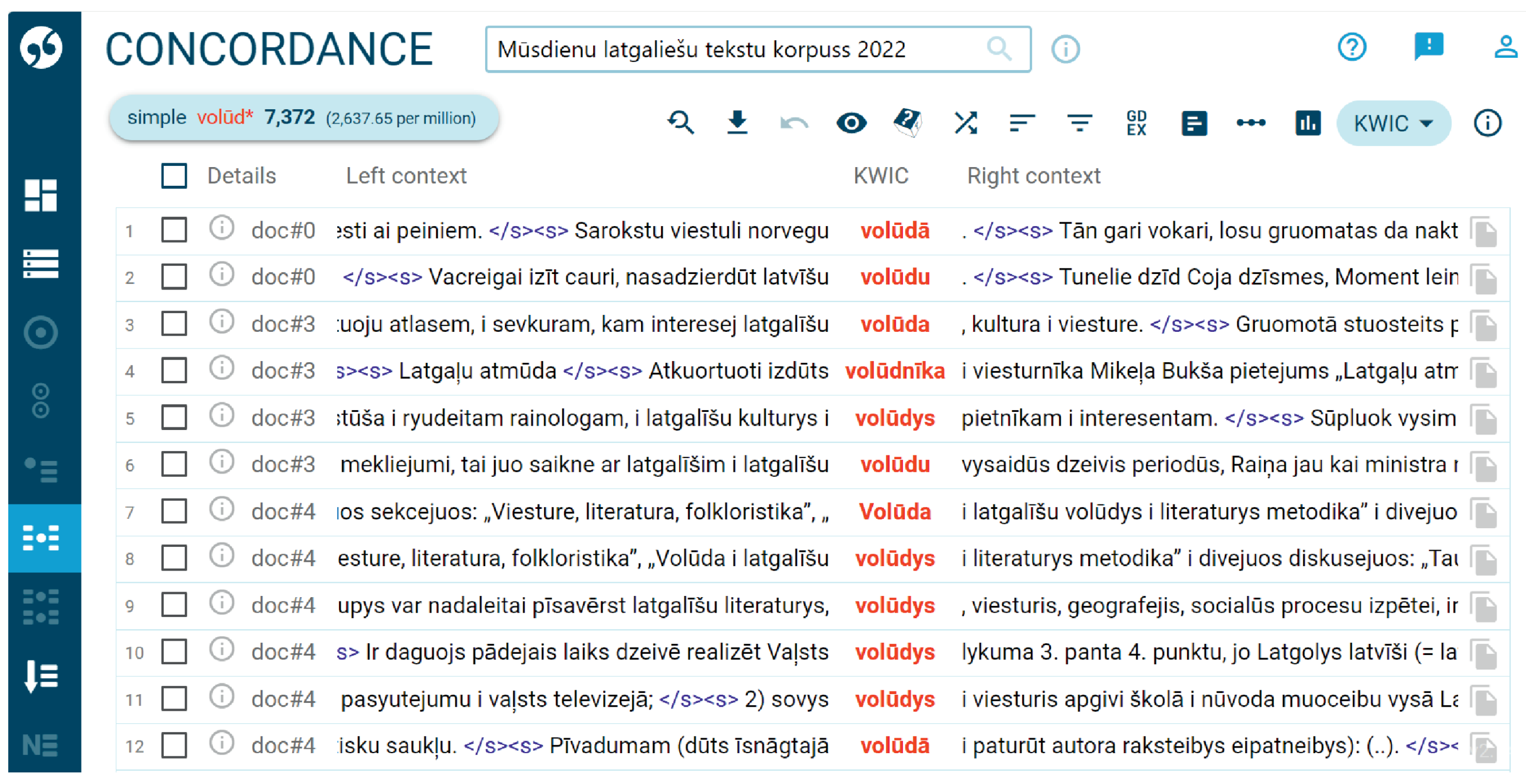
CORPORA OF CONTEMPORARY LATGALIAN TEXTS AND SPEECH

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

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



CONCORDANCE

Mūsdienu latgaliešu tekstu korpuss 2022

simple volūd* 7,372 (2,637.65 per million)

