

Mūsdienu latgaliešu valodas korpusa (MuLa) „jaunā elpa”

Anna Briška anna.briska@rta.lv,
Rēzeknes Tehnoloģiju akadēmija

Konference "Latviešu valodas digitālie resursi un rīki izglītībai", Zoom 02.06.2021.



Latvijas Zinātnes
padome

Projekts “Humanitāro zinātņu
digitālie resursi:
integrācija un attīstība”
(Nr. VPP-IZM-DH-2020/1-0001)

Galvenie jautājumi priekšlasījumā

- Kas ir Mūsdienu latgaliešu tekstu korpus "MuLa"?
- MuLa 2012 un MuLa 2022 salīdzinājums.
- Korpusa lietotāju anketēšanas rezultāti.
- Jaunā sadaļa – MuLa runas korpus.

Latgaliešu tekstu korpus “MuLa”

- Izveidots laikā no 2010. līdz 2012. gadam, kopš 2020.gada oktobra papildināšana: gan ar tekstiem, gan ar runas korpusu.
- Apjoms: 1 miljons vārdlietojumu.
- Teksti, kas publicēti laika posmā no 1988. līdz 2012. gadam (papildināšana ar tekstiem līdz 2022.gadam).
- <http://hipilatlit.ru.lv/lv/> un <http://www.korpuss.lv/id/MuLa>
- Veidotāji: Rēzeknes Tehnoloģiju akadēmija (Rēzeknes Augstskola) un Latvijas Universitātes Matemātikas un informātikas institūts, projekta partneris 2010.-2012.gadā arī Vītauta Dižā Universitāte Lietuvā.

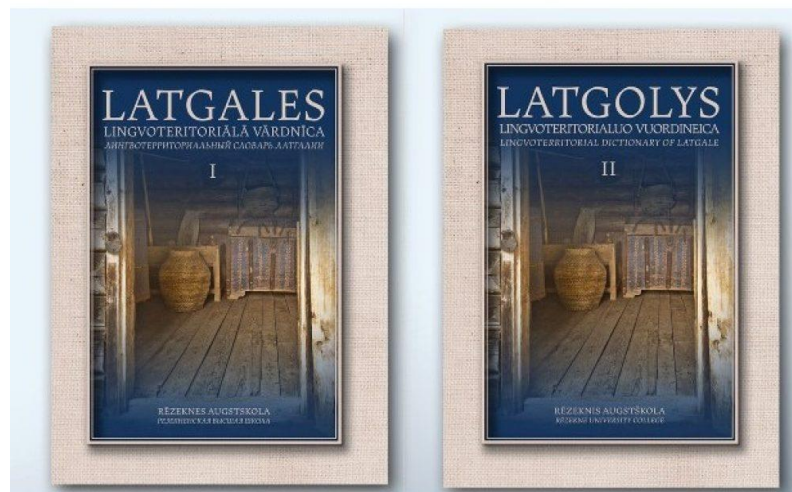


MuLa 2012 → MuLa 2022

	MuLa 2012	MuLa 2022 plānots
Izstrādes laiks	2010 – 2012	2020 - 2022
Tekstu datējums	1988 – 2012	2012 - 2022
Tekstu veidi		
Informatīvie	35 % → 36,89 %	40 %
Literārie	30 % → 40,45 %	40 %
Tehniski speciālie	35 % → 22, %	20 %
Paraugu lielums	13 – 33200 vārdlietojumu	Inf. un tehn. spec. teksti līdz 2 500 vārdlietojumu Lit. teksti līdz 20 000 vārdlietojumu

Iekļaujamo tekstu autoru raksturojums

- MuLa 2012: ap 390 autoru, entuziasti.
- MuLa 2022: latgalisko tekstu radīšana ir arī algots darbs (galvenokārt projektos):
 - lakuga.lv (~4500 publicētu rakstu 2013-2021),
 - RTA izdotā "Lingvoteritoriālā vārdnīca" (300 šķirķļu LV, Ltg, RU, EN),
 - Latvijas Radio Latgales studija (LSM.lv),
 - Žurnāls "A 12",
 - Reģionālā prese: "Latgales Laiks", "Vietējā Latgales Avīze" u.c.



