10.35539/LTNC.2022.0047.I.A.R.D.B.S.N.G.M.G.A.S.K.S.244.263

Specializēta latviešu valodas runas korpusa un izrunas vārdnīcas izveide vizuālās diagnostikas izmeklējumu lingvistiskai analīzei un sistemātiskai transkribēšanai

Ilze Auziņa, Roberts Darģis, Baiba Saulīte, Normunds Grūzītis, Mikus Grasmanis, Andrejs Spektors, Kaspars Stepanovs

Šis darbs ir daļa no ERAF praktiskas ievirzes pētījumu projekta "Latviešu valodas runas atpazīšana un sintēze medicīnas lietojumiem" (Nr. 1.1.1.1/18/A/153) rezultātiem sinerģijā ar Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi" projektu Nr. VPP-IZM-DH-2020/1-0001.

Atslēgvārdi: valodas resursi, automātiskā runas atpazīšana, dabiskās valodas apstrāde, mašīnlasāma vārdnīca, medicīnas valoda

Ievads

Latvijā dažādās medicīnas nozarēs, t. sk. diagnostiskajā radioloģijā, plaši tiek izmantotas mūsdienīgas medicīnas tehnoloģijas, tomēr izmeklējumu rezultātu apraksti un epikrīzes vēl aizvien tiek sagatavotas pilnībā manuāli, izmantojot diktofonu centru pakalpojumus. Taču, pastāvīgi pieaugot diagnostisko izmeklējumu skaitam, ārstiem un pacientiem izmeklējumu rezultātu apraksti jāgaida pat vairākas dienas, ja vien rezultāti nav ļoti steidzami. Ir nepieciešams jauns un ilgtspējīgs risinājums un tehniskā infrastruktūra, kas nodrošinātu latviešu valodas runas tehnoloģiju integrēšanu medicīnā un ļautu atvieglot un paātrināt izmeklējumu rezultātu aprakstu ieguvi.

LU Matemātikas un informātikas institūtā (LU MII), sadarbojoties ar Rīgas Austrumu klīniskās universitātes slimnīcu (RAKUS), tiek īstenots Eiropas Reģionālās attīstības fonda praktiskas ievirzes pētījums "Latviešu valodas runas atpazīšana un sintēze medicīnas lietojumiem". Projekta mērķis ir radīt būtiskus latviešu valodas resursus (t. sk. specializētu latviešu valodas runas korpusu un izrunas vārdnīcu) un tehnoloģiju komponentes automātiskai runas atpazīšanai un runas sintēzei medicīnas jomā, kā arī demonstrēt šo resursu un tehnoloģiju izmantošanu, izstrādājot un aprobējot izmeklējumu diktēšanas un transkribēšanas platformu diagnostiskās radioloģijas vajadzībām.

Platforma būs pielāgota diviem lietojuma scenārijiem. Pirmajā scenārijā paredzēts, ka noteiktiem radioloģijas izmeklējumiem, piemēram, rentgenam un ultrasonogrāfijai, kuru apraksti ir relatīvi vienkārši un kuru diktātu automātiska atpazīšana ir precīzāka (salīdzinājumā ar, piemēram, datortomogrāfijas un magnētiskās rezonanses aprakstiem), ārsti izmantos diktēšanas platformu. Tajā pēc diktāta automātiskas atpazīšanas viņi paši tekstu varēs rediģēt un apstiprināt, tādējādi ātri iegūstot jau gatavu izmeklējuma aprakstu. Savukārt otrajā scenārijā paredzēts, ka diktofonu centra darbinieki pārskatīs un rediģēs automātiski atpazītus diktātus, nevis rakstīs visu no jauna.

Līdzīgi kā daudzās citās valodās, arī latviešu valodai pēc pietiekami liela vispārīga runas korpusa izveides (Pinnis et al. 2014) ir izstrādātas vairākas automātiskas runas atpazīšanas (*Automatic Speech Recognition*; turpmāk tekstā – ASR) sistēmas (Salimbajevs & Strigins 2015; Znotiņš et al. 2015). Tomēr šādas sistēmas nav izmantojamas medicīnā (t. sk. radioloģijā), jo tās ir paredzētas vispārīgam lietojumam un nespēj pietiekami precīzi atpazīt specifiskus tekstus, konkrēti – radioloģisko izmeklējumu un epikrīžu diktātus.

Lai varētu izstrādāt medicīnā izmantojamu ASR sistēmu, ir jāsagatavo specializēti valodu resursi un jāveic papildu priekšapstrādes un pēcapstrādes darbības:

- valodas modeļa izstrādei ir nepieciešams anonimizētu medicīnisku ziņojumu, galvenokārt radioloģisko izmeklējumu aprakstu un epikrīžu, tekstu korpuss (sk. 1. nodaļu);
- ortogrāfiski transkribēts un anonimizēts runas korpuss, kas nepieciešams ASR sistēmas novērtēšanai un teksta priekšapstrādes un pēcapstrādes likumu izstrādei (sk. 2. nodaļu);
- medicīnas terminu (t. sk. latīnisko), saīsinājumu, nosaukto entitāšu, kā arī vispārlietojamās leksikas izrunas vārdnīca (sk. 3. nodaļu);
- datu priekšapstrādes un pēcapstrādes likumi, kuri nodrošina, ka ziņojuma tekstā automātiski tiek izvērsti saīsinājumi, cipari tiek pierakstīti ar vārdiem u. tml. un automātiski transkribētajā tekstā tiek iekļauti atbilstoši apzīmējumi, saīsinājumi u. tml. (sk. 4. nodaļu).

Akustiskā modeļa pielāgošana medicīnas tekstiem var nedaudz ietekmēt ASR transkripciju precizitāti, tomēr ASR precizitāti galvenokārt ietekmē pielāgotais valodas modelis un specializētā izrunas vārdnīca (Blackley et al. 2019).

Šajā pētījumā specializēto korpusu un izrunas vārdnīcas izejas materiāls ir anonimizēti RAKUS radioloģijas izmeklējumu arhīva dati: audioieraksti (radioloģisko izmeklējumu un epikrīžu audio ieraksti) un teksti (radioloģisko izmeklējumu un epikrīžu apraksti).

Teksta korpusa izveide

Lai izstrādātu medicīnas jomas korpusus, kas izmantoti par pamatu tālākiem pētījumiem un sarežģītāku resursu izstrādei, šajā projektā pielāgotas esošās datu kopas. Korpusos iekļauti dati no lielākās slimnīcas Latvijā, t. i., RAKUS, radioloģijas izmeklējumu arhīva.

