10.35539/LTNC.2022.0047.A.J.S.A.K.226.243

Mūsdienu latgaliešu valodas runas korpusa izveide mazāk lietoto valodu dokumentēšanas kontekstā

Angelika Juško-Štekele,
Antra Kļavinska

Raksts tapis Valsts pētījuma programmas "Humanitāro zinātņu digitālie resursi" projektā "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001).

Atslēgvārdi: korpuslingvistika, reprezentativitāte, korpusa dizains, metadati, transkripcija, konvencijas

Ievads

Mazāk lietoto valodu dokumentēšanai pēdējā laikā tiek veidoti dažādi rakstītu un mutvārdu tekstu korpusi, audio, videoieraksti, tiek veikti korpusā balstīti pētījumi. Atšķirībā no rakstu valodas, runātā valoda ir spontānāka un dabiskāka, turklāt nestandartizēti valodas varianti atklāj valodas daudzveidību un sniedz ieskatu pārmaiņās, kas notiek valodā laikā un telpā, kā arī saskarē ar citām valodām un valodas variantiem. Salīdzinot ar standartizētas rakstu valodas korpusu, reģionālā valodas varianta runas korpusa izveide ir daudz komplicētāka pat lielākajās pasaules valodās, piemēram, angļu valodā (Anderson et al. 2007; Anderwald, Wagner 2007), un vēl jo vairāk tādām valodām kā latviešu vai latgaliešu valoda.

Pēc 2011. gada tautas skaitīšanas datiem, latgaliešu valodu ikdienā lieto 164 510 Latvijas iedzīvotāju, kas ir apmēram 8 % no kopējā Latvijas iedzīvotāju skaita. Pēc UNESCO datiem latgaliešu valoda ar 150 00 uzrādītiem lietotājiem 2013. gadā atzīta par vienu no pasaules apdraudētajām valodām, kas pakļauta riskam (*Vulnerable*), jo visas paaudzes joprojām lieto mutvārdu formu, bet valodas ilgtspēja ir nopietni apdraudēta (Moseley 2020), tāpēc ka gados jaunāko valodas lietotāju skaits samazinās. Plašs pasākumu loks reģionālo vai mazāk lietoto valodu aizsargāšanai un lietošanas veicināšanai ietverts Eiropas Reģionālo vai minoritāšu valodu hartā (pieņemta 1992. gadā)[1]. Savukārt Eiropas Parlamenta 2013. gada 11. septembra rezolūcija par Eiropas apdraudētajām valodām un lingvistisko daudzveidību Eiropas Savienībā uzsver, ka efektīvus rīkus Eiropas apdraudēto valodu prasmes, izplatīšanas, mācīšanas un saglabāšanas veicināšanai var nodrošināt jaunās tehnoloģijas, multivides platformas, kas ietver atbalstu gan saturam, gan lietojumprogrammām[2].

Ievērojot ES direktīvas un ieteikumus reģionālo un riskam pakļauto valodu saglabāšanai, izpētei un attīstībai, kā arī Valsts valodas politikas nostādnes 2021.–2027. gadam attiecībā uz daudzveidīgu tekstu korpusu izstrādi, publiskošanu tīmeklī un pieejamību, Rēzeknes Tehnoloģiju akadēmijas pētnieku grupa valsts pētījumu

1 Eiropas Reģionālo vai minoritāšu valodu harta. Pieejams: https://m.likumi.lv/doc.php?id=209969 [skatīts 25.10.2021.].

2 Eiropas Parlamenta 2013. gada 11. septembra rezolūcija par Eiropas apdraudētajām valodām un lingvistisko daudzveidību Eiropas Savienībā (2016/C 093/07) Eiropas Savienības Oficiālais Vēstnesis. 2016. gada 9. marts, Nr. 59., 52.–58.lpp. Pieejams: https://eur-lex.europa.eu/legal-content/LV/TXT/PDF/?uri=OJ:C:2016:093:FULL&from=ES [skatīts 25.10.2021.].

programmas projektā "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001)" 2020. gadā uzsākusi darbu pie mūsdienu latgaliešu valodas runas korpusa (MuLaR) izveides, un tas paredzēts latgaliešu valodas dokumentēšanai, apguvei, studiju un pētniecības vajadzībām.

Dialektu un izlokšņu daudzveidības dēļ runas korpusa izveide ir laikietilpīgs process, kas ietver daudz manuāla darba, lai transkribētu un apstrādātu datus, pirms tie iegūst korpusa formu (Vuković 2021). Neraugoties uz to, saskatāmas daudzas apdraudēto, mazāk lietoto valodu dokumentēšanas priekšrocības, veidojot runas korpusus: pieejamība un ilgtspējība; liels oriģinālu, digitāli arhivētu, transkribētu, anotētu un kataloģizētu datu apjoms; daudzfunkcionalitāte, aptverot potenciāli dažādas tēmas, tekstu žanrus; kontekstualizācija, kas iespējama, izmantojot metadatus; iespēja atspoguļot dažādos laika posmos dokumentētus datus u. tml. (Rießler, Wilbur 2017: 60).

Raksta mērķis ir, lietojot zinātniskās literatūras referatīvās analīzes un salīdzinošo metodiku, identificēt un analizēt problēmjautājumus, kas nozīmīgi MuLaR izveides procesā, proti:

1) kā izveidot reprezentatīvu latgaliešu valodas runas korpusu, ņemot vērā latgalisko izlokšņu daudzveidību;
2) kādu metodoloģiju izmantot dabiskas, spontānas valodas dokumentēšanai (jaunu datu vākšana, pastāvošo ierakstu izmantošana);
3) kā veikt runas ierakstu transkribēšanu, lai pēc iespējas precīzāk atklātu mutvārdu teksta iezīmes;
4) kā izveidot atvērtā piekļuvē pieejamu, ērti izmantojamu runas korpusa platformu.

Eiropas reģionālo valodu un dialektu runas korpusi

Pēc EP oficiālās informācijas, ES ir vairāk nekā 60 vietējo reģionālo un mazākumtautību valodu, kurās runā apmēram 40 miljoni cilvēku (Pasikowska-Schnass 2016: 1). Eiropas vienotajā valodas resursu un tehnoloģiju infrastruktūrā CLARIN pieejami 129 runas korpusi, no tiem 118 veidoti ar paralēlu audioierakstu un tā transkripcijas funkciju, 11 – tikai ar transkribētiem tekstiem[3]. CLARIN platformā atrodam, piemēram, katalāņu valodas fonoprosodisku runas korpusu (PhonCAT), kas satur 8719 minūšu spontānas runas ierakstus no 234 Barselonā dzīvojošiem katalāņu valodā runājošiem vīriešiem vecumā no 5 līdz 45 gadiem. Igauņu dialektu korpuss satur fonētiski transkribētus dialogus 1 284 000 vārdu apjomā, kas

3 CLARIN Spoken Corpora. Introduction (2021). Available: https://www.clarin.eu/resource-families/spoken-corpora [accessed 14.11.2021.].

ierakstīti ilgākā laika periodā, sākot no 1938. gada līdz 20. gadsimta 70. gadiem. Pie masīvākajiem dialektu korpusiem jānosauc, piemēram, Freiburgas angļu valodas dialektu korpuss (FRED), kas ietver Britu salu iedzīvotāju ķeltisko dialektu morfosintaktiskās variācijas, ieraksti veikti no 2000. līdz 2005. gadam. FRED korpuss ietver 2,5 miljonus vārdlietojumu, 300 stundu ierakstus no 372 intervējamām personām 163 dažādās apdzīvotās vietās, kas administratīvi atbilst 9 dialektu apgabaliem (Hernández 2006).

