10.35539/LTNC.2022.0047.S.M.A.B.N.N.208.225

Latgaliešu valodas korpuss citu Eiropas mazāk lietoto valodu kontekstā: korpusa raksturojums, lietojums un potenciālā iespējošana

Sanita Martena, Anna Briška,
Nikole Naua (*Nicole Nau*)

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: latgaliešu valoda, korpuss, latgaliešu valodas korpuss, MuLa, reģionālās valodas, Eiropas mazāk lietotās valodas

Ievads

2012. gadā tika izveidots pirmais mūsdienu latgaliešu valodas korpuss – MuLa (http://www.korpuss.lv/id/MuLa), kurā tika iekļauti latgaliešu rakstu valodā publicētie teksti. Korpusa apjoms ir 1 milj. vārdlietojumu, un korpusa tekstu lielāko daļu veido literārie un informatīvie teksti, kas tapuši laika posmā no 1988. līdz 2012. gadam. Kopš 2020. gada ir atjaunots darbs ar MuLa, atlasot korpusam tekstus, kas publicēti pēc 2012. gada, plānojot tajā vairāk iekļaut zinātnisku, populārzinātnisku un periodikas tekstu fragmentus.

Valodas korpusi sākotnēji tika izstrādāti valodnieku vajadzībām; arī mūsdienās korpusi galvenokārt tiek izmantoti lingvistiskiem pētījumiem, taču citi iemesli, kāpēc valodas korpuss ir nepieciešams, var būt pat svarīgāki. Korpuss ir izejas punkts valodas tehnoloģijām un rīkiem (piemēram, tādiem kā pareizrakstības pārbaudītāji), korpusa dati ir pamats mūsdienu vārdnīcu sagatavošanai, un korpusi tiek izmantoti, izstrādājot valodu mācību materiālus. Mazāk lietotajās valodās, tādām kā latgaliešu, korpusam ir svarīga loma valodas dokumentēšanā vēl jo vairāk tad, ja tā saturs ir speciāli sagatavots, atlasīts un tajā ir ietverti mazāk pieejami avoti un valodas kopienai nozīmīgi teksti. Tātad korpuss kalpo ne tikai pētniekiem, bet arī valodas runātāju kopienai.

Ņemot vērā iepriekš minēto, raksts sniedz ieskatu mūsdienu latgaliešu valodas korpusa MuLa pilnveides un attīstīšanas procesā, īpašu uzmanību pievēršot korpusa pašreizējai un potenciālajai mērķauditorijai un salīdzinot korpusu ar citiem Eiropas mazāk lietoto valodu korpusiem.

Veidojot rakstu, tika izvirzīti trīs galvenie pētījuma jautājumi:

- kas ir raksturīgs Eiropas mazāk lietoto valodu korpusiem,
- kāds būs šobrīd papildinātais un pilnveidotais mūsdienu latgaliešu valodas korpuss, salīdzinot ar 2012. gadā plašākai sabiedrībai publicēto korpusu;
- kas ir līdzšinējais MuLa korpusa lietotājs, kādiem nolūkiem korpuss ir izmantots, un kā korpuss varētu tikt lietots nākotnē.

Iepriekš minētie pētījuma jautājumi ir noteikuši raksta struktūras izveidi. Raksta pamatā ir trīs galvenās nodaļas: pirmajā nodaļā ir dots ieskats citu Eiropas mazāk lietoto valodu korpusu piedāvājumā, lai noteiktu situāciju ar MuLa korpusa izstrādi plašākā Eiropas kontekstā. Otrajā nodaļā ir piedāvāts īss MuLa 2012. gada versijas rakstu-rojums un salīdzinājums ar šobrīd izstrādājamo jauno versiju. Trešās

nodaļas pamatā ir 2021. gada pavasarī veiktās aptaujas rezultāti; aptauja tika veikta ar mērķi izzināt latgaliešu valodas korpusa lietotāju profilu un viņu redzējumu korpusa uzlabošanā un lietošanā. Raksta noslēgumā ir dots pētījuma kopsavilkums un secinājumi par MuLa korpusa funkcionalitāti un izmantošanas iespējām nākotnē.

1. Mazāk lietoto valodu korpusi

Elektroniskie valodas korpusi, kuros ir liels rakstīto tekstu fragmentu skaits, kas reprezentē kādu valodu vai valodas paveidu, tiek izstrādāti jau kopš 20. gs. 60. gadiem. Paralēli ir attīstījusies arī korpuslingvistika – valodniecības nozare, kas pēta korpusu sastādīšanu un izmantošanu valodas struktūru analīzei (Sampson, McCarthy 2007; McEnery, Hardie 2012: 71). Sākotnēji pētījumi tika koncentrēti uz angļu valodu un citām valodām, kam ir liels runātāju un arī pētnieku skaits, bet kopš 90. gadu beigām tiek uzsvērta korpusu nozīme tieši mazāk lietoto valodu dokumentēšanā un pētīšanā, kas vienlaikus nozīmē arī šo valodu attīstību (McEnery, Ostler 2002; Ostler 2009; Cox 2011; Knight et al. 2021). Tāpat arvien vairāk paplašinās un kļūst daudzveidīgāks korpusu izmantošanas lauks.

21. gs. sākumā UNESCO darba grupa, pievēršot uzmanību daudzu pasaules valodu apdraudētībai, izvirzīja 9 galvenos kritērijus valodu vitalitātes jeb dzīvotspējas pakāpes novērtēšanai. Pēc šiem kritērijiem ir iespējams noteikt valodas pašreizējo funkcionalitāti, kā arī prognozēt tās nākotni (UNESCO 2017). Starp 9 kritērijiem divi gan tiešā, gan netiešā veidā ir saistīti ar korpusu lomu valodas funkciju nodrošināšanā un paplašināšanā. Pirmais no šiem kritērijiem ir valodu dokumentēšana (*Type and Quality of Documentation*), uzsverot steidzamību valodu dokumentēšanā, lai noteiktas valodas runātāju kopiena varētu formulēt specifiskus uzdevumus valodas attīstīšanai un lai identificētais valodas materiāls (teksti un audiovizuālais materiāls) būtu pietiekami apjomīgs un daudzveidīgs, izstrādājot gramatikas izdevumus, vārdnīcas u. tml. Otrais kritērijs ir – jaunu jomu un mediju attīstīšana (*Response to New Domains and Media*), ar to saprotot valodas lietojumu gan nu jau tādos tradicionālajos medijos kā televīzija, radio, gan arī valodas daudzveidīgu lietojumu interneta vidē, tīmeklī pieejamos medijos par visdažādāko saturu un ar daudzveidīgu valodas variantu lietojumu. Valodu korpusi šajā ziņā ir neatsverams tehnoloģiju laikmeta resurss gan valodu dokumentēšanai un izpētei, gan paši var kļūt par vērtīgu krātuvi dažādu lingvistisko avotu, mācību līdzekļu un citu materiālu izstrādei.

