



UNIVERSITY OF
LATVIA

Infrastructure for Latvian Corpora Development and Application

Roberts Dargis

Supervisor: Prof. Guntis Bārzdiņš

Doctoral Thesis Defense

University of Latvia Faculty of Science and Technology

October 17, 2025

The Challenge: A Digital Gap for Latvian

- Natural Language Processing (NLP) and Digital Humanities are driven by high-quality, language-specific corpora.
- Lower-resource languages like Latvian lag behind due to the limited availability of structured and annotated corpora.
- While many machine learning tools are multilingual, they require extensive language-specific data for effective performance.
- By focusing on the development of an infrastructure tailored to Latvian corpora, this research fills a significant gap.

Objectives

Research aim: to foster the seamless integration of the Latvian language into modern NLP tools and corpus-based studies.

- 1. Develop a tailored infrastructure** to streamline the creation, annotation, and management of Latvian corpora.
- 2. Construct diverse corpora** in structured, interoperable formats with multi-layered, standardized annotations.
- 3. Establish robust digital tools** that facilitate corpus-based research in digital humanities and political science.

Hypotheses

1. Automated pipelines will reduce manual effort in corpus creation while maintaining high accuracy.
2. Annotated corpora will improve the performance of NLP models for Latvian.
3. An accessible infrastructure will enhance corpus-based research across disciplines.

Main Results

- Automated Annotation & Data Pipelines
 - Designed an integrated infrastructure for corpus creation, annotation, and management.
 - Reduced manual work and improved data quality through automation.
- Development of Diverse Latvian Corpora
 - Created multiple specialized corpora: parliamentary, error-annotated, and speech corpora.
 - Addressed key resource gaps and enabled new digital humanities research.
- Integration into Unified Infrastructure (Korpuss.lv)
 - Standardized and merged multiple corpora from various domains.
 - Enhanced NLP applications and large-scale linguistic research.

Practical Significance and Approbation

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

- **Validation Through High-Impact Projects:** methods and tools implemented in 9 national and international projects.
- **Academic Recognition:** adopted across interdisciplinary projects, positive reception in academia with frequent citations.
- **Advancement in Language Technology:** provides a standardized, scalable platform for Latvian linguistic data, boosts language technology and digital humanities research, enhances discoverability and analysis of corpora.

Publication of the Research Results

- Total publications:
 - 37 publications (majority Scopus-indexed)
 - 3 Q1 journal articles
- Core publications in thesis:
 - 14 selected publications (13 Scopus indexed)
 - main author: 10 (author's contribution: 60%-80%)
- The core publications detail the design, implementation, and application of the Latvian corpora infrastructure.
- Collaborative effort, with the author playing a major role.
- Conducted at IMCS UL.

An aerial photograph of a city, likely New York City, showing a dense urban landscape with numerous buildings and trees. The image is overlaid with a semi-transparent blue filter. At the bottom of the image, there is a horizontal bar composed of several colored segments: red, purple, blue, teal, and green. The text is centered in the upper half of the image.

Infrastructure and Methodology for Corpora Development

Latvian Parliamentary Corpora

- First published in 2016 in plain text with speaker annotations and metadata. The latest versions includes data from 1993 to 2022.
- In 2018 enriched with multiple annotation layers, including morphosyntactic annotations, named entity recognition and automated translations into English.
- Encoded into linked data representation, extending on EuroParl framework. The dataset comprises approximately 4.9 million RDF triples.
- Part of ParlaMint corpus - a comparable and interoperable collection of parliamentary debates from 29 European countries and autonomous regions.

Speech Corpora

- Speech corpora are vital for advancing language technologies like ASR and TTS, while also enriching linguistic research and diverse applications in digital humanities.
- Corpora described in thesis:
 - LATE Conversational Corpus (LATE-sarunas) - includes recordings and orthographic transcriptions of private conversations.
 - LATE Media Speech Corpus (LATE-mediji) - includes recordings of broadcasts from Latvian public media, created using automatically aligned subtitles with audio as a starting point.
 - Radiology Speech Corpus (LVMED) – created using real dictation recordings from hospital transcription center archives.

Speech Corpora Transcription

- Directly impacts usability, interpretability, and application in research and NLP technology.
- Key Considerations:
 - Pronunciation variations, mispronunciations, self-repairs, false starts.
 - Non-verbal sounds, unclear speech, and physiological noise.
 - Orthographic style: numbers as words vs digits, abbreviations, multilingual markers.
- 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.

BasluTalka.lv: Crowdsourcing Speech Corpus

- **Objective:** Collect large, open speech corpora with minimal manual transcription.
- **Motivation:** Existing corpora were restricted due to copyright and personal data issues.
- **Strategy:** Culturally resonant landing page *BalsuTalka.lv* that directs users to Mozilla Common Voice.
- **Campaign period:** May 4, 2023 – March 2024
- **Impact:**
 - From 18 recorded and 14 validated hours from 321 speakers
 - To 277 recorded and 223 validated hours from 5,712 speakers

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.
- An error-annotated corpus development schema and platform were designed and validated through the creation of two corpora:
 - **Corpus of the Tests of the State Language Proficiency Testing (VVPP):** the corpus consists of 900 successfully passed tests from the State Language Proficiency Testing (Certification).
 - **Latvian Language Learner Corpus (LaVA):** corpus contains 1,015 essays written by foreign students at Latvian higher education institutions.