Attēli no lakuga.lv

lekļaujamo tekstu raksturojums

- MuLa 2012: teksti dažādās ortogrāfijās:

Tuoļuok sekoj ari aproksti par tū , ka **muote** nagatavieja nikuo bārna pīdzimšonai
Reigā ir breineigi kopī , tī dzeivoj **Muote** Latveja ar vaiņuku golvā , laikam napracāta
Jī auguši pi tāva ar muoti seši , **muote** vysim vucejuse puoterus , svātdīnēs
seikom kuojeņom jau aizaveļ . Vokorā **muote** savad klāvā , sasauc perekli . Jau tyka

rykšōtōjus saceikstem Reigas hipodromā . **Mōte** Anna – dorbūs čakla un dzili ticeiga
Tai kai Aloizs beja izskateigs , cara **mōte** Marija Fjodorovna jam pīvērse uzmaneibu
dzejnīks , roksti nu nikō ... Ak vai ! Dzejas **mōte** sovīm " skūlānim " vuceja pavysam
vērteibas , dzimtaine , gimine – tāvs , **mōte** , bārni ... i pošreizejōs Latvijas

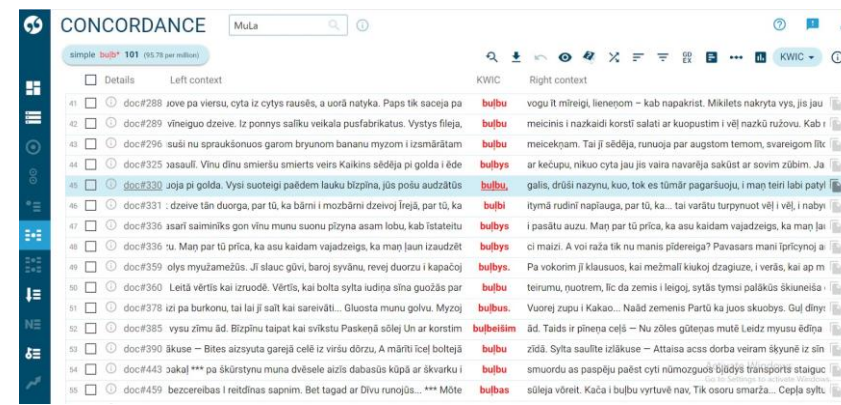
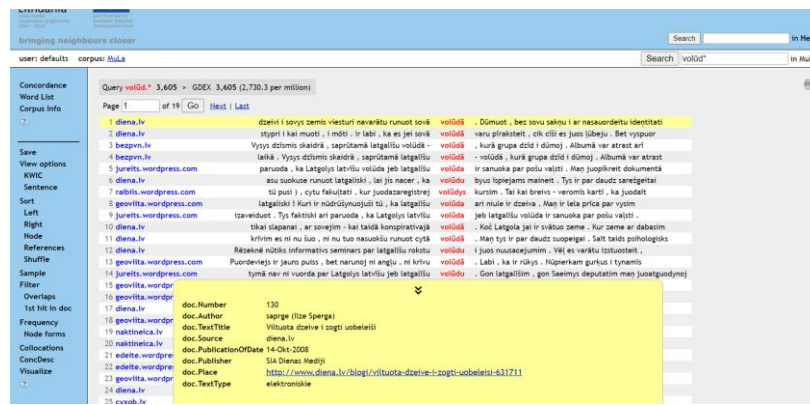
jaušama blov . Labeiba beja vai kleitas , Kū **mote** globoja reiz ? Klūgas un solmi veitī

- MuLa 2022:

- Ir apzinātas ~95 izdotas grāmatas (kopš 2012.g. novembra), kurās visi teksti vai daļa ir latgaliešu rakstu valodā.
- Lielākā daļa tekstu ir rediģēti pēc 2007.gadā pieņemtajiem ortogrāfijas noteikumiem, līdz ar to ir iespējas automatizētai tekstu marķēšanai nākotnē.