Medicīnas valodu reprezentējošais teksta korpuss veidots, izmantojot anonimizētus izmeklējumu, operāciju un epikrīžu aprakstus no RAKUS arhīva. Teksta korpusa izveide ietver trīs galvenos soļus:

1. teksta izguve;
2. teksta anonimizēšana;
3. teksta normalizēšana.

Šobrīd runas un teksta datu plūsma RAKUS ir šāda:

- radiologi, ārsti un viņu palīgi diktē izmeklējumu vai epikrīzi, izmantojot dažādas ierīces un kanālus, t. sk. telefona zvanus;
- diktāta audioieraksti tiek iesniegti diktofonu centrā;
- audioieraksti tiek manuāli pārrakstīti, izveidojot *Microsoft Word* dokumentus, kuru pamatā ir iepriekš definētas dokumentu veidnes;
- sagatavotie dokumenti tiek nosūtīti ārstiem apstiprināšanai;
- apstiprinātie dokumenti tiek pievienoti arhīvam.

Lai izvairītos no sensitīvu datu nonākšanas ārpus ārstniecības iestādes, pirmie divi soļi veikti RAKUS IT infrastruktūrā. Teksta segmenti izgūti no vairāk nekā 100 tūkstošiem dažādu veidu medicīnisko izrakstu, kas RAKUS tikuši sagatavoti pēdējo astoņu gadu laikā. Sagatavoto izrakstu skaits katru gadu pieaug un šobrīd sasniedz ap 15 tūkstošiem izrakstu gadā.

Teksta izguve

Ārstniecības iestādēs izveidotajos *MS Word* dokumentos bieži ir vairākas savstarpēji pakārtotas tabulas un dažādi teksta formatējuma elementi, piemēram, treknraksts, slīpraksts, pasvītrojums. Teksta izguvē ir svarīgi saglabāt teksta dalījumu teikumos un rindkopās, jo vairāki teksta analīzes rīki (piemēram, morfoloģiskās marķēšanas un sintaktiskās analīzes rīki) paļaujas uz korektu teksta dalījumu teikumos un nekorekts dalījums var būtiski ietekmēt šo rīku precizitāti. Dalījums teikumos un rindkopās tiek izmantots, arī lai apmācītu valodas modeli, kas automātiski ģenerētās diktātu transkripcijas sadala teikumos un rindkopās.

Jāņem vērā arī tas, ka medicīniskajos izrakstos ir dažādi īsi segmenti bez pieturzīmēm, piemēram, sadaļu virsraksti, tabulu, kolonnu, rindu un citu teksta lauku nosaukumi. Ja visu tekstu apvienotu vienā un dalītu teikumos tikai pēc pieturzīmēm, šādi īsie segmenti nekorekti tiktu pievienoti teikumiem, kas ir aiz tiem. Pretēja rakstura problēma rastos, ja katru teksta segmentu uzskatītu par atsevišķu rindkopu. Šādā gadījumā īsie teksta segmenti, t. sk. specifiski noformētie segmenti (treknraksts, kursīvs u. c.), tiktu uzskatīti par atsevišķām rindkopām un daudzi teikumi tiktu sadalīti vairākās neloģiskās, īsās rindkopās.

Teksta izguvei no RAKUS izrakstu arhīva tika izstrādāts pielāgots rīks, kas segmentus izgūst, balstoties uz dokumenta iekšējo XML datu struktūru (Microsoft Word Open XML jeb DOCX formāts, kas RAKUS izrakstos tiek lietots kopš 2010. gada).

Teksta anonimizēšana

Nākamais būtiskais solis ir izgūtā teksta anonimizēšana, lai izvairītos no personas datu (vārds, uzvārds, personas kods, adrese u. c.) iekļaušanas korpusā, pat ja tam nav plānota publiska

piekļuve. Kaut gan teksts ir automātiski anonimizēts, šis korpuss joprojām tiek uzskatīts par potenciāli sensitīviem datiem, jo korpusa lielā apjoma dēļ to nav iespējams manuāli caurskatīt un pārbaudīt.

Izgūtais teksta apjoms ir pietiekami liels tālākajiem pētījumiem (medicīnas valodas modelēšanai), tāpēc teksta anonimizēšana tika izvirzīta par augstāku prioritāti nekā teksta saglabāšana, proti, ja bija vismazākās šaubas, ka teksta segments varētu saturēt personas datus, tas tika izņemts no korpusa, pat ja pastāvēja liela varbūtība, ka šajā segmentā personas datu, visticamāk, nemaz nav.

Teksts tika anonimizēts segmentu (aptuveni: rindkopu) līmenī. Ja segmentā konstatēts kaut viens vārds, kas varētu norādīt uz potenciāliem personas datiem, viss segments tika izņemts no teksta korpusa. Papildus tam pēc tekstu normalizēšanas (skat. nākamo sadaļu) visas atlikušās rindkopas teksta korpusā ir sadalītas teikumos, bet teikumi ir sakārtoti alfabētiskā secībā. Tādējādi ir pilnībā sajaukta rindkopu un teikumu sākotnējā secība: pat ja kāds izteikums satur kādus personas datus, tiem nav pieejams plašāks konteksts.

Teksta normalizēšana

Tekstu korpusa normalizēšana notiek vairākos posmos. Vispirms automātiski tiek labotas tipiskās drukas un pareizrakstības kļūdas. Pirmajā korpusa apstrādes iterācijā šādu kļūdu labošana vēl nebija iespējama, bet pēc pirmās iterācijas tika izgūts tekstu korpusa vārdformu biežumsaraksts un biežāk lietotās vārdformas, par kuru pareizrakstību un izrunu nebija iespējams pārliecināties automātiski (sk. 3. nodaļu), un tās tika caurskatītas manuāli, tostarp kļūdainajām vārdformām norādot korektu rakstību (sk. 3. nodaļu). Rakstības normalizēšana ilustrēta ar diviem ļoti izplatītiem piemēriem RAKUS korpusā:

1) termins “dupleksdoplerogrāfija” izrakstos tiek lietots visdažādākajos rakstības variantos, t. sk. ar pareizrakstības kļūdām: “dupleks-doplerogrāfija”, “dupleks doplerogrāfija”, “duplekss doplerogrāfija”, “duplex-doplerogrāfija”, “duplex – doplerogrāfija” utt. – tie visi (visās formās) tika automātiski normalizēti kā “dupleksdoplerogrāfija” (attiecīgajās formās);
2) savukārt termins “folikuls” tiek bieži lietots ar vienu vai divām pareizrakstības kļūdām: “folīkuls”, “folikuļi”, “folīkuļi” utt. – kļūdainās formas korpusā tika automātiski izlabotas.