Veidojot MuLaR, tiek ņemti vērā labās prakses piemēri citu mazāk lietoto valodu un dialektu runas korpusu izstrādē. Viens no tādiem labi strukturētiem un ērti lietojamiem resursiem ir Spišas dialekta korpuss (*Korpus Spiski*) – mutvārdu tekstu kolekcija, kas dokumentē Polijas Spišas reģiona iedzīvotāju runu. Korpuss ir paredzēts ikvienam reģionālās valodas un kultūras entuziastam, lietotājiem ir brīva piekļuve teksta transkripcijām un audio ierakstiem. Teksti ir pierakstīti ortogrāfiskajā transkripcijā, izmantojot poļu alfabētu, tāpēc tos var lasīt un meklēt arī lietotāji, kas nepārzina fonētisko transkripciju. Meklētājprogramma nodrošina dažādas meklēšanas kategorijas, piemēram, lemmas, dažādas gramatiskās formas, atbilstošus audioieraksta segmentus un metadatus. Leksika, kas raksturīga šim reģionam, ir skaidrota vārdnīcā. Teksti ir morfoloģiski marķēti, tas zinātniekiem sniedz iespējas veikt pētījumus par gramatisko sistēmu, vārddarināšanas procesiem, sintaksi un leksiku. Augstas kvalitātes audioieraksti ir piemēroti fonētikas pētījumiem, savukārt metadati (par runātāju vecumu, dzimumu, dzīvesvietu, izglītību, tautību) sniedz iespējas sociolingvistikas pētījumiem (Grochola-Szczepanek et al. 2019).

Korpusa mērķa un tehnoloģisko risinājumu ziņā MuLaR ir saistoša arī čehu valodas dialektu korpusa DIALEKT pieredze. DIALEKT veidotāji to piesaka kā viegli izmantojamu korpusu, kas piemērots ne tikai dialektoloģijas pētniekiem, bet arī valodniecības robežzinātņu speciālistiem, skolotājiem un citiem interesentiem. DIALEKT iekļauti gan vēsturiski dati no 20. gadsimta 50.–80. gadu ekspedīcijām, kas jau iestrādāti valodniecības atlantos u. c. zinātniskos izdevumos, gan jaunāki ieraksti, kas neparedz striktu apjoma ierobežojumu, bet, līdzīgi kā MuLaR iecerē, ir orientēti uz regulāru datu papildināšanu, nepretendējot uz stingru datu kopu balansu. Tomēr DIALEKT saglabā dialektu apgabalu principu, kas redzams korpusa metadatos. Ierakstu transkribēšanai korpusā tiek izmantots komplekss ortogrāfiski dialektoloģiskais princips ar dialektoloģiskā principa pārsvaru, kas pamatots ar tradicionālām dialektoloģiskā pieraksta konvencijām Čehijā (Komrsková et al. 2017).

Iestrādes

Lai nodrošinātu mūsdienīgus digitālos resursus un rīkus latgaliešu valodas dokumentēšanai, pētniecībai un apguvei, tiek veidots Mūsdienu latgaliešu valodas tekstu korpuss (MuLa). Patlaban pieejamajā versijā noteiktās proporcijās ievietoti latgaliešu rakstu valodā publicēti (1988–2012) teksti ar metadatiem, apjoms – viens miljons vārdlietojumu. Projektā

"Humanitāro zinātņu digitālie resursi: integrācija un attīstība" korpuss tiek pilnveidots un papildināts ar jaunākiem tekstiem, kā arī tiek veidots jauns mūsdienu latgaliešu valodas runas korpuss (MuLaR)[4].

Ar vienošanās līgumu MuLaR tiek integrēti projektā "Triangulation Approach for Modelling Convergence with a High Zoom-In Factor" (TriMCo[5]) un projektā "Diachronic Typology of Differential Argument Marking"[6] transkribētie latgaliešu runas teksti. TriMCo projekta empīriskās bāzes vajadzībām, izmantojot ELAN programmatūru, tika veidots multilingvāls Dialektu korpuss ar krievu, baltkrievu, lietuviešu un latgaliešu tekstiem, kas projektu ierobežotā laika un cilvēkresursu trūkuma dēļ nevainagojās ar lietotājiem pieejamu latgaliešu runas datni.

MuLaR tiek izmantotas arī Rēzeknes Tehnoloģijas akadēmijas zinātnisko grantu konkursā finansētā projekta "Austrumu diasporas latgalisko izlokšņu ierakstu datubāzes izveide" iestrādes (projekta partneris – Krasnojarskas latviešu biedrība "Dzintars"). Projektā tika izveidots publiski pieejams interneta resurss (https://reiti.rta.lv/), kur ievietoti lauka pētījumos Sibīrijā (galvenokārt Krasnojarskas un Tomskas apgabalā) no latgalisko izlokšņu runātājiem iegūtie videoieraksti, kā arī transkribēti audioierakstu teksti. Projekta tīmekļvietnē ir īstenotas meklēšanas iespējas ar atslēgvārdiem, pieejami metadati (respondenta dzīvesvieta, vecums, dzimums, izglītība), kā arī karte, kurā apzīmētas intervēšanas vietas. Ne visi lauka pētījumos iegūtie ieraksti projekta laikā tika transkribēti un ievietoti datubāzē, šis darbs tiek turpināts MuLaR izstrādes procesā.

Reprezentatīva latgaliešu runas korpusa problemātika

Reprezentativitāte ir viens no plašāk un daudzveidīgāk skatītajiem korpuslingvistikas jautājumiem, kas visbiežāk korelē ar korpusa līdzsvarotības un datu ieguves jautājumiem. Berlīnes Universitātes lingvistikas profesors Anatols Stefanovičs (*Anatol Stefanowitsch*) valodas korpusa reprezentativitāti saista ar tā līdzsvarotību un blakus korpusa autentiskumam un apjomam atzīst to par vienu no būtiskākajām valodas korpusa pazīmēm (Stefanowitsch 2020: 22–23), kas ir saistīta ar korpusa mērķi un plānoto lietojumu. MuLaR saskaņā ar projekta sākotnējo ieceri ir paredzēts latgaliešu valodas dokumentēšanas, pētniecības, studiju un apguves vajadzībām. Šim nolūkam MuLaR reprezentativitātes kritēriju lokā tiek vērtēti: aptverošs

4 Skat šajā izdevumā Sanitas Martenas, Annas Briškas un Nikoles Nauas rakstu "Latgaliešu valodas korpuss citu Eiropas mazāk lietoto valodu kontekstā: korpusa raksturojums, lietojums un potenciālā iespējošana".

5 Prof. Dr. Björn Wiemer, German Research Foundation, Johannes-Gutenberg-University Mainz, projekta numurs: WI 1286/16-1.

