Latgalian Language Corpora and Other Digital Resources in the Context of European Lesser-Used Languages

Sanita Martena, Antra Kļavinska, Ilga Šuplinska
Rēzekne Academy of Technologies

6th Baltic Summer School of Digital Humanities
Large Language Models and Small Languages

Rīga, National Library of Latvia, 24.07.2024

Overview

1. Latvian (**lvs and ltg**) in the context of languages of the world.
   Latgalian: a short sociolinguistic insight.
2. European lesser-used languages: why does it matter?
3. Latgalian: research directions Sanita (1.-3.p. – 25 min)
4. Latgalian language corpora and other digital resources and tools Antra 30 min
5. Latgalian in education Ilga 5 min

Break 10 min

Workshop Ilga 40 min

Summing up / Q&A Sanita + Ilga, Antra 10 min

1 Languages of the World

7,164 languages

Latvian

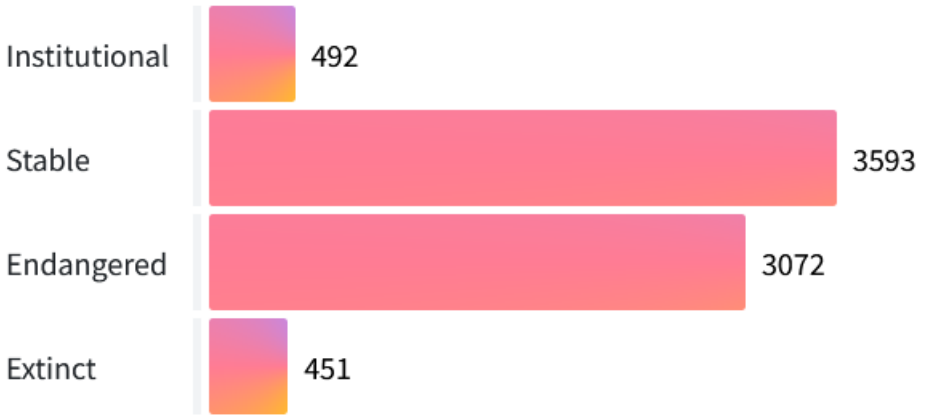
Summary

Latvian is classified as a “macrolanguage” in the ISO 639 standard and is assigned to [lav] as its three-letter code. Macrolanguages were introduced into the standard in order to reconcile the fact that in some usage contexts the entity represented by the three-letter code is deemed to be a single language, while in other usage contexts it is subdivided into two or more individual languages, each of which has its own code. The individual languages that make up this macrolanguage are listed below.

