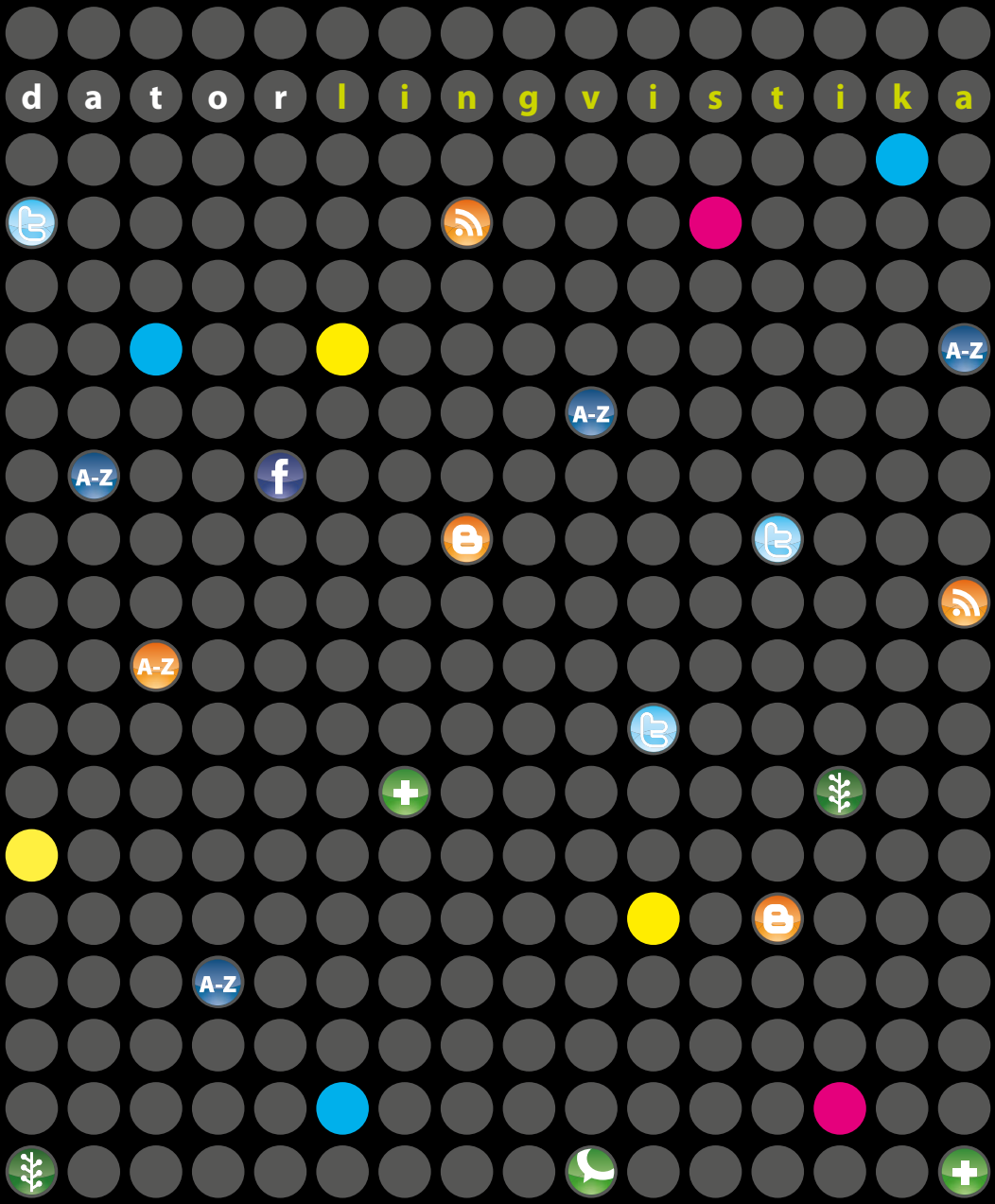
Latviešu valoda
digitālajā vidē



datorlingvistika
A-Z
A-Z
A-Z
A-Z

Latviešu valoda
digitālajā vidē

Informatīvi izglītojoša
semināru cikla materiāli

Rīga 2012

UDK 811.174:004
La 802

Latviešu valoda digitālajā vidē: datorlingvistika. Informatīvi izglītojoša semināru cikla materiāli [tiešsaiste]. Rakstu krājums. Rīga : LVA, 2012. 186 lpp.
Pieejams: *http://www.valoda.lv/Petijumi/Valodas_situacijas_izpete/mid_510.*
ISBN 978-9984-815-57-2

Redaktore *Dr. philol.* I. Auziņa
Korektore Marta Balode
Recenzente *Dr. philol.* T. Gornostaja
Maketētājs Mārtiņš Zunde

Latviešu valodas aģentūra

Latviešu valodas aģentūra sadarbībā ar Latvijas Universitātes Sociālo zinātņu fakultātes Kognitīvo zinātņu un semantikas centru, Rēzeknes Augstskolu, Ventspils Augstskolu un Liepājas Universitāti 2011. gadā organizēja informatīvi izglītojošu semināru ciklu **„Latviešu valoda digitālajā vidē"**. Semināru cikla mērķis bija apzināt Latvijā pēdējos gados īstenoto datorlingvistikas jomā, kā arī iepazīstināt ar jaunākajiem pētījumiem datorlingvistikā un tās saskares zinātņu jomās, popularizēt datorvides piedāvātās iespējas valodas izpētes jomā. Krājumā apkopoti raksti, kas iepazīstina ar referātos paustajām idejām un pētījumu rezultātiem.

Sadarbības augstskolas

KZSC

Latvijas Universitātes Sociālo zinātņu fakultāte

Latvijas Universitāte Anno 1919

Rēzeknes Augstskola

Ventspils Augstskola

ISBN 978-9984-815-57-2

Saturs

Priekšvārds

Mūsdienu informācijas tehnoloģiju laikmetā dators ir neaizstājams palīgs. Tādējādi arvien nozīmīgāka loma praktisko (un ne tikai) zinātņu vidū ir datorlingvistikai.

Latviešu valodas aģentūra sadarbībā ar Latvijas Universitātes Sociālo zinātņu fakultātes Kognitīvo zinātņu un semantikas centru (KZSC), Rēzeknes Augstskolu, Ventspils Augstskolu un Liepājas Universitāti 2011. gadā organizēja informatīvi izglītojošu semināru ciklu **„Latviešu valoda digitālajā vidē".**

Semināru cikla mērķis bija apkopot un popularizēt Latvijā pēdējos gados īstenoto datorlingvistikā, iepazīstināt studentus, valodas pētniekus, kā arī interesentus ar jaunākajiem pētījumiem datorlingvistikā un tās saskares zinātņu jomās, informēt par datorvides piedāvātajām iespējām un rosināt izmantot valodas izpētes jomā dažādas datubāzes, rīkus un programmas.

Krājumā sniegts ieskats datorlingvistikas attīstībā Latvijā, aktuālās tulkošanas un terminoloģijas problēmās, aplūkoti jautājumi par valodas funkcionēšanu digitālajā vidē, resp., dažādos sociālajos tīklos, skatīti korpuslingvistikas jautājumi, runāts par datorprogrammu izstrādi un to lietošanu pētnieciskos nolūkos.

Ceram, ka rakstu krājums būs noderīgs plašam lietotāju lokam: ne tikai filoloģijas un datorzinātnes studentiem, skolotājiem un docētājiem, valodu pedagogiem un tulkotājiem, bet arī valodniekiem, informācijas un komunikācijas tehnoloģiju speciālistiem un daudziem citiem interesentiem.

Informatīvi izglītojošais izdevums ļauj iepazīties ar paveikto četrās Latvijas augstskolās, kā arī ar šo augstskolu nākotnes iecerēm.

Liels paldies par sadarbību LU KZSC un tā direktoram *Dr. phil.* J. Šķilteram, Rēzeknes Augstskolai un Filoloģijas katedras vadītājai *Dr. philol.* I. Šuplinskai,

Ventspils Augstskolai un Lietišķās valodniecības centra direktoram, Anglistikas katedras vadītājam *Dr. philol.* J. Baldunčikam, Liepājas Universitātei un zinātnes prorektorei *Dr. philol.* I. Ozolai. Liels paldies arī visiem referentiem un publikāciju autoriem, krājuma atbildīgajai redaktorei *Dr. philol.* I. Auziņai, recenzentei *Dr. philol.* T. Gornostajai un visiem, kas jebkādā veidā atbalstījuši un sekmējuši gan semināru cikla norisi, gan rakstu krājuma izstrādi un izdošanu.

Dr. philol. **Inita Vītola**
Latviešu valodas aģentūras
Valodas attīstības daļas vadītāja

Ievads

Kas ir datorlingvistika?

Gunta Nešpore

Datorlingvistika ir valodniecības un datorikas robežzinātne, kas pēta valodas automātiskas analīzes iespējas, veidojot elektroniskus valodas resursus un valodas apstrādes rīkus.

Datorlingvistikas mērķis ir formāli precīzi (datoram „saprotamā" formā) vai statistiski aprakstīt valodu un tajā izteikto nozīmi, lai varētu automātiski izgūt no teksta saturu vai izteikt tekstā vajadzīgo saturu. Šāda mērķa sasniegšana pagaidām ir problemātiska, jo prasa ļoti pilnīgu visu valodas līmeņu (arī semantikas) formālu aprakstu. Tomēr virzība uz to notiek – dators tiek izmantots dažādās nozarēs un dažādos valodas analīzes līmeņos.

Jau tagad datorlingvistikā ir izdarīts ļoti daudz, arī latviešu valodā daudzi no mums ikdienā saskaras ar datorlingvistikas pētījumu rezultātiem – tie ir dažādi rīki un resursi, piemēram, pareizrakstības pārbaudītāji, mašīntulkošanas rīki (piemēram, *translate.google.com* piedāvātais tulkotājs, kurā iekļauta arī latviešu valoda), instrukcijas latviešu valodā *GPS* navigācijas sistēmās; daudzi valodnieki un citu humanitāro zinātņu pārstāvji ir pazīstami ar valodas korpusiem, vēl plašāka sabiedrība izmanto elektroniskās vārdnīcas. Lai arī valodas analīzes rīku un resursu radīšanā (tātad datorlingvistikā) liela nozīme ir datorikai, rezultāts ir nenovērtējams palīgs arī humanitāro zinātņu pārstāvjiem, īpaši valodniekiem, un tā izmantošanai bieži vien nav nepieciešamas specifiskas zināšanas. Tāpat kā vārdnīcu lietotājiem nav jābūt leksikogrāfiem, arī valodas apstrādes rīku un resursu lietotājiem nav jābūt datorlingvistiem. Jebkuram valodas pētniekam ir

vērts iepazīties ar piedāvātajiem resursiem, lai taupītu savu laiku un enerģiju un bieži vien arī nonāktu pie objektīvākiem valodas datiem un precīzākām to interpretācijas iespējām.

Tiek šķirti vairāki datorlingvistikas virzieni, kas cits citu papildina un mijiedarbojas. Daži no tiem tiks raksturoti sīkāk.

Korpuslingvistika pēta valodas vai citas parādības (kas ietvertas teksta saturā – vēsture, sabiedriskā doma, politiskie procesi un to atspoguļojums, piemēram, Saeimas plenārsēdēs), izmantojot lielu daudzumu elektronisku tekstu – valodas korpusu.

Valodas korpuss ir liels, strukturēts, elektroniski glabājams un apstrādājams tekstu kopums; tajā atrodami reāli, „neuzspodrināti" valodas dati. Valodas korpusā var ātri un precīzi atrast visdažādākos kontekstus un situācijas, kādās tiek lietota kāda valodas vienība.[1]

Atkarībā no lietošanas mērķa un tajā iekļautajiem tekstiem ir vairāku veidu valodas korpusi.

Tekstu korpuss vai runas korpuss. Kā jau no nosaukuma izriet, tekstu korpusā ir atrodami teksti. Visbiežāk tas ir paredzēts rakstu valodas izpētei. Runātas valodas izpētei ir paredzēts runas korpuss, tā pamatā ir valodas audioieraksti, kas papildināti ar to atšifrējumu un citu informāciju, piemēram, datiem par runātāju vecumu, dzimumu, dzimto valodu.

Vienvalodas korpuss vai paralēlais (divvalodu vai daudzvalodu) korpuss. Vienvalodas korpusu veido vienas valodas teksti, bet paralēlo korpusu – teksti un to tulkojums citā valodā. Paralēlā korpusa teksti mēdz būt sastatīti – tādā gadījumā ir norādītas atbilsmes starp teikumiem (vai citām teksta vienībām) oriģinālteksta un tā tulkojumā. Šāds korpuss ir izmantojams, piemēram, tulkošanas ekvivalentu pētīšanā, leksikogrāfijā un cita veida sastatāmajā analīzē.

Vispārīgais korpuss vai specializētais (speciāls) korpuss. Vispārīgie korpusi nav ierobežoti tematiski vai pēc kādas citas pazīmes, tie kaut kādā mērā atspoguļo valodas stāvokli attiecīgajā laika posmā. Specializētā korpusā savukārt apkopoti teksti pēc kādas papildu pazīmes. Tas var būt ierobežots tematiski (piemēram, kādas zinātnes nozares korpuss), ģeogrāfiski (piemēram, kādas izloksnes vai dialekta korpuss), pēc runātāju vecuma (piemēram, jauniešu vai bērnu valodas korpuss) vai pēc kādas citas pazīmes.

[1] Par latviešu valodā pieejamajiem valodas korpusiem un citiem resursiem sk. N. Grūzīša rakstā un nodaļā „Latviešu valodas digitālie resursi un rīki".

Jebkura veida valodas korpuss var būt marķēts jeb anotēts, t. i, korpusa datiem var tikt pievienota visdažādākā papildu informācija – metadati. Tie var raksturot korpusā iekļautās teksta vienības (gramatiskās formas, teikumus u. tml.) vai tekstu un tā autoru (uzrakstīšanas laiks, autora vārds, teksta tēma, izdevējs u. tml.). Korpusa anotācija paver lietotājam daudz plašākas meklēšanas iespējas korpusā, kā arī citas korpusa izmantošanas iespējas. Piemēram, ja korpuss ir marķēts morfoloģiski, tajā var automātiski atrast visas vārdformas, kas atbilst kādai morfoloģiskai pazīmei – var meklēt vārdformas pēc dzimtes, skaitļa, locījuma, deklinācijas, konjugācijas u. tml. Sintaktiski marķētā korpusā var atrast, piemēram, visus teikuma priekšmetus, vienlīdzīgus teikuma locekļus, adverbiālus izteicējus u. tml.

Parasti korpusa lielumu mēra vārdlietojumos, tas ir kopējais korpusā iekļauto vārdformu (un citu vienību) skaits. Vienīgi sintaktiski marķēto korpusu lielumu skaita teikumos, un tas parasti ir mazāks nekā nemarķēts vai morfoloģiski marķēts korpuss.

Ir dažāda lieluma valodas korpusi. Mūsdienās par vispārīga tekstu korpusa minimālo apjomu varētu uzskatīt vienu miljonu vārdlietojumu, lai arī bieži vien tie ir lielāki, bet mūsdienu lielākie korpusi (angļu valodai) ir pat miljardu vārdlietojumu lieli.

Par sava veida specializēto korpusu var uzskatīt arī citus elektroniskos resursus, piemēram, elektroniskās vārdnīcas, jo arī tās ir noteiktā veidā sakārtots tekstu krājums. No otras puses, ideālā gadījumā vārdnīca ir arī korpusa izmantošanas rezultāts, t. i., vārdnīca ir veidota, izmantojot valodas korpusu, kurā redzami iespējamie valodas vienības lietošanas gadījumi un ar automātiski iegūtas statistikas palīdzību nosakāms tipiskais un retāk sastopamais valodā.

Viena no datorlingvistikas centrālajām darbības jomām ir **gramatikas analīze**. Tā vajadzīga, piemēram, pareizrakstības pārbaudes rīku izveidei, tekstu korpusa marķēšanai, meklēšanas rīku izveidei (lai interneta meklētājs piedāvātu ne tikai meklēto vārdformu, bet arī citas attiecīgā vārda formas, automātiski jānosaka vārda pamatforma, lai konstatētu tā locīšanas paradigmu). Gramatikas analīze vajadzīga arī tālākai valodas apstrādei, piemēram, mašīntulkošanā, runas tehnoloģijās, diskursa analīzē.

Izveidot automātisku gramatikas analizatoru nozīmē uzrakstīt likumus, pēc kuriem dators automātiski var noteikt, piemēram, vārdformas morfoloģiskās pazīmes vai teikuma sintaktisko struktūru. Lai to izdarītu, jāatceras, ka ir precīzi un formāli jādefinē arī tas, kas cilvēkam šķiet pašsaprotams, skaidri jānošķir forma un saturs – cilvēks valodas vienības formu un saturu parasti uztver kā vienotu veselumu, bet datorprogramma to nespēj. Viena no galvenajām problēmām gramatikas analīzē un datorlingvistikā vispār ir daudznozīmība. Piemēram, cilvēks,

redzot teikumu *Es ceļu māju*, visticamāk pat neiedomāsies, ka forma *māju* var būt ne tikai lietvārda forma, bet arī verba *māt* tagadnes 1. personas forma, savukārt automātiskas gramatikas analīzes programmai ir vajadzīgi formāli (ar nozīmi nesaistīti) nosacījumi, kā izvēlēties vajadzīgo vārdformas interpretāciju.[2] Lai tos aprakstītu, bieži vien jāprecizē arī tradicionālais latviešu valodas gramatikas apraksts. Piemēram, analizējot dažādu tekstu sintaksi, bieži jāsaskaras ar sarežģītām un ne vienmēr gramatiski un stilistiski perfektām formām, kas nav apskatītas latviešu valodas sintakses teorijā, tomēr tekstā ir jāapraksta.

Mašīntulkošana ir datorlingvistikas daļa, kurā nepieciešama visu valodas līmeņu automātiska analīze. Lai arī datorlingvistika kā nozare aizsākās tieši ar domām par automatizētu tulkošanu, tā vēl aizvien ir grūts un līdz galam neatrisināts uzdevums. Lai arī dažādām valodām pieejamās mašīntulkošanas kvalitāte atšķiras, dators jau tagad daudzās valodās var palīdzēt aptuveni saprast kāda teksta saturu, izveidot teksta tulkojuma melnrakstu, ar datora palīdzību diezgan kvalitatīvi var tulkot kādas saturiski un gramatiski ierobežotas jomas tekstus (piemēram, laika prognozes, finanšu ziņas vai tml.). Skaidrs arī, ka ar datora palīdzību var labāk vai sliktāk pārcelt citā valodā teksta saturu, bet ne zemtekstus vai stila nianses. Nevajadzētu no mašīntulkošanas gaidīt, piemēram, daiļliteratūras tulkojumus, bet izmantot tās iespējas, ko mašīntulkošana šobrīd var piedāvāt.

Runas tehnoloģijas nodarbojas ar runātās valodas apstrādi. Šķir divas galvenās runas tehnoloģijas daļas: runas atpazīšanu un runas sintēzi. Runas atpazīšanas uzdevums ir runātu tekstu (audio) pārvērst rakstītā tekstā, bet runas sintēze nodarbojas ar rakstīta teksta pārvēršanu skaņā. Lai to izdarītu, ir jāapraksta valodas skaņu akustiskās īpašības, fonētiski fonoloģiskie procesi un prosodija. Latviešu valodā labāk attīstītas ir runas sintēzes tehnoloģijas.

Līdzīgi kā mašīntulkošana arī runas sintēze mēdz būt saturiski ierobežota un neierobežota (universāla). Pirmā ir piemērota, piemēram, telekomunikāciju, telefonbanku un citās ierobežota dialoga sistēmās, bet otrā – universāla izmantojuma sistēmās, piemēram, vājredzīgajiem un neredzīgajiem cilvēkiem, datorizētā valodas mācīšanā. Saturiski ierobežotās runas sintēzes sistēmas darbojas precīzāk, runātā teksta kvalitāte ir labāka, bet tās var nedarboties ārpus tām paredzētās tematikas. Savukārt universālās sistēmas var sintezēt jebkura satura tekstu, lai arī, iespējams, sliktākā kvalitātē nekā kādai jomai īpaši paredzētas sistēmas.

Lai izveidotu universālu runas sintēzes sistēmu, jāveic dažādu līmeņu teksta apstrāde: sākot no teksta sagatavošanas automatizētai fonētiskai transkribēšanai, beidzot ar audiosignāla ģenerēšanu un prosodijas modelēšanu. Parasti runas

[2] Par daudznozīmības problēmu plašāk sk. N. Grūzīša rakstā.

sintēzē izmanto divas skaņas apstrādes un glabāšanas tehnoloģijas – konkatenāciju jeb savirknēšanu un uz likumiem balstītu pieeju. Savirknēšanas gadījumā tiek izmantoti gatavi skaņu segmenti (piemēram, fonēmas, fonēmu varianti, difoni, zilbes), bet uz likumiem balstītā runas sintēzē tiek izmantoti likumi, kas raksturo fonēmu savstarpējo ietekmi un attieksmes un pēc kuriem skaņu veido speciālas ierīces.

Runas tehnoloģijas saista runāto valodu ar pārējiem valodas apstrādes virzieniem. Ja runātais teksts automātiski tiek pierakstīts, to var apstrādāt tālāk tāpat kā jebkuru citu tekstu (gramatikas analīze, informācijas izgūšana no teksta), savukārt – ja dators „prot" nolasīt rakstītu tekstu, lietotājs var runāta teksta formā saņemt automātiski atlasīto un tekstā formulēto informāciju. Līdz ar to var izveidot ar balsi vadāmas dialogu sistēmas, ko var izmantot, piemēram, viedtelefonos, dažādos uzziņu pakalpojumos, biļešu rezervēšanas sistēmās. Latviešu valodā ir izveidotas vairākas runas sintēzes sistēmas, tiek strādāts arī ar runas atpazīšanu, taču pagaidām plaši lietojamu sistēmu nav.

Pastāv arī citas datorlingvistikas nozares un virzieni, kas šeit netiek aplūkoti sīkāk, bet plašāku informāciju par valodas tehnoloģiju pamatiem, kā arī par pašreizējo stāvokli Latvijas datorlingvistikā var atrast ieteicamās literatūras sarakstā. Datorlingvistikas mācību grāmatas latviešu valodā vēl nav, tāpēc tiek piedāvāta tematiski plaša mācību grāmata angļu valodā.

Ieteicamā literatūra

Mācību grāmata

Jurafsky, Daniel, and James, H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall, 2009

Par datorlingvistikas pētījumiem latviešu valodā un latviešu valodas tekstu korpusa izmantošanu

1. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA).* Riga, 2011. Pieejams: *http://dspace.utlib.ee/dspace/handle/10062/16955*
2. *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 2010.
3. Languages in the European Information Society. Latvian. META-NET White Paper Series. Budapest, 2011. Pieejams: *http://www.meta-net.eu/whitepapers/download/meta-net-languagewhitepaper-latvian.pdf*
4. Andronova, E., Andronovs, A. Latviešu valodas korpuss un tā izmantošana. No: *Valodas prakse: vērojumi un ieteikumi*. Populārzinātnisku rakstu krājums, Nr. 6. Rīga : Latviešu valodas aģentūra, 2011, 41.–57. lpp.
5. Levāne-Petrova, K. Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē No: *Vārds un tā pētīšanas aspekti* : rakstu krājums, Nr. 15(1). Liepāja : LiePA, 2011, 187.–193. lpp.

Normunds Grūzītis

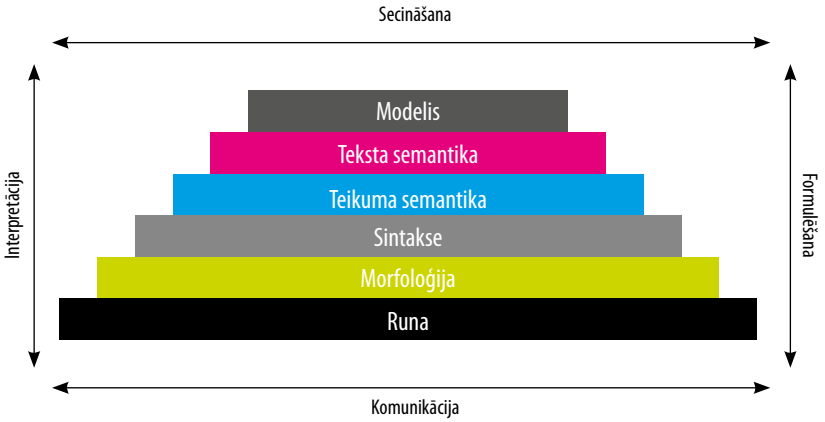
Datorlingvistikas pētījumi LU Matemātikas un informātikas institūtā

Ievads

Kopš savas darbības pirmsākumiem 1989. gadā LU Matemātikas un informātikas institūta (LUMII) Mākslīgā intelekta laboratorijas (AILab) galvenais pētījumu virziens ir bijusi datorlingvistika. Datorlingvistikas centrālais mērķis ir dabiskās valodas (t. i., runas vai teksta) formā izteiktas nozīmes formāla reprezentācija – šādas reprezentācijas (modeļa) automātiska iegūšana no dotās runas/teksta (t. i., valodas analīze) un otrādi – runas/teksta ģenerēšana (sintēze) saskaņā ar doto modeli (domai). Taču, lai nonāktu līdz šim mērķim, vispirms ir jārod optimāli risinājumi komplicētām, savstarpēji saistītām problēmām vairākos zemākos valodas analīzes līmeņos (sk. 1. attēlu).

Optimāls risinājums, no vienas puses, var nozīmēt kompromisu meklēšanu, lai efektīvi formalizētu dažādas neregulāras valodas parādības. No otras puses, tas nereti veicina mazāk pētītu valodniecības jautājumu izpēti un valodniecības teorijas aspektu sistematizēšanu. Liela nozīme šajā gadījumā ir apjomīgam, reprezentatīvam un, vēlams, arī attiecīgi marķētam valodas korpusam: statistiska analīze šādos empīriskos datos var palīdzēt atklāt šķietami „neregulāro" parādību regulārās pazīmes un kopsakarības.

1. attēls. **Dabiskās valodas analīzes līmeņi un to loma saziņas procesā**



AILab darbība latviešu valodas datorlingvistikas pētījumos un izstrādēs notiek vairākos virzienos (Milčonoka et al. 2004; Grūzītis et al. 2004; Bārzdiņš u. c. 2006; Skadiņa et al. 2010), aptverot praktiski visus analīzes un sintēzes līmeņus:

– gramatikas analīzes un sintēzes metodes un rīkus,
– rakstītās un runātās valodas korpusus (uzkrāšanu un marķēšanu),
– mašīnlasāmās vārdnīcas un leksiskās ontoloģijas (semantiskos tīklus),
– likumbāzēto un statistisko mašīntulkošanu,
– semantiskās analīzes un diskursa reprezentācijas metodes,
– kontrolētas dabiskās valodas un formālas valodas mijiedarbību,
– runas sintēzi un (eksperimentālu) analīzi.

Mašīntulkošana ir viens no tipiskākajiem un uzskatāmākajiem datorlingvistikas lietojumiem, kas konceptuāli ietver pilnu valodas analīzes ciklu.[1] Taču plaši lietotajā un pretrunīgi vērtētajā statistiskajā mašīntulkošanā (SMT) tikai salīdzinoši nesen ir sākts izmantot t. s. faktorētos modeļus, kuros papildus virspusējai n-grammu statistikai tiek ņemta vērā arī gramatiskā informācija (Skadiņa, Brālītis 2009; Skadiņš et al. 2010). Tiesa, latviešu valodai izstrādātajos SMT modeļos pagaidām ir iekļauta tikai morfoloģiskā informācija (atbilstoši pieejamajiem anotētajiem tekstu korpusiem), taču LUMII patlaban norit aktīvs darbs pie sintaktiski anotēta latviešu valodas paraugkorpusa izveides (sk. 2.1. nodaļu), kas drīzā nākotnē dos iespēju eksperimentēt arī ar sintaktiskiem valodas modeļiem.

[1] Idejiski: tulkojot domu pret domu.

Lai arī mūsdienu praksē ar dabiskās valodas apstrādi saistītos lietojumos dažādu objektīvu iemeslu dēļ līdz padziļinātai semantiskajai analīzei lielākoties (pagaidām) nemaz nenonāk, arī teksta gramatiskā (morfoloģiskā un sintaktiskā) analīze sniedz nozīmīgas iespējas informācijas meklēšanā, analīzē un tulkošanā.

Daudznozīmība

Galvenā problēma dabiskās valodas automātiskā analīzē ir tās izteiktā daudznozīmība visos analīzes līmeņos: no morfoloģisko pazīmju līdz pat diskursa referentu noteikšanai. Lai arī tas ļoti apgrūtina valodas deterministiskas (iepriekš paredzamas jeb viennozīmīgas) analīzes iespējas un vispārīgā gadījumā to faktiski padara neiespējamu, daudznozīmība ir fenomenāls līdzeklis, kas valodas lietošanu padara ērtu cilvēkam. Vārdu, vārdformu un sintaktisko konstrukciju klāsts valodā ir galīgs, taču to kombinēšanas un interpretācijas iespējas ir teorētiski bezgalīgas. To uzskatāmi ilustrē vārdu izplatība, piemēram, Britu nacionālajā tekstu korpusā (BNC), kurā 7500 biežāk lietotie vārdi pārklāj 90 % vārdlietojumu (Atkins, Rundell 2008). Līdzīga situācija ir novērojama arī līdzsvarotajā mūsdienu latviešu valodas korpusā (sk. 2.1. nodaļu), kurā 7500 biežāk sastopamās vārdformas pārklāj 73 % vārdlietojumu. Taču korekts salīdzinājums un biežumsaraksta plašāka izmantošana būs iespējama tad, kad latviešu valodas korpuss būs morfoloģiski viennozīmīgi nomarķēts[2], norādot arī pamatformas.