2000. gadā tika publicēts raksts ar programmatisku nosaukumu "Korpuslingvistikas jaunā dienas kārtība: strādāt ar visām pasaules valodām" (*A new agenda for corpus linguistics: Working with all of the world's languages*) (McEnery, Ostler 2000). Raksta autori sniedza īsu kopsavilkumu par korpusu sagatavošanu pasaules valodās 20. gadsimtā. 20. gadsimta beigās daudzām valodām joprojām vēl nebija korpusu, un autori uzskatīja, ka šī situācija ir jāmaina, jo korpusiem būs arvien svarīgāka loma valodas attīstīšanā. Rakstā tika minētas dažādas valodu kategorijas, kuru valodu korpusu izstrāde kavējas, un aplūkotas atsevišķas problēmas katrā no kategorijām (McEnery, Ostler 2000: 405–410). Tika norādīts, ka kategorijas

raksturo šādas pazīmes: runātāju skaits, valodas statuss (piemēram, valsts valoda, atzīta minoritāšu vai reģionālā valoda, dialekts), runātāju statuss (pamatiedzīvotāji/pirmiedzīvotāji vai imigranti); atsevišķa kategorija ir zīmju valodas.

Aplūkojot situāciju Eiropā, var konstatēt, ka 21. gadsimta pirmajā desmitgadē korpusi ir izstrādāti valodām, kam Eiropas mērogā ir samērā maz runātāju, piemēram, islandiešu valodai (korpusa izstrāde 2006.–2010. g., tagadējais apjoms 25 miljoni vārdlietojumu; avots: *Tagged Icelandic Corpus*) vai latviešu valodai (Levāne-Petrova 2012). Savukārt otrajā desmitgadē korpusi tiek veidoti tādām Eiropas valodām, kuras tiek iekļautas citās valodu kategorijās (Mc Enery un Ostler 2000). Tās ir oficiāli atzītas reģionālās vai minoritāšu valodas, vai arī valodas un dialekti, kam attiecīgajā valstī nav noteikta statusa. Runātāju skaita ziņā šīs valodas stipri atšķiras. Piemēram, katalāņu un galisiešu valodai, kurām Spānijā attiecīgajos reģionos ir oficiālās valodas statuss, ir vairāki miljoni runātāju, savukārt ziemeļsāmu valodā, kam Norvēģijā ir īpašs pirmiedzīvotāju valodas statuss, runā mazāk par 25 000 cilvēku, bet lejas sorbu valodā (minoritātes valoda Vācijā) ir tikai 7000 runātāju. Pie valodām, kuru runātāju skaits ir starp 100 000 un 1 miljonu, pieder basku valoda Spānijā, velsiešu valoda Lielbritānijā, rietumfrīzu valoda Nīderlandē (visām trim ir oficiāls statuss savā reģionā), kā arī silēziešu valoda Polijā un astūriešu valoda Spānijā (nevienā no abām valstīm valodām nav piešķirts kāds oficiālas valodas statuss).

1. tabulā ir dota pamatinformācija par dažu iepriekš minēto mazāk lietoto valodu tekstu korpusiem. Informācijas avoti ir šo korpusu tīmekļvietnes (sk. avotu sarakstu) un raksti: Bartels 2012 (par DTK), Kulik 2018 (par KŚM), Viejo u. c. 2008 (par ESLEMA).

Valoda (valsts)	Statuss	Korpusa nosaukums, pieejams no ... (gads)	Korpusa apjoms (vārdlietojumi)
Basku (Spānija)	Oficiālā valoda reģionā	ETC; 2013. g.	355 miljoni
Astūriešu (Spānija)	Nav statusa	ESLEMA; 2008. g.	(nav pabeigts)
Lejas sorbu (Vācija)	Minoritātes valoda	DTK; 2010. g.	23 miljoni
Velsiešu (Lielbritānija)	Oficiālā valoda reģionā	CorCenCC; 2020. g.	11 miljoni
Ziemeļsāmu (Norvēģija)	Pirmiedzīvotāju valoda	SIKOR; 2015. g.	8,9 miljoni
Silēziešu (Polija)	Nav statusa	KŚM; 2018. g.	2 miljoni
Latgaliešu (Latvija)	Valsts valodas paveids	MuLa; 2012. g.	1 miljons

1. tabula. Piemēri ar Eiropas mazāk lietotajām valodām, to statusu, korpusu izveides gadu un apjomu.

Iegūstot informāciju par Eiropas mazāk lietoto valodu korpusiem, raksta autores ir secinājušas, ka runātāju skaits nekorelē ar korpusa pieejamību vai tā apjomu. Svarīgāks ir tas, vai valodai ir finansiāls un institucionāls atbalsts. Šajā ziņā valodas ar noteiktu statusu ir labākā situācijā, jo tām bieži ir savi institūti ar pastāvīgu finansējumu, savukārt citām mazāk

lietotajām valodām un dialektiem šāda atbalsta var trūkt vai tas tiek piešķirts neregulāri (sk. Viejo u. c. 2008 par astūriešu valodu). Tikpat nozīmīgas ir arī atsevišķu pētnieku vai valodas aktīvistu iniciatīvas. Tā, piemēram, kašubu valodai, kam Polijā ir reģionālās valodas statuss un valsts atbalsts, tekstu korpuss tikai tagad top (Pomierska, Stanulewicz 2019), bet oficiāli neatzītajai silēziešu valodai korpuss ir tapis, pateicoties viena cilvēka entuziasmam (Kulik 2018).

Tālāk rakstā ir aplūkotas trīs pazīmes, pēc kurām var raksturot jebkuru tekstu korpusu, tajā pašā laikā īpaši nozīmīgas tās ir tieši mazāk lietoto valodu korpusu raksturošanā: (i) ko korpuss ietver (kādu žanru, kāda laikposma, kāda stila tekstus); (ii) kādam nolūkam tas ir domāts, kādi lietojumi tikuši paredzēti, plānojot korpusu; (iii) kā tas ticis sastādīts (vairāk manuāli vai vairāk automātiski). Visi minētie jautājumi ir savstarpēji cieši saistīti.