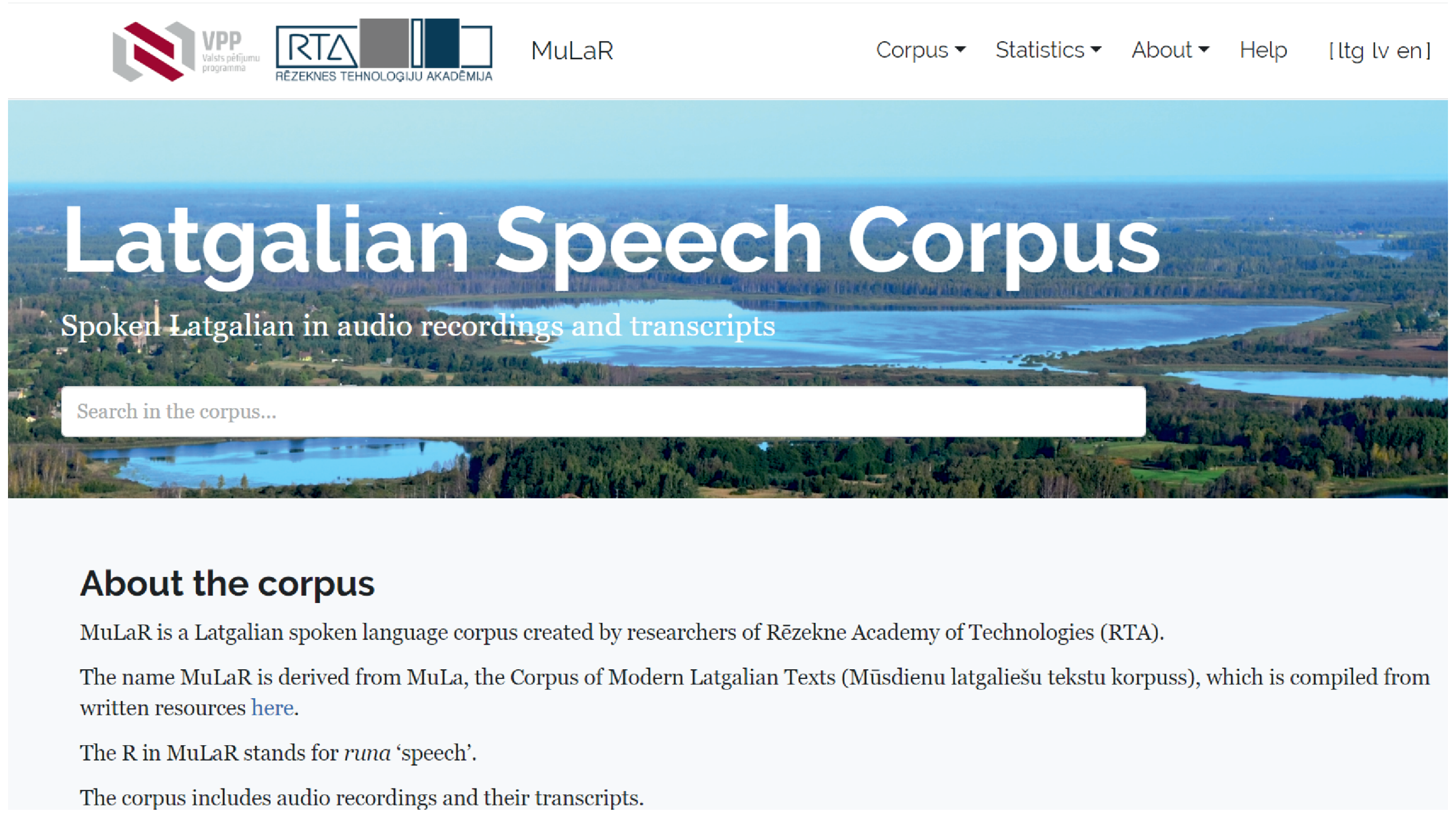
KWIC

	Details		Left context	KWIC	Right context
1	☐	doc#0	ɛsti ai peiniem. </s><s> Sarokstu viestuli norvegu	volūdā	. </s><s> Tān gari vokari, losu gruomatas da nakt
2	☐	doc#0	</s><s> Vacreigai izīt cauri, nasadzierdūt latvīšu	volūdu	. </s><s> Tunelie dzīd Coja dzīsmes, Moment leiп
3	☐	doc#3	:uoju atlasem, i sevkuram, kam interesej latgalīšu	volūda	, kultura i viesture. </s><s> Gruomotā stuosteits p
4	☐	doc#3	s><s> Latgaļu atmūda </s><s> Atkuortuoti izdūts	volūdnīka	i viesturnīka Mikeļa Bukša pietejums „Latgaļu atm
5	☐	doc#3	ɪtūša i ryudeitam rainologam, i latgalīšu kulturys i	volūdys	pietnīkam i interesentam. </s><s> Sūpluok vysim
6	☐	doc#3	mekliejumi, tai juo saikne ar latgalīšim i latgalīšu	volūdu	vysaidūs dzeivis periodūs, Raiņa jau kai ministra ı
7	☐	doc#4	ıos sekcejuos: „Viesture, literatura, folkloristika", „	Volūda	i latgalīšu volūdys i literaturys metodika" i divejuo
8	☐	doc#4	esture, literatura, folkloristika", „Volūda i latgalīšu	volūdys	i literaturys metodika" i divejuos diskusejuos: „Taı
9	☐	doc#4	upys var nadaleitai pīsavērst latgalīšu literaturys,	volūdys	, viesturis, geografejis, socialūs procesu izpētei, ir
10	☐	doc#4	s> Ir daguojs pādejais laiks dzeivē realizēt Vaļsts	volūdys	lykuma 3. panta 4. punktu, jo Latgolys latvīši (= la
11	☐	doc#4	pasyutejumu i vaļsts televizejā; </s><s> 2) sovys	volūdys	i viesturis apgivi školā i nūvoda muoceibu vysā Lɛ
