work.

Additionally, the corpus information page lists key publications associated with the corpus, offering insights into corpus development, methodologies, and significant findings. Such references are instrumental in understanding the research context. Furthermore, the page provides detailed citation information to ensure proper acknowledgment in academic work, including one highlighted publication and a data reference if the corpus is published in the CLARIN repository.

Proper citation is essential for the authors of the corpus, as it enhances their citation-based metrics and indexes. Corpora are often developed as part of funded research projects, where various key performance indicators (KPIs) are used to evaluate the impact of the project. One such KPI is the number of studies utilizing the corpus. Proper citation helps to identify and document these studies, thereby demonstrating the relevance and impact of the corpus. This, in turn, aids authors in securing further funding.

Federated content search (FCS) is an efficient way to locate relevant corpora for a study. FCS allows researchers to search for specific words or phrases across multiple corpora simultaneously. Search queries can be made on the token or lemma layers. The search query supports wildcard symbols for single and multiple characters, as well as the *OR* operator.

3.3.2 NoSketch Engine

NoSketch Engine[2] is a robust corpus management and analysis tool widely used by linguists, lexicographers, and digital humanities researchers (Kilgarriff et al., 2014). It facilitates the exploration of large text corpora by enabling complex queries that extract detailed insights into word usage, syntactic patterns, and semantic relationships.

One of the fundamental features of NoSketch Engine is its concordance tool, which generates a list of occurrences of a word or phrase within its context in a corpus. This allows researchers to examine how specific terms are used across various texts, time periods, or genres. The tool supports both basic searches and advanced queries through a Corpus Query Language (CQL), enabling users to filter results by attributes like wordform, lemma, or grammatical tags.

The platform also provides frequency lists, ranking words or phrases by their occurrence in a corpus. These lists are instrumental in identifying thematic trends and comparing linguistic features across various contexts. Researchers can generate lists for individual words, lemmas, or phrases, and refine them further based on specific search criteria. This functionality allows comparisons of linguistic patterns across genres, text types, or time periods. For instance, by analyzing adjectives used to describe men and women in a literary corpus, researchers can uncover patterns in gender representation. Frequency lists can also be applied to identify word usage trends in specific domains, such as tracking the prevalence of a term like ”inflation” in parliamentary debates.

Another notable functionality is the timeline feature, which visualizes word or phrase frequencies over time. This capability is particularly valuable for diachronic studies, enabling researchers to trace language evolution, monitor shifts in public discourse, or track the prominence of specific topics over time. Timelines are generated by plotting absolute or relative word frequencies, and they include filters to exclude periods with insufficient data, ensuring accurate interpretations. For example, timelines can reveal the historical rise and fall of terms like ”crisis” in Latvian news articles, highlighting changes in societal focus.

3.3.3 CLARIN-LV

The Common Language Resources and Technology Infrastructure (CLARIN) is a central component of an European initiative designed to provide sustainable access to a wide array of digital language resources and tools. Latvia is part of this initiative. The CLARIN-LV repository hosts metadata and most of the corpora are also available for download.

[2]http://www.sketchengine.eu

3.3.4 Impact

The domain name *Korpuss.lv* was initially registered in 2007 to provide information about the first version of the Balanced Corpus of Modern Latvian. In May 2018, the platform evolved into an index of multiple corpora, launching with a collection of ten corpora. The name *Latvian National Corpora Collection* (LNCC) was officially adopted in November 2022. The first corpus within the CLARIN repository was published in July 2020, and citation guidelines were introduced in January 2023. As of now, the LNCC hosts 39 corpora, 29 of which are published in the CLARIN repository. Over the past year, the LNCC platform has recorded 6,600 users and 33,000 page views.

To assess the academic impact of the LNCC, a systematic analysis was conducted using the Google Scholar search engine[3]. Search terms were enclosed in quotation marks to ensure exact matches. The analysis revealed that *Korpuss.lv* has been cited in over 200 scholarly works since 2020. Despite the relative novelty of the name, and the term *Latvian National Corpora Collection* has appeared in 18 English-language publications and 8 Latvian-language publications.

We recommend that authors cite LNCC resources using the designated publication or CLARIN data reference when available. However, instances of direct citation using *Korpuss.lv* URLs persist. Specifically, URLs linking to corpus information pages on *Korpuss.lv* have been referenced in 37 scholarly works, while CLARIN URLs have been cited in 81 scholarly works.

Using CLARIN URLs is particularly advantageous, as corpus codes and names in Latvian or English are often too generic, leading to false positives in search results. Standardized citations improoves precision and facilitate the accurate identification and attribution of resources in scholarly contexts.