Mazāk lietotās rakstu valodas bieži netiek lietotas visās sfērās, kurās funkcionē Eiropas valstu oficiālās jeb valsts valodas. Līdz ar to tāda korpusa sastādīšana, kas būtu gan līdzsvarots, gan reprezentatīvs, var būt īpaši problemātiska, tomēr korpusu veidotāji meklē un atrod tam dažādus risinājumus. Tāpat arī dažām Eiropas reģionālajām vai mazākumtautību valodām ir bagāta vēsture, tomēr mūsdienās tās samērā maz tiek lietotas rakstu formā. Ja galvenais korpusa mērķis ir valodas dokumentēšana un avotu, tostarp vēsturisko, bāzes veidošana, tad korpusa līdzsvarotība ir mazāk svarīga. Korpusā tiek iekļauts pēc iespējas vairāk veselu, pilna apjoma tekstu, nevis tikai to fragmenti, kā tas mēdz būt nacionālajos korpusos. Piemēram, lejas sorbu korpus*ā* (DTK, sk. 1. tabulu) plānots ietvert visus lejas sorbu valodā publicētos (drukātos) tekstus; pašlaik tajā ir iekļauti gandrīz visi teksti, kas publicēti periodā no 1848. līdz 1937. gadam, un ir iesākta jaunāku tekstu iekļaušana (Bartels 2012: 13; Bartels 2020). Savukārt Skotijas gēlu valodas korpuss *Corpas na Gàidhlig* ietver 317 izvēlētus tekstus no 18. gadsimta līdz mūsdienām. Šādi korpusi var būt īpaši noderīgi kā datubāze vārdnīcu izstrādei, kā arī valodas attīstības pētniekiem; mazāk noderīgi tie ir mūsdienu valodas apguvējiem vai pētniekiem un tikai nosacīti ir lietojami valodas tehnoloģijas rīku izstrādei.

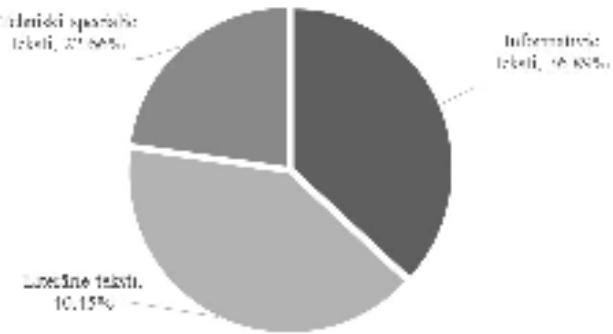
Jaunais velsiešu valodas korpuss CorCenCC ir tapis ar mērķi atspoguļot mūsdienu velsiešu valodu niansētā un plašā lietojumā, lai korpuss kalpotu dažādām tā lietotāju grupām (pētniekiem, pašai valodas lietotāju kopienai, valodas apguvējiem un citiem velsiešu valodas interesentiem). Atkarībā no korpusa lietotāju grupas ir uzsvērti arī tā izmantošanas daudzveidīgie mērķi: valodnieciskiem pētījumiem, savas valodas izlokšņu izzināšanai, valodas mācībām, vārdnīcu un digitālo rīku izstrādei (Knight et al. 2020: 3). Šie mērķi ir noteikuši iekļaujamo tekstu izvēli. Korpusa veidotāji ir centušies sameklēt tādus paraugus, kas būtu tipiski mūsdienu velsiešu valodas lietojumam autentiskās valodas lietojuma jomās, ņemot vērā, ka tā ir bilingvāla sabiedrība, kurā velsiešu valoda funkcionē paralēli angļu valodai, bet ne visās jomās (Knight et al. 2020: 6). Korpusā tika iekļauti dažādu žanru drukātie teksti, kā arī teksti no elektroniskajiem medijiem (tīmekļvietnes, emuāri, e-pasta vēstules, īsziņas); trešais korpusa satura komponents ir runas paraugu atšifrējumi (Knight u. c. 2020: 7–9). Minētā vīzija par korpusa CorCenCC plašo lietojumu ir redzama arī korpusa prezentācijā internetā, kas ar lietotājiem draudzīgo dizainu uzrunā gan profesionāļus, gan ikvienu valodas interesentu. Pieejami ir arī rīki valodas mācībām, bet perspektīvā ir tēzaura izveide, kas ļaus atrast noteikta semantiskā lauka vārdus un teicienus (Knight et al. 2020).

Velsiešu valodas korpuss CorCenCC noteikti kalpos par paraugu un iedvesmas avotu mazāk lietoto (tomēr dzīvo un dzīvotspējīgo) valodu korpusu izveidei nākotnē. Autori par savu pieredzi un secinājumiem ir arī publicējuši grāmatu (Knight et al. 2021). Tomēr ir skaidrs (to atzīst arī paši autori), ka tikai nedaudzas Eiropas mazāk lietotās valodas var cerēt uz tādiem finansiāliem un cilvēku resursiem, kādi vajadzīgi šādam lielam projektam. Resursu trūkuma dēļ arvien biežāk korpusi tiek sastādīti (pus)automātiski, izmantojot programmatūras, kas sameklē tekstus internetā. Šī metode izmantota arī ar vairāk lietotām un labi dokumentētām valodām, lai iegūtu lielāku daudzumu datu. Piemēram, latviešu valodas korpuss lvTenTen14, kas ir pieejams platformā *sketchengine.eu*, šobrīd ir vislielākais latviešu valodas korpuss, tam ir 530 367 474 vārdlietojumi (par TenTen sērijas korpusiem sk. Jakubíček u. c. 2013). Mazu resursu valodām (*low resource languages*) automātiskā korpusa izveide dažreiz ir vienīgais veids, kā tikt pie korpusa (Scannel 2007; Jauhiainen et al. 2020). Taču tieši šīm valodām tāda metode var radīt tehniskas problēmas (piemēram, ar automātisko valodas atpazīšanu) un būtiskus trūkumus: internetā brīvi pieejamiem tekstiem ne vienmēr ir laba kvalitāte, un tie nepietiekami atspoguļo rakstu valodas žanru un stilu daudzveidību (Jauhiainen et al. 2020; Millour, Fort 2020). Automātiski no interneta resursiem sastādītie korpusi var būt piemēroti valodas tehnoloģijas rīku izstrādei (Bernhard u. c. 2021), taču valodas attīstīšanai un apguvei, tāpat arī valodas dokumentēšanai un vārdnīcu izstrādei labāk der manuāli sagatavotie korpusi, kuros ir rūpīgi atlasīti un pārbaudīti teksti vai to fragmenti.

2. Latgaliešu valodas tekstu korpuss

Pirmo latgaliešu valodas tekstu korpusu tika sākts veidot pirms aptuveni 10 gadiem (2011–2012), kad Rēzeknes Augstskolā ar Latvijas–Lietuvas pārrobežu sadarbības programmas atbalstu tika realizēts projekts "HipiLatLit" ("Humanitārās izglītības pētniecības infrastruktūras izveide Austrumlatvijā, Lietuvā"). Projekta laikā, sadarbojoties Rēzeknes Augstskolai, Vītauta Dižā universitātei Kauņā un Latvijas Universitātes Matemātikas un informātikas institūtam, kā viens no rezultātiem tika izveidots Mūsdienu latgaliešu tekstu korpuss – MuLa. Korpusa apjoms bija 1 milj. vārdlietojumu, un tajā tika iekļauti Latvijā latgaliešu rakstu valodā publicēti teksti laika posmā no Atmodas (1988) līdz 2012. gadam. Tas tika izveidots kā speciālais valodas korpuss, iekļaujot tekstus noteiktās proporcijās atbilstoši latgaliešu rakstu valodas lietojuma specifikai.