Saskaņā ar Zifa likumu (Zipf 1949) jebkura vārda sastopamības biežums ir apgriezti proporcionāls šī vārda pozīcijai valodas vārdu biežumsarakstā (attiecībā pret kādu tekstu korpusu). Tas nozīmē, ka visbiežāk lietotais vārds tekstos sastopams divreiz biežāk nekā otrs biežākais vārds, trīsreiz biežāk nekā trešais biežākais vārds utt. Citiem vārdiem sakot, salīdzinoši maz vārdu tiek lietoti bieži – lielākā daļa tiek lietoti reti. No tā varam secināt divas lietas. Pirmkārt, dažādos kontekstos vārdi tiek intensīvi atkalizmantoti, radot pamatu plašai leksiskajai daudznozīmībai, un, otrkārt, ir nepieciešams ļoti liels[3] tekstu korpuss, lai tas būtu pietiekami reprezentatīvs attiecībā uz reti lietotiem vārdiem un to nozīmēm.

Teorētiski katrs nākamais valodas analīzes līmenis palīdz atrisināt daudznozīmību iepriekšējā līmenī: morfoloģisko daudznozīmību lielā mērā novērš sintaktiskā analīze, sintaktisko daudznozīmību – leksiskā un sintaktiskā semantika, bet semantisko daudznozīmību – konteksts un t. s. fona jeb ontoloģiskās, jeb pasaules zināšanas.

---

[2] Pirmais miljons pusautomātiski, pēc tam automātiski, lietojot mašīnmācīšanās metodes.

[3] Salīdzinājumam: minētais BNC korpuss aptver 100 milj. vārdlietojumu, taču mūsdienu korpuslingvistikā tiek runāts jau par miljardus vārdlietojumu lieliem, no tīmekļa resursiem kompilētiem korpusiem (Kehoe & Gee 2007).

Kā piemēru apskatīsim vienkāršu teikumu: *Es ceļu māju.* Ja nošķirti analizējam katra vārda morfoloģiskās pazīmes, iegūstam šādus vienkāršotas analīzes variantus:

- *es* – vienskaitļa pirmās personas vietniekvārds;
- *ceļu* – verbs īstenības izteiksmē, tagadnē, pirmajā personā (*celt*) vai lietvārds vienskaitļa akuzatīvā (*ceļš*), vai lietvārds daudzskaitļa ģenitīvā (*ceļš* vai *celis*);
- *māju* – lietvārds vienskaitļa akuzatīvā (*māja*) vai lietvārds daudzskaitļa ģenitīvā (*māja*), vai verbs īstenības izteiksmē, tagadnē, pirmajā personā (*māt*).

Sintaktiskā analīze samazina morfoloģiskās analīzes variantu un to iespējamo kombināciju skaitu, taču arī tā pati par sevi diemžēl nedod viennozīmīgu rezultātu:

- *es* $_{\text{SUBJEKTS}}$ + *celt* $_{\text{PREDIKĀTS}}$ + *māja* $_{\text{OBJEKTS;}}$
- *es* $_{\text{SUBJEKTS}}$ + *ceļš* $_{\text{OBJEKTS}}$ + *māt* $_{\text{PREDIKĀTS.}}$

Lai izvēlētos korekto sintaktiskās analīzes variantu, ir nepieciešamas:

- leksiskās semantikas zināšanas, piemēram, vārds *māja* ir hiponīms vārdam *celtne*, bet vārds *celt* vienā nozīmē ir sinonīms vārdam *būvēt*, citā – hiponīms vārdam *pārvietot*;
- zināšanas par verbu sintaktisko valenci – tipiskajiem paplašinātājiem, piemēram, verbs *māt* ir intransitīvs, tātad tiek lietots kopā ar prepozicionālu konstrukciju (piemēram, atbild uz jautājumu *ar ko?*), nevis tiešo objektu (*ko?*);
- sintaktiskās semantikas (semantiskās valences) zināšanas – kādas semantiskās kategorijas (tipiskās semantiskās lomas) verbs piesaista, piemēram, verbs *māt* piesaista INSTRUMENTU, bet ne PACIENSU (darbības izjutēju);
- fona zināšanas, piemēram: *Tas, ko kāds būvē, ir celtne. Ikviena celtne ir nekustamais īpašums. Tas, ko kāds pārvieto, ir kustamais īpašums. Kustamais īpašums nav nekustamais īpašums.*

Visi šie semantiskās analīzes aspekti kopā palīdz izvēlēties dotajam teikumam ne tikai atbilstošāko sintaktisko interpretāciju, bet arī semantisko interpretāciju, nozīmes *celt – pārvietot* vietā izvēloties *celt – būvēt*. Cilvēks ar šādiem daudznozīmības risināšanas uzdevumiem lielākoties[4] tiek galā ļoti efektīvi un precīzi, taču automātiskā gramatiskajā un semantiskajā analīzē tas ir nopietns šķērslis – lingvistisko un pasaules zināšanu formalizēšanas grūtību dēļ. Jāatzīmē, ka kopš 2005. gada LUMII ir veikti vairāki pētījumi un eksperimentālas izstrādes (galvenokārt Valsts pētījumu programmu ietvaros), aptverot minētos semantiskās analīzes aspektus.

[4] Kā rāda pieredze, veidojot sintaktiski un semantiski anotētus tekstu korpusus, pat valodnieki vienu un to pašu teikumu/vārdu mēdz interpretēt dažādi (Brants 2000; Chklovski & Mihalcea 2003). Nereti iemesls tam ir nepietiekams konteksts vai pārāk detalizēta nozīmju kopa (skaidrojošā vārdnīca), arī atklātie jautājumi sintaksē.

1. LATVIEŠU VALODAS GRAMATISKĀ ANALĪZE UN SINTĒZE

Automatizētas morfoloģiskās analīzes problemātika pamatā ir atrisināta daudzām pasaules valodām, t. sk. latviešu. LUMII ir izstrādāti un nekomerciālai lietošanai brīvi pieejami vairāki latviešu valodas morfoloģiskie analizatori un sintezatori, kā arī formālie leksikoni, uz kuriem tie balstās. Elastīgs analizators, kas satur bagātīgu vārddarināšanas likumu kopu (Paikens 2007) un tiek izmantots, piemēram, *SemTi-Kamols* sintaktiskajā analizatorā (sk. 1.1. nodaļu), kā arī projektos ārpus LUMII, ir pieejams kā programmatūras bibliotēka[5], savukārt Latviešu literārās valodas vārdnīcas leksikonā balstīts analizators/sintezators ir pieejams kā automatizēti izmantojama tīmekļa pakalpe.[6]

Taču pastāv divas problēmas, kuru risināšana nav tik ļoti sarežģīta, cik darbietilpīga. Pirmkārt, tas ir leksikona izmērs. Esošie leksikoni aptver 50–60 tūkstošus vārdu[7], taču neierobežota teksta analīzei būtu nepieciešams 2–3 reizes lielāks leksikons, iekļaujot tajā arī retāk lietotus vārdus, nemaz nerunājot par nozarspecifiskiem tekstiem un attiecīgu terminoloģiju. Otrkārt, morfoloģiskais analizators neanalizē vārdlietojuma kontekstu – dotai vārdformai tiek piedāvāti visi iespējamie analīzes varianti. Tas, kas bieži vien ir nepieciešams praksē, ir morfoloģiskais tageris, kas, izmantojot lokālus likumus un/vai statistiku, ar lielu varbūtību spēj noteikt pareizo analīzes variantu attiecīgajā konteksta „logā".

Statistiska tagera apmācīšanai ir nepieciešams manuāli morfoloģiski marķēts (pārbaudīts) korpuss, vēlams līdzsvarots un gana apjomīgs, ņemot vērā, ka latviešu valoda ir izteikti fleksīva – lai dažādu veidu morfosintaktiskie šabloni būtu pietiekami bieži reprezentēti. LUMII ir izveidots šāda tagera prototips[8], kas ir pieejams arī kā tīmekļa pakalpe.[9] Tas ir balstīts uz atsevišķiem, nelieliem, manuāli marķētiem korpusiem (kopā ~60 000 vārdlietojumu) un iepriekšminētā morfoloģiskā analizatora. Novērtējums rāda, ka tagera precizitāte ir 70–90 % atkarībā no teksta veida (tā līdzības treniņkorpusa tekstiem), reti lietotu īpašvārdu un citu neatpazītu vārdu biežuma u. tml. Praktiskiem lietojumiem šāda precizitāte ir nepietiekama[10]: pieņemot, ka vidējais teikuma (teikuma daļas) garums ir 10 vārdi, katrā teikumā (daļā) tiek pieļautas vidēji divas kļūdas, kas turpmāk rada būtiskas kļūdas sintaktiskajā analīzē. Ja interesē tikai vārdšķiras noteikšana, tad tagera precizitāte ir virs 95 %, kas ir labs rādītājs noteiktos praktiskos lietojumos, piemēram, precīzākā informācijas meklēšanā un atlasē. Tagera precizitāti zināmā

[5] http://www.semti-kamols.lv/ (Rīki > Morfoloģiskais analizators)
[6] http://valoda.ailab.lv/ws/morph/
[7] T. sk. minimāli nepieciešamo informāciju par vārdu locīšanas paradigmām.
[8] http://eksperimenti.ailab.lv/tagger/
[9] http://valoda.ailab.lv/ws/tagger/
[10] Pasaules pieredze rāda, ka morfoloģiskā tagera precizitātei būtu jābūt vismaz 95 %.

mērā iespējams uzlabot, attīstot apmācības algoritmu un n-grammu statistiku kombinējot ar sintaktiskiem likumiem, taču galvenais stūrakmens, lai būtiski uzlabotu precizitāti, ir apjomīgs treniņkorpuss.[11]

1.1. Atkarību gramatikā balstīts hibrīds gramatikas modelis

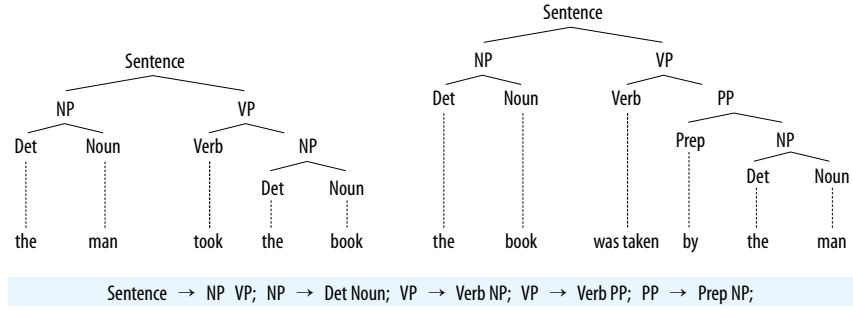
Tā kā gramatiskās nozīmes latviešu valodā tiek izteiktas galvenokārt ar fleksijām (galotnēm), vārdu secība teikumā ir sintaktiski brīva – teikuma locekļu noteikšanu nodrošina vārdu locīšanas paradigmas un to saskaņošana teikumā. Taču pat sintētiskās valodās vārdu secība ir tikai relatīvi brīva: to ierobežo gan analītisko formu lietošana, gan semantiskais saistījums (sk. 1.2. nodaļu).

Sintaktiskajā analīzē var izšķirt divas galvenās pieejas: frāžu struktūras (PSG – *phrase structure grammar*) pieeju (Chomsky 1956) un atkarību gramatikas (*dependency grammar*) pieeju (Tesnière 1959) (sk. 2. un 3. attēlu). Atšķirīgās pieejas piedāvā principiāli atšķirīgus skatījumus un metodiku ar attiecīgām priekšrocībām un trūkumiem. Analītiskajām valodām (piemēram, angļu), kurās vārdu secība ir strikti noteikta, tradicionāla ir pieeja no augšas uz leju: ievērojot vārdu lineārās secības, teikumi tiek dalīti frāzēs (~vārdkopās) un to sastāvdaļās (t. sk. rekursīvās) saskaņā ar produkcijas likumu kopu. Turpretim fleksīvo, sintētisko valodu gadījumā tradicionāla ir pieeja no apakšas uz augšu: tiek noteikti teikuma locekļi un pakārtojuma attieksmes starp tiem atbilstoši valodā pieļaujamajām sintaktiskajām funkcijām. Tiesa, praksē analītisko un sintētisko valodu dalījums netiek īpaši ievērots: sintētisku valodu formalizēšanai nereti tiek izvēlēta *PSG* pieeja, savukārt, parādoties efektīviem, korpusā balstītiem analizatoriem, analītisko valodu gadījumā populāra ir kļuvusi arī atkarību gramatikas pieeja.

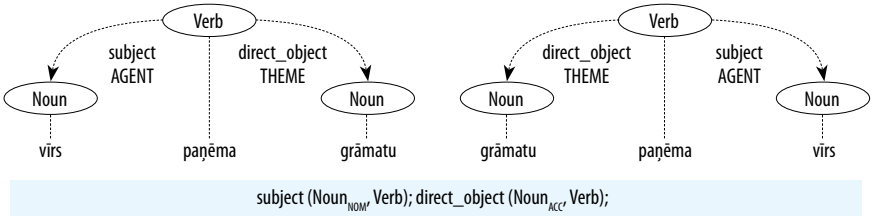
Vienas vai otras pieejas priekšrocības nenosaka tikai valodas tips. No vienas puses, atkarību gramatika piedāvā salīdzinoši elastīgākas analīzes iespējas: ar vienkāršāku likumu (funkciju) kopu, nespecificējot vārdu lineāro secību, iespējams pārklāt vienlīdz plašu valodas apakškopu. No otras puses, atkarību gramatika nav īsti piemērota valodas ģenerēšanai, kas ir *PSG* pieejas stiprā puse. Taču 1. attēla kontekstā svarīgāks aspekts ir sintaktiskās analīzes koku tālākas atainošanas iespējas semantiskās interpretācijas modelī. Šajā ziņā priekšrocība ir atkarību gramatikas pieejai, kuras centrā ir verbs (izteicējs) un tā argumentu struktūra, kas tiek kodēta tiešā veidā – ar pakārtojuma attieksmēm (atkarībām). Turklāt verbs atrodas teikuma centrā ne tikai sintaktiski, bet arī semantiski – kā notikums (situācija, darbība). Līdz ar to šādas pakārtojuma attieksmes tiešā veidā var attēlot semantiskajās attieksmēs (sk. 3. attēlu).

[11] Kā rāda pasaules pieredze, fleksīvajai valodai būtu nepieciešams vismaz miljons vārdlietojumu.

2. attēls. **Angļu valodas teikumu vienkāršots attēlojums *PSG* gramatikā. Darāmās un ciešamās kārtas analīzes koki atšķiras, kompensējot formālo atšķirību trūkumu morfoloģijas līmenī**



3. attēls. **Latviešu valodas teikumu vienkāršots attēlojums atkarību gramatikā. Latviešu valodā nav jāmaina kārta, lai mainītu vārdu secību – analīzes koks abos gadījumos sakrīt (tiek lietotas tās pašas funkcijas), atšķiras tikai verba paskaidrotāju lineārā secība**



Atkarību pieejā balstītās gramatikas lielākoties tiek realizētas, ļoti vienkāršojot oriģinālo L. Tenjēra pieeju (Tesnière 1959), lingvistisko precizitāti un dabiskās valodas īpatnības upurējot par labu efektīviem analīzes algoritmiem. Tādējādi katrs vārds (gan palīgnozīmes, gan pilnnozīmes jeb patstāvīgas nozīmes) faktiski tiek uzskatīts par patstāvīgu teikuma locekli, kas ir iesaistīts atsevišķā pakārtojuma attieksmē.[12] Turklāt netiek ņemts vērā, ka valodās ar brīvu vārdu secību vārdu secība pat tīri sintaktiski nav absolūti brīva.[13] Ņemot vērā šos aspektus, Valsts pētījumu programmas informācijas tehnoloģijās ietvaros LUMII ir izstrādāts hibrīds, atkarību gramatikā balstīts sintaktiskās analīzes modelis, saukts par *SemTi-Kamols* gramatiku (Bārzdiņš et al. 2007; Nešpore et al. 2010)

[12] Piemēram, analītiskā verba forma „bija paņēmis" tiek mākslīgi sadalīta divos teikuma locekļos, tādējādi palīgverbs paskaidro patstāvīgās nozīmes verbu vai otrādi.

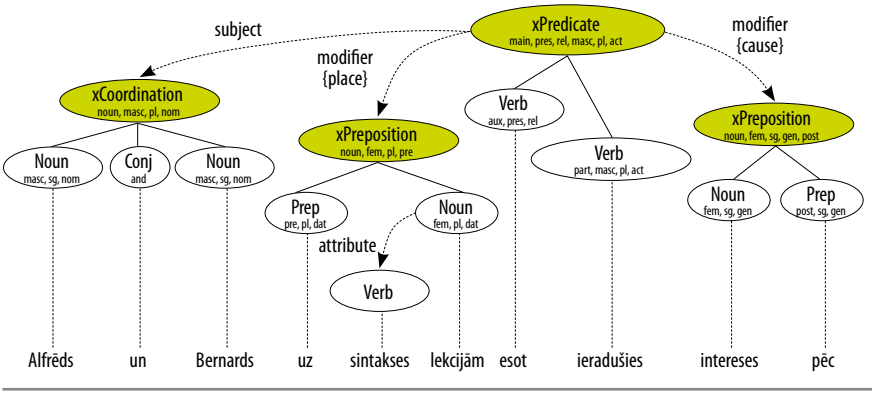
[13] Piemēram, konstrukcijā „uz <kā?>" prievārda pozīcija vienmēr ir pirms lietvārda.

L. Tenjēra strukturālajā sintaksē ir izšķirti trīs pamatjēdzieni: sintaktiskās attieksmes, junkcijas un transference. Analīzes koka virsotņu apzīmēšanai L. Tenjērs papildus ir ieviesis kodola jēdzienu. Tās ir sintaktiski nedalāmas vienības, ko veido viens centrālais elements (pilnnozīmes vārds) un, neobligāti, viens vai vairāki palīgvārdi un kas tiek iegūtas morfosintaktiskās transferences rezultātā. Pakārtojuma attieksmes līdz ar to figurē kodolu līmenī, nevis teikuma virsējās struktūras (vārdu) līmenī. Savukārt junkcijas apvieno tādus elementus, kas atrodas nevis pakārtojuma, bet sakārtojuma (horizontālās) attieksmēs (galvenokārt vienlīdzīgi teikuma locekļi un teikuma daļas). Visbeidzot transference ir operācija, kuras izpildes laikā kodola funkcionālais jeb palīgvārds izmaina pilnnozīmes vārda oriģinālo kategoriju (funkciju).

Piedāvātajā *SemTi-Kamols* hibrīdajā modelī ir realizēts Tenjēra sintaktiskā kodola jēdziens, lietojot mehānismu, kas ir ļoti tuvs transferences operācijai. Turklāt kodola jēdziens ir paplašināts, ietverot tajā arī junkcijas. Lai to īstenotu, ir ieviests komplekso jeb „x-vārdu" jēdziens (sk. 4. attēlu). Tas ir līdzeklis, kas, abstrahējot analītiskās formas un vienlīdzīgus teikuma locekļus, kalpo kā tilts starp *PSG* un vienkāršotu atkarību gramatiku.

Transferences rezultātā analīzes kokā tiek izveidotas jaunas, mākslīgas virsotnes (kompleksi vārdi) ar savām morfosintaktiskajām pazīmēm, kas lielākoties tiek mantotas no tās veidojošajiem elementiem, bet var atspoguļot arī papildu informāciju, kas attiecas uz x-vārdu kopumā un kas var būt nepieciešama atbilstošās (-o) pakārtojuma funkcijas (-u) piemērošanā. Rekursīvi aizvietojot visas analītiskās formas un vienlīdzīgos teikuma locekļus ar x-vārdiem, tiek iegūta vienkārša teikuma struktūra.

4. attēls. **Vienkāršots teikuma analīzes koks hibrīdajā, atkarību gramatikā balstītajā modelī**



SemTi-Kamols modelis var tikt izmantots dažādu sintaktisko konstrukciju adekvātai attēlošanai, t. sk. analītiskām verbu formām (piemēram, veidojot saliktos laikus un konstrukcijas ar modāliem verbiem), prepozicionālām konstrukcijām un sakārtojuma attieksmju aprakstīšanai. Šis modelis var tikt paplašināts arī attiecībā uz pakārtotām un sakārtotām teikuma daļām: palīgteikumu (caur tā centrālo locekli – izteicēju) parasti var piesaistīt kādam no virsteikuma locekļiem kā paskaidrotāju, savukārt vienlīdzīgas teikuma daļas var traktēt analoģiski kā vienlīdzīgus teikuma locekļus (arī šajā gadījumā par saknēm ņemot šo teikuma daļu izteicējus). Taču praksē, likumbāzētā ceļā automātiski analizējot teikumu, šādu konstrukciju atpazīšana radītu ievērojamu sintaktisko daudznozīmību.[14] Tāpēc automātiskā analizatora gramatika pašlaik ir ierobežota, nepārsniedzot vienkārša paplašināta teikuma (vai salikta teikuma daļas) robežas (Bārzdiņš et al. 2007), bet paralēli *SemTi-Kamols* modelis ir attīstīts arī pilnu teikumu aprakstīšanai (Pretkalniņa et al. 2011), un pašlaik tiek veikta pilnā modeļa aprobēšana neierobežotos tekstos, veidojot sintaktiski anotētu latviešu valodas paraugkorpusu (sk. 2.1. nodaļu). Šo korpusu ar laiku varēs izmantot arī statistiska pilnu teikumu parsētāja apmācīšanai.

1.2. Kontrolētā latviešu valoda

Ņemot vērā latviešu valodas sintaktiski un semantiski anotēta tekstu korpusa pašreizējo trūkumu, LUMII ir veikti nopietni pētījumi un iestrādes arī t. s. kontrolētas dabiskās valodas jomā, aplūkojamo latviešu valodas apakškopu sašaurinot gan sintaktiski, gan semantiski. Tādējādi tiek nodrošinātas valodas deterministiskas analīzes iespējas, vienlaikus saglabājot ierobežotās valodas praktisku lietojamību noteiktās sfērās un noteiktiem mērķiem.

Eksistē dažādas visnotaļ izsmalcinātas kontrolētās dabiskās valodas (*CNL – controlled natural language*), lielākoties veidotas kā angļu valodas apakškopas (Wyner et al. 2010). *CNL* parasti nodrošina tās teikumu/tekstu automātiskas interpretācijas iespējas kādā no formālajām valodām, piemēram, pirmās kārtas loģikā vai tās apakškopā – aprakstošajā loģikā. Lai gan loģikā balstīta *CNL* ir intuitīvs un šķietami neformāls zināšanu atainošanas līdzeklis, kontrolētā valoda savā būtībā ir tikpat izteiksmīga kā atbilstošais formālisms un tās interpretācija ir iepriekš paredzama.

Ja *CNL* mērķis ir plašāks – ierobežota mašīntulkošana un dažādi multilingvāli lietojumi, kur interpretācija nenotiek formālas valodas līmenī, bet tikai tulkošanas ekvivalentu līmenī, tad gramatiskie ierobežojumi var nebūt tik strikti. Jebkurā

[14] Piemēram, formāli būtu grūti nošķirt vienlīdzīgus izteicējus no salikta sakārtota teikuma daļām.

gadījumā sintaktiskie ierobežojumi *CNL* lietotājam galvenokārt „uzspiež" tikai precizitāti, konsekvenci un parasti arī vienu un to pašu „pareizo" sintaktiskās analīzes koka variantu (neatkarīgi no konteksta). Taču būtiskākais ierobežojums ir viennozīmīgais leksikons – pilnnozīmes vārdi parasti netiek interpretēti. Tādēļ kontrolētās valodas parasti tiek lietotas vienlaikus tikai viena domēna (nozares) ietvaros, pieņemot, ka tādā gadījumā terminoloģija un tulkošanas ekvivalenti ir viennozīmīgi.

LUMII ir izstrādāts kontrolētās latviešu valodas prototips, kas zināšanu reprezentācijā un formālajā loģikā nepieredzējušiem nozaru ekspertiem un vidusmēra galalietotājiem nodrošina iespēju „lasīt" un „rakstīt" nozarspecifiskas ontoloģijas (Grūzītis, Bārzdiņš 2011), un paralēli tiek strādāts pie ievērojami vispārīgākas gramatikas un atbilstošiem analīzes/sintēzes rīkiem, uz kuru pamata vēlāk būs iespēja elastīgi un salīdzinoši ātri nodrošināt dažādus konkrētus kontrolētās latviešu valodas lietojumus.

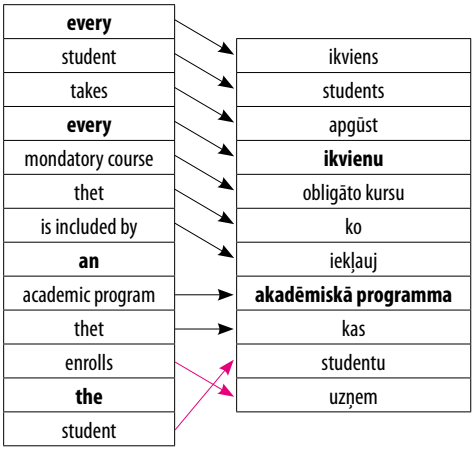
Latviešu valodas ierobežošana, padziļināta pētīšana un formalizēšana noved pie interesantiem secinājumiem, no kuriem vienu apskatīsim sīkāk.

Analītiskā angļu valoda ar noteiktu vārdu secību, vienkāršu morfoloģiju un konsekventu artikulu lietojumu ievērojami atvieglo *CNL* teikumu precīzu interpretēšanu (piemēram, loģikas aksiomu un likumu formā). Taču latviešu valodas gadījumā precīza tulkošana uz formālu valodu vai angļu valodu (un otrādi) nav vienkārša pat ierobežotas valodas gadījumā. Problēmas rada, piemēram, anaforas, kas nominālu vārdkopu (*NP*) gadījumā angļu valodā ir atpazīstamas pēc noteiktā artikula lietojuma, bet latviešu valodā analoģiski „marķieri" vispārīgā gadījumā netiek lietoti. Līdz ar to jautājums ir, kā tulkošanas laikā nepazaudēt, kas ir jaunā informācija (potenciālie antecedenti) un kas ir zināmā informācija (anaforas).

Lai arī latviešu valodā teorētiski būtu iespējams uzspiest mākslīgu „artikulu" lietošanu, izmantojot nenoteiktos un norādāmos vietniekvārdus, tas valodu padarītu nedabisku. Tomēr tas nenozīmē, ka latviešu valodā (un citās sintētiskajās valodās) nav formālu pazīmju, kas palīdz noteikt teikuma informatīvo struktūru; šādas pazīmes ir, bet vispārīgā gadījumā tās teikuma virsējā struktūrā parādās netieši.

Izrādās, ka pastāv sakarība starp zināmo un jauno informāciju un vārdu secību teikumā, ko var formulēt kā teikuma tematiskās un rematiskās daļas mijiedarbību (*TFA – topic focus articulation*) (Hajičová 1993). Tēma ir tas, par ko tiek runāts (zināmais), bet rēma – jaunais, kas tiek pateikts par tēmu. Lai arī neierobežotā latviešu valodā (un ne tikai latviešu valodā) teikuma informatīvā struktūra ne vienmēr tiek atspoguļota ar sistemātiskām izmaiņām vārdu secībā, tika izvir-

5. attēls. **Likums, kas paralēli verbalizēts angļu un latviešu valodā. Sastatījums ilustrē, kā izmaiņas vārdu secībā ietekmē teikuma informatīvo struktūru; angļu valodā mainās sintaktiskās konstrukcijas (piemēram, darāmās kārtas vietā tiek lietota ciešamā kārta)**



zīta hipotēze, ka ierobežotas, loģikā balstītas latviešu valodas gadījumā izmaiņas neitrālajā vārdu secībā var tikt izmantotas kā drošas, intuitīvi apmierināmas formālās pazīmes, tādējādi kompensējot „trūkstošos" artikulus (Grūzītis 2010). Vienkāršotās *TFA* metodes pamatprincips ir šāds: visas *NP*, kas atrodas pirms izteicēja, pieder teikuma tematiskajai daļai, bet izteicējs un visas *NP*, kas seko pēc tā – rematiskajai daļai (sk. 5. attēlu).

Lai pārbaudītu šo hipotēzi praksē, tika izstrādāta ierobežotas latviešu valodas gramatika.[15] Gramatikas analīze pilnībā tika balstīta teikuma informatīvās struktūras analīzē, nodrošinot precīzu, automātisku tulkošanu uz kontrolēto angļu valodu un aprakstošo loģiku (Grūzītis, Bārzdiņš 2011; Grūzītis 2010). Izvēloties reprezentatīvu piemēru kopu (aptverot dažādu veidu un sarežģītības aksiomas un likumus), tika veikta dažādu nozaru pārstāvju anketēšana (piedalījās gandrīz 80 respondenti) (Grūzītis et al. 2010). Aptaujas rezultāti[16] apstiprina, ka vienkāršos vārdu secības nosacījumus dzimtās valodas runātājs apmierina intuitīvi un *TFA* metode līdz ar to ir piemērota zināmās un jaunās informācijas (t. i., koreferenču) automātiskai noteikšanai kontrolētās latviešu valodas teikumos. Turklāt aptauja atspoguļo arī vairākus citus būtiskus aspektus. Piemēram, bagātīgā morfoloģija dod plašas iespējas, kā vienu un to pašu nozīmi var teikumā sintaktiski dažādi realizēt (piemēram, apzīmētāja palīgteikumu var aizstāt ar ekvivalentu

[15] Tiešsaistes demonstrācija un izstrādātās gramatikas ir pieejamas: *http://valoda.ailab.lv/cnl/*
[16] *http://goo.gl/3CdDj* (Aptaujas rezultāti ierobežotai lietuviešu valodai: *http://goo.gl/2q8Gu*)

apzīmētāju: *kurss, kas iekļauts akadēmiskajā programmā → akadēmiskajā programmā iekļautais kurss*). Savukārt nenoteiktie un norādāmie vietniekvārdi („artikulu" lomā) noteiktās situācijās ir vēlami, jo tie var atvieglot teikuma lasāmību un līdz ar to interpretāciju. Tādējādi ir izveidots paralēlu gramatiku komplekts, kas nodrošina iespējami labāku (dabiskāku) kontrolētās latviešu valodas teikumu sintēzi (t. sk. pārfrāzēšanu), vienlaikus nodrošinot lietotāja noformulēto teikumu elastīgu analīzi, akceptējot dažādas pieļaujamās sintaktiskās alternatīvas un nenoteikto/norādāmo vietniekvārdu patvaļīgu lietošanu.