6 Dr. Ilja A. Seržant, Incoming Fellowship Programme, Grant Agreement Number: 291784, Zukunftskolleg, University of Konstanz.

mūsdienu latgaliešu valodas lietojuma areāls, autentiski teksti, kas fiksēti dialoga, poliloga vai stāstījuma situācijā un papildināti ar pētniecībai saistošiem valodas lietotāju sociodemogrāfiskiem metadatiem.

Lai nodrošinātu reprezentatīvu mūsdienu latgaliešu runātās valodas aptveri, ir būtiski fiksēt autentisku valodas lietojumu visā latgaliešu valodas lietošanas areālā. Tipoloģiski iespējams izdalīt vismaz četrus šādus areālus: Latgales kopiena (Ltg), pārējo Latvijas novadu (pārnovadu) kopiena (Lv), Austrumu diaspora Krievijā, Baltkrievijā u. c. (Ad) un Rietumu diaspora ASV, Kanādā, Zviedrijā, Austrālijā, Vācijā, Polijā, Lietuvā u. c. (Rd). Pēc 2011. gada tautas skaitīšanas datiem, ikdienā latgaliešu valodu lieto 35,5 % no Latgales iedzīvotājiem (Ltg), 9,2 % no iedzīvotājiem Rīgā un Pierīgā (LvRP), 4,8 % iedzīvotāju Vidzemē (LvV), 4,3 % Zemgalē (LvZ) un 1,5 % Kurzemē (LvK). Latgaliešu valodas lietošanas aptvere Austrumu diasporā un Rietumu diasporā bez papildu statistikas datiem nav precīzi nosakāma, tāpēc MuLaR nevar pretendēt uz pilnīgi precīzu datu līdzsvaru. Atsaucoties uz Ziemeļarizonas Universitātes profesora Daglasa Baibera (*Douglas Biber*) viedokli, MuLaR pētnieku grupa pieņem, ka korpusu reprezentativitāte ir ne tik daudz atkarīga no filigrānām skaitliskajām proporcijām, cik no tā, vai korpusā iekļautie teksti atklāj iespējami pilnīgu ainu par valodas lietojumu (Biber 1993).

MuLaR izstrādes pirmajā posmā iekļaušanai korpusā tiek apstrādāti 20 stundu audioieraksti, no kuriem 4 stundas ir Austrumu diasporas (Sibīrijas) audioieraksti, 16 stundas – ieraksti, kas veikti Latgales teritorijā. Balstoties uz aprēķina, ka dominējošais latgaliešu valodas lietotāju skaits ir Latgales iedzīvotāji, MuLaR Latgales ierakstu apjomam ilgtermiņā jānodrošina iespējami adekvāts balanss ar latgaliešu valodas ierakstiem pārējos Latvijas novados, kas pēc tautskaites datiem ir nosacīti nedaudz vairāk nekā puse (55,7 %) no Latgales kopienas. Austrumu un Rietumu diasporas ieraksti attiecīgi veido atlikušo nosacīto pusi. Tādējādi MuLaR valodas teritoriālās aptveres balansam iespējams izmantot formulu: MuLaR*h*100% = 50%*h*Ltg + 26%*h*Lv + 12%*h*Ad + 12%*h*Rd, kur Ltg – Latgales kopienas audioieraksti, Lv – pārnovadu kopienas audioieraksti, Ad – Austrumu diasporas audioieraksti, Rd – rietumu kopienas audioieraksti, *h* – mainīgā audioieraksta stundu daļa. Turpmākajā korpusa izstrādes gaitā balanss var mainīties, jo sevišķi Austrumu un Rietumu diasporas sadaļās, kur apjoma palielināšanās varētu nebūt būtiska, saglabājot reprezentativitātes principu, cik tas iespējams. Iespējami pilnīgam balansam Latgales kopienas un pārnovadu kopienas sadaļās tiek veidota sava iekšējā struktūra, kas, izmantojot tautskaites datus, sīkāk sadalās pēc latgaliešu valodas lietotāju administratīvi teritoriālā pārklājuma.

Latgaliešu valodas lietotāju administratīvi teritoriālais sadalījums savā ziņā korelē ar korpusa izstrādē aktuālu jautājumu par latgaliešu valodas izlokšņu skaitu, kas Latgalē ir apmēram 70. Atsaucoties uz valodnieces Annas Stafeckas pētījumiem, MuLaR pētnieku grupa pieturas pie atzinuma, ka Latgalē izlokšņu robežas nosacīti atbilst viena pagasta teritorijai (pēc 1939. g. administratīvi teritoriālā iedalījuma) (Stafecka 2017: 58). Līdz ar to MuLaR reprezentativitātes nodrošināšanai nav plānots piemērot izlokšņu kritēriju, atstājot iespēju pētniekiem pētīt izlokšņu iezīmes pēc administratīvi teritoriālā dalījuma un pētniecībai pieejamiem teicēju metadatiem.

Būtisks jautājums MuLaR izstrādes procesā ir runātāju metadati: dzimums, vecums, dzīvesvieta, kas tiek dokumentēti ar mērķi padarīt korpusu saistošu turpmākiem sociolingvistikas pētījumiem, jo sevišķi, analizējot latgaliešu valodas lietotāju skaitu un pašas valodas īpatnības dažādās vecuma grupās un to ietekmi uz iespējamo latgaliešu valodas lietojumu nākotnē. Jau 2011. gada tautskaites dati rāda, ka vairumā Latgales novadu latgaliešu valodā runājošo skaits samazinās vecumā līdz 19 gadiem un palielinās vecumā virs 70 gadiem, atsevišķos novados (Baltinavas, Rugāju) vecumā virs 90 gadiem pat sasniedzot 100 %. Balstoties uz tautskaites datiem par Latgales iedzīvotāju – ikdienas latgaliešu valodas lietotāju – sadalījumu pa vecuma grupām, Latgales kopienas interviju ierakstus, ievērojot iekšējo teritoriālo proporciju, plānots strukturēt šādi: 15 % ieraksti no iedzīvotājiem vecumā līdz 19 gadiem, 45 % – no 20 līdz 65 gadiem, 35 % – vecumā virs 65 gadiem. Pārējos Latvijas novados piemērojamā statistika, kas veidota, pamatojoties uz tautskaites datiem, ir šāda: 10 % – vecumā virs 19 gadiem, 50 % – vecumā no 20 līdz 65 gadiem, 40 % – vecumā virs 65 gadiem. Austrumu un Rietumu diasporas kopienās informantu vecuma līdzsvaru nav plānots īstenot, jo nav drošu datu par populācijas apjomu un vecuma struktūru.

Teritoriālā kopiena	Ltg	Lv	Ad	Rd	
Ierakstu apjoma proporcija	50 %	26 %	12 %	12 %	
Informantu dzimumu struktūra	S:V	50:50	50:50	50:50	50:50
Informantu vecuma struktūra	0–19	15	10		
	20–64	45	50		
	65+	50	40		
Teksta žanrs	Intervijas, dialogi, polilogi, stāstījumi				

1.tabula. MuLaR līdzsvarotības koncepcija.