Nākamais solis pēc dažāda veida rakstības un segmentācijas kļūdu labošanas un teksta normalizēšanas ir teksta automātiska sadalīšana teikumos un tekstvienībās (vārdformas, akronīmi un saīsinājumi, simboli un apzīmējumi, interpunkcija). Strukturāli sadalītajam tekstam tika veikta morfoloģiskā analīze, izmantojot LU MII atvērtā pirmkoda rīkkopu NLP-PIPE (Paikens et al. 2013; Znotiņš & Cīrule 2018). Tādējādi tekstu korpuss ir papildināts ar informāciju par vārdlietojumu pamatformām un morfoloģiskajām pazīmēm, kas ir noderīgi korpusa tālākā lingvistiskajā analīzē, piemēram, pamatformu pārbaudei un lietojuma biežuma salīdzināšanai citos latviešu valodas korpusos un leksiskas resursos.

Trešais būtiskais normalizēšanas apakšsolis ir tekstu korpusa automātiska izvēršana (verbalizēšana): skaitļu, datumu, dažādu saīsinājumu un apzīmējumu pārrakstīšana pilnos vārdos, kontekstuāli saskaņojot locījumus. Papildus tam gadījuma vietās tiek verbalizētas arī pieturzīmes (piemēram, komats tiek aizvietots ar "komats", punkts – ar "punkts" vai "nākamajā teikumā") un strukturālais noformējums (piemēram, rindkopu pārnesumi – "jaunā rindā" u. tml.; sadaļu virsraksti – "secinājumos rakstām", "slēdzienā [lūdzu] rakstām" u. tml.). Tādējādi izvērstajā tekstu korpusā tiek reprezentētas arī tipiskās instrukcijas un to biežāk lietotie, alternatīvie verbalizācijas varianti, ko ārsti līdzšinējā darbplūsmā mēdz norādīt diktofonu centra operatoriem. Tas tiek darīts, lai teksta korpuss būtu pēc iespējas līdzīgāks runas korpusam (sk. 2. nodaļu) un lai runas atpazīšanas sistēmas valodas modelis pēc iespējas labāk reprezentētu to, kā apraksti un izraksti tiek diktēti. Šis apakšsolis un tam pretējais virziens (transkripciju automātiska savēršana) detalizētāk ir aprakstīts 4. nodaļā.

Teksta normalizēšanai un tipisko kļūdu novēršanai ir būtiska loma turpmākajā valodas statistiskā modeļa izveidē un izmantošanā. Lai gan visas pareizrakstības kļūdas izlabot nav iespējams, svarīgi ir novērst tipiskās. Tā kā RAKUS korpuss ir ļoti apjomīgs (vairāk nekā 100 milj. tekstvienību), manuāli analizētas un automātiski labotas tikai tās kļūdas un nekonsekvences, kuru biežums korpusā ir vismaz 1000 (t. i., relatīvais biežums vismaz 0,00001). Par retākiem gadījumiem: tiek pieņemts, ka valodas modelim, kas runas atpazīšanas sistēmai palīdz ģenerēt diktātu automātiskās transkripcijas, būs tendence izvēlēties pareizākus rakstības variantus, t. i., tiek pieņemts, ka normalizētajā korpusā pastāv spēcīga korelācija starp statistiski biežākiem un leksiski pareizākiem vārdlietojumu kontekstiem (n-grammām). Lai gan praksē tā ne vienmēr notiek, sistēmas prototipa testēšanā iesaistītie ārsti norāda, ka automātiskajās transkripcijās bieži vien vārdi tiek pierakstīti pareizāk, nekā viņi tos ir diktējuši.

Runas korpusa izveide

Lielais teksta korpusa apjoms ļoti labi atspoguļo rakstītā teksta dažādību. Diemžēl ar to vien ir par maz, lai izveidotu izrunas vārdnīcas. Tekstā ir sastopami daudzi svešvārdi un saīsinājumi, kuru izruna nav viennozīmīga. Lai sīkāk varētu pētīt izrunas nianses, ir nepieciešams runas korpuss, tāpēc projektā tika izstrādāts 30 stundu ortogrāfiski transkribēts runas korpuss.

Runas korpusa izstrāde sastāv no vairākiem soļiem:

1. runas datu atlase;
2. runas datu automatizēta apstrāde un anonimizēšana;
3. automātiski atpazītā teksta pārskatīšana, rediģēšana un marķēšana.

Runas datu atlase

Runas dati tika ņemti no pēdējo sešu mēnešu diktofonu centra arhīva. Lai nodrošinātu korpusa kvalitāti, tajā netika iekļauti ieraksti ar zemu kvalitāti. Korpusā netika iekļauti diktāti, kas veikti pa telefonu, ieraksti ar skaļu fona troksni vai ieraksti, kurus ierunājis cilvēks ar spēcīgu akcentu un daudzām izrunas kļūdām.

Lai nodrošinātu runas korpusa reprezentativitāti, proporcionāli tika atlasīti dažādu runātāju ieraksti. Korpusā tika iekļauti ieraksti no 71 runātāja, atlasot ierakstus vienmērīgi no visa perioda.

Automātiska runas datu apstrāde un anonimizēšana

Lai paātrinātu runas korpusa izveidi, no izrakstiem izgūtais teksts tika automātiski sastatīts ar ierakstu. Automātiskais sastatījums nav precīzs, tāpēc automātiski atpazītais teksts tika pārskatīts un manuāli rediģēts.

Audiodatu anonimizēšana ir konceptuāli un tehniski sarežģītāka nekā teksta anonimizēšana. Lai anonimizētu audiofailu (diktātu), vispirms būtu jāiegūst tā automātiska transkripcija un jānosaka, kuros laika intervālos audiofailā ir pieminēti personas dati, attiecīgos segmentus no audiofaila pēc tam izgriežot. Tomēr audiodatu anonimizēšanā nedrīkstam paļauties uz automātiskām transkripcijām, kas ir salīdzinoši neprecīzas tieši attiecībā uz personas datiem (jo šāda veida dati ir apzināti izņemti no valodas modeļa un izrunas vārdnīcas; sk. 1. un 3. nodaļu). Šāda pieeja – anonimizēt tikai atsevišķus, īsus audio segmentus, balstoties automātiskajās transkripcijās, – nenodrošinātu pietiekamu anonimizēšanas precizitāti.