2. LATVIEŠU VALODAS RESURSI

Kā jau tas tika iezīmēts, datorlingvistikā lietotās metodes var iedalīt divās principiālās pieejās: likumbāzētajā, kas saskaņā ar lingvistiski pamatotām teorijām tiecas precīzi aprakstīt valodu kā sistēmu, un statistiskajā, kas, izmantojot salīdzinoši virspusējas statistikas un mašīnmācīšanās metodes, mēģina iegūt valodas modeļa aproksimāciju. Statistika zināmā mērā ļauj izvairīties no nepieciešamības pēc padziļinātas zināšanu formalizēšanas un ļauj vieglāk nodrošināt plašāku valodas pārklājumu. Taču, jo augstāks analīzes līmenis tiek iesaistīts, jo padziļinātākas (strukturālākas) zināšanas ir nepieciešamas.

Praktiskiem lietojumiem visperspektīvākā ir hibrīda pieeja, apvienojot zināšanās bāzētos un statistiskos līdzekļus. Taču jebkurā gadījumā ir nepieciešami resursi, kas formāli reprezentē valodu un zināšanas par to. Šādi resursi ir gan gramatiski un semantiski anotēti tekstu korpusi, gan dažādu veidu mašīnlasāmās vārdnīcas (ne tikai skaidrojošās un tulkojošās, bet arī morfoloģiskās un valenču vārdnīcas), kā arī formālās gramatikas (sk. 1. nodaļu).

2.1. Tekstu korpusi

Ar dažādu veidu, laiku un stilu latviešu valodas tekstu kolekciju digitalizēšanu AILab ir nodarbojusies jau izsenis. No tiem lielākie un brīvi pieejamie ir latviešu folkloras materiāli un latviešu literatūras klasiķu darbi.[17]

No 1995. gada līdz 2001. gadam LUMII piedalījās ES projektā *TELRI* (*Trans-European Linguistic Resources Infrastructure*), kura laikā tika sākta paralēlo tekstu uzkrāšana un marķēšana. Sadarbojoties ar Tulkošanas un terminoloģijas centru, ir uzkrāti angļu un latviešu valodas paralēlie teksti, kuri tiek izmantoti mašīntulkošanas sistēmu izstrādē un testēšanā.

[17] http://valoda.ailab.lv/

Līdz ar Latviešu valodas aģentūras atbalstu kopš 2007. gada pakāpeniski tiek strādāts pie mūsdienu latviešu valodas korpusa izveides[18], savukārt Valsts pētījumu programmas (VPP) informācijas tehnoloģijās un projekta *CLARIN* (Skadiņa et al. 2010) ietvaros ir sagatavoti vairāki papildu korpusi.

Līdzsvarots mūsdienu latviešu valodas tekstu korpuss (~3,5 milj. vārdlietojumu), kompilēts no drukātiem un elektroniskiem materiāliem, kas radīti pēc 1990. gada. Papildus ir sagatavota arī korpusa morfoloģiski marķētā versija, taču jāuzsver, ka marķēšana tika veikta pilnībā automātiski, izmantojot iepriekš minēto *SemTi-Kamols* gramatisko analizatoru – katram vārdlietojumam ir pievienots viens vai vairāki morfoloģiskās analīzes varianti, kurus sintaktiskais analizators šaurākā vai plašākā kontekstā atzinis par iespējamiem.

Latvijas Republikas 5.–9. Saeimas sēžu stenogrammas (~20 milj. vārdlietojumu) ar metadatiem. Korpuss ir strukturāli marķēts, norādot informāciju par runātājiem, sēdēm u. tml.

Latviešu valodas tīmekļa korpuss (~100 milj. vārdlietojumu), kas satur ~700 000 tīmekļa lapu, kas tikušas publicētas pirms 2005. gada (Džeriņš, Džonsons 2007). Arī šis korpuss ir automātiski morfoloģiski marķēts, taču atšķirībā no līdzsvarotā korpusa katram vārdlietojumam ir saglabāts tikai viens analīzes variants. Lai arī, kvalitatīvi vērtējot, šāds risinājums nav pieņemams, ņemot vērā korpusa apjomu, kvantitatīvi tam ir savs praktiskais lietojums.

Jāpiemin arī Latviešu valodas seno tekstu korpuss[19] (~1 milj. vārdlietojumu), kurā ir apkopoti 16.–18. gs. teksti[20] un kurā ir saglabāta (nomarķēta) tekstu oriģinālā struktūra un izklājums: veicot vaicājumus korpusā, katram vārdlietojuma piemēram ir piesaistīta precīza „adrese" oriģinālajā tekstā (piemēram, grāmata → nodaļa → pants vai lappuse → rindiņa) un nepieciešamības gadījumā ir iespējams atvērt attiecīgo konteksta logu (lappusi/rindiņu vai pantu).

Par pieejamajiem latviešu valodas korpusiem un to izmantošanu pēdējā laikā ir tapušas vairākas publikācijas (Andronovs, Andronova 2011; Levāne-Petrova 2011). Par korpusu izmantošanas iespējām LUMII regulāri rīko labi apmeklētus seminārus. Turklāt, kā liecina statistika, izmantojot korpusa pārlūkprogrammu *Bonito*, tiek veikts vairāk nekā 16 000 pieslēgšanās sesiju gadā. Piedāvātos korpusus izmanto gan Latvijas skolas, augstskolas un universitātes, tulkotāji u. c., gan arī vairākas ārzemju universitātes.

[18] *http://www.korpuss.lv/*
[19] *http://www.korpuss.lv/senie/*
[20] Seno tekstu korpuss pakāpeniski tiek papildināts ar jauniem digitalizētiem un apstrādātiem avotiem.

Jaunākais darbības virziens latviešu valodas korpusu attīstīšanā attiecas uz sintaktiski marķēta korpusa izveidi. Šāda korpusa izveidei VPP „Nacionālā identitāte" ietvaros ir izstrādāts *SemTi-Kamols* gramatikas modeļa *PML* (*Prague Markup Language*) profils (Pretkalniņa et al. 2011), kas ļauj izmantot tādu *PML* formātu atbalstošu rīku kā *TrEd*, kas savukārt ir pārbaudīts praksē, veidojot pasaulē vadošo Prāgas atkarību korpusu (*Prague Dependency Treebank*). Šobrīd ir izveidots integrētu rīku un metodikas kopums[21], un ar Latvijas Zinātnes padomes (LZP) atbalstu tas ir aprobēts, sagatavojot sintaktiski anotētu latviešu valodas paraugkorpusu (~200 teikumu) (Pretkalniņa, Levāne-Petrova 2011). Nākotnē ir jāsasniedz ~50 000 teikumu liels korpuss[22], kas būs ļoti nozīmīgs resurss ne tikai gramatikas pētījumiem, bet arī korpusā balstīta statistiskā analizatora izveidei, faktorēto mašīntulkošanas sistēmu attīstīšanai u. c. Turklāt sintaktiski marķēts korpuss ir nepieciešamais sākumpunkts tālākai semantiski marķēta korpusa izveidei.

2.2. Vārdnīcas

AILab ir īstenojusi un turpina īstenot vairākus latviešu valodas skaidrojošo vārdnīcu (gan mūsdienu, gan vēsturisku) digitalizēšanas projektus. Trīs apjomīgākie ir LLVV (Latviešu literārās valodas vārdnīca), SV (Skaidrojošā vārdnīca) un MEV (K. Mīlenbaha un J. Endzelīna Latviešu valodas vārdnīca). Tās ir gan saturiski, gan funkcionāli atšķirīgas vārdnīcas, kas ir pieejamas tiešsaistē.[23]

LLVV (8 sējumi, >64 000 šķirkļu) mašīnlasāmā versija tika izveidota Valsts pētījumu programmas *Letonika* laikā (2006.–2009. gadā). Mašīnlasāmās vārdnīcas versija saturiski pilnībā atbilst oriģinālam (neskaitot nebūtiskas tehniskas atšķirības), taču tā satur detalizētu strukturālu marķējumu, kas ļauj pēc vajadzības vārdnīcu reorganizēt[24] un vizuāli dažādi noformēt. Jāatzīmē, ka ar mašīnlasāmu versiju (atšķirībā no vienkārši digitalizētas vārdnīcas) tiek saprasts strukturālais marķējums, kas apraksta, kādus elementus šķirklis satur, nevis instruē, kā tas būtu vizuāli jāataino (sk. 6. attēlu). Balstoties uz vizuālā noformējuma kopsakarībām, pāreja no vārdnīcas digitālās (ieskenētās) versijas uz mašīnlasāmo ir veikta pusautomātiski – ar vēlāku manuālu pārbaudi un korekcijām (Pretkalniņa, Millere 2010).

SV izveides mērķis ir nodrošināt vispusīgu skaidrojumu ikvienam latviešu valodas tekstos sastopamam (individuālam) vārdam. Pašlaik vārdnīca satur vairāk nekā 173 tūkstošus šķirkļu, kuru sastādīšanai izmantoti vismaz 140 avoti, taču

[21] *http://eksperimenti.ailab.lv/tred/*
[22] Piemēram, Prāgas PDT korpusā un vācu valodas TIGER korpusā ir >50 000 teikumu.
[23] *http://www.tezaurs.lv/*
[24] Piemēram, piedāvāt saturiski vienkāršotu vārdnīcas mobilo versiju: *http://tezaurs.lv/wap/*

6. attēls. **LLVV šķirkļa vienkāršots piemērs digitālā un mašīnlasāmā formā**
(OCR: optical character recognition, RTF: rich text format, XML: extensible markup language)

Šķirklis digitālā formā (OCR → RTF)	**Šķirklis mašīnlasāmā formā (RTF → XML)**
ābece, -es, dsk. ģen. -ču, s. 1. *Mācību grāmata lasīšanas apgūšanai; pirmmācības grāmata.* Bilžu a. Gaiļa ābece – *ābece ar gaiļa attēlu uz vāka.* Mācīties ābeci – *mācīties lasīt.* Morzes ābece – *Morzes sistēmas telegrāfa zīmju kopums*; *Morzes alfabēts*. □ Viņa apsolīja atnest.. savu veco ābeci un man ierādīt [lasīšanu]. Tā bija veca gaiļa ābece, lai gan nebija vairs ne gaiļa, ne pirmās lapas. Birznieks-Upītis IIa, 110. ◊ Ķīniešu ābece – *saka par ko grūti saprotamu*. 2.…	<s> <v> <vf>ābece</vf> <gram>-es, dsk. ģen. -ču, s.</gram> </v> <n nr="1"> <def>Mācību grāmata lasīšanas apgūšanai.</def> <def>Pirmmācības grāmata.</def> <piem> <t>Bilžu ābece.</t> </piem> <piem> <t>Gaiļa ābece.</t> <n><def>Ābece ar gaiļa attēlu uz vāka.</def></n> </piem> ... <fraz> <t>Ķīniešu ābece.</t> <n><def>Saka par ko grūti saprotamu.</def></n> </fraz> </n> ... </s>

tā tiek periodiski papildināta, ņemot vērā arī biežāk meklētos, bet neatrastos vārdus, kas meklēšanas sistēmā tiek automātiski reģistrēti. SV struktūra atšķirībā no LLVV ir vienkāršāka, piemēram, nav doti vārdlietojumu piemēri. Piemēru pievienošana visu vārdu visām nozīmēm, par pamatu ņemot korpusa datus, būtu ļoti noderīga, taču ārkārtīgi darbietilpīga, turklāt esošā līdzsvarotā korpusa apjoms ne tuvu nav pietiekams retu vārdu nozīmju aptveršanai.[25] Taču ilgtermiņā latviešu valodas leksikogrāfijā noteikti būtu jāpāriet no salīdzinoši preskriptīvās pieejas uz deskriptīvu pieeju, t. i., korpusā balstītu vārdnīcu izstrādi, kur primārais atskaites punkts ir vārdlietojumu piemēri.[26] Tas ļautu veidot sistemātiskāku nozīmju dalījumu, ņemt vērā vārdu un nozīmju biežumu u. tml., kas ir ļoti būtiski ne tikai valodas apguvē, bet arī efektīvākā automatizētā analīzē.

[25] Pasaules pieredze rāda, ka tam nepieciešams korpuss ar 100 milj. vārdlietojumu.

[26] Interesanti, ka vienīgā autoram zināmā latviešu valodas vārdnīca, kas pašlaik tiek veidota, pilnībā balstoties korpusā, ir Latviešu valodas vēsturiskā vārdnīca (*http://tezaurs.lv/lvvv/*).

MEV elektroniskā versija ir pakāpeniski izstrādāta un attīstīta ilgākā laika periodā. Vārdnīcas šķirkļu struktūra ir ļoti komplicēta, un lietoto rakstzīmju klāsts ir ļoti plašs. Tās pirmā nozīmīgākā versija tapa ar Latviešu fonda atbalstu (2000.–2002. gadā), padarot MEV pieejamu tīmeklī. Otra nozīmīgāka versija kļuva pieejama 2004. gadā, kad ar LR Izglītības un zinātnes ministrijas (IZM) atbalstu vārdnīca tika konvertēta *Unicode* kodējumā, ļaujot adekvāti atainot vārdnīcā izmantotās rakstzīmes (Nešpore u. c. 2006). Tiesa, pat *Unicode* standarts nepārklāj dažas ļoti reti izmantotas rakstzīmju un diakritisko zīmju kombinācijas.

AILab piedāvāto tīmekļa vārdnīcu popularitāte ir liela: ik gadu vidēji 200 000 apmeklējumu laikā tiek veikti vairāk nekā 500 000 šķirkļu pieprasījumi. Pēdējo trīs gadu laikā AILab vārdnīcas ir izmantotas ~120 valstīs (90 % Latvijā un 10 % citās valstīs).

Tomēr esošo vārdnīcu digitalizēšana un sagatavošana mašīnlasāmā formā nav vienīgais AILab mērķis šajā jomā. Pētot dažādas attieksmes starp vārdu nozīmēm, kas netieši, bet relatīvi sistemātiski parādās vārdnīcās (vārdu skaidrojumos), eksperimentāli ir attīstītas metodes skaidrojošajās vārdnīcās doto zināšanu formalizēšanai semantiska tīkla formā, kas ir viens no aktuālajiem valodas resursiem teksta semantiskajā analīzē (Grūzītis u. c. 2007).

3. EKSPERIMENTĀLI LIETOJUMI

3.1. Mašīntulkošana

Mašīntulkošanas pētījumi AILab tika sākti 1994. gadā, izstrādājot likumos balstītas latviešu-angļu-latviešu tulkošanas sistēmas prototipu *LATRA* (Greitāne 1997). Tā ir interlingvas (starpniekvalodas) sistēma, kas izmanto Lundas universitātes mašīntulkošanas sistēmas *SWETRA* interlingvas reprezentāciju un ir integrējama tajā. Sākotnēji sistēma, līdzīgi kā *SWETRA*, tulkoja laika ziņu un biržas informācijas tekstus, bet vēlāk sistēma pielāgota juridisku tekstu tulkošanai un papildināta ar ierobežotu semantisko komponentu, izmantojot *SIMPLE* ontoloģiju (Skadiņa 2003). Likumbāzētās tulkošanas pieeja LUMII tika attīstīta līdz 2004. gadam, iestrādājot *LATRA* sistēmā semantiskās analīzes moduļus un pielāgojot to jauniem domēniem (nozarēm), piemēram, ES dokumentu tulkošanai (Skadiņa et al. 2010).

2005. gadā ar LZP atbalstu LUMII tika sākti pētījumi statistiskās mašīntulkošanas (SMT) metožu piemērotības novērtēšanā un ir izstrādāts angļu–latviešu statistiskās tulkošanas sistēmas prototips (Skadiņa, Brālītis 2007). Sistēmas apmācīšanai

ir izmantots paralēlais tekstu korpuss (*JRC-ACQUIS*[27]). Sistēmas novērtējums, tulkojot ES dokumentus un izmantojot *BLEU* metriku, ir līdzvērtīgs citām tā laika mašīntulkošanas sistēmām, kas orientētas tulkošanai uz fleksīvo valodu (Skadiņa et al. 2010).

2009. gadā tika sākti pētījumi faktorēto metožu izmantošanā SMT[28] (Skadiņa, Brālītis 2009; Khalilov et al. 2010). Faktorētā modeļa datu sagatavošanai ir izmantots 1. nodaļā minētais statistiskais morfoloģiskais marķētājs. Sistēmai, kurā kā faktori izmantoti vārda pamatforma un vārda morfoloģiskā informācija latviešu valodā, pašlaik ir iegūti labākie rezultāti. Veicot sistēmas pielāgošanu ES juridisko tekstu tulkošanai, izdevās būtiski uzlabot *BLEU* novērtējumu: 50,81 punkts. Taču, mēģinot SMT sistēmu izmantot atšķirīgu tekstu tulkošanā, tulkošanas kvalitāte būtiski pasliktinās (tikai 10–11 *BLEU* punkti), jo sistēma nav pietiekami apmācīta darbam ar šādiem tekstiem. Tāpēc sākti pētījumi, kā automātiski atrast paralēlos tekstus tīmeklī, kas ļautu paplašināt sistēmas lietojamību. Pirmais uzlabojums (*+3 BLEU* punkti) iegūts, būtiski palielinot valodas modelim izmantoto latviešu valodas tekstu daudzumu (Skadiņa 2010).

Pašlaik tiek veidots sintaktiski anotēts paraugkorpuss (sk. 2.1. nodaļu), kas tiks izmantots SMT sistēmas valodas modeļa papildināšanai ar sintaktisko komponentu.

3.2. Runas tehnoloģijas

1990. gadu vidū AILab tika veikti pirmie eksperimenti runas sintēzē un atpazīšanā. Turpmākie pētījumi pamatā tika attīstīti tieši sintēzes (*text-to-speech*) virzienā, izmantojot konkatenācijas jeb segmentu savirknēšanas metodi (Auziņa 2003; Auziņa 2004a; Auziņa 2004b). Taču atsevišķi nelieli projekti tika īstenoti arī saistībā ar runas atpazīšanu, piemēram, 2004. gadā tika izstrādāta programmatūra, ar kuras palīdzību iespējams atpazīt izolētas fonēmas (~34), fonēmu savienojumus un īsus vārdus.

2008. gadā tika apkopotas un būtiski pilnveidotas līdzšinējās iestrādes, ar SIA „Lattelecom BPO" atbalstu uzbūvējot jaunu runas sintēzes sistēmu[29] (Pinnis, Auziņa 2010). Lai arī sistēma tika īpaši pielāgota laika ziņu sintēzei (lietošanai zvanu centros), tā uzrāda labus rezultātus arī neierobežotu tekstu sintēzē. Turpmāko pētījumu uzdevums būtu uzlabot runas prosodijas modelēšanu, kas

[27] http://optima.jrc.it/Acquis/
[28] http://smtdemo.ailab.lv/
[29] http://runa.ailab.lv/

ir viens no runas sintēzes kvalitātes stūrakmeņiem, taču tam ir nepieciešams (marķēts) runātās valodas korpuss.

Ar SIA „Lattelecom BPO" atbalstu tika izveidota arī eksperimentāla runas atpazīšanas sistēma, kas spēj atpazīt 60 bieži lietotus vārdus neatkarīgi no cilvēka balss īpatnībām. Arī runas atpazīšanas sistēmas attīstīšanai ļoti būtisks priekšnosacījums ir runas korpuss. Savukārt specifisku dialogu sistēmu un balss–balss tulkošanas sistēmu veidošanā augstas precizitātes nodrošināšanai būtiska komponente ir lietojumam un nozarei atbilstoša kontrolētās valodas gramatika (sk. 1.2. nodaļu).

2001. gadā tika sākta runātās valodas korpusa izveide, digitalizējot un transkribējot vairākus runas materiālus. Kopš 2010. gada sadarbībā ar SIA „Desol" un SIA „Valodu vēstniecība" ir savākts ievērojams apjoms sarunu valodas ierakstu (~1 milj. vārdlietojumu), kas šobrīd ir atšifrēti (transkribēti) un kuros ir marķētas runai raksturīgās pazīmes, veidojot specializētu korpusu. Nākotnē korpuss būtu jāpapildina ar fonētisko un prosodisko transkripciju.

Nobeigums

Kopš 2008. gada ar LR IZM atbalstu LUMII piedalās Eiropas pētniecības infrastruktūras sadarbības projektā *CLARIN*[30] (*Common Language Resources and Technology Infrastructure*). Šī projekta un iniciatīvas mērķis ir novērst pašreizējo sadrumstalotību valodas resursu un rīku jomā un izveidot integrētu, paplašināmu un sadarbību veicinošu pētniecības infrastruktūru, kas ļautu viegli piekļūt un izmantot valodas resursus un tehnoloģijas gan humanitāro un sociālo zinātņu pētījumos, gan datorlingvistikas lietojumā.

Ņemot vērā konceptuālu un organizatorisku jautājumu risināšanu un situācijas apkopošanu Latvijā, dalība *CLARIN* ir stimulējusi LUMII virzību ceļā uz latviešu valodas resursu un rīku pieejamību, standartizāciju un atvērtību. Domājot par pieejamību ir nodrošinātas vairākas tīmekļa pakalpes (*web services*) attālinātai un automatizētai rīku un resursu izmantošanai[31]: morfoloģiskais analizators un sintezators, vārdu un teikumu dalītājs, statistisks morfoloģiskais marķētājs (tageris), runas sintezators un LLVV mašīnlasāmajā formātā. Tāpat ir veicināta sadarbība ar Latvijas akadēmisko tīklu LANET akadēmiskās identitāšu federācijas izveidē, kas nākotnē atvieglos piekļuvi valodas resursiem, kuru izmantošana ir ierobežota vienīgi akadēmiskiem mērķiem. Tas attiecas uz resursiem, kurus piedāvā ne tikai

[30] http://www.clarin.lv/
[31] http://valoda.ailab.lv/ws/

Latvijas institūcijas (piemēram, MEV, mūsdienu valodas korpuss u. c. problemātiski autortiesību objekti), bet arī citas Eiropas pētniecības institūcijas.

Vienlaikus ar tīmekļa pakalpju nodrošināšanu ir veikta arī vairāku izmantoto datu formātu standartizācija, izvērtējot un izmantojot gan *de iure* standartus, tādus kā *ISO LMF* (*Lexical Markup Framework*), *ISO MAF* (*Morpho-syntactic Annotation Framework*) un *ISOcat* datu kategoriju (terminoloģijas) reģistru, gan *de facto* standartus, piemēram, *MULTEXT-East* morfoloģiskās marķēšanas formātu un *TCF* (*Text Corpus Format*), kas tiek izmantots populārajā *WebLicht*[32] infrastruktūrā sadarbībai starp dažādām (t. sk. dažādu valodu) tīmekļa pakalpēm. Iestrādes standartizācijas jomā ir aktīvi jāturpina attīstīt gan darbā ar latviešu valodas korpusiem, gan vārdnīcām.

Visbeidzot atvērtība nozīmē, ka brīvi pieejami (nekomerciāliem mērķiem) pakāpeniski kļūst ne tikai LUMII rīki, bet arī to pirmkods – plašā nozīmē ietverot resursus (piemēram, morfoloģisko leksikonu). Runājot par rīkiem – ar *GPL licenci*[33] (*General Public License*) ir pieejams *SemTi-Kamols* sintaktiskā analizatora pirmkods, bet ar laiku ir paredzēts šādi sagatavot un licencēt arī morfoloģisko analizatoru, tageri, kontrolētās valodas analīzes un sintēzes līdzekļus utt. Attiecībā uz resursiem – ar *Creative Commons Attribution-NonCommercial-ShareAlike* licenci[34] ir pieejams morfoloģiskais leksikons, bet nākotnē līdzīgi varētu tikt licencētas arī atsevišķas mašīnlasāmās vārdnīcas un tekstu korpusi.

Nobeigumā autors vēlas pateikties kolēģiem – Ingunai Skadiņai, Ilzei Auziņai, Andrejam Spektoram un Baibai Saulītei – par palīdzību šī raksta tapšanā.

[32] *http://weblicht.sfs.uni-tuebingen.de/*
[33] *http://www.gnu.org/licenses/gpl.html*
[34] *http://creativecommons.org/licenses/*

Literatūra

1. Andronova, E., Andronovs, A. Latviešu valodas korpuss un tā izmantošana. No: *Valodas prakse: vērojumi un ieteikumi*. Populārzinātnisku rakstu krājums Nr. 6. Rīga : Latviešu valodas aģentūra, 2011, 41.–57. lpp.
2. Atkins, B. T. S., Rundell, M. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008.
3. Auziņa, I. Segmentu izvēle runas sintēzei. No: *Linguistica Lettica.* Rakstu krājums, Nr. 12, Rīga, 2003, 61.–72. lpp.
4. Auziņa 2004a – Auziņa, I. Datorizēta latviešu valodas fonētiskās transkribēšanas sistēma. No: *Vārds un tā pētīšanas aspekti:* rakstu krājums, Nr. 8. Liepāja : LiePA, 2004, 224.–232. lpp.
5. Auziņa 2004b – Auziņa, I. Latviešu valodas grafēmas-fonēmas atbilsmju likumu sistēma. No: *LZA Vēstis*, Sērija A. Rīga, 2004, 11.–16. lpp.
6. Bārzdiņš, G., Grūzītis, N., Nešpore, G., Saulīte, B. Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In: *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, 2007, pp. 13–20.
7. Bārzdiņš, G., Grūzītis, N., Kudiņš, R., Nešpore, G., Spektors, A. Latviešu valoda semantiskajā tīmeklī. No: *LZA Vēstis*, 60. sējums, 6. numurs, A daļa. Rīga, 2006, 26.–42. lpp.
8. Brants, T. Inter-Annotator Agreement for a German Newspaper Corpus. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 107–112.
9. Chklovski, T., Mihalcea, R. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2003.
10. Chomsky, N. Three models for the description of language. In: *IRE Transactions on Information Theory*, Vol. 2, No. 3, 1956, pp. 113–124.
11. Džeriņš, J., Džonsons, K. Harvesting National Language Text Corpora from the Web. In: Proceedings *of the 3rd Baltic Conference on Human Language Technologies*. Kaunas, 2007, pp. 87–94.
12. Greitāne, I. Mašīntulkošanas sistēma LATRA. No: *LZA Vēstis* Nr. 3/4, 1997, 1.–6. lpp.
13. Grūzītis, N. Word Order Based Analysis of Given and New Information in Controlled Synthetic Languages. In: *Proceedings of the 1st Workshop on the Multilingual Semantic Web* (at WWW 2010), Raleigh (NC), CEUR, Vol. 571, 2010, pp. 29–34.
14. Grūzītis, N., Nešpore, G., Saulīte, B. Verbalizing Ontologies in Controlled Baltic Languages. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 2010, pp. 187–194.
15. Grūzītis, N., Nešpore, G., Saulīte, B. Hierarhisku attieksmju izgūšana no latviešu valodas skaidrojošās vārdnīcas. No: *Vārds un tā pētīšanas aspekti :* rakstu krājums, Nr. 11. Liepāja : LiepU, 2007, 147.–159. lpp.
16. Grūzītis, N., Auziņa, I., Bērziņa-Reinsone, S., Levāne-Petrova, K., Milčonoka, E., Nešpore, G., Spektors, A. Demonstration of Resources and Applications at the Artificial Intelligence Laboratory, IMCS, UL. In: *Proceedings of the 1st Baltic Conference on Human Language Technologies,* Riga, 2004, pp. 38–42.
17. Grūzītis, N., Bārzdiņš, G. Towards a More Natural Multilingual Controlled Language Interface to OWL. In: *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, Oxford (UK), 2011, pp. 335–339.
18. Hajičová, E. *Issues of Sentence Structure and Discourse Patterns*. Prague: Charles University, 1993.

19. Kehoe, A., Gee, M. New corpora from the web: making web text more 'text-like'. In: *Towards Multimedia in Corpus Studies, Studies in Variation, Contacts and Change in English*, Pahta P., Taavitsainen I., Nevalainen T., Tyrkkö J. (Eds.) Vol. 2, 2007.

20. Khalilov, M., Fonollosa, J., Skadiņa, I., Brālītis, E., Pretkalniņa, L. Towards Improving English-Latvian Translation: A System Comparison and a New Rescoring Feature. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, 2010, pp. 1719–1725.

21. Levāne-Petrova, K. Morfoloģiski marķēta valodas korpusa izmantošana valodas izpētē. No: *Vārds un tā pētīšanas aspekti* : rakstu krājums, Nr. 15 (1). Liepāja : LiepU, 2011, 187.–193. lpp.

22. Milčonoka, E., Grūzītis, N., Spektors, A. Natural Language Processing at the Institute of Mathematics and Computer Science: Ten Years Later. In: *Proceedings of the 1st Baltic Conference on Human Language Technologies,* Riga, 2004, pp. 6–11.

23. Nešpore, G., Grūzītis, N., Andronova, E., Spektors, A. K. Mīlenbaha un J. Endzelīna Latviešu valodas vārdnīcas pilnveidota elektroniskā versija. No: *Letonikas pirmais kongress. Valodniecības raksti*. Rīga : LZA, 2006, 241.–249. lpp.

24. Nešpore, G., Saulīte, B., Bārzdiņš, G., Grūzītis, N. Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications,* Vol. 219, IOS Press, 2010, pp. 233–240.

25. Paikens, P. Lexicon-Based Morphological Analysis of Latvian Language. In: *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, Kaunas, 2007, pp. 235–240.

26. Pinnis, M., Auziņa I. Latvian Text-to-Speech Synthesizer. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications,* Vol. 219, IOS Press, 2010, pp. 69–72.