Languages

ltg Latgalian

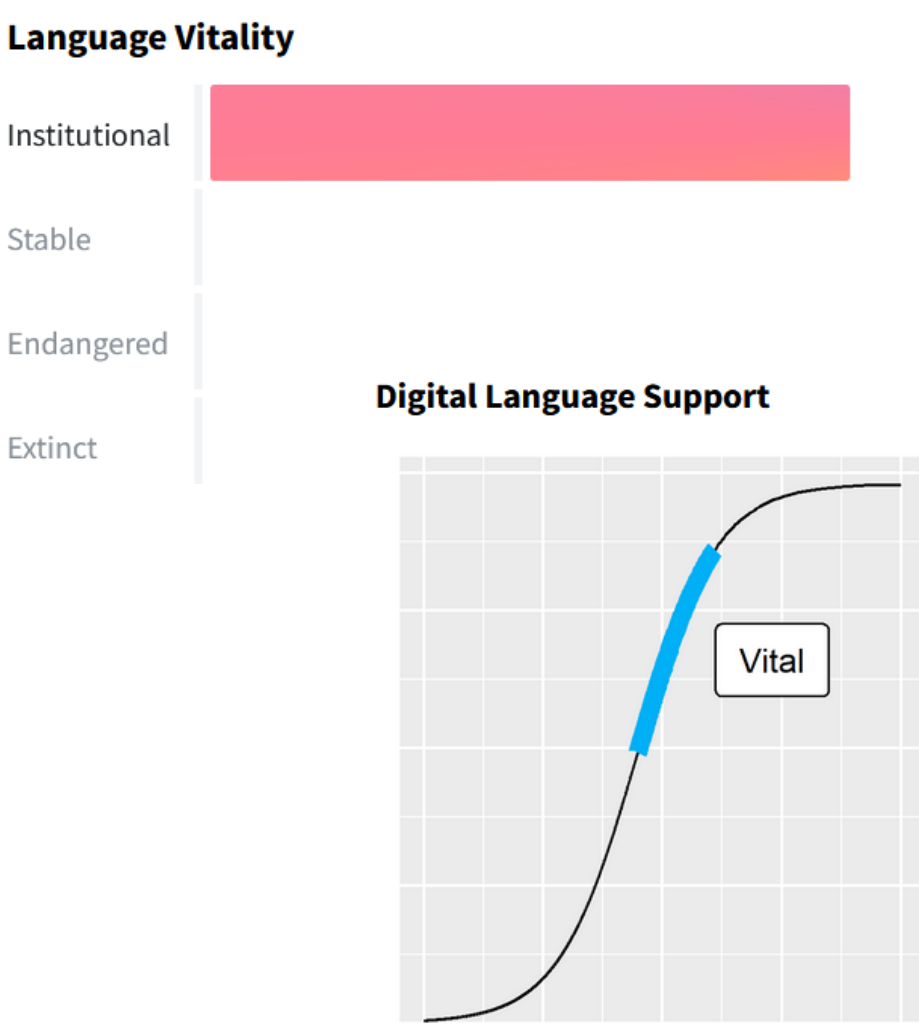
lvs Standard Latvian



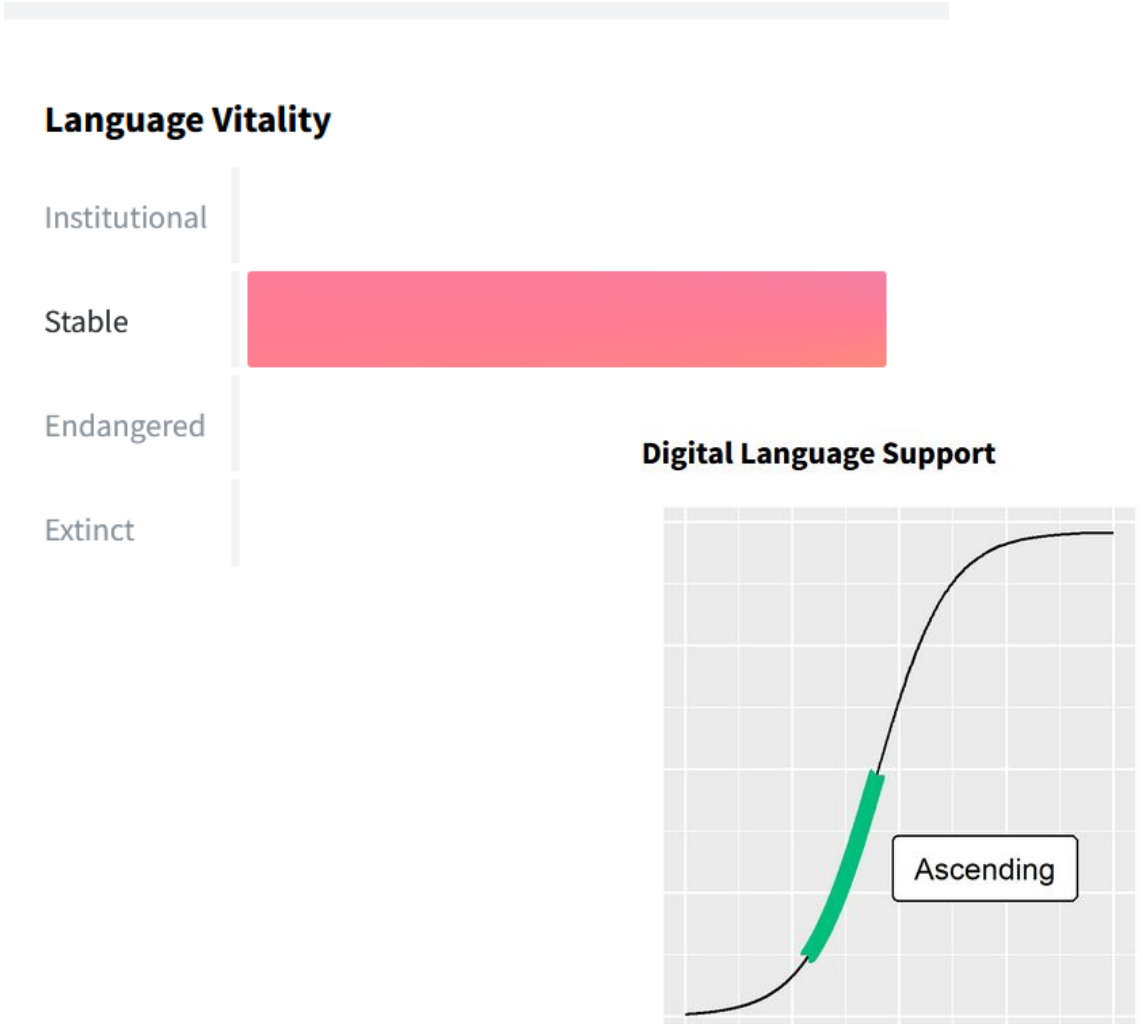
Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Language Vitality: Standard Latvian and Latgalian

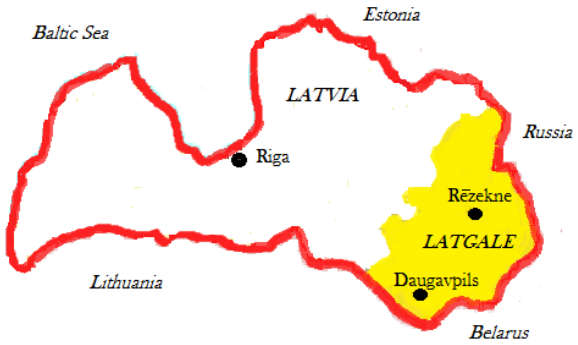
Standard Latvian (lvs)

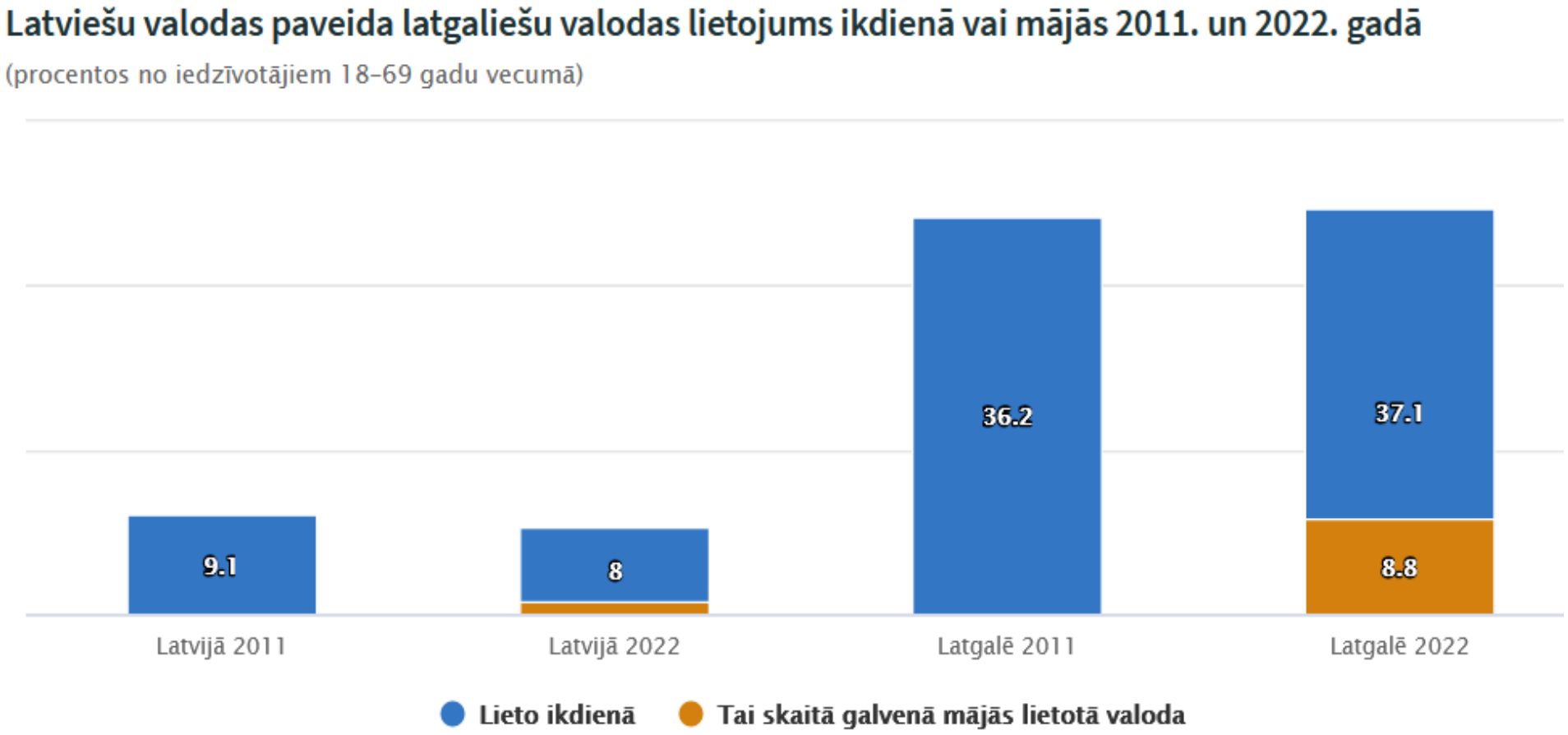


Latgalian (ltg)