Vairumā gadījumu personas dati tiek diktēti audioieraksta sākumā, tāpēc tika izmantota vienkāršāka un anonimizēšanas ziņā uzticamāka metode, lai arī šādā veidā audiofails bieži vien tiek apgriezts vairāk, nekā tas būtu nepieciešams. Automātisko transkripciju vietā tika izmantoti secīgi teksta segmenti no anonimizētā, bet alfabētiski nesajauktā tekstu korpusa (sk. 2.1. nodaļu). Izmantojot anonimizētā teksta un audioieraksta automātisku sastatījumu, atlasītajiem audiofailiem automātiski tika nogriezts viss nesastatītais sākums un beigas.

Manuāla runas datu transkribēšana un marķēšana

Radioloģisko izmeklējumu un epikrīžu apraksti tiek sastatīti ar atbilstošu audio ierakstu, kuram, izmantojot ASR sistēmu, tiek piedāvāts aptuvens atšifrējums. Klausoties diktāta ierakstu, diktāta transkripcija tiek pierakstīta atbilstoši izstrādātajām vadlīnijām. Diktāta transkribēšanas režīmā tiek pārbaudīts, vai transkripcijā ir pierakstīts viss, kas diktātā tiek teikts, un vai tiek pierakstīts tā, kā tas tiek pateikts (sk. 1. attēlu).

Runas korpusā visi automātiski atpazītie teksti ir pārskatīti un pierakstīti t. s. ortogrāfiskajā transkripcijā. Ortogrāfiskā transkripcija ir burtiska sacītā atveide rakstos mašīnlasāmā formā,

1. attēls. Diktāta transkribēšanas saskarne.

ievērojot valodas ortogrāfijas principus (Oostdijk & Boves 2008: 644; Schiel & Draxler 2004; Goedertier et al. 2000: 909). Atveidojot runātās valodas piemērus rakstu formā, tiek saglabāts runas fakta autentiskums, t. i., tiek rakstīts tieši tā, kā tiek teikts, paturot visus vārdu atkārtojumus, ja nepieciešams, parādot arī vārdu izrunas atkāpes no literārās valodas normas. Teksti tiek pierakstīti ortogrāfiskajā transkripcijā, izmantojot korpusa transkribēšanas vadlīnijas, kas balstītas uz iepriekšējo pieredzi (Pinnis et al. 2014) un papildinātas atbilstoši specializētā runas korpusa vajadzībām, piemēram, norādot atbilstošu terminu, saīsinājumu izrunu. Transkripcijā netiek norādītas pauzes – klusuma pauzes un aizpildītās pauzes (*āā, ēē, em khm* u. tml.), tāpat netiek norādīti neverbālie elementi, piemēram, ieelpa un izelpa, cilvēka radīts fizioloģisks troksnis – mēles klakšķināšana, šņaukāšanās, čāpstināšana u. tml. Šī informācija jau ir uzkrāta, izmantojot "Latviešu valodas runas atpazīšanas korpusu" (Pinnis et al. 2014), un tiek izmantota ASR sistēmu izstrādē.

Ortogrāfiski marķējot runas audiosignālu, tiek ievēroti vairāki ortogrāfiskās transkripcijas pamatprincipi: (1) runātais tiek pierakstīts burtiski, norādot arī vārdu atkārtojumus, pārteikšanos u. tml.; (2) pēc iespējas tiek ievērota interpunkcija un pareizrakstība; (3) norādītas būtiskas atkāpes no pareizrunas normām, piemēram, *Aknu izmērs četrdesmit [čēesmit] milimetri, tālāk rakstām [rakstam]*; (4) visi vārdi, ja vien tie nav īpašvārdi vai akronīmi, tiek rakstīti ar mazajiem burtiem; (5) cipari, saīsinājumi tiek rakstīti ar vārdiem atbilstoši to izrunai audioierakstā.

Nekorekta izruna

Nekorektai vārdu izrunai var būt dažādi iemesli, piemēram, latviešu valoda nav runātāja dzimtā valoda, runā vērojama latviešu valodas izlokšņu ietekme. Tas var būt saistīts arī ar īpatnēju atsevišķu vārdu izrunu, kas raksturīga katram indivīdam. Samērā bieži neprecīzi tiek izrunāti skaitļa vārdi, piemēram, *četrdesmit [čēesmit], piecdesmit [pieesmit], divi [div]*. Atkāpes no literārās valodas normām vērojamas arī patskaņu kvantitātē, t. sk. noteiktās darbības vārdu formās, piemēram *rakstām [rakstam]*, kā arī patskaņu un līdzskaņu kvalitātē, piemēram, *dorsāls [dorzāls], patoloģija [patologija].*

Saīsinājumi un svešvārdi

Ja vārds tiek izrunāts kā saīsinājums (t. sk. mērvienība) vai svešvārds (latīnisms, zāļu nosaukums u. tml.), tad tā izruna tiek likta kvadrātiekavās, piemēram:

USG [ū es gā] veikta trešajā martā
RTG [er tē gā] gramma krūškurvim taisnā projekcijā
Spinālā kanāla salīdzinoša stenoze C3 [cē trīs] C7 [cē septiņi] zonā –
uzskatāmāk C6 [cē seši] C7 [cē septiņi] dextra [dekstra].

Ja terminu vai nosaukumu veido vairāki vārdi, izruna tiek norādīta aiz katra vārda, kuram tiek precizēta izruna, piemēram:

Lietot Deferasirox [deferasiroks] Mylan [milan] reizi dienā
..truncus [trunkus] coeliacus [celiakus] līmenī..
distālais gals vena cava [kava] superior [supērior] projekcijā

Ja mērvienību apzīmējumus (piemēram, *cm, s, g, mg, mm, diam.*) izrunā kā pilnu vārdu, tad raksta pilnu vārdu, piemēram:

Necelt smagumus vairāk par pieciem kilogramiem aptuveni mēnesi.
Necelt smagumus vairāk par pieci kilogrami aptuveni mēnesi.
..postinfarkta zona ar diametru divdesmit divi reiz piecpadsmit milimetri.