27. Pretkalniņa, L., Levāne-Petrova, K. Preparatory Work for Latvian Treebank. In: *Proceedings of the International Conference on Corpus Linguistics*, St. Petersburg, 2011 (pieņemts publicēšanai).

28. Pretkalniņa, L., Nešpore, G., Levāne-Petrova, K., Saulīte, B. A Prague Markup Language Profile for the SemTi-Kamols Grammar Model. In: *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA),* Riga, 2011, pp. 303–306.

29. Pretkalniņa L., Millere I. Adaptive Authomatic Mark-up Tool for Legacy Dictionaries. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 2010, pp. 147–153.

30. Skadiņa, I., Brālītis, E. Experimental Statistical Machine Translation System for Latvian. In: *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, Kaunas, 2007, pp. 281–286.

31. Skadiņa, I., Brālītis, E. English-Latvian SMT: knowledge or data? In: *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA), NEALT*, Vol. 4, 2009, pp. 242–245.

32. Skadiņa, I. Electronic Dictionaries and Multilingual Information Society. In: *Terminology and Technology Transfer in the Multilingual Information Society*, Termnet Publisher, 2003, pp. 140–146.

33. Skadiņa, I. Angļu-latviešu statistiskās mašīntulkošanas sistēmas izveide: metodes, resursi, un pirmie rezultāti. In: *XI starptautiskā baltistu kongresa „Baltu valodu pagātne, tagadne un nākotne" referātu tēzes*. Rīga : LU, 2010, 166.–168. lpp.

34. Skadiņa, I., Auziņa, I., Grūzītis, N., Levāne-Petrova, K., Nešpore, G., Skadiņš, R., Vasiļjevs, A. Language Resources and Technology for the Humanities in Latvia. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 2010, pp. 15–22.

35. Skadiņš, R., Goba, K., Šics, V. Improving SMT for Baltic languages with factored models. In: *Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 2010, pp. 125–132.

36. Tesnière, L. *Éléments de syntaxe structurale*. Klincksieck, Paris, 1959 (Tulkojums krievu valodā: Теньер Л. *Основы структурного синтаксиса*. Ред. В.Г. Гак. Москва, Прогресс, 1988).

37. Wyner, A., Angelov, K., Barzdins, G., Damljanovic, D., Davis, B., Fuchs, N., Hoefler, S., Jones, K., Kaljurand, K., Kuhn, T., et al. In: *On Controlled Natural Languages: Properties and Prospects*. Revised papers of the Workshop on Controlled Natural Language (CNL 2009), LNAI, Vol. 5972, Springer, 2010, pp. 281–289.

38. Zipf, G. K. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA : Addison-Wesley, 1949.

Latviešu valoda sociālajos tīklos

Līva Brice

Viens otram skolotājs: tviteris kā latviešu valodas spodrinātājs digitālajā vidē

Vispirms bija čivināšana un vīterošana, tad *twīti*, līdz visbeidzot tika nolemts, ka *Twitter* latviski ir *tviteris*, tātad viens tvitera ieraksts ir *tvīts* un rakstīt tvītu ir *tvītot*. (LZA TK ITTEA 2009) Mikroblogu vietnes nosaukuma latviskošana labi ataino latviešu valodas lietojumu digitālajā vidē, sākot ar *galveni* un *kājeni* datorprogrammu tulkojumos, līdz rusismiem un anglismiem čata istabās un interneta vietņu nosaukumos.

Latviešu valodas lietojumu apzināt interneta vidē ir salīdzinoši grūti, jo liela daļa komunikācijas ir privāta, un varam tikai iedomāties, ka oficiālajos e-pastos tiek izmantota gramatiski un stilistiski pareiza valoda, bet vēstulēs draugiem.lv, iespējams, slengs. Mikrobloga vietne tviteris šādā ziņā ir unikāla, jo nodrošina platformu privātai informācijai un saziņai, padarot to publiski pieejamu ikvienam. Tas ļauj piekļūt dažādu vecuma, nodarbošanās un dzīvesvietas tvitera lietotājiem un viņu valodas lietojumam.

Twitter laikam ir domāts valodām, kuru gramatikā ir maz komatu.
Latvietis, kurš cenšas rakstīt gramatiski pareizi, knapi iekļaujas 140 zīmēs.[1]

[1] Turpmāk tekstā atšķirīgā fontā ievietoti piemēri – tvīti no analizējamajiem datiem.

Komunikācija tviterī

Tviteris ir 2006. gada jūlijā izveidota mikroblogošanas vietne. Pēc uzbūves tas ir hibrīds, kas sevī apvieno četrus citus saziņas veidus: blogus, e-pastus, īsziņas un tūlītēju ziņojumapmaiņu. Katrs no šiem atšķirīgajiem veidiem daļēji ir konstruējis tvitera komunikācijas specifiku – mikroblogu galvenā doma ir informēšana par sev aktuālo, gluži kā blogiem, tikai īsākā, koncentrētākā veidā. Caur mikroblogu, tāpat kā caur jebkuru sociālo tīklu, tiek uzturēta saziņa, un veidojas lietotāju tīkls, bet, kas svarīgi, tvitera lietotājam nav jāseko tam, kurš seko viņam, respektīvi, saites tīklā var būt vienpusējas. Tas liek papildus izvērtēt radītās informācijas saturu.

Tviteris funkcionāli ir ļoti vienkāršs – lietotāja uzdevums ir pateikt 140 zīmēs, kas, viņaprāt, būtu jāzina citiem, un šīs 140 zīmes ir lietotāja sevis prezentācija. Tajās var ietilpt atbildes (@), retvīti (RT – *retweet*) – kāda cita tvīta pārpublicēšana, norāde uz privāto ziņu (DM – *direct message*), hipersaites, bildes, video, bet viss teksts nedrīkst būt garāks par 140 zīmēm.

Tvitera lietotāju profilu veido tvīti, fona attēls, profila attēls un trīs pamata lauki – bio (īsa informācija par tvitera lietotāju), atrašanās vietas un interneta adreses. Var tikt izmantoti dažādi papildu lauki lietotāja profilā. Galvenais tvitera lietotāja sevis prezentēšanas līdzeklis ir valoda.

Runājot par latviešu valodu un tviteri, ir gan pozitīvi, gan negatīvi aspekti to mijiedarbībā. 140 zīmes ierobežo gan domas plašumu, gan vārdu daudzumu, tajā pašā laikā liekot domāt konstruktīvāk un izvairīties no liekvārdības.

Kā galvenais negatīvais aspekts ir jāmin tendence rakstīt **bez pieturzīmēm**, jo katrs komats vai punkts ir zīme, tātad paliek mazāk vietas, kur izpausties.

Tā kā viena tvīta tekstuālā informācija ir ierobežota, lai ekonomētu zīmju skaitu, plaši tiek izmantoti **saīsinājumi**, izlaižot vārdā patskaņus, īsinot vietniekvārdus un skaitlisku informāciju aizstājot ar cipariem, piemēram, *vnk* (vienkārši), *kkas* (kaut kas), *1dien* (pirmdien). Tāpat ne tikai digitālās valodas lietojumā, bet arī ikdienā ienāk aizvien vairāk **svešvalodu frāžu saīsinājumi** – *FYI* (*for your information* – tavai informācijai), *LOL* (*laugh out loud* – smieties skaļā balsī), *ASAP* (*as soon as possible* – pēc iespējas ātrāk). Un arī tvitera lietotājiem kā jebkurai sabiedrības daļai ir savs **žargons**. Tipisks tvīts, kurā atrodami visi minētie mīnusi, izskatās šādi:

> Lūdzu RT to info ja nē tad man asap DM kko atsūtīt.

Pozitīvi ir tas, ka tviteris izskauž liekvārdību, jo lietotājs cenšas koncentrēti paust savu domu, kā arī, mēģinot iekļauties 140 zīmēs, pārlasa tekstu, tā, iespējams, sekojot vairāk līdzi valodai. Tomēr svarīgākais tvitera pozitīvais aspekts ir kopiena, kurā tās biedriem ir svarīgi, kā viņus redz citi.

Sociālo tīklu pētniece Dana Boida (*Danah Boyd*), neaplūkojot visu interneta vidi un tās tendences, bet gan koncentrējoties uz sociālo tīklošanas vietņu veidoto sabiedrību, runā par „tīklotu publiku" (*networked publics*)[2] un apgalvo, ka sociālie tīkli ar savu specifiku ir izveidojuši atsevišķu sabiedrību. Viņa tīkloto sabiedrību definē kā „sabiedrību, kuru restrukturizē tīklotās tehnoloģijas", tādā veidā šīs sabiedrības darbības vieta ir konstruēta ar tīkla tehnoloģiju palīdzību, un iedomātā kopības sajūta rodas cilvēku, tehnoloģijas un prakses krustpunktā. (Boyd 2011, 39) Neaplūkojot tīkloto sabiedrību kā vienu veselumu, bet gan sadalot to mazākās vienībās, respektīvi, konkrētās sociālās tīklošanas vietnēs, šajā gadījumā tviteri, var aplūkot tvitera kopienu kā sabiedrību, kurā tās dalībnieki ietekmē viens otru, līdz ar to arī valodas lietojumu un izvēlēto stilu.

Analīze

Pastāvot drukātajiem plašsaziņas līdzekļiem, radio un TV, par valodas lietojumu bija jāuztraucas tikai sadzīvē un konkrētās grupās. Ienākot internetam indivīda ikdienā, valodas lietojumu un attīstības tendences kļuvis aizvien grūtāk pārraudzīt. Bet, lai valoda tiktu kopta un sekotu līdzi laikam, svarīgi ir izprast, kā jaunajos interneta komunikācijas veidos tiek lietota latviešu valoda, vai tās lietojumam tiek pievērsta uzmanība un vai lietotāji ir ieinteresēti kopt latviešu valodu. Tādēļ par šī pētījuma mērķi tika izvirzīts noskaidrot, vai un kā par latviešu valodu tiek runāts tviterī.

Tviteris pēc savas uzbūves un informācijas organizēšanas veida ir ļoti parocīgs jebkāda veida tekstuālai analīzei, jo pētniekam ir pieejams plašs datu apjoms digitālā veidā. Ikviens interneta lietotājs, izmantojot tvitera vai tā aplikāciju piedāvātās iespējas, var veikt visvienkāršāko kontentanalīzi.

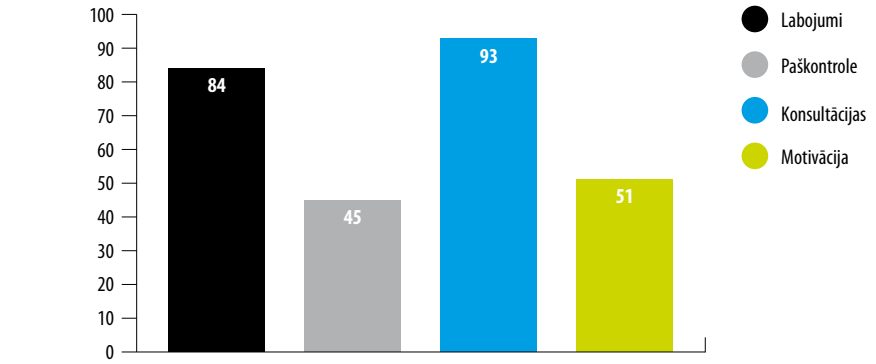
Šajā darbā tika veikta tvītu kontentanalīze laika posmā no 2010. gada janvāra līdz 2011. gada martam. Izmantojot *Google* reāllaika meklētāju, tvītos tika meklēti vārdi „gramatika", „pareizi", vārdu savienojums „latviešu valoda". Kopā tika atlasīti 273 tvīti, kuri, balstoties uz to saturu, tika sagrupēti četrās kategorijās (sk. 1. attēlu).

Analīzes mērķis bija noskaidrot, vai, cik daudz un kādā kontekstā digitālajā vidē tiek pievērsta uzmanība valodas lietojumam. Tāpat ir būtiski atcerēties, ka, runājot par tvitera kā kopienas ietekmi, lietotāju rīcību, piemēram, valodas lietojuma jautājumos, Latvijas tvitera lietotāji nav grupēti vai klasificēti. Nav veikti oficiāli pētījumi par to, kas ir Latvijas tvitera lietotājs, tomēr daļa tā lietotāju tiek uzskatīti par sava veida inteliģenci.

---

[2] D. Boidas piedāvātais termins *Networked publics* nav saistīts ar M. Kastela vai J. Van Dijka *Network Society*. D. Boidas termins darbā tiek tulkots kā *tīklotā sabiedrība*, ar to domājot sabiedrību, kas ir veidojusies sociālajos tīklos.

1. attēls. **Tvītu sadalījums kategorijās**



Tomēr tā kā tviterī ir vērojama pudurošanās, citiem vārdiem sakot, slāņošanās un dažādu grupu veidošanās, tad jāatceras, ka rakstā paustās atziņas un tendences nav attiecināmas uz visu tvitera kopienu, bet gan konkrētu tās daļu, kurai ir svarīga piederība kopienai un sevis prezentēšana. Raksturot un tuvāk analizēt precīzu aplūkotās tvitera sabiedrības daļu, neveicot lietotāju aptauju, nav iespējams.

Konsultācijas

Skaitliski lielāko kategoriju (34 %) veido konsultācijas. Kā iepriekš minēts, tviterī norisinās viens–pret–daudziem komunikācija, tā radot iespēju lietotājam ar vienu tvītu iegūt daudzu cilvēku viedokli kādā jautājumā. Sekotājiem bieži tiek uzdoti jautājumi par latviešu valodas gramatiku, piemēram, precizējot, kā tiek lietots kāds konkrēts vārds. Tas bieži vien izraisa diskusijas, kurās lietotāja sekotāji, balstoties uz savām zināšanām vai izmantojot uzziņu literatūru, izskaidro, kāpēc konkrētajā situācijā ir jālieto tieši tas vārds vai vārdforma.

Kā saka pareizi no lv val gramatikas viedokļa – 16 gadus veca, 16 gadus jauna, vai 16 gadīga ??????????????

gramatiski pareizi ir Kankuna vai Kankūna?

Rakstu tekstu un domāju: vārds „tostarp" rakstiski izskatās dīvaini, tomēr sarunvalodā lietots. Vai tas ir gramatiski korekts?

Šādi jautājumi latviešu vidē rosina gramatikas atkārtošanu un diskusijas, bet būtiski, ka tā ir arī iespēja tiem, kuriem latviešu valoda nav dzimtā valoda, pavaicāt

dzimtās valodas runātājiem, nevis sarežģītām grāmatām, kā un, galvenais, kāpēc kāds vārds ir jālieto vai kāds likums jāievēro.

> tie kas saprot latviešu valodas gramatiku, kā ir pareizi: „karte dot iespēju" vai „karte dod iespēju"?

Aizvien biežāk sociālajās tīklošanas vietnēs, tajā skaitā tviterī, tiek diskutēts par jaunvārdiem, piemēram, 2011. gada februārī plaši tika apspriests, vai vārds *krējumelis* ir piemērots, un tika piedāvāti varianti, kā vēl dēvēt krējuma izstrādājumu.

Tviteris kalpo par konsultāciju vietu lietotājiem, tā pilnīgojot arī viņu zināšanas par valodām.

Labojumi

Procentuāli otrais lielākais tvītu veids bija labojumi (31 %) – kāda esoša uzrakstīta tvīta labojums. Tiek izlabots konkrēts vārds, norādot uz gramatiski pareizo rakstību:

> *Gramatiski* pareizi – „viltusprofils"…

Tiek skaidroti interpunkcijas likumi, piemēram, paskaidrots komata lietojums:

> man garšo rums ;) Pirms „lai" nevajadzēja komatu :D Labi, labi - man vnk patīk, ka viss ir gramatiski pareizi :)

Kā arī tiek izlabots vārds un paskaidrots, kāpēc kāds vārds ir jālieto konkrētajā formā:

> Gramatiski pareizi būtu „Ulžu", bet labskanības dēļ pieņemts lietot „Uldu". „Kūts" daudzskaitļa ģenitīvā jau arī skan dīvaini :)

Lai norādītu uz kļūdām, tiek izmantota sarakstes forma, tvītus rakstot kā atbildi (@) vai tajā minot lietotāja vārdu, kurš ir kļūdījies.

Ļoti aktīvi gramatikas kļūdām tiek sekots līdzi sabiedrībā pazīstamu cilvēku, uzņēmumu, organizāciju un partiju profilos, pievēršot uzmanību faktam, ka šis ir publisks izteiksmes veids un ir jāseko līdzi, lai valodas lietojums būtu korekts. Tā kā analīzes periodā ietilpa arī vēlēšanu laiks, tad daudzi labojumi tika veltīti partiju un deputātu kandidātu tvitera kontiem, norādot, ka tviteris ir tikpat svarīgs publisks izteiksmes veids kā intervija žurnālā. Publicējot gramatiski nepareizu tvītu, iepriekšminētie saņēma nosodošus komentārus no tvitera lietotājiem, minot, ka *tautas kalpiem* ir jāmāk rakstīt pareizā latviešu valodā.

Labojumu lielais īpatsvars, kā arī vēlēšanu atblāzma tviterī parādīja, ka lietotāji paši cenšas uzturēt *tīru* un gramatiski pareizu latviešu valodu.

Motivācija

Ja iepriekšējās kategorijas bija balstītas vairāk uz savstarpēju, tiešu lietotāju komunikāciju, tad paškontrole un motivācija ir saistīta ar lietotāja tēlu kopienā.

Motivācija var tikt saistīta ar jau iepriekš minēto iekļaušanos grupā, sevi prezentējot atbilstoši grupas vēlmēm. Tvitera gadījumā, runājot par valodu, sevis prezentēšanas vietā precīzāk būtu lietot terminu „sava iespaida pārvaldīšana" (*impression management*), kas aplūko, kā citu viedokļi un veids, kā tie redz indivīdu, ietekmē indivīda uztveri un veidu, kā mēs izvēlamies sevi prezentēt citiem. (Chester, Bretherton 2007, 225)

Alise Mārvika (*Alice E. Marwick*) un D. Boida uzsver, ka tvitera lietotāji raksta kognitīvi konstruētai auditorijai, iedomātai grupai, kura, iespējams, nemaz nelasa lietotāja radīto saturu. (Marwick, Boyd 2010, 2) Arī Luisa Anna Šeita (*Lois Ann Scheidt*) skaidro, ka iedomātā auditorija eksistē tikai rakstītajā tekstā caur stilistiskām un lingvistiskām izvēlēm (Scheidt 2006, 197–201) un tvitera valodas lietojuma gadījumā šīs izvēles ir nozīmīgas auditorijas acīs.

Lasu gramatiski nepareizos twitterdraudziņu „tvītus" un gribu sev acis izskrāpēt.

un vēl esmu sapratusi to, ka daudz simpātiskāki man šķiet cilvēki, kas māk „runāt" – gan skaisti, gan gramatiski pareizi.

Mēģināju atrast kādu, kam sekot, bet, ****, tik daudz idiotu, kuri pat nevar gramatiski pareizi uzrakstīt vienkāršus vārdus.

Man patīk lasīt saturīgus un gramatiski kvalitatīvus tvītus. Tagad, draugi, nokaunieties par to, uz ko liekat manām acīm skatīties! Labunakt!

Iepriekš piedāvātie piemēri t. s. (pareizrakstības) motivācijas tvītiem parāda, kāda ir citu lietotāju attieksme pret gramatiski nepareiziem tvītiem, tomēr jāatceras, ka motivācija un arī paškontrole darbojas tikai tādā gadījumā, ja indivīdam ir svarīga iekļaušanās tvitera kopienā un/vai savs tēls.

Paškontrole

Tvitera lietotājs nevar noteikt savu tvītu auditoriju, jo, ja vien viņš nav izvēlējies uzlikt aizsardzību – liegt publisku pieeju tvītiem, atļaujot tos lasīt tikai lietotāja apstiprinātiem lietotājiem, tad tie ir pieejami ikvienam interneta lietotājam. Šāda tvītu pieejamība ikvienam var ietekmēt konkrētā indivīda reālās dzīves tēlu, piemēram, darba devējs pirms darba intervijas aplūko lietotāja tvitera profilu, kas var likt apšaubīt kandidāta spējas un zināšanas. Šī iemesla dēļ tvitera lietotājiem

ir ļoti augsta paškontrole, kas tvītotājiem liek atvainoties par gramatikas kļūdām vai skaidrot, kāpēc tādas ir pieļautas, tādā veidā it kā nepasliktinot savu tēlu.

> Esmu šausmīgākais i-ego-ideālists. Ja kaut kas manā profilā daudzajās e-vidēs nav gramatiski pareizi, tad cenšos to ātri labot. Kārtīgais…

> Izlasot savu pēdējo ierakstu saprotu, ka jaiet cucēt pārāk liels nogurums un nespēja salikt vienu teikumu gramatiski pareizi :D ar sniegu!

> Un es jau 5 reizes pārlasīju savu tikko iečivināto tekstu un nespēju saprast vai tas vispār ir gramatiski pareizi uzrakstīts!

Bērnijs Hogans (*Bernie Hogan*) paškontroli skaidro kā līmeni, līdz kādam publicētā informācija ir normatīvi pieņemama. Viņš to dēvē par zemāko kopsaucēju (*lowest common denominator*) (Hogan 2010, 383), kas ir publiski izpaužamā informācija, kuru indivīds publicēs, nekaitējot savam tēlam.

Kā iepriekš aplūkots, tviterī tiek laboti neprecīzi tvīti un notiek konsultācijas par to, kā pareizi veidot teikumu, kā pareizi rakstīt. Par pareizrakstības normu neievērošanu tvitera lietotāji, kuriem ir svarīgs savs tēls, atvainojas.

> Žonglēju starp 3 valodām, tapēc atvainojiet, ja tweeti ir dramatiski, gramatiski iedragāti…

Nobeigums

Kultūras ministre Sarmīte Ēlerte, runājot par latviešu valodu digitālajā vidē, ir teikusi: „Svarīgi, lai internetā dzīvotu latviešu valoda visā savā būtībā un dziļumā!" (LTV 2010) Šāds apgalvojums liek pievērst uzmanību faktam, ka svarīgi ir ne tikai lietot pareizu latviešu valodu, bet arī diskutēt par to. Pateicoties datorā pieejamajām pareizrakstības pārbaudēm, indivīdam vairs nav jāatceras, kā pareizi rakstāms vārds *istaba* vai pirms kuriem saikļiem liekams komats, jo dators to izdara lietotāja vietā. Tieši tāpēc diskusijas, kopīgas *prāta vētras* par kāda vārda lietojumu un gramatikas grāmatu „pāršķirstīšana", lai atbildētu uz kāda tvītu, ir svarīgs aspekts, lai valoda digitālajā vidē dzīvotu (sk. 2 attēlu). Tviteris vairāk nekā citi interneta komunikācijas rīki parāda latviešu valodas lietošanas kultūru digitālajā vidē un ļauj to analizēt.

Ir svarīgi saprast, ka tviteris kā komunikācijas rīks ir starp privāto un publisko komunikācijas formu. Latvijas gadījumā izpratne par tviteri kā publisko komunikācijas rīku nav līdz galam nostiprinājusies, līdz ar to tas tiek izmantots vairāk privātai saziņai un personīgu domu, emociju paušanai. Tieši šī iemesla dēļ lieto-

2. attēls. **Latviešu valodas lietošanas aspekti tvitera vidē**

tājiem ir vairāk jādomā gan par saturu, gan par valodas lietojumu, jo vēstījumi ir fiksēti rakstītā formā un ar tvitera tehniskajām iespējām var tikt viegli izplatīti ārpus tvitera robežām.

Latviešu valodas kvalitātes un kultūras tviterī salīdzināšana ar *ielas* valodu būtu nepareiza, tomēr valodas lietojuma jautājums ir tieši saistīts ar lietotāja vēlmi iekļauties konkrētajā grupā, kā arī ar sevis prezentēšanu, jo lietotāja konts var tikt saistīts ar personas reālo dzīvi.

Kā parādīja tvītu analīze, tvitera lietotāju vidū ir tādi, kuriem ir svarīga gramatiski pareizas latviešu valodas lietošana digitālajā vidē un kuri runā par valodas kvalitāti. Tas liek domāt, ka konkrētais sociālais tīkls, tā formāts un lietotāji var stiprināt un kopt latviešu valodu, kā arī kļūt par interesantiem un informācijas bagātiem pētījuma objektiem digitālās latviešu valodas pētniecībā.

Mācība – cilvēki retvītos tikai gramatiski pareizus tweetus -> twitteris attīsta pareizrakstību.

Literatūra

1. Boyd, D. Social Network Sites as Networked Publics: Affordances, Dynamics and Implications. In: Papacharissi, Z. (ed.). *Networked Self: Identity, Community and Culture on Social Network Sites*. New York: Routledge, 2011, pp. 39–58.
2. Chester, A., Bretherton, D. Impression Management and Identity Online. In: Oxford *Handbook of Internet Psychology*. Oxford : Oxford University Press, 2007.
3. Hogan, B. The Presentation of Self in the Age of Social Medi: Distinguishing Performances and Exhibitions Online. In: *Bulletin of Science, Technology & Society*, 30(6), 2010, pp. 377–386.
4. LTV raidījums „Labrīt, Latvija!" [tiešsaiste]. Rādīts 2010. gada 16. novembrī. [Skatīts 2011. gada 3. martā.] Pieejams: *http://www.youtube.com/watch?v=Et8Xoq1Dz5Y*
5. *LZA TK ITTEA. Twitter turpmāk būs jāsauc tviteris* [tiešsaiste]. Raksts publicēts 2009. gada 2. novembrī. [Skatīts 2011. gada 3.martā.] Pieejams: *http://datuve.lv/raksts/2990/Twitter_turpmak_bus_jasauc_tviteris/*
6. Marwick, A. E., Boyd, D. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. In: *New Media & Society,* 20(10), 2010, pp. 1–20.
7. Scheidt, L. A. Adolescent Diary Weblogs and the Unseen Audience. In: Buckingham D, Willett R. (ed). In: *Digital Generations: Children, Young People and New Media.* New Jersey: Lawrence Erlbaum Associates. 2006, pp. 193–210.

Uldis Bojārs

Sociālā tīmekļa satura apkopošana un analīze

Pēdējā laikā cilvēki arvien aktīvāk izmanto sociālā tīmekļa pakalpojumus (blogus, sociālos tīklus, mikroblogošanas servisus u. c.). Šis raksts apskata sociālā tīmekļa izpēti, koncentrējoties uz tās sākumposmu – satura vākšanu un tā sākotnējo analīzi. Raksta pamatā ir autora pieredze mikroblogu servisa *Twitter*[1] ziņu vākšanā un analīzē projekta „Nacionālā identitāte digitālajā vidē" ietvaros.

Ievads

Sociālais tīmeklis ir tīmekļa lietojumu kopums, kuru galvenā vērtība ir lietotāju piedalīšanās, to radītais saturs un savstarpējā sadarbība (O'Reilly 2005). Atšķirībā no parastām tīmekļa lapām lietotāji var piedalīties sociālajā tīmeklī, pievienojot savu saturu (piemēram, blogos), atzīmējot viņus interesējošo saturu (tīmekļa grāmatzīmju servisos) vai komunicējot ar citiem lietotājiem (sociālo tīklu servisos). Labs piemērs sociālā tīmekļa popularitātei ir sociālais tīkls *Facebook*, kura lietotāju skaits 2011. gadā sasniedza 750 miljonus.[2]

Sociālais tīmeklis ļauj pētniekiem vērot cilvēku uzvedību virtuālajā vidē un automātiski apkopot un analizēt lielus datu apjomus, tādēļ tas ir labs pētniecības datu avots. Īpaši piemēroti izpētei ir atvērtā tipa sociālā tīmekļa servisi, kuru saturs ir tīmeklī publiski pieejams.[3] To priekšrocības ir divējādas: 1) to saturu var uzskatīt par publisku komunikāciju interneta vidē; 2) publiski pieejamu saturu ir vienkāršāk apkopot un pētīt.

Tviteris (*Twitter*) ir pasaulē lielākais mikroblogošanas serviss, kurā lietotāji var publicēt īsas, līdz 140 simbolu garas ziņas (kā arī apmainīties ar šīm ziņām savā starpā). 2011. gadā tvitera lietotāju skaits sasniedza 200 miljonus.[4] Tvitera servisam ir raksturīgs liels ziņu skaits, gandrīz momentāna ziņu publicēšana, turklāt tā saturs ir „nefiltrēts" (t. i., ziņas pirms publicēšanas tajā netiek īpaši filtrētas vai

[1] Šajā rakstā, atkarībā no konteksta, tiek lietots īpašvārds, firmas un mikroblogu servisa nosaukums *Twitter* vai LZA Terminoloģijas komisijas ieteiktais latviskojums „tviteris".

[2] http://latimesblogs.latimes.com/technology/2011/07/watch-facebooks-new-product-announcement-live.html

[3] *Facebook* ir piemērs aizvērtā tipa servisam, jo lielākā daļa lietotāju radītā satura tajā nav publiski pieejama un ir redzama vienīgi ierobežotam lietotāju lokam.

[4] http://www.blogherald.com/2011/07/17/twitter-infographic-zero-to-200-million-users-in-five-years/

rediģētas). Tādējādi tajā izpaužas tieša, nepastarpināta un publiska lietotāju komunikācija par viņiem aktuālām tēmām.[5]

Tvitera analīzi vēl vieglāku padara tvitera piedāvātie programmatūras interfeisi (*API*), kuri ļauj automātiski iegūt un apstrādāt tajā esošo informāciju.

Pētījuma apraksts

Rakstā ir apkopota latviešu valodā pieejamo tvitera ziņu vākšanas un analīzes pieredze, kas gūta, analizējot komunikāciju tvitera vidē 10. Saeimas vēlēšanu laikā. Turpmākajās nodaļās tiek aplūkoti datu vākšanas un analīzes tehniskie aspekti. Detalizēta 10. Saeimas vēlēšanu tvitera ziņu analīze ir pieejama atsevišķā šim pētījumam veltītā rakstā par politisko ziņu pragmatiku tvitera komunikācijā (Šķilters et al. 2011).