“MuLa” pielietojuma iespējas

- Platformās pieejamie rīki:
 - Vārdu saraksts (word list),
 - Konkordance – meklējamais vārds/vārda daļa kontekstā.
- Kur un kā var (varētu) izmantot?
 - Izglītībā – valodu apguvē.
 - Zinātnē – vārdnīcu izstrāde, pētījumi par valodu, kultūru, vēsturi.
 - Uzņēmējdarbībā – zīmolu izstrāde, produktu apraksta veidošana.
 - Žurnālistikā – satura veidošana.
 - Privāti – redzesloka paplašināšana, izklaide.



Kur “MuLa” ir izmantots līdz šim?

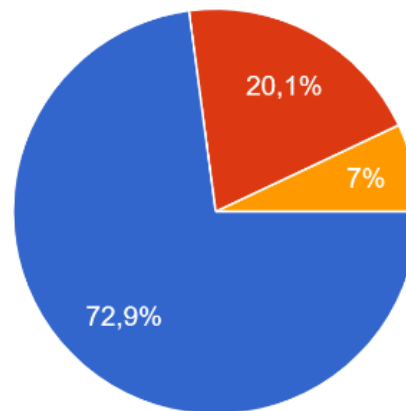
- Korpusā balstīti pētījumi, piemēram:
 - Anna Briška (2013). Mūsdienu latgaliešu rakstu valodas korpuss: izveide un izmantošanas iespējas. Maģistra darbs. RTA.
 - Nicole Nau (2014). Differential object marking in Latgalian. Grammatical Relations and their Non-Canonical Encoding in Baltic. John Benjamins Publishing Company, pp 207 – 255.
- Izglītībā:
 - Latviešu valodas rokasgrāmata (valodasrokasgramata.lv)
 - Nicole Nau (2019). Bringing Latgalian into schools: visions, needs and obstacles. 13th Conference on Baltic Studies in Europe. <http://naunicol-e.home.amu.edu.pl/wp-content/uploads/2019/09/NauCBSE2019Gdansk.pdf>
- Atsevišķu pētnieku rakstos:
 - Antra Kļavinska (etnonīmi), Sandra Ūdre (frazēoloģija, leksika), Baiba Saulīte.

Lietotāju anketēšana 2021

- Laiks: 2021.gada marts – aprīlis.
- Uzdevumi:
 - **noskaidrot**, cik lielā mērā un kādos nolūkos korpuss tiek lietots,
 - **apzināt**, ar kādām grūtībām saskaras korpusa lietotāji.
- Respondenti: 214.

Cik labi pazīstat latgaliešu tekstu korpusu MuLa?

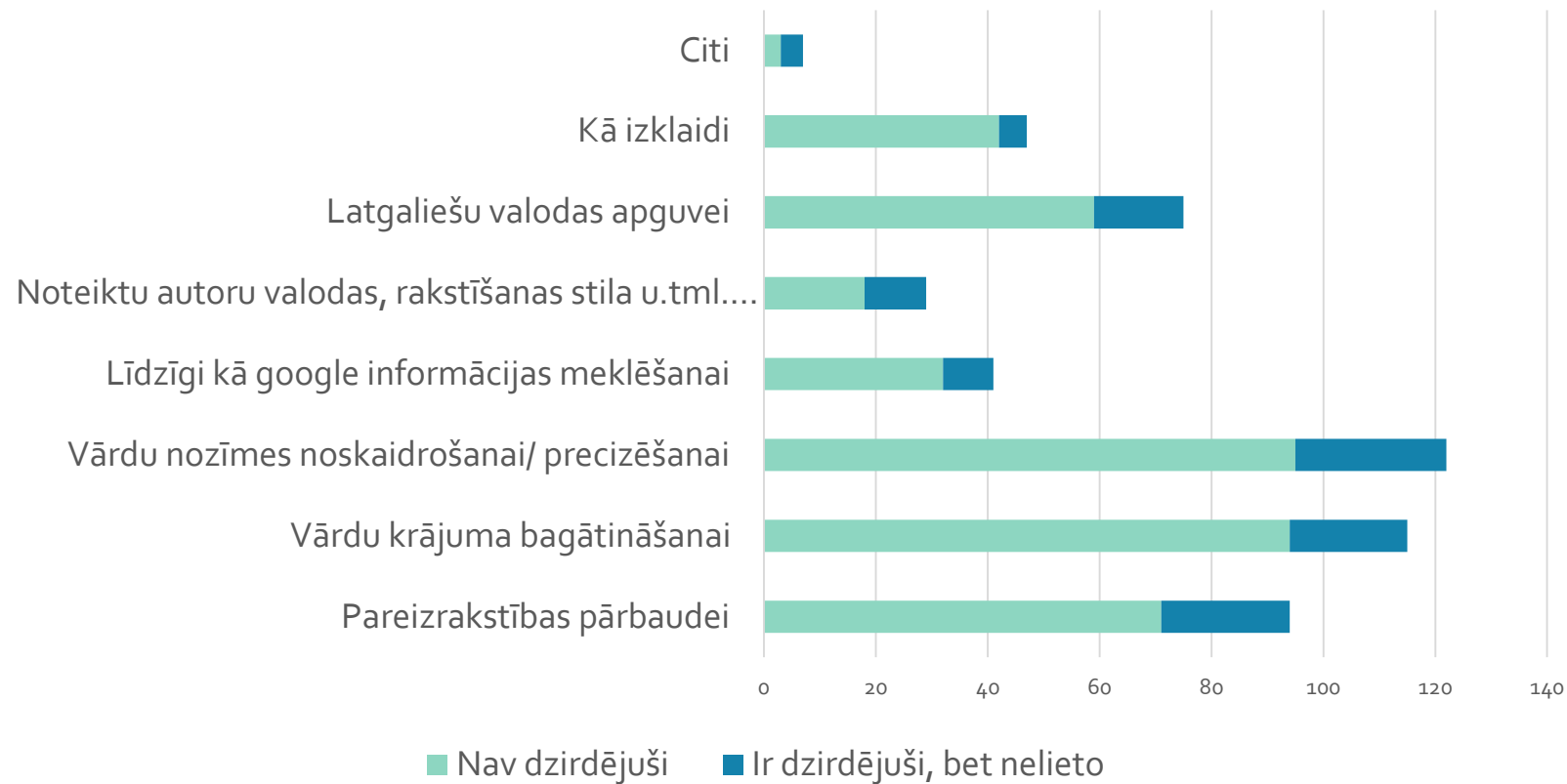
214 atbildes



- Pirmo reizi dzirdu – kas tas tāds?
- Esmu dzirdējis/-usi par to, bet nelietoju
- Zinu, esmu lietojis/-usi

Korpusa pielietojums

Kā Jums šķiet, kā nākotnē latgaliešu tekstu korpuss varētu
Jums noderēt?



Korpusa pielietojums

- Kādam nolūkam korpusu izmanto tā lietotāji (15), Top3:
 - **Meklēju piemērus, kas parāda, kā lieto vārdu - 13**
 - **Meklēju vārdu kontekstā, lai labāk saprastu tā nozīmi - 8**
 - **Gribēju pārbaudīt rakstību - 7**
 - Pētīju konkrētu valodas parādību zinātniskam darbam - 4
 - Vēlējos pārbaudīt kādus faktus - 4
 - Ieguvu kultūrvēsturisku informāciju - 3
 - Gribēju uzzināt, kā loka vārdu - 1
 - Veicu skolotāja/as uzdotus uzdevumus - 0
 - Veidoju savu vārdnīcu - 0
 - Cits -0.

Ko tālāk?

Trūkumi	Risinājumi
Korpuss ir par mazu	Paplašināšana (jauni teksti, dažādi periodi – teksti latgaliski ir saglabājušies kopš 1753. gada)
Nevar atrast visas viena vārda formas uzreiz	Marķēšana
Nevar meklēt pēc gramatiskajām kategorijām (piemēram, pēc darbības vārda laika)	Marķēšana
Korpusa platformā pārāk maz palīdzības (informācijas), kā var lietot korpusu	Instrukcijas, video pamācības (ir plānots kā projekta rezultāts) dažādām mērķgrupām
Nesaprotu platformas uzrakstus un aprakstus angļu valodā	Instrukcijas, video pamācības dažādām mērķgrupām, platforma latviešu/latgaliešu valodā

“MuLa” runas korpuss

- Plānotais apjoms 10 stundas runas transkriptu (ELAN programmā).
- Runas avoti:
 - Folkloras un etnolingvistiskās ekspedīcijas (pēc 2012. gada);
 - Sibīrijas latgaliešu diasporas ekspedīcijas (2017-2018);
 - Radio un TV raidījumi;
 - Lauka pētījums 2021;
 - TriMCo korpusa dati.



“MuLa” runas korpuss

- Platformās pieejamie rīki:
 - Vārdu saraksts (word list);
 - Statistika – vārdlietojumu skaits katram teicējam;
 - Meklētājs – pamata (basic), gan pēc regulārajām izteiksmēm (COL).
- Meklētāja rezultāti:
 - Runas ieraksta fragments (audio);
 - Runas atšifrējums (teksts);
 - Konteksta paplašināšana.

Recording	Transcription	
▶ 0:00 / 0:00 — 🔊 ⋮	varbyut ka tys beja eistīs laiks ka ka ka varbyut jau tī vyss beja izsmalts un un gribējēs vēl kū jaunu	▼ □
▶ 0:00 / 0:00 — 🔊 ⋮	i daīt tys laiks , apsamaucīs i zīm , snīga daudz pi myusu beja	▼ □
▶ 0:00 / 0:00 — 🔊 ⋮	nu , soka : " gaileit , gaileit . . . " vysod pa - kab laiks . . . laikā vysu . . . i pasaceļtīs , i laikā dorbus sadareit , i vysa kū , kab napuorgulēt . paldis jam pasoka	▼ □
▶ 0:00 / 0:00 — 🔊 ⋮	<i>F124-1929-Cibla:</i> bet nikai muotis naatzyna ▶ muote kai nu ka soka vysur ka jau bārns tys na - nalaikā teik ▶ atdeve bārnu nomā tymā ▶ i tagad jei bārnu nomā i auga a kara laiks atguoja kara laikā redzi ▶ kurai sīvai sīvītei nabeja bārna ▶ jēme iz vysaidom iz šahtom iz vysurīni mobilizēja syutēja speciāli naprasēja ▶ mobilizēja syutēja a - ha tod nū - nūskrēja ▶	▲ □

Attēlā: meklētā vārda “laiks” rezultāts “MuLa” runas korpusa pilotversijā

Rezumējot

“MuLa” paplašinātās versijas jaunieguvumi:

- Vienotā ortogrāfijā veidoti teksti pēc 2007. gada (lielāka iespēja morfoloģiskajai marķēšanai nākotnē), kurus rada autori ar zināšanām un izglītību latgaliešu rakstu valodas jautājumos.
- Saņemta atgriezeniskā saite no lietotājiem un potenciālajiem lietotājiem.
- Video lekciju cikls, instrukcijas potenciālajiem lietotājiem – kā sasniedzamais rezultāts projekta laikā.
- Runas korpuss.

Paldies par
uzmanību!