Ja svešvārds vai saīsinājums tiek izrunāts tieši tā, kā pierakstīts, izrunu kvadrātiekavās nepievieno, piemēram, *luteum, TUR*.

Matemātiskie simboli

Arī matemātiskie simboli tiek rakstīti vārdiem, piemēram:

sin > dxt → sinistra vairāk kā dextra [dekstra]

Xefo 8 mg x 1 pēc vajadzības → Xefo [ksefo] astoņi miligrami reiz viens pēc vajadzības

Cipari un datumi

Ciparus (arī decimāldaļskaitļus, daļskaitļus) raksta ar vārdiem tā, kā tie tiek izrunāti:

centrālajā zonā ar kopējo apjomu LL [el el] seši komats septiņi centimetri, AP [ā pē] pieci komats četri centimetri un CC [cē cē] septiņi cm [cē em]

Arī datums tiek rakstīts tā, kā tas tiek izrunāts:

Pacientam plānots konsīlijs nulle pirmajā nulle otrajā divtūkstoš deviņpadsmitajā.
Pacientam plānots konsīlijs divtūkstoš deviņpadsmitā gada pirmajā februārī.
Izmeklējums veikts devītā pirmā divtūkstoš divdesmitā gadā.

Speciāli simboli

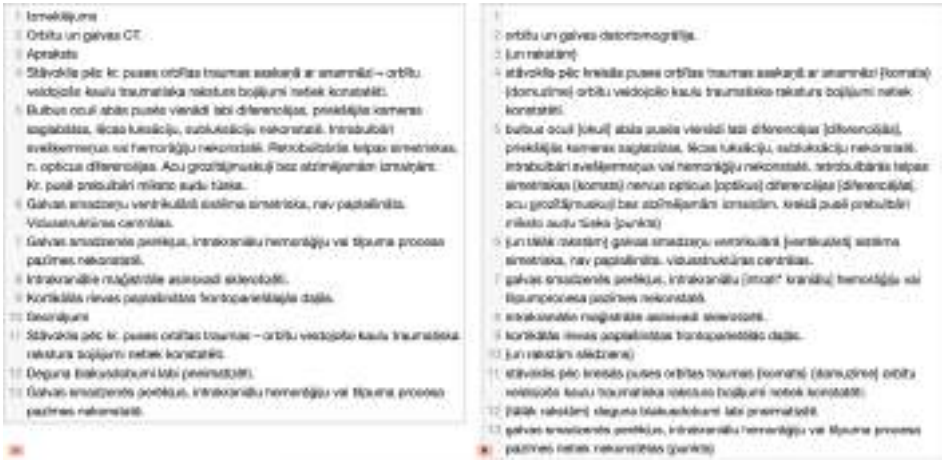
Speciāli simboli tiek izmantoti, lai parādītu aprautus vārdus un nesaprotamu runu, stostīšanos. Ja vārdam netiek izrunātas beigas, tā vietā tiek likta zvaigznīte. Pareizā forma šajos gadījumos netiek pierakstīta, piemēram, *Lab* kreisajā pusē neliels izaugums.* Nesaprotama pārteikšanās un stostīšanās, ko nevar transkribēt kā vārdu(s) vai aprautu(s) vārdu(s), tiek atzīmēta ar restīti, piemēram, *pirmajā divtūkstoš divdesmitajā # {kur izmeklējums uzrakstām}*.

Formatēšanas komandas

Transkripcijā tiek norādītas arī teksta formatēšanas komandas, kas tiek marķētas, pamatojoties uz iepriekšējo darbu pie vispārējas nozīmes latviešu valodas diktēšanas korpusa izstrādes (Pinnis et al. 2016). Ārstu norādes par formatējumu, apraksta sākumu, gaitu vai beigām ir liktas figūriekavās, piemēram, *{nākamajā rindā lūdzu rakstām}, {komats}, {rakstām tā}, {tas arī viss}, {paldies}, {paldies par šo visu}, {pirms slēdziena jāpieraksta, ka}, {pie slēdziena pierakstām}*.

2. attēlā parādīti sastatīti teksta un runas korpusa fragmenti: kreisajā pusē (A) dots teksta fragments no izmeklējumu arhīva, savukārt labajā pusē (B) – ortogrāfiski transkribēts diktāta fragments. Ortogrāfiskajā transkripcijā ir izvērsti saīsinājumi (*CT → datortomogrāfija, Kr. → kreisā* u. c.), norādītas formatēšanas komandas (*{un rakstām}, {komats}, {domuzīme}, {un tālāk rakstām}* u. c.) un precizēta terminu un vārdformu izruna (*oculi [okuli], opticus [optikus], diferencējas [diferencējās]* u. c.).

Ortogrāfiskā transkripcija tiek izmantota, lai kvantitatīvi novērtētu runas atpazīšanas sistēmas precizitāti, izstrādātu medicīnas jomai pielāgotāku akustisko modeli, papildinātu izrunas vārdnīcu ar vārdu, akronīmu, simbolu u. c. tekstvienību izrunas alternatīvām. Runas



2. attēls. Sastatīti references teksta
(A) un runas korpusa (B) fragmenti.

korpuss ļauj novērtēt, cik precīzi darbojas visi ASR komponenti: akustiskais modelis, izrunas vārdnīca un valodas modelis.

Diktātu kategorizēšana

Atšifrējot diktātus, par katru audioierakstu tiek norādīta papildu informācija – izmeklējuma veids (modalitāte) un radioloģijas nozare (sk. 1. tabulu). Šobrīd korpusā iekļauti 900 dažādu veidu izmeklējumu apraksti: datortomogrāfijas izmeklējumi (CT), magnētiskās rezonanses izmeklējumi (MR), mamogrāfijas izmeklējumi (MG), rentgenogrammu apraksti (CR) un ultrasonogrāfijas izmeklējumi (US), kuru sagatavošanai tiek bieži izmantots RAKUS diktofonu centrs. Pašlaik dati tiek kategorizēti manuāli, un šo informāciju plānots izmantot, lai apmācītu automātisku izmeklējuma modalitāšu un nozaru klasifikatoru, kā arī lai noteiktu, kura izmeklējumos lietotā leksika ir kopīga visām modalitātēm un radioloģijas nozarēm, bet kura – specifiski katrai no modalitātēm un nozarēm. Tas ļaus arī papildināt izrunas vārdnīcu ar informāciju par vārdu lietojumu.