Speciālā latgaliešu valodas korpusa tekstu veidu iedalījums tika aizgūts no čehu nacionālā korpusa izstrādātājiem (Čermák 2011). Informatīvo tekstu grupā tika iekļauti raksti laikrakstos, avīzēs, žurnālos, interneta portālos, blogos, kā arī ziņas un ceļojumu apraksti. Tehniskie un specifiskie teksti tika iedalīti zinātniskajos rakstos, reliģiskajos tekstos un mācību līdzekļos. Pie daiļliteratūras tekstu kopas tika pievienota latgaliski publicētā proza un dzeja. Sākotnēji tika plānots, ka tehniskie un specifiskie teksti veidos 35 % no visa korpusa apjoma, bet darba procesā tika konstatēts, ka reāli pieejamo tekstu proporcija atšķiras no plānotās, un tie bija tikai ~22 %



1. attēls. MuLa korpusa vārdlietojumi (%) dažādos teksta veidos (Briška 2013: 32).

(sk. 1. attēlu). Savukārt literāro tekstu daļa no plānotajiem 30 % tika palielināta līdz ~40 %, bet informatīvie teksti – nepilni 37 % vārdlietojumu plānoto 35 % vietā (Briška 2013: 13, 32).

Pēc 10 gadu pārtraukuma, Valsts pētījumu programmas (VPP) projekta "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (2020–2022) laikā, latgaliešu valodas tekstu korpuss MuLa tiek papildināts ar tekstiem, kas ir publicēti pēc 2012. gada. Tāpat paralēli tiek veidots pilnīgi jauns – latgaliešu mutvārdu runas korpuss, kurā transkribēti Latgales iedzīvotāju un Sibīrijas latgaliešu runas ieraksti (vairāk sk. šajā izdevumā: Juško-Štekele, Kļavinska).

VPP projektā MuLa korpusa izstrādei tika izvirzīti šādi mērķi: precizēt iepriekšējo korpusa versiju, novēršot atsevišķas nepilnības un dažu tekstu dubultošanos; papildināt MuLa ar jauniem tekstiem, kas veidotu kopumā 1 milj. vārdlietojumu (palielinot korpusa apjomu līdz 2 milj. vārdlietojumu); saglabāt tekstu līdzsvarotību, apzinot visus iespējamos latgaliski publicētos dažādu žanru tekstus drukātā versijā un elektroniski pieejamos materiālus digitālajā vidē; korpusam atlasīt pēc iespējas rediģētus un publicētus (arī e-vidē pieejamus) tekstus.

Mūsdienu latgaliešu rakstu valodā, īpaši neformālās valodas lietojuma situācijās (sociālo tīklu vietnēs, privātā komunikācijā u. c.) un arī uzrakstos publiskajā telpā ir vērojama lingvistiskā varietāte. Tā kā valodas lietotāji latgaliešu valodu pārsvarā ir apguvuši tās mutvārdu formā, sazinoties ģimenē, un rakstīšanas prasmi nav mācījušies (Martena, Marten, Šuplinska 2022), tad teksti tiek rakstīti nevis ievērojot ortogrāfijas noteikumus, bet pēc izrunas, intuīcijas, atbilstoši vietējai izloksnei u. tml. Domājot par korpusa ilgtermiņa izmantošanas iespējām, tika atlasīti un korpusā iekļauti tikai rediģēti un publicēti teksti.

Kā izskatās papildinātā MuLa korpusa versija, salīdzinot ar korpusu, kas tapa pirms desmit gadiem? Pēc kādiem kritērijiem un kuri teksti ir atlasīti 2022. gada versijai?

Šobrīd, salīdzinot ar situāciju pirms 10 gadiem, latgaliešu valodā vairāk ir pieejami zinātniskie un populārzinātniskie teksti, kā arī masu mediji (drukātie un digitālajā vidē). Piemēram, regulāri kopš 2012. gada sešas reizes gadā iznāk žurnāls "A12", portālā "LaKuGa" ("Latgalīšu Kulturys Gazeta") kopš 2013. gada katru mēnesi tiek publicēti vidēji 45 raksti (vidējais rādītājs laika posmā no 2013. gada janvāra līdz 2021. gada oktobrim). Līdz ar to 2022. gada korpusa versijā pieaug zinātnisko un populārzinātnisko, kā arī periodikas tekstu īpatsvars.

Korpusam no periodikas ir atlasīti teksti latgaliešu valodā no tādiem laikrakstiem un žurnāliem kā "Latgales Laiks", "A12", no digitālajā vidē pieejamā latgaliešu kultūras ziņu portāla *lakuga.lv* un no Latvijas sabiedrisko mediju portāla *lsm.lv*.

No zinātniskajiem un populārzinātniskajiem tekstiem šobrīd visvairāk tekstu ir sagatavots no Latgales lingvoteritoriālās vārdnīcas (Šuplinska 2012), kura tika izdota 2012. gadā. Atlasot vārdnīcas tekstus, tika ievērota šķirkļu autoru un viņu dzimumu proporcionalitāte, tematika un joma (ģeogrāfija, vēsture, folklora, valoda un literatūra u. c.), kā arī pašu šķirkļu satura daudzveidība (pilsētas, kultūras reālijas, notikumi, izdevumi, personības, dievības u. tml.). Starp šīs žanru grupas tekstiem korpusā ir iekļauti arī zinātniskie raksti, kas latgaliski publicēti tādos izdevumos kā "Via Latgalica" un "Latgalistikys kongresu materiali". Tāpat korpusā ir iekļauts arī populārzinātnisks izdevums "Latgalīša CV", "Latgalistikys kongresa materiali" (tēzes), LR Satversme, kas ir pieejama arī latgaliešu valodā u. c.

Mūsdienās ir pieaudzis arī daiļliteratūras klāsts latgaliešu valodā, un no tekstu pieejamības viedokļa tie ir iegūstami elektroniski, sazinoties gan ar pašiem autoriem, gan izdevniecībām (pirms 10 gadiem daudz laika tika veltīts tekstu ieskenēšanai no grāmatām, žurnāliem, laikrakstiem un jo īpaši – tekstu pārbaudīšanai). Līdz ar to atjaunotajā MuLa versijā ir redzama lielāka daiļliteratūras tekstu daudzveidība, jo ir iekļauti arī jaunu literātu teksti.

Starp korpusam atlasītajiem daiļliteratūras tekstiem ir gan teksti, kas rakstīti oriģinālvalodā (latgaliski), gan arī tie, kas pārveidoti latgaliski no latviešu literārās valodas vai tulkoti no citām valodām, piemēram, Kārļa Skalbes pasaka "Kaķīša dzirnavas", Antuāna de Sent-Ekziperī stāsts "Mazais princis", Lūisa Kerola darbs "Alises piedzīvojumi Brīnumzemē". Ir ievērots arī žanriskais princips, atlasot korpusam gan dzejoļu krājumus, gan prozas tekstus, gan arī atsevišķus dramaturģijas darbus (publicēto lugu starp visiem daiļliteratūras žanriem latgaliešu valodā ir vismazāk).