MuLaR izstrādes posmi

Teorētiskajā literatūrā tiek definēti runas korpusa izveides posmi. Piemēram, Pols Tompsons (*Paul Thompson*) piedāvā četrus posmus: 1) datu vākšanas posms, kurā nepieciešams apspriest gan audio/video ierakstīšanas tehniskos aspektus, gan kontekstuālās informācijas vākšanu un dalībnieku piekrišanas ieguvi; 2) transkribēšanas posms; 3) marķēšanas un anotēšanas posms, kurā transkripcijas tiek pārveidotas mašīnlasāmā formātā, kā arī tiek veikta marķēšana; 4) piekļuves posms, kurā uzsvars vispirmām kārtām likts uz iespējamo korpusa lietojumu, proti, vai tas būs pieejams citiem pētniekiem un kādā formā to var darīt pieejamu, ja to vēlas (Thompson 2005). Dažkārt runas korpusa izstrāde tiek reducēta trīs

pamatposmos: 1) runas ierakstu veikšana, 2) transkribēšana un kodēšana, 3) pārvaldība un analīze (Adolphs, Knoght 2010: 40).

MuLaR ir pirmais latgaliešu valodas runas korpuss Latvijā, RTA pētniekiem līdz šim nav bijis pieredzes runas korpusu izveidē, tāpēc šajā gadījumā kā pirmais būtu izceļams plānošanas posms, kas Tompsona sniegtajā pirmā posmā aprakstā attiecināts galvenokārt uz tehniskajiem aspektiem un metadatiem. MuLaR korpusa plānošanas posms sākās ar citu runas korpusu izpēti, korpusa dizaina izvēli, konsultēšanos ar pieredzējušiem ārzemju kolēģiem, piemērotas transkribēšanas datorprogrammas izvēli un apguvi, iekļaujot tajā arī diskusijas par konvencijām. MuLaR korpusa izveide patlaban tiek turpināta šādos posmos: datu ieguve (interviju ierakstīšana un arhivēšana); datu apstrāde (transkribēšana); korpusa platformas izveide (piekļuves posms).

Datu ieguve

MuLaR ir iecerēts kā multimodāls spontānas runas datu kopums. Primārais uzdevums ir transkribēt jau iepriekš veiktos audioierakstus, kā arī organizēt jaunu datu ieguvi. Korpusa izveides sākumposmā tiek izmantoti Rēzeknes Tehnoloģiju akadēmijā organizētu lauka pētījumu materiāli un plašsaziņas līdzekļu ieraksti. Korpusa metadatos patlaban definēti trīs tekstu pamatavoti:

1) dažādos Latgales novados 2009.–2021. gadā ierakstītas intervijas, tostarp "TriMCo" un "Diachronic Typology of Differential Argument Marking" projektu dati (jau iepriekš transkribētās intervijas tiek pārbaudītas un rediģētas saskaņā ar MuLaR konvencijām);

2) Sibīrijā dzīvojošo latgaliešu runas ieraksti (lauka pētījumi 2017.–2018. gadā, kas līdz šim daļēji bija ievietoti austrumu diasporas latgalisko izlokšņu datubāzē (https://reiti.rta.lv/);

3) TV ieraksti (2018.–2020. gadā Latgales Reģionālajā televīzijā veiktas intervijas).

2021. gadā korpusa datu ieguves nolūkā tika organizēts lauka pētījums Aglonā. Tā laikā aptaujāti 22 informanti (3 – vecumā līdz 19 gadiem, 4 – vecumā virs 65 gadiem, 15 – vecumā no 20 līdz 65 gadiem), iegūti ieraksti 15 stundu garumā. MuLaR izveides turpmākajā gaitā plānots veikt papildu datu ieguvi pārējos Latvijas novados un Rietumu diasporā.

Lai nodrošinātu atbilstību Akadēmiskā godīguma vispārējām vadlīnijām un Eiropas Parlamenta un ES Padomes regulas "Par fizisku personu aizsardzību" prasībām, pirms intervijas veikšanas tiek iegūta mutiska (ierakstot audioierakstā) vai rakstiska informēta piekrišana (intervējot bērnus – vecāku vai aizbildņu piekrišana) datu sniegšanai un audioieraksta veikšanai. Rakstisku informēto piekrišanu paraksta ne tikai datu subjekts, bet arī pētnieks, kas piedalās intervijā (Tauginienė et al. 2019).

Konfidencialitātes un personas datu aizsardzības nolūkā intervējamo personu un intervētāju dati tiek šifrēti, ievērojot ES Vispārīgajā datu aizsardzības regulā definēto datu pseidonimizāciju, lai personas datus nebūtu iespējams saistīt ar konkrētu datu subjektu bez papildu informācijas izmantošanas. MuLaR datu uzglabāšanas politika paredz, ka šāda papildu informācijai tiks uzglabāta atsevišķi, nodrošinot tādu tehniskas un organizatoriskas drošības pakāpi, lai personas dati netiktu saistīti ar konkrētu datu subjektu[7].

Ja personas runā lietoti uz viņu pašu vai kādu citu personu vērsti jutīgi dati, kas dod iespēju identificēt minēto personu, šādi teksta fragmenti korpusā netiek iekļauti. Katram audioierakstam un transkripcijas failam ir pievienoti metadati: ziņas par ieraksta vietu un laiku, ieraksta ilgumu, informanta dzimumu un vecumu. Korpusa izstrādē stingri tiek ievērots datu minimizēšanas princips, kas paredz, ka personas dati tiek apstrādāti tikai tam paredzētajam mērķim un tam nepieciešamajā apjomā.

Papildus runas ierakstīšanas procesam ir svarīgi dokumentēt informāciju par procesa dalībniekiem, runas situāciju, proti, metadatus jeb "datus par datiem". Kā atzīst digitālo humanitāro zinātņu eksperts Lū Burnards (*Lou Burnard*), metadati precizē kontekstu, tādējādi ļaujot sasaistīt doto teksta paraugu ar tā izcelsmi. Lietojot valodas korpusu, neizbēgami rodas jautājumi par datu precizitāti un autentiskumu, un bez metadatiem pētnieks nevar atbildēt uz šiem jautājumiem (Burnard 2005: 31). Metadati ir ļoti svarīgi korpusam, lai varētu sasniegt reprezentativitātes, līdzsvara un viendabīguma standartus (Adolphs, Knoght 2010: 42).