Modalitāte		Biežums	Nozare	Biežums
CT	Datortomogrāfija	489	Onkoloģiskā radioloģija	425
MR	Magnētiskā rezonanse	129	Torakālā radioloģija	130

MG	Mamogrāfija	113	Krūts radioloģija	117
CR	Rentgenogramma	109	Neiroradioloģija	67
US	Ultrasonogrāfija	60	Sirds un kardiovaskulārā radioloģija	59
			Gastrointestinālā un abdominālā radioloģija	54
			Uroģenitālā radioloģija	22
			Muskuloskeletālā radioloģija	15
			Galvas un kakla radioloģija	9
			Neatliekamā radioloģija	2

1. tabula. Modalitāšu un nozaru
īpatsvars (reprezentācija) runas korpusā.

Izrunas vārdnīca

Pēc teksta izvēršanas un automātiski iegūtās runas transkripcijas pārskatīšanas nākamais solis ir izrunas vārdnīcas izveide. Izrunas vārdnīcu veido tekstvienību (vārdu, saīsinājumu un abreviatūru u. tml.) saraksts un norādes par to izrunu, kas runas atpazīšanas ASR sistēmai būtu jāatpazīst, kā arī automātiski ģenerēta fonētiskā transkripcija. Tā drīzāk ir datu kopa, nevis izrunas vārdnīca tradicionālā izpratnē, kur katrai vārdnīcā iekļautajai vienībai ir pievienota minimāla gramatiskā informācija un fonētiskā transkripcija.

Izrunas vārdnīcā iekļaujamo vārdu saraksts tiek iegūts no tekstu korpusa: teksti tiek marķēti un no tiem tiek izgūtas visas unikālās tekstvienības. Sarakstā iekļautās tekstvienības var iedalīt četrās grupās:

- latviešu valodas vārdi, kas tiek izrunāti atbilstoši latviešu valodas pareizrunas normām (lielākā daļa vārdu),
- kļūdaini uzrakstīti vārdi, kurus nevajadzētu iekļaut izrunas vārdnīcā,
- saīsinājumi un simboli, kas jāizvērš, norādot to izrunu vai lasījumu,
- vārdi citās valodās (ne latviešu), piemēram, medicīnas termini latīņu valodā, zāļu nosaukumi, kuru izruna latviešu valodā atšķiras no rakstības – tiem jānodrošina fonētiskā transkripcija, izmantojot latviešu valodas fonēmas.

Kopumā specializētajā tekstu korpusā ir 1,8 miljoni tekstvienību, no kurām aptuveni 1,1 miljonam tekstvienību ir vismaz viens burts, tādēļ tās var uzskatīt par izrunas vārdnīcā iekļaujamām vienībām. Manuāla visu vārdu iedalīšana iepriekš minētajās četrās grupās būtu ļoti laikietilpīgs uzdevums, tāpēc tika izmantoti vairāki latviešu valodas resursi un rīki, lai sagrupētu lielāko daļu vārdu.

Vispirms vārdi, kas bija atrodami lielākajā tiešsaistes latviešu valodas vārdnīcā *Tēzaurs.lv* (Spektors et al. 2016), tika atzīmēti kā standartvalodas vārdi. Vārdu locījumi tika atpazīti tikai tiem vārdiem, kuriem *Tēzaurs.lv* ir norādītas locīšanas paradigmas. Specializētajā medicīnas jomas korpusā bija sastopami arī tādi vārdi, kuri ir *Tēzaurs.lv*, bet kuriem nav norādītas locīšanas paradigmas. Lai varētu automātiski atpazīt vēl vairāk vārdu, šiem vārdiem locīšanas paradigmas tika pievienotas manuāli. Šī informācija savukārt tika integrēta *Tēzaurs.lv*.

Lai atpazītu vārdus un frāzes, kas, iespējams, ir personas dati, ir izmantots atvērtā pirmkoda latviešu valodas nosaukto entitāšu atpazinējs – daļa no jau pieminētās NLP-PIPE rīkkopas (Znotiņš & Cīrule 2018). Īpašvārdi, kas uzskatāmi par personas datiem (vārds, uzvārds, adrese), tiek pusautomātiski atlasīti un izmantoti kā papildu filtrs teksta anonimizēšanā.

Vārdi, kas netika klasificēti automātiski, bija jāpārskata un jāsagrupē manuāli. Tekstvienības, kas teksta korpusā sastopamas vismaz 1000 reižu (aptuveni 14 tūkst.), tika atlasītas manuālai pārskatīšanai. Vārdi, kuru nav *Tēzaurs.lv* un kuri manuāli tika atzīmēti kā standartvalodas vārdi, ir pievienoti kandidātu sarakstam iekļaušanai *Tēzaurs.lv*. To vidū ir arī diktātos lietoto standartvalodas vārdu abreviatūras, piemēram, vārda *ultrasonogrāfija* abreviatūra USG, vārdu savienojuma *magnētiskā rezonanse* abreviatūra MG, vārda *rentgenogramma* abreviatūra RTG.

Atlasītajām tekstvienībām izrunas vārdnīcā tiek norādīti izrunas varianti, ja tie atšķiras no vispārpieņemtajām izrunas normām, vai tekstvienību (gk. abreviatūru vai citvalodu vārdu) lasījums. Manuāli pārskatot tekstvienību sarakstu, tiek norādīts, 1) vai vārds ir svešvalodā (pazīme – [svešvaloda], 2) vai vārds ir īpašvārds (pazīme – [personas dati]), 3) vai tas ir saīsinājums, abreviatūra vai matemātisks simbols (pazīme – [saīsinājums]), 4) vai vārds ir kļūdaini uzrakstīts (pazīme – [drukas kļūda]. Ja sarakstā iekļautajai tekstvienībai bez plašākas apkaimes nav iespējams norādīt izrunu vai saprast, vai tas ir saīsinājums, īpašvārds u. tml., tiek pievienota pazīme [konteksts], piemēram, *l, e, g, pac.*

Pazīme [personas dati] pievienota tad, ja tekstvienība ir īpašvārds – vārds, uzvārds, apdzīvotas vietas nosaukums u. tml., kas savukārt liek pārbaudīt, vai tā nav daļa no personas datiem, kam korpusā nevajadzētu parādīties.

Ja tekstvienībai ir pievienota pazīme [saīsinājums], tad parasti tiek norādīts gan lasījums/izruna, gan izvērstā forma, piemēram, *PVN* [pē vē en] – *pievienotās vērtības nodoklis*, *RAKUS* [rakus] – *Rīgas Austrumu klīniskās universitātes slimnīca*, *EKG* [ē kā gā] – *elektrokardiogramma*.

Vārdnīcā netiek izmantota fonētiskā transkripcija. Fonētiskā izruna starptautiskajā fonētiskajā alfabētā (IPA) tiek automātiski ģenerēta, izmantojot Tēzaurs API[1].

Izrunas vārdnīca un informācija par tekstvienību biežumu tiek izmantota valodas modeļa izstrādē. Reti sastopami vai kļūdaini vārdi un tekstvienības, kas potenciāli ir daļa no personas datiem, netiek iekļauti valodas modelī.

1 https://api.tezaurs.lv

Tekstvienība	Tips	Izruna / lasījums / tekstvienība bez drukas kļūdām	Izvērsta forma
ķīmijterapija	parasts vārds		
Spirix	svešvaloda	spiriks	
EKG	saīsinājums	ē kā gā	elektrokardiogramma
RAKUS	saīsinājums	rakus	rīgas austrumu klīniskās universitātes slimnīca
Neu	nepieciešams konteksts		
mkmol	saīsinājums	mikromoli	
arstēšanas	drukas kļūda	ārstēšanas	
Kalniņa	personas dati		
g	nepieciešams konteksts		
Valmiera	personas dati		
Mydocalm	svešvaloda	midokalm	

2. tabula. Izrunas vārdnīcas fragments.

Teksta izvēršanas un savēršanas gramatika

Tekstu korpusa (anonimizētā RAKUS arhīva) izvēršanai un transkripciju savēršanai tiek izmantota līdzīga pieeja, kāda tika izmantota radioloģijas izmeklējumu diktātu transkribēšanas sistēmas izveidē igauņu valodai (Paats et al. 2018) – kontekstuāli formālas un skaitļojamas gramatikas likumi. Gramatikas likumi tika definēti, izmantojot paralēlos tekstus, kas tika iegūti runas korpusa izveides rezultātā: oriģinālie apraksti un tiem atbilstošās segmentu līmenī sastatītās ortogrāfiskās transkripcijas. Valodas modelis automātiskajai runas transkribēšanai tiek apmācīts uz izvērstā (verbalizētā) tekstu korpusa datiem, savukārt automātiskās transkripcijas tiek automātiski savērstas (deverbalizētas), izmantojot pretējus gramatikas likumus un tādējādi būtiski samazinot transkripciju pēcrediģēšanai nepieciešamo darba apjomu. Gramatikas likumu principi parādīti 3. attēlā.

```
1 CT krūškurvja   orgāniem ,
2 CT krušu kurvja organiem komats
3 CT krūšu kurvja orgāniem ,

1 pēc Ultravist 300      - 100  ml         i/v          ievades
2 pēc Ultravist trīsimt    simt mililitru intravenozas ievades
3 pēc Ultravist 300        100  ml         i/v          ievades

1 nieres konkrements 1,89                             cm         / diam.
2 nieres konkrements viens komats astoņdesmit deviņi centimetri   diametrā
3 nieres konkrements 1,89                             cm         / Ø
```

3. attēls. Fragmenti no sastatītā teksta un runas korpusa, kas ilustrē izvēršanas un savēršanas gramatikas lomu. Rindu apzīmējumi: 1 – fragments no teksta korpusa (references teksts); 2 – atbilstošais ortogrāfiskās transkripcijas fragments no runas korpusa (izvēršanas gramatikas mērķis); 3 – sagaidāmais automātiskās transkribēšanas sistēmas galarezultāts (savēršanas gramatikas mērķis).

4. attēls. Izmeklējumu diktātu automātiskās transkribēšanas un manuālās pēcrediģēšanas sistēmas prototips: kreisajā pusē – automātiski atpazītā runa; labajā pusē – automātiski pēcapstrādāta transkripcija.

Teksta divvirzienu pārrakstīšanas gramatika sastāv no vairākiem simtiem likumu, kas apraksta bezkonteksta un kontekstuālu teksta šablonu atpazīšanu un apstrādi – izvēršanu vai savēršanu. Tā kā latviešu valoda ir morfoloģiski bagāta, izvēršanas gramatikai ir jānodrošina vārdformu kontekstuāla sintaktiska saskaņošana, savukārt savēršanas gramatikai jāatpazīst vārdi locījumos.

Izstrādātās runas automātiskās transkribēšanas un manuālās pēcrediģēšanas sistēmas prototips ilustrēts 4. attēlā. Šim automātiski transkribētajam un savērstajam MR izmeklējuma diktātam bija nepieciešami vien daži vienkārši manuāli labojumi. Dalījums teikumos un rindkopās, kā arī interpunkcija ir sistēmas ģenerēta; divās vietās manuāli tika precizēts, bet interpunkciju nebija nepieciešams koriģēt. Akronīmi MR un PSA tika korekti automātiski transkribēti ar izrunas vārdnīcu. Akronīms TUR-P tika transkribēts neprecīzi; ņemot to vērā, ir pilnveidota tekstu korpusa normalizēšana un papildināta izrunas vārdnīca. Skaitļu, mērvienību (*cm, cm3*), simbolu (*x*) u. c. atslēgvārdu (*k/v, l/m*) automātiska atpazīšana un saīsināšana veikta korekti. Formatējuma instrukcija "un rakstām lūdzu slēdzienu" atpazīta un automātiski interpretēta, izveidojot slēdziena sadaļu.

Secinājumi

Rakstā iezīmēta komplicētu specializēto korpusu izstrāde un adaptācija, kas ir nozīmīga latviešu valodas korpuslingvistikas resursu paplašināšanai. Specializētais teksta un runas korpuss ir ne vien ļāvis izstrādāt automatizētu sistēmu vizuālās diagnostikas izmeklējumu transkribēšanai, bet arī veicinās medicīnas valodas sistemātiskāku izpēti un lietojumu.