Lai pētītu ar 10. Saeimas vēlēšanām saistītās ziņas, par pētījuma pamatu tika izraudzīta tvitera lietotāju kopa, kurai varētu būt svarīga nozīme vēlēšanu komunikācijā (politiskās partijas, politiķi, deputātu kandidāti; mediju organizācijas un citi lietotāji, kuri raksta par politiku un vēlēšanām; aktīvākie Latvijas tvitera lietotāji). Apkopojot šajā kopā ietilpstošo 1377 lietotāju vienas nedēļas laikā (28.09.2010.–04.10.2010.) radītās ziņas, tika iegūts tvitera ziņu korpuss, kuru veido 50 032 vienības (tvitera ziņas jeb tvīti). Vēlāk tika izveidots paplašināts korpuss, kurā ir sešu nedēļu dati (23.09.2010.–03.11.2010.) un vairāk nekā 238 000 ziņu. Tālāk rakstā tiek izmantoti piemēri no šeit aprakstītā vienas nedēļas tvitera ziņu korpusa.

Rakstā aprakstītās metodes ir piemērojamas arī citiem sociālā tīmekļa datu vākšanas un analīzes projektiem, tomēr daļa no norādēm, piemēram, datu vākšanas kritēriju izvēle var mainīties atkarībā no katra projekta specifikas.

Datu vākšana

Vispirms ir jāizvēlas datu vākšanas principi. Galvenie jautājumi, uz kuriem pētniekiem jāatbild, ir:

- vācamās informācijas veids (tvitera ziņas, sociālā tīkla informācija u. c.);
- informācijas atlases kritēriji (piemēram, pēc atslēgas vārdiem, noteiktas lietotāju kopas radītas ziņas, visas ziņu straumes izlase).

[5] Tvitera lietotājiem ir iespēja ierobežot savu ziņu pieejamību, tādēļ daļa ziņu nav publiski redzamas. Tomēr lielākā daļa tvitera ziņu ir publiskas.

Šajā rakstā tiek apskatīta tvitera ziņu vākšana. Pirmkārt, atkarībā no dotā uzdevuma ir jāizvēlas informācijas atlases kritēriji. Biežāk izmantotie datu vākšanas kritēriji ir: a) visas ziņas, kas atbilst dotajiem atslēgas vārdiem; b) visas noteiktas lietotāju kopas radītās ziņas; c) visu tvitera ziņu izlase.

Twitter piedāvā vairākus *API* (*Application programmimg interface*, programmatūras interfeisus) automātiskai informācijas iegūšanai:

- *REST API*[6] – sniedz informāciju par tvitera ziņām, lietotāju kontiem un citiem sistēmas pārziņā esošiem objektiem;
- *Search API*[7] – meklēšanas interfeiss, kas sniedz informāciju par meklēšanas kritērijiem atbilstošām ziņām;
- *Streaming API*[8] – „straumēšanas" interfeiss, kas sniedz nepārtrauktu pieeju gandrīz momentānai (*real-time*) informācijai par pēdējām ziņām, kas atbilst dotajiem nosacījumiem.

Lielākajai daļai pētniecības uzdevumu ir piemērots *Streaming API*, kurš nodrošina nepārtrauktu kritērijiem atbilstošo ziņu saņemšanu. Šim *API* izsaukumam tiek nodots saraksts ar datu vākšanas parametriem, kas ir atkarīgi no konkrētā uzdevuma un visbiežāk ir tvitera kontu saraksts (*follow* parametrs) vai atslēgas vārdu saraksts (*track* parametrs). Tviteris nodrošina arī gadījuma ziņu izlasi (~1 % ziņu visiem lietotājiem un ~10 % ziņu lietotājiem, kam ir *Gardenhose* pieeja) un iespēju meklēt ziņas pēc ģeogrāfiskās atrašanās vietas, tomēr tas nav ērti izmatojams, lai atrastu latviešu ziņas, jo ļoti nelielu daļu no kopējās tvitera ziņu straumes veido teksti latviešu valodā.

Pētījuma vajadzībām visatbilstošākā bija iespēja sekot noteiktas lietotāju kopas publicētajām ziņām. Sākotnējā sarakstā tika iekļauti 179 lietotāji, kuriem varētu būt svarīga nozīme priekšvēlēšanu un vēlēšanu laika komunikācijā (politiskās partijas, politiķi, deputātu kandidāti; mediju organizācijas un citi lietotāji, kuri raksta par politiku un vēlēšanām; aktīvākie Latvijas tvitera lietotāji). Saraksts tika papildināts 1) ar lietotājiem, kuri pārpublicēja sākotnējā sarakstā esošo lietotāju ziņas vai sarunājās ar tiem; 2) ar citiem kontiem, kurus ieteica nozares speciālisti. Tādējādi sarakstā tika ietverti 1377 tvitera lietotāju konti.

Lai varētu izmantot tvitera *API*, ir nepieciešama atbilstoša tvitera klienta bibliotēka, kas spēj sazināties ar serveri un nodrošina tvitera ziņu saņemšanu. Šim nolūkam tika izmantota *Tweepy* bibliotēka[9] (*Python* programmēšanas valodai) un izveidota programma, kas saņemtās ziņas saglabā katru savā failā *Twitter JSON* (*JavaScript Object Notation*, *JavaScript* objektu pieraksts) ziņu formātā. Tā kā izvēlētā programmēšanas

[6] http://dev.twitter.com/pages/api_overview
[7] http://dev.twitter.com/doc/get/search
[8] http://dev.twitter.com/pages/streaming_ap
[9] http://code.google.com/p/tweepy/

valoda nodrošina darbu ar *Unicode* tekstiem, nebija grūtību saņemt un apstrādāt latviešu un citu valodu tekstus. Sarežģījumi, kas radās komunikācijā ar tvitera *Streaming API*, tika novērsti, izmainot programmā iebūvētos aiztures laikus.

Pēdējais datu apkopošanas darbu posms ir savākto datu integritātes pārbaude. Tā kā savienojums ar tvitera serveriem reizēm mēdz pārtrūkt, tad ir jāpārbauda, vai savāktajā datu kopā nav pārtraukumu. Ja tādi ir, tad tie ir laikus jānovērš, savācot trūkstošos datus, izmantojot tvitera *API* līdzekļus.

Datu vākšanas posmu var apiet, ja ir iespēja izmantot citu savāktos datus. Viena no šādām datu kopām ir mūsu savāktais 10. Saeimas vēlēšanu korpuss (Šķilters et al. 2011). Lielāku latvisko tvitera ziņu korpusu ir savākusi tvitera datu analīzes kompānija SIA „Soon". Šis korpuss ietver lielāko daļu Latvijas tvitera lietotāju kontu, un 2011. gada februārī tajā bija informācija par vairāk nekā 41 tūkstoti lietotāju[10], no kuriem 12 778 bija aktīvi tvitera lietotāji.[11]

Datu priekšapstrāde

Pirms uzsākot tālāku analīzi, jāveic saņemto datu priekšapstrāde. Priekšapstrādes procedūras ir atkarīgas no tā, kāda veida datu analīze tiks veikta. Datu analīzes uzdevumus var iedalīt divās grupās:

a) satura analīze, kurā tiek analizēts ziņu teksts;
b) strukturālā un metadatu analīze, kurā tiek pētīti tvitera ziņās ietvertie un tām pievienotie metadati (informācija par minētajiem tvitera kontu vārdiem u. c).

Abu veidu analīzes uzdevumiem ir jāsagatavo dati tiem piemērotos formātos un jāizstrādā rīki darbam ar datiem. Strukturālā un metadatu analīze neprasa īpašus sagatavošanās darbus (tviteris jau piegādā strukturētus datus par katru ziņu vienību), tādēļ tās galvenais priekšapstrādes uzdevums ir datu saglabāšana analīzei piemērotā rīkā. Tā kā tiek analizēti lieli datu apjomi, informāciju par tvitera ziņām ir ieteicams ielādēt datubāzē. Tika izraudzīta *MongoDB*[12] datubāzu vadības sistēma, kura nodrošina darbu ar *JSON* datu ierakstiem (tas ir viens no galvenajiem datu formātiem, ko izmanto tviteris) un kuru ar *PyMongo*[13] bibliotēkas palīdzību ir iespējams izmantot *Python* programmēšanas valodā.

Satura analīzei domātā ziņu teksta priekšapstrādei izmantota *NLTK* (*Natural Language Toolkit*) bibliotēka[14] (Bird et al. 2009) un raksta autora izstrādāta

[10] https://twitter.com/so_on/status/33160124864544768
[11] https://twitter.com/so_on/status/33160786746671104
[12] http://www.mongodb.org/
[13] http://api.mongodb.org/python/1.11/
[14] http://www.nltk.org/

programmatūra. Datu priekšapstrādē ietilpa saņemto ziņu nolasīšana, teksta sadalīšana vārdlietojumos (*tokens*), to attīrīšana un atslēgvārdu aizvietošana.[15] Datu priekšapstrādes laikā tika nolasītas *JSON* formātā esošās tvitera ziņas, sadalot tās vārdos, izmantojot *NLTK* bibliotēkas *WhitespaceTokenizer*. Pēc tam tika dzēstas palikušās pieturzīmes (*. , : ; ! ?*). Tā kā *NLTK* bibliotēka nenodrošina automātisku *Unicode* starptautisko simbolu (t. sk. latviešu valodas) apstrādi, *UTF-8* kodējumā esošie simboli tika pārvērsti *Unicode* formā, izmantojot *Python decode* funkciju.

Latviešu valodas tvitera tekstu analīzi apgrūtina vairāki faktori. Pirmkārt, interneta vidē latviešu valodas vārdus mēdz pierakstīt gan ar pareizām diakritiskajām zīmēm (*ā*), gan tās transliterējot (*aa*), gan ignorējot (*a*). Otrkārt, latviešu valoda ir fleksīvā valoda, resp., vārdi funkcionē dažādos locījumos. Tādēļ pirms tekstu analīzes jānovērš dažādi vārdu pieraksta veidi.

Mūsu rīcībā nebija automātiskā vārdu morfoloģiskā analizatora un lemmatizētāja, kas varētu vārdus pārvērst pamatformā, tādēļ tika izveidota tabula pētījumam svarīgāko vārdu dažādu formu aizvietošanai ar vienu atslēgvārdu, piemēram, „vēlēju" tika aizvietots ar „C-vēlēt", kur „C-" ir atslēgvārda pazīme. Šī metode tika izmantota arī interesējošo biežāk sastopamo terminu un nosaukumu dažādu rakstību apvienošanai (piemēram, partiju nosaukumiem un to saīsinājumiem).

Tvitera ziņu īpatnība ir liels daudzums atsauču uz citiem tvitera kontiem (piemēram, *@pedejapartija*) un uz tīmekļa adresēm (*http://…*). Šīs valodas vienības tika aizvietotas ar atslēgvārdiem „C-nick" un „C-http". Vēlēšanu tvitera korpusā parādījās arī dažādu procentuālo rezultātu minējumi (piemēram, 35 %), kuri tika aizvietoti ar atslēgvārdu „C-%".

Pēc priekšapstrādes beigām visas tvitera ziņas tika apkopotas un saglabātas tālākai analīzei piemērotos formātos. Satura analīzes uzdevumiem parasti tika izmantoti teksta faili ar visu ziņu tekstiem (ar un bez atslēgvārdu aizvietošanas). Ziņu teksti tika izmantoti arī konkordances rīka izveidē.

Satura analīze

Ziņu saturu var analizēt gan kvantitatīvi (statistiski), gan kvalitatīvi. Kvalitatīvā analīze pēta ziņas saturu kopumā un klasificē to atbilstoši noteiktiem kritērijiem.

Tvitera pētījumā tika veikta noskaņas (*sentiment*) analīze, t. i., tika noteikta tvītu attieksme pret ziņas subjektu. Ziņas klasificētas pēc attieksmes: pozitīvas,

[15] Atslēgvārdu aizvietošanas rezultātā dažādas vārda lietojuma formas tika aizvietotas ar īpašu atslēgvārdu.

neitrālas vai negatīvas. Tā kā latviešu valodai nav pieejami tādi rīki kā *LIWC*[16], noskaņas analīze tika veikta manuāli, katru ziņu individuāli apskatot un klasificējot. Šīs analīzes rezultāti ir atrodami pētījumā „The Pragmatics of Political Messages in Twitter Communication" (Šķilters et al. 2011).

Kvantitatīvās analīzes laikā pētīts tvitera ziņu korpuss kopumā, nosakot tā vārdu lietojuma biežumu un vārdu kolokācijas jeb ciešus vārdu savienojumus. Pēc datu priekšapstrādes veikšanas valodas vienību biežumu un kolokācijas varēja noteikt, izmantojot *NLTK* bibliotēku (Bird et al. 2009).

Vienas nedēļas tvitera ziņu korpusa vārdlietojumu biežuma analīzes rezultāti ir apskatāmi 1. tabulā. Tajā ir iekļautas valodas vienības, kuras tekstā ir sastopamas vairāk nekā 1000 reižu un kuru garums pārsniedz trīs simbolus. Daļa no šīm valodas vienībām sākas „C-" un ir atslēgvārdu aizvietošanas rezultāts (piemēram, „vēlēšanu" tika aizvietots ar „C-vēlēšanas"). Valodas vienību sarakstā bija sastopamas arī simbolu virknes, kas ir raksturīgas virtuālajai videi un pilda tajā īpašas funkcijas. Šādas 1. tabulā iekļautās simbolu virknes apzīmē atsauces uz citiem tvitera kontiem (*C-@nick*), esošo ziņu pārpublicēšanas jeb *retweet* pazīmes (*RT*), tīmekļa saišu pieminējumi (*C-http*) un saīsinājumi, kas pauž rakstītāja emocijas jeb emotikoni (piemēram, :) apzīmē smaidu vai pozitīvu attieksmi).

Kolokāciju analīzē apskatītas divu vārdu kolokācijas (*bigrams*), kurās vārdi atrodas blakus viens otram, un noteikts to biežums un kolokāciju reitings (nozīmīgums). Kolokāciju nozīmīguma noteikšanai tika izmantots *PMI (Pointwise Mutual Information)* kritērijs (Manning, Schutze 2003).[17]

1. tabula. **Valodas vienību biežums**

C-Balsot	2839	**C-@nick**	30424
C-Vēlēšanas	2696	**C-http**	13384
C-Latvija	2466	**RT**	9556
C-Labs	2150	**:)**	5152
C-Būs	2129	**:D**	3087
tikai	1626	**;)**	816
bija	1453	**:(**	430
C-Diena	1441	**:))**	383
C-PLL	1260	**:P**	196
C-%	1216	**;D**	172
šodien	1048		

[16] *LIWC* (*Linguistic Inquiry and Word Count*) rīks ir izmantojams detalizētai angļu valodas tekstu analīzei, tai skaitā teksta noskaņas (sentiment) analīzei.

[17] *NLTK* bibliotēkā realizētie kolokāciju reitinga kritēriji ir atrodami *http://nltk.googlecode.com/svn/trunk/doc/howto/collocations.html*

Tā kā korpusā daudziem vārdiem un to savienojumiem bija līdzīgs biežums, daudzām kolokācijām ir vienādas *PMI* vērtības. Tādēļ šis reitings izmantots, lai atrastu nozīmīgākās kolokācijas, un tālākajā pētījuma gaitā atlasītas tās kolokācijas, kuras ir interesantas konkrētā pētījuma kontekstā. (Šķilters et al. 2011)

Strukturālā un metadatu analīze

Strukturālajā un metadatu analīzē ir apskatīts birku (*hashtags*) un tvitera ziņu pārpublicējumu (*retweets*) lietojums.

Tvitera birkas ir simbolu virknes, kas sākas ar # un tiek izmantotas, lai iezīmētu tvitera ziņas un ļautu viegli atrast visas ziņas, kuras lieto noteiktu birku. Piemērs birku lietojumam ir dažādas tēmas, notikumi un konferences, kuras atspoguļojošo tvitera ziņu atzīmēšanai tiek izmantotas atbilstošas birkas (piemēram, birka *#icwsm* atbilst *ICWSM* konferencei[18]). Vienas nedēļas korpusā birkas tika lietotas ~4,5 % ziņu. Korpusā kopā tika atrasts 750 dažādu birku, tomēr lielākā daļa no tām tika lietota tikai vienu reizi.

Korpusā biežāk lietoto birku dinamiku (lietojuma izmaiņas vairāku dienu laikā) var apskatīt 2. tabulā. Uzskatāmības dēļ tabulā ir parādītas tās birkas, kuru tematika tieši vai netieši ir saistīta ar pētījuma tematiku – politiku un

2. tabula. **Biežāk lietotās birkas**

Birka	30-Sep	01-Oct	02-Oct	03-Oct	04-Oct
#ir	27	34	21	32	
#pietiek	7	10	16	10	5
#pll	5				
#politika	5				
#politsports	5				
#velesanas		12	346	94	5
#cieti		8	16		
#fail		9	5		9
#sleptareklama		5			
#nobalsoju			71		
#twibbon			60		
#vēlēšanas			35	11	
#velesanas2010			7		

[18] http://www.icwsm.org/2011/papers.php

10. Saeimas vēlēšanām. Deviņas no desmit populārākajām birkām vēlēšanu dienā (02.10.2010.) saistītas ar politiku un parādās tabulā. Visplašāk lietotā birka šajā laika periodā ir *#velesanas*, kura parādās dienu pirms vēlēšanām un sasniedz ļoti augstu popularitāti 2. un 3. oktobrī. Divas dienas pēc vēlēšanām šīs birkas popularitāte jau ir neliela, tā parādot tvitera virtuālās sabiedrības ātro reakciju uz notikumiem un īso uzmanības laiku. Stabila popularitāte bija birkām *#ir* un *#pietiek* (4 no 5 dienām), kuras ir saistītas ir žurnālistiku un aktuālo notikumu atspoguļošanu.

Pārpublicēto ziņu analīze balstās uz to, ka tviteris ir ieviesis vispārpieņemtu notāciju ziņu pārpublicēšanai[19] un nodrošina iespēju esošās ziņas pārpublicēt ar vienu pogas spiedienu. Tvitera ziņu pārpublicēšana ir noderīgs materiāls pētniecībai (Boyd et al. 2010). Aptuveni 16 % (8001 ziņa) no tvitera korpusa ziņām ir pārpublicētas.[20] Kopumā lietotāji bija pārpublicējuši 4726 oriģinālās ziņas, no kurām:

- 73,1 % ziņu ir pārpublicētas 1 reizi;
- 13,8 % ziņu ir pārpublicētas 2 reizes;
- tikai 1 % ziņu ir pārpublicētas vismaz 10 reižu.

No 20 visvairāk pārpublicētajām ziņām 70 % ir tieši saistītas ar vēlēšanām, 10 % ir netieši saistītas, bet atlikušajiem 20 % ziņu nav saistības ar vēlēšanām. Lielākā daļa ar vēlēšanām saistīto ziņu (8 no 14) bija satīriska rakstura ziņas par politiķiem un politiskajām partijām.

Birku un ziņu pārpublicēšanas analīzes tehniskā daļa tika realizēta, no tvitera saņemtās ziņas (pilnu ziņu struktūru *JSON* formātā) ievietojot *MongoDB* datubāzē[21]. Šī datubāze uzglabā datus *JSON* formātā, kas sakrīt ar izplatītāko tvitera ziņu datu formātu. Birku un pārpublicējumu izpētei tika veikti atbilstoši datubāzes vaicājumi.

Kopsavilkums

Sociālā tīmekļa servisi ir noderīgs un interesants izpētes datu avots. Šajā rakstā tika apskatīta mikroblogu servisa *Twitter* datu analīze, sākot no datu vākšanas un turpinot ar savākto ziņu satura un strukturālo analīzi.

[19] Pārpublicētās ziņas tiek sāktas ar tekstu „RT @nick" (bez pēdiņām, @nick aizvietojot ar oriģinālās ziņas autora tvitera konta vārdu), kuram seko sākotnējās ziņas teksts. Reizēm pārpublicēto ziņu autori tās maina, ziņas saīsinot vai tām pievienojot savus komentārus.

[20] Šeit un tālāk pētījumā par pārpublicētām ziņām tiek uzskatītas tādas, kurām tviteris ir piešķīris pārpublicētās ziņas pazīmi. Tādas ir ~90 % no ziņām, kurās tekstu „RT @".

[21] *http://www.mongodb.org/*

Latviešu valodas tekstu izpēti apgrūtina tas, ka trūkst morfoloģiskās analīzes rīku un tekstu padziļinātai analīzei (piemēram, noskaņojuma analīzei). Šādu rīku izveide un pieejamība padarītu vienkāršāku turpmāko sociālā tīmekļa satura izpēti.

Šis darbs ir viens no pirmajiem tvitera vides latviskās komunikācijas izpētē. Interesanti turpmākās pētniecības virzieni ir gan tālāka sociālā tīmekļa lietojumu analīze, gan arī latviešu valodas lietojuma virtuālajā vidē izpēte.

Pateicības

Šis darbs tika veikts ar valsts pētījumu programmas „Nacionālā identitāte" atbalstu.

Literatūra

1. Bird, S., Klein E., and Loper, E. *Natural Language Processing with Python*. O'Reilly, 2009. Pieejams: *http://www.nltk.org/book*
2. Boyd, D., Golder, S., & Lotan, G. *Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS-43*. IEEE: Kauai, HI, 2010.
3. Manning, C. D., Schutze, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 2003.
4. Tim O'Reilly. *What is Web 2.0: Design patterns and business models for the next generation of software.* Social Science Research Network Working Paper Series, 2005. Pieejams: *http://ssrn.com/abstract=1008839*
5. Šķilters, Jurģis, Kreile, Monika, Bojārs, Uldis, Brikše, Inta, Pencis, Jānis, Uzule, Laura. The Pragmatics of Political Messages in Twitter Communication. *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts (MSM 2011), Heraklion, Crete, 2011*. Pieejams: *http://ceur-ws.org/Vol-718/paper_18.pdf*

Jānis Pencis

Valodas atklātie konceptuālie tīklojumi tvitera komunikācijā: politisko partiju apvienību un to līderu identitāšu piemērs

Ievads

Pētījums analizē struktūras dinamiskos vienumos – identitātes kopienās (kopienās, ko vieno kopīgs piederības un atšķirības fenomens). (Brewer 1991, 475–482) Pētījumā ir atklātas indivīda un grupas kategorizācijas fenomena nesakritības politiskā diskursa un tā polarizācijas efektu gadījumā. Empīriskā materiāla analīzei ir izmantotas gan statistikas, gan satura analīzes metodes. Saskaņā ar Kurtu Levinu (Lewin 1936) un balstoties uz empīriskiem rezultātiem, ir iespējams identificēt vairākus sociālās struktūras raksturojumus: sociālās struktūras ir daļa–veselais struktūras ar iekšējām un ārējām savstarpējām atkarībām, sociālās struktūras ir atkarīgas no perspektīvas, un tās ir mērķorientētas. Taču tajā pašā laikā pastāv nozīmīgas atšķirības starp kopienām fiziskajā un virtuālajā vidē. Šajā pētījumā ir pieņemts, ka visām kopienām galvenais raksturojums ir aktivitāte sociālo saišu struktūrā. Sociālās struktūras ir vairāk vai mazāk aktīvas, un saitēm, kas uzrāda aktivitāti, nav jābūt fiziskām.

Pētījuma metodoloģija un dizains

Pētījumā izmantotajā datu kopā ir 31 612 tvitera ziņojumi četras dienas pirms 10. Saeimas vēlēšanām (no 28.09.2010. līdz 1.10.2010.). Šobrīd nav pieejami publiski dati par kopējo tvitera lietotāju skaitu Latvijā, taču pēc mediju ekspertu sniegtās informācijas kopējais skaits varētu būt ap 40 000 lietotāju (2010. gada novembrī). (Rutule 2010) Datu kopa ir izveidota, izmantojot tvitera *Streaming* programmatūras interfeisu. Lai apkopotu tvitera ziņojumus, atbildes ziņojumus (izmantojot *@atbilde* apzīmējumu), kā arī ziņojumu pārpublicējumus attiecīgajā laika periodā, sākotnēji tika atlasīti tvitera lietotāji no Latvijas, kā arī tie tvitera lietotāji, kas raksta latviešu valodā. Izvēloties tvitera lietotājus, kuru radītais saturs ir saistīts ar vēlēšanu tematiku, manuāli tika atlasīti 179 indivīdi, kas 1) ir politiskās partijas, to apvienības un 10. Saeimas deputātu kandidāti; 2) žurnālisti, politikas analītiķi un citi cilvēki, kas aktīvi diskutē par politiku un vēlēšanām; kā arī 3) cilvēki, kas ir aktīvi Latvijas tviterī. Tā kā pētījuma mērķis ir izraudzīties vismaz 1000 tvitera lietotāju, kuru ziņojumus

varētu apkopot izvēlētajā laika periodā, sākotnējā datu kopa tika paplašināta, 1) apkopojot ziņojumus no sākotnējās kopas; 2) identificējot jaunus tvitera lietotājus, kas šajos ziņojumos ir minēti; 3) atsijājot tvitera lietotājus, kas nav saistīti ar Latvijas tematiku; kā arī 4) atkārtojot šo procesu vēlreiz. Tādējādi tika atlasīti 1377 tvitera lietotāji, kuru ziņojumi ir apkopoti šī pētījuma vajadzībām. Atlasītos tvitera lietotājus savā ziņā var uzskatīt par viedokļa līderiem Latvijā. Pētījumā tvitera lietotāju un tviteru ziņu izlase veidota mērķtiecīgi, izvairoties no ikdienišķām sarunām, iekļaujot ziņas, kas saistītas ar šī pētījuma interesēm – politiku un identitātes veidošanu. Šāda apzināti izvēlēta datu kopa palielina pētījuma temata analīzes precizitāti.

Apkopotie tvitera ziņojumi ir apstrādāti ar *NLTK* (*Natural Language Toolkit*) instrumentiem. Tvitera ziņojumu apstrādi veido: 1) datu kopas attīrīšana no tvitera ziņojumu dublikātiem vai dzēstiem ziņojumiem; 2) pilna tvitera ziņojumu datu struktūras saglabāšana metadatu analīzei; 3) tvitera ziņojumu sadalīšana analizējamos elementos (vārdos u. c.); 4) dažādi rakstīta viena un tā paša vārda vai frāzes rakstības aizvietošana ar atslēgvārdiem.

Tā kā latviešu valodai nav izstrādātas programmatūras, kas ļautu automatizēti aizvietot dažādos vārdu locījumus un rakstības variantus (piemēram, interneta komunikācijā aizvietojot garumzīmes ar diviem patskaņiem bez garumzīmēm), šajā pētījumā ir izstrādāts aizvietošanas saraksts. Pilnas latviešu valodas morfoloģijas aizvietošanas kartes izstrāde ir pārlieku laikietilpīga un ārpus pētījuma jomas, turklāt šajā gadījumā ir jākonsolidē valodas vienības, kas ir svarīgas pētījuma kontekstā (Saeimas vēlēšanas). Tādēļ sarakstā ir ietvertas vissvarīgākās valodas vienības, piemēram, partiju un politiķu vārdi, ar vēlēšanām saistīti vārdi (piemēram, *balsot, vēlēšanas*). Tāpat tiek aizvietoti visi procentu skaitļi (piemēram, 35 %) un saites ar atslēgvārdiem *C-%* un *C-http*.

Tvitera ziņojumu apstrāde ir sagatavojusi empīrisko materiālu tā tālākai analīzei. Pētījumā tiek analizēts tvitera ziņojumu saturs. Analīzes kategorijas sākotnēji ir formulētas, izmantojot kvalitatīvo satura analīzi, taču vēlāk tiek izmantoti arī kolokāciju un konkordanču analīzes rezultāti. Satura analīzei tiek izmantoti teksti no visiem datu kopā ietvertajiem tvitera ziņojumiem. Atkarībā no analīzes uzdevuma ir izmantots gan teksts pirms atslēgvārdu aizvietošanas, gan pēc aizvietošanas. Būtiskākie teksta apstrādes veidi satura analīzes fāzē ir konkordances rindas (lai redzētu vārdu tā oriģinālajā kontekstā), vārdu biežuma analīze, kā arī kolokāciju analīze. Kolokāciju reitingu noteikšanai ir izmantota *PMI* formula (*Pointwise Mutual Information metric*), neņemot vērā kolokācijas, kuru biežums ir mazāks par trīs (Manning, Schutze 2003).

Satura analīzes rezultāti: kognitīvās distances starp grupām un to indivīdiem jeb politisko partiju apvienībām un to līderiem

Izmantojot Centrālās vēlēšanu komitejas datus, ir izveidots visu 1234 Saeimas deputātu kandidātu saraksts[1] un meklētas to reprezentācijas četras dienas pirms vēlēšanām atlasītajos tvitera ziņojumos. Sākumā ir atlasītas visas kolokācijas, kurās ir kāds no kandidātu uzvārdiem. Lai nošķirtu gadījumus, kad viens un tas pats uzvārds ir attiecināms uz vairākiem kandidātiem, ir izmantota konkordances rindu analīze. Kopumā atlasītajās kolokācijās ir identificēti 44 kandidātu uzvārdi. Tas nozīmē, ka tikai 3,56 % no visiem kandidātiem ir reprezentēti tvitera komunikācijā (atlasītajā datu kopā) četras dienas pirms vēlēšanām. Turpinājumā ir uzskaitīts, cik daudz kolokāciju parādās ar katru uzvārdu. Vidēji katrs kandidāta uzvārds četras dienas pirms vēlēšanām veido 4,68 kolokācijas; turpmākajā analīzē ir ietverti tie uzvārdi, kas ir ar statistiski nozīmīgu kolokāciju skaitu ($n \geq 4{,}68$). Tie ir apkopoti pirmajā tabulā (sk. 1. tabulu).

Līdzīgi kandidātu analīzei 2. tabula parāda, cik daudz kolokāciju analizētajā datu kopā ir atrastas katram partiju nosaukumam.

