10.35539/LTNC.2022.0047.I.S.I.A.R.D.A.V.312.327

CLARIN valodas resursu un rīku pētniecības infrastruktūra humanitārajām un sociālajām zinātnēm

Inguna Skadiņa, Ilze Auziņa,
Roberts Darģis, Arnis Voitkāns

CLARIN Latvija nacionālā mezglpunkta izveidi un uzturēšanu atbalsta Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi" projekts "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001) un Eiropas Reģionālās attīstības fonda projekts "Latvijas Universitāte un institūti Eiropas pētniecības telpā – ekselence, aktivitāte, mobilitāte, kapacitāte" (Nr. 1.1.1.5/18/I/016).

Atslēgvārdi: pētniecības infrastruktūras, digitālās humanitārās zinātnes, valodas resursi, latviešu valoda, zināšanu infrastruktūra, atvērtā zinātne, FAIR

Ievads

Lai arī digitālu valodas resursu un rīku izveide uzsākta drīz pēc pirmo datoru parādīšanās, nepieciešamība veidot digitālas valodas resursu krātuves, kas būtu pieejamas ikvienam pētniekam neatkarīgi no to atrašanās vietas, aktualizējas gadsimtu mijā līdz ar straujo interneta uzplaukumu un digitālos avotos balstītas pētniecības attīstību.

Iemesli, kāpēc nepieciešams veidot valodas resursu krātuves un pētniecības infrastruktūras, ir vairāki. No vienas puses, digitāla valodas resursu un rīku pētniecības infrastruktūra ir pamatnosacījums digitālo humanitāro zinātņu pētījumiem. No otras puses, uzkrājoties arvien vairāk digitālajiem valodas resursiem, rodas nepieciešamība šos pētniecības datus padarīt pieejamus plašam pētnieku lokam un nodrošināt to saglabāšanu ilgtermiņā (t. i., pasargāt no pazušanas pēc kāda konkrēta projekta vai iniciatīvas beigām). Visbeidzot, jāpiemin otrais Gūtenberga efekts (valodas, kas nebūs vai būs nepietiekami pārstāvētas digitālajā vidē, ir apdraudētas un var ar laiku izzust), kas kā iespējamais apdraudējums valodas pastāvēšanai tika aktualizēts jau 20. gadsimta 90. gados.

Apzinoties nepieciešamību veidot valodas resursu un rīku krātuves, novērst digitālo valodas resursu sadrumstalotību un veicināt to saglabāšanu ilgtermiņā, 2006. gadā vairāku valstu pētnieku grupa nāca klajā ar iniciatīvu veidot Vienotu valodas resursu un tehnoloģiju infrastruktūru (CLARIN – *Common Language Resources and Technology Infrastructure*; Váradi et al. 2008). Šī iniciatīva aizrāva daudzus Eiropas zinātniekus – apvienojot spēkus, tika veidoti valodas resursu un rīku pārskati, izstrādāta tehniskā specifikācija Eiropas līmeņa repozitoriju sistēmas izveidei, apzināti un pētīti licencēšanas modeļi, kā arī definēts Eiropas pētnieciskās infrastruktūras juridiskais pamats un pārvaldības modeļi. 2012. gada 29. februārī, pēc tam, kad Eiropas Savienība (ES) bija izveidojusi Eiropas pētniecības infrastruktūru (ERIC) juridisko pamatu, deviņas ES valstis nodibināja CLARIN ERIC. Latvija CLARIN ERIC iestājās 2016. gada vasarā (Skadiņa et al. 2020).

Pašlaik CLARIN apvieno 25 valstis: 22 valstis ir pilntiesīgas ERIC locekles, bet trīs valstīm ir novērotājas statuss. CLARIN mērķi – vienotu piekļuvi valodas resursiem un rīkiem – infrastruktūra realizē ar vairāk nekā 60 centriem, kuri apvienoti virtuālā tīklā un ļauj zinātniekiem, izmantojot vienotu pierakstīšanos, piekļūt valodas resursiem un rīkiem neatkarīgi no viņu atrašanās vietas. CLARIN virtuālajā

valodas resursu krātuvē (VLO – *Virtual Language Observatory*)[1] pašlaik reģistrēti vairāk nekā 1,25 miljoni valodas resursu un rīku.

CLARIN piedāvā ne tikai ilgtermiņa risinājumus un tehnoloģijas digitālo valodas datu un rīku izvietošanai, bet arī konsultē un atbalsta pētniekus, kas veic valodas datos balstītus pētījumus.

Šī raksta mērķis ir iepazīstināt digitālo humanitāro zinātņu pētniekus ar CLARIN ERIC un CLARIN Latvijas mezglpunkta (CLARIN-LV) piedāvātajām iespējām, sniedzot ieskatu gan pētniecības infrastruktūrā apkopotajos valodas resursos Latvijā un Eiropā, gan dažādu valodas resursu un rīku kopās (*Resource Families*[2]). Rakstā arī sniegti praktiski ieteikumi valodas resursu deponēšanai un licencēšanai, kā arī repozitorijā ievietoto valodas datu citēšanai.

CLARIN ERIC

CLARIN ir digitāla infrastruktūra, kas piedāvā datus, rīkus un pakalpojumus valodas resursos balstītas pētniecības atbalstam. CLARIN piedāvā:

- ērti lietojamus valodas resursus un rīkus (skat. nodaļas "Vienotā valodas resursu un rīku krātuve" un "Valodas resursu saimes") – piedāvājot rīkus valodas datu izpētei, analīzei un saglabāšanai, CLARIN nodrošina infrastruktūru pētījumiem, kuru pamatā ir digitālie valodu resursi,
- zināšanu infrastruktūru (skat. nodaļu "CLARIN-LV zināšanu infrastruktūra") – zināšanu infrastruktūras mērķis ir nodrošināt, lai CLARIN konsorcijā pieejamās zināšanas būtu sasniedzamas,
- finansējumu (skat. nodaļu "CLARIN atbalsts sadarbībai valodas resursu izveidē") – CLARIN veicina pētniecību dažādās jomās, regulāri izsludinot konkursus projektiem, kas sekmē CLARIN infrastruktūras resursu izmantošanu un paplašināšanu.

CLARIN infrastruktūras un centru darbību nodrošina CLARIN dalībvalstu nacionālie konsorciji. Nacionālos konsorcijus parasti veido universitātes, institūti, bibliotēkas un publiskie arhīvi. Katrā dalībvalstī ir vismaz viens centrs (repozitorijs), kurā tiek apkopoti valodas resursi un rīki.

1 CLARIN *Virtual Language Observatory* (VLO), https://vlo.clarin.eu/

2 CLARIN *Resource families*, https://www.clarin.eu/resource-families

Vienotā valodas resursu un rīku krātuve

CLARIN virtuālā valodas resursu krātuve ir katalogs, kurā apkopoti CLARIN dalībvalstu centru dokumentēto valodas resursu un rīku (VRR) metadati. VLO nodrošina plašas meklēšanas iespējas, ļaujot atrast nepieciešamo starp vairāk nekā 1,25 miljoniem valodas datu ierakstu.

Kad nepieciešamais VRR ir atrasts, lietotājs tiek novirzīts uz to CLARIN repozitoriju, kurā tas noglabāts. Repozitorijs nodrošina piekļuvi VRR atbilstoši tā licencēšanas nosacījumiem, t. i., brīvpieejas resursi ir lejupielādējami vai pārlūkojami, bet ar licenci aizsargātiem resursiem var piekļūt, izmantojot vienotu pierakstīšanos.

VLO ir apkopota informācija ne tikai par CLARIN centros reģistrētajiem valodas resursiem un rīkiem, bet arī par citu iniciatīvu valodas resursiem. Piemēram, nozīmīgu daļu latviešu valodas ierakstu veido vairākas Europeana[3] Latvijas Nacionālās bibliotēkas kolekcijas (*Europeana – National Library of Latvia*). Pašlaik VLO apkopota informācija par 2555 latviešu valodas resursiem un rīkiem no 32 dažādiem repozitorijiem un kolekcijām[4]. VLO saturs tiek atjaunots divas reizes nedēļā.

Valodas resursu kopas

Digitālo humanitāro, humanitāro un sociālo zinātņu pētniekiem noderīgi ir CLARIN resursu kopu (*Resource Families*) iniciatīvas sagatavotie pārskati par CLARIN infrastruktūrā pieejamajām valodas resursu un rīku saimēm. Pašlaik ir sagatavoti pārskati par 12 korpusu saimēm (datorizētas saziņas korpusi, akadēmisko tekstu korpusi, vēsturisko tekstu korpusi, valodas apguvēju korpusi, literārie korpusi, manuāli marķētie korpusi, multimodālie korpusi, laikrakstu korpusi, paralēlie korpusi, parlamentāro debašu korpusi, atsauces korpusi (*Reference corpora*), runas korpusi), piecām leksisko resursu saimēm un četrām rīku saimēm (normalizācijas rīki, nosaukto entitāšu atpazīšanas rīki, morfoloģiskās analīzes un lemmatizācijas rīki, noskaņojuma analīzes rīki).

CLARIN atbalsts sadarbībai valodas resursu izveidē

CLARIN ERIC finansiāli atbalsta dažādas iniciatīvas, kuru mērķis ir sekmēt sadarbību CLARIN valstu konsorciju starpā CLARIN statūtos minēto uzdevumu veikšanai.

Viena no iniciatīvām ir jaunu valodas resursu izveide, kas sekmīgi realizēta projektā *ParlaMint* (Erjavec et al. 2022). Projekta mērķis bija izveidot daudzvalodu parlamentāro debašu korpusu ar standartizētiem metadatiem, vienotu datu un marķējuma formātu. Korpusā (Erjavec et al. 2021) iekļauti 17 valstu parlamentu dati, tajā skaitā arī Latvijas Republikas

3 https://www.europeana.eu/en/item/97/418759

4 VLO skatīts 05.05.2022.

Saeimas sēžu stenogrammas. Visi dati ir automātiski morfoloģiski un sintaktiski marķēti, pievienots arī nosaukto entitāšu marķējums. Dati ir sagatavoti TEI (*Text Encoding Initiative*) formātā un ir pieejami gan meklēšanai *noSketchEngine* platformā[5], gan lejupielādei no CLARIN.SI repozitorija ar lingvistisko marķējumu[6] un bez tā[7].

CLARIN-LV valodas resursu un rīku krātuve kā pētniecības datu repozitorijs

CLARIN-LV repozitorijs ir viens no vairāk nekā 60 virtuālajiem valodas resursu centriem. Tā mērķis ir sekmēt digitālo humanitāro zinātņu pētījumus, apkopojot latviešu un citu Latvijā pētītu valodu resursus un rīkus, nodrošinot VRR pieejamību ilgtermiņā un rūpējoties par to izmantošanu atbilstoši licences nosacījumiem.

Valodas resursi

Valodas resursu uzkrāšana un dokumentēšana CLARIN-LV repozitorijā tika sākta 2020. gada vasarā, pēc CLARIN-LV repozitorija izveides un stabilizēšanas un sākotnējās valodas resursu un rīku apzināšanas. Pašlaik (t. i., 2022. gada maijā) repozitorijā iekļauti 35 valodas resursi – 22 korpusi, 11 leksikoni un 2 rīki. Vairāku valodas resursu metadati ir sagatavošanā vai iesniegti izvērtēšanai. Repozitorijā ir gan monolingvālas, gan multilingvālas datu kopas. Lai arī lielākā daļa pašlaik CLARIN-LV repozitorijā iekļauto valodas resursu veidoti LU MII Mākslīgā intelekta laboratorijā, krātuvei pievienoti pirmie citu institūciju – LU Literatūras, folkloras un mākslas institūta, LU Latviešu valodas institūta, Rēzeknes Tehnoloģiju akadēmijas, Ventspils Augstskolas un Latviešu valodas aģentūras – izstrādātie resursi. Šobrīd visi CLARINLV repozitorijā iekļautie valodas resursi un rīki ir brīvpieejas – tos var izmantot pētniecībā caur pārlūkprogrammu, korpusu pārvaldības rīku *noSketchEngine* vai kādu citu platformu, vairāk nekā 10 datu kopas ir lejupielādējamas un brīvi izmantojamas atbilstoši licencei.

Citēšana

Būtiski pieaugot digitālo datu daudzumam un to lietojumam, kvalitatīvu datu kopu, it īpaši valodas resursu, izveide ir nozīmīgs pētniecības rezultāts, kuru, līdzīgi kā publikāciju, ir

5 http://www.clarin.si/noske/

6 https://www.clarin.si/repository/xmlui/handle/11356/1432

7 https://www.clarin.si/repository/xmlui/handle/11356/1431

1. attēls. CLARIN-LV sagatavotās "Pandēmijas dienasgrāmatu" datu kopas (Reinsone et al. 2021) atsauce tekstuālā un *bibtex* formātā. Attēlā redzams arī ieraksta unikālais identifikators http://hdl.handle.net/20.500.12574/48.

nepieciešams citēt. Datu pieejamību un to citēšanu sekmē atvērtās zinātnes (*Open Science*) un FAIR principu[8] ieviešana Eiropā un Latvijā.

Iespēju citēt datu kopas piedāvā arī CLARIN repozitoriji. Katrai repozitorijā dokumentētai datu kopai vai rīkam tiek piekārtots unikāls identifikators (PID – *persistent identificator*), kas ļauj to viennozīmīgi identificēt un citēt (skat. 1. attēlu). Citēšanai nepieciešamo atsauci repozitorijs ģenerē automātiski no metadatiem, kurus sagatavojis valodas resursa iesniedzējs (deponētājs). Tas ne tikai atvieglo citēšanas procesu resursa izmantotājam, bet arī palīdz iegūt informāciju, cik pētnieku šo resursu ir citējuši (izmantojuši), kas bieži ir būtiski finansējuma iegūšanai resursa uzturēšanai un attīstīšanai. Repozitorijs arī uzskaita resursa skatījumu un lejupielāžu skaitu (skat. nodaļu "Populārākie valodas resursi").

Automātiski veidotie citējumi, kā arī valodas resursu ieraksti CLARIN katalogā sekmē datu kopas pamanāmību un lietojumu ne tikai Latvijā, bet arī Eiropā un citur pasaulē. Piemēram, *Google Scholar* atsauces uz CLARIN-LV un citos CLARIN repozitorijos apkopotajiem valodas resursiem iekļauj autoru profilos. *Google Scholar* atsauce uz Universālo atkarību datu kopu (*Universal Dependencies 2.5*), kas iekļauta LINDAT/CLARIAH-CZ repozitorijā un kurā ietverts arī sintaktiski marķēts latviešu valodas korpuss, ietver gan repozitorija nosaukumu, gan unikālo identifikatoru, gan publicēšanas gadu un citējumu skaitu (skat. 2. attēlu).

8 FAIR – *Findability, Accessibility, Interoperability, and Reuse of digital assets*, skat. https://www.go-fair.org/fair-principles/

2. attēls. *Google Scholar* atsauce uz Universālo atkarību datu kopām.

Deponēšana

Viens no CLARIN pētniecības infrastruktūras pamatuzdevumiem ir valodas resursu un rīku uzglabāšana ilgtermiņā. Atbilstoši CLARIN ERIC statūtiem katrā dalībvalstī jābūt vismaz vienam sertificētam datu centram, kas tehniski var nodrošināt drošu datu uzglabāšanu un ar to uzturēšanu saistītus pakalpojumus.

CLARIN repozitorijos tiek uzglabāti valodas resursi, kurus veidojuši zinātnieki dažādu projektu laikā. Svarīgi, ka šie resursi ir uzticami, t. i., tos pēc projekta beigām var izmantot citi pētnieki. Tāpēc CLARIN (un CLARIN-LV kā dalībnieks) piedāvā zinātniekiem valodas resursus deponēt, t. i., nodot ilgtermiņa uzglabāšanā.

Datu iesniegšanu parasti veic zinātnieks (kāds no autoriem) vai zinātniskās institūcijas nozīmēts datu pārvaldnieks. Lai iesniegums nebūtu anonīms un dati būtu uzticami, iesniedzējam jāpieslēdzas CLARIN repozitorijam, izmantojot savu akadēmiskā lietotāja kontu (skat. arī nodaļu "Licencēšana un vienotā pierakstīšanās"). Pēc pieslēgšanās CLARIN repozitorijam iesniedzējs sagatavo valodas resursa vai rīka metadatus (informāciju par valodas resursu vai rīku), aprakstot iesniedzamo datu kopu. Viens no svarīgākajiem metadatu laukiem ir VRR autori (veidotāji), šis lauks ļauj atsaukties uz datu kopu un tās autoriem līdzīgi kā publikācijas gadījumā. Tāpat iespējams norādīt publikāciju, kurā datu kopa vai rīks aprakstīts. Ne mazāk būtiska informācija ir finansējuma avots. Norādītie metadati var tikt izmantoti meklēšanai un valodas resursu atlasīšanai.

3. attēls. Latviešu valodas sintaktiski marķētu tekstu korpusu (*Latvian Treebank*) versijas CLARIN-LV repozitorijā.

Pēc metadatu sagatavošanas deponējamo valodas resursu vai rīku iespējams augšupielādēt CLARIN repozitorijā, norādot tā izmantošanas nosacījumus (licenci). Iesniedzot datus CLARIN repozitorijam, iesniedzējs arī dod tiesības CLARIN-LV repozitorijam izplatīt datus atbilstoši licences nosacījumiem[9].

Sagatavotie metadati un pievienotā datu kopa tiek publicēta tikai pēc tam, kad to pārskatījis CLARIN-LV repozitorija datu pārvaldnieks, kura uzdevums ir pārliecināties par iesniegto metadatu un pievienoto datu kvalitāti un ticamību.

Laika gaitā datu kopas var tikt mainītas un papildinātas. CLARIN repozitoriji neatbalsta izmaiņas iesniegtajos datos (izņēmums var būt kļūdas un neprecizitātes datu kopas metadatos), bet piedāvā iespēju veidot vairākas datu kopas versijas. Tas ļauj zinātniekiem izvēlēties pētījumam piemērotāko versiju. Piemēram, ja nepieciešams kādu pētījumu atkārtot vai veikt rezultātu salīdzināšanu, piemērotāka var izrādīties kāda vecāka datu kopas versija (skat. 3. attēlu).

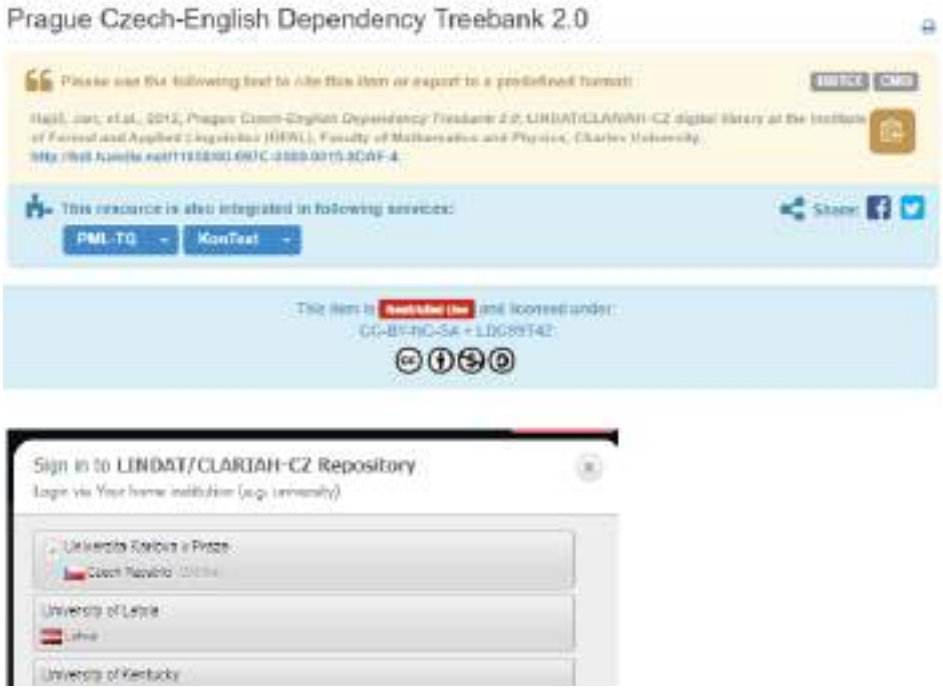
Tā kā valodas resursu krātuves un deponēšana ir jaunums Latvijā, 2021. gada sākumā tika rīkots praktiskais seminārs par valodas resursu deponēšanu, kura dalībnieki praksē iepazinās ar deponēšanas gaitu, kā arī tika izveidota deponēšanas instrukcija[10], kuru var izmantot ikviens pētnieks, kas vēlas iesniegt pētniecības datu kopas ilgtermiņa uzglabāšanai CLARIN-LV repozitorijā.

Licencēšana un vienotā pierakstīšanās

Līdz ar atvērtās zinātnes kustības izplatību Latvijā un pasaulē arvien aktuālāka kļūst datu kopu autorība, to licencēšana un tajā iekļauto datu autortiesību un privātuma jautājumi. Lai šos jautājumus risinātu, CLARIN ir izveidota Juridisko jautājumu komiteja, kas izstrādā vadlīnijas un ieteikumus ar datu pārvaldību saistītos jautājumos un seko līdzi aktualitātēm.

9 Deponēšanas process detalizēti aprakstīts repozitorija lapā: https://repository.clarin.lv/repository/xmlui/page/deposit

10 https://www.clarin.lv/attachments/CLARIN%202020_resursu_iesniegsana.pdf



4. attēls. Vienotās pierakstīšanās piemērs piekļuvei LUNDAT/CLARIAH-CZ repozitorija *Prague Czech-English Dependency Treebank 2.0*.

Katrai CLARIN datu kopai, kas iesniegta repozitorijā, tiek norādīta licence, ar kādu zinātnieki un interesenti var to izmantot. Licenci datu kopai izvēlas datu kopas iesniedzējs (parasti zinātnieks vai zinātnieku grupa, kas šo valodas resursu veidojusi). Tā kā zinātnieki ne vienmēr pārzina dažādu licenču specifiku, ir izveidots CLARIN licenču ieteicējs, kas, lietotājam atbildot uz vairākiem jautājumiem, palīdz izvēlēties piemērotāko licenci.

Tā kā CLARIN atbalsta atvērto zinātni[11], tad datu kopām ieteikts izmantot *Creative Commons* licences. Šai licenču grupai pievienotās pazīmes apraksta licencēšanas nosacījumus. Biežāk lietotās pazīmes (atribūti) ir: BY – prasa norādīt datu kopas autorību un ļauj to brīvi izmantot, SA – nosaka, ka atvasinātās datu kopas jāizplata ar tādu pašu licenci kā oriģināls, NC – liedz komerciālu lietojumu, ND – liedz veidot atvasinātus un adaptētus darbus. Vairums latviešu valodas resursu un rīku, kuri kataloģizēti CLARIN VLO, ir publiski pieejami.

11 https://www.clarin.eu/eosc

Item/Handle	Number of views
Tēzaurs.lv 2020 (Spektors, Andrejs et al) (20.500.12574/5)	185 [+bots: 188]
LVBERT - Latvian BERT (Znotiņš, Artūrs) (20.500.12574/43)	125 [+bots: 127]
Latvian Treebank v2.5 (Rituma, Laura et al) (20.500.12574/51)	117 [+bots: 117]
Latvian AMR Sembank (Znotiņš, Artūrs et al) (20.500.12574/40)	112 [+bots: 113]
LUIS: data collection for task oriented dialogue system creation (Gruzīte, Nešpore-Bērzkalne et al)...	104 [+bots: 104]

5. attēls. Pieci visvairāk skatītie valodas resursi
2022. gada aprīlī.

Lai gan *Creative Commons* licence ir ļoti populāra, jo ļauj datus brīvi izmantot gan pētniecībai, gan vairumā gadījumu arī jaunu darbu izveidei, ne vienmēr ar datu kopām var tik brīvi rīkoties. Lai kontrolētu šādu datu kopu lietojumu, CLARIN repozitoriji izmanto vienotu pierakstīšanos ierobežota lietojuma datu kopu lietojuma kontrolei. Vienotā pierakstīšanās (Latvijā tiek realizēta, izmantojot LAIFE identitāšu federāciju) nodrošina gan datu kopu izmantošanu atbilstoši licencei (to lietotājs digitāli apstiprina pirms lietošanas uzsākšanas), gan arī iespēju izmantot ierobežotas piekļuves datu kopas kādā no CLARIN ERIC repozitorijiem. Piemēram, LINDAT CLARIN repozitorijā Prāgā ir apkopoti daudzu valodu, t. sk. latviešu valodas, sintaktiski marķēti korpusi. Vairums (84 no 109) resursu ir izmantojami ar CC-BY-NC-SA licenci, bet, piemēram, Prāgas čehu-angļu atkarību sintaktiski marķētā korpusa 2.0 versija[12] tiek izplatīta ar papildu ierobežojumiem, tāpēc pirms lejupielādes ar vienoto pierakstīšanos tiek identificēts datu lietotājs (skat. 4. attēlu).

Populārākie valodas resursi un rīki

Valodas resursu uzkrāšana un dokumentēšana CLARIN-LV repozitorijā tika sākta 2020. gada vasarā, vispirms iekļaujot LU MII Mākslīgā intelekta laboratorijas populārākos valodas resursus. Pēc praktiskā semināra 2021. gada ziemā valodas resursus sāka reģistrēt arī citas zinātniskās organizācijas. 2021. gada vasaras sākumā repozitorijs tika papildināts ar populārākajiem valodas korpusiem, bet vēlāk ar leksikoniem.

Trīs visvairāk skatītie valodas resursi ir *Tēzaurs.lv* (Spektors et al. 2019), "Līdzsvarotais latviešu valodas korpuss" (Levāne-Petrova un Darģis 2018) un "Sintaktiski marķētais latviešu valodas korpuss v2.5" (Rituma et al. 2020), kas visi ir reģistrēti repozitorija izveides pirmsākumos.

12 http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

Tā kā dažādi valodas resursi un rīki repozitorijā iekļauti atšķirīgā laikā, precīzāku informāciju par populārākajiem resursiem sniedz statistika mēneša griezumā (skat. 5. attēlu). Redzams, ka lietotājus interesē gan valodas korpusi, gan leksikoni. Tāpat vērojama interese par mašīnlasāmām datu kopām un modeļiem. Tomēr interese par valodas apstrādes rīkiem ir mazāka, ko varētu skaidrot ar to, ka CLARIN lietotāji galvenokārt ir humanitāro zinātņu pētnieki, kurus vairāk interesē valodas resursi, kas ir ērti izmantojami, nevis iespēja veidot savus risinājumus, izmantojot repozitorijā iekļautos datus.

Analizējot šo statistiku, ir būtiski atcerēties, ka tā norāda tikai valodas resursa pamanāmību CLARIN-LV repozitorijā, bet neatspoguļo kopējo datu kopas lietojumu (piem., korpusa lietojumu, izmantojot *SketchEngine* platformu).

CLARIN-LV zināšanu infrastruktūra

CLARIN pētniecības infrastruktūra nav tikai valodas resursu un rīku repozitorijs, bet arī zināšanas, ko uzkrājuši daudzie CLARIN centri un nacionālie konsorciji. Zināšanu apmaiņa ietver darbības, kas saistītas ar lietotāju iesaistīšanu un izglītošanu, piemēram, mācības (darbsemināri, lekcijas), diskusijas un zināšanu izplatīšanas pasākumi (konferences, semināri), kā arī ikdienas atbalsts, izmantojot zināšanu centrus.

Pakalpojumi humanitāro zinātņu pētnieku atbalstam – lietotāju iesaiste, izglītība un konsultācijas

Lai iepazīstinātu Latvijas humanitāro un sociālo zinātņu pārstāvjus ar CLARIN resursiem un rīkiem, mācītu, kā izmantot latviešu valodas resursus, pēdējo gadu laikā ir organizēti vairāki semināri.

2018. gada sākumā seminārs "Valodas resursi un rīki digitālajām humanitārajām zinātnēm" tika organizēts, lai demonstrētu LU MII Mākslīgā intelekta laboratorijā izstrādātos valodas rīkus un resursus. Seminārā plašāka publika pirmo reizi pēc Latvijas pievienošanās CLARIN ERIC tika iepazīstināta ar infrastruktūras nacionālajiem un starptautiskajiem mērķiem, un semināra dalībnieki tika aicināti aktīvi piedalīties CLARIN ekspertu tīkla izveidošanā Latvijā. Seminārā pulcējās humanitāro un sociālo zinātņu pārstāvji, tostarp filologi, žurnālisti, politologi, tulki, bibliotekāri, vēsturnieki.

Papildus ir organizēti vairāki praktiski semināri, kuros pētnieki iepazīstināti ar vairākiem latviešu valodas korpusiem – "Līdzsvaroto mūsdienu latviešu valodas tekstu korpusu" (LVK2018), "Latviešu valodas sintaktiski marķēto korpusu", "Latviešu valodas apguvēju korpusu" (LaVA).

Informācija par CLARIN resursiem un rīkiem ir integrēta Latvijas Universitātes Humanitāro zinātņu fakultātes maģistra studiju programmas kursā "Ievads datorlingvistikā". Kursa laikā studenti tiek iepazīstināti ar pētniecības infrastruktūru, tiek aicināti izpētīt CLARIN repozitoriju – atrast saviem pētījumiem noderīgus resursus un rīkus – un pastāstīt

par saviem atklājumiem seminārā. Savukārt topošajiem skolotājiem Latvijas Universitātes Pedagoģijas, psiholoģijas un mākslas fakultātē tiek piedāvāts kurss "Valodas korpuss izglītības procesā". Par valodas resursiem ir stāstīts arī demonstrāciju seminārā "Digitālie resursi novadu savdabības atklāšanai, izglītībai un kultūrai".

Zināšanu infrastruktūra

Lai palīdzētu zinātniekiem neapjukt milzīgajā CLARIN valodas resursu daudzveidībā un sekmētu zināšanu apmaiņu un pārnesi, CLARIN ir izveidoti vairāk nekā 20 zināšanu centri[13] (*knowledge centres*, *K-Centres*, K-centri). K-centri konsultē ne tikai par kādu konkrētu valodu vai modalitāti (runātā valoda, zīmju valoda u. c.), bet arī par valodas apstrādes rīkiem, dažādiem datu kopu veidiem u. c. ar valodas pētniecību saistītiem jautājumiem.

CLARIN-LV kopā ar Helsinku Universitāti (FIN-CLARIN), Trumses Universitāti (CLARINO) un Vītauta Dižā Universitāti (CLARIN-LT) apvienojušies zināšanu centrā SAFMORIL[14], kas sniedz atbalstu digitālo humanitāro zinātņu pētniekiem, valodniekiem un valodu tehnoloģiju izstrādātājiem, kuri analizē un apstrādā morfoloģiski bagātu valodu datus. Pašlaik CLARIN-LV ar partneriem gatavo pieteikumu leksikogrāfijas zināšanu centra ELEXIS izveidei.

CLARIN-LV sniedz konsultācijas par latviešu valodas resursiem un rīkiem. Līdz šim biežāk uzdotie jautājumi saistīti ar valodas korpusiem, tajos lietoto marķējumu un meklēšanas iespējām. Tāpat vairākiem pētniecības centriem sniegti padomi un praktiska palīdzība valodas korpusu izveidē.

Veiksmīga sadarbība izveidojusies ar Latvijas Nacionālo bibliotēku, palīdzot uzstādīt *noSketchEngine* platformu un apmācot resursu sagatavošanā un ievietošanā. Gan LNB, gan citas zinātniskās organizācijas kā servisu aktīvi izmanto latviešu valodas apstrādes rīkkopu NLP-PIPE[15] (Znotiņš 2015), kas piedāvā tekstu morfoloģisko marķēšanu, sintaktisko parsēšanu un nosaukto entitāšu atpazīšanu.

Iedvesmojoties no CLARIN vienotā (*federated*) satura meklēšanas servisa CLARIN-FCS[16], kas ļauj meklēt valodas korpusos neatkarīgi no tā darbināšanas vietas, CLARIN-LV ir izveidojis vienotu meklēšanu Latvijas korpusos *noSketchEngine* platformā[17] (skat. 6. attēlu). Šobrīd meklēšana tiek veikta tikai LU MII korpusos, bet tiek strādāts arī pie LNB korpusu pievienošanas meklēšanas sistēmai.

13 https://www.clarin.eu/content/knowledge-centres

14 https://www.clarin.eu/blog/tour-de-clarin-clarin-knowledge-centre-systems-and-frameworks-morphologically-rich-languages

15 http://nlp.ailab.lv/

16 https://www.clarin.eu/content/federated-content-search-clarin-fcs

17 http://www.korpuss.lv/search

6. attēls. Vienotā meklēšana latviešu valodas korpusos.

CLARIN-LV nākotnes uzdevumi un loma Latvijas digitālo humanitāro zinātņu attīstībā

Pēdējo piecu gadu laikā CLARIN-LV ir kļuvis par stabilu un atpazīstamu valodas resursu krātuvi Latvijā un Eiropā. CLARIN-LV ļauj zinātniekiem izmantot repozitorijā reģistrētos resursus, sniedz konsultācijas par to lietojumu. CLARIN-LV mērķis ir kļūt par tādu starptautiski integrētu nacionālā līmeņa valodas resursu un rīku infrastruktūru, kas pilnvērtīgi nodrošina digitālo humanitāro zinātņu, datorlingvistikas un valodu tehnoloģiju pētnieku kopienu vajadzības un sekmē pilnvērtīgu latviešu valodas dzīvi digitālajā laikmetā.

Šī mērķa sasniegšanai nepieciešams turpināt darbu pie repozitorija satura, papildinot to ar Latvijā pētniecības projektos tapušiem valodas resursiem un rīkiem. Svarīga ir arī valodas resursu un rīku pielāgošana pētnieku vajadzībām. Jau pašlaik citos CLARIN centros, piemēram, Čehijā, Islandē, Polijā un Slovēnijā, CLARIN nav tikai repozitorijs un kompetences centrs, bet arī digitālajām humanitārajām zinātnēm pielāgoti pakalpojumi un vide pētniecībai. Pirmie pielāgotie risinājumi tiek veidoti Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi" projektā "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001), kur latviešu valodas rīkkopa NLP-PIPE tiek pielāgota izmantošanai Latvijas Nacionālās bibliotēkas tekstu apstrādē.

Nākotnē ir svarīgi turpināt ciešu DHZ pārstāvju, datorlingvistu un datorzinātnieku sadarbību jaunos pētniecības projektos un Valsts pētījumu programmās, tādējādi nodrošinot gan DHZ tālāku uzplaukumu, gan sekmējot valodas resursu izveidi un pilnveidi. Nepieciešams sistemātisks un mērķtiecīgs valsts atbalsts ilgtspējīgiem pētījumiem valodas resursu un rīku izveidē un uzturēšanā, kas nodrošinātu valodu līdztiesību ikdienas lietojumā un digitālajā vidē.

Vēl arvien novērojama būtiska zināšanu un izpratnes plaisa starp valodas resursu un rīku veidotājiem un to lietotājiem – humanitāro un sociālo zinātņu pētniekiem –, kas kavē CLARIN pētniecības infrastruktūras pilnvērtīgu izmantojumu. Viens no soļiem, kas ļautu palielināt savstarpēju sapratni un sadarbību, ir sekmīga Latvijas Atveseļošanas un noturības mehānisma plāna ieviešana[18], kurš paredz augsta līmeņa digitālo prasmju apguvi valodu tehnoloģiju jomā. Sekmējot Latvijas atvērtās zinātnes stratēģijas ieviešanu, jāturpina darbs pie licencēšanas, īpaši atvērtās piekļuves, un FAIR principu popularizēšanas zinātnieku vidū un atbalsta CLARIN-LV infrastruktūrā.

18 https://likumi.lv/ta/id/322858-par-latvijas-atveselosanas-un-noturibas-mehanisma-planu

Bibliogrāfija

Erjavec, Tomaž, Ogrodniczuk, Maciej, Osenova, Petya, Ljubešič, Nikola, Simov, Kiril, Pančur, Andrej, Rudolf, Michał, Kopp, Matyáš, Barkarson, Starkaður, Steingrímsson, Steinþór, Çöltekin, Çagrı, de Does, Jesse, Depuydt, Katrien, Agnoloni, Tommaso, Venturi, Giulia, Calzada Pérez, D de Macedo, Luciana, María, Navarretta, Costanza, Luxardo, Giancarlo, Coole, Matthew, Rayson, Paul, Morkevičius, Vaidas, Krilavičius, Tomas, Dargis, Roberts, Ring, Orsolya, van Heusden, Ruben, Marx, Maarten, Fišer, Darja, (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Erjavec, Tomaž et al. (2021). *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1.*, Slovenian language resource repository CLARIN.SI. Available: http://hdl.handle.net/11356/1432 [accessed 09.05.2022.].

Levāne-Petrova, Kristīne, Darģis, Roberts (2018). *Balanced Corpus of Modern Latvian (LVK2018)*. CLARIN-LV digital library at IMCS, University of Latvia. Available: http://hdl.handle.net/20.500.12574/11 [accessed 09.05.2022.].

Rituma, Laura, Pretkalniņa, Lauma, Saulīte, Baiba, Nešpore-Bērzkalne, Gunta, Grūzītis, Normunds (2020). *Latvian Treebank v2.5*. CLARIN-LV digital library at IMCS, University of Latvia. Available: http://hdl.handle.net/20.500.12574/10 [accessed 09.05.2022.].

Reinsone, Sanita, Ļaksa-Timinska, Ilze, Jaudzema, Justīne (2021). *Corpus of Latvian Pandemic Diaries 2020–2021*. CLARIN-LV digital library at IMCS, University of Latvia. Available: http://hdl.handle.net/20.500.12574/48 [accessed 09.05.2022.].

Skadina, Inguna, Auzina, Ilze, Gruzitis, Normunds, Znotins, Arturs (2020). Clarin in Latvia: From the preparatory phase to the construction phase and operation. *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*.

Spektors, Andrejs, Pretkalniņa, Lauma, Grūzītis, Normunds, Paikens, Pēteris, Rituma, Laura, Saulīte, Baiba (2019). *Tēzaurs.lv 2020*. CLARIN-LV digital library at IMCS, University of Latvia. Available: http://hdl.handle.net/20.500.12574/9 [accessed 09.05.2022.].

Váradi, Tamás, Krauwer, Steven, Wittenburg, Peter, Wynne, Martin, Koskenniemi, Kimmo (2008). CLARIN: Common Language Resources and Technology Infrastructure. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Znotiņš, Artūrs (2015). *NLP-PIPE: Latvian NLP Tool Pipeline*. CLARIN-LV digital library at IMCS, University of Latvia. Available: http://hdl.handle.net/20.500.12574/4 [accessed 09.05.2022.].

CLARIN Language Resources and Technology Infrastructure for the Humanities and Social Sciences

Inguna Skadiņa, Ilze Auziņa,
Roberts Darģis, Arnis Voitkāns

Keywords: research infrastructures, digital humanities, language resources, Latvian language, knowledge infrastructure, open science, FAIR

Established in 2012, CLARIN research infrastructure aims to maintain an infrastructure to support the sharing, use and sustainability of the language data and tools for research in the social sciences and humanities (SSH). In Latvia, after joining CLARIN ERIC in 2016, work on creation of CLARIN research infrastructure started in 2018. This paper aims to provide overview of CLARIN infrastructure and introduce the researchers of Latvian digital humanities to the opportunities offered by CLARIN ERIC and in particular CLARIN Latvia (CLARIN-LV) node.

At first, we introduce to the fundamental elements of CLARIN ERIC – Virtual Language Observatory, Language Resource Families and funding and cooperation mechanisms. Then, we provide an overview of CLARIN-LV repository and language resources and tools documented. We explain citation mechanisms and provide practical recommendations for depositing and licensing. The core values of CLARIN are very closely aligned with the FAIR data principles, therefore we stress the importance of long-term preservation of research outcomes (including language resources and tools) and explain CLARIN role in supporting open science and FAIR principles. Furthermore, mechanisms for the knowledge sharing and user support are presented. It includes CLARIN knowledge centres, targeted seminars, university level courses and individual consultations. Finally, paper outlines future goals of CLARIN-LV and the next steps to implement them. The main directions include extension and adaptation of repository content, long-term cooperation with SSH researchers, support for higher level education in language technologies and support for implementation of open science principles.