Līdzšinējie eksperimenti, pielāgojot runas sintēzes un runas atpazīšanas sistēmas medicīnas lietojumiem, ir pierādījuši nozarei specifisku datu nozīmi. Vispārīgu ASR sistēmu var pielāgot specifiskai jomai, konkrēti – radioloģijai, izmantojot specializētas datu kopas – specializētu teksta un runas korpusu, izrunas vārdnīcu. Papildus korpusiem, kas nepieciešami valodas modeļa un akustiskā modeļa izstrādei, būtiska ir arī teksta izvēršanas un savēršanas gramatika, kas ļauj samazināt automātiski ģenerēto izmeklējumu un slimību aprakstu manuālās pēcapstrādes apjomu. Šobrīd tiek pārskatīta runas korpusa datu ortogrāfiskā transkripcija un vienādots pieraksts, kas palīdzēs objektīvāk novērtēt (kvantitatīvi) runas atpazīšanas precizitāti, kā arī ļaus pilnveidot izrunas vārdnīcu.

Turpmākais darbs ir saistīts ar automatizētās diktēšanas platformas pilnveidi un tās efektivitātes novērtēšanu reālos apstākļos. Jau sākotnējā prototipa testēšanas fāzē esam novērojuši, ka automātiskās transkripcijas kopumā veidojas konsekventākas nekā diktofonu centra dažādu operatoru sagatavotās transkripcijas. Automātiskās transkripcijas, protams, nav bez kļūdām, taču atpazīšanas precizitāte un pēcrediģēšanas darba apjoms ir ļoti atkarīgi arī no diktēšanas stila. Tāpēc radiologiem tiek izstrādāti ieteikumi diktātu ierunāšanai, lai iegūtu pēc iespējas kvalitatīvāku rezultātu.

Kopumā būtiskākais jautājums un kritērijs sistēmas izmantošanai praksē ir: vai automātisko transkripciju pēcrediģēšana, izmantojot specializēto redaktoru, kas nodrošina arī automātisku teksta un audio sastatīšanu, ir ātrāka un ērtāka nekā pilnībā manuāla izmeklējuma apraksta sagatavošana. Uz šo jautājumu jau šobrīd ir pozitīva atbilde, jo daļa radiologu, kas piedalās sistēmas testēšanā, būtu gatavi turpmāk vismaz daļu no izmeklējumu transkripcijām rediģēt paši, negaidot rindā uz diktofonu centru un vienlaikus samazinot diktofonu centra noslodzi.

Blackley, Suzanne V., Huynh, Jessica, Wang, Liqin, Korach, Zfania, Zhou, Li (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association*, No. 26(4), pp. 324–338.

Goedertier, Wim, Goddijn, Simo, Martens, Jean-Pierre (2000). Orthographic Transcription of the Spoken Dutch Corpus. *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC), Athens, pp. 909–914.

Oostdijk, Nelleke, Boves, Lou (2008). Preprocessing speech corpora: Transcription and phonological annotation. Lüdeling, Anke, Kytö, Merja (eds.). *Corpus Linguistics: An International Handbook*. Volume I, Berlin, New York: W. de Gruyter, pp. 642–663.

Paats, Andrus, Alumäe, Tanel, Meister, Einar, Fridolin, Ivo (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,* No. 31(5), pp. 615–621.

Paikens, Pēteris, Rituma, Lauma, Pretkalniņa, Lauma (2013). Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA), pp. 267–277.

Pinnis, Mārcis, Auziņa, Ilze, Goba, Kārlis (2014). Designing the Latvian Speech Recognition Corpus. *Proceedings of the 9th edition of the Language Resources and Evaluation Conference* (LREC), Reykjavik, Iceland: ELRA, pp. 1547–1553.

Pinnis, Mārcis, Salimbajevs, Askars, Auziņa, Ilze (2016). Designing a Speech Corpus for the. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC), Portorož, Slovenia: ELRA, pp. 775–780.

Salimbajevs, Askars and Strigins, Jevgenijs (2015). Latvian speech-to-text transcription service. *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (INTERSPEECH), pp. 772–723.

Schiel, Florian and Draxler, Christoph (2004). The Production of Speech Corpora. Version 2.5, 01.06. Available: http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/Tp1.html [accessed 12.04.2022.].

Spektors, Andrejs, Auziņa, Ilze, Darģis, Roberts, Grūzītis, Normunds, Paikens, Pēteris, Pretkalniņa, Lauma, Rituma, Laura, Saulīte, Baiba (2016). Tezaurs.lv: the largest open lexical database for Latvian. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC), pp. 2568–2571.

Znotiņš, Artūrs, Cīrule, Elita (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies – The Baltic Perspective*, Vol. 307, pp. 183–189.

Znotiņš, Artūrs, Polis, Kaspars, Darģis, Roberts (2015). Media monitoring system for Latvian radio and TV broadcasts. *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (INTERSPEECH), pp. 732–733.

Summary

Development of a Specialized Latvian Speech Corpus and Pronunciation Dictionary for the Linguistic Analysis and Systematic Transcription of Visual Diagnostic Examinations

Ilze Auziņa, Roberts Darģis, Baiba Saulīte, Normunds Grūzītis, Mikus Grasmanis, Andrejs Spektors, Kaspars Stepanovs

Keywords: speech corpus, pronunciation dictionary, medical terminology, digital language resources, automatic speech recognition, natural language processing, post-editing

The Laboratory of Artificial Intelligence (AiLab) at the Institute of Mathematics and Computer Science of the University of Latvia (IMCS UL) in cooperation with the Riga East University Hospital (REUH) has developed the RUTA:MED platform for automated transcription of medical audio recordings. This was done within an ERDF-funded industry-driven research project aimed at developing specific Latvian speech recognition systems for the medical domain.

This paper describes the creation of Latvian language resources for the medical domain focusing on digital imaging to develop a medical speech recognition system for Latvian. The language resources include a pronunciation dictionary, a text corpus for language modelling, and an orthographically transcribed speech corpus for the (i) adaptation of the acoustic model, (ii) evaluation of the speech recognition accuracy, (iii) development and testing of rewrite rules for automatic text conversion to the spoken form and back to the written form.

Experiments to date in adapting speech synthesis and speech recognition systems to medical applications have demonstrated the importance of industry-specific data. The general ASR system can be adapted to a specific field, namely radiology, using specialized data sets – a specialized text and speech corpus, pronunciation dictionary. In addition to the corpora required for the development of the language model and the acoustic model, the grammar of text expansion and compression is also important, which allows to reduce the amount of manual post-processing of automatically generated examinations and disease descriptions.