1 Latgalian: a Short Sociolinguistic Insight

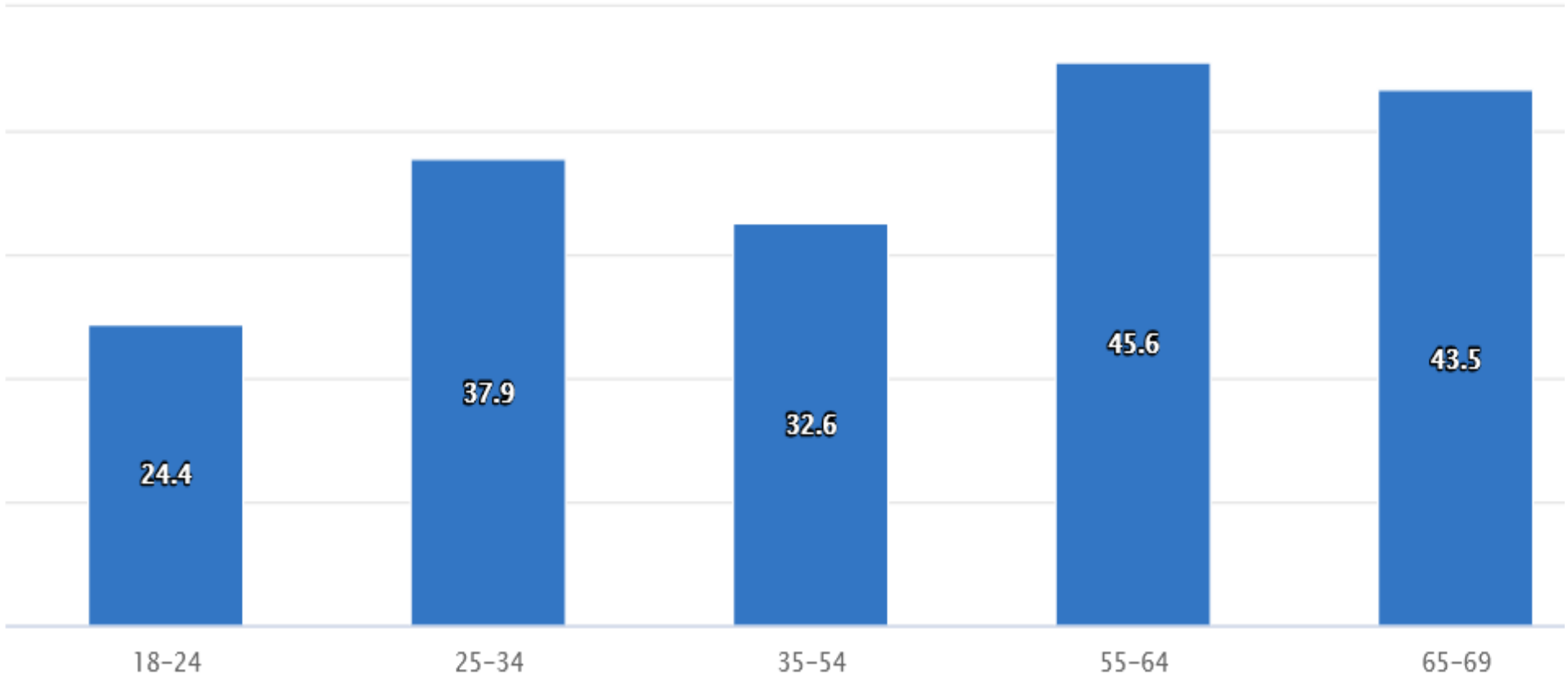




https://stat.gov.lv/lv/statistikas-temas/izglitiba-kultura-zinatne/izglitibas-limenis/preses-relizes/21078-latgaliesu

Latgales iedzīvotāju (18–69 gadi), kuri ikdienā vai mājās lieto latgaliešu valodu, īpatsvars pa vecuma grupām, 2022

(procentos no iedzīvotāju skaita attiecīgajā grupā)



https://stat.gov.lv/lv/statistikas-temas/izglitiba-kultura-zinatne/izglitibas-limenis/preses-relizes/21078-latgaliesu

https://jakubmarian.com/map-of-languages-and-language-families-of-europe/

2 European lesser-used languages: why it matters

Labelling issues:

- Lesser-used languages
- Endangered languages
- Regional languages
- Collateral languages
- Contested languages

Why does it matter?

- Part of identity, uniqueness
- An additional point of view on the world
- Richer linguistic repertoire, pluriliteracy
- Linguistic and cultural diversity
- Language rights, equality

3 Latgalian: Research directions (I)

- Economic and non-economic value of Latgalian
- Laypersons‘ perception of Latgalian and it‘s status (Folk Linguistics)
- Literacies (incl. corpus literacy) in regional collateral languages
- Language attitudes and societal discourses

Special Issue of the Journal of Multilingual and Multicultural Development on *Literacy (Development) in Collateral Regional Languages of Europe*

Routledge
Taylor & Francis Group

Journal of Multilingual and Multicultural Development

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rmmm20

When the parish turns into a classroom: teaching pluriliteracies in regional collateral languages through place-based education (the case of Latgalian)

Sanita Martena & Heiko F. Marten

Developing corpus literacy: A perspective of Latgalian language and cultural studies

Angelika Juško-Štekele & Antra Kļavinska

Social values of an added literacy: the case of Latgalian

Nicole Nau & Tomasz Wicherkiewicz

3 Latgalian: Research directions (II)

- Latgalian in Linguistic landscapes (LL)
- Latgalian in socio-onomastic perspective (survey about public attitudes towards bilingual road signs (April, 2024))

Which municipalities, communities, or individuals have been the most active in implementing the Latgalian written language in public signage? (91 responses)

Kārsava / Kuorsova, 2014

Cibla / Cybla, 2021

Photo by lakuga.lv

Nautrēni / Nautrāni, 2020

3 Latgalian: Research directions (III)

- Language education
- Recipients and writers in contemporary Latgalian literature
- Educational games:

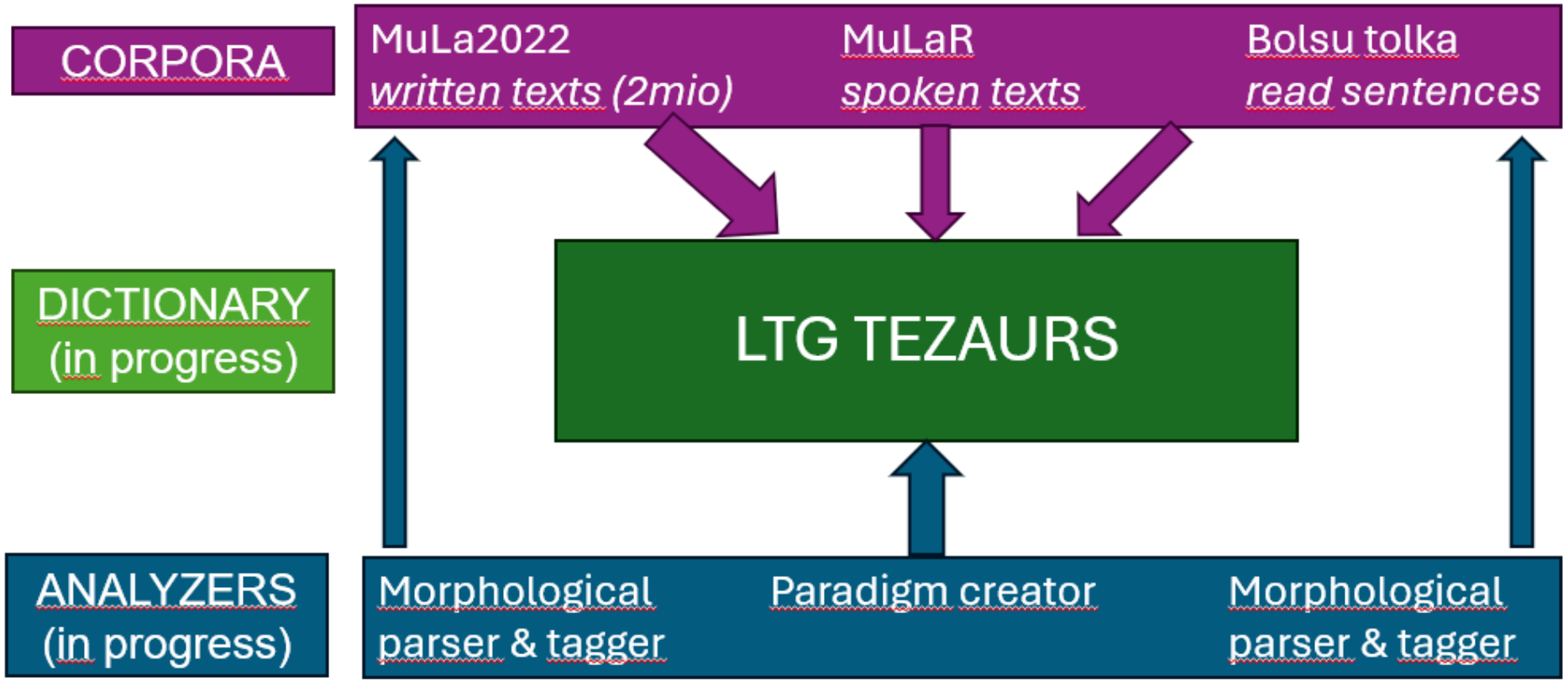
Games | futureofmuseums.eu (2014)

Iepazīsimies (rta.lv) (2017)

4 Lesser-used languages and language corpora (Martena, Nau, Briška 2021)

Language (State)	Status	Name of the Corpus, available from ... (year)	Size of the Corpus (words)
Basque (Spain)	Official Regional Language	ETC; 2013	355 million
Asturian (Spain)	No Status	ESLEMA; 2008	Not finished
Law Sorbian (Germany)	Minority Language	DTK; 2010	23 mil
Welsh (UK)	Official Regional Language	CorCenCC; 2020	11 mil
Northern Sámi (Norway)	Indigenous (autochthon) Language	SIKOR; 2015	8,9 mil
Silesian (Poland)	No Status	KŚM; 2018	2 mil
Latgalian (Latvia)	Historic Variety of the Latvian Language	MuLa; 2012 MuLa; 2022	2 mil

4 Latgalian language corpora and other digital resources and tools



(Nau, 2024)

Nacionālā korpusu kolekcija
korpuss.lv

CLARIN-LV

RTA
RĒZEKNES TEHNOLOĢIJU AKADĒMIJA

Mākslīgā intelekta laboratorija
LU MII

4 DEVELOPMENT OF CORPORA OF LATGALIAN

Opportunities and strengths:

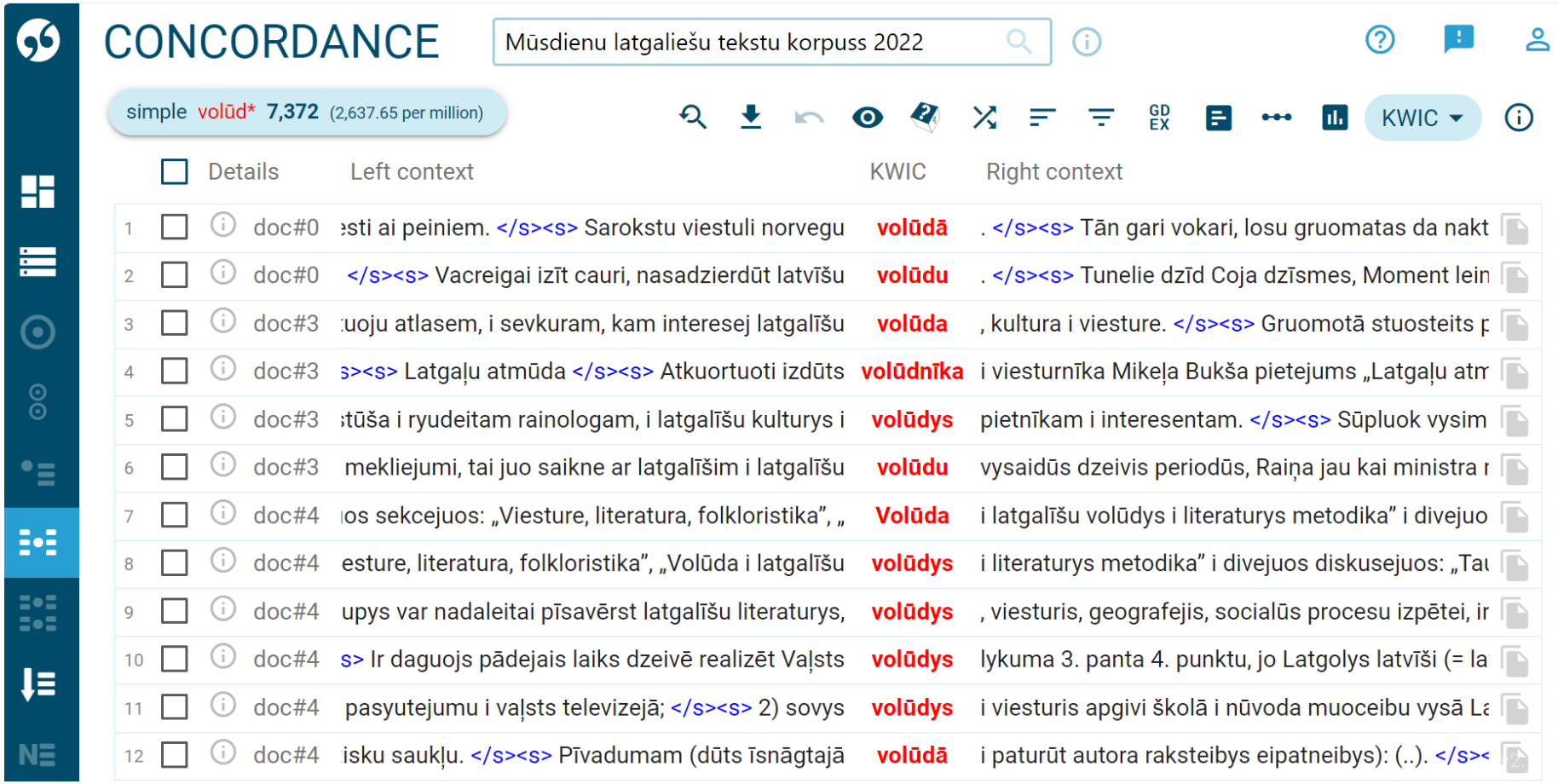
- new exciting ways of exploring Latgalian,
- long-term, stable cooperation with other European regional language research centers,
- fundraising, new projects, crowdsourcing,
- adherence to FAIR data principles and inter-institutional cooperation in data management,
- raising awareness of the value of the Latgalian and promoting it in the local community, especially among young people.