Viena gada laikā kopš jaunās VPP sākuma (pamatā – 2021. gadā) papildinātajai MuLa versijai kopumā ir iegūts aptuveni 565 000 vārdlietojumu. Starp tiem ir dažādu žanru mūsdienu teksti (publicēti pēc 2012. gada) ar metadatiem: periodika ~385 000 vārdlietojumu, daiļliteratūra ~106 000 vārdlietojumu, zinātniskie un populārzinātniskie teksti ~74 000 vārdlietojumu.

3. MuLa korpusa lietotāju profils un redzējums par korpusa izmantošanu

2021. gada pavasarī šī raksta autores veica anketēšanu par latgaliešu valodas tekstu korpusa "MuLa" lietošanu. Mērķgrupas, kuras īpaši tika uzrunātas, bija baltu valodu pētnieki Latvijā un ārzemēs, studenti, RTA tīmekļvietnes apmeklētāji, projekta "HipiLatLit" laikā 2012. gadā rīkoto korpusa lietošanas mācību dalībnieki (studenti, pētnieki, skolotāji), latviešu un latgaliešu valodas skolotāji, portāla "LaKuGa" veidotāji un lasītāji. Anketu varēja aizpildīt arī ikviens cits interesents. Anketēšanas mērķis bija noskaidrot, cik lielā mērā un kādos nolūkos korpuss tiek lietots, kā arī apzināt, ar kādām grūtībām saskaras korpusa lietotāji, lai turpmāk korpusu varētu pilnveidot un paplašināt tā lietojuma iespējas.

Anketa tika publicēta kā *Google* veidlapa, un no 2020. gada 17. marta līdz 20. aprīlim tika saņemtas pavisam 214 respondentu atbildes. Tika apzināti gan tie cilvēki, kuri korpusu lieto (grupa "Zina un lieto"), gan tie, kuri zina par korpusu, bet dažādu iemeslu dēļ nelieto (grupa "Zina un nelieto"), gan tie, kuriem korpuss ir kaut kas jauns (grupa "Nezina un nelieto"). Zīmīgi, ka procentuāli visvairāk savā grupā (86 % gadījumu) latgaliešu valodu privātajā sfērā lieto tie, kuri par korpusu nav dzirdējuši, turpretī MuLa korpusa lietotāji tikai 53 % gadījumu lieto latgaliešu valodu privātajā sfērā, toties šajā grupā par to ir lielāka zinātniskā un profesionālā interese. Savukārt tie, kuriem ir pedagoģiska interese par latgaliešu valodu, procentuāli visvairāk ir tajā respondentu grupā, kuri par MuLa korpusu zina, ir dzirdējuši, bet to nelieto.

Pirmā grupa "Zina un lieto"

Starp 214 respondentiem ir tikai 15 cilvēki, kuri korpusu lieto, no tiem deviņas sievietes vecumā no 32 līdz 52 gadiem un seši vīrieši vecumā no 32 līdz 69 gadiem.

Astoņiem respondentiem no 15 ir zinātniska interese par latgaliešu valodu, astoņi lieto valodu arī privātajā jomā, seši lieto latgaliešu valodu savā darbā, bet tikai viens respondents šajā grupā ir ar pedagoģisku interesi par latgaliešu valodu. Korpusa lietotāji to izmanto, lai meklētu piemērus, kas parāda, kā var lietot kādu vārdu, lai meklētu vārdu kontekstā, lai labāk saprastu tā nozīmi vai lai pārbaudītu pareizrakstību. Četri respondenti ir lietojuši korpusu, pētot konkrētu valodas parādību zinātniskam darbam.

Lai arī kopumā korpusa saturs un izmantošanas iespējas tika vērtēti kā labi (11 respondentu), tomēr starp būtiskākajiem korpusa trūkumiem tika minēti šādi: korpuss ir par mazu (7 respondenti), korpusa platformā ir pārāk maz informācijas par to, kā lietot korpusu (6 respondenti), nevar meklēt pēc gramatiskajām kategorijām (7 respondenti), nevar atrast visas viena vārda formas uzreiz (5 respondenti). Komentāros pavīdēja arī vēlme meklēt korpusā informāciju pēc noteikta teksta žanra, *lai salīdzinātu kādas vienības lietošanu dažādos žanros.*

Korpusa lietotāji anketā sniedza arī plašus komentārus par to, kas būtu uzlabojams. Apkopojot teikto, var secināt, ka ar korpusa platformu respondenti nav apmierināti divu

iemeslu dēļ: pirmkārt, platforma nav saprotama lietotājam, jo tā ir tikai angļu valodā (*tā nav latviskota (latgaliskota*)), otrkārt, nav plašāka apraksta ne par pašu korpusu, ne par tā lietošanu (*varbūt būtu nepieciešama saskarnes atvieglota (light) versija neprofesionāliem lietotājiem; citreiz īsti nesaprotu, kā meklēt, kā tur darboties, pietrūkst pacietības, un metu mieru. Lai gan gribētu prast ar to strādāt labāk*).

Otrā grupa "Zina un nelieto"

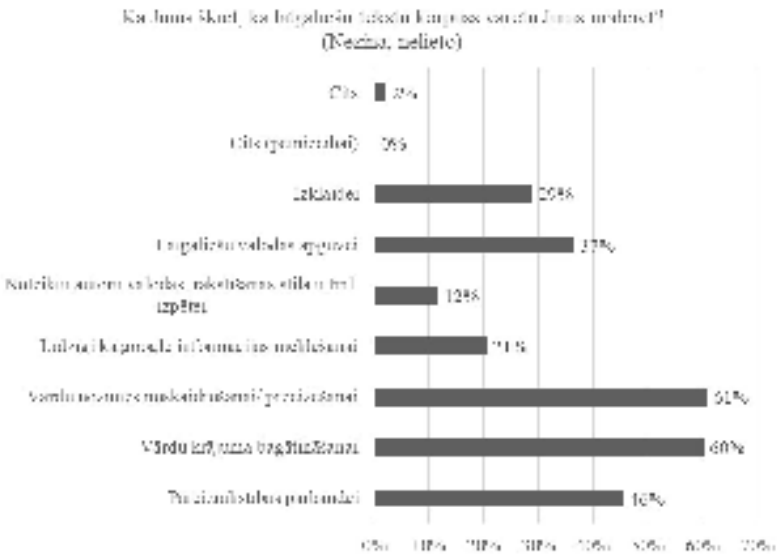
Viena piektā daļa jeb 43 aptaujas dalībnieki ir tie, kuri par MuLa korpusu ir dzirdējuši, taču to nelieto. Galvenie informācijas kanāli, kur respondenti ir uzzinājuši par korpusu, ir skola vai augstskola (37 %), sociālie mediji un draugi, kolēģi (30 %), platforma *korpuss.lv* (30 %). Pieci respondenti ir pat piedalījušies korpusu lietošanas mācību kursos, tomēr MuLa korpusu nelieto. Kā galvenie iemesli, kāpēc respondenti nelieto korpusu, ir minēti: nav vajadzības (84 %), nezina, ko īsti ar to darīt (19 %), kā arī – kā ar to rīkoties (12 %). Uz jautājumu, vai ir lietoti citi valodu korpusi, 70 % respondentu ir atbildējuši apstiprinoši (populārākie ir latviešu valodas, krievu valodas un angļu valodas korpusi).

