The Latgalian TEZAURS and the challenges of linguistic diversity

Tool developed by a group at Rēzekne Academy of Technologies (Antra Kļavinska, Sanita Martena, Nicole Nau, Ilga Šuplinska, and Anna Briška). Presented by Nicole Nau.

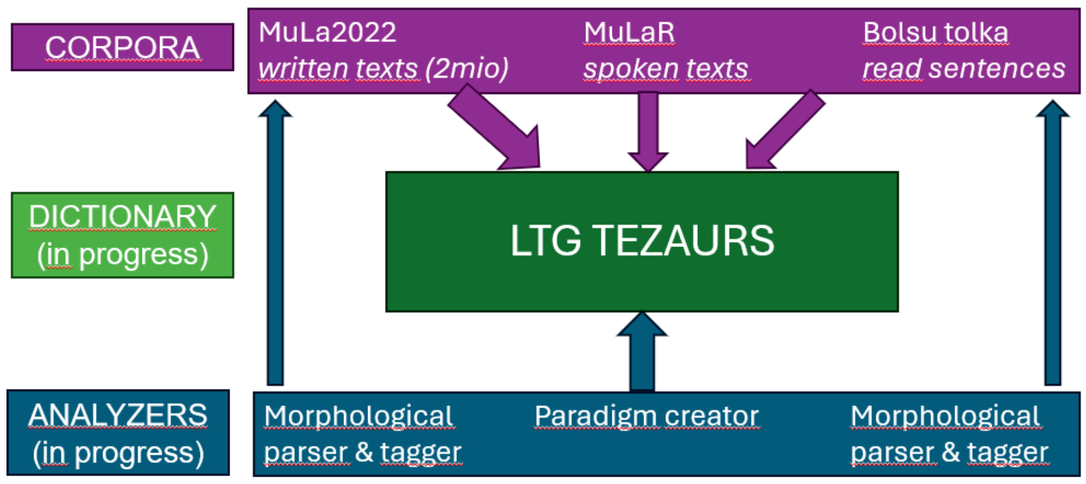
Why Latgalian?

- Latgalian is a lesser-used but vital language from Eastern Latvia with its own writing tradition.
- In Latvia, Latgalian and Standard Latvian are seen as varieties of one language – Latvian.
- Latgalian and Standard Latvian are…
  - so different that each needs its own set of resources and tools,
  - so similar that tools for Latgalian may partly be based on tools developed for Standard Latvian.

LTG TEZAURS: Three goals

- To **document** the lexicon of modern written Latgalian as well as Latgalian dialects, based on existing dictionaries and evidence from corpora.
- To support the **maintenance** of Latgalian by giving reliable information about:
  - word meaning,
  - word use in authentic texts,
  - spelling and inflection according to current standards,
  - pronunciation.
- To create a lexical database as an integral part of a **digital ecosystem** comprising corpora, dictionaries, morphological parsers and taggers, and future applications.

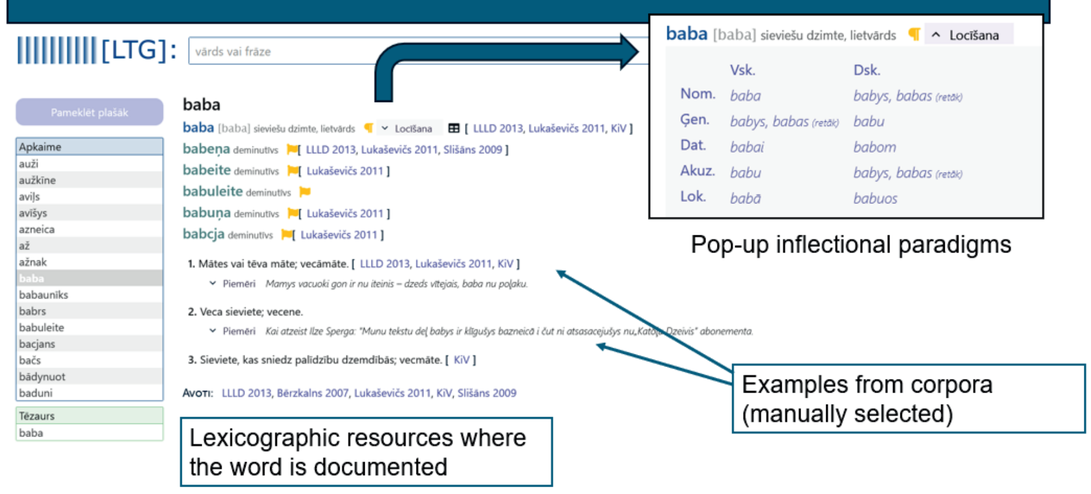
Digital resources and tools for Latgalian



How is the tezaurs created

- A group at Rēzekne Academy of Technologies (RTA) is developing a prototype with 100+ entries within the Latvian project DHELI (*Towards the Development of an Open and FAIR Digital Humanities Ecosystem in Latvia*).
- We are using the existing platform and dictionary-writing system of the Latvian *tēzaurs* (tezaurs.lv).
- We cooperate with the Artifical Intelligence Laboratory of the University of Latvia (LU MII AiLab), who is responsible for the technical side of the project.
- Crucial for our project is the successful cooperation among linguists, experts on Latgalian culture, and speech technologists.

What does an entry contain?



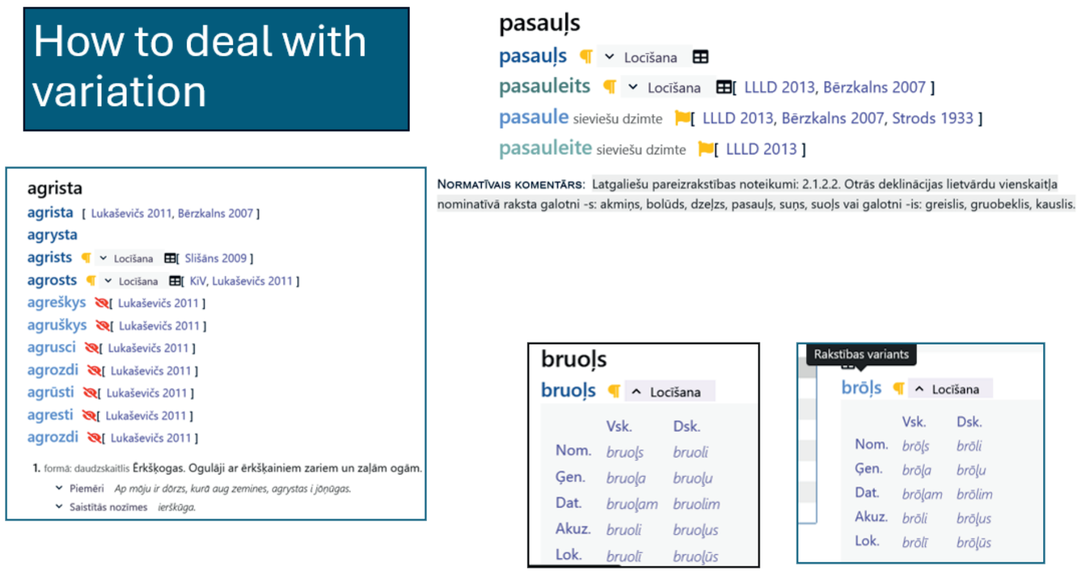
Other features

- Meaning explanations in Standard Latvian containing a translation equivalent enable the use as a bilingual dictionary.
- Semantic relations: linking synonyms
- Derivational relations: linking derived words with their base
- Pronunciation by real speakers, using audios collected by Bolsu tolka (under Common Voice)

Main challenge: inner diversity of Latgalian

- Variation in spelling and inflection **licensed** by the current standard (should be shown to users)
  - How to display in a user-friendly way?
  - Which are the preferred forms – from the point of view of users and that of language planners?
- **Non-licensed** variation (dialect forms, individual spelling, mistakes) – numerous instances in the corpora
  - All wordforms should be analyzable for lemmatization and tagging.
  - Paradigms for users should contain only licensed forms.
  - Unlicensed forms will show up in authentic examples.

How to deal with variation



This presentation was supported by the National Research Programme “Digital Humanities” project “Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia” (No. VPP-IZM-DH-2022/1-0002).

RĒZEKNES TEHNOLOĢIJU AKADĒMIJA

VPP-IZM-DH-2022/1-0002

Latvijas Zinātnes padome