Challenges:

- **linguistic**:
  - insufficient data (texts, audio recordings),
  - subjectivity factor in manual transcription of audio recordings,
  - variation as a challenge for the morphological tagging of corpora,
- **extralinguistic:**
  - lack of financial and human resources for the planned and sustainable digitization;
  - lack of interest / information among potential users,
  - lack of corpus literacy.

MŪSDIENU LATGALIEŠU TEKSTU KORPUSS (MuLa)
CORPUS OF CONTEMPORARY LATGALIAN TEXTS

Website: https://korpuss.lv/id/MuLa2022



MYUSU DĪNU LATGALĪŠU RUNYS KORPUSS
CORPUS OF CONTEMPORARY LATGALIAN SPEECH

Website: https://mularkorpuss.rta.lv

MuLaR Corpus Statistics About Help [ltg lv en]

0:00 / 0:00

a ka īrauga tāvs sasītu tiuleņ giva rūkuos mītu

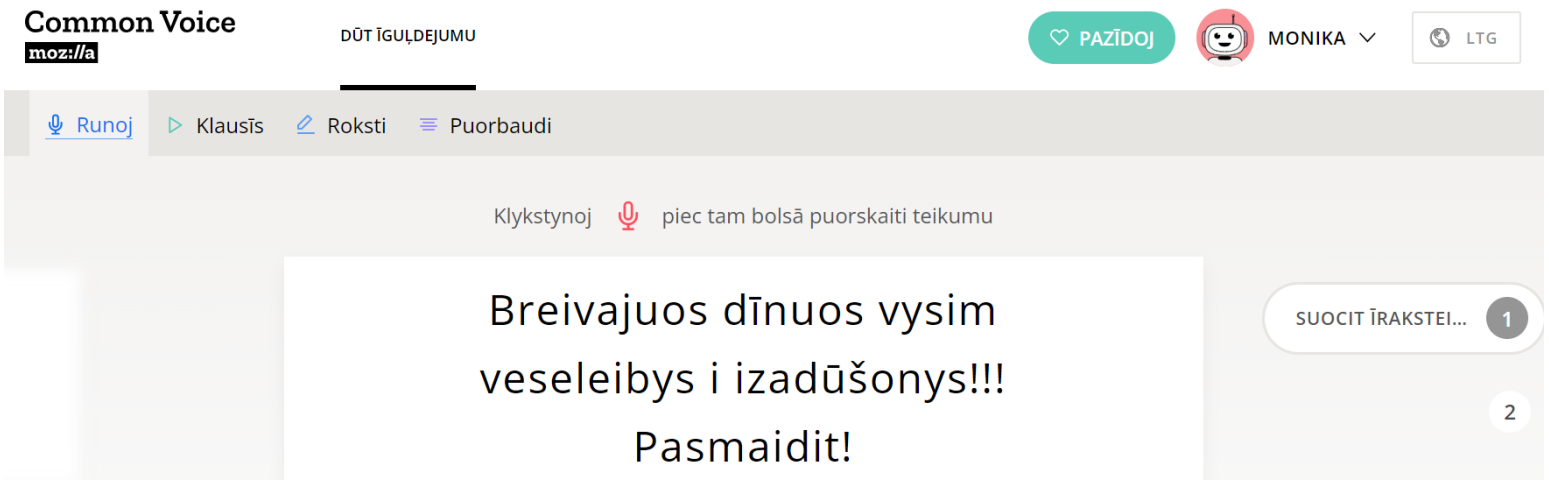
Person ID: F124-1929 **Year of birth:** 1929 **Sex:** f **Place:** Cibla
Source: L

0:00 / 0:00

F124-1929: mobilizēja syutēja a - ha tod nū - nūskrēja no tuo i atgruo - i pajēme meitu vot i navajadzēja jai iz tom šahtom braukt
io8-F: a kai jius nūklivot Krīvejī ?
F124-1929: nu tāvs bāga nu vuocīšim nabeja vēļ partejī īleids bet jau kandidatūs pats bogotuokīs i pats komunists beja īstuostēja kai nu tāvs beja jis redzi tymā

Person ID: F124-1929 **Year of birth:** 1929 **Sex:** f **Place:** Cibla
Source: L

https://korpuss.lv/en/id/BolsuTolka



BALSUTALKA.LV=

Bolsutolka.lv Speech Corpus (Common Voice 17.0)

The speech corpus includes sentences in Latgalian, read by different speakers of Latgalian dialects. The Mozilla Common Voice platform is used for data collection. Part-of-speech tagging and lemmatization has been done manually in this Latgalian corpus.

speech (9) specialised (28) latgalian (3) morphology (33) manually annotated (6)

Corpus size	24 hours (130k tokens)
Data period	2023–2024

The Latgalian TEZAURS

https://darbaversija.ltg.tezaurs.lv/

[LTG]

Pameklēt plašāk

Apkaime

avīze
azarmale
azars
azarts
azneica
až
ažnak
baba
babaunīks
babris
babrs
babuleite
bacjans
bačs
bāda

Tēzaurs

baba

baba

baba [baba] sieviešu dzimte, lietvārds ¶ Locīšana [LLLD 2013, Lukaševičs 2011, KiV]

babeņa sieviešu dzimte, deminutīvs, lietvārds ¶ Locīšana [LLLD 2013, Lukaševičs 2011, Slišāns 2009]

babeite sieviešu dzimte, deminutīvs, lietvārds ¶ Locīšana [Lukaševičs 2011]

babuleite sieviešu dzimte, deminutīvs, lietvārds ¶ Locīšana

babuņa sieviešu dzimte, deminutīvs, lietvārds ¶ Locīšana [Lukaševičs 2011]

babcja deminutīvs [Lukaševičs 2011]