12	☐	doc#4	isku saukļu. </s><s> Pīvadumam (dūts Īsnāgtajā	volūdā	i paturūt autora raksteibys eipatneibys): (..). </s><

- The current version **MuLa 2022** contains balanced samples of Latgalian texts published 1988–2021 with accompanying metadata about the author, place and year of the publication, as well as information about the type and genre of the text.
- Construction periods: 2011 – 2013; 2020 – 2022.
- Current size: 2,000,000 words, 2,800,000 tokens.
- Publishers: Rezekne Academy of Technologies, AiLab IMCS University of Latvia.
- Authors: Anna Briška, Ilze Ziņģe, Roberts Darģis, Krtistīne Pokratniece, Sanita Martena, Antra Kļavinska.
- Funding: National Research Programme project "Digital Resources for Humanities: Integration and Development" (No. VPP-IZM-DH-2020/1-0001).
- Corpus management platform: NoSketch Engine.
- The corpus is included in the Latvian National Corpora Collection (LNCC).

MYUSU DĪNU LATGALĪŠU RUNYS KORPUSS
CORPUS OF CONTEMPORARY LATGALIAN SPEECH

Website: *https://mularkorpuss.rta.lv*



Latgalian Speech Corpus

Spoken Latgalian in audio recordings and transcripts

About the corpus

MuLaR is a Latgalian spoken language corpus created by researchers of Rēzekne Academy of Technologies (RTA).

The name MuLaR is derived from MuLa, the Corpus of Modern Latgalian Texts (Mūsdienu latgaliešu tekstu korpuss), which is compiled from written resources here.