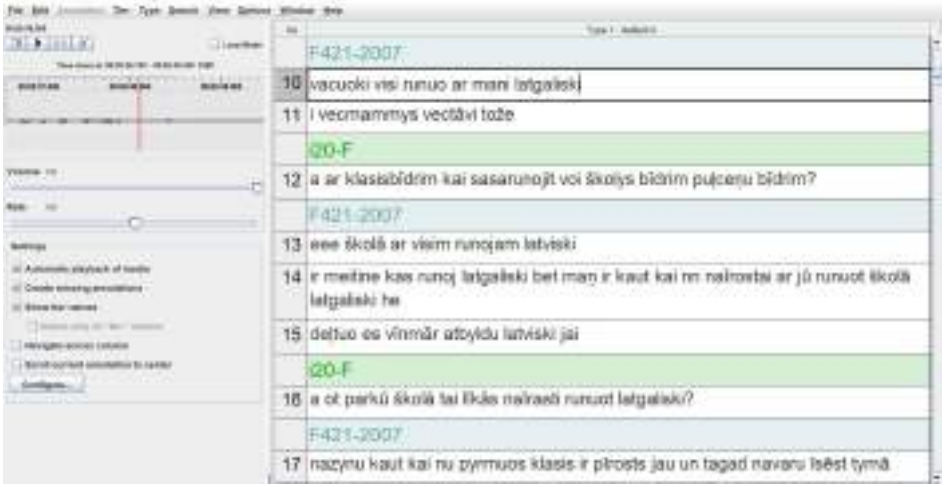
MuLaR plānošanas un izveides sākumposmā darba grupa ir vienojusies par metadatu kategorijām, piemēram, par to, kā tiks ģenerēti failu nosaukumi, kā tiks kodēti runātāju un intervētāju nosaukumi, kuri metadati būs izmantojami tikai korpusa administrēšanā, kuri būs publiski pieejami korpusā u. tml.

Datu apstrāde

Uzsākot plānot transkribēšanas procesu, vispirmām kārtām jāizlemj, kāds transkripcijas veids tiks izmantots: fonētiskā, prosodiskā, ortogrāfiskā transkripcija vai to kombinācija. Tas lielā mērā atkarīgs no runas korpusa mērķa (Thompson 2005).

Freiburgas angļu valodas dialektu korpusa (*FRED*) veidotāji, analizējot dialektu mutvārdu tekstu transkribēšanas problēmas runas korpusa vajadzībām, pamato ortogrāfiskās transkripcijas priekšrocības, salīdzinot ar fonētisko transkripciju. Ja konkrētā reģionā pastāv vairāki dialekti un izloksnes, būtu sarežģīti atainot visas fonētiskās iezīmes vienkopus, turklāt tas ir

7 *Eiropas Parlamenta un Padomes Regula (ES) 2016/679 (2016) par fizisku personu aizsardzību attiecībā uz personas datu apstrādi un šādu datu brīvu apriti un ar ko atceļ Direktīvu 95/46/EK.* Pieejams: https://eur-lex.europa.eu/legal-content/LV/TXT/HTML/?uri=CELEX:32016R0679&from=LV [skatīts 14.11.2021.].



1. attēls. Transkribēšana ELAN programmā.

arī ļoti laikietilpīgs process. Specifisku fonētiskās transkripcijas zīmju lietojums rada arī acīmredzamas tehniskas problēmas (fonētiskās transkripcijas lietojums datorrakstā, neiespējama visa veida marķieru meklēšana, jo pētniekam nebūtu ne jausmas, kādas formas meklēt). Būtu jāizmanto tāda transkripcija, kas ir dabisks kompromiss starp pieraksta detalizāciju (dziļumu) un pārklājumu (plašumu). Korpusa prasībām atbilstošāka ir ortogrāfiskā transkripcija, jo tam jābūt mašīnlasāmam, nodrošinot vieglu piekļuvi, dažādas meklēšanas iespējas un salīdzināmību ar citiem korpusiem (Anderwald, Wagner 2007).

Ņemot vērā citu runas korpusu izveides pieredzi, MuLaR korpusa veidotāji ir vienojušies mutvārdu tekstus atveidot latgaliešu valodas ortogrāfiskajā transkripcijā. Tā kā korpusa platformā plānots sasaistīt transkribēto tekstu ar audiofailu, ortogrāfiskā transkripcija atvieglo meklēšanas iespējas, tajā pašā laikā neizslēdz iespēju veikt pētījumus fonētikā, fonoloģijā, kā arī dialektoloģijā.

Transkribēšanai izmantota ELAN programmatūra, kas izstrādāta Maksa Planka Psiholingvistikas institūtā Neimegenā (Nīderlandē). ELAN ir rīks, kas paredzēts audio un video ierakstu transkribēšanai, nodrošina audioierakstu segmentēšanu, vairākus veidus, kā skatīt anotācijas, katrs skats ir savienots un sinhronizēts ar multivides laika skalu[8].

8 ELAN (Version 6.2) [Computer software]. (2021). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Rindas	• Katram teicējam un katram intervētājam ir sava rinda, rindas nosaukums ir teicēja vai intervētāja kods. Pirmās pakāpes rindās tiek rakstīts tikai runas atšifrējums. Komentāri, skaidrojumi iespējami asociētajās rindās, piemēram, norāde uz vārdu vai izteikumu citā valodā.
Ortogrāfija	• Tiek lietota mūsdienu latgaliešu valodas ortogrāfija, bet runas pieraksts netiek koriģēts atbilstoši latgaliešu rakstības noteikumiem[9], piemēram, vārds 'vējš' var tikt pierakstīts kā *viejs, viejš, vējs* u. tml. (atbilstoši izrunai). • Lielie sākumburti netiek lietoti segmenta sākumā, bet tiek lietoti īpašvārdos, piemēram, *Leldīne, Zīmyssvātki, Dīvs.* • Skaitļi rakstāmi ar vārdiem. • Svarīgākie pieraksta nosacījumi (t. sk. problēmgadījumi). • Divskanis [uo] tiek pierakstīts ar diviem burtiem: *uo,* piemēram, *muosa.* Garais patskanis [ō] tiek atveidots ar diviem burtiem *oo,* burts *ō* netiek lietots. • Līdzskaņu burtus nemīkstina, kad aiz tiem seko patskaņi *e*, *ē*, *i*, *ī* vai divskaņi *ei*, *ie*, *iu*, arī aizguvumos no citām valodām, piemēram, no krievu valodas: *posle* (nevis *posļe*). Ja skaidri saprotams, ka runātājs izrunā kādu vārdu vai frāzi latviešu literārajā valodā, tad līdzskani šajā pozīcijā mīkstina, piemēram, *atsevišķi gadījumi* (latgaliski būtu *atseviški gadejumi*). • Latgaliešu valodas sistēma lietojama arī īpašvārdos: *Andeni, Antonina* (nevis *Andeņi, Antoņina*). • Mīkstu līdzskaņu burtus *ķ, ģ, ņ, ļ* raksta pirms patskaņiem *a, ā, o, u, ū* un divskaņiem *au, uo*, citu patskaņu mīkstinājumu apzīmē ar *j*, piemēram, *cjoce, acjom.* • Plato patskani apzīmē ar *e, ē* (attiecīgi arī divskani *ei*), pat ja tas ir ļoti plats un tā izruna līdzinās [*a*], piemēram, *zeme* (nevis *zjamja*), *meita* (nevis *mjaita*). • Pozicionālās skaņu pārmaiņas pierakstā netiek atspoguļotas, piemēram, *atsasoka* (nevis *acasoka*). • Ja informanta runā lietotas saīsinātās formas, tās tā arī tiek pierakstītas nepārveidojot, piemēram, *guo, guoa, vāg, tān, pasceja.*
Interpunkcija	• Pieturzīmes netiek lietotas, izņemot jautājuma zīmes, kas tiek lietotas īstā jautājuma teikuma beigās, piemēram, *a sātys veinu taisiejot?* • Pēdiņas tiek lietotas nosaukumos ar īpašvārdisku nozīmi, piemēram, *gruomota "Syltuo mola".*
Prosodija	• Pauzes netiek norādītas (ja pauze ir gara, tā veido atsevišķu segmentu, ELAN programma automātiski mēra segmenta garumu). • Uzsvara vieta netiek fiksēta. Ja runā uzsvara vieta atšķiras no normas, to atzīmē asociētajā rindā.
Vārdu fragmenti	• Tiek rakstīta vārda daļa, izlaistā vārda fragmenta vietā tiek lietota defise. Piemēram, *runoj latv- latgališu volūdā.* • Aiz aprautās daļas ir atstarpe, piemēram, *s- s- smējās.*
Abreviatūras	• Ja tiek izrunāta vispārzināma abreviatūra (gan burtu, gan zilbju), tā tiek pierakstīta ar burtiem: *PVN* (izrunā *pē vē en*), arī tad, ja izrunā abreviatūru citā valodā, piemēram, *PR* [izrunā *pī ār*]. • Ja tiek lietots nezināms saīsinājums, tiek pierakstīts pēc izrunas, piemēram, *pē pē kā*.
Neskaidrs teksts	• Tiek lietots apzīmējums [*unint*.]
Neverbālie elementi	• Apstiprinājuma signāls: *mhm*. • Pildītas pauzes, vilcināšanās, piemēram: *eee*, *aaa.* • Smiekli starp vārdiem: *haha, ha ha ha, he he, hihi* (atbilstoši skanējumam). • Smiekli vienlaicīgi ar vārdiem netiek atzīmēti.