1. Mātes vai tēva māte; vecāmāte. [LLLD 2013, Lukaševičs 2011, KiV]
   - Piemēri *Mamys vacuoki gon ir nu iteinis – dzeds vītejais, baba nu poļaku.*
   - Saistītās nozīmes

2. Veca sieviete; vecene.
   - Piemēri *Kai atzeist Ilze Sperga: "Munu tekstu deļ babys ir klīgušys bazneicā i čut ni atsasacejušys nu „Katōļu Dzeivis" abonementa.*
   - Saistītās nozīmes

3. Sieviete, kas sniedz palīdzību dzemdībās; vecmāte. [KiV]

Avoti: LLLD 2013, Bērzkalns 2007, Lukaševičs 2011, KiV, Slišāns 2009

Rādīt paslēptos piemērus

Rediģēt

Pievienot šķirkli

Semantiskās saites

Korpusu piemēri

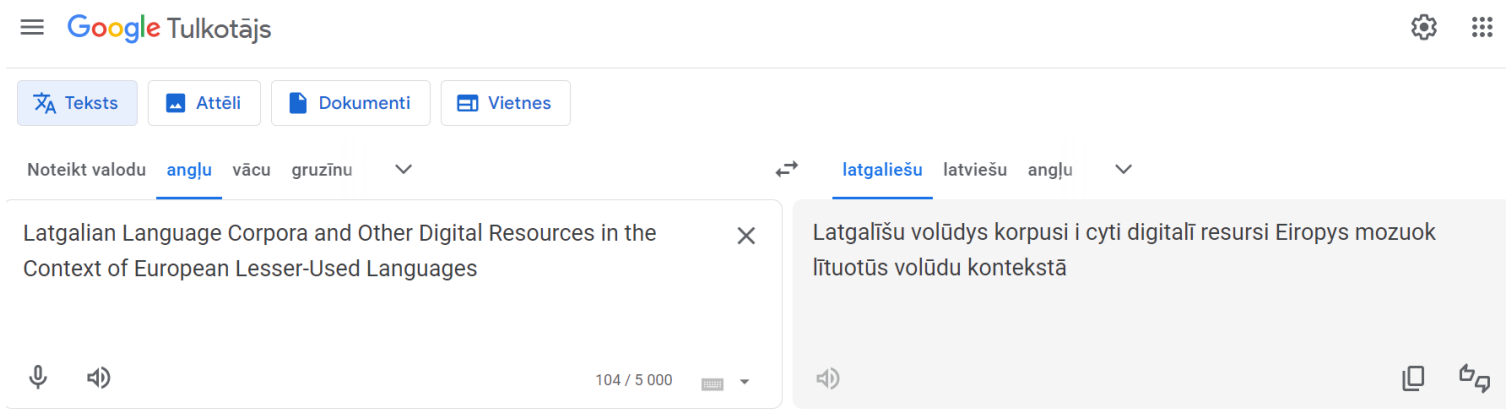
Nejaušs šķirklis

Vaicājumi

Other digital resources and tools for Latgalian:

- Online Dictionaries: Latvian-Latgalian http://vuordineica.lv/

  Lithuanian-Latvian-Latgalian http://hipilatlit.ru.lv/dictionary/lv/dictionary.html
- Latgalian Spelling Tool: https://va.hugo.lv/proofingen
- Pronunciation App: https://fryske-akademy.nl/fa-apps/trainer/
- Latgalian Verbs | Latvian Apps
- Google Translate:



5 Latgalian in Education

mercator
European Research Centre on Multilingualism and Language Learning
hosted by FRYSKE AKADEMY

https://www.mercator-research.eu/fileadmin/mercator/documents/regional_dossiers/latgalian_in_latvia_2nd.pdf

Video Lectures

Latvīšu literaruo volūda,
latgalīšu rokstu volūda,
gramatika,
volūdu korpusi,
školāns - aktivs gramatikys atkluojiejs

Korpusprateiba i cytys digitaluos prasmis, īpazeistūt Latgolai rakstureigūs personvuordus (prīkšvuordus)

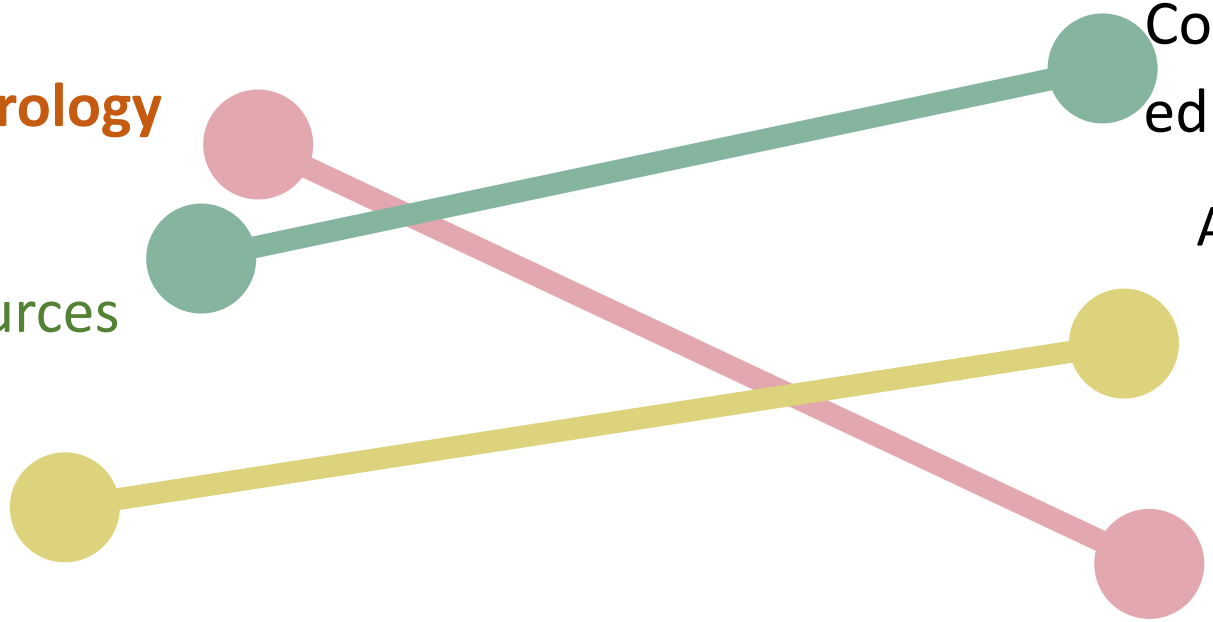
https://ltg.korpuss.rta.lv/en/video_lectures/

Latgalian language / regional studies for school, 32 sets of materials, 2 sets of methodology