Infrastructure for Error Annotation

- 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.
- Morphological annotation is done semi-automatically.
- Error annotation is completely automatic with rule-based system, using morphological annotations.

An aerial photograph of a city, likely Cambridge, Massachusetts, showing a dense cluster of buildings and green spaces. The image is overlaid with a semi-transparent blue filter. At the bottom, there is a horizontal bar composed of several colored segments: red, purple, blue, teal, and green. The text 'Infrastructure for Digital Humanities' is centered in white, bold, sans-serif font.

Infrastructure for Digital Humanities

Korpuss.lv

- Background
 - Korpuss.lv launched in 2007 for the Balanced Corpus of Modern Latvian.
 - Expanded in 2018 to host multiple corpora (10 at launch, 42 currently).
 - Renamed Latvian National Corpora Collection (LNCC) in 2022.
- Platform Highlights
 - User-friendly interface with corpus cards, filters, and sorting tools to aid corpora discoverability.
 - Corpus information page includes extended metadata with related publications, citation guidelines and links to external resources.
 - The user interface is available in Latvian and English.
 - Federated content search across multiple corpora.

Latvian National Corpora Collection (LNCC)

- Aims to unify multiple Latvian corpora into a standardized collection.
- Covers diverse text types and genres: news, social media, blogs, books, scientific texts, debates, and essays.
- A collaborative effort by the Digital Humanities and Language Technology communities in Latvia.
- Currently includes 42 corpora developed by 14 institutions.
- Most datasets are automatically tokenized and morphologically tagged.

NoSketch Engine: Corpus Analysis Tool

- Open-source web platform for exploring large text corpora.
- Used by linguists, lexicographers, and digital humanities researchers.
- Core Features:
 - **Concordance Tool** – Displays word/phrase use in context, supports advanced CQL searches.
 - **Frequency Lists** – Ranks words or phrases by occurrence, reveals linguistic and thematic trends.
 - **Timeline Analysis** – Visualizes word frequency changes over time, useful for diachronic studies.

The Common Language Resources and Technology Infrastructure (CLARIN)

- CLARIN is a central component of an European initiative designed to provide sustainable access to a wide array of digital language resources and tools.
- The CLARIN-LV repository hosts metadata and most of the corpora are also available for download, some behind academic login.
- Each resource has special permanent URL.
- Integrated into Virtual Language Observatory (VLO) - a faceted search and discovery tool that allows users to explore, search, and access linguistic resources.

Timeline

- 2018: Index of corpora
- 2020: First corpus published in CLARIN repository
- 2022: Officially named Latvian National Corpora Collection (LNCC)
- 2023: Citation guidelines introduced
- Current status:
 - 42 corpora
 - 29 corpora in CLARIN
 - Past year 6,600 users, 33,000 page views

Academic Impact since 2020 (Google Scholar Analysis)

- Korpus.lv cited in 200+ scholarly works
- “Latvian National Corpora Collection” appears in:
 - 18 English-language publications
 - 8 Latvian-language publications
- Direct citations:
 - 37 works use Korpus.lv URLs
 - 81 work use CLARIN URLs

An aerial photograph of a city, likely Cambridge, Massachusetts, showing a dense cluster of buildings and green spaces. The image is overlaid with a semi-transparent blue filter. At the bottom, there is a horizontal bar with several colored segments: red, purple, blue, teal, and green. The text "What happened to the hypotheses?" is centered in white, bold font.

**What happened
to the hypotheses?**

First hypothesis

- Automated annotation tools and structured data pipelines will reduce the manual effort required for corpus creation while maintaining high accuracy.
- Confirmed through quantitative and qualitative evaluations of the described tools. For instance, in the error annotation pipeline, automated modules effectively identified and categorized various error types, relieving annotators of time-intensive tasks. Achieving over 90% accuracy on key layers demonstrates that the introduced tools successfully balance efficiency and reliability.

Second hypothesis

- The availability of annotated corpora will improve the development and performance of NLP models tailored for the Latvian language.
- Validated through measurable gains in Latvian NLP applications trained on these resources. The enriched corpora enabled the advancement of models such as Latvian BERT, text-to-speech, and ASR systems, which exhibited higher accuracy and contextual understanding — confirming that comprehensive linguistic data directly enhance model performance.

Third hypothesis

- A structured and easily accessible infrastructure for Latvian corpora will significantly enhance corpus-based research methodologies across various disciplines.
- Confirmed through its adoption in digital humanities, political science, and linguistics, the unified open-access platform has reduced technical barriers and enabled broader research participation. Its use in interdisciplinary projects and citations in over 200 studies demonstrate that user-oriented design and open data formats effectively expand the scope and impact of corpus-based research.

An aerial photograph of a city, likely Cambridge, Massachusetts, featuring a large, multi-story university building with many windows. The image is overlaid with a dark blue gradient. The text "Thank you!" is centered in white, bold, sans-serif font. At the bottom of the image, there is a horizontal bar with several colored segments: red, purple, blue, teal, and green.

Thank you!