2. tabula. MuLaR korpusa tekstu transkribēšanas pamatprincipi[10].

9 Valsts valodas centra Latviešu valodas ekspertu komisijas Latgaliešu ortogrāfijas apakškomisijas lēmums Nr.1. 2007. gada 28. septembrī. "Par Latgaliešu rakstības noteikumiem". Pieejams: https://likumi.lv/ta/id/164904-par-latgaliesu-rakstibas-noteikumiem [skatīts 12.11.2021.].

10 Pateicamies Adama Mickeviča universitātes Poznaņā (Polijā) profesorei Nikolei Nauai (*Nicole Nau*) par konsultācijām un ieteikumiem transkribēšanas konvenciju izstrādē.

Oriģinālie audiofaili pirms transkribēšanas tiek apstrādāti: 1) tiek izgriezti fragmenti, kas satur nevajadzīgas detaļas, piemēram, personas datus, mūziku, runu citā valodā (izņemot atsevišķus vārdus vai teikumus) u. tml., 2) liela apjoma faili tiek sadalīti mazākos (10–30 minūtes); 3) lai atšifrēšana būtu vieglāka, faili tiek konvertēti WAV formātā.

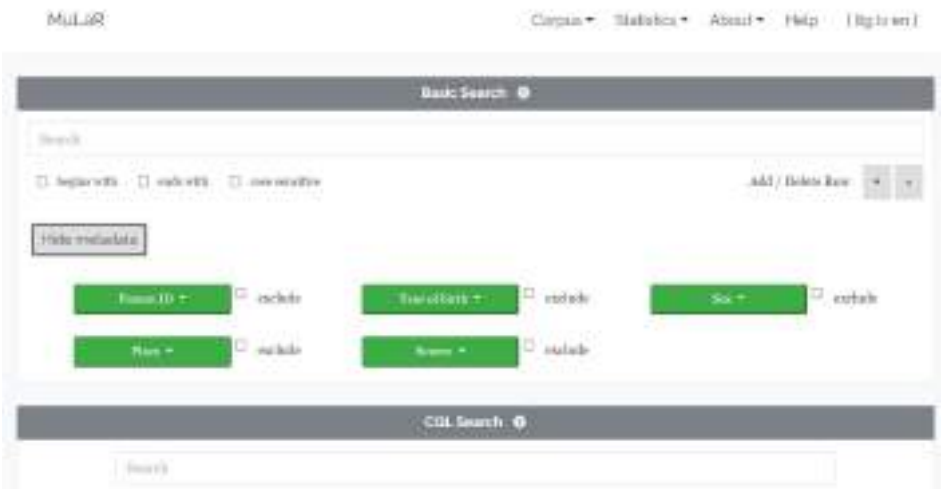
Strādājot ELAN programmā, katram runātājam tiek piešķirta sava rinda, kurā ieraksta runas atšifrējumu. Rindas tiek segmentētas, ņemot vērā loģiskās izteikuma robežas, intonatīvās vienības, runas pauzes. Līdz ar to segmentu garums variējas. Katram runātajam (gan intervētājiem, gan informantiem) tiek piešķirts kods, lai atvērtajā piekļuvē pieejamais teksts būtu pseidonimizēts, bet projekta dalībnieki varētu viņus identificēt un saistīt ar metadatiem. Piemēram, runātāju kodi *M400-1971* vai *F406-1960*, kur redzama norāde uz dzimumu, kārtas numuru un dzimšanas gadu. Intervētāja kodā ir redzams kārtas numurs un norāde uz dzimumu, piemēram, *i01-F* vai *i04-M* (sk. 1. attēlu).

Sekojot citu ortogrāfiski transkribētu runas korpusu piemēriem (Anderwald, Wagner 2007; Grochola-Szczepanek 2019; Pinnis et al. 2014 u. c.), latgaliešu runas korpusa veidotāji ir izstrādājuši savas tekstu transkribēšanas konvencijas (sk. 2. tabulu). Darba grupā daudz diskutēts, cenšoties veidot tādu pierakstu, kas būtu saprotams plašam interesentu lokam, un tajā pašā laikā tiecoties pēc iespējas precīzāk atspoguļot mutvārdu teksta (tostarp individuālās runātāja) iezīmes. Diskusiju rezultātā nonākts pie lēmuma mutvārdu tekstus nepārveidot standartizētā latgaliešu literārajā rakstu valodā. Problēmas sagādā prosodijas atveidojuma detalizācijas pakāpes noteikšana transkribēšanas procesā, fonētisko un morfoloģisko variantu daudzveidības (dažādu latgalisko izlokšņu vai individuālo runas īpatnību) atspoguļojums. Tas, iespējams, var apgrūtināt meklēšanu korpusā, tāpēc plānots veidot tādu korpusa platformu, kas piedāvātu pēc iespējas plašākas iespējas meklēt atsevišķas vārda daļas. Ja, veicot transkribēšanu, tiek fiksēts kāds problēmgadījums, tas tiek apspriests darba grupā, tādējādi konvenciju saraksts pakāpeniski tiek papildināts vai precizēts. Lai pēc iespējas precīzāk sekotu konvencijām un novērstu atkāpes no tām, pirms ievietošanas korpusa platformā katru transkribējumu vēlreiz pārbauda un, ja nepieciešams, rediģē cits darba grupas dalībnieks.

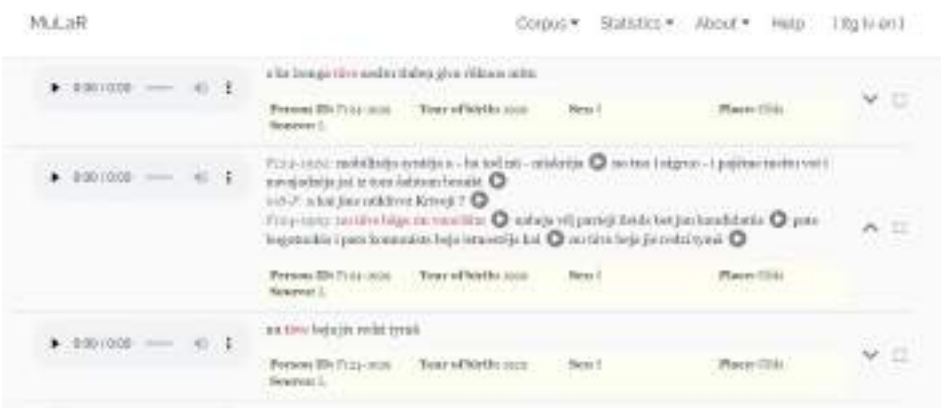
Korpusa platformas izveide

Kā minēts iepriekš, viens no runas korpusiem, ko MuLaR korpusa veidotāji ir atzinuši par vērā ņemamu paraugu, ir Polijas Spišas dialekta korpuss. Šī korpusa platformas veidotāji ir izmantojuši tīmekļa saskarni SpoCo.

SpoCo veidotāji to raksturo kā viegli uzturamu, efektīvu sistēmu valodas datu meklēšanai tīmeklī, kas piemērota dialektu runas korpusiem un salāgojama ar ELAN transkribēšanas programmu. Platformas lietojuma vienkāršība ir svarīga, lai atvieglotu pētniecību un pieejamību plašam lietotāju lokam, tostarp tiem, kam ir ierobežoti tehniskie un finanšu resursi. Sākotnēji SpoCo tika izstrādāta konkrēta krievu valodas dialekta runas korpusa mērķim, taču



2. attēls. MuLaR pilotkorpusa saskarne.



3. attēls. Meklēšanā izgūstamo datu paraugs MuLaR pilotkorpusā.

tā ir viegli pielāgojama un izmantojama citām valodām un dialektiem. Patlaban šai saskarnei ir izveidotas divas versijas, kuru galvenā atšķirība ir vērojama to uzbūves tehnoloģijā, taču abām versijām ir kopīga lielākā daļa funkciju: lietotāju pārvaldība, meklēšana korpusā, koriģēšanas iespējas, pilna teksta pārlūkošana (Waldenfels, Woźniak 2016).

MuLaR pilotkorpuss ir veidots uz SpoCo platformas otrās versijas bāzes, šī versija tika veidota tieši Spišas dialektu korpusam. Platforma piedāvā meklēšanas iespējas, ierakstot pilnu

vārdu, vārda sākuma vai beigu daļu. Ir iespējams ierobežot meklēšanu, metadatu sadaļā atlasot datus pēc avota (Latgalē, Austrumu diasporā vai medijos iegūti runas dati), pēc runātāju dzimuma, vecuma, dzīvesvietas, kā arī viena informanta runā (sk. 2. attēlu).

Meklēšanas rezultāts ir redzams segmenta līmenī, vajadzības gadījumā ir iespēja paplašināt kontekstu. Paralēli pieejams audiofails un transkripcija (sk. 3. attēlu).

Secinājumi un ieceres

Aspektam, ka MuLaR ir pirmais latgaliešu valodas runas korpuss Latvijā, ir gan pozitīvās, gan negatīvās puses. Pozitīvās: pētnieki var radoši izpausties, pētīt un izvēlēties viņiem saistošāko korpusa dizainu, datorprogrammas, iegūt pieredzi no ārzemju kolēģiem, pētniekiem u. tml.; negatīvās: darbs pie runas korpusa ir komplicēts un laikietilpīgs.

Ļoti svarīgs ir plānošanas posms, lai nodrošinātu korpusa reprezentativitāti, lai tiktu savākta un precīzi dokumentēta visa korpusa izmantošanai būtiskā informācija (vienošanās ar informantiem, vajadzīgie metadati, audiofaili). Svarīgi ir plānot, kāds būs tehniskais nodrošinājums, lai tiktu veikti kvalitatīvi audioieraksti un droša datu glabāšana. Jau plānošanas posmā ir svarīgi izlemt, kāda programmatūra tiks izmantota tekstu transkribēšanai un korpusa platformas izveidei (meklēšanas iespējas, iespējas transkribēto tekstu sasaistīt ar audiofailiem). Nozīmīgs darbs tiek veikts, izstrādājot mutvārdu tekstu transkribēšanas konvencijas, ir jāizdara izvēle starp plašumu un dziļumu, proti, starp lielu mutvārdu tekstu datu apjomu un detalizētu datu anotāciju. Ilgstošākais un darbietilpīgākais posms ir tekstu transkribēšana (vienas minūtes audioieraksta atšifrēšana atkarībā no pētnieka pieredzes prasa laiku līdz pat vienai stundai). Tā kā transkribēšanu veic galvenokārt trīs projekta dalībnieki, ļoti svarīgi ir saglabāt augstu konsekvences līmeni transkribēšanas procesā.

Līdz 2022. g. beigām plānots pabeigt lauka pētījumos iegūto audioierakstu transkribēšanu un pēc izvērtēšanas integrēt tos MuLaR pilotkorpusā. Projekta noslēgumā paredzēts veikt plānveidīgu MuLaR pilotkorpusa validāciju un tā darbības efektivitātes izvērtēšanu. 2021. gada decembrī ir apstiprināts valsts pētījumu programmas "Letonika latviskas un eiropeiskas sabiedrības attīstībai" projekts "Latviešu valodas daudzveidība laikā un telpā", kurā viens no uzdevumiem ir turpināt MuLaR izstrādi, balstoties projekta "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" īstenošanas gaitā gūtajā pieredzē. Projektā, kurš turpināsies līdz 2024. gada decembrim, plānota jaunu latgaliešu runas datu ieguve (intervijas iepriekš neaptvertos Latvijas novados, radio un TV ieraksti), tekstu transkribēšana un ievietošana mutvārdu korpusā. Vienlaikus tiek plānoti arī turpmākie korpusa pilnveides posmi, izvērtējot transkribēto tekstu morfoloģiskās marķēšanas iespējas.

Bibliogrāfija

Adolphs, Svenja, Dawn, Knoght (2010). Building a spoken corpus: what are the basics? O'Keeffe, Anne, McCarthy, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. London, New York: Routledge Taylor & Francis group, pp. 38–52.

Anderson, Jean, Beavan, Dave, Kay, Christian (2007). SCOTS: Scottish corpus of texts and speech. Beal, Joan, Corrigan, Karen, Moisl, Hermann (eds.). *Creating and digitizing language corpora. Volume 1: Synchronic databases*. Basingstoke: Palgrave Macmillan, pp. 17–34. Available: https://core.ac.uk/download/pdf/110386.pdf [accessed 09.11.2021.].

Anderwald, Lieselotte, Wagner, Susanne (2007). FRED – The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. Beal, Joan, Corrigan, Karen, Moisl, Hermann (eds.). *Creating and digitizing language corpora. Volume 1: Synchronic databases*. Basingstoke: Palgrave Macmillan, pp. 35–53. Available: https://citeseerx.ist.psu.edu/viewdoc/download?-doi=10.1.1.519.4935&rep=rep1&-type=pdf [accessed 09.11.2021.].

Biber, Douglas (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, No. 8 (4), pp. 243–257. Available: http://otipl.philol.msu.ru/media/biber930.pdf [accessed 09.11.2021.].

Burnard, Lou (2005). Metadata for Corpus Work. Wynne, Martin (ed.). *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, pp. 30–46. Available: http://ota.ox.ac.uk/documents/creating/dlc/ [accessed 09.11.2021.].

Grochola-Szczepanek, Helena, Górski L. Rafal, von Waldenfels, Ruprecht, Woźniak, Michał (2019). Korpus języka mówionego mieszkańców Spisza. *LingVaria*, No. 14(27), s. 165–180. Available: https://w.akademicka.pl/ojs/lv/article/view/727/712 [accessed 14.11.2021.].

Hernández, Nuria (2006). *User's Guide to F R E D Freiburg Corpus of English Dialects.* Freiburg: English Dialects Research Group, Albert-Ludwigs-Universität. Available: https://freidok.uni-freiburg.de/fedora/objects/freidok:2489/datastreams/FILE1/content [accessed 14.11.2021.].

Komrsková, Zuzana, Kopřivová, Marie, Lukeš, David, Poukarová,-Petra, Goláňová, Hana (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný casopis*, No. 68(2), pp. 219–228. Available: https://doi.org/10.1515/jazcas-2017-0031 [accessed 14.11.2021.].

Moseley, Christopher (ed.) (2010). *Atlas of the World's Languages in Danger*, 3rd edn. Paris, UNESCO Publishing. Available: http://www.unesco.org/culture/en/endangeredlanguages/atlas [accessed 14.11.2021.].

Pasikowska-Schnass, Magdalena (2016). Regional and minority languages in the European Union. *European Parliamentary Research Service*. September 2016. Available: https://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589794/EPRS_BRI(2016)589794_EN.pdf [accessed 14.11.2021.].

Pinnis, Mārcis, Auziņa, Ilze, Goba, Kārlis (2014). Designing the Latvian Speech Recognition Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik: European Language Resources Association (ELRA), pp. 1547–1553. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/284_Paper.pdf [accessed 09.11.2021.].

Pošeiko, Solvita (2016). Latgaliešu valodas attīstība. Lauze, Linda (zin. red.). *Valodas situācija Latvijā: 2010–2015. Sociolingvistisks pētījums*. Rīga: LVA, 173.–194. lpp.

Rießler, Michael, Wilbur, Joshua (2017). Documenting endangered oral histories of the Arctic: A proposed symbiosis for endangered language documentation and oral history research, illustrated by Saami and Komi examples. Kasten, Erich, Roller, Katja, Wilbur, Joshua (eds.). *Oral History Meets Linguistics.* Fürstenberg, Havel: Kulturstiftung Sibirien, pp. 31–64. Available: https://www.researchgate.net/publication/314577167_Documenting_endangered_oral_histories_of_the_Arctic_A_proposed_symbiosis_for_language_documentation_and_oral_history_research_illustrated_by_Saami_and_Komi_examples [accessed 14.11.2021.].

Stefanowitsch, Anatol (2020). *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press. Available: https://books.google.lv/books?id=3ZHeDwAAQBAJ&printsec=frontcover&hl=lv&-source=gbs_ge_summary_r&-cad=0#v=onepage&q&f=true [accessed 14.11.2021.].

Tauginienė, Loreta, Ojsteršek, Milan, Foltýnek, Tomáš, Marino, Franca, Cosentino, Marco, Gaižauskaitė, Inga, Glendinning, Irene, Sivasubramaniam, Shiva, Razi, Salim, Ribeiro Laura, Odiņeca, Tatjana,

Trevisiol, Oliver (2019). *Akadēmiskā godīguma vispārējās vadlīnijas.* Pieejams: https://www.academicintegrity.eu/wp/wp-content/uploads/2019/10/RED_Guidelines_RTU_VS_amended_v2.pdf [skatīts 09.11.2021.].

Thompson, Paul (2005). Spoken language corpora. Martin Wynne (ed.). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, pp. 59–70. Available: http://ota.ox.ac.uk/documents/creating/dlc/ [accessed 09.11.2021.]

von Waldenfels, Ruprecht, Woźniak, Michał (2016). SpoCo – a simple and adaptable web interface for dialect corpora. Kupietz, Marc, Geyken, Alexander (eds.). *Journal for Language Technology and Computational Linguistics. Corpus Linguistic Software Tools*, No. 31 (1), pp. 155–170. Available: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6218/file/Journal_for_language_technology_and_computational_linguistics_1_2016.pdf [accessed 14.11.2021.].

Vuković, Teodora (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. *Lang Resources & Evaluation*, No. 55, pp. 731–756. Available: Representing variation in a spoken corpus of an endangered dialect: the case of Torlak | SpringerLink [accessed 25.10.2021.].

Creation of Contemporary Latgalian Speech Corpus in the Context of Documenting Lesser Used Languages

Angelika Juško-Štekele, Antra Kļavinska

Keywords: corpus linguistics, representativeness, corpus design, metadata, transcription, convention

According to data of UNESCO, in 2013, Latgalian language with 150,000 users was recognised as one of the world's endangered and vulnerable languages, as all generations still use the oral form, but the sustainability of the language is seriously jeopardised, since the number of young language users decreases. Pursuant to the EU directives and recommendations for preservation, research and development of regional and endangered languages, as well as the Guidelines for the State Language Policy 2021–2027 regarding development, disclosure on the web and accessibility of varied text corpus, in 2020, a group of researchers of the Rēzekne Academy of Technologies in the Project of State Research Programme *Digital Resources of Humanities: Integration and Development* (No. VPP-IZM-DH-2020/1-0001) started its work on the development of the Contemporary Latgalian Speech Corpus (MuLaR) aimed at the documentation, research, studies and acquisition of Latgalian.

The aim of the article is to identify and analyse the issues that are important in the process of creating MuLaR, applying the referential analysis of the scientific literature and comparative methodology. In turn, applying the analytical-synthetic method and based on the experience accumulated by the corpus creators, there was developed an initial model for the corpus architectonics and technological solutions, covering such issues as ensuring a representative Latgalian speech corpus, bearing in mind the territorial distribution of Latgalian language communities and diversity of Latgalian patois; the most appropriate methods to document natural, spontaneous language: collection of new data, opportunities to use the existing recordings (interviews, TV, radio broadcasts, field research data collections), other databases (reiti.rta.lv); understanding metadata; ethical aspects of the speech corpus; transcribing (software, conventions to reveal the features of spoken text as accurately as possible); creation of an accessible, easy-to-use open-access platform, using the experience of creating oral speech corpuses for lesser-used languages / dialects in other countries. The article declares the main challenges for the corpus development after the initial validation of the corpus data, including in relation to the morphological tagging possibilities of the corpus.