Šīs grupas respondentu komentāros (kopā 10 komentāri) tika minēts gan potenciālais latgaliešu valodas korpusa lietotājs (*korpuss vajadzīgs latgaliešu valodas pētniekiem, arī lietotājiem*), gan izteiktas idejas, kas būtu darāms, lai korpusu varētu izmantot vairāk, un šie komentāri ir diezgan līdzīgi iepriekšējās – korpusa lietotāju – grupas viedokļiem. Tomēr šajā grupā vēlmes ir izteiktas precīzāk (*varētu kādu vebināru vai attālinātu semināru izveidot, popularizējot un izskaidrojot, kā praktiski to varētu lietot*), kā arī – attiecinot uz sabiedrības informēšanu kopumā (*plašāk informēt sabiedrību par korpusa izveidi un mērķi*).

Trešā grupa "Nezina un nelieto"

Trešā grupa – respondenti, kuri nav dzirdējuši par latgaliešu valodas korpusu, – bija vislielākā, pavisam 156 respondenti. Šai grupai pamatinformācija par MuLa korpusu, iespējams, tika iegūta, aizpildot anketu (tajā bija dota arī saite uz pašu korpusu, kuru anketas aizpildīšanas laikā varēja aplūkot).

Šajā grupā, salīdzinot ar abām iepriekšējām grupām, procentuāli visvairāk bija atbilžu, kurās bija norādīts, ka korpuss varētu noderēt arī izklaidei (sk. 2. attēlu). Tas ļauj izteikt pieņēmumu, ka potenciālais korpusa lietotājs sagaida arī kaut ko atraktīvu. Tie, kuri jau ir kaut ko zinājuši par korpusiem un lietojuši citu valodu korpusus, salīdzinoši vairāk ir norādījuši, ka latgaliešu tekstu korpuss varētu noderēt noteiktu autoru valodas, rakstīšanas stila izpētei, nekā tie, kuri neko nav dzirdējuši par korpusiem, attiecīgi 26 % un 12 % katrā grupā (sk. 2. attēlu). Tomēr populārākās atbildes uz jautājumu, kur latgaliešu tekstu korpuss varētu noderēt, bija trīs: vārdu nozīmes precizēšanai, vārdu krājuma bagātināšanai, pareizrakstības pārbaudei (sk. turpat). Visi trīs aspekti ir ļoti nozīmīgi jebkuras valodas apguvē, līdz ar to šis viedoklis būtu jāņem vērā, piemēram, izstrādājot uzdevumus latgaliešu valodas apguvei. Tos varētu



2. attēls. Otrās un trešās grupas respondentu redzējums par latgaliešu tekstu korpusa izmantošanas iespējām.

lietot gan skolēni un skolotāji, gan arī pieaugušie valodas apguvēji, tādā veidā ļaujot prognozēt, ka potenciālais korpusa lietotājs nākotnē varētu būt arī valodas apguvējs.

Apkopojot komentāru sadaļā rakstīto šajā grupā, kurā bija visvairāk respondentu (156 no 214), var minēt trīs būtiskākos – un tieši izglītojošos – aspektus. Pirmais – anketas aizpildīšana kā sevis informēšana par MuLa korpusu un pateicība par šo iespēju (*Paldies par MuLa vietnes informēšanu! Brīnišķīga iespēja, ar kuru noteikti iepazīšos. Pirmo reizi ieraudzīju, ņemšu vērā – paldies! Brīnišķī*gs korpuss, paldies par darbu!). Otrais – ierosinājumi korpusa popularizēšanai un cilvēku izglītošanai (*Noderētu plašāks info par šo lietotni. Lai iespējami vairāk parādās par to informācijas. Tā ir ļoti laba iespēja cilvēkiem uzlabot savas latgaliešu valodas zināšanas* u. c.). Trešais aspekts ir cieši saistīts ar raksta pirmajā nodaļā minētajiem UNESCO kritērijiem valodas dzīvotspējas nodrošināšanai, starp kuriem valodas dokumentēšana, reaģēšana uz jaunajiem medijiem un izmantošana dažādās tehnoloģijās ir īpaši nozīmīgi mazāk lietoto valodu kontekstā. Arī respondenti, komentējot MuLa korpusu, ir norādījuši uz tā nozīmi pašas valodas popularizēšanā, attīstīšanā un apguvē (*Ļoti jauki, ka Latgaliešu valoda tiek izzināta. Vai ir plānots popularizēt*

latgaliešu valodu? Tas noteikti palīdzētu saglabāt valodas dzīvību tālākajā nākotnē. Vēlētos, lai latgaliešu valodu neaizmirstu un mācītu arī interesentiem. Paļdis par latgalīšu volūdys cīneišonu, ceļšonu gaismā, jaunu asneņu deidzēšonu! Prīks.).

Nobeigums un secinājumi

Mazāk lietoto valodu korpusi ir nozīmīgi gan pašas valodas dokumentēšanā un attīstīšanā, gan ir vērtīgs resurss lingvistisko un lingvodidaktisko avotu izstrādē, vienlaicīgi arī popularizējot mazāk lietoto valodu lietojumu. MuLa korpuss nedz apjoma, nedz izmantošanas iespēju ziņā nav salīdzināms ar tādiem reģionālo valodu korpusiem kā basku, velsiešu vai ziemeļsāmu valodas korpusiem. Tajā pašā laikā tas ir rūpīgi izveidots latgaliešu rakstu valodas reprezentatīvo tekstu korpuss.

Apjoma ziņā MuLa korpuss (2022. gada rudenī ir plānoti 2 milj. vārdlietojumu) ir salīdzināms, piemēram, ar silēziešu valodas korpusu Polijā. MuLa ir ticis un tiek veidots manuāli, atlasot dažādu žanru tekstus un rūpējoties par žanru un tekstu daudzveidību, kā arī par valodas kvalitāti. Korpusa lielākais trūkums ir – tas nav lemmatizēts un nav morfoloģiski marķēts, t. i., kāda vārda formas nevar atrast ar vienu vaicājumu, bet jāmeklē katra forma atsevišķi.