Kā liecina minētie pētījuma dati, gan PLL, gan „Vienotība" ir politisko partiju apvienības ar nozīmīgu, augstu pieminēšanas reitingu tvītos. Turklāt arī šo apvienību premjerministra kandidātiem – Aināram Šleseram (PLL) un Valdim Dombrovskim („Vienotība") – ir vieni no augstākajiem pieminēšanas reitingiem. Taču, neskatoties uz šīm līdzībām, katrai partijai un to kandidātiem ir atšķirīgi vēlēšanu rezultāti – „Vienotība" ir uzvarējusi vēlēšanās, iegūstot 33 parlamenta vietas, un V. Dombrovskis

1. tabula. **Kolokāciju skaits deputātu kandidātiem četras dienas pirms vēlēšanām**

Štokenbergs	6
Zemītis	6
Urbanovičs	7
Loskutovs	7
Līdaka	10
Āboltiņa	10
Dombrovskis	17
Šķēle	18
Šlesers	47

[1] 10. Saeimas vēlēšanas 2010. gada 2. oktobrī: kandidātu saraksts un statistika. Centrālās vēlēšanu komitejas mājaslapa. Sk. 4. jan., 2011: http://www.cvk.lv/cgi-bin/wdbcgiw/base/komisijas2010.cvkand10.sak

2. tabula. **Kolokāciju skaits partijām un to apvienībām četras dienas pirms vēlēšanām**

Par prezidentālu republiku (PPR)	2
Pēdējā partija (LP)	3
Par cilvēktiesībām vienotā Latvijā (PCTVL)	3
Zaļo un Zemnieku savienība (ZZS)	5
Saskaņas centrs (SC)	10
Visu Latvijai–TB/LNNK (VL/TB–LNNK)	12
Vienotība (Vienotība)	34
Par labu Latviju (PLL)	41

ir apstiprināts par premjerministru, savukārt PLL ir ieguvusi tikai 8 vietas parlamentā.[2] Tas liek pievērsties detalizētākai šo politisko indivīdu un grupu identitātes veidošanas analīzei pirmsvēlēšanu tvitera ziņojumu politiskajā kategorizācijā. Pirmkārt, ir aprēķinātas 10 kolokācijas ar visaugstāko reitingu katram no četriem atslēgvārdiem – attiecīgo kandidātu uzvārdiem (A. Šlesers un V. Dombrovskis), kā arī attiecīgo apvienību nosaukumiem (PLL un „Vienotība"). Otrkārt, ir izmantota konkordances rindu analīze, lai atklātu semantisko bagātību katrai kolokācijai. Saskaņā ar konkordances analīzi tematiski visbagātīgākā komunikācija tvitera vidē ir V. Dombrovskim – viņa TOP 10 kolokāciju kopa atklāj 29 politiskos tematus. Tematu konotācija ir noteikta manuāli katrā tvitera ziņojumā atsevišķi un vispārināta saskaņā ar trīs kategorijām (pozitīvi, neitrāli un negatīvi). Piemēram:

Draugi, rīt es došos vēlēt par Dombrovski, jo es uzticos viņa profesionālismam. (pozitīvi)

Šlesers apšauba sociālo mediju objektivitāti. (neitrāli)

Dombrovskis: korupcijas protežē vai reketieris? (negatīvi)

3. tabulā ir apkopots procentuālais sadalījums, cik daudz politisko tematu ziņas subjektam ir ar pozitīvu, neitrālu vai negatīvu konotāciju.

3. tabula. **Atslēgvārdu tematu konotācija četras dienas pirms vēlēšanām**

TOP 10 kolokāciju kopa Kandidāts (apvienība)	**Tematu skaits**	**Pozitīvi**	**Neitrāli**	**Negatīvi**
Dombrovskis	29	27,59 %	68,97 %	3,45 %
Šlesers	12	25,00 %	41,67 %	33,33 %
„Vienotība"	11	0,00 %	54,55 %	45,45 %
PLL	14	21,43 %	21,43 %	57,14 %

[2] Par 10. Saeimas deputātu kandidātu sarakstiem nodotais derīgo vēlēšanu zīmju skaits un kandidātu sarakstu iegūtais vietu skaits 10. Saeimā. Centrālās vēlēšanu komisijas mājaslapa. Sk. 13. jūn., 2011: http://web.cvk.lv/pub/public/29763.html

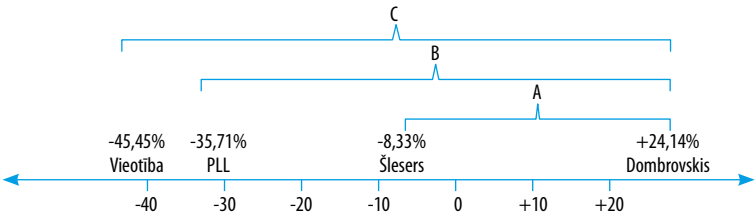
Kā redzams šajā tabulā, indivīds un grupa ir kategorizēti līdzīgi Šlesera un PLL gadījumā – abi vairāk ir saistīti ar negatīviem tematiem nekā pozitīviem. V. Dombrovska un „Vienotības" gadījums ir citāds: indivīds ir pārsvarā kategorizēts ar pozitīviem vai neitrāliem tematiem, savukārt grupa – ar negatīviem vai neitrāliem.

Atņemot negatīvo tematu procentuālo daļu no pozitīvo tematu procentuālās daļas, katram kandidātam un apvienībām ir aprēķināts konotācijas reitings. Rezultātā „Vienotības" reitings ir -45,45 %, PLL -35,71 %, A. Šlesera (PLL) -8,33 %, bet V. Dombrovska („Vienotība") +24,14 %. Saskaņā ar šo statistiku var pieņemt, ka grupas un tās indivīdu identitātes veidošana var ietvert nozīmīgas nesakritības starp kognitīvajām un fiziskajām distancēm. Konkrētajā gadījumā kognitīvā distance starp V. Dombrovski un „Vienotību" ir lielāka nekā fiziskā. Respektīvi, starp V. Dombrovski un viņa pārstāvēto apvienību „Vienotība" ir tuva fiziskā distance, jo V. Dombrovskis ir „Vienotības" biedrs – daļa no vienuma, un pretēji – „Vienotība" ir daļa no V. Dombrovska politiskās karjeras. Taču neskatoties uz to, šis pētījums uzrāda, ka starp šiem diviem politiskajiem ziņu subjektiem eksistē kognitīvi liela distance analizēto tvitera lietotāju vidū. Šī distance ir attēlota ar līniju C 1. attēlā.

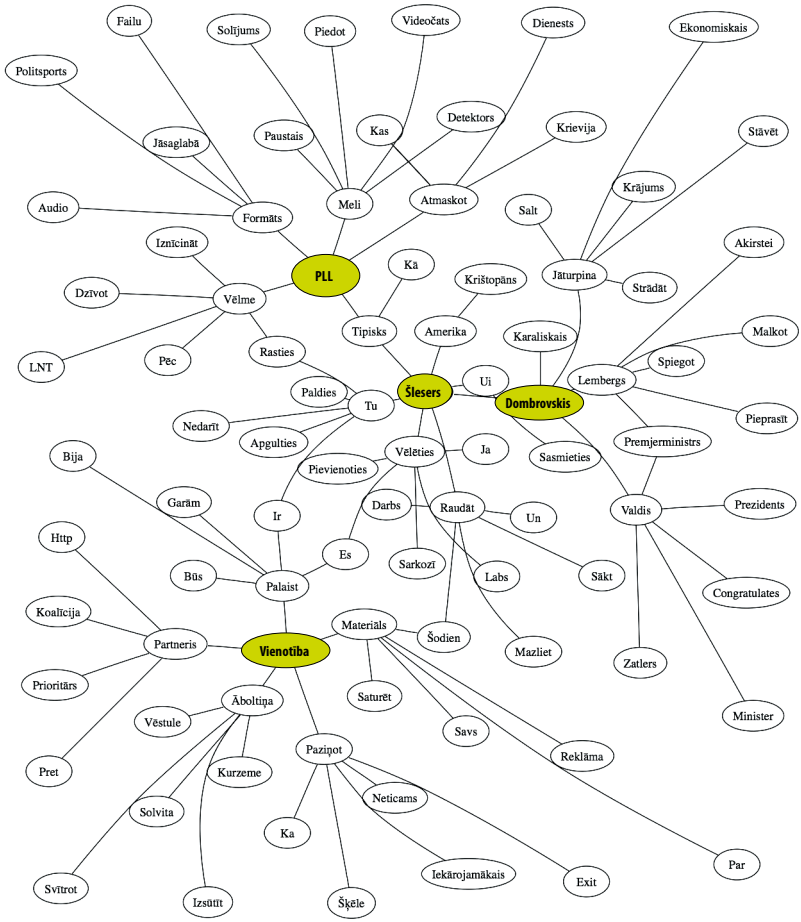
1. attēlā uz vienas nepārtrauktas līnijas ir novietoti četri politiskie ziņu subjekti atbilstoši iepriekš aprēķinātajiem konotācijas reitingiem – politisko partiju apvienība „Vienotība", politisko partiju apvienība PLL, kā arī šo apvienību premjerministra kandidāti A. Šlesers (PLL) un V. Dombrovskis („Vienotība"). Kā redzams attēlā, kognitīvā distance analizētās mērķpublikas mentālajos procesos starp V. Dombrovski un viņa apvienību „Vienotība" (C) ir lielāka nekā kognitīvā distance starp V. Dombrovski un viņa konkurentiem A. Šleseru (A) un PLL (B).

Šī pētījuma dati liecina, ka analizētā mērķpublika vērtē V. Dombrovski kā daudz pozitīvāku un svarīgāku politisko darbinieku nekā viņa pārstāvēto politisko partiju apvienību „Vienotība". Tas šķiet loģiski, jo V. Dombrovska karjera ir daudz ilglaicīgāka nekā „Vienotības" pastāvēšana. Turklāt „Vienotības" vēlēšanu kampaņa fokusējās uz V. Dombrovski kā svarīgāko vēlētāju ieguvumu. Tiek pieņemts, ka

1. attēls. **Kognitīvās distances starp politiskajiem ziņu subjektiem**



2. attēls. **Valodas atklātais konceptuālais tīklojums tvitera komunikācijā atlasītajiem politisko ziņu subjektu atslēgvārdiem**



tas ir viens no svarīgākajiem faktoriem šo nesakritību veidošanā. Neskatoties uz negatīvajiem tvitera ziņojumiem, politisko partiju apvienība „Vienotība" uzvarēja vēlēšanās un ieguva 33 parlamenta vietas. Rezultātā daļa (indivīds) kognitīvi ir kļuvusi daudz svarīgāka nekā veselums (grupa).

2. attēlā apkārt katram no četriem atlasītajiem atslēgvārdiem – „Vienotība", V. Dombrovskis, A. Šlesers un PLL – ir dotas piecas kolokācijas ar visaugstāko

reitingu. Valodas atklātais konceptuālais tīklojums tvitera komunikācijā ir paplašināts, pievienojot katram nākamajam atslēgvārdam piecas kolokācijas ar visaugstāko reitingu.

Dažiem no atslēgvārdiem datu kopā ir mazāk par piecām kolokācijām (piemēram, *atmaskot*), savukārt citiem atslēgvārdiem ir kopīgas kolokācijas (piemēram, *Šlesers* un *Dombrovskis*). Rezultātā var secināt, ka valodas atklātajā konceptuālajā tīklojumā tvitera komunikācijā eksistē gan tiešas, gan pastarpinātas saites starp politiskajiem ziņu subjektiem. Iepriekš veiktie reitingu un kognitīvo distanču aprēķini korelē ar šajā attēlā novērojamiem saišu attālumiem – jo pastarpinātākas ir saites valodas atklātajā konceptuālajā tīklojumā tvitera komunikācijā starp politiskajiem ziņu subjektiem, jo lielāka ir kognitīvā distance starp tiem analizētās mērķpublikas mentālajos procesos. Tādēļ 2. attēlā ir novērojamas līdzīgas distances kā 1. attēlā: A – vismazākā starp V. Dombrovski un A. Šleseru, B – lielāka par A – starp V. Dombrovski un PLL, kā arī C – vislielākā starp V. Dombrovski un „Vienotību".

Tas ļauj formulēt šādu hipotēzi: (1) jo tematiski dažādāka un (2) biežāka ir komunikācija, kā arī (3) jo vairāk komunikācijas kanālu ir izmantots, lai minētu grupas indivīdu pozitīvi, jo lielāka iespēja, ka indivīds kļūst kognitīvi nozīmīgāks par grupu un rada pārmaiņas grupas svarīguma uztverē.

Šī hipotēze atbilst sākotnējam pieņēmumam, ka jebkura veida kopienas galvenais raksturojums ir aktivitāte (šajā gadījumā – komunikatīva aktivitāte) sociāli saistītā struktūrā. Sociālās struktūras ir vairāk vai mazāk aktīvas, un saites, kas uzrāda šo aktivitāti, ne vienmēr ir fiziskas.

Secinājumi

Šī pētījuma rezultātu vispārinājumi ļauj izvirzīt hipotēzi, ka sociālo grupu un indivīdu uzvedībai ir atšķirīgas funkcijas, kas ir balstītas personu un to apkārtējās vides mijiedarbībā. Ja aplūko vispārēju sociālās vides situāciju, ir jānošķir šīs divas atšķirīgās funkcijas.

Šie secinājumi tādējādi saskan ar Levina teoriju divos aspektos: (a) identitātes grupu struktūru veidošanā nošķirot atšķirīgās funkcijas grupām kā veselumiem un to veidojošajiem indivīdiem un (b) paplašinot šo teoriju ar empīriska virtuālo kopienu un sociālo tīklu komunikācijas (tvitera) pētījuma rezultātiem.

Identitātes kopienu analīze parāda dažādas daļa–veselais attiecības un atkarības starp apakšgrupām un to veidojošajiem indivīdiem, veidojot perspektīvus kategorizācijas efektus. Turklāt ātrās un dinamiskās funkcionālās atkarības, saspīlējumi

un spēki rada fleksiblas struktūras sociālās identitātes grupām. Tādējādi Levina dzīves telpa ir iekšēji un ārēji savstarpēji atkarīga struktūra. Iekšējā struktūrā eksistē saites, kas vieno indivīdus to identitātes grupas veidošanā. Iekšējā struktūra ietver dažādus pārākuma efektus – grupas līderi ir daudz prominentāki nekā grupas perifērijas locekļi. Šajā valodas atklātajā konceptuālajā tīklojumā tvitera komunikācijā eksistē arī ārējās saites, kas vieno savā starpā dažādas grupas, tādējādi arī ārējām grupām nosakot konkrētās identitātes grupas struktūru. Svarīgi ir tas, ka saitēm, kas vieno grupas, nav jābūt fiziskām. Jebkuru sociālo vienību (arī virtuālo identitāšu kopienu) svarīgākā iezīme ir to aktivitāte. Protams, grupas var atšķirties pēc to aktivitātes, taču šāda aktivitāte nav fiziska iezīme.

Pateicības

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu projektā „Atbalsts doktora studijām Latvijas Universitātē".

Literatūra

1. Brewer, M. B. The social self: On being the same and different at the same time. In: *Personality and Social Psychology Bulletin*, 1991, 17, pp. 475–482.
2. Brewer, M. B., Gardner, W. Who is this "we"? Levels of collective identity and self representations. In: *Journal of Personality and Social Psychology*, 71, 1996, pp. 83–93.
3. Fillmore, C. J. Frame semantics and the nature of language. In: *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Vol. 280, 1976, pp. 20–32.
4. Haythornthwaite, C. Social networks and online community. In: Joinson, A., McKenna, K., Postmes, T., Reips, U.-D. (Eds.). *The Oxford Handbook of Internet Psychology*. Oxford : Oxford University Press, 2007, pp. 121–137.
5. Langacker, R. W. *Grammar and Conceptualization.* Berlin : Mouton de Gruyter, 2000.
6. Lewin, K. *Principles of Topological Psychology.* New York : McGraw-Hill, 1936.
7. Lewin, K. *Resolving Social Conflicts & Field Theory in Social Science.* Washington, DC : American Psychological Association, 1997.
8. Rutule, E. *Pētījums: arī Latvijā interneta lietotāji biežāk komunicē sociālajos tīklos, nevis klātienē. Sociālo mediju monitorings un vortāls WebRadar* [tiešsaiste]. Publicēts 2010. gada 1. novembrī. [Skatīts 2011. gada 13. jūnijā] Pieejams: *http://www.webradar.lv/2010/11/petijums-ari-latvija-interneta-lietotaji-biezak-komunice-socialajos-tiklos-nevis-klatiene/*
9. Smith, E. R., Henry, S. An in-group becomes part of the self: Response time evidence. In : *Personality and Social Psychology Bulletin,* 22, 1996, pp. 635–642.

Latviešu valodas resursi

Eduards Cauna

Dzejnieka valodas vārdnīcas datorversijas izveide

Ventspils Augstskolas Lietišķās valodniecības centrā 2010.–2011. gadā tika realizēts dzejnieka valodas vārdnīcas pilotprojekts. Tā laikā tika apkopota rakstnieka valodas vārdnīcu veidošanas pieredze Latvijā un pasaulē un aprobēta nepieciešamā datorlingvistikas programmatūra. Tika iegūts dzejas korpuss, ko veido 1283 dzejoļi. Studenti tika iepazīstināti ar mūsdienīgām teksta apstrādes un analīzes iespējām. Projekta laikā tika radīts īpaši dzejai piemērots korpusa rīks *rkorpuss*.

Rakstnieku valodas vārdnīcas Latvijā un pasaulē

Rakstnieka valodas vārdnīca tiek uzskatīta par vērtīgu materiālu rakstnieka valodas pētīšanai, jo ļauj analizēt konkrētā rakstnieka valodas raksturīgākās īpatnības, vārdu un to kombināciju (kolokāciju, t. sk. frazeoloģismu) izvēli, kā arī rakstnieka ieguldījumu konkrētās valodas leksikas (vārdu krājuma) papildināšanā. Cita būtiska un praktiska šādas valodas vārdnīcas loma ir neoloģismu, kā arī reto vai īpatnēji lietoto vārdu skaidrojums. Tādējādi ikvienam lasītājam šādā vārdnīcā iespējams atrast sev nepazīstama vārda nozīmi.

Pasaulē rakstnieka valodas vārdnīcas ir veidotas daudziem pazīstamiem rakstniekiem. Svarīgākie piemēri ir Homērs, Viljams Šekspīrs, Gēte, Dante,

Moljērs, Migels de Servantess, Henriks Ibsens, Marks Tvens, Aleksandrs Puškins, Fjodors Dostojevskis un Karels Čapeks.

Centieni apkopot kāda viena konkrēta rakstnieka leksiku pārskatāmā formā ir zināmi jau vairākus gadsimtus. Tomēr vārdnīcu skaits ir salīdzinoši neliels, jo to izveidi kavē teksta apstrādes grūtības. Piemēram, lai izveidotu korektu rakstnieka valodas vārdnīcu, vajag tā vai citādi fiksēt *visus* vārdus *visos* šā autora tekstos, ņemot vērā vārdformas un rakstības variantus. Šādu vārdnīcu sauc par pilno rakstnieka valodas vārdnīcu.

1956. gadā iznāca pirmais Aleksandra Puškina valodas vārdnīcas sējums (Словарь языка Пушкина 2000). Tas atstāja lielu iespaidu uz latviešu valodnieku priekšstatiem par to, kā šāda tipa vārdnīcai ir jāizskatās. Vārdnīcas veidotāji centušies godprātīgi atklāt svarīgākās nozīmju nianses A. Puškina valodā, palīdzot lasītājiem izprast senus, retus vai īpatnēji lietotus vārdus, tomēr tās būtisks trūkums ir morfoloģiskā raksturojuma neesamība.

Arī Latvijā ir bijuši vairāki mēģinājumi analizēt rakstnieku valodu un veidot atsevišķos darbos atrodamās leksikas vārdnīcas:

- Rainis (poēma „Ave Sol" (Kļumele 1955); „Ave Sol", dzejoļu krājumi „Tie, kas neaizmirst" un „Gals un sākums" (Эглите 1988));
- Aspazija (krājums „Sarkanās puķes" (Augstkalne, 1957));
- J. Alunāns (krājums „Dziesmiņas" (Cēbere 1958));
- A. Alunāns (luga „Paša audzināts" (Mežaraups 1959, Mežaraups 1963));
- R. Blaumanis (stāsts „Nezāle" (Juknevičs 1959); stāsts „Bailes no laimes" (Gasiņa 1959));
- V. Plūdonis (poēma „Atraitnes dēls" (Zeltiņa, 1959));
- A. Upīts (romāns „Zaļā zeme" (Petre 1987, Petre 1990));
- 19. gs. 90. gadu latviešu dzeja (10 autoru darbi (Štekele 1994)).

Tāpat ir veikta E. Veidenbauma, E. Treimaņa-Zvārguļa un Zvanpūša dzejas valodas statistisko parametru izpēte. Pētīta arī citu rakstnieku valoda. Tomēr tie visi ir galvenokārt studentu diplomdarbi, kuros pētīts kāda rakstnieka viens darbs, un tāpēc nesniedz pilnīgu priekšstatu par rakstnieka valodu.

Būtisks pagrieziens rakstnieka valodas vārdnīcu veidošanā notika līdz ar Karela Čapeka vārdnīcas izveidi Čehu nacionālā korpusa institūtā (Čermák 2008, 323–332). Tās radīšanai pirmoreiz tika izmantoti korpuslingvistikas paņēmieni, veicot rūpīgu morfoloģisko marķēšanu un kolokāciju analīzi. Tas pirmoreiz pasaulē ļāva iegūt matemātiski pamatotus rezultātus, salīdzinot konkrēta rakstnieka valodu ar vispārīgo valodu. Tādējādi bija iespējams atklāt K. Čapeka leksikas īpatnības un vārdus, kuriem viņš devis priekšroku (to biežums ir lielāks nekā vispārīgajā

valodā). Kolokāciju analīze ļāva iegūt raksturīgākos vārdu savienojumus, kam ir īpaša nozīme piemēru atlasē.

Tekstu sagatavošana

Tekstu skenēšana. Projektā piedalījās 56 studenti, kas tika sadalīti 17 grupās, katrai izvēloties savu dzejoļu krājumu. Projekta dalībniekiem tika dots uzdevums veikt atlasīto dzejas krājumu skenēšanu, optisko atpazīšanu un tekstu pārlasīšanu (korektūru), lai iegūtu tekstu datorversiju, kas saturiski pilnīgi atbilst oriģinālam. Skenēšana bija jāveic 600 dpi izšķirtspējā pustoņu (*grayscale*) režīmā, lai nodrošinātu teksta atpazīšanu arī grāmatās ar maziem burtiem.

Tekstu atpazīšana. Pēc skenēto failu sakārtošanas un pārbaudes tika veikta tekstu atpazīšana, izmantojot *ABBYY FineReader*. Šī programma ir pazīstama ar savu augsto atpazīšanas precizitāti, tomēr tās veiksmīga darbība ir atkarīga no skenējuma kvalitātes, kas dažos gadījumos bija nepietiekama (instrukcijai neatbilstoša izšķirtspēja, teksta izzušana grāmatas locījumu vietās u. tml.). Tāpat atpazīšanas precizitāti ietekmēja programmā iekļautais leksikons, kas dažkārt neļāva korekti atpazīt retāk lietojamos vārdus (piemēram, „bakurētaino seju" bija pārtapis par „baku rētai no seju").

Tekstu pārlasīšana. Atpazītie teksti tika saglabāti *MS Word* formātā (*doc*) un pārlasīti *Word* vidē. Tas ļāva lietot pazīstamos teksta apstrādes un pareizrakstības pārbaudes rīkus, atjaunot rindu formatējumu (ja kāda rinda atpazīšanas laikā bija saplūdusi ar iepriekšējo), kā arī sadalīt dzejoļus pa lappusēm.

Tekstu sagatavošana marķēšanai. Marķētājs spēj analizēt tikai neformatētu (*text only*) tekstu, kas saglabāts *utf-8* kodējumā. Šim nolūkam tika izmantots *Windows* programma *Notepad*, katru dzejoli kopējot no *Word* uz *Notepad* un saglabājot kā atsevišķu teksta failu. Saskaņā ar izveidoto darbplūsmas modeli, katram teksta failam tika piešķirts savs numurs (piemēram, *003.txt*), bet dzejoļa autors un nosaukums glabāti atsevišķā *Excel* failā. Tādējādi iegūtos 1283 dzejoļus vēlāk bija iespējams korpusā importēt automātiski.

Morfoloģiskā marķēšana

Tekstu morfoloģiskai marķēšanai tika izmantots LU Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijā izstrādātais tekstu korpusu marķēšanas rīks „Marķētājs 1.1 (alfa versija)" (*annotator-r742* ar leksikona versiju 739) (*TKMR*). Tā ir *Java* vidē lokāli lietojama programma, kas salīdzina analizējamā tekstā atrodamos vārdus ar iebūvētajā leksikonā norādītajiem visās formās un

locījumos. Analīze notiek vārdu pa vārdam, un katram vārdam tiek piedāvāti visi iespējamie varianti. Rīka grafiskā saskarne ļauj izvēlēties piemērotāko atbilsmi vai arī ierakstīt savu variantu, izmantojot *MULTEXT-East* morfoloģisko pazīmju kopu (Instrukcija). Tāpat ir iespējama jauna vārda ievade, norādot paradigmu.

Darbs auditorijā parādīja, ka rīks ir ātri apgūstams un lietotājiem grūtības nesagādā. Tomēr relatīvi nelielais leksikons bieži radīja situācijas, kad vajadzīgā vārda leksikonā nebija. Tā kā jaunu vārdu pievienošana ir neērta un vairākumam lietotāju arī nesaprotama, leksikonā neiekļautie vārdi visbiežāk tika marķēti kā nezināmi elementi (reziduāļi). Leksikonā nebija atrodami tādi vārdi un vārdformas kā *vajaga, raugās, dziest, sāpe, visviens, nava, skatiens, trakais, skrej, šad, jāsaka, dziļumā, burenieks, tiecas, blāv, vantis, dzejnieki, lillā, gulstas, audziet, kuro, šorīt, dvēsle, zobens, arīdzan, tveras, slēpties, dubļi, pretmīla, liktens, hobijs, ziedlapas, radītdziņa, pārteicies, attiecas, nodziest, štrunts, mīlē, sajaucas* u. c.

Jaunievadītie vārdi tiek saglabāti tikai uz lokālā datora un nav pieejami citiem lietotājiem. Ja viņiem iznāk sastapties ar to pašu vārdu uz cita datora, tas jāievada no jauna. Kā jau tika atzīmēts, vairākumā gadījumu šādi vārdi tika atzīmēti kā reziduāļi. Kopējais reziduāļu skaits korpusā ir 5520, kas pie kopējā elementu skaita 122798 veido 4,5 %. Tātad darba procesā marķētājam nebija pazīstams aptuveni katrs divdesmit otrais vārds. Tomēr salīdzinot programmai pazīstamās vārdformas ar nepazīstamajām, korpusā kopumā ir 14,5 % vārdformu, kas marķētājam nav zināmas (aptuveni katrs septītais vārds).

Minētie piemēri arī ilustrē dažas problēmas, ar kurām jāsaskaras, marķējot reālās valodas tekstus ar marķētāju, kas būvēts, balstoties uz literārās (normētās) valodas paradigmām. Tādi vārdi kā *vajaga, nava, burenieks, dvēsle, liktens* u. tml. neatbilst normai, un marķētājs uzskata *liktens* un *liktenis* par diviem dažādiem vārdiem, kaut arī faktiski tie ir tikai rakstības varianti. Līdzīga situācija veidojas, kad marķētājam nākas mācīt divas paradigmas vārdam *akmens* (ģen. *akmens*, bet ļoti bieži *akmeņa*). Ir skaidrs, ka reālās valodas marķēšanas rīkos būtu jābūt iespējai norādīt rakstības variantus. Tas ļautu pēc tam pētīt, kā, piemēram, *liktens/liktenis* lietots dažādu autoru darbos, izvairoties no situācijas, kad tautasdziesmās un citos darbos tiek lietotas literārajā valodā „neeksistējošas" formas (piemēram, „liku bēdu zem *akmeņa*").

Korpusa programmatūra *rkorpuss*

Marķētais teksts tika apkopots korpusā, kuru uztur īpaši šim nolūkam autora izveidotā *rkorpuss* programmatūra. Tā veidota uz *LAMP* platformas (*Linux, Apache, MySQL, php*), un to raksturo ērta lietotāja saskarne latviešu valodā un pieejamība ar pārlūkprogrammu (*Firefox, Chrome* u. tml.) no jebkura datora, kas pievienots internetam. Lai nodrošinātu autortiesību ievērošanu, materiāli pieejami tikai mācību un pētnieciskiem mērķiem un tikai ierobežotam autentificētu lietotāju skaitam.

Korpusa lietotājiem ir pieejami tradicionālie līdzekļi – meklēšana pēc vārda vai tā daļas (izmantojot aizstājzīmes) vai pēc lemmas, kā arī lemmu biežuma saraksti (grupēti pēc vārdšķiras) un *KWIC* (*keyword in context*) jeb konkordances skatījums. Atlasi var veikt pēc autora un/vai krājuma. Īpaši noderīga ir iespēja no *KWIC* pāriet uz oriģinālteksta plašākai konteksta izpētei, ko citas korpusa programmas parasti nepiedāvā. (Sk. 1. attēlu.)

Studentos lielu interesi izraisīja iespēja salīdzināt dažādu dzejnieku darbos biežāk lietotos vārdus. Īpaši interesants izrādījās lietvārdu reģistrs, kas labi atklāja katra dzejnieka iemīļotās tēmas. (Sk. 2. attēlu.)