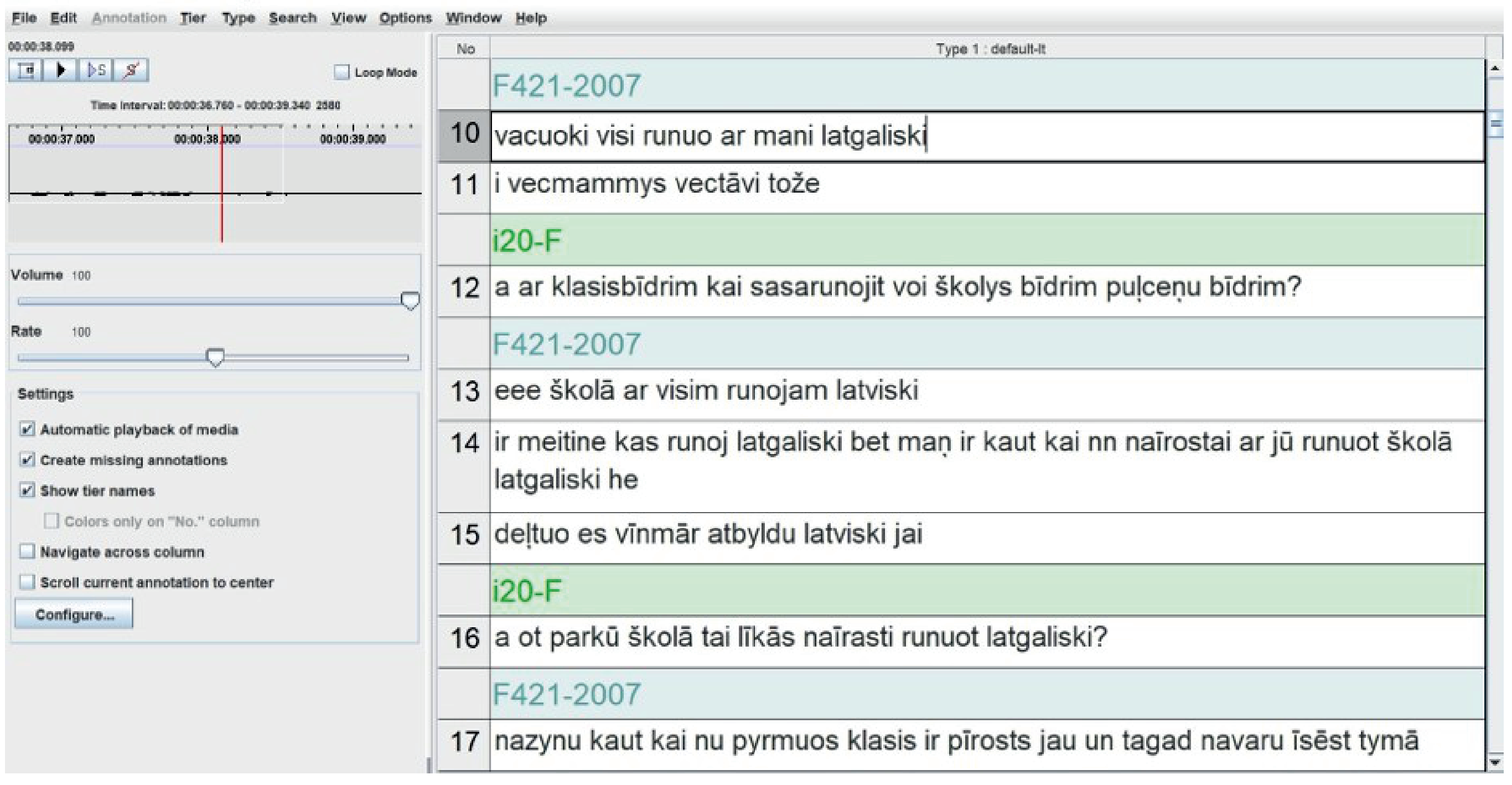
The R in MuLaR stands for *runa* 'speech'.

The corpus includes audio recordings and their transcripts.

- Construction periods: 202 – 2022, 202 – 2024.
- Current size: 23 hours of audio recordings, with transcribed text totalling 187296 tokens.
- Publisher: Rezekne Academy of Technologies.
- Authors: Sanita Martena, Nicole Nau, Antra Kļavinska, Angelika Juško-Štekele, Armands Kociņš-Kūceņš, Ausma Sprukte, Anna Briška, Ingars Gusāns, Laura Mazure.
- Funding: National Research Programme project "Digital Resources for Humanities: Integration and Development" (No. VPP-IZM-DH-2020/1-0001); National Research Programme project "Diversity of Latvian in Time and Space" (No. VPP-LETONIKA-2021/4-0003).
- Web Interface: SpoCo.

Metadata:

- location and time of the recording,
- the duration of the audio segment,
- the speaker's gender and age.