MuLa korpuss jau gandrīz 10 gadus ir pieejams jebkuram interesentam. Tomēr aptaujas dati rāda, ka līdz šim tas ir ticis lietots ļoti maz (galvenokārt – zinātniskiem un profesionāliem nolūkiem), taču valodas runātāju kopienā par to ir interese. Iepazīstinot ar korpusu un tā iespējām skolotāju semināros (2022. gada februārī un jūnijā), Latgales kongresā (2022. gada aprīlī) u. c. pasākumos, tiek uzdoti jautājumi un ierosinājumi, kas liecina par nepieciešamību popularizēt korpusu vairāk un mērķtiecīgāk. Līdz šim korpusa lietotāji to ir izmantojuši vārda nozīmes noskaidrošanai un savas izpratnes padziļināšanai, aplūkojot vārda lietojuma kontekstu, kā arī pareizrakstības pārbaudei. Veicot lingvistiskos pētījumus, korpuss tiek izmantots latgaliešu valodas leksikas vai gramatikas izzināšanai, taču šādu pētnieku un pētījumu ir ļoti maz.

Respondentu atbildes ļauj ieskicēt arī MuLa korpusa trūkumus, kuru novēršana potenciāli varētu paplašināt korpusa lietotāju skaitu. Lietotājam draudzīgāks korpuss, kā rāda aptaujas dati, būtu, ja tiktu veikti gan saturiski, gan tehniski uzlabojumi. Satura ziņā korpuss būtu jāpaplašina kvantitatīvi (tas šobrīd jau tiek darīts), tāpat būtu jāuzlabo, jāvariē meklēšanas iespējas korpusā (piemēram, pēc vārda pamatformas (lemmas), pēc morfoloģiskām pazīmēm, pēc tekstu žanriem). Gan satura, gan tehnisko uzlabojumu ziņā jādomā par korpusa vietnes aprakstu un lietošanas pamācībām latviešu un/vai latgaliešu valodā. Jāpiedāvā vienkāršas, bet precīzas instrukcijas, piemēram, video formātā, kā rīkoties ar korpusu, kā arī jāsniedz pamata informācija par pašu korpusu.

Aptaujas dati atklāj tieši reģionālo valodu un līdz ar to arī korpusu izmantošanas specifiku: reģionālo valodu funkcionalitāte, salīdzinot ar nacionālajām valodām, ir šaurāka, tās vairāk

tiek lietotas neformālos kontekstos, diezgan nestabila ir arī šo valodu vieta oficiālajās izglītības sistēmās. Aptaujas dalībnieku otrajā grupā ("Zina un nelieto") no 43 respondentiem 70 % ir atzinuši, ka nezina, ko tieši varētu darīt ar latgaliešu valodas korpusu, lai gan citu valodu korpusus (piemēram, latviešu, angļu) viņi lieto. Starp respondentiem ir vērojama arī pedagoģiskā interese mācīt latgaliešu valodu, tomēr korpuss valodas apguves procesā gandrīz netiek izmantots.

Aptaujas dati un līdzšinējā saziņa ar dažādām mērķauditorijām (studentiem, skolēniem, skolotājiem, pētniekiem) ļauj izdarīt secinājumus, ka korpusa funkcionalitātes paplašināšana ir plānojama trīs virzienos: valodu izglītības, mediju satura bagātināšanas un pētniecības virzienā. Šī mērķa sasniegšanai ir nepieciešama korpusa popularizēšana un sabiedrības informēšana kopumā, bet vēl nozīmīgāka ir iesaistīšanās praktiskā darbā un dialogā ar latgaliešu valodas skolotājiem, ar studentiem – nākamajiem latviešu valodas skolotājiem un Latgales mediju darbiniekiem. Mērķauditorijai, kas nav valodnieki, ir nepieciešami konkrēti, praktiski piemēri, kas atklāj korpusu kā vērtīgu resursu valodas un satura iepazīšanai un apguvei. Viņiem jāpalīdz pamanīt tās iespējas, ko piedāvā korpuss, piemēram, iepazīt daudzveidīgu valodas lietojumu autentiskās situācijās (intervijās, emuāros u. tml.). Tāpat arī humanitāro un sociālo zinātņu (mediju) pētniekiem korpuss var kļūt par žanru, satura un valodas stila ziņā bagātu tekstu krātuvi individuālo pētījumu veikšanai. Savukārt valodniekiem darbā ar korpusu vēl aizvien aktuāls un neatrisināts ir lemmatizācijas un morfoloģiskās marķēšanas jautājums, kas īpaši nozīmīgi būtu, piemēram, vārdnīcu izstrādē.

Avoti (tekstā minētie valodu korpusi)

CorCenCC (velsiešu valodas korpuss). Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M. and Scannell, K. (2020). *CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh*. Cardiff University. http://doi.org/10.17035/d.2020.0119878310 [skatīts 02.07.2022.].

Corpas na Gàidhlig (skotu gēlu valodas korpuss). *Corpas na Gàidhlig, Digital Archive of Scottish Gaelic (DASG)*. University of Glasgow. https://dasg.ac.uk/corpus/ [skatīts 05.07.2022.].

DTK (lejas sorbu valodas korpuss). Serbski institut (n.d.). *Dolnoserbski tekstowy korpus*. https://www.niedersorbisch.de/korpus/ [skatīts 05.07.2022.].

ESLEMA (astūriešu valodas korpuss). *Eslema Corpus de la llingua asturiana*. https://eslema.it.uniovi.es/corpus/busqueda.html [skatīts 01.07.2022.].

ETC (basku valodas korpuss). Euskara Institutua / Basque Language Institute (2013-2021). *Egungo Testuen Corpusa (ETC)/ Corpus of Contemporary Basque (ETC)*. https://www.ehu.eus/etc/ [skatīts 05.07.2022.].

KŚM (silēziešu valodas korpuss). Kulik, Grzegorz (2018). *Korpus Ślōnskij Mowy*. https://korpus.silling.org/query?corpname=silesian [skatīts 05.07.2022.].

MuLa (mūsdienu latgaliešu tekstu korpuss). LU Matemātikas un informātikas institūts, Rēzeknes Tehnoloģiju akadēmija (2011–2013). http://www.korpuss.lv/id/MuLa [skatīts 01.07.2022.].

SIKOR (ziemeļsāmu valodas korpuss). Giellatekno - Saami Language Technology, UiT The Arctic University of Norway and The Divvun group at UiT The Arctic University of Norway, 2015, *SIKOR North Saami free corpus*, Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/100 [skatīts 01.07.2022.].

Tagged Icelandic Corpus (islandiešu valodas korpuss). Stofnun Árna Magnússonar í íslenskum fræðum / Árni Magnússon Institute for Icelandic Studies (2006-). *MÍM Mörkuð Íslensk Málheild / The Tagged Icelandic Corpus*. https://clarin.is/en/resources/mim/ [skatīts 05.07.2022.].

Bibliogrāfija

Bartels, Hauke (2012). Massnahmen zur Dokumentation des Niedersorbischen. *Slavia Occidentalis*, No. 69, pp. 7–22.