Lingvoculturology

Digital resources

Students experience



Corpus of modern Latgalian texts, educational games, linguistic map

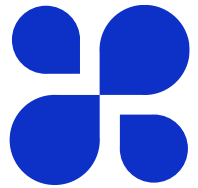
Associative, exploratory, self-guided learning approach; different studying spaces

Language in culture, culture in language; inner regularities of the language

https://www.tavaklase.lv/

Digital resources

1. https://latgalesdati.du.lv/ (the oldest resource (1993): calendar with birthdays dates, memorial dates and major events)
2. http://datatest.lnb.lv/digitala_biblioteka/laikraksti/index3.htm (Gaisma (1905-1906), Drywa (1908-1917), Zīdūnis (1921-1940), Latgolas Škola (1921-1938), Latgolas Vōrds (1919-1940), Jaunais Vōrds (1931-1940), Sauleite (1927-1940), Latgolas Bolss (1943-1944, 1955-1985), Dzeive (1948-2000))
3. www.lakuga.lv (Latgalian Culture Newspaper)
4. http://futureofmuseums.eu/lg/ (Linguoterritorial dictionary of Latgale)
5. https://www.lu.lv/filol/latgalistica/index.htm (materials for Latgalistics)
6. http://www.ltgasoc.lv/pakalpojumi/vuiceibu-materiali/
7. http://hipilatlit.ru.lv/dictionary/lv/dictionary.html
8. http://www.korpuss.lv/id/MuLa
9. http://iepazisimies.rta.lv/ (Famous personalities (33) from Latgale)
10. http://www.lingvistiskakarte.lv/info/5 (Latgalian linguistic events, books , personalities)



Digital resources

1. https://latvianapps.com (Latgalian Verbs)
2. https://oratastic.eu/latgalian-for-beginners/ (Latgalian language training, basic level)
3. https://enciklopedija.lv/skirklis/171760-Antons-Rupainis (~10 articles about Latgalian literature, linguistic)
4. https://mularkorpuss.rta.lv/#!/
5. https://owlplus.eu/latgalian/chapter/10-materiali/ https://www.tavaklase.lv/ (learning materials for school, video)

Education

https://owlplus.eu/module/

The languages of the OWL+ project

For this purpose, the OWL+ project gathers researchers and educators who work on four different autochthonous minority languages: South Saami in Norway (also spoken in Sweden), West Frisian in the Netherlands, Mirandese in Portugal, and Latgalian in Latvia.

Project "Ownership and Leadership: Pathway for (Endangered) Languages' Use in School (OWL+)" (2022–2025) partners: Friske Academy, Netherlands, Frisian), The Interdisciplinary Centre for Social and Language Documentation, Portugal, Mirandese), Nord University, Norway, South Saami), RTA, Latvia, Latgalian and Tallinn University, Estonia, technical solution and support

WEST FRISIAN

More than 400 000 people speak West Frisian in the province of Fryslân (NL). It is the province's second official

MIRANDESE

In the region of Miranda, on the Portuguese-Spanish border, the main language is Mirandese. Althoug it is prevalent with most families in the area, many people feel insecure about speaking the language in public. Historically, the

LATGALIAN

The Latgalian language, predominantly spoken in Latgale, Latvia, has faced oppression in the past. Over centuries

SOUTH SAAMI

With only 500 to 600 speakers, South Saami is an endangered minority language spoken in parts of Norway and Sweden. It is one of several Saami languages, with great historical and cultural values

Workshop

RTA
RĒZEKNES TEHNOLOĢIJU AKADĒMIJA

owl+

Co-funded by the Erasmus+ Programme of the European Union

LATGALIAN LANGUAGE E-LEARNING PLATFORM

- **Provides knowledge** about the Latgalian language, education and culture
- Offers interactive and **creative tasks** in Latgalian and in English
- Encourages you to think about who we are, where we belong, and **about our identities**
- Access to **all the major resources** currently available on Latgalian language or for learning it.

Check the platform!

https://owlplus.eu/latgalian/

The Latgalian language platform (S. Martena, H.F. Marten, I. Gusāns, I.Šuplinska)

1) provides knowledge about Latgalians who have made a difference – in former times and today, in and outside of Latgale; Latgalian culture (literature and music) and education; the Latgalian language (including some words and other language examples), its history and aspects of its use today;

2) offers interactive and creative tasks (get some information on how to count, listen, speak, and write in Latgalian; be able to compare the vocabulary of Standard Latvian and Latgalian; understand how to activate knowledge in other languages (e.g. English or Russian) in order to complete specific tasks and manage in different language situations, etc.);

3) encourages you to think about who we are, where we belong, and about our identities; builds attitudes towards the Latgalian language and allows you to understand how we can spread information about Latgale, Latgalians, and the Latgalian language to other people - wherever in the world they may be,

4) of particular importance to the user – the "Materials" section contains shortcuts to all the major resources currently available on Latgalian language or for learning it.

Latgalian – Simple Book Publishing (owlplus.eu)

Practical exercises

1.

1.1. If you attended a school in Latgale, add the year of foundation or the time of existence of your school.

1.2.If you didn't go to school in Latgale, then take the role of the author of the first Latgalian dictionary and find the Latin and Latgalian/Latvian words for Polish!

2. If necessary, read the material on the legal status of the Latgalian language in Latvia and test your knowledge by taking the test!

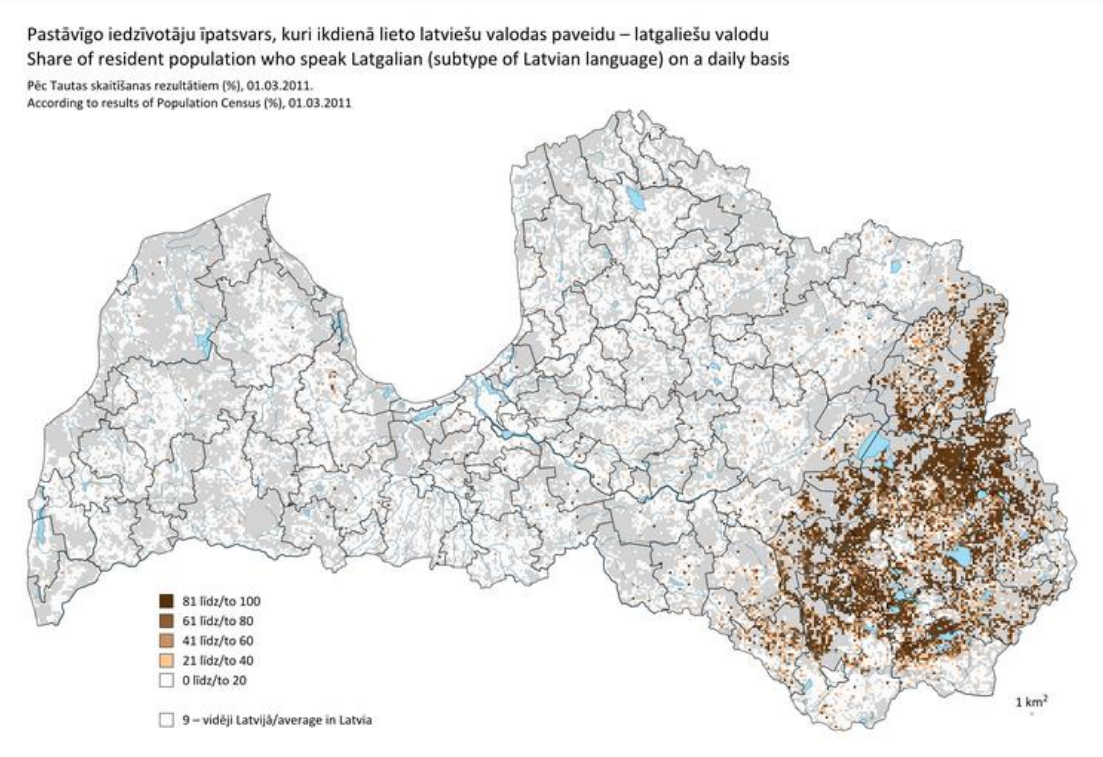
3. See the "Materials" section and the "Latgale data" resource for information on 24 July. There you will find an ordinance on the teaching of the state language in 1922. Compare the information with the situation today. How has it changed?

4. In the same chapter, check out the educational game "Let's get acquainted". Can you name how many prominent female personalities are in this educational game? Describe one of them!

Sum-up and Q/A

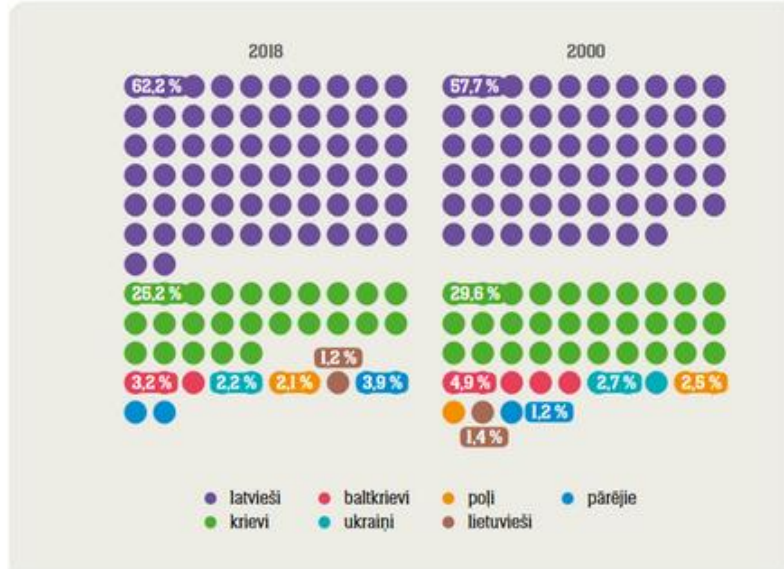
1. Look at the map and answer the following questions:

- 1) Where in Latvia is Latgalian used most often on an everday basis – in which towns and which counties / parishes?
- 2) What do you think – why do some people who live outside Latgale continue to use Latgalian? In which situations do they do so?



2. Which other languages do people in Latvia speak often at home? Name these languages in Latgalian!

https://owlplus.eu/latgalian/chapter/frisian-and-other-minority-languages-in-the-netherlands/



2. att. **Iedzīvotāju skaits pēc tautības.**
Dati: CSP nacionālais sastāvs

4. Discuss with a friend or in a group with other classmates:

- 1) What are the most common ethnicities of people who live in Latvia?
- 2) How has the proportion of ethnic groups in Latvia changed throughout almost 20 years (between 2000 and 2018)? What do think – why have there been such changes?
- 3) Try to guess what the proportion of different ethnic groups will be in Latvia in 2025. What will be different than today? Find good reasons and arguments for your prognosis!
- 4) Is ethnic group or nationality the same as language? For example, would a person who identifies himself as Polish always speak Polish, at home and in other situations?
- 5) Which are the languages spoken most often by people in Latvia?
- 6) Why is there no category “Latgalian” in the graph on ethnic groups in Latvia? (If you don’t have an idea of how to answer, look at the previous parts of this e-book)

References

➢ **About Latgalian corpora and corpus literacy**:

- Dargis, R. et al. (2024). BalsuTalka.lv - Boosting the Common Voice Corpus for Low-Resource Languages. *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, 2080-2085. https://aclanthology.org/2024.lrec-main.187.pdf
- Juško-Štekele, A.; Kļavinska, A. (2022). Creation of Contemporary Latgalian Speech Corpus in the Context of Documenting Lesser Used Languages,Mūsdienu latgaliešu valodas runas korpusa izveide mazāk lietoto valodu dokumentēšanas kontekstā. Letonica 47, 2026-243. https://lulfmi.lv/files/letonica/47/12-creation-of-contemporary-latgalian-speech-corpus-in-the-context-of-documenting-lesser-used-languages.pdf
- Juško-Štekele, A.; Kļavinska, A. (2024). Developing corpus literacy: A perspective of Latgalian language and cultural studies. *Journal of Multilingual and Multicultural Development.* DOI: 10.1080/01434632.2024.2359020
- Kļavinska, A. (2022). Standard Latgalian Spell-Checking in the Digital Environment: Text Corpora Capabilities,LATGALIEŠU RAKSTU VALODAS PAREIZRAKSTĪBAS PĀRBAUDE DIGITĀLAJĀ VIDĒ: TEKSTU KORPUSA IESPĒJAS. *Linguistica Lettica*. Vol. 30, pp. 306-322. https://doi.org/10.22364/lingualet.30
- Martena, S., Briška, A., Naua, N. (2022). The Corpus of Latgalian in the Context of Other Lesser Used Languages of Europe: Characterization, Usage and Potential Latgaliešu valodas korpuss citu Eiropas mazāk lietoto valodu kontekstā: korpusa raksturojums, lietojums un potenciālā iespējošana. *Letonica 47*, 208-2025. https://lulfmi.lv/files/letonica/Letonica_47.pdf
- Martena, S., Marten, H.F. (2024). When the parish turns into a classroom: teaching pluriliteracies in regional collateral languages through place-based education (the case of Latgalian), Journal of Multilingual and Multicultural Development
- Multimedia/learning materials https://ltg.korpuss.rta.lv/en/multimedia_-_learning_materials/

➢ **About Latgalian Tezaurs**:

- Nau, N. (2024) The Latgalian TEZAURS and the challenges of linguistic diversity. Poster presentation.Baltic DH Forum. https://drive.google.com/file/d/14Rxwj3MOZzuNDgT52MxSBodcnVpHXD-M/view
- Tezaurs.ltg (Thesaurus) – Open Lexical Database for the Latgalian language https://www.rta.lv/news/tezaurs-ltg-thesaurus-open-lexical-database-for-the-latgalian-language-25022024

➢ **About world languages:**

- Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com