[3]Google Scholar: https://scholar.google.com/

4. Conclusion

This thesis has presented a comprehensive study on the development and application of an infrastructure for Latvian corpora, addressing key challenges in corpus creation, annotation, and accessibility. Through the implementation of structured methodologies, digital tools, and real-world applications, this research has significantly contributed to the advancement of natural language processing (NLP) and corpus-based research in Latvian linguistics and digital humanities. In doing so, this thesis has addressed the stated research hypotheses and demonstrated their validity through practical implementations and results.

The first hypothesis, namely that ”automated annotation tools and structured data pipelines will reduce the manual effort required for corpus creation while maintaining high accuracy”, has been confirmed through extensive quantitative and qualitative evaluations of the described tools. For instance, in the error annotation pipeline, automated modules systematically identified and categorized spelling, lexical, morphological, and punctuation errors, alleviating annotators from time-intensive tasks. Assessments of these automated tools demonstrated consistently high annotation accuracy – exceeding 90% on key layers – validating that the introduced technology effectively balances efficiency and reliability.

Furthermore, the enriched, annotated corpora produced through this infrastructure have proven instrumental in advancing Latvian NLP applications. The improved performance of tailored NLP models – evidenced by the successful deployment of models such as Latvian BERT, text-to-speech systems, and specialized automatic speech recognition (ASR) systems – substantiates the second hypothesis that ”the availability of annotated corpora will improve the development and performance of NLP models tailored for the Latvian language”. By providing comprehensive and high-quality linguistic data, the infrastructure has facilitated the development of Latvian NLP models, ensuring that the Latvian language remains relevant in the digital era.

Finally, the third hypothesis, which proposed that ”a structured and easily accessible infrastructure for Latvian corpora will significantly enhance corpus-based research methodologies across various disciplines”, has been substantiated by real-world applications in digital humanities, political science, and linguistics. The creation of an accessible, unified platform for Latvian corpora has lowered the technical barriers to corpus usage, allowing researchers without specialized computational knowledge to conduct complex textual analyses. The successful adoption of these resources in interdisciplinary projects and their citation in over 200 scholarly works underscore the infrastructure’s long-term value and confirm that user-oriented design and open data formats expand the scope of corpus-based research.

In summary, this research has successfully achieved its core goal of seam-

lessly integrating Latvian language data and NLP tools into modern research and application frameworks. By accomplishing these objectives, the thesis strengthens the use of Latvian language resources in scientific and digital innovation while also offering new insights into corpus creation, annotation, and utilization in broader contexts. This research not only enhances the accessibility and usability of Latvian linguistic data but also lays the foundation for future advancements in NLP applications and corpus-based research methodologies.

Bibliography

Alfaifi, A., Atwell, E., and Hedaya, I. (2014). Arabic learner corpus (alc) v2: a new written and spoken corpus of arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014*, volume 2, pages 77–89. Kobe International Communication Center.

Alumäe, T., Paats, A., Fridolin, I., and Meister, E. (2017). Implementation of a Radiology Speech Recognition System for Estonian Using Open Source Software. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2168–2172.

Auzina, I., Kaija, I., and Levane-Petrova, K. (2020). Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā. *Valoda: nozīme un forma*, 11:7–26.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of the 23rd INTERSPEECH Conference*, pages 2278–2282.

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., and Birch, A. (2017). Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ekgren, A., Gyllensten, A. C., Gogoulou, E., Heiman, A., Verlinden, S., Öhman, J., Carlsson, F., and Sahlgren, M. (2022). Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language resources and evaluation*, 46(1):131–142.

Erjavec, T., Kopp, M., Ljubesic, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Coltekin, C., Korzinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, S., Bartolini, R., Bel, N., Perez, M. C., Dargis, R., and e. al. (2024). Parlamint ii: advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubesic, N., Simov, K., Pancur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrimsson, S., Coltekin, C., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Perez, M. C., de Macedo, L., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevicius, V., Krilavicius, T., Dargis, R., Ring, O., van Heusden, R., Marx, M., and Fiser, D. (2023). The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Gilquin, G., De Cock, S., and Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain (Louvain-La-Neuve).

Granfeldt, J., Nugues, P., Persson, E., Thulin, J., Ågren, M., and Schlyter, S. (2006). Cefle and direkt profil: A new computer learner corpus in french l2 and a system for grammatical profiling. In *LREC-2006, The fifth international conference on Language Resources and Evaluation*, pages 565–570. ELRA.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., et al. (2009). *International Corpus of Learner English*. Presses universitaires de Louvain Louvain-la-Neuve.