Bartels, Hauke (2020). Das niedersorbische Globalkorpus als Ziel einer ganzheitlichen Konzeption zum Aufbau von Textkorpora. *LĚTOPIS. Zeitschrift für sorbische Sprache, Geschichte und Kultur*, 2020, No. 2, pp. 3–44

Bernhard, Delphine; Ligozat, Anne-Laure; Bras, Myriam; Martin, Fanny; Vergez-Couret, Marianne; Erhart, Pascale; Sibille, Jean; Todirascu, Amalia; Boula de Mareüil, Philippe; Huck, Dominique (2021). Collecting and annotating corpora for three under-resourced languages of France: Methodological issues. *Language Documentation & Conservation*, No. 15, pp. 316–357 Available: https://scholarspace.manoa.hawaii.edu/handle/10125/74645 [accessed 13.11.2021.].

Briška, Anna (2013). *Mūsdienu latgaliešu rakstu valodas korpuss: izveide un izmantošanas iespējas*. Maģistra darbs. Nepublicēts. Rēzekne: Rēzeknes Augstskola.

Cox, Christopher (2011). Corpus linguistics and language documentation: challenges for collaboration. Newman, John, Baayen, Harald, Rice, Sally (eds.). *Corpus-based studies in language use, language learning, and language documentation*. Leiden: Brill, pp. 239–264.

Čermák, František (2011). The Case of The Czech National Corpus: Its Design and Development. Gozdz-Roszkowski, S. (ed.). *Explorations across Languages and Corpora*. Frankfurt: P. Lang, pp. 29–44. Available: https://nanopdf.com/download/1-general-remarks-esky-narodni-korpus_pdf [accessed 13.11.2021.].

Jakubíček, Miloš, Kilgarrif, Adam, Kovář, Vojtěch, Rychlý, Pavel, Suchomel, Vít (2013). The TenTen Corpus Family. *7th International Corpus Linguistics Conference CL*, pp. 125–127. Available: https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf [accessed 13.11.2021.].

Jauhiainen, Heidi, Jauhiainen, Tommi, Lindén, Krister (2020). Building web corpora for minority languages. *Proceedings of the 12th Web as Corpus Conference* (LREC 2020), Marseille, 11–16 May 2020, pp. 23–32. Available: http://www.lrec-conf.org/proceedings/lrec2020/workshops/WAC-II/index.html [accessed 13.11.2021.].

Knight, Dawn, Morris, Steve, Fitzpatrick, Tess (2021). *Corpus Design and Construction in Minoritised Language Contexts: The National Corpus of Contemporary Welsh*. London: Palgrave.

Knight, Dawn, Morris, Steve, Fitzpatrick, Tess, Rayson, Paul, Spasić, Irena, Thomas, Enlli Môn (2020). *The National Corpus of Contemporary Welsh: Project Report* | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. arXiv:2010.05542, October 2020.

Kulik, Grzegorz (2018). Korpus Ślōnskij Mōwy – prymiera. *Wachtyrz* 19.12.2018. Available: https://wachtyrz.eu/korpus-slonskij-mowy-prymiera/ [accessed 12.11.2021.].

Levāne-Petrova, Kristīne (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. *Baltistica* VIII Priedas 2012, 89.–98. lpp.

Martena, Sanita, Marten, Heiko, Šuplinska, Ilga (2022). *Latgalian. The Latgalian language in education in Latvia.* 2nd Edition. The Netherlands: Mercator, Fryske Akademy.

McEnery, Tony, Hardie, Andrew (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, Tony, Ostler, Nicholas (2000). A new agenda for corpus linguistics – working with all of the world's languages. *Literary and Linguistic Computing*, No. 15(4), pp. 403–420.

Millour, Alice, Fort, Karën (2020). Text Corpora and the Challenge of Newly Written Languages. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, Marseille: European Language Resources association, pp. 111–120. Available: https://aclanthology.org/2020.sltu-1.15/ [accessed 12.11.2021.].

Ostler, Nicholas (2009). Corpora of less studied languages. *Corpus linguistics: an international handbook*. Lüdeling Anke, Kytö, Merja (eds.). Berlin, New York: Walter de Gruyter, pp. 457–483.

Pomierska, Justyna, Stanulewicz, Danuta (2019). Tworzenie korpusu języka kaszubskiego – dostępne źródła zdigitalizowanych tekstów. *Acta Cassubiana*, No. 21, pp. 159–178

Sampson, Geoffrey, McCarthy, Diana (2007). *Corpus linguistics: Readings in a widening discipline*. London: Continuum, pp. 1–2.

Scannel, Kevin (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental*, No, 4, pp. 5–15.

Šuplinska, Ilga (zyn. red.) (2012). *Latgolys lingvoteritorialuo vuordineica. Lingvoterritorial Dictionary of Latgale*. II. Rēzekne: Rēzeknis Augstškola.

UNESCO (2017). *Endangered languages. A methodology for assessing language vitality and endangerment.* Available: http://www.unesco.org/new/en/culture/themes/endangered-languages/language-vitality/#topPage [accessed 28.10.2021.].

Viejo, X.; Sauri, R; Neira, A. (2008). Eslema. Towards a corpus for Asturian. *Collaboration: Interoperability between people in the creation of language resources for less-resourced languages. A SALTMI workshop. LREC 2008*. Marrakesh.

Summary

The Corpus of Latgalian in the Context of Other Lesser Used Languages of Europe: Characterization, Usage and Potential

Sanita Martena, Anna Briška, Nicole Nau

Keywords: Latgalian, corpus, corpus of Latgalian, MuLa, regional languages, lesser-used languages of Europe

The article reports on the current process of completing and further developing the corpus of contemporary written Latgalian (MuLa). It gives an overview over the sources and the principles of compiling the corpus and compares it with corpora of other lesser used languages of Europe. In addition, it analyses the profile of current and potential users, users' experience with the corpus so far and their wishes for the future. The first version of the corpus MuLa, which has been publicly available since 2012, contains 1 million running words. It will be enlarged to 2 million words in the extended and corrected version prepared 2020–2022.

Corpora of lesser used languages play an important role in the documentation and development of the language. They are also a valuable resource for the preparation of linguistic tools and teaching materials. While MuLa does not have either the size or the functionality of corpora of such well-cared-for European regional languages as Basque, Welsh, or Sami, the fact that a corpus exists and is being further developed puts Latgalian in a better position than some other regional languages in Europe. Due to a shortage of financial and human resources, corpora of lesser used languages are often compiled from sources automatically gathered from the Internet. MuLa, in turn, is still compiled manually, which allows higher control of register diversity and balance, as well as linguistic quality. The corpus is not tagged, but it is accessible to everyone.

Data about the usage and users of MuLa since 2012 have been collected with an online questionnaire answered by 214 respondents. The study shows that the corpus has been used very little, mostly by researchers and a few other professionals. On the other hand, many respondents expressed a potential interest and ideas about potential uses of the corpus for learning about Latgalian as well as further developing their linguistic skills. In order for the corpus to be used more broadly, promotion and the spread of information within society is indispensable. Still more important is cooperation and a constant dialogue with teachers and university students.