1. attēls. **Konkordances vaicājumam *rude.**

Nr.	Konteksts pa kreisi	Vārds	Konteksts pa labi
1	ilgi stāv zem sniega . Vien	rudenos	, kad gleznāk atviz dienas un
2	iet ciema dievanamā . Kā vēla	rudens	bite medus kamā es reibstu ,
3	pēc tikliem brauc un mudas un	rudenos	kā sīki lāsumi plaukst māras sila
4	stikla . Tālu aiz meža vītīs	rudena	siens , un smarža vijīgi mikla
5	kaisma zīmēs uz zaļganā ūdens pirmās	rudena	vētras stāstu - un pusnaktī brāsma
6	cita ... un tagad jau gaisā	rudens	jausma , plaukst piekāpu mežos lāsmaini
7	un mūsu vēlā , mūsu skumjā	rudens	mīla negaidot iedegas kaismā un maiguma
8	ko vētra nolauzusi . No palsas	rudens	saules krauja silst un smilšu strāvas
9	asums kā ciesā , dzīve mūžīgam	rudenim	rada . Vien sapnim ziedonis jāievada
10	pat nava tādas slavas kā taviem	rudeniem	un taviem pavasariem . Es tevi
11	draugs - dziesmu par aizmirstu jūrmalu	rudens	vējā , gaisi kad sagriežas smilšu
12	klaidu kāpu , kas aizbērs vēlā	rudenī	sētu un māju . Vētras ,
13	, drīz buras zaudē plīvi un	rudens	negaisi tām irdās šķiedras plūc -
14	kroplus zarus kliež . Vien retu	rudeni	tam sīku ogu nasta , ko
15	atdeva slieksni par pagalvi zemu pelēkām	rudens	zālēm , jumti noslēdza draudzību klusu
16	. Un tikai reti naktis ,	rudens	naktīs , kad bākas sargs nedus

2. attēls. **Lietvārdu biežuma saraksts**

Līvzeme (Memento * Treji deviņi, 2001) | Fricis Dziesma

Izvēlieties vārdšķiru: lietvārds

Ievadiet vārdu: rude* vai pamatformu:

Vārds kontekstā | Vārdu biežums | Iziet

Lietojumu skaits	Vārds pamatformā	Lietotās formas
45	vilnis	viļņi, viļņu, vilni, viļņus, viļņiem, vilnis, viļņos
45	jūra	jūras, jūrai, jūru, jūra
45	krasts	krastiem, krastu, krastā, krasts, krasti, krastus, krastam
42	smilts	smiltīs, smilšu, smiltīm
41	ūdens	ūdeņu, ūdeni, ūdens, ūdeņi, ūdeņus, ūdeņiem
40	laiva	laivas, laivā, laivu, laivām, laivai
34	kāpa	kāpu, kāpas, kāpām, kāpai
27	zvejnieks	zvejnieki, zvejnieku, zvejnieks, zvejnieka, zvejniek, zvejniekiem
26	bura	buras, buru, burām, bura
26	dzīve	dzīvē, dzīvei, dzīves, dzīvi, dzīvēm
24	priede	priedes, priedēm, priežu, priede, priedei
23	gads	gadi, gadu, gad, gada, gadiem, gadus
23	līvzeme	līvzeme, līvzemes
22	roka	rokās, rokām, roka, roku
21	diena	dienām, dienas, diena, dienu
18	vasara	vasara, vasaras, vasaru
18	tīkls	tīklu, tīkli, tīkliem, tīklā, tīklus, tīklos
17	rudens	rudeņos, rudens, rudenim, rudeņiem, rudenī
17	nakts	nakts, naktii, naktīs, naktī
16	celš	celos, cela, celam, celus, celu, celš
16	acs	acis, acīm, acu
16	drauga	drauga
15	kuģis	kuģi, kuģu, kuģus, kuģis
15	selga	selgu, selgā, selgas, selgai
14	mākonis	mākoņu, mākoni, mākoņa, mākoņus, mākonis
14	joma	jomas, jomai
13	joms	jomā, jomu
13	vakars	vakara, vakaru, vakars, vakaros
13	saule	saules, saule, sauli
13	mūžs	mūžus, mūža, mūžs, mūžu, mūžam
12	putas	putu, putas

Šķirkļu izveide

Viens no rakstnieka valodas vārdnīcas uzdevumiem ir paskaidrot reti lietotus vai pat rakstnieka izdomātus vārdus. Vēsturiski ir bijuši dažādi uzskati, kā šādam skaidrojošam šķirklim būtu jāizskatās. Pilotprojekta laikā studenti veidoja šķirkļus, kurus veidoja šādas daļas:

- šķirkļavārds,
- vārdšķira,
- skaidrojums (neobligāts),
- piemēri (1–3).

Katrai grupai bija jāizveido šķirkļi trīs lietvārdiem, trīs darbības vārdiem un trīs īpašības vārdiem. Studenti tika aicināti pievērst uzmanību vārdu īpatnējam, no ikdienas valodas atšķirīgam lietojumam, ja tāds bija sastopams. Skaidrojumu izveidei bija atļauts izmantot „Latviešu valodas vārdnīcas" un „Latviešu literārās valodas vārdnīcas" elektronisko versiju (*www.tezaurs.lv*). Tomēr nereti skaidrojums nav vajadzīgs, ja ir pietiekami daudz piemēru.

- **bogs** *n.* – *sanāk krastmalā pircēju bogs*
  – *sievas stāv bogā gar krastu*
- **kaisma** *n.* – *deg sārta zelta kaismā ģeorģīnes*
  – *kaisma zīmēs uz zaļganā ūdens*
- **rudens** *n.* (*ģen.* rudeņa) – *Kā vēla rudens bite medus kamā es reibstu*
  – *Tālu aiz meža vītīs rudeņa siens*
- **suinīties** *v.* – *krevēs un pušumos nosuinītām palēverainām ausīm*

Rezultāti

- Pilotprojekta laikā tika digitalizēti, morfoloģiski izanalizēti un ievietoti korpusā 46 autoru 1283 dzejoļi, iegūstot 122 798 teksta elementus (*tokens*), 105 375 vārdus, 26 999 leksēmas, 11 942 lemmas.
- 56 studenti ieguva priekšstatu par tekstu korpusa izveides posmiem, praktiski piedalījās tā izveidē, apguva morfoloģiskās marķēšanas rīku un iemācījās izmantot dzejas analīzei korpusa rīku *rkorpuss*. Tāpat viņi ieguva jaunas zināšanas par valodas datorizētas analīzes iespējām, par attiecībām starp literatūras un literāro (normēto) valodu.
- Tika izveidots korpusa rīks *rkorpuss*, kas īpaši piemērots dzejas materiāliem.

Secinājumi un darāmie darbi

- Lai iegūtu kvalitatīvu korpusa materiālu, visi teksta sagatavošanas posmi prasa rūpību un precīzu instrukciju ievērošanu. Nākotnē to būtu vēlams uzticēt nelielai, labi apmācītai grupai (2–3 cilvēki). Atpazīto tekstu salīdzināšana ar oriģinālu prasa īpašas iemaņas, tāpēc šim nolūkam vēlams piesaistīt profesionālus korektorus.
- Studenti ļoti labprāt piedalās ar valodu saistītos izpētes projektos, ja tie notiek datorvidē.
- Jāizmanto cits rīks morfoloģiskai marķēšanai, kurā leksikona papildinājumi nokļūtu uzreiz centrālā datubāzē un kas būtu elastīgāks reālās valodas aprakstīšanai. Jāveic teorētiska analīze, kā rīkoties ar rakstības variantiem (*jūra/jūŗa*, *dvēsele/dvēsle*, *liktens/liktenis*) un starpparadigmu formām (*dizaineris/dizainers*, *akmens/akmeņa*).

- Jāpilnveido korpusa rīks *rkorpuss*, pievienojot algoritmus kolokāciju analīzei un moduli pusautomātiskai šķirkļu ģenerēšanai ar šķirkļavārdu, vārdšķiru un raksturīgākajiem piemēriem.

Pateicības

Autors izsaka pateicību Maijai Baltiņai un Dainai Vucānei par organizatorisko un pedagoģisko ieguldījumu projekta izpildes laikā.

Literatūra

1. Augstkalne, R. *Aspazijas dzejoļu krājuma "Sarkanās puķes" valodas vārdnīca.* Diplomdarbs. Rīga : Latvijas Valsts universitāte,1957.
2. Mežaraups, A. Par A. Alunāna lugas "Paša audzināts" valodas vārdnīcu. No: *P. Stučkas Latvijas Valsts universitātes zinātniskie raksti*, 35. sējums. Rīga, 1963, 233.–254. lpp.
3. Cēbere, A. *Jura Alunāna "Dziesmiņu" valodas vārdnīca*. Diplomdarbs. Rīga : Latvijas Valsts universitāte, 1958.
4. Čermák, František. An Author's Dictionary: The Case of Karel Čapek. *Proceedings of the XIII Euralex International Congress*, Barcelona 2008, 323–332.
5. Gasiņa, A. *R. Blaumaņa stāsta "Bailes no laimes" valodas vārdnīca.* Diplomdarbs. Rīgā : P. Stučkas Latvijas Valsts universitāte, 1959.
6. Juhnēvičs, M. *Rūdolfa Blaumaņa stāsta "Nezāle" valodas vārdnīca*. Diplomdarbs Rīga : P. Stučkas Latvijas Valsts universitāte, 1959.
7. Kļumele, V. *Raiņa poēmas "Ave Sol!" valodas vārdnīca*. Diplomdarbs. Rīga : Latvijas Valsts universitāte, 1955.
8. *Latviešu valodas tekstu korpusu morfoloģiskās un sintaktiskās marķēšanas rīks. Instrukcija.* [Skatīts 2011. gada 9. oktobrī.] Pieejama: http://www.semti-kamols.lv/doc_upl/Instrukcija.pdf.
9. Mežaraups, A. *A. Alunāna viencēliena "Paša audzināts" valodas patstāvīgo vārdu vārdnīca un palīgvārdu un izsauksmes vārdu vokabulārijs*. Diplomdarbs. Rīga : P. Stučkas Latvijas Valsts universitāte, 1959.
10. Petre, Biruta. Dažas atziņas par A.Upīša romāna "Zaļā zeme" vārdu semantiku un tās atspoguļojumu vārdnīcās. No: *Valodas aktualitātes – 1989*. Rīga, 1990, 107.–111. lpp.
11. Petre, Biruta. Galvenās Andreja Upīša romāna "Zaļā zeme" leksikas analīzes problēmas. No: *LPSR Zinātņu akadēmijas Vēstis*, Nr. 9. Rīga, 1987. – 134.–145. lpp.
12. Štekele, Baiba. *19. gadsimta 90. gadu latviešu dzejas valodas datorfonds.* Maģistra darbs. Latvijas Universitāte. Rīga, 1994.
13. TKMR – Tekstu korpusu marķēšanas rīks. [Skatīts.2011. gada 9. oktobrī.] Pieejams: http://www.semti-kamols.lv/?sadala=217.
14. Zeltiņa, V. *V. Plūdoņa poēmas "Atraitnes dēls" valodas vārdnīca*. Diplomdarbs. Rīga : P.Stučkas Latvijas Valsts universitāte, 1959.
15. Словарь языка Пушкина: в 4 т. / Отв. ред. акад. АН СССР В. В. Виноградов. 2-е изд., доп. / Российская академия наук. Ин-т рус. яз. им. В. В. Виноградова. М.: Азбуковник, 2000. http://www.slovari.ru/default.aspx?s=0&p=230. [Skatīts 10.09.2011.]
16. Эглите, Г. Г. *Лексикографический анализ словарного состава поэзии Райниса.* Автореферат диссертации на соискание ученой степени кандидата филологических наук. АН Латвийской ССР, Институт языка и литературы им. А. Упита. Р.,1988.

Juris Baldunčiks, Jānis Naglis

Ventspils Augstskolas Lietišķās valodniecības centra latviešu valodas resursi internetā

Ventspils Augstskolas Lietišķās valodniecības centra pētniecības virzieni un latviešu valodas resursu izstrādes nepieciešamība

Ventspils Augstskolas (VeA) Lietišķās valodniecības centra (LVC) galvenais darbības mērķis ir veikt zinātnisko pētniecību lietišķās valodniecības jomā, kā arī sekmēt tulkošanas nozares attīstību Latvijā. Minētā mērķa sasniegšanai VeA LVC ir izvirzīti šādi uzdevumi:

1) veikt pētījumus specializācijas virzienos:
   a) tulkošanas teorija un prakse, starpkultūru komunikācija,
   b) terminoloģija un nozarvalodas,
   c) valoda un tulkošana e-vidē,
   d) leksikogrāfijas teorija un prakse,
   e) salīdzināmā un kontrastīvā lingvistika;
2) veicināt maģistra un doktora līmeņa speciālistu gatavošanu savas zinātniskās darbības virzienos;
3) piedalīties VeA studiju programmu īstenošanā tulkojumzinātnes nozarē.

Izanalizējot mūsdienu latviešu valodniecības, kā arī tulkojumzinātnes un tulkojumvēstures stāvokli un šķēršļus, kas traucē to attīstību, kā būtiski trūkumi minami, pirmkārt, valodnieku darbošanās vairākās valodniecības apakšnozarēs bez pietiekami plašas valodas dotumu bāzes un, otrkārt, ļoti vāji izstrādāts Latvijas tulkošanas nozares attīstības dokumentējums. Šie negatīvie faktori būtiski ietekmē gan zinātnisko pētījumu un akadēmisko (bakalaura, maģistra) darbu kvalitāti, gan mācībgrāmatu un interneta materiālu informatīvo pilnīgumu un faktu drošumu. Lai vismaz daļēji novērstu šīs nepilnības, VeA LVC sāka ilgtermiņa pētniecības darbu trīs jomās: 1) valodu kontakti un aizguvumi latviešu valodā; 2) latviešu terminoloģijas vēsture un mūsdienu attīstība; 3) tulkošanas teorija un prakse Latvijā un starpkultūru komunikācija (1944–2011).

Šo pētniecības jomu svarīga sastāvdaļa ir galaprodukta izveide tīmeklim piemērotā formā, lai pētījumu rezultāti (arī starprezultāti) būtu pieejami visplašākajai auditorijai.

Valodas resursu materiāls un tā primārā apstrāde

Lai arī nekādi neizdodas iegūt kaut nelielu finansējumu minētajiem LVC pētījumu virzieniem, LVC notiek darbs pie zinātniskām tēmām, kas saistītas ar plaša latviešu rakstu valodas materiāla ieguvi ar mērķtiecīgas izlases ekscerpēšanas metodi (korpusa veidošanai ar avotu pilnīgu skenēšanu LVC nav līdzekļu). Darba galvenais mērķis un paredzamais rezultāts ir hronoloģiski sakārtotu autentisku valodas izrakstu jeb mikrotekstu kopas, kuras ļauj izsekot kādas leksēmas vai termina lietojumam latviešu valodā gadsimtu gaitā. Izraksti (ekscerpti) iegūti no vārdnīcām, dažādām grāmatām, brošūrām, reklāmlapām un citiem publicētiem sīkdarbiem, kā arī periodiskajiem izdevumiem. Ir izmantotas visas nozīmīgākās 17.–20. gs. pirmās puses vārdnīcas, oriģinālā un tulkotā literatūra, tomēr daiļliteratūra pagaidām veido ļoti nelielu daļu. Šie valodas dotumi nav leksikogrāfiski noformēti, respektīvi, nav izveidota vārdnīca tradicionālā izpratnē, tomēr ir veikta primārā apstrāde, piemēram, mikrotekstu korpusā „Aizguvumi latviešu valodā" ir nošķirti homonīmi, iezīmēts vārda polisēmiskais raksturs, dotas norādes uz galveno formālo variantu.

Latviešu valodas resursu izveide ir ilgs un darbietilpīgs process, īpaši jau finansējuma trūkuma apstākļos, tomēr sabiedrībai tie nepieciešami jau tagad. Tāpēc lietotājiem tiek nodotas jau pirmās iestrādnes, kas, lai arī nepabeigtas, tomēr būs noderīgas informācijai un zinātniskajam darbam un ļaus spriest par ieplānoto darbu kopapjomu un nozīmi.

„Aizguvumi latviešu valodā" (projekta vadītājs prof. J. Baldunčiks) ir alfabētiskā kārtībā sakārtots aizguvumu reģistrs, kurā iekļauti visi pilnīgi vai daļēji asimilētie aizguvumi (identismi un puskalki), kas sastopami latviešu rakstu avotos. Senākie reģistrējumi ir no 16. gadsimta beigām, jaunākie – 20. gadsimta 20.–30. gados. Rakstības ziņā teksti iespēju robežās saglabā oriģināla iezīmes. Tomēr tehnisku problēmu dēļ tā sauktais asais (pārsvītrotais) **s** aizstāts ar **ş**, bet fraktūras burtu alfabēta pēdējais burts – ar **3**. Izmantojot šos resursus nākotnē kādu lielāku un komplicētāku datubāzu veidošanai, tehniskā pielāgošana, šķiet, neradīs nekādus būtiskus sarežģījumus.

Aizguvumu reģistra šķirkļa uzbūves ilustrācijai var izmantot kopu **ādere**. Šķirkļa galvu veido: 1) leksēmas pamatforma pustrekniem burtiem; 2) norāde uz divām pamatnozīmēm (vai pietiekami atšķirīgām nozīmes niansēm); 3) norāde iekavās uz pirmās reģistrācijas gadu (17.–18. gs. vārdnīcām dots arī šifrs, piemēram, Elg 1683). Šai informācijai seko hronoloģiski sakārtots ekscerptu kopums. Avoti šifrēti, pamatā izmantojot literatūrā tradicionāli pieņemtos saīsinājumus (pilns saīsinājumu saraksts būs pieejams projekta noslēgumā), periodiskajiem izdevumiem norādīts numurs un lappuse. Sk. 1. tabulu.

1. tabula. **Vārda *ādere* šķirklis**

ādere 1/2 (Elg 1683)

Bićie źył – Aderkuſtenaßan	Elg 1683, 11
ahdere – Ader	St 1761, 13
Ader – dſihşle .. Blut Ader – ahdere, aşşins-dſihşle..	St 1789, 27
Tad wiņņam Aşşins (Ahdere) us Rohku jalaiſch..	Hag 1790, 9
..no tahm uhdens=ahderehm dabbu..	St 1796, 95
..şarkanas ahderes şchur tur rahdahs..	LGG 1797, 1, 40
..ſelta un şudraba ahderes atraſt.	LGG 1797, 2, 139
..aşşinis weegli 3aur ahderehm tekk..	LGG 1798, 1, 17
..ar şmilktehm, mahlu, glihſdu, klintu, radſehm, akmiņeem un uhdens ahderehm..	Mīl 1803, 31/33
..kas ahderu laidejeem, şahls puhşchlotajeem un pareggoņeem pakkaļ tekk.	LA 1822, 30, 3
..ahderi laiſdams ar radſiņeem no baltaja şkahrda.	LA 1823, 49, 2
..ahdere jalaiſch..	Adolfi 1837, 72
Eekşch ſtahdu eekşchahm tahdas reeres irr, ko warr tahm ahderehm lihdſinaht, kas dſihwneekeem şawâs meeşâs irr.	TLD 1839, 39, 154
..kas ahderes un ſarnas irr aisşpundejis..	LA 1840, 3, 10
Nowahri 2 kahlus wehſchu .. notihri wiņņus, tà ka tahs kahjas paleek klaht, teem maſakeem lauſ gluſchi nohſt un isņemm to melno ahderi ahrâ.	Lpg 1851, 122
..kas aşşinis israida 3aur şirds=şiteenu=ahderi pa wişşu meeşu.	MV 1857, 21, 166
..kur pahrdohs wişşadas Eņģlenderu dſelſu, mişşiņa, kappera un tehrauda leetas un rihkus, kà .. ahderu un şchreppu 3irrişchus..	MV 1858, 31, 248
..kaut arri aşşinis mannâs ahderês kà leddus şaşaltu..	Konter 1858, 12
Semme usņemm leetu un, kahdâ dſiļļumâ, eet gan ka 3aur ſemmes ahdereem, gan eekşch dohbjumeem şalaşşahs.. / ..kur eet tee minerala=ahderi jeb dſihşli.. / ..dſenn aşşinis 3aur şmalkahm truhbiņahm, 3aur aşşins=dſihşlehm jeb ahderehm..	Kav 1860, 6; 36; 127
Stahdijumu eekşchpuşşê irr tahdas daļļas, kas lohpu ahderehm lihdſinajamas..	MV 1861, 30, 239
..kreişajâ puşşê ne bij wişşas ahderes puşchu, bet tik ahda eeşkrambeta.	Bekm 1862, 39
..iſ weenas lihdſ diwi tuļļi reşnas (3aurmehrâ) dſelſu truhbas, kas teek ſemmê eedſihta, un prohti tik dſiļļi, kamehr uhdens=ahdere irr aiſşneegta.	BV 1870, 40, 315
3aur to, ka wiņņa katrâ minnutê 3–10 reiſ maſak pukſt, ta leek aşşinim rettak pa ahderehm şkreet..	BV 1870, 43, 340
Жила – dśihksla, ahdere, zihpsla – Sehne, Ader	KLVV 1872, 141

Жилка – dśihkliņa, ahderite; schķeedris – Aederchen; Faser	KLVV 1872, 141
Жилосечиво – ahderdśelśis – Schnepper	KLVV 1872, 141
Прожилок – dśihsla, ahdere (mineraļôs) – Ader	KLVV 1872, 489
Ahdere – die Ader; pulş-ahdere (auch dſihwibas-dſihşla, rohkas pulkſtenitis) – der Puls..	LVV 1872, 2
..augoşchâ mehneşî ir ahdere jalaiſch, mati un nagi jagreeſch..	Darbs 1875, 27, 326
ahdere – жила – Ader	LKVV 1879, 1
Ader – ahdere, aşins dſihşle; die güldene ~, ſelta ahdere Bergader – metala ahdere oder ſtruhkla kalnôs	VLV 1880, 27; 119
Nikotins pahreet ļoti ahtri aşinîs un kad to taişni eeşchļah3 kahdâ aşins=traukâ (ahderê), tad ſibeņa ahtrumâ ari ir wiņa nahwiga darboşchanàs redſama..	Rota 1884, 10, 115
arterija – pulşa adere.	SV 1886, 26
Ahdere, *şk*. dſihşla, arterija. Ahderes fiſtula, *şk*. dſihşlas fiſtula. Ahderu meſglis, *sk*. dſihşlu meſglis (Aderkropf).	KV 1891, 22
Aorta – galwenà aşiņu ahdere, kuŗa şahkas no şirds kreişàs daļas. No wiņas şahkas wişas arterijas, t. i. tahs ahderes, kuŗas wada pa meeşu aşinis..	KV 1891, 87
Noſeedſneeks, ſinams, domaja, ka wiņa rokas ahderes (dſihşlas) pahrgreeſtas un ka tagad wiņa aşinis tek te3edamas.	Vārpas II 1895, 23
..atraſtas ſelta ahderes..	MVM 1899, 1, 74
No aşiņu şpeedeena wenas iſplehşchas.. Şchahdi paplaşchinatas wenas dehwè par „krampju ahderèm".. Krampju ahderes wisbeeſchak eemetas apakşchſtilbôs, şewişchķi pee şeeweetèm..	Straut 1899, 49
Ahdere *şk*. dſihşlas.	KV 1904, 30
..wişu labako Daņu ſirgu ahderês rinķo „Oppenheima" aşinis..	RAv 1909, 150, 1
ahdere – die Ader.	LVV 1914, 10
ādere (gleich estn. āder entlehnt), Ader in weiterem Sinne, die Sehnen (dzīslas) einschliessend, Etn. II, 64; āderi laist (nach dem Deutsch), eine Ader lassen; āderi cirst dass.; laidīsim tikai kādam jēram āderi vaļā, wollen wir ein Lamm schlachten, Kaudz. M. 36	ME 1923, 236

Homonīmi iekļauti reģistrā kā atsevišķi šķirkļi ar arābu cipariem augšrakstā, piemēram, **aktīvs**[1] (darbīgs), **aktīvs**[2] (bilances daļa), **aktīvs**[3] (darāmā kārta). Turklāt šķirkļa galvā var tikt iekļautas šādas papildnorādes: 1) norāde uz paralēlformas šķirkli, piemēram, **abate** sk. arī *abtese* (**abtese** sk. arī *abate*), vai uz neatvasinātās formas šķirkli, piemēram, **aizstellēt** sk. arī *stellēt* (šķirkļa **stellēt** galvā ar hipersaiti būs doti visi atvasinājumi, resp., puskalki); 2) norāde pie formas varianta uz galveno formu, kur izveidota ekscerptu kopa, piemēram, **akrostihons** sk. *akrostihs*; 3) norāde sk. *Anglicismi* (informācija par konkrēto vārdu jāskatās monogrāfijā „Anglicismi latviešu valodā" (Rīga : Zinātne, 1989); 4) norāde sk. arī

Anglicismi (daļa informācijas par konkrēto vārdu atrodama šajā reģistrā, bet cita jāskatās iepriekš minētajā monogrāfijā).

A burta sadaļā izstrādāti vairāk nekā 1700 aizguvumu; tie galvenokārt ir t. s. identismi (piemēram, *abonents*, *afiša*, *āmurs*), kā arī puskalki (morfoloģiskie un morfoloģiski derivatīvie), piemēram, *aerokamanas*, *acetilsalicilskābe*, *augustinieši*, *aizkorķēt*, *apbūvēt*. Lai izvairītos no okazionālismiem, reģistrā iekļauti tikai tie aizguvumi, kas fiksēti vismaz divos avotos. Izņēmuma kārtā doti vārdi ar vienu reģistrējumu, ja tie lietoti rakstu avotos arī pēc Otrā pasaules kara.

Šāda mikrotekstu kopuma izveidē izšķiramas vairākas stadijas. Korpusa sākotnējais pamats bija projekta vadītāja kartotēka, kas tapusi 80. un 90. gados. Kartotēkas šķirkļavārdi tika ierakstīti datorā programmas *MicrosoftWord* tabulā, lai būtu iespējams to tālākveidot, jaunus mikrotekstus ievadot tieši šajā tabulā (sk. 2. tabulā ierakstus šķirklī *landhofmeistars* un *landtāgs*), nevis turpinot papīra kartīšu rakstīšanu.

2. tabula. **Ieraksti šķirklī *landhofmeistars* un *landtāgs***

landhofmeistars (1819)	
Şchee kungi bija ar wahrdeem: 3eenigs Landopmeiftera kungs, Barons Renne..	Laun 1819, 19
landmaršals (1823) *sk. arī* maršals	
landrāts (1857)	
landsmanis (1823)	
landšafte (1822)	
landšturms (1871)	
landtāgs (1857)	
Widſemmes muiſchneeku landtags tikka wakkar .. atklahts.	BV 1870, 1, 1

3. tabula. **Burta *B* kopuma fragments**

briljants[1] (1856)	
Brillanta gredſenus..	MV 1856, 15, 113
..dahrgeem şlihpeteem akmiņeem, kà dimanteem (jeb briļļjanteem)..	V 1857, 43, 345
..ar brillanta akmiņeem..	MV 1862, 24, 185
..brangs rohſchu=şarkans brilliants..	LA 1863, 13, 78
..briljanti (dahrgi akmiņi) şpihguļoja..	MV 1869, 32, 253
..Ķeiſereene tai dahwajuşe ſelta aproh3i ar şmaragdeem un briljanteem.	BV 1871, , 26
Бриліант – briljants (dargs, slihpets dimants) – Brillant, geschliffener Diamant	KLVV 1872, 26
..laiſtijàs wişadôs briļļantôs.	Aus 1885, 7, 431
..leeli briljanti..	Jrūd 1900, 40
Par briljantu şau3 şlihpetu dimantu..	Jrze 1903, 42
Briljants – noşlihpets dimants, kas laiſtas.	SV 1911, 57
briljants[2] (1885)	
Wakarâ pretîm pilei bija briljanta uguņoşchana.	Rota 1885, 34, 405
brilles (Lj 1685)	
Brille – Augenspiegel, Brille	j 1685, 20
brille – die Brille, iſt beſſer als brillis	Wel 1828, 24
br lles..	Mag 1832, 39
..brille muhşu wahjahm a33im palihgâ nâk..	MV 1856, 11, 84
Semneeks: „Es aimirşu mahjâ brilli, un bes brilles newarru laşşiht..	DunB 1868, 3, 11
..ar brilli us degguna..	DunB 1868, 13, 51
..penşene jeb uſ deguna uſşprauſchama brille peekahrta pee kruhtim..	BV 1883, 184, 2

Pēc kartotēkas pārrakstīšanas tabulā šķirkļi kļūst ļoti atšķirīgi, jo dažādiem vārdiem kartotēkā uzkrājies atšķirīgs ekscerptu skaits. Nelielu burta *B* kopuma fragmentu sk. 3. tabulā.

Tālākajā darba gaitā šķirkļi tiek papildināti ar nozīmīgāko 17.–18. gs. vārdnīcu materiālu, sākot ar G. Manceļa 1638. gada *Lettus* un šai vārdnīcai pievienoto frazeoloģijas un sarunu daļu (M 1638). No 19. gs. avotiem obligāti ekscerpējamie ir Kr. Valdemāra vārdnīcas (KLVV 1872, LKVV 1879, KLV 1890), K. Ulmaņa un J. Neikena (LVV 1872), G. Bražes (1880), Konversācijas vārdnīca (KV 1891–1898), kā arī 1878. un 1886. gada svešvārdu vārdnīcas (SV 1878, SV 1886). 20. gs. pirmajā pusē obligāti ekscerpējamie avoti ir Konversācijas vārdnīca (KV, sākumdatējums – 1904. gads), J. Dravnieka vācu–latviešu vārdnīca (VLV 1910), svešvārdu vārdnīcas

4. tabula. **Vārda *agresija* šķirklis**

agresija (1914)	
agreşija – uſbrukums.	SV 1914, 16
agresija – l. *aggressio*, uzbrukums.	SV 1926, 3
Agresija (lat. *aggressiō*) – uzbrukums.	KV 1927, 109

(SV 1904, SV 1906, SV 1908, SV 1911, SV 1912, SV 1914, SV 1926, SV 1934), K. Mīlenbaha un J. Endzelīna vārdnīca (ME, sākumdatējums – 1923. gads) un tās papildinājumi (EH, sākumdatējums – 1934. gads), kā arī Latviešu konversācijas vārdnīca (sākumdatējums – 1927. gads). Bet izlases kārtībā tiek ekscerpētas arī vairākas citas vārdnīcas. Citu avotu izmantojums netiek reglamentēts. Jau citētais vārds *ādere* ilustrē korpusa šķirkli pēc vārdnīcu materiāla pievienošanas. Tomēr ne visos šķirkļos ir tik daudz materiāla, piemēram, vārdam *agresija* pagaidām ir tikai trīs ieraksti (sk. 4. tabulu).

Reģistrā „Latviešu terminoloģijas attīstība" materiāls dalīts pēc nozarēm (fizika, ķīmija, valodniecība u. c.), bet atsevišķu terminu kopas kārtotas pēc vēsturiski onomasioloģiskā principa – hronoloģiskā secībā vienkopus doti visi kāda jēdziena apzīmējumi latviešu valodā. Pašlaik lielākās iestrādes ir valodniecības terminoloģijā, samērā plaši materiāli apkopoti arī dabas zinātņu, transporta un teātra nozarē. Ilustrācijai var noderēt mikrotekstu kopums *ūdeņradis* no ķīmijas apakšnozares (sk. 5. tabulu).

5. tabula. **Vārda *ūdeņradis* mikrotekstu kopums**

ūdeņradis	**Autors/avots, gads**
Uhdens, ko wişşi paſihſtam, naw 3its nekas, kà uhdens gaişs, kas ar şkahbes gaişu şajau3ees.	K. Lepevičs, 1852, 123
..no tşchetrahm grunts=leetahm, no şkahbraſcha, uhdenraſcha, ohgles un şlahpraſcha..	K. Lepevičs, MV 1857, 3,31
Uhdenainis irr gaişs bes pehrwes, şmarſchas un şmakas..	J. Alunāns/SDP I, 1860, 50
Udenî şaweenoti diwi mehri ugguns gahſa ar weenu mehru şkahbekļa..	H. Kavals/LA 1865, 3, 20
Şchihs gruntsleetas eedallahs şchahdâ wihſe: .. şkahbeklis, uhdenraddis, şlahpeklis..	PA 1865, 3, 23
Uhdenraddons (U) tà pat beſ pehrwes, şmakka un gahrşchas..	Kr. Dinsbergs/BV 1869, 25, 197
..pildija weenu puhşli ar ohgļu şkahbjuma gahſi un wirs şcha ohtru ar uhdens buhşchanas gahſi..	P. Sternmanis, 1869, 90
..if şlahpekļa un uhdenekļa (Wasserstoff) şaweenoşchanahs.	BV 1871, 35, 276
Uhdenradis, ģidrogens – Wasserstoff	KLVV 1872, 50

..uhdeņradim* un ļahpeklim lihdſinajahs.. *Uhdeņradis ir twaiks, 3aur kuŗa ķimiķu jeb tà ļakot eekļchķigu ļaweenoļchanohs ar ļkahbekļa twaiku 3eļahs uhdens.	BV 1873, 39, 303
Ļkahbeklis, ohgļradis un uhdensradis ir elementi.. Tee wahrdi ogļrads (Kohlenstoff) un uhdensradis (Wasserstoff) naw riktigi atwaļinati.. Riktigak buhtu jatei3 „ogļu weela" un „uhdens weela", jeb, kà ari daſchur laļams, „ogļains" un „uhdeņains".	A. Dīriķis/BV 1875, 25, 198
..uhdensweela (Wasserstoff)..	BV 1876, 18, 141
Uhdenradis – водород, водотвор..	LKVV 1879, 234
..58) hidroģens, uhdensweela..	J. Bankins 1880, 59
..3aur ļawu ogļu=weelas, uhdeņa=weelas un ļkahbekļa ļaturu..	ĪTK, 1882, 73
..2 apjomus (wolumus) uhdenekļa un 1 apjomu skahbekļa..	SDP 1893, 73
..pee ļlahpekļa, uhdeņweelas..	Austr 1896, 10, 944
Водород – uhdenradis, hidrogens	KLV 1903, 31
Hidrogens – uhdens weela jeb uhdeni radoļcha weela	SV 1904, 101
..peem. uhdeņradis, dſelſs, ļehrs..	DzV 1909, 232, 6
Hidrogens – uhdeņradis	SV 1911, 141
водород – Wasserstoff – ūdeņradis	ZTV 1922, 72

Tulkošanas teorijas un prakses materiāls un tā primārā apstrāde

Šim pētniecības virzienam ir starpdisciplinārs raksturs: apzināt un sistematizēt informāciju par tulkošanas nozari, tulkojumiem un tulkotājiem laika posmā no 1944. līdz 2011. gadam, kas vienlaikus aptver tulkošanas teorijas un prakses, valodniecības un starpkultūru komunikācijas aspektus.

Lai sasniegtu izvirzītos mērķus, viss pētniecības darbs būtu veicams trīs virzienos ar konkrētiem uzdevumiem katrā no tām: 1) ar tulkošanas teoriju un praksi saistītās publikācijas: jāapkopo un jāsistematizē publikācijas, kas veltītas tulkošanas teorijai un/vai praksei, jāizveido to pilnu tekstu (kopiju) digitālā datubāze; 2) tulkojumu bibliogrāfija: jāapkopo un jāsistematizē visu perioda tulkojumu bibliogrāfija; 3) tulkošanas speciālistu un organizāciju datubāze un profesionālo dzīvesstāstu videoarhīvs: jāapkopo un jāsistematizē informācija par visiem tulkotājiem, tulkojumzinātniekiem un tulkojumkritiķiem, kā arī terminologiem, kas cieši saistīti ar tulkošanas nozari. Šādam kompleksam pētījumam būtu nepieciešams samērā liels finansējums, tāpēc pašreizējā situācijā jāizvēlas alternatīva, resp., svarīgākais – ar tulkošanas teoriju un praksi saistīto publikāciju apzināšana un apstrāde.

Darbs ietver šādus uzdevumus: 1) visu iespējamo bibliogrāfisko rādītāju un uzziņu avotu apzināšana uz izpēte; 2) iegūto publikāciju analīze, īsa anotēšana un šķirošana pēc žanriem un tulkošanas problēmām; 3) autoru enciklopēdiskā rādītāja veidošana; 4) to svarīgāko nozares notikumu un norišu apzināšana pasaulē, kuras tiks izmantotas hronoloģiskajam sastatījumam.

Darba rezultāts ir vairāki bibliogrāfijas apkopojumi un rādītāji: 1) tulkojumzinātnes un tulkojumkritikas bibliogrāfija; 2) ar tulkošanas teoriju un praksi saistīto publikāciju pilnu tekstu (kopiju) datubāze. Pašlaik ir ieskenētas aptuveni 140 publikācijas (1944–1968), kas tiks izvietotas mājaslapā. Tiek apzinātas un sagatavotas nākamā perioda (1969–1991) publikācijas.

Latviešu valodas standartizācijas materiāls un tā primārā apstrāde

Šajā sadaļā tiek apzināta latviešu valodas standartizācijas vēsture no J. Alunāna līdz 20. gs. 30. gadiem: bibliogrāfija, pilni publikāciju teksti vai fragmenti, kas saistīti ar visdažādākajām valodas standartizācijas un attīstības problēmām. Pašlaik ir sagatavoti aptuveni 10 teksti no 19. gadsimta, gandrīz pilnībā saglabāts to autentiskums, tikai fraktūras burti aizstāti ar mūsdienu burtiem.

VeA LVC mājaslapa kā valodas resursu lietošanas punkts

Ņemot vērā plānoto resursu lielo apjomu, jāizveido ērti lietojama un tajā pašā laikā ātri funkcionējoša resursa lietotāja saskarne jeb interfeiss. Vizuālajam noformējumam nevajadzētu būt neizteiksmīgam vai uzbāzīgam, svarīgākais, lai tas būtu viegli uztverams un būtiskākie komponenti pietiekami izcelti. Resursu navigācijai un informācijas izvietojumam jābūt intuitīvam, t. i., informācijai jābūt izvietotai pēc iespējas pārskatāmāk, kā arī tā nedrīkst būt pārāk sīki strukturēta, jo tas varētu radīt neērtības vai pat lietotāju apjukumu.

Valodas resursiem jābūt arī salīdzinoši viegli uzturamiem, jo resursu papildināšanu veiks cilvēki, kas nav saistīti ar informācijas tehnoloģijām. Tātad projektā jābūt integrētai satura pārvaldības sistēmai, ar kuras palīdzību ir iespējams viegli rediģēt mājaslapas saturu un resursus.

Ņemot vērā, ka mājaslapā izvietojamo valodas resursu formāti ir dažādi (teksts, tabulas, faili ar *.doc* un *.pdf* paplašinājumiem), kā arī tēmas ir dažādas, resursi jāsadala pa nodaļām, kuru nosaukumi skaidri raksturotu to saturu.

Mājaslapa un resursi tika veidoti, izmantojot tīmekļa programmēšanas valodas (*HTML, CSS, PHP, JavaScript (jQuery), MySQL*). Projekts tika izvietots uz Ventspils Augstskolas servera.

Resursu funkcionalitāte un attīstības iespējas

Tika izveidota Lietišķās valodniecības centra mājaslapa, kas ir izvietota uz Ventspils Augstskolas servera. Mājaslapas adrese ir *http://lvc.venta.lv*. Mājaslapā ir atrodamas gan informatīvas sadaļas (*Ziņas, Par mums*), gan dažādi resursi (*Latviešu valodas resursi, Tukošanas teorija un prakse, Uzziņu literatūra)*. Mājaslapas karte:

- Ziņas (jaunākās LVC ziņas, informācija par resursu atjauninājumiem)
- Latviešu valodas resursi
  – Aizguvumi latviešu valodā
  – Latviešu terminoloģijas attīstība
- Tulkošanas teorija un prakse (skenēti materiāli par tulkošanas teoriju un praksi Latvijā)
- Uzziņu literatūra (tematiskie teksti par dažādām valodniecības problēmām)
- Par mums
  – Gada pārskats
  – Projekti

Mājaslapa realizēta uz satura pārvaldības sistēmas *Wordpress* bāzes (sk. 1. attēlu).

Wordpress ir mūsdienīgs blogu dzinējs, kurš izstrādāts, uzsverot vizuālo noformējumu, tīmekļa standartus un ērtu lietojamību. Tas ir bezmaksas, atvērtā koda rīks, kas darbināms uz tīmekļa servera. *Wordpress* galvenokārt ir paredzēts blogu veidošanai, bet tikpat ērti un viegli uz tā bāzes var tikt realizēta jebkura cita veida mājaslapa. Satura pārvaldības sistēma ļauj ātri un ērti pievienot, labot vai dzēst lapas saturu.

Wordpress platforma ir ļoti izplatīta, tāpēc internetā ir pieejama plaša informācija, kā arī platformas funkcionalitāti var viegli papildināt, pievienojot dažādus gatavus moduļus.

1. attēls. ***Wordpress* satura pārvaldības sistēma**

No resursiem šobrīd mājaslapā ir pieejami „Aizguvumi latviešu valodā". Pagaidām datubāzē vēl tiek ievadīti vārdi, kas sākas ar burtiem *A* un *Ā*, bet jau tuvākajā laikā pakāpeniski tiks pievienoti vārdi, kas sākas ar burtiem *B*, *C* utt.

Atverot resursu „Aizguvumi latviešu valodā", lapas augšējā daļā (zem navigācijas joslas) parādās alfabēta josla. Alfabētiskā sadalījuma pieejamība atkarīga no tā, vai datubāzē ir ievadīti vārdi, kas sākas ar attiecīgās sadaļas burtu. Un zem tās parādās divas kolonnas ar pogām (sk. 2. attēlu).

Katra poga apzīmē lappusi ar vārdiem, piemēram, ja pogas nosaukums ir „aberācija – abolicija", tas nozīmē, ka, uzklikšķinot uz šīs pogas, atvērsies kolonna ar vārdiem, sākot ar vārdu „aberācija" un beidzot ar vārdu „abolicija" (sk. 3. attēlu). Lapas augšpusē un apakšpusē parādās lappušu joslas.

Uzklikšķinot uz kādas rindas vai vārda, atsevišķā panelī parādās attiecīgā vārda avotu tabula (sk. 4. attēlu).

Tabulu var aizvērt, uzklikšķinot uz attiecīgā vārda vai rindas vai arī atverot kāda cita vārda avotu tabulu.

Kad lapas administrators ir autorizējies mājaslapas satura pārvaldības sistēmā, resursā „Aizguvumi latviešu valodā" parādās papildu iespējas – jaunu vārdu pievienošanas forma (sk. 5. attēlu), rediģēšanas forma (sk. 6. attēlu) un vārdu dzēšanas iespēja (sk. 7. attēlu).

Iepriekš aprakstītais resurss „Aizguvumi latviešu valodā" programmēts, izmantojot *AJAX* tehnoloģiju – tiek ielādēta tikai tā lapas daļa, kurā jāatjauno informācija, nevis tiek pārlādēta visa lapa. Šāda pieeja prasa mazāk laika, lai ielādētu nepieciešamo informāciju lapā. Arī lietotājam ir ērtāk un patīkamāk lietot lapu ar dinamisku saturu.

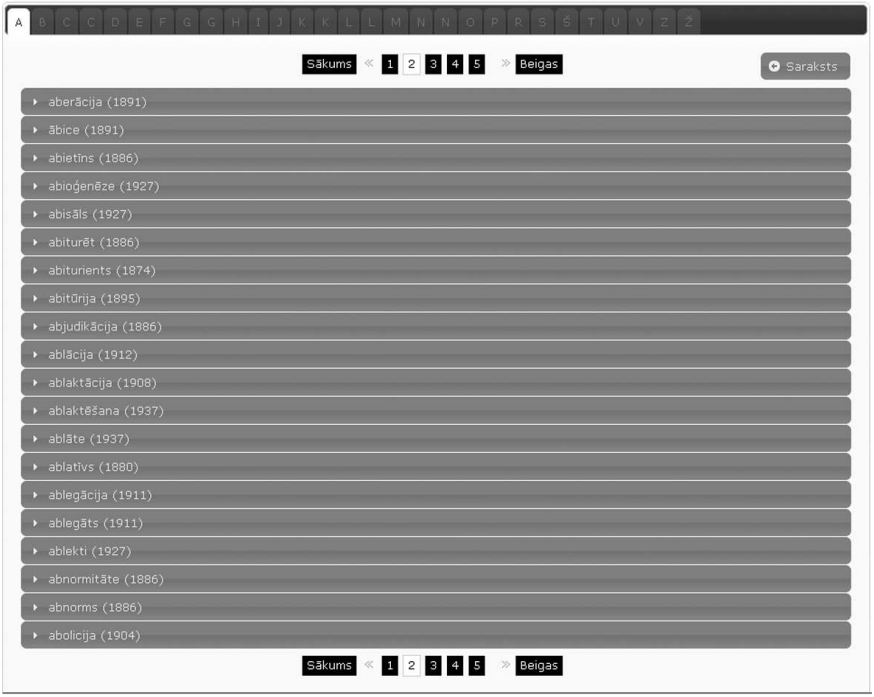
Taču *AJAX* tehnoloģijai ir arī savi mīnusi. Pirmkārt, paralēli datu ielādei lapā adreses joslā neveidojas adrese uz kādu konkrētu datu izkārtojumu lapā. Līdz ar to zūd iespēja norādīt ar hipersaiti uz kādu konkrētu resursu. Otrkārt, *Google* un citu meklēšanas dziņu „roboti", kas „apstaigā" mājaslapas un ievāc informāciju, „neredz" šādu dinamiska tipa saturu. Līdz ar to mēs nevaram atrast meklēšanas dziņos šo informāciju, kas tiek ielādēta lapā dinamiski (izmantojot *JavaScript*).

Par laimi, šīm problēmām ir atrasts risinājums, kas tiek pakāpeniski ieviests projektā.

Projekts joprojām atrodas izstrādes režīmā. Tiek novērstas dažādas kļūdas un tehniskas nepilnības. Tuvākajā nākotnē LVC mājaslapā mainīsies gan vizuālais noformējums, gan funkcionalitāte, lai nodrošinātu lietotājiem patīkamāku un ērtāku resursu apguvi.

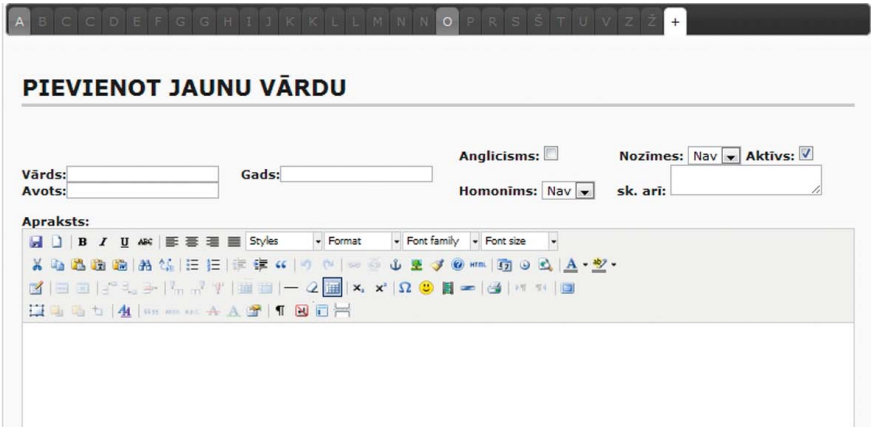
2. attēls. **Vārdu alfabētiskais sadalījums**

3. attēls. **Resurss „Aizguvumi latviešu valodā"**

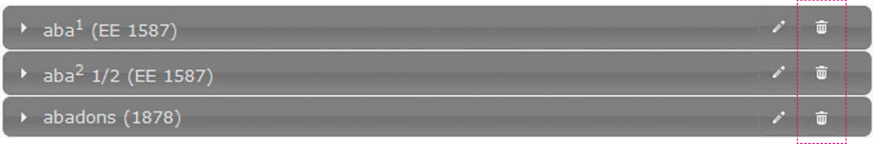


4. attēls. **Vārda *abietīns* avotu tabula**

5. attēls. **Jauna vārda pievienošanas forma**



6. attēls. **Administratoram pieejamās rediģēšanas un dzēšanas pogas**



7. attēls. **Vārda *abandonēt* dzēšana**

Valda RUDZIŠA

Datorizētā tulkošana tulkošanas studijās un praksē

Ievads

Tulkošanas speciālistu sagatavošana šobrīd ir uzskatāma par vienu no prioritārajiem virzieniem izglītībā, jo modernās sabiedrības sekmīga attīstība lielā mērā ir atkarīga no informācijas pieejamības, un šai informācijai jābūt pieejamai valsts valodā. Latvijas iekļaušanās Eiropas Savienībā ir ietekmējusi ne vien valsts ārējos sakarus un ekonomiku, bet būtībā visas sabiedrības dzīves jomas, tādēļ visdažādākajās nozarēs paplašinās un padziļinās starptautiskie kontakti. Līdz ar to aizvien lielāku nozīmi iegūst tulkošanas pakalpojumi. Jebkuras iestādes vai uzņēmuma veiksmīga darbība globālajā informācijas sabiedrībā ir atkarīga no sekmīgas komunikācijas un savlaicīgas informācijas apmaiņas. Jo īpaši svarīgi tas ir uzņēmumiem, kuri vēlas nodrošināt savu konkurētspēju starptautiskajā tirgū, piedāvājot inovatīvus produktus. Kā liecina statistika, apmēram 80 % uzņēmumu ir saistīti ar darbošanos starptautiskajā tirgū. Arī Latvijas eksportspēja zināmā mērā ir atkarīga no tulkošanas pakalpojumu kvalitātes, ja ņemam vērā, ka daudzi Latvijas ražojošie uzņēmumi pārsvarā nodarbojas ar savas produkcijas eksportu.

Svarīga ir ne tikai paša produkta izstrāde, bet arī attiecīgās dokumentācijas un reklāmas materiālu sagatavošana. Lai atbilstīgais produkts tiktu atpazīts un par

to rastos interese ne tikai vietējā, bet arī starptautiskā tirgū, ļoti būtiska nozīme ir t. s. produkta lokalizācijai. Ar šo procesu nodarbojas tulkošanas speciālisti, ne tikai tulkojot tekstus svešvalodā, bet arī pielāgojot tos mērķauditorijas (svešās kultūras) īpatnībām.

Izstrādājot produkta piedāvājumu interneta vidē, nepieciešamas specifiskas zināšanas un prasmes, piemērojot specifiskas tulkošanas programmatūras. Šādas programmatūras izmantošana nodrošina ātru un kvalitatīvu informācijas pieejamību produkta mērķauditorijai, nodrošinot vienveidīgu terminoloģijas lietojumu, kas nepārprotami atvieglo komunikāciju nozares speciālistu vidū. Šādas programmatūras izmantošana samazina arī uzņēmuma izmaksas, jo ar programmatūras palīdzību īsā laikā iespējams nekļūdīgi atjaunot nepieciešamo informāciju. Svarīgs konkurences veicināšanas faktors, kā jau minēts, ir pēc iespējas ātra un nevainojama jaunā produkta tehniskās dokumentācijas sagatavošana gan vietējā valodā, gan svešvalodās.

Pētījumi liecina, ka tulkojamo tekstu apjoms nemitīgi pieaug, it īpaši tajās jomās un valodu kombinācijās, kurās specializējas Ventspils Augstskolas Tulkošanas studiju fakultāte: ekonomika, jurisprudence, lietišķā komunikācija.

Tulkošanas tehnoloģijas nozīme tulkošanas studijās

Mūsdienu apstākļos starptautisko uzņēmumu efektīva darbība aizvien vairāk ir atkarīga no informācijas svešvalodā. Līdz ar to tulkošana ir kļuvusi par svarīgu ekonomikas faktoru. Nereti tulkojamā informācija ir ļoti apjomīga, un uzņēmumiem tā jāiegūst ātri. Šādu pasūtījumu izpilde nav iespējama bez modernās tulkošanas tehnoloģijas (tulkošanas rīku) izmantošanas un kvalificētu speciālistu iesaistīšanas ne tikai tulkošanas, bet arī terminoloģijas, teksta redakcionālās apstrādes un tulkošanas projektu pārvaldībā. Tulkošanas tehnoloģijas rīku lietošana nodrošina efektīvāku tulkošanas procesu (ātrums, precizitāte, zemākas izmaksas), konsekventu terminoloģijas lietošanu, tulkošanas atmiņas uzkrāšanu un izmantošanu turpmākiem tulkojumiem.

Sabiedrībai ir nepieciešami speciālisti, kuri spēj profesionāli organizēt apjomīgus tulkošanas projektus mūsdienu prasībām atbilstošā tulkošanas birojā, izmantojot modernos tulkošanas rīkus, piemēram, elektroniskās tulkošanas atmiņas (*translation memories*), terminoloģijas elektroniskās datubāzes (*MultiTerm* u. c.), mašīntulkošanas programmatūras atsevišķiem teksta veidiem un citus elektroniskos resursus. Līdz ar to svarīga tulkošanas studiju programmas sastāvdaļa ir kursi, kuros studējošie apgūst prasmes un iemaņas tulkošanas rīku izmantošanā tulkošanas procesā.

Tulkošanas biroja konkurētspēja ir atkarīga no moderna aprīkojuma un prasmēm šo aprīkojumu izmantot savā profesionālajā darbībā. Ņemot vērā minētās tehnoloģijas attīstības tendences, Ventspils Augstskolas Tulkošanas studiju fakultāte ir pievērsusies tādu tulkošanas speciālistu sagatavošanai, kuri spēj a) profesionāli organizēt apjomīgus tulkošanas projektus mūsdienu prasībām atbilstošā tulkošanas birojā, izmantojot modernus tulkošanas rīkus; b) tulkot dažādu produktu dokumentāciju, tirgzinības materiālus un citu papildinformāciju, kā arī ar produktu saistītos, internetā publicētos materiālus un pielāgot mērķvalodas kultūras īpatnībām, ko sauc par lokalizāciju; c) prast noformēt dažādu formātu elektroniskos dokumentus, izveidot terminoloģijas elektroniskās datubāzes un elektroniskās tulkošanas atmiņas atbilstoši pasūtītāja interesēm; d) darboties tulkošanas projektā ne tikai kā tulkotājiem, bet arī kā terminologiem, redaktoriem un speciālistiem, kas organizē elektronisko tulkošanas atmiņu un terminoloģijas datubāzu pārvaldību un uzturēšanu, sniedz profesionālas konsultācijas uzņēmuma terminoloģijas datubāzu un tulkošanas sistēmu izveidē un uzturēšanā.

Populārākās tirgū pieejamās tulkošanas programmatūras ir, piemēram, *SDL TRADOS, MemoQ, Wordfast, Deja Vu, TRANSIT, ACROSS*. *SDL TRADOS* tiek uzskatīta par līderi globālajā informācijas pārvaldības sistēmā, piemēram, Eiropas Savienības institūcijās pārsvarā izmanto šo programmatūru. Arī Ventspils Augstskolā mācību mērķiem tiek izmantots *SDL TRADOS*. Taču augsti kvalificētam tulkošanas speciālistam jāorientējas arī citu ražotāju tulkošanas programmatūru piedāvājumos, tādēļ studiju procesā cenšamies sniegt informāciju ne tikai par uzņēmuma *SDL TRADOS* tulkošanas rīkiem, bet arī citu uzņēmumu ražojumiem, lai jaunais speciālists prastu novērtēt, kurš no rīkiem viņa darba specifikai būtu vispiemērotākais. Pēdējā laikā praktiķu vidū populāra ir arī tulkošanas programma *MEMOQ*, kuru izstrādājis 2004. gadā dibinātais Ungārijas uzņēmums *KILGRAY. Translation Technologies*.

Pēdējos divdesmit trīsdesmit gados ir būtiski mainījusies tulkošanas darba vide. Tulkošanas speciālista darbs nav iedomājams bez specifiskām zināšanām valodas tehnoloģiju jomā, prasmēm tās izmantot savā profesionālajā darbībā. Latvijā diemžēl joprojām pastāv uzskats, ka tulkotāja galvenais darbarīks ir vārdnīcas un dators, kurā tiek ievadīts tulkojums. Par to liecina fakts, ka tulkošanas studijas joprojām tiek ierindotas humanitāro zinātņu nozarē.

Tādējādi netiek ņemts vērā, ka tulkošana ir nozare, kurai nepieciešami apjomīgi ieguldījumi atbilstošas infrastruktūras izveidei, kā arī kvalitatīvu rezultātu sasniegšanai nepieciešams darbs nelielās grupās. Studiju laikā ir svarīgi regulāri sadarboties ar tulkošanas biroju pārstāvjiem, lai studējošie zinātu par jaunākajām tendencēm tulkošanas tirgū, par jaunāko tulkošanas tehnoloģiju izmantošanu

tulkošanas procesā, jo, kā zināms, tehnoloģijas pēdējos gados attīstās īpaši strauji. Arī pasniedzējiem ir jābūt ļoti elastīgiem, lai spētu savus studiju kursus nemitīgi papildināt un uzlabot saistībā ar modernās informācijas sabiedrības attīstības tendencēm.

Modernas, mūsdienu prasībām atbilstošas tulkošanas studijas ir uzskatāmas par starpdisciplināru jomu, kas lielā mērā pārklājas ar informācijas tehnoloģijas nozari. Datorlingvistikas rašanās ir cieši saistīta ar mašīntulkošanas attīstību, kas vēl aizvien ir viens no datorlingvistikas virzieniem līdzās daudziem citiem. Svarīgs šīs nozares praktiskais aspekts ir arī tulkošanas tehnoloģijas izmantošana, optimizējot tulkošanas procesu.

Mašīntulkošana versus datorizētā tulkošana

Interesanti, ka pēdējos gados liela uzmanība tiek pievērsta nevis datorizētajai tulkošanai, bet gan mašīntulkošanai. Tas ir arī saprotams, jo globalizācijas apstākļos informācijas apmaiņas apjoms starptautiskajā publiskajā telpā strauji aug un valodas barjeras tiek uzskatītas par šķērsli šīs informācijas pieejai. Eiropas Savienība savā teritorijā ir nodrošinājusi četras brīvības, proti, preču, kapitāla, pakalpojumu un personu brīvu kustību, taču šo brīvo apriti traucē valodas barjera, un ir nepieciešami risinājumi šo barjeru mazināšanai vai novēršanai. Lielas cerības šai sakarā tiek liktas uz mašīntulkošanas attīstības iespējām nākotnē, jo, kā zināms, tieši pēdējos gados tā piedāvā daudzus inovatīvus risinājumus.

Kaut arī mašīntulkošana nav šā raksta tēma, tomēr būtu vērts precizēt, ko saprotam ar šo terminu, jo praksē mašīntulkošana un datorizētā tulkošana nereti tiek jaukta, kā arī mašīntulkošanas vietā lietots termins „automātiskā tulkošana". *Valodniecības pamatterminu skaidrojošajā vārdnīcā* mašīntulkošana ir definēta kā „tulkošana, kuru pilnībā, cilvēkam šajā procesā neiejaucoties, veic dators" (VPSV 2007, 227). Mašīntulkošana cilvēkus ir fascinējusi jau datoru rašanās pirmsākumos. Pirmā mašīntulkošanas programma tika izstrādāta 20. gadsimta 50. gados Džordžtaunas Universitātē ASV, kaut gan interese par automātiskās tulkošanas iespējām radās jau līdz ar pirmo datoru izstrādi. Pirmā datorspeciālistu izstrādātā mašīntulkošanas programma tika demonstrēta 1954. gada 7. janvārī. Tās pamatā bija visai vienkārša sistēma – tā balstījās uz noteiktu gramatikas algoritmu un vārdnīcu ar 250 šķirkļiem; tulkojamais teksts bija veltīts dažādām tēmām, piemēram, matemātikai, politikai, jurisprudencei, ķīmijai u. c. (Hutchins 1954)

Sākumā valdīja zināma eiforija, jo datorspeciālisti, attīstot ideju par automatizētu tulkošanu bez cilvēka palīdzības, bija pārliecināti, ka tulkotāja profesija drīzumā nebūs nepieciešama, šo speciālistu pilnībā aizstