
LATVIJAS
UNIVERSITĀTE

Promocijas darba
kopsavilkums

Roberts Darģis

# INFRASTRUKTŪRA LATVIEŠU VALODAS KORPUSU IZSTRĀDEI UN LIETOJUMIEM

Rīga 2025

LATVIJAS
UNIVERSITĀTE

EKSAKTO ZINĀTŅU UN TEHNOLOĢIJU FAKULTĀTE

**Roberts Darģis**

# INFRASTRUKTŪRA LATVIEŠU VALODAS KORPUSU IZSTRĀDEI UN LIETOJUMIEM

Promocijas darba kopsavilkums

Zinātniskā doktora grāda (Ph.D.) iegūšanai
Zinātnes nozare – datorzinātne un informātika
Apakšnozare – datoru un sistēmu programmatūra

Rīga 2025

Promocijas darbs izstrādāts Latvijas Universitātes Matemātikas un informātikas institūtā Mākslīgā intelekta laboratorijā laika posmā no 2016. gada līdz 2025. gadam.

Darba forma: publikāciju kopa datorzinātņu un informātikas nozarē, datoru un sistēmu programmatūras apakšnozarē.

Darba zinātniskais vadītājs: *Dr. sc. comp.,* prof. **Guntis Bārzdiņš**, Latvijas Universitāte.

Darba recenzenti:

1. *Dr. sc. comp.* **Kārlis Čerāns**, Latvijas Universitāte;
2. *Dr. sc. comp.* **Matīss Rikters**, Nacionālais progresīvās rūpniecības zinātnes un tehnoloģiju institūts, Japāna;
3. *Dr. sc. comp.* **Jurgita Kapočiūte-Dzikiene**, Vitauta Dižā Universitāte, Lietuva.

Promocijas darba aizstāvēšana notiks 2025. gada 17. oktobrī Latvijas Universitātes Datorzinātnes un informātikas nozares un Elektrotehnikas, elektronikas, informācijas un komunikāciju tehnoloģiju nozares promocijas padomes atklātā sēdē Raiņa bulvārī 19.

Ar promocijas darbu un tā kopsavilkumu var iepazīties Latvijas Universitātes Bibliotēkā Rīgā, Raiņa bulvārī 19.

LU Datorzinātnes un informātikas nozares un Elektrotehnikas,
elektronikas, informācijas un komunikāciju tehnoloģiju zinātņu nozares
promocijas padomes priekšsēdētāja *Dr. sc. comp.* **Inguna Skadiņa**

Promocijas padomes sekretāre **Sintija Siliņa**



ISBN 978-9934-36-424-2
ISBN 978-9934-36-425-9 (PDF)

# Anotācija

Šis promocijas darbs veltīts latviešu valodas korpusu infrastruktūras izveidei un lietojumiem. Pētījumā aplūkota nepieciešamība pēc strukturētiem un viegli pieejamiem lingvistiskajiem datiem, kas noder gan dabiskās valodas apstrādei, gan korpusos balstītiem pētījumiem digitālajās humanitārajās zinātnēs, politikas zinātnē un lingvistikā. Pētījuma laikā ir izstrādāti rīki un metodes korpusu veidošanai, marķēšanai un to pieejamības nodrošināšanai.

Promocijas darbs ir veidots kā tematiski saistītu publikāciju kopa, kas aptver vairākus projektus un praktiskus lietojumus, demonstrējot izstrādātās infrastruktūras efektivitāti. Iegūtie rezultāti ir nozīmīgi dabiskās valodas apstrādes lietojumprogrammās, valodas modelēšanā un starpdisciplināros akadēmiskajos pētījumos.

# Saturs

# 1. Ievads

Pēdējā desmitgadē dabiskās valodas apstrādes (NLP) un digitālo humanitāro zinātņu jomas ir strauji attīstījušās. Valodu specifiski korpusi ir kļuvuši par stūrakmeni gan akadēmiskos pētījumos, gan tehnoloģiju inovācijās. Resursiem bagātām valodām ir ievērojami sasniegumi NLP jomā, bet valodas, kurām ir mazāk resursu, NLP tehnoloģiju attīstībā joprojām atpaliek. Efektīvu valodas modeļu izstrādi kavē visaptverošu korpusu trūkums, tādējādi ierobežojot latviešu valodas tehnoloģisko progresu.

## 1.1 Pētījuma problēmas aktualitāte

Šī pētījuma nepieciešamība izriet no tā, ka latviešu valoda nav pietiekami pārstāvēta datorlingvistikā un dabiskās valodas apstrādes lietojumos. Šo promocijas darbu motivē atziņa, ka mūsdienās dati ir būtiskākais aspekts dabīgās valodas apstrādes risinājumu izstrādē. Lielākā daļa populāro mašīnmācīšanās rīkkopu ir paredzētas daudzvalodu lietojumam, bet to efektivitātes paaugstināšanai ir nepieciešams liels apjoms kvalitatīvu datu.

Pievēršoties infrastruktūras izveidei, kas īpaši pielāgota latviešu valodas korpusiem, šis pētījums ne vien veicina latviešu valodas iekļaušanu mūsdienīgos NLP rīkos, bet arī atvieglo starpdisciplinārus pētījumus datorlingvistikā, politikas zinātnē un digitālajās humanitārajās zinātnēs.

## 1.2 Pētījuma mērķis un uzdevumi

Šī pētījuma galvenais mērķis ir integrēt latviešu valodu mūsdienīgos NLP rīkos un veicināt korpusos balstītus pētījumus. Lai sasniegtu šo mērķi, pētījumā ir izvirzīti vairāki specifiski uzdevumi.

- **Izveidot pielāgotu infrastruktūru**
  Izstrādāt un ieviest specializētu infrastruktūru, kas optimizē latviešu valodas korpusu izveidi, marķēšanu un pārvaldību. Tā ietver intuitīvas saskarnes un automatizētus darba plūsmas risinājumus, kas samazina manuālā darba apjomu, vienlaikus nodrošinot augstu datu kvalitāti un konsekvenci.
- **Konstruēt strukturētus un sadarbspējīgus korpusus**
  Veidot daudzveidīgus korpusus ar standartizētām daudzslāņu marķēšanas shēmām, kas nodrošina, ka lingvistiskie dati ir viegli pieejami un sadarbspējīgi ar starptautiskajiem formātiem, tādējādi uzlabojot to izmantojamību gan valodniecības pētījumos, gan mūsdienīgos NLP lietojumos.
- **Izstrādāt daudzpusīgus digitālos rīkus**

Attīstīt lietotājiem ērtas digitālās platformas, kas atvieglo korpusos balstītus pētījumus. Šie rīki ir paredzēti, lai pētnieki digitālajās humanitārajās zinātnēs un politikas zinātnēs varētu efektīvi analizēt latviešu valodas datus plašākā pētniecības un lietojuma kontekstā.

## 1.3 Hipotēzes

- Automātiski marķēšanas rīki un strukturētas datu plūsmas samazinās korpusa izveidei nepieciešamo manuālo darbu, vienlaikus saglabājot augstu precizitāti.
- Marķētu korpusu pieejamība veicinās latviešu valodai pielāgotu NLP modeļu izstrādi un uzlabos to veiktspēju.
- Strukturēta un viegli pieejama latviešu valodas korpusu infrastruktūra būtiski veicinās korpusos balstītu pētniecību dažādās disciplīnās.

## 1.4 Pētījuma metodes

Šajā darbā izmantotās pētījuma metodes tika rūpīgi izvēlētas, lai sasniegtu izvirzītos mērķus – izstrādāt latviešu valodas korpusu izveides, marķēšanas un lietošanas infrastruktūru un rīkus. Pētījumā tika apvienotas gan kvalitatīvās, gan kvantitatīvās metodes, lai nodrošinātu, ka izstrādātie rīki un metodes ir ne tikai tehniski pamatotas, bet arī efektīvas praktiskos lietojumos.

- **Literatūras apskats**
  Tika veikta visaptveroša esošās zinātniskās literatūras un pieejamo tehnoloģiju izpēte, lai izveidotu stabilu izpratni par korpusu izstrādi, marķēšanas metodēm un dabiskās valodas apstrādi. Šis pārskats sniedza ieskatu aktuālajās pieejās, palīdzēja identificēt esošo metožu trūkumus un veidoja pamatu latviešu valodai pielāgotu metožu izstrādē.
- **Kvantitatīvā un kvalitatīvā novērtēšana**
  Infrastruktūras veiktspēja tika rūpīgi izvērtēta, izmantojot noteiktus rādītājus. Tika analizēti galvenie veiktspējas indikatori, piemēram, marķējuma precizitāte, apstrādes efektivitāte un kļūdu samazināšana, lai noteiktu automatizēto procesu efektivitāti salīdzinājumā ar manuālajām metodēm. Kvantitatīvā analīze sniedza skaidru pamatu uzlabojumu izvērtēšanai un salīdzināšanai ar esošajiem risinājumiem. Papildus tika veikta kvalitatīvā novērtēšana – kļūdu analīze un manuāla pārbaude, lai nodrošinātu lingvistisko precizitāti un datu konsekvenci.
- **Kontrolēti eksperimenti**
  Tika izstrādāti vairāki kontrolēti eksperimenti, lai novērtētu atsevišķu infrastruktūras komponentu ietekmi. Salīdzinot dažādas automatizēto marķētāju un datu pārveidošanas algoritmu versijas kontrolētos apstākļos, pētījumā tika identificētas visefektīvākās pieejas latviešu valodas lingvistisko īpatnību apstrādei un kļūdu samazināšanai.

- **Praktiski lietojumi**
  Infrastruktūras vispārējā piemērotība praktiski tika pārbaudīta dažādos projektos, tostarp parlamenta debašu korpusos, apguvēju valodas korpusos un runas korpusos. Šo praktisko lietojumu rezultāti apstiprināja infrastruktūras stabilitāti un vispārēju pielietojamību dažādās jomās, tostarp lingvistikā, digitālajās humanitārajās zinātnēs un dabiskās valodas apstrādē.
- **Iteratīva izstrāde un pielāgošana**
  Korpusa infrastruktūras izstrāde un ieviešana tika veikta iteratīvi. Prototipi tika izstrādāti, testēti un nepārtraukti pilnveidoti, pamatojoties uz eksperimentu rezultātiem un ekspertu atsauksmēm. Šāda pieeja nodrošināja sistēmas attīstību un ļāva ātri veikt uzlabojumus, reaģējot uz tehniskajiem izaicinājumiem un lietotāju prasībām.

Kopā šīs pētniecības metodes nodrošināja stabilu pamatu latviešu valodas korpusu izveides, marķēšanas un pētniecības infrastruktūras izstrādei, novērtēšanai un nepārtrauktai uzlabošanai.

## 1.5 Galvenie darba rezultāti

Promocijas darba galvenais rezultāts ir daudzpusīgs pētījumu kopums, kas būtiski sekmē latviešu valodas korpusu izveidi, pārvaldību un lietojumu.

- **Automatizētas marķēšanas un datu apstrādes plūsmas**
  Izstrādātā un ieviestā visaptverošā infrastruktūra atvieglo korpusu veidošanu, marķēšanu un pārvaldību. Pielāgotu rīku un metožu izstrāde ievērojami samazina nepieciešamību pēc manuāla darba un nodrošina augstu datu kvalitāti.
- **Dažādu latviešu valodas korpusu izstrāde**
  Šajā pētījumā izveidoti vairāki specializēti korpusi, tostarp parlamenta debašu korpusi, apguvēju korpusi un dažādi runas korpusi. Šie resursi aizpilda esošās nepilnības latviešu valodas resursos un paver iespējas jauniem pētījumiem digitālajās humanitārajās zinātnēs.
- **Dažādu korpusu integrācija**
  Pētījumā ir standartizēti dažādu jomu korpusi, un tie ir apvienoti vienotā infrastruktūrā – Korpuss.lv. Tas ievērojami sekmē latviešu valodas apstrādes rīku attīstību un sniedz plašas iespējas korpusos balstītiem pētījumiem digitālajās humanitārajās zinātnēs un datorlingvistikā.

Šo rezultātu pamatā ir apjomīgi pētījumi, kas veikti Latvijas Universitātes Matemātikas un informātikas institūta (LU MII) Mākslīgā intelekta laboratorijā (AiLab). Pētījums ir komandas darba rezultāts, kura īstenošanā autors veicis nozīmīgu ieguldījumu. Autora personīgais ieguldījums ir detalizēti aprakstīts promocijas darbā.

## 1.6 Praktiskā nozīme un rezultātu aprobācija

Papildus teorētiskajam ieguldījumam šim pētījumam ir arī ievērojama praktiskā nozīme.

- **Aprobācija pētniecības projektos**
  Šajā promocijas darbā izstrādātās metodes un rīki veiksmīgi ieviesti vairākos nacionāla un starptautiska mēroga projektos, kas apliecina pētījuma rezultātu praktisko nozīmi. Atbalsts pētījumiem iegūts augstas konkurences projektu konkursos, ko finansē Eiropas Reģionālās attīstības fonds, Latvijas Zinātnes padome un dažādas valsts pētījumu programmas.
- **Akadēmiskais novērtējums**
  Infrastruktūras aprobāciju apliecina tās plašais izmantojums starpdisciplināros projektos un pozitīvās atsauksmes akadēmiskajā vidē. Pastāvīga rezultātu citēšana zinātniskajās publikācijās apstiprina darba nozīmīgumu un ilgtermiņa ietekmi.
- **Valodas tehnoloģiju attīstība**
  Šis pētījums būtiski uzlabo piekļuvi latviešu valodas datiem, jo piedāvā standartizētu un mērogojamu platformu. Tas ne vien veicina valodas tehnoloģiju attīstību, bet arī atbalsta pētniecību digitālajās humanitārajās zinātnēs, jo korpusi ir vieglāk atrodami un izmantojami.

Šī promocijas darba rezultāti ir aprobēti vairāku modeļu apmācībā.

- **Ilvars – latviešu valodas vīriešu balss runas sintēzes modelis**[1] – neironu tīkla modelis runas sintēzei latviešu valodā. Modelis apmācīts, izmantojot VITS rīkkopu uz 25 stundu audiogrāmatu korpusa, kas ierunāts vīrieša balsī. API pieejams izmantošanai akadēmiskiem un nekomerciāliem mērķiem.
- **Latviešu BERT bāzes modelis**[2] – BERT valodas modelis, kas trenēts uz latviešu valodas datiem, izmantojot maskētās valodas modelēšanas un nākamā teikuma prognozēšanas uzdevumus (Znotins and Barzdins, 2020).
- **Plaša pielietojuma latviešu runas atpazīšanas modelis**[3] – latviešu valodai pielāgots *whisper-large-v3* modelis, kurā izmantotas divas runas datu kopas: latviešu valodas daļa no Common Voice 19.0 un jaunākais LATE-mediju korpuss.
- **Plaša pielietojuma latgaliešu runas atpazīšanas modelis**[4] – latgaliešu valodai pielāgots *whisper-large-v3* modelis, kurā izmantota latgaliešu valodas daļa no Common Voice 19.0.

[1] Ilvars – latviešu valodas vīriešu balss runas sintēzes modelis – https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/89

[2] Latviešu BERT bāzes modelis – https://huggingface.co/AiLab-IMCS-UL/lvbert

[3] Plaša pielietojuma latviešu ASR modelis – https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-lv-late-cv19

[4] Plaša pielietojuma latgaliešu ASR modelis – https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-latgalian-2503

## 1.7 Pētnieciskie projekti

Šajā promocijas darbā aprakstītie pētījumi ir īstenoti un aprobēti dažādos nozīmīgos projektos, kas uzskatāmi apliecina izstrādātās infrastruktūras un metožu daudzpusību un praktisko pielietojumu.

**Eiropas Reģionālās attīstības fonda investīciju programmas *Atbalsts starptautiskās sadarbības projektiem pētniecībā un inovācijās* pētījumu projekts *Latvijas Universitāte un institūti Eiropas pētniecības telpā – ekselence, aktivitāte, mobilitāte, kapacitāte* (2018–2022) (1.1.1.5/18/I/016)**

Šī projekta mērķis bija veicināt Latvijas pētniecības institūciju konkurētspēju, mobilitāti un kapacitāti Eiropas Pētniecības telpā. Galvenās aktivitātes ietvēra augsta līmeņa projektu pieteikumu sagatavošanu, konferenču organizēšanu un dalību tajās, kā arī sadarbības stiprināšanu Eiropas pētniecības infrastruktūrās.

Tiešā saistībā ar promocijas darbu tika uzsākta Latvijas valodas resursu repozitorija CLARIN-LV izveide, ieviests noSketchEngine rīks un sagatavoti korpusi standartizētos formātos. Šie sākotnējie risinājumi ir pamatā turpmākai korpusu infrastruktūras attīstīšanai.

**Latvijas Zinātnes padomes Fundamentālo un lietišķo pētījumu projekts *Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums* (LaVA) (2018–2021) (lzp-2018/1-0527)**

Šī projekta mērķis bija izveidot pētniecības ietvaru latviešu valodas apguves izpētei, izstrādājot valodas apguvēju korpusu. Galvenie uzdevumi ietvēra kļūdu marķēšanas metodoloģijas izstrādi, korpusa infrastruktūras izveidi un apguvēju kļūdu kvantitatīvo un kvalitatīvo analīzi.

Promocijas darbā tika izstrādāta kļūdu marķēšanas shēma, automatizētā apstrādes darbplūsma un korpusa pārvaldības rīki. Dalība šajā projektā ne tikai uzlaboja marķēšanas procesa precizitāti un efektivitāti, bet arī apliecināja promocijas darbā izstrādātās metodoloģijas lietderību.

**Eiropas Reģionālās attīstības fonda programmas *Praktiskas ievirzes pētījumi* pētniecības projekts *Latviešu valodas runas atpazīšana un sintēze medicīnas lietojumiem* (RUTA:MED) (2019–2022) (1.1.1.1/18/A/153)**

Projekta mērķis bija izstrādāt specializētus latviešu valodas resursus, lai pielāgotu runas atpazīšanas un sintēzes tehnoloģijas medicīnas jomai, īpaši radioloģijai.

Promocijas darbā tika veikts metodoloģiskais un infrastruktūras pētījums MediSpeech korpusa izstrādei un tika izveidota transkripcijas un marķēšanas darbplūsma, kas pielāgota medicīnisko izmeklējumu diktātiem, sistemātiski risinot ar nozares specifisko valodu saistītās tehniskās problēmas.

**Valsts pētījumu programmas *Humanitāro zinātņu digitālie resursi* pētījumu projekts *Humanitāro zinātņu digitālie resursi: integrācija un attīstība* (2020–2022) (VPP-IZM-DH-2020/1-0001)**

Šī projekta galvenais mērķis bija veicināt digitālo humanitāro zinātņu resursu attīstību un integrāciju Latvijā. Novēršot resursu sadrumstalotību un veicinot starpinstitucionālo sadarbību, tika uzlabota digitālo resursu pieejamību un izmantojamību.

Projektā tika izstrādāta Korpuss.lv platformas pirmā versija, kas ir centrālais promocijas darba temats. Promocijas darbā piedāvātā infrastruktūra un metodes bija nozīmīgas platformas veidošanā. Šī infrastruktūra ne tikai apvieno dažādus latviešu valodas korpusus, bet arī veicina starpdisciplinārus pētījumus digitālajās humanitārajās zinātnēs.

**Eiropas Savienības investīciju programmas *Apvārsnis 2020* pētījumu projekts *SELMA – Stream Learning for Multilingual Knowledge Transfer* (2021–2023) (957017)**

Šajā starptautiskajā projektā tika izveidota daudzvalodu atvērtā pirmkoda platforma, kas spēj apstrādāt lielu datu apjomu. Platformas mērķis bija palīdzēt plašsaziņas līdzekļu pārraudzītājiem un žurnālistiem efektīvi analizēt lielu informācijas plūsmu, kā arī veidot audiovizuālo saturu, papildinot to ar transkripciju, tulkojumu un subtitriem, tādējādi padarot to pieejamāku.

Sadarbības gaitā apstiprinājās promocijas darbā izstrādāto datu kopu un marķēšanas formātu lietderība progresīvu NLP modeļu un daudzvalodu satura apstrādes sistēmu izveidē.

**Valsts pētījumu programmas *Letonika latviskas un eiropeiskas sabiedrības attīstībai* pētījumu projekts *Mūsdienu latviešu valodas izpēte un valodas tehnoloģiju attīstība* (LATE) (2022-2024) (VPP-LETONIKA-2021/1-0006)**

Projekts bija vērsts uz modernās latviešu runas izpēti, tostarp tās gramatiskajiem, fonētiskajiem un fonoloģiskajiem aspektiem. Šī mērķa sasniegšanai bija nepieciešams izstrādāt atbilstošus valodas resursus un rīkus.

Promocijas darbā tika izveidota runas korpusa infrastruktūra un transkripcijas metodoloģija, kas tika ieviesta un aprobēta LATE-mediji un LATE-sarunas korpusu izstrādē. Savstarpējā mijiedarbība starp promocijas darba pētījumu un šo projektu ir ļāvusi būtiski uzlabot gan valodas tehnoloģiju attīstību, gan korpusu izstrādes stratēģiju.

**ES Atveseļošanas un noturības mehānisma *Augsta līmeņa digitālo prasmju apguves nodrošināšana* pētījumu projekts *Valodu tehnoloģiju iniciatīva* (2023–2026) (2.3.1.1.i.0/1/22/I/CFLA/002)**

Iniciatīvas mērķis ir izstrādāt un pilnveidot liela mēroga valodas modeļus, gramatikas un leksikonus gan vienvalodas, gan daudzvalodu audiovizuālo datu ap-

strādei, vienlaikus veidojot resursus un rīkus, lai palīdzētu izstrādātājiem un lietotājiem apgūt valodas tehnoloģijas.

Promocijas darbā sākotnēji izstrādātā Korpuss.lv platforma tika aktīvi izmantota akadēmiskiem pētījumiem. Tas ne tikai sniedza vērtīgu atgriezenisko saiti, kas ļāva būtiski uzlabot platformu, bet arī uzlaboja akadēmisko vidi, atbalstot pētniecības projektus un studiju kursus digitālajās humanitārajās zinātnēs un datorlingvistikā. Infrastruktūras izmantošana izglītībā apstiprināja platformas praktisko lietojamību, tās efektivitāti topošo ekspertu apmācībā un tās lomu starpdisciplināru pētījumu veicināšanā valodas tehnoloģijās.

**Latvijas Zinātnes padomes Fundamentālo un lietišķo pētījumu projekts *Biežākās kļūdas latviešu valodā: korpusā balstīta kļūdu analīze un teksta labošana* (Norma) (2024–2026) (lzp-2023/1-0481)**

Projekta mērķis ir izveidot daļēji automātiski marķētu latviešu valodas kā dzimtās valodas runātāju kļūdu korpusu, kurā tiks dokumentētas, labotas un skaidrotas biežākās latviešu valodas kļūdas. Kļūdu korpuss tiks izmantots, lai izstrādātu pilnīgāku gramatikas pārbaudītāju, kas norāda ne tikai uz tehniskām neprecizitātēm un vienkāršākām pareizrakstības vai interpunkcijas kļūdām, bet arī uz teikuma konstrukciju izveides kļūdām.

Šajā projektā tiek paplašināta un aprobēta iepriekšējos projektos gūtā pieredze un izveidotā kļūdu korpusu izstrādes metodika. Iepriekšējos projektos tika izstrādāti valodas apguvēju kļūdu korpusi, bet šis projekts ir balstīts uz dzimtās valodas runātāju kļūdu korpusu, kurā sastopamas sarežģītākas konstrukciju un stila kļūdas.

## 1.8 Pētījuma rezultātu publikācijas

Promocijas darbs ir veidots kā publikāciju kopa. Promocijas darba autors ir piedalījies 37 publikāciju izstrādē, no kurām lielākā daļa ir iekļautas Scopus datubāzē un trīs no tām ir iekļautas Q1 līmeņa žurnālos. Promocijas darbā ir iekļautas 14 rūpīgi atlasītas pamatpublikācijas, no kurām 13 ir indeksētas Scopus. Promocijas darba autors ir galvenais autors 10 no šīm publikācijām.

Šajā darbā iekļautajās pamatpublikācijās detalizēti aprakstīta latviešu valodas korpusu infrastruktūras izstrāde, ieviešana un lietojums, atspoguļojot tās attīstību no sākotnējās koncepcijas līdz praktiskai lietošanai. Šīs publikācijas veido promocijas darba kodolu un sniedz padziļinātu ieskatu metodoloģijā, eksperimentos un iegūtajos rezultātos.

1. **R. Dargis**, G. Rabante-Busa, I. Auzina and S. Kruks, *ParliSearch – A system for large text corpus discourse analysis*, Human Language Technologies – The Baltic Perspective, Vol. 289, IOS Press. (2016). Scopus, WoS

2. **R. Dargis**, I. Auzina, U. Bojars, P. Paikens and A. Znotins, *Annotation of the Corpus of the Saeima with Multilingual Standards*, Proceedings of the 2018

ParlaCLARIN Workshop. (2018).

3. **R. Dargis**, I. Auzina and K. Levane-Petrova, *The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners*, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pp. 4111–4115. (2018). Scopus, WoS

4. U. Bojars, **R. Dargis**, U. Lavrinovics and P. Paikens, *LinkedSaeima: a Linked Open Dataset of Latvia's Parliamentary Debates*, Proceedings of the 15th SEMANTiCS Conference, Vol. 11702, Springer, pp. 50–56. (2019). Scopus

5. **R. Dargis**, N. Gruzitis, I. Auzina and K. Stepanovs, *Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian*, Human Language Technologies – The Baltic Perspective, Vol. 328, IOS Press, pp. 135–141. (2020). Scopus, WoS

6. **R. Dargis**, I. Auzina, K. Levane-Petrova and I. Kaija, *Detailed Error Annotation for Morphologically Rich Languages: Latvian Use Case*, Human Language Technologies – The Baltic Perspective, Vol. 328, IOS Press, pp. 241–244. (2020). Scopus, WoS

7. **R. Dargis**, I. Auzina, K. Levane-Petrova and I. Kaija, *Quality Focused Approach to a Learner Corpus Development*, Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 392–396. (2020). Scopus, WoS

8. N. Gruzitis, **R. Dargis**, V. Lasmanis, G. Garkaje and D. Gosko, *Adapting Automatic Speech Recognition to the Radiology Domain for a Less-Resourced Language: The Case of Latvian*, Intelligent Sustainable Systems, Vol. 333, Springer, pp. 267–276. (2022). Scopus

9. **R. Dargis**, I. Auzina, I. Kaija, K. Levane-Petrova and K. Pokratniece, *Corpus Based Self-Assessment Platform for Latvian Language Learners*, Baltic Journal of Modern Computing, Vol. 10(3), pp. 392–401. (2022). Scopus, WoS

10. **R. Dargis**, I. Auzina, I. Kaija, K. Levane-Petrova and K. Pokratniece, *LaVA – Latvian Language Learner corpus*, 13th Language Resources and Evaluation Conference (LREC), pp. 727–731. (2022). Scopus, WoS

11. B. Saulite, **R. Dargis**, N. Gruzitis, I. Auzina, K. Levane-Petrova, L. Pretkalnina, L. Rituma, P. Paikens, A. Znotins, L. Strankale, K. Pokratniece, I. Poikans, G. Barzdins, I. Skadina, A. Baklane, V. Saulespurens and J. Ziedins,

*Latvian National Corpora Collection – Korpuss.lv*, 13th Language Resources and Evaluation Conference (LREC), pp. 5123–5129. (2022). Scopus, WoS

12. **R. Dargis**, A. Znotins, I. Auzina, B. Saulite, S. Reinsone, R. Dejus, A. Klavinska and N. Gruzitis, *BalsuTalka.lv – Boosting the Common Voice Corpus for Low-Resource Languages*, Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 2080–2085. (2024). Scopus

13. **R. Dargis** and B. Saulite, *Korpuss.lv – a Versatile Platform for Digital Humanities*, Baltic Journal of Modern Computing, Vol. 12(4), University of Latvia, pp. 636–645. (2024). Scopus, WoS

14. I. Auzina, N. Gruzitis, **R. Dargis**, G. Rabante-Busa, D. Gosko, J. Vempers, R. Kivkucans and A. Znotins, *Recent Latvian Speech Corpora for Linguistic Research and Technology Development*, Baltic Journal of Modern Computing, Vol. 12(4), University of Latvia, pp. 646–658. (2024). Scopus, WoS

Autors ir piedalījies arī vairāku starptautiski recenzētu publikāciju izstrādē, kas pastarpināti saistītas ar promocijas darba tēmu un demonstrē plašāku pētījuma ietekmi uz datorlingvistiku un digitālajām humanitārajām zinātnēm.

1. **R. Dargis** and A. Znotins, *Baseline for keyword spotting in Latvian broadcast speech*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

2. I. Auzina, M. Pinnis and **R. Dargis**, *Comparison of rule-based and statistical methods for grapheme to phoneme modelling*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

3. G. Garkaje, E. Zilgalve and **R. Dargis**, *Normalization and automatized sentiment analysis of contemporary online Latvian Language*, Human Language Technologies – The Baltic Perspective, Vol. 268, IOS Press. (2014). Scopus, WoS

4. A. Znotins, K. Polis and **R. Dargis**, *Media monitoring system for Latvian radio and TV broadcasts*, Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). (2015). Scopus, WoS

5. I. Auzina, K. Levane-Petrova, G. Rabante-Busa, **R. Dargis** and A. Fabregas, *Designing an annotated longitudinal Latvian children's speech corpus*, Human Language Technologies – The Baltic Perspective, Vol. 289, IOS Press. (2016). Scopus, WoS

6. A. Spektors, I. Auzina, **R. Dargis**, N. Gruzitis, P. Paikens, L. Pretkalnina, L. Rituma and B. Saulite, *Tezaurs.lv: the largest open lexical database for Latvian*, Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC). (2016). Scopus, WoS

7. **R. Dargis** and I. Auzina, *Towards a Modern Text-to-Speech System for Latvian*, Human Language Technologies – The Baltic Perspective, Vol. 307, IOS Press, pp. 26–29. (2018). Scopus, WoS

8. O. Urek, A. Vulane, **R. Dargis**, A. Taurina, T. Zirina and H. G. Simonsen, *Latvian CDI: methodology, developmental trends and cross-linguistic comparison*, Journal of Baltic Studies, Vol. 50(3), Routledge, pp. 285–305. (2019). Scopus, WoS

9. I. Auzina, **R. Dargis** and K. Levane-Petrova, *Latviešu valodas apguvēju kļūdu analīze: pareizrakstības kļūdas*, Vārds un tā pētīšanas aspekti, LiePA, pp. 220–227. (2019).

10. N. Gruzitis, **R. Dargis**, L. Rituma, G. Nespore-Berzkalne and B. Saulite, *Deriving a PropBank Corpus from Parallel FrameNet and UD Corpora*, Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet, pp. 63–69. (2020).

11. **R. Dargis**, P. Paikens, N. Gruzitis, I. Auzina and A. Akmane, *Development and Evaluation of Speech Synthesis Corpora for Latvian*, Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6633–6637. (2020). Scopus, WoS

12. **R. Dargis**, K. Levane-Petrova and I. Poikans, *Lessons Learned from Creating a Balanced Corpus from Online Data*, Human Language Technologies - The Baltic Perspective, Vol. 328, IOS Press, pp. 127–134. (2020). Scopus, WoS

13. I. Auzina, I. Kaija, K. Levane-Petrova, K. Pokratniece and **R. Dargis**, *Latviešu valodas apguvēju korpusa (LaVA) izmantošana pētniecībā un mācību uzdevumu izstrādē*, Latviešu valodas apguve. XIII Starptautiskais baltistu kongress, LiePA, pp. 142–161. (2021).

14. L. Skestere and **R. Dargis**, *Agenda-Setting Dynamics during COVID-19: Who Leads and Who Follows?*, Social Sciences, Vol. 11(12), pp. 556. (2022). Q1, Scopus, WoS

15. I. Skadina, I. Auzina, **R. Dargis** and A. Voitkans, *CLARIN valodas resursu*

*un rīku pētniecības infrastruktūra humanitārajām un sociālajām zinātnēm*, Letonica, Vol. 47, pp. 312–327. (2022). Scopus

16. I. Skadina, I. Auzina, **R. Dargis**, E. Lasmanis and A. Voitkans, *CLARIN-LV: Many Steps till Operation*, CLARIN Annual Conference, pp. 9–13. (2022).

17. A. Znotins, **R. Dargis**, N. Gruzitis, G. Barzdins and D. Gosko, *RUTA:MED – Dual Workflow Medical Speech Transcription Pipeline and Editor*, Natural Language Processing and Information Systems, Vol. 13286, Springer, pp. 209–214. (2022). Scopus, WoS

18. I. Auzina, **R. Dargis**, B. Saulite, N. Gruzitis, M. Grasmanis, A. Spektors and K. Stepanovs, *Specializēta latviešu valodas runas korpusa un izrunas vārdnīcas izveide vizuālās diagnostikas izmeklējumu lingvistiskai analīzei un sistemātiskai transkribēšanai*, Letonica, Vol. 47, pp. 244–262. (2022). Scopus

19. I. Auzina, **R. Dargis**, I. Kaija, K. Levane-Petrova and K. Pokratniece, *Valodas korpusu izmantošana latviešu valodas uzdevumu automātiskā ģenerēšanā*, Letonica, Vol. 47, pp. 264–282. (2022). Scopus

20. B. Saulite, I. Auzina and **R. Dargis**, *Nacionālā korpusu kolekcija Korpuss.lv*, Linguistica Lettica, Vol. 31(1), LU Latviešu valodas institūts, pp. 202–223. (2023). Scopus

21. T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubesic, K. Simov, A. Pancur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrimsson, C. Coltekin, J. de Does, K. Depuydt, T. Agnoloni, G. Venturi, M. C. Perez, L. de Macedo, C. Navarretta, G. Luxardo, M. Coole, P. Rayson, V. Morkevicius, T. Krilavicius, **R. Dargis**, O. Ring, R. van Heusden, M. Marx and D. Fiser, *The ParlaMint corpora of parliamentary proceedings*, Language Resources and Evaluation, Vol. 57, Springer, pp. 415–448. (2023). Q1, Scopus, WoS

22. **R. Dargis**, G. Barzdins, I. Skadina, N. Gruzitis and B. Saulite, *Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams*, Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, pp. 289–293. (2024). Scopus

23. T. Erjavec, M. Kopp, N. Ljubesic, T. Kuzman, P. Rayson, P. Osenova, M. Ogrodniczuk, C. Coltekin, D. Korzinek, K. Meden, J. Skubic, P. Rupnik, T. Agnoloni, J. Aires, S. Barkarson, R. Bartolini, N. Bel, M. C. Perez, **R. Dargis** and e. al., *ParlaMint II: advancing comparable parliamentary corpora across Europe*, Language Resources and Evaluation. (2024). Q1, Scopus, WoS

Šīs publikācijas kopumā ne tikai apstiprina pētījuma metodoloģiju, bet arī izceļ tā ilgtermiņa praktisko un akadēmisko nozīmi.

# 2. Korpusi

Šajā nodaļā sniegts pārskats par korpusiem, kas izveidoti, izmantojot 3. nodaļā aprakstīto infrastruktūru un metodoloģiju. Šie korpusi ir būtiska rīku un metodoloģiju aprobācijas sastāvdaļa un ir par pamatu padziļinātai infrastruktūras un metodoloģisko aspektu izpētei 3. nodaļā, kas ir šī promocijas darba galvenais mērķis.

## 2.1 Nacionālā korpusu kolekcija (NKK)

Latviešu valodas korpusi arvien biežāk tiek izmantoti lielo valodas modeļu priekšapmācībai, piemēram, LVBERT (Znotins and Barzdins, 2020), LitLatBERT (Ulcar et al., 2021) un GPT2-LV (Plenert, 2021). Lielo valodas modeļu bezpiemēru (*zero-shot*) mācīšanās spēja ir atkarīga ne tikai no modeļa lieluma un datu apjoma, bet arī no apmācības korpusos atrodamo enciklopēdisko zināšanu kvalitātes un plašuma. Šis faktors ir veicinājis "GoodData" koncepcijas attīstību (Press, 2021). Revolucionārais valodas modelis GPT-3 (Brown et al., 2020) tika apmācīts, izmantojot 750GB lielu galvenokārt angļu valodas "GoodData" datu kopumu, savukārt GPT-SW3, kas paredzēts salīdzināšanā "mazajai" zviedru valodai, tika apmācīts ar 100GB teksta kolekciju (Ekgren et al., 2022). Nacionālā korpusu kolekcija (NKK), kuras šībrīža apjoms sasniedz gandrīz 10GB un aptver plaša spektra saturu, ir nozīmīgs solis latviešu valodas "GoodData" izveidē. Tas ir kritiski svarīgi augstas kvalitātes latviešu valodas lielo valodas modeļu apmācībai, kas ir nākotnes dabiskās valodas izpratnes (NLU) un ģenerēšanas (NLG) pamatā.

Tā kā latviešu valodai ir salīdzinoši maz resursu, maz ticams, ka viens korpuss kādreiz būs tik liels un kvalitatīvs, lai apmācītu tik lielus valodas modeļus kā GPT-3 angļu valodai. Turklāt neviens atsevišķs līdzsvarots latviešu valodas tekstu korpuss nevar būt pilnībā pietiekams mūsdienu leksikogrāfijas un gramatikas pētījumiem. Joprojām nav pietiekami aptverti jauni tekstu veidi un avoti, piemēram, lietotāju veidots saturs un sarunvaloda, kā arī specializētas jomas.

NKK mērķis ir apvienot vairākus korpusus vienotā kolekcijā ar standartizētu formātu. Mērķis ir aptvert dažādus latviešu valodas lietojumus un iekļaut visus galvenos teksta veidus un žanrus, piemēram, ziņas, sociālos medijus, emuārus, grāmatas, zinātniskos tekstus, debates un esejas, sabalansējot gan kvalitāti, gan apjomu. Lai sasniegtu šo mērķi, ir nepieciešams nepārtraukts, daudzu institūciju un vairāku projektu darbs, ko atbalsta Latvijas digitālo humanitāro zinātņu un valodas tehnoloģiju kopiena. Pašlaik NKK ir 39 korpusi, ko izstrādājušas 13 institūcijas.

Gandrīz visi NKK korpusi ir automātiski morfoloģiski marķēti, izmantojot LU MII izstrādāto atvērtā pirmkoda morfoloģisko marķētāju, kas sasniedz 92,7% pilnas morfoloģiskās marķēšanas precizitāti un 97,6% lemmatizācijas precizitāti

(Paikens et al., 2013; Paikens, 2016). Latviešu valodas marķēšanas shēma, kas izstrādāta un pilnveidota LU MII gadu gaitā, ir pozicionāla shēma, kas atvasināta no MULTEXT-EAST standarta (Erjavec, 2012) un ir pielāgota latviešu valodas īpatnībām (Paikens et al., 2024).

## 2.2 Latvijas parlamentārie korpusi

Šī nodaļā aprakstīti centieni uzlabot Latvijas parlamentāro datu izmantošanas iespējas digitālo humanitāro zinātņu un politikas zinātnes pētniekiem.

Šis pētījums bija par pamatu Latvijas parlamentārā korpusa iekļaušanai *ParlaMint* korpusā – salīdzināmā un savstarpēji sadarspējīgā 29 Eiropas valstu un autonomo reģionu parlamentāro debašu kolekcijā (Erjavec et al., 2023, 2024).

### *2.2.1 Saeimas korpuss*

Saeimas korpuss (Auziņa et al., 2019) pirmo reizi tika publicēts 2016. gadā. Sākotnējā versija bija pieejama vienkāršā teksta formā ar norādēm par runātājiem un ar citiem metadatiem.

Korpusa dati tika iegūti no Saeimas mājaslapas[1], kur tiek publicēti visu Saeimas sēžu stenogrammu pilnie teksti HTML formātā. Teksti tika pusautomātiski apstrādāti, lai identificētu runātājus un to maiņu. Teksts tika segmentēts izteikumos, nodrošinot, lai katrs izteikums satur tikai viena runātāja runu. Korpuss ietver datus no 1993. līdz 2022. gadam.

### *2.2.2 Automātiskā marķēšana ar plaši izmantotiem standartiem*

Pieaugot korpusu pieejamībai dažādās valodās, kļuva skaidrs, ka nemarķēti teksti nav pietiekami salīdzināmiem pētījumiem. Korpusu papildināšana ar informācijas slāņiem, kas balstās uz plaši izmantotiem standartiem, veicina starpvalodu pētījumus, neprasot pētniekiem zināt visas iesaistītās valodas.

Saeimas korpuss tika papildināts ar vairākiem marķēšanas slāņiem:

- morfosintaktisko informāciju lingvistiskai analīzei, ietverot lemmas, morfoloģiskās pazīmes un sintaktiskās atkarības;
- mašīntulkojumiem angļu valodā;
- nosaukto entitāšu pieminējumiem, kas sasaistīti ar *Wikidata* zināšanu bāzi.

Morfosintaktiskais marķējums ietver lemmas, vārdšķiras, morfoloģiskās pazīmes un sintaktiskās atkarības, ievērojot universālo atkarību (UD) standarta formātu. Teksti tiek automātiski morfoloģiski marķēti, izmantojot CMM balstītu marķētāju (Paikens, 2016). Sintaktiskās atkarības tiek noteiktas ar neironu tīklos balstītu atkarību marķētāju (Znotins, 2016), kas trenēts uz latviešu valodas UD korpusa versiju 2.1 (Pretkalnina et al., 2016).

[1] Saeimas mājaslapa – https://www.saeima.lv/lv/transcripts/category/21

Runātais teksts tika mašīntulkots angļu valodā, izmantojot neironu tīklos balstītu sistēmu (Barone et al., 2017). Mašīntulkojumi atvieglo kvantitatīvu analīzi un uzlabo meklēšanas iespējas starptautiskos pētījumos. Tomēr, tā kā šādi tulkojumi var nebūt precīzi, kvalitatīviem pētījumiem ieteicams izmantot profesionālu tulkojumu.

Nosaukto entitāšu atpazīšanai un saistīšanai tika izstrādāta speciāla sistēma, balstoties uz iepriekšējiem pētījumiem par ziņu korpusu analīzi (Paikens, 2016). Šajā pieejā kā entitāšu zināšanu bāze tiek izmantoti strukturēti *Wikidata* dati (Ismayilov et al., 2018). Lai nodrošinātu precīzu entitāšu sasaisti ar *Wikidata* identifikatoriem, tā tika papildināta ar vārdu locījumiem un automātiski tika ģenerēti personvārdu, saīsinājumu un organizāciju nosaukumu varianti.

### *2.2.3 LinkedSaeima*

*LinkedSaeima* ir Saeimas korpusa saistīto datu (*linked data*) reprezentācija. Tā ietver strukturētu informāciju par parlamenta sēdēm un tajās minētajām nosauktajām entitātēm ar saitēm uz *Wikidata* identifikatoriem. Saistītie dati nodrošina parlamentāro debašu strukturētu attēlojumu, definējot parlamenta sēžu objektus, to īpašības un savstarpējās saiknes.

*LinkedSaeima* korpuss tika izveidots, pārveidojot Saeimas korpusu saistīto datu formātā līdzīgi kā *Europarl* modelī. Tas ietver strukturētus metadatus un saites uz *Wikidata*. Kopumā datu kopums sastāv no aptuveni 4,9 miljoniem RDF ierakstu, kas nodrošina paplašinātas meklēšanas iespējas un integrāciju ar globālajām saistīto datu infrastruktūrām.

Datu kopums ietver 497 221 runu (izteikumu) no 1 293 parlamenta sēdēm. Šīs runas ir teikuši 690 runātāji, kuri ieņēmuši 162 dažādus amatus, un tās satur 392 530 atsauces uz 2 998 unikālām *Wikidata* entitātēm. Datu kopums sniedz strukturētu informāciju par vairākām objektu klasēm.

- Sēde (*lpv_eu:SessionDay*) – parlamenta plenārsēde, kas parasti ietver vairākas runas.
- Runa (*lpv_eu:Speech*) – individuāls izteikums, ko sēdes laikā teicis viens runātājs.
- Runātājs (*lpv:Speaker*) – persona, kas uzstājas ar runu.
- Amats (*lpv:PoliticalFunction*) – runātāja politiskais amats runas teikšanas brīdī (piemēram, Ministru prezidents). Viena persona dažādos laika posmos var ieņemt vairākus amatus.

*LinkedSaeima* datu modelī ievērota *LinkedEP* projekta struktūra un izmantota *LinkedPolitics* vārdnīca, kas šajā darbā tiek apzīmēta ar prefiksiem *lpv* un *lpv_eu* (Van Aggelen et al., 2017).

Šī korpusa galvenā inovācija, salīdzinot ar *LinkedEP*, ir nosaukto entitāšu informācijas iekļaušana. Tā tiek attēlota ar īpašību *schema:mentions*, kas sasaista entitātes ar atbilstošajiem *Wikidata* identifikatoriem. Turklāt runātāju amatiem tiek piešķirti unikāli URI identifikatori *schema:mentions*, kas ļauj veikt vaicājumus

(piemēram, atlasīt runas, kuras teikuši ārlietu ministri) un sasaistīt amatus ar ārējām datu kopām. Runātāju amati (*lpv:PoliticalFunction*) var arī saturēt saites uz attiecīgajām *Wikidata* entitātēm.

## 2.3 Runas korpusi

Pieprasījums pēc daudzveidīgiem runas korpusiem strauji pieaug. Tie ir nepieciešami gan dažādām atvērtā pirmkoda un komerciālām runas tehnoloģijām, gan moderniem pētījumiem valodniecībā un digitālajās humanitārajās zinātnēs. Valodas tehnoloģiju attīstībā korpusi ir nepieciešami, lai uzlabotu un novērtētu runas atpazīšanas (ASR) un runas sintēzes (TTS) modeļus. Valodniecībā runas korpusi noder fonētikas, prosodijas, morfoloģijas, sintakses, semantikas un pragmatikas izpētē.

Šajā sadaļā ir aprakstīti vairāki konceptuāli atšķirīgi latviešu valodas runas korpusi, kuru izstrādē autors ir sniedzis nozīmīgu ieguldījumu.

### *2.3.1 LATE sarunu korpuss (LATE-sarunas)*

LATE sarunu korpusā (Auziņa et al., 2024b) ir iekļautas privātas sarunas, intervijas, publisku uzstāšanos, piemēram, konferenču, ieraksti un to atšifrējumi ortogrāfiskajā transkripcijā.

Katram audioierakstam pievienoti metadati, tostarp runātāja dzimums un vecuma grupa (12–15, 16–25, 26–50, 51–75, 76+), kā arī informācija par runas formu: dialogs vai monologs, spontāna vai sagatavota runa.

Korpusā ir 35 stundu ieraksti no vairāk nekā 300 runātājiem. Datu kopa tiek izplatīta CLARIN-LV repozitorijā ar CLARIN ierobežoto licenci.

### *2.3.2 LATE plašsaziņas līdzekļu runas korpuss (LATE-mediji)*

LATE mediju korpuss (Auziņa et al., 2024a) satur Latvijas sabiedrisko mediju raidījumu ierakstus, kas aptver gan spontānu, gan sagatavotu runu.

Runas dati tiek transkribēti atbilstoši latviešu valodas ortogrāfiskajiem standartiem, ievērojot pieturzīmes un citas gramatikas normas.

Korpuss sastāv no 70 stundām ierakstu, kuros piedalās vairāk nekā 250 runātāju. Datu kopa tiek izplatīta CLARIN-LV repozitorijā ar CLARIN akadēmisko licenci.

### *2.3.3 Radioloģijas runas korpuss (LVMED)*

Lai gan Latvijā plaši tiek izmantotas modernas medicīniskās tehnoloģijas, medicīniskie atzinumi joprojām tiek sagatavoti manuāli. Lielākās veselības aprūpes iestādes šo dokumentu sagatavošanai uztur iekšējos transkripcijas centrus vai izmanto ārpakalpojumus. Pieaugot diagnostikas izmeklējumu skaitam, gan ārsti, gan pacienti nereti saskaras ar vairāku dienu aizkavēšanos atzinumu saņemšanā. Turklāt

transkripcijas pakalpojumi ir dārgi, kas padara tos nepieejamus reģionālajām veselības aprūpes iestādēm.

Universālas ASR sistēmas, kas apmācītas uz visaptverošiem runas un teksta korpusiem, to specifiskās valodas dēļ nav piemērotas medicīniskajiem atzinumiem. Vārdu kļūdu līmenis (WER) ir pārāk augsts, lai šādas sistēmas būtu efektīvi izmantojamas, tāpēc būtiska nozīme ir domēnam pielāgotai ASR sistēmai.

Iedvesmojoties no veiksmīgas igauņu pieredzes ASR sistēmas ieviešanā radioloģijā (Paats et al., 2018; Alumäe et al., 2017), arī latviešu valodai tika izstrādāti ASR pielāgošanai nepieciešamie resursi, tai skaitā 30 stundu LVMED korpuss (Auziņa et al., 2022), kurā audioieraksti manuāli precīzi transkribēti. Korpuss veidots no reāliem ierakstiem, kas atlasīti no slimnīcu transkripciju centru arhīviem.

Ieraksti veido 10% no aptuveni 300 stundu arhīva, un tie atlasīti, balstoties uz runātājiem un akustisko apstākļu dažādību. Audioieraksti tika filtrēti pēc kvalitātes, atmetot zemas kvalitātes ierakstus. Augstas kvalitātes ierakstus, ja nepieciešams, var mākslīgi degradēt, taču pretējais nav iespējams. Rezultātā korpuss ietver aptuveni 70 unikālus runātājus.

Lai korpuss būtu līdzsvarots un daudzpusīgs, katram runātājam tika atlasīts aptuveni vienāds ierakstu skaits. Tā kā pat viena runātāja ierakstīšanas apstākļi var atšķirties, ierakstu atlase tika veidota maksimāli garā periodā, nodrošinot vienmērīgus intervālus starp ierakstiem.

Marķēšanas vadlīnijas medicīnisko diktātu transkribēšanai tika pielāgotas balstoties uz pieredzi, kas iegūta vispārīga latviešu valodas runas korpusa izveidē (Pinnis et al., 2014, 2016).

### *2.3.4 BalsuTalka.lv un BolsuTalka.lv*

No 2023. gada vidus līdz 2024. gada vidum ievērojami tika paplašināta daudzvalodu korpusa *Common Voice*[2] (CV) latviešu valodas daļa un tās dažādība. Tas tika panākts, īstenojot valsts mēroga sabiedrisko iniciatīvu *BalsuTalka.lv*. Šajā kampaņā rūpīgi atlasītu tekstu korpusu ierunāja tūkstošiem dažādu vecuma grupu un tautību cilvēku gan Latvijā, gan diasporā. 2023. gada beigās kampaņa veiksmīgi tika paplašināta – tika pievienota arī latgaliešu valoda, kas līdz tam CV korpusā nebija pārstāvēta.

Pirmais solis CV runas korpusa paplašināšanā ir ierunājamo tekstu papildināšana. Pirms kampaņas latviešu valodas CV korpusā bija aptuveni 7 000 teikumu, kas galvenokārt iegūti no filmu subtitriem. Lai palielinātu dažādību un paplašinātu ierakstīto runas paraugu skaitu, korpuss tika papildināts līdz gandrīz 30 000 teikumu. Teksta korpusā tika iekļauti dažādu žanru teksti ar daudzveidīgu leksiku no dažādiem funkcionālajiem stiliem. Galvenā uzmanība tika pievērsta tam, lai teksti būtu viegli lasāmi un ekspresīvi – tika iekļauti jautājuma teikumi, izsaukuma teikumi, dialogi un sarunvalodas fragmenti.

Dati tika papildināti, ievērojot tēmu dažādību (piemēram, ziņu virsraksti, receptes), leksisko un teikumu konstrukciju daudzveidību. CV 18.0 versijā ir ie-

[2] *Common Voice* – https://commonvoice.mozilla.org

kļautas 293 stundas latviešu valodā un tās ierunājuši 6 086 runātāji. No tām ir pārbaudītas 244 stundas.

Lai izveidotu CV korpusu latgaliešu valodai, *Mozilla* CV lietotāja saskarne tika lokalizēta un platformā tika iesniegts sākotnējais 5 000 teikumu kopums latgaliski. Izvēles kritērijos ietilpa tekstu atbilstība latgaliešu rakstu valodai, fonētiskā un intonatīvā daudzveidība (stāstījuma, jautājuma, izsaukuma teikumi) un satura dažādība. Teksta fragmenti no vārdnīcām, īsi dialogi un frazeoloģismi no daiļliteratūras un populārzinātniskiem darbiem tika pievienoti manuāli.

Latgaliešu "Bolsu tolka" kampaņa tika organizēta kā paplašinājums latviešu "Balsu talkai", radot sinerģiju starp abām iniciatīvām. Pašreizējais latgaliešu CV tekstu korpuss satur gandrīz 10 000 teikumu, savukārt CV 18.0 versijā latgaliešu valodā ir iekļautas 27 ierakstītas stundas (321 runātājs) un 25 pārbaudītas stundas. *Common Voice* korpusi ir brīvi pieejami gan pētniecībai, gan komerciālai lietošanai.

## 2.4 Korpusi ar kļūdu marķējumu

Korpusi ar kļūdu marķējumu parasti tiek veidoti no valodas apguvēju radītiem tekstiem, kas sistemātiski tiek analizēti, lai identificētu, labotu un kategorizētu lingvistiskās kļūdas.

Valodas apguvēju korpusi tiek veidoti un analizēti jau vairāk nekā 25 gadus. Angļu valodai ir izstrādāti daudzi korpusi (Granger et al., 2009; Gilquin et al., 2010), līdzīgi resursi kļūst arvien populārāki arī citām valodām, piemēram, franču (Granfeldt et al., 2006), zviedru (Volodina et al., 2019), norvēģu (Tenfjord et al., 2006), nīderlandiešu (Lemmens and Perrez, 2010), japāņu (Gries and Adelman, 2014), arābu (Alfaifi et al., 2014), ķīniešu (Wang et al., 2015), portugāļu (Mendes et al., 2016) un citām valodām.

Pieaug arī pieprasījums pēc latviešu valodas kā svešvalodas apguves. Latviešu valodu kā svešvalodu māca ne tikai augstskolās Latvijā, bet arī vairāk nekā 20 universitātēs visā pasaulē (Šalme, 2011; Laizane et al., 2018). Tādēļ korpusos balstīti mācību materiāli ir būtiski ārvalstu studentiem, kas apgūst latviešu valodu gan Latvijā, gan ārvalstīs. Valodas apguvēju korpusi ir svarīgi ne tikai studentu valodas izpētei, bet arī izglītībā izmantojumu lietotņu izstrādei.

Kļūdu korpusa izstrādes principi un platforma tika radīta un aprobēta, veidojot divus apguvēju korpusus. Detalizēta informācija par marķēšanas shēmu un procesu ir sniegta 3.2. nodaļā.

### *2.4.1 Valsts valodas prasmes pārbaudes darbu korpuss (VVPP)*

Sākotnējā kļūdu marķēšanas pieeja tika izstrādāta un aprobēta VVPP korpusa (Auziņa et al., 2018) izveidē. Šis korpuss tika izveidots Latviešu valodas aģentūras pētījumā *Latviešu valodas prasmes kvalitāte: valsts valodas prasmes pārbaudes rezultāti*.

Korpusu veido veiksmīgi nokārtoti valsts valodas prasmes pārbaudes darbi, kuros tiek vērtēts pretendenta valsts valodas apguves līmenis. Korpusā ir apkopoti

900 valsts valodas pārbaudes darbu rakstītprasmes testi: no katra valodas apguves līmeņa (A1, A2, B1, B2, C1, C2) korpusā iekļauti 150 darbi. Pēc pārbaudes nokārtošanas pretendentam tiek izsniegts valsts valodas prasmes sertifikāts, kas nepieciešams nodarbinātībai un pastāvīgās uzturēšanās atļaujas iegūšanai. Diemžēl šis korpuss nav publiski pieejams privātuma apsvērumu dēļ.

### *2.4.2 Latviešu valodas apguvēju korpuss (LaVA)*

Kļūdu marķēšanas pieeja tālāk pilnveidota LaVA korpusa (Auziņa et al., 2021) izstrādes laikā. Šis korpuss tika izveidots Latvijas Zinātnes padomes Fundamentālo un lietišķo pētījumu projektā *Latviešu valodas apguvēju korpusa izveide: metodes, rīki un izmantojums*.

LaVA korpuss satur 1 015 esejas (190 000 tekstvienību un 790 000 rakstzīmju, neskaitot atstarpes), kuras rakstījuši ārzemju studenti Latvijas augstākajās mācību iestādēs. Šie studenti apgūst latviešu valodu kā svešvalodu pirmo vai otro semestri, parasti sasniedzot A1 (vai, iespējams, A2) valodas apguves līmeni.

Vieni no svarīgākajiem jautājumiem, kas jārisina, pirms sāk vākt šāda tipa datus, ir autortiesības un personas datu aizsardzība. Lai autorus informētu par, kā viņu radītie teksti tiks izmantoti un lai iegūtu viņu nepārprotamu piekrišanu šādai datu izmantošanai (Kaija and Auzina, 2020), tika izstrādāta strukturēta anketa un piekrišanas veidlapa. Veidlapa ir izdrukāta uz vienas A4 lapas puses, un tai ir trīs daļas: informatīva vēstule, piekrišanas apliecinājums un metadatu anketa, kurā iekļauta informācija par autoru: dzimums, vecums, dzimtā valoda, citu valodu zināšanas un tas, cik ilgi tiek apgūta latviešu valoda. Otrā lapas puse ir atstāta tukša, lai autori ar roku varētu uzrakstīt savas esejas.

Biežākās eseju tēmas ir: *Es un mana ģimene*, *Mana ikdiena* un *Manas studijas*. Lai aizsargātu personu datus, studenti tiek aicināti rakstīt izdomātu informāciju.

Darbi tika ievākti no piecām universitātēm: Rīgas Stradiņa universitātes (87%), Rēzeknes Tehnoloģiju akadēmijas (4%), Latvijas Universitātes (3%), Liepājas Universitātes (3%) un Latvijas Kultūras akadēmijas (3%).

# 3. Infrastruktūra un metodes

Šajā nodaļā ir aprakstīta infrastruktūra un metodes, uz kurām balstās iepriekšējā nodaļā minēto latviešu valodas korpusu izstrāde. Turpmākajās nodaļās sīkāk aplūkots izstrādes process, marķēšanas shēmas un rīki, kas atvieglo korpusu izveidi un apstrādi. Tajās arī parādīts, kā šie komponenti palīdz risināt problēmas, kas saistītas ar korpusu izstrādi morfoloģiski bagātai valodai ar ierobežotiem valodas resursiem, kāda ir latviešu valoda. (Auziņa et al., 2021).

## 3.1 Runas korpusi

Liela, rūpīgi atlasīta runas korpusa izveide ir laikietilpīgs uzdevums. Lai sasniegtu konkurētspējīgus rezultātus, mūsdienu automātiskās runas atpazīšanas (ASR) sistēmu izstrādei pietiek ar mazāku datu apjomu, tāpēc ka var pielāgot lielus, iepriekš apmācītus daudzvalodu modeļu, piemēram, XLS-R (Babu et al., 2022), Whisper (Radford et al., 2023) un MMS (Pratap et al., 2024).

Šajā nodaļā apskatītas dažādas pieejas, kas lietotas vairāku konceptuāli atšķirīgu latviešu runas korpusu izstrādē. Šie korpusi tiek izmantoti gan datos balstītai pētniecībai digitālajās humanitārajās zinātnēs, gan ASR modeļu izstrādei.

### *3.1.1 Runas korpusu transkripcija*

Transkripcijas formāts ir būtisks runas korpusu aspekts, jo tas ietekmē korpusu interpretāciju un pielietojumu dažādās pētniecības un tehnoloģiju izstrādes jomās. Transkripcijas formāta izvēle var ievērojami ietekmēt analīzes dziļumu un mašīnmācītu modeļu precizitāti.

Lai izveidotu visaptverošu transkripcijas formātu, ir jāpieņem vairāki lēmumi, jo tajā jāiekļauj dažādi informācijas slāņi. Labi definēts transkripcijas formāts nosaka, kā tiek atspoguļotas izrunas atšķirības, teksta formatējums un runas īpatnības (piemēram, pārteikšanās, nepabeigti vārdi un nepareiza izruna). Ortogrāfiskās transkripcijas pieraksts būtiski ietekmē datu lietojamību. Piemēram, skaitļu rakstīšana vārdos (”divdesmit pieci”) pretstatā cipariem (”25”) ietekmē gan lasāmību, gan automātisko runas atpazīšanu. Tāpat transkripcijas vadlīnijās ir jānorāda, vai vārdi, kas izrunāti citā valodā, jāatzīmē atsevišķi, kas ir īpaši svarīgi daudzvalodu korpusos. Papildu izaicinājumus rada saīsinājumi, jo to izruna var atšķirties no rakstiskās formas. Piemēram, lēmums par to, vai transkribēt ”ASV” kā izrunu ”ā es vē” vai kā vienu vārdu, ietekmē ASR apmācību. Papildu sarežģījumus rada diktēšanas korpusi, jo tie parasti ietvert dažādas komandas kā ”punkts”, ”komats” vai ”jauna rinda”, kas skaidri jānošķir no dabiski izrunātas interpunkcijas, lai izvairītos no pārpratumiem.

Visu runas korpusu, kas aprakstīti 2. nodaļā, transkripcija balstīta uz vieniem un tiem pašiem principiem. Transkripcijās tiek norādīti neverbālie elementi, neskaidra runa un fizioloģiskie trokšņi. Transkripcija tiek veikta vienā slānī, kur papildu informācija tiek iezīmēta, izmantojot īpašu marķējumu. Tas atsevišķos gadījumos ļauj iegūt specifiskus transkripcijas līmeņus.

Kvadrātiekavās norāda novirzes no standarta izrunas normām (piemēram, ”lasām [lasam]”; ”interesanti [intresanti]”), kā arī uz saīsinājumu un svešvārdu izrunu (piemēram, ”SIA [si ā]”; ”ZZS [zē zē es]”; ”Rail [reil] Baltica [boltik]”). Tajās atzīmē arī saīsinātus vārdus (piemēram, ”četrdesmit [čēesnt]”), kā arī skaitļu izrunu, kas prasa sintaktiski saistīto formu saskaņošanu (piemēram, ”7,8 [septiņi komats astoņi] grami” – nominatīvs; ”līdz 1940. [tūkstoš deviņsimt četrdesmitajam] gadam” – datīvs).

LVMED diktēšanas korpusā izpildāmās komandas tiek apzīmētas figūriekavās (piemēram, ”{punkts}”, ”{jauna rinda}”).

Korpusiem, kam iepriekš bija pieejams rakstisks teksts, ir saglabāts atsevišķs transkripcijas slānis, kas ļauj veikt turpmākus pētījumus. LVMED korpusā ir arī pacientiem sniegto medicīnisko atzinumu teksti. Tas dod iespēju analizēt un apmācīt modeļus par to, kā transkripcija tiek rediģēta līdz pabeigtam tekstam, kāds nepieciešams pacientu dokumentācijā. LATE-mediji korpusā ir iekļauti konkrētu segmentu subtitri, kas ļauj analizēt, kā tiek saīsināts runātais teksts un kāda informācija tiek izlaista, lai atbilstu subtitru vadlīnijām.

Segmenti, kuros ir neskaidra vai pārprotama izruna, ASR apmācību var ietekmēt negatīvi. Tomēr šie segmenti ir īpaši vērtīgi specifiskiem lingvistiskajiem pētījumiem. Detalizēta ortogrāfiskā transkripcija ļauj veikt padziļinātu lingvistisko analīzi un filtrēt augstas kvalitātes datus ASR apmācībai.

### *3.1.2 Balsu talka*

Liela apjoma atvērtu runas korpusu pieejamība mazāk izplatītām valodām ir ierobežota. Ilgu laiku tas attiecās arī uz latviešu valodu. Lai gan eksistēja vairāki slēgti latviešu runas korpusi ar vairāk nekā 100 stundu apjomu, kas tika izmantoti automātiskās runas atpazīšanas (ASR) modeļu apmācībai, 2023. gada sākumā bija pieejami tikai daži nelieli atvērtie korpusi. Lielākais no tiem bija latviešu valodas ”Common Voice” (CV) 13.0, kurā bija 18 stundas ierakstu.

Pateicoties nacionāla mēroga brīvprātīgas iesaistes iniciatīvai ”Balsu talka”, kuru kopīgi organizēja vairākas institūcijas, latviešu valodas CV korpusā gada laikā tika desmitkāršots gan datu apjoms, gan runātāju skaits. Līdzīgs projekts tika veiksmīgi īstenots arī latgaliešu valodai, kuru lieto aptuveni 150 000 cilvēku un kuru mūsdienās uzskata par apdraudētu. Šo iniciatīvu mērķis bija ne tikai paplašināt datu kopas, bet arī veicināt to daudzveidību – aptvert plašu runātāju loku, dažādus izrunas variantus, tekstu žanrus, stilus, intonācijas, gramatiskās konstrukcijas un vārdu krājumu.

Plānojot runas datu vākšanu, īpaša uzmanība tika pievērsta tam, kā vislabāk veicināt sabiedrības iesaisti un nodrošināt savākto datu ilgtermiņa ietekmi. Tika

apsvērti divi galvenie varianti: izveidot pielāgotu platformu vai izmantot esošu, globālas iniciatīvas pārvaldītu platformu. Globāla platforma spētu nodrošināt lielāku ilgtermiņa ietekmi, bet caur to varētu būt sarežģītāk uzrunāt vietējo sabiedrību. Savukārt pielāgota platforma ļautu veiksmīgāk iesaistīt sabiedrību un stiprinātu kultūras saikni, bet prasītu ievērojamu ieguldījumu izstrādē un varētu ierobežot starptautisko atpazīstamību.

Galu galā tika izvēlēta hibrīda pieeja, kas apvieno abu variantu priekšrocības – tika izveidota pielāgota mājaslapa ar atmiņā paliekošu domēna nosaukumu "BalsuTalka.lv", kur lietotāji tika tālāk novirzīti uz CV platformu[1].

Šī mērķtiecīgā kampaņa veicināja plašu līdzdalību gan Latvijā, gan diasporā. Kampaņas "BalsuTalka" nosaukums un dizains atspoguļoja spēcīgu latvisko identitāti. Kampaņa balstījās uz vietējiem medijiem un kultūrai tuvām vērtībām, tika izmantoti nacionālie simboli, lai rezonētu ar auditoriju un veicinātu iesaisti. Pielāgotā sākumlapa skaidri aicināja piedalīties, tādējādi palīdzot izvairīties no neskaidrībām, kādas var rasties, ja lietotāji uzreiz tiktu novirzīti uz lielu starptautisku platformu.

CV platforma nodrošināja projekta tehnisko pamatu – uzticamu, mērogojamu infrastruktūru ar stingriem privātuma un drošības standartiem. Šī standartizētā daudzvalodu datu krātuve ir plaši atzīta starptautiskajā dabiskās valodas apstrādes kopienā, tādēļ latviešu valodas dati ir viegli pieejami un savietojami ar plašākiem daudzvalodu pētījumiem un valodu modeļiem.

Iniciatīva tika oficiāli uzsākta 2023. gada 4. maijā – Latvijas neatkarības atjaunošanas dienā. Sākotnēji šī kampaņa tika plaši reklamēta sabiedriskajā televīzijā, radio un tiešsaistes platformās. Vēlāk viena no veiksmīgākajām reklāmas kampaņām norisinājās populārajā latviešu valodas tiešsaistes vārdnīcā Tezaurs.lv, trīs mēnešu laikā iesaistot vairāk nekā 16 000 dalībnieku.

Pirms iniciatīvas sākuma latviešu valodas CV korpusā (13.0 versija, 2023. gada marts) bija 18 ierakstītas un 14 pārbaudītas stundas no 321 runātāja. Sešu mēnešu laikā 15.0 versijā (2023. gada septembris) šie rādītāji pieauga līdz 165 ierakstītām un 88 pārbaudītām stundām no 2 773 runātājiem. Līdz 2024. gada martam 17.0 versijā jau bija 277 ierakstītas un 223 pārbaudītas stundas no 5 712 runātājiem, ierindojot latviešu valodu starp vadošajām valodām dalībnieku skaita ziņā attiecībā pret kopējo dzimtās valodas runātāju skaitu.

Vienlaikus tika uzsākta arī latgaliešu valodas CV korpusa izveide. Tika rūpīgi atlasīti un iesniegti pirmie 5 000 teikumi, ievērojot latgaliešu valodas ortogrāfijas standartus un nodrošinot fonētisko un intonāciju dažādību. Līdz CV 17.0 versijas izlaišanai bija ierakstīti gandrīz 10 000 latgaliešu teikumu, iegūstot 24 pārbaudītas runas stundas no 250 runātājiem.

Šī iniciatīva skaidri apliecina, ka lokalizētas un kultūrai piesaistītas brīvprātīgā darba iniciatīvas runas datu vākšanā var veiksmīgi iesaistīt sabiedrību. Jaunie korpusi jau šobrīd sniedz būtisku ieguldījumu ASR tehnoloģiju attīstībā un valodniecībā.

[1] https://commonvoice.mozilla.org

## 3.2 Kļūdu marķēšanas infrastruktūra

Kļūdu marķēšana ir būtiska, lai izprastu un uzlabotu latviešu valodas apguvēju rakstītprasmi. Šim nolūkam izstrādātā infrastruktūra nodrošina sistemātisku valodas kļūdu marķēšanu tekstos, ko rakstījuši latviešu valodas apguvēji.

Tā kā latviešu valoda ir morfoloģiski bagāta valoda ar plašu formu un vārddarināšanu sistēmu, kā arī relatīvi brīvu vārdu secību, kļūdu marķēšana ir sarežģīta. Izstrādātā marķēšanas shēma ir piemērota arī citām valodām ar līdzīgām īpašībām – brīvu vārdu secību un bagātu morfoloģiju.

Lai atbalstītu marķēšanas procesu, ir izstrādāta infrastruktūra, kas atvieglo apguvēju tekstu digitalizāciju, labošanu un morfoloģisko marķēšanu. Šī infrastruktūra ir aprobēta, izveidojot divus nozīmīgus korpusus: *Valsts valodas prasmes pārbaudes darbu korpusu* (VVPP) un *Latviešu valodas apguvēju korpusu* (LaVA).

Turpmākajās sadaļās tiek detalizēti aprakstītas šīs infrastruktūras sastāvdaļas, tostarp korpusa izveides process un kļūdu marķēšanas shēma, kas nodrošina latviešu valodas apguvēju kļūdu sistemātisku kategorizāciju.

### *3.2.1 Korpusa izveides process*

Korpusa izveides process sastāv no četriem galvenajiem soļiem:

1. datu digitalizācija;
2. teksta labošana;
3. morfoloģiskā marķēšana;
4. kļūdu marķēšana.

Katru darbību neatkarīgi veic divi marķētāji, bet neatbilstību gadījumā galīgo lēmumu pieņem trešais, neatkarīgais marķētājs. Kļūdu tipi tiek noteikti automātiski, pamatojoties uz morfoloģisko marķējumu un sastatījumu starp oriģinālo un laboto tekstu.

Lielākā daļa darbu ir rakstīti ar roku un tiem ir nepieciešama digitalizācija. Tikai daži jaunākie darbi COVID-19 attālināto mācību dēļ ir iesniegti digitālā formātā. Tekstu digitalizācijas posmā divu marķētāju vienprātības līmenis rakstzīmju līmenī ir 97,4%.

Teksta labošanas posmā oriģinālais teksts tiek koriģēts atbilstoši gramatikas normām, balstoties uz pieņemto mērķhipotēzi (Auzina et al., 2020). Rakstzīmju līmenī marķētāju vienprātības līmenis šajā posmā sasniedz 96,8%.

Gan oriģinālais, gan labotais teksts tiek morfosintaktiski marķēts. Sākotnējā marķējumu versija tiek ģenerēta, izmantojot LU MII morfoloģisko marķētāju (Paikens, 2016), vēlāk to manuāli pārskata divi marķētāji. Morfosintaktiskais marķējums ietver vārdšķiru, lemmu un citas latviešu valodai raksturīgas morfoloģiskās pazīmes. Papildus tiek norādīta atsevišķa versija oriģinālajam vārdam ar izlabotām pareizrakstības kļūdām. Sastatījums starp oriģinālo un laboto vārdu tiek izmantots kļūdu analīzē, lai automātiski noteiktu, kādus kļūdu veidus pieļāvuši apguvēji. Bez sastatījuma nebūtu iespējams atšķirt pareizrakstības kļūdas no formveidošanas

vai vārddarināšanas kļūdām. Kopējais marķētāju vienprātības līmenis visos slāņos ir 92,5%, ar individuālu precizitāti robežās no 95,5% oriģinālajā marķējumā līdz 99,3% laboto lemmu līmenī.

Pēdējais korpusa izveides posms ir automātiska kļūdu marķēšana. Pašlaik kļūdas tiek klasificētas sešās kategorijās, bet, tā kā šis solis ir automatizēts, tajā var veikt izmaiņas pēc vajadzības. Šīs sešas kategorijas ir: pareizrakstības kļūdas, locīšanas un formveidošanas kļūdas, leksikas kļūdas, interpunkcijas kļūdas, sintakses kļūdas un kombinētās kļūdas.

Katrā darbības solī nepieciešamais laiks tiek automātiski mērīts korpusa platformā. Ja marķētājs ilgāk par 15 sekundēm neveic nekādas darbības, taimeris tiek apturēts un atsāk skaitīt laiku, tiklīdz tiek konstatēta aktivitāte (peles kustība vai tastatūras ievade). Katram solim nepieciešamais vidējais laiks ir apkopots 3.1. tabulā. Viena eseja vidēji tiek apstrādāta 45 minūtēs, jo sākotnējie posmi jāizpilda divas reizes.

| Solis | Ātrums (rakstzīmes minūtē) | Vidējais laiks solim (minūtes) |
|---|---:|---:|
| Sākotnējā digitalizācija | 128.9 | 6.1 |
| Gala digitalizācija | 370.6 | 2.1 |
| Sākotnējā teksta korekcija | 177.8 | 4.4 |
| Gala teksta korekcija | 274.1 | 2.8 |
| Sākotnējā marķēšana | 95.1 | 8.4 |
| Gala marķēšana | 117.1 | 6.6 |

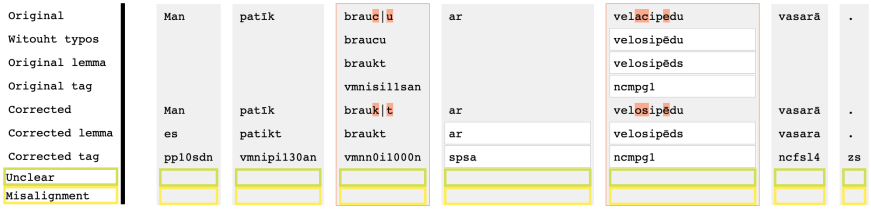
*Tabula 3.1*
**Laika patēriņš korpusa izveides soļos**

### *3.2.2 Kļūdu marķēšanas shēma*

Pētījumā izmantotajā kļūdu marķēšanas shēmā kļūdas netiek marķētas tieši. Tā vietā detalizēti kļūdu kodi tiek ekstrapolēti no piecām viegli identificējamām īpašībām: oriģinālās tekstvienības bez pareizrakstības kļūdām, oriģinālās lemmas, oriģinālā teksta morfoloģiskā marķējuma, labotās lemmas un labotā teksta morfoloģiskā marķējuma (skat. 3.1. attēlu).

Pareizrakstības kļūdas tiek noteiktas automātiski, salīdzinot oriģinālo tekstvienību ar tā laboto versiju. Rakstzīmju līmeņa sastatījums kopā ar likumbāzētu sistēmu ļauj precīzi identificēt nepareizi lietotus rakstzīmju pārus. Šī informācija ir noderīga gan kvalitatīvajiem pētījumiem (jo nodrošina detalizētas meklēšanas iespējas), gan kvantitatīvajai analīzei (grupējot kļūdas).

Morfoloģiskais marķējums sniedz plašu informāciju, tostarp informāciju par vārdšķiru. Interpunkcijas kļūdas ir viegli identificējamas – ja labotā tekstvienība




*Att. 3.1:* **Kļūdu marķēšanas saskarne**

atšķiras no oriģinālās un abas tekstvienības ir pieturzīmes, kļūda tiek fiksēta kā pieturzīmju kļūda.

Leksikas kļūdas tiek konstatētas, ja labotās tekstvienības lemma atšķiras no oriģinālās lemmas.

Visas pārējās kļūdas tiek klasificētas kā gramatikas kļūdas, kuru detalizētai analīzei tiek izmantots morfoloģiskais marķējums.

Papildus iespējams marķēt divas īpašas situācijas – *neskaidrs teksts* un *sastatījuma kļūda*. *Sastatījuma kļūda* tiek lietota gadījumos, kad automātiskā oriģinālā un labotā teksta sastatīšana ir neprecīza, savukārt *neskaidrs teksts* tiek lietots, kad marķēšana ir neviennozīmīga vai kad kļūdu ietekmē plašāks konteksts, bet to nevar atspoguļot pašreizējā shēmā. Šādi gadījumi tiek rūpīgi pārskatīti, lai uzlabotu marķēšanas sistēmu un, ja nepieciešams, ieviestu sintakses līmeņa kļūdu marķēšanu.

## 3.3 Infrastruktūra digitālajām humanitārajām zinātnēm

Valodas pētniecības mērķis ir izprast valodas struktūru, funkcijas un lietojumu, veicot sistemātiskus pētījumus. Valodas korpuss ir strukturēts rakstītu tekstu, transkribētu runas vai video ierakstu kopums. Korpusi nodrošina stabilu empīrisko bāzi gan kvalitatīvai, gan kvantitatīvai analīzei, uzlabojot lingvistisko pētījumu precizitāti un dziļumu (McEnery and Hardie, 2012). Izmantojot korpusus, pētnieki savos pētījumos var panākt lielāku objektivitāti un atkārtojamību. Valodai nemitīgi attīstoties, korpusu nozīme šo pārmaiņu dokumentēšanā un analīzē tikai pieaug, padarot tos par neaizstājamu valodniecības resursu.

Neskatoties uz daudzajiem ieguvumiem, korpusu efektīvu izmantošu ierobežo vairāki izaicinājumi. Daudzi korpusi pētniekiem nav viegli pieejami, tāpēc ka tie ir sarežģītā formātā vai to analīzei ir nepieciešama specializēta programmatūra. Nereti tas prasa īpašas tehniskās prasmes, ierobežojot pētniekus, kuriem šādu zināšanu trūkst.

Šajā sadaļā aprakstīts, kā nodrošināta korpusu pieejamība, lai veicinātu korpusos balstītus pētījumus digitālajās humanitārajās zinātnēs un valodniecībā.

### *3.3.1 Korpuss.lv*

*Korpuss.lv* tika izveidots kā galvenais piekļuves punkts *Nacionālajai korpusu kolekcijai* (NKK).

Vietnes lietotāja saskarne (UI/UX) ir veidota tā, lai lietotājiem atvieglotu dažādu korpusu izpēti. Vietne ir pieejama gan latviešu, gan angļu valodā. Sākumlapā ir pieejams korpusu saraksts ar īsu informāciju par katru korpusu (korpusa kartīte). Korpusu sarakstu var filtrēt un kārtot pēc dažādiem parametriem, lai pētnieki vieglāk varētu atrast viņu pētījumam vispiemērotākos korpusus.

Katrā korpusa kartītē ir norādīta galvenā informācija, tostarp korpusa kods, pilns nosaukums, datu publicēšanas periods, apjoms, izstrādātāji un meklēšanas poga (ja ir pieejama meklēšana tiešsaistē). Noklikšķinot uz kartītes, lietotājs tiek novirzīts uz detalizētu korpusa informācijas lapu.

Korpusa informācijas lapā ir pieejama plašāka informācija un saistītie resursi, tostarp saites uz korpusa mājaslapu, meklēšanas saskarni, vārdu biežumu sarakstu un CLARIN repozitoriju. Šie resursi atvieglo piekļuvi pētniekiem un veicina korpusu izmantošanu zinātniskajos darbos.

Korpusa informācijas lapā ir norādītas svarīgākās publikācijas, kas saistītas ar korpusu un kas sniedz ieskatu tā izstrādē, metodoloģijā un galvenajās atziņās. Šādas atsauces ir būtiskas, lai izprastu pētījuma kontekstu.

Turklāt lapā ir sniegta detalizēta citēšanas informācija, tostarp viena galvenā publikācija un atsauce uz datiem, ja korpuss ir publicēts CLARIN repozitorijā. Šī informācija nodrošina, ka akadēmiskajos darbos tiek izmantota pareiza korpusa atsauce.

Pareiza citēšana ir būtiska korpusa autoriem, jo tā uzlabo viņu citējumu skaitu un indeksēšanas rādītājus. Korpusi bieži tiek izstrādāti pētījumu projektos, kuros projektu ietekme tiek izvērtēta, izmantojot dažādus veiktspējas rādītājus (KPI), lai novērtētu projekta ietekmi. Viens no šādiem KPI ir pētījumu skaits, kuros korpuss ir izmantots. Pareiza citēšana palīdz identificēt un dokumentēt šos pētījumus, demonstrējot korpusa nozīmi un ietekmi, kas savukārt palīdz autoriem piesaistīt finansējumu jaunu resursu izstrādei.

Vienotā meklēšana (FCS) ir efektīvs veids, kā atrast atbilstošus korpusus pētījumam. FCS ļauj pētniekiem meklēt konkrētus vārdus vai frāzes vairākos korpusos vienlaikus. Meklēšanas vaicājums var tikt veikts gan tekstvienību, gan lemmu līmenī. Meklēšanas vaicājums atbalsta aizstājējzīmju simbolus vienai un vairākām rakstzīmēm, kā arī *OR* operatoru.

Ļoti būtiska Korpuss.lv funkcija ir vienotā meklēšana (FCS), kas ļauj meklēt vairāk nekā 35 korpusos vienlaikus. FCS ir efektīvs veids, kā pētījumam atrast atbilstošus korpusus, jo palīdz iegūt statistisku pārskatu par interesējošo vārdu vai frāzi un piedāvā saites uz katra konkrētā korpusa konkordancēm. Tā kā korpusi ir morfoloģiski marķēti, FCS var meklēt gan tekstvienību, gan lemmu līmenī, turklāt meklēšanas vaicājums atbalsta aizstājējzīmju simbolus vienai un vairākām rakstzīmēm, kā arī *OR* operatoru.

### 3.3.2 NoSketch Engine

NoSketch Engine[2] ir daudzfunkcionāls korpusu pārvaldības un analīzes rīks, ko plaši izmanto lingvisti, leksikogrāfi un digitālo humanitāro zinātņu pētnieki (Kilgarriff et al., 2014). Tas atvieglo lielu tekstu korpusu izpēti, ļaujot veikt sarežģītus pieprasījumus un iegūt detalizētu pārskatu par vārdu lietojumu, sintaktiskajiem modeļiem un semantiskajām attieksmēm.

Viena no būtiskākajām NoSketch Engine funkcijām ir konkordances rīks, kas ģenerē sarakstu ar vārda vai frāzes lietojumiem kontekstā. Šī iespēja ļauj pētniekiem analizēt, kā konkrēti termini tiek lietoti dažādos tekstos, laika posmos vai žanros. Rīks atbalsta gan vienkāršus meklējumus, gan sarežģītus vaicājumus, izmantojot korpusa vaicājumu valodu (*Corpus Query Language* – CQL), kas ļauj lietotājiem filtrēt rezultātus pēc dažādiem parametriem, piemēram, pēc vārda formas, lemmas vai morfoloģiskajām pazīmēm.

Platformā ir pieejami arī biežuma saraksti, kas ļauj sakārtot vārdus vai frāzes pēc to sastopamības korpusā. Šie saraksti ir īpaši noderīgi, lai noteiktu tematiskās tendences un salīdzinātu lingvistiskās iezīmes dažādos kontekstos. Pētnieki var iegūt sarakstus atsevišķiem vārdiem, lemmām vai frāzēm un tos tālāk precizēt, izmantojot specifiskus meklēšanas kritērijus. Šī funkcionalitāte ļauj salīdzināt lingvistiskos modeļus dažādos žanros, tekstu tipos vai laika posmos. Piemēram, analizējot īpašības vārdus, kas tiek izmantoti, lai aprakstītu vīriešus un sievietes literārajā korpusā, pētnieki var atklāt dzimumu reprezentācijas tendences. Biežuma sarakstus var izmantot, arī lai izsekotu vārdu lietojuma tendencēm noteiktās jomās, piemēram, analizējot termina ”inflācija” izplatību Saeimas debatēs.

Vēl viena nozīmīga funkcija ir laika līnijas vizualizācija, kas attēlo vārdu vai frāžu lietošanas biežumu laika gaitā. Šī iespēja ir īpaši vērtīga diahroniskajos pētījumos, jo tā ļauj izsekot valodas attīstībai, analizēt izmaiņas publiskajā diskursā vai pētīt konkrētu tēmu popularitātes dinamiku. Laika līnijas tiek veidotas, attēlojot absolūtos vai relatīvos biežuma rādītājus. Piemēram, laika līnijas var atklāt, kā vārda ”krīze” lietojums laika gaitā mainījies Latvijas ziņu rakstos, atspoguļojot sabiedrības interešu maiņu.

### 3.3.3 CLARIN-LV

Kopīgo valodas resursu un tehnoloģiju infrastruktūra (CLARIN) ir nozīmīga Eiropas iniciatīvas daļa, kas nodrošina ilgtspējīgu piekļuvi plašam digitālo valodas resursu un rīku klāstam. Latvija ir šīs iniciatīvas dalībniece. CLARIN-LV repozitorijā tiek glabāti metadati, un lielākā daļa korpusu ir pieejami arī lejupielādei.

### 3.3.4 Ietekme

Domēna vārds Korpuss.lv sākotnēji tika reģistrēts 2007. gadā, lai sniegtu informāciju par *Mūsdienu latviešu valodas līdzsvarotā korpusa* pirmo versiju. 2018. gada

[2]http://www.sketchengine.eu

maijā tika izveidota jauna lapas versija un lapa tika pārveidota par korpusu kolekcijas indeksu, kurā sākotnēji bija desmit korpusu. 2022. gada novembrī platforma tika nosaukta par *Nacionālo korpusu kolekciju*. Pirmais korpuss CLARIN repozitorijā tika publicēts 2020. gada jūlijā, un citēšanas vadlīnijas tika ieviestas 2023. gada janvārī. Šobrīd NKK ietver 39 korpusus, no kuriem 29 ir publicēti CLARIN repozitorijā. Pēdējā gada laikā NKK platformu ir apmeklējuši 6 600 lietotāji un tie kopā apskatījuši 33 000 sadaļas.

Lai novērtētu NKK akadēmisko ietekmi, tika veikta sistemātiska analīze, izmantojot *Google Scholar* meklētājprogrammu[3]. Meklēšanas vaicājumi tika ievietoti pēdiņās, lai nodrošinātu precīzu atbilstību. Analīze atklāja, ka kopš 2020. gada Korpuss.lv ir citēts vairāk nekā 200 zinātniskajos darbos. Lai gan nosaukums *Nacionālā korpusu kolekcija/Latvian National Corpora Collection* ir salīdzinoši jauns, tas ir parādījies 18 angļu valodā publicētos darbos un 8 latviešu valodas publikācijās.

NKK resursus autoriem tiek ieteikts citēt, izmantojot atbilstošo publikāciju vai CLARIN datu atsauci, ja tāda ir pieejama, tomēr joprojām pastāv tiešas citēšanas gadījumi, izmantojot Korpuss.lv tīmekļa adreses. Konkrētāk, saites uz korpusu informācijas lapām vietnē Korpuss.lv ir minētas 37 zinātniskajos darbos, savukārt CLARIN saites ir citētas 81 darbā.

CLARIN saišu izmantošana ir īpaši ieteicama, jo korpusu kodi un nosaukumi latviešu vai angļu valodā bieži ir pārāk vispārīgi, tāpēc meklēšanas rezultāti var būt kļūdaini. Standartizētas atsauces uzlabo precizitāti un atvieglo resursu identificēšanu un citēšanu akadēmiskajos darbos.

[3] *Google Scholar* – https://scholar.google.com/

# 4. Secinājumi

Šajā darbā ir sniegts visaptverošs pētījums par latviešu valodas korpusu infrastruktūru un lietojumiem, risinot būtiskus izaicinājumus korpusu izveidē, marķēšanā un lietojumos. Izmantojot strukturētas metodes, digitālos rīkus un praktiskas lietojumprogrammas, šis pētījums ir būtiski veicinājis dabiskās valodas apstrādes (NLP) un korpusos balstītas pētniecības attīstību latviešu valodniecībā un digitālajās humanitārajās zinātnēs. Praktiski īstenojot un izvērtējot iegūtos rezultātus, ir apstiprinājušās pētījumā izvirzītās hipotēzes.

Pirmo hipotēzi, ka "automātiski marķēšanas rīki un strukturētas datu plūsmas samazinās korpusa izveidei nepieciešamo manuālo darbu, vienlaikus saglabājot augstu precizitāti", apstiprina plaši kvantitatīvi un kvalitatīvi novērtējumi. Piemēram, kļūdu marķēšanas procesā automatizētie moduļi sistemātiski identificēja un klasificēja pareizrakstības, leksikas, morfoloģiskās un interpunkcijas kļūdas, būtiski samazinot manuāli nepieciešamo darba apjomu. Veikto eksperimentu rezultāti uzrādīja konsekventi augstu marķēšanas precizitāti, galvenajos līmeņos pārsniedzot 90%. Tas apliecina, ka ieviestajās tehnoloģijās ir atrasts labs līdzsvars starp efektivitāti un uzticamību.

Turklāt ar šīs infrastruktūras palīdzību izveidotie un marķētie korpusi ir pierādījuši savu nozīmīgumu latviešu valodas NLP lietojumos. Pielāgoto NLP modeļu veiktspējas uzlabošanos apliecina tādu modeļu kā latviešu valodas BERT, runas sintēzes un specializēto automātiskās runas atpazīšanas (ASR) sistēmu veiksmīga ieviešana. Tas savukārt apstiprina otro hipotēzi, ka "marķētu korpusu pieejamība veicinās latviešu valodai pielāgotu NLP modeļu izstrādi un uzlabos to veiktspēju". Nodrošinot plašu un kvalitatīvu korpusu kopu, izveidotā infrastruktūra ir būtiski sekmējusi latviešu valodas NLP modeļu attīstību, kas palīdz nodrošināt, ka latviešu valoda pilnvērtīgi tiek izmantota digitālajā laikmetā.

Visbeidzot, trešo hipotēzi, ka "strukturēta un viegli pieejama latviešu valodas korpusu infrastruktūra būtiski veicinās korpusos balstītu pētniecību dažādās disciplīnās", apstiprina praktiski lietojumi digitālajās humanitārajās zinātnēs, politikas zinātnēs un datorlingvistikā. Brīvi pieejamas un vienotas latviešu valodas korpusu platformas izveide ir mazinājusi tehniskos šķēršļus korpusu izmantošanā, tāpēc pētnieki bez specializētām programmēšanas zināšanām var veikt padziļinātus korpusos balstītus pētījumus. Šo resursu veiksmīgā integrācija starpdisciplināros projektos un to citēšana vairāk nekā 200 zinātniskajos darbos apliecina infrastruktūras ilgtermiņa nozīmi un pierāda, ka lietotājiem ērts dizains un atvērti datu formāti paplašina korpusos balstītas pētniecības iespējas.

Pētījumā ir veiksmīgi sasniegts galvenais mērķis – integrēt latviešu valodas datus un NLP rīkus mūsdienu pētniecībā un lietojumos –, tādēļ promocijas darbs ir ne tikai stiprinājis latviešu valodas resursu izmantošanu zinātnē un valodas teh-

noloģijās, bet arī sniedzis jaunu izpratni par korpusu veidošanu, marķēšanu un izmantošanu plašākā kontekstā. Šis pētījums ne vien uzlabo latviešu valodas datu pieejamību un izmantojamību, bet ir arī pamats turpmākai NLP lietojumu un korpusos balstītās pētniecības metožu attīstībai.

# Literatūra


Alfaifi, A., Atwell, E., and Hedaya, I. (2014). Arabic learner corpus (alc) v2: a new written and spoken corpus of arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014*, volume 2, pages 77–89. Kobe International Communication Center.

Alumäe, T., Paats, A., Fridolin, I., and Meister, E. (2017). Implementation of a Radiology Speech Recognition System for Estonian Using Open Source Software. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2168–2172.

Auzina, I., Kaija, I., and Levane-Petrova, K. (2020). Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā. *Valoda: nozīme un forma*, 11:7–26.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of the 23rd INTERSPEECH Conference*, pages 2278–2282.

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., and Birch, A. (2017). Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ekgren, A., Gyllensten, A. C., Gogoulou, E., Heiman, A., Verlinden, S., Öhman, J., Carlsson, F., and Sahlgren, M. (2022). Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language resources and evaluation*, 46(1):131–142.

Erjavec, T., Kopp, M., Ljubesic, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Coltekin, C., Korzinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, S., Bartolini, R., Bel, N., Perez, M. C., Dargis, R., and e. al. (2024). Parlamint ii: advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubesic, N., Simov, K., Pancur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrimsson, S., Coltekin, C., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Perez, M. C., de Macedo, L., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevicius, V., Krilavicius, T., Dargis, R., Ring, O., van Heusden, R., Marx, M., and Fiser, D. (2023). The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Gilquin, G., De Cock, S., and Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Presses universitaires de Louvain (Louvain-La-Neuve).

Granfeldt, J., Nugues, P., Persson, E., Thulin, J., Ågren, M., and Schlyter, S. (2006). Cefle and direkt profil: A new computer learner corpus in french l2 and a system for grammatical profiling. In *LREC-2006, The fifth international conference on Language Resources and Evaluation*, pages 565–570. ELRA.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., et al. (2009). *International Corpus of Learner English*. Presses universitaires de Louvain Louvain-la-Neuve.

Gries, S. T. and Adelman, A. S. (2014). Subject Realization in Japanese Conversation by Native and Non-native Speakers: Exemplifying a New Paradigm for Learner Corpus Research. In *Yearbook of Corpus Linguistics and Pragmatics 2014*, pages 35–54. Springer.

Ismayilov, A., Kontokostas, D., Auer, S. r., Lehmann, J., and Hellmann, S. (2018). Wikidata through the eyes of dbpedia. *Semantic Web*, 9(4):493–503.

Kaija, I. and Auzina, I. (2020). Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection. In *Selected papers from the CLARIN Annual Conference 2019*, pages 41–47.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1:7–36.

Laizane, I. et al. (2018). The understanding of the concepts of first language, second language and foreign language outside of latvia. In *Rural Environment. Education. Personality.(REEP). Proceedings of the International Scientific Conference (Latvia)*, number 11. Latvia University of Life Sciences and Technologies.

Lemmens, M. and Perrez, J. (2010). On the use of posture verbs by French-speaking learners of Dutch: A corpus-based study. *Cognitive Linguistics*, 21(2).

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

Mendes, A., Antunes, S., Jansseen, M., and Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. In *Proceedings of the Tenth Language Resources and Evaluation Conference–LREC'16*, pages 3207–3214. European Language Resources Association.

Paats, A., Alumäe, T., Meister, E., and Fridolin, I. (2018). Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models. *Journal of Digital Imaging*, 31(5):615–621.

Paikens, P. (2016). Deep Neural Learning Approaches for Latvian Morphological Tagging. In *Human Language Technologies – The Baltic Perspective*, volume 289. IOS Press.

Paikens, P., Pretkalnina, L., and Rituma, L. (2024). A computational model of latvian morphology. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, page 221.

Paikens, P., Rituma, L., and Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 267–277, Oslo, Norway.

Pinnis, M., Auzina, I., and Goba, K. (2014). Designing the Latvian speech recognition corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1547–1553, Reykjavik, Iceland.

Pinnis, M., Salimbajevs, A., and Auzina, I. (2016). Designing a speech corpus for the development and evaluation of dictation systems in Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 775–780, Portoroz, Slovenia.

Plenert, A. (2021). gpt2-lv. In *Hugging Face*. https://huggingface.co/aidan-plenert-macdonald/gpt2-lv edition.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Press, G. (2021). Andrew Ng Launches A Campaign For Data-Centric AI. *Forbes*, Jun 16, 2021.

Pretkalnina, L., Rituma, L., and Saulite, B. (2016). Universal dependency treebank for latvian: A pilot. In *Human Language Technologies - The Baltic Perspective*, volume 289. IOS Press.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Šalme, A. (2011). *Latviešu valodas kā svešvalodas apguves pamatjautājumi*. Latviešu valodas aģentūra.

Tenfjord, K., Meurer, P., and Hofland, K. (2006). The ask corpus-a language learner corpus of norwegian as a second language. In *LREC*, volume 6, pages 1821–1824.

Ulcar, M., Zagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., and Robnik-Sikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *CoRR*, abs/2107.10614.

Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the european parliament as linked open data. *Semantic Web*, 8(2):271–281.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., et al. (2019). The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Wang, M., Malmasi, S., and Huang, M. (2015). The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Znotins, A. (2016). Word embeddings for latvian natural language processing tools. In *Human Language Technologies - The Baltic Perspective*, volume 289. IOS Press.

Znotins, A. and Barzdins, G. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.

# Resursi


Auziņa, I., Darģis, R., Bojārs, U., Paikens, P., Znotiņš, A., and Rābante-Buša, G. (2019). Lr 5.–13. saeimas sēžu stenogrammu korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/50.

Auziņa, I., Darģis, R., Levāne-Petrova, K., Auziņa, A., Saulīte, B., Ļaksa Timinska, I., Gailīte, E., Nešpore-Bērzkalne, G., Rābante-Buša, G., Pokratniece, K., and Klints, A. (2024a). Late plašsaziņas līdzekļu korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/114.

Auziņa, I., Darģis, R., Levāne-Petrova, K., Pokratniece, K., and Vēvere, D. (2018). Valsts valodas prasmes pārbaudes darbu korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/49.

Auziņa, I., Darģis, R., Rābante-Buša, G., Timinska-Ļaksa, I., Gailīte, E., and Auziņa, A. (2024b). Late sarunu korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/113.

Auziņa, I., Kaija, I., Levāne-Petrova, K., Pokratniece, K., and Darģis, R. (2021). Latviešu valodas apguvēju korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/42.

Auziņa, I., Saulīte, B., Akmane, A., Millere, E., Naļivaiko, I., Stepanovs, K., Darģis, R., and Grūzītis, N. (2022). Radioloģisko izmeklējumu transkripciju korpuss. CLARIN-LV digitālā bibliotēka. http://hdl.handle.net/20.500.12574/67.