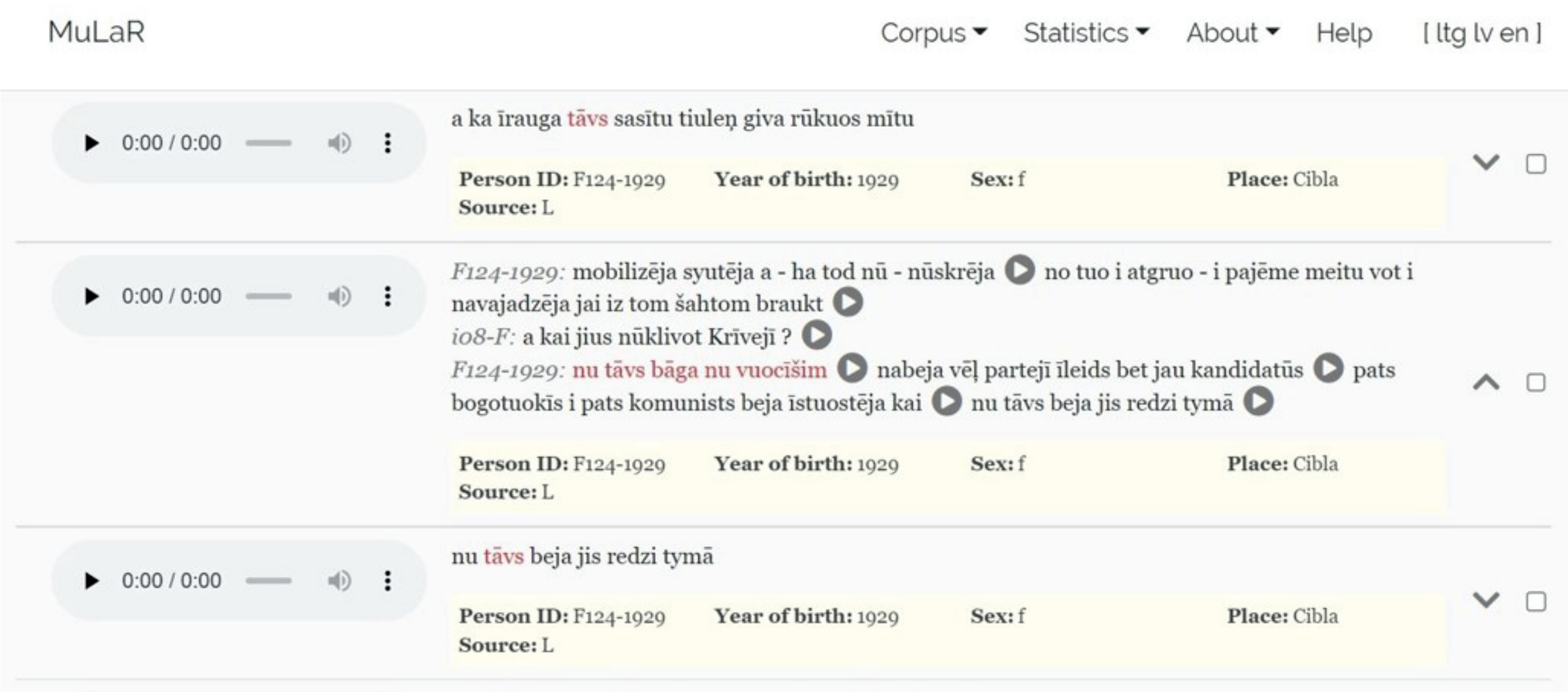


☑ Automatic playback of media
☑ Create missing annotations
☑ Show tier names
☐ Colors only on "No." column
☐ Navigate across column
☐ Scroll current annotation to center

F421-2007
10 vacuoki visi runuo ar mani latgaliski
11 i vecmammys vectāvi tože
i20-F
12 a ar klasisbīdrim kai sasarunojit voi školys bīdrim pulceņu bīdrim?
F421-2007
13 eee školā ar visim runojam latviski
14 ir meitine kas runoj latgaliski bet maņ ir kaut kai nn naīrostai ar jū runuot školā latgaliski he
15 deļtuo es vīnmār atbyldu latviski jai
i20-F
16 a ot parkū školā tai līkās naīrasti runuot latgaliski?
F421-2007
17 nazynu kaut kai nu pyrmuos klasis ir pīrosts jau un tagad navaru īsēst tymā

Content:

- **audio recordings:**
  - interviews conducted during field work in Latgale and Siberia (2009–2021),
  - TV and radio broadcasts (2018–2023),
- **transcriptions:**
  - made using the ELAN software,
  - orthographic transcription that preserves dialeçt features.



a ka īrauga tāvs sasītu tiuleņ giva rūkuos mītu

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

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 parteijī īleids bet jau kandidatūs pats bogotuokīs i pats komunists beja īstuostēja kai nu tāvs beja jis redzi tymā

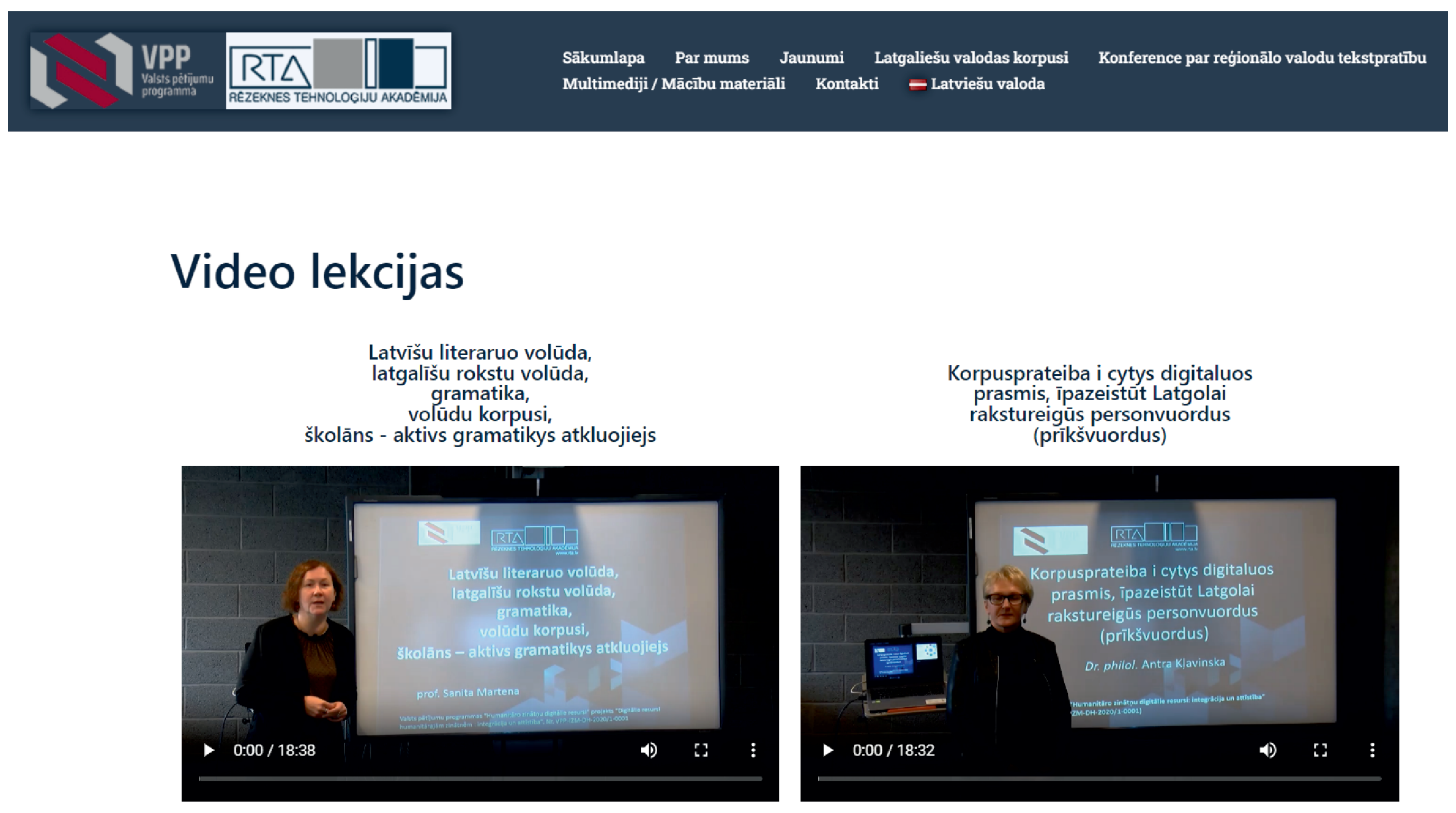
Person ID: F124-1929 **Year of birth:** 1929 **Sex:** f **Place:** Cibla
Source: L

nu tāvs beja jis redzi tymā

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

DEVELOPMENT OF CORPUS LITERACY

- Study courses "Language corpora in the educational process", "Language technologies" for RTA teacher program students.
- Video lectures for Latgalian language teachers about the possibilities of using corpora.
  Website: *https://ltg.korpuss.rta.lv/video_lekcijas/*


VPP Valsts pētījumu programma
RTA RĒZEKNES TEHNOLOĢIJU AKADĒMIJA
Sākumlapa
Par mums
Jaunumi
Latgaliešu valodas korpusi
Konference par reģionālo valodu tekstpratību
Multimediji / Mācību materiāli
Kontakti
Latviešu valoda
0:00 / 18:38
0:00 / 18:32
prof. Sanita Martena
Dr. philol. Antra Kļavinska

Video lekcijas

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

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

DEVELOPMENT OF CORPORA OF LESSER-USED LANGUAGE

Opportunities and strengths:

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

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.

This presentation was supported by the National Research Programme project "Diversity of Latvian in Time and Space (LADI)" (No. VPP-LETONIKA-2021/4-0003). Presented by Antra Kļavinska.

RTA RĒZEKNES TEHNOLOĢIJU AKADĒMIJA

VPP Valsts pētījumu programma

Latvijas Zinātnes padome