Gries, S. T. and Adelman, A. S. (2014). Subject Realization in Japanese Conversation by Native and Non-native Speakers: Exemplifying a New Paradigm for Learner Corpus Research. In *Yearbook of Corpus Linguistics and Pragmatics 2014*, pages 35–54. Springer.

Ismayilov, A., Kontokostas, D., Auer, S. r., Lehmann, J., and Hellmann, S. (2018). Wikidata through the eyes of dbpedia. *Semantic Web*, 9(4):493–503.

Kaija, I. and Auzina, I. (2020). Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection. In *Selected papers from the CLARIN Annual Conference 2019*, pages 41–47.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1:7–36.

Laizane, I. et al. (2018). The understanding of the concepts of first language, second language and foreign language outside of latvia. In *Rural Environment. Education. Personality.(REEP). Proceedings of the International Scientific Conference (Latvia)*, number 11. Latvia University of Life Sciences and Technologies.

Lemmens, M. and Perrez, J. (2010). On the use of posture verbs by French-speaking learners of Dutch: A corpus-based study. *Cognitive Linguistics*, 21(2).

McEnery, T. and Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mendes, A., Antunes, S., Jansseen, M., and Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. In *Proceedings of the Tenth Language Resources and Evaluation Conference–LREC'16*, pages 3207–3214. European Language Resources Association.

Paats, A., Alumäe, T., Meister, E., and Fridolin, I. (2018). Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models. *Journal of Digital Imaging*, 31(5):615–621.

Paikens, P. (2016). Deep Neural Learning Approaches for Latvian Morphological Tagging. In *Human Language Technologies – The Baltic Perspective*, volume 289. IOS Press.

Paikens, P., Rituma, L., and Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 267–277, Oslo, Norway.

Pinnis, M., Auzina, I., and Goba, K. (2014). Designing the Latvian speech recognition corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1547–1553, Reykjavik, Iceland.

Pinnis, M., Salimbajevs, A., and Auzina, I. (2016). Designing a speech corpus for the development and evaluation of dictation systems in Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 775–780, Portoroz, Slovenia.

Plenert, A. (2021). gpt2-lv. In *Hugging Face*. https://huggingface.co/aidan-plenert-macdonald/gpt2-lv edition.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Press, G. (2021). Andrew Ng Launches A Campaign For Data-Centric AI. *Forbes*, Jun 16, 2021.

Pretkalnina, L., Rituma, L., and Saulite, B. (2016). Universal dependency treebank for latvian: A pilot. In *Human Language Technologies - The Baltic Perspective*, volume 289. IOS Press.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Šalme, A. (2011). *Latviešu valodas kā svešvalodas apguves pamatjautājumi*. Latviešu valodas aģentūra.

Tenfjord, K., Meurer, P., and Hofland, K. (2006). The ask corpus-a language learner corpus of norwegian as a second language. In *LREC*, volume 6, pages 1821–1824.

Ulcar, M., Zagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., and Robnik-Sikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *CoRR*, abs/2107.10614.

Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the european parliament as linked open data. *Semantic Web*, 8(2):271–281.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., et al. (2019). The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Wang, M., Malmasi, S., and Huang, M. (2015). The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Znotins, A. (2016). Word embeddings for latvian natural language processing tools. In *Human Language Technologies - The Baltic Perspective*, volume 289. IOS Press.

Znotins, A. and Barzdins, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.

Resources

Auziņa, I., Darģis, R., Bojārs, U., Paikens, P., Znotiņš, A., and Rābante-Buša, G. (2019). Corpus of the saeima (the parliament of latvia). CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/50.

Auziņa, I., Darģis, R., Levāne-Petrova, K., Auziņa, A., Saulīte, B., Ļaksa Timinska, I., Gailīte, E., Nešpore-Bērzkalne, G., Rābante-Buša, G., Pokratniece, K., and Klints, A. (2024a). Late-media. CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/114.

Auziņa, I., Darģis, R., Levāne-Petrova, K., Pokratniece, K., and Vēvere, D. (2018). Corpus of the tests of the state language proficiency testing. CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/49.

Auziņa, I., Darģis, R., Rābante-Buša, G., Timinska-Ļaksa, I., Gailīte, E., and Auziņa, A. (2024b). Late-conversational. CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/113.

Auziņa, I., Kaija, I., Levāne-Petrova, K., Pokratniece, K., and Darģis, R. (2021). Latvian language learner corpus. CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/42.

Auziņa, I., Saulīte, B., Akmane, A., Millere, E., Naļivaiko, I., Stepanovs, K., Darģis, R., and Grūzītis, N. (2022). Latvian radiology speech corpus. CLARIN-LV digital library. http://hdl.handle.net/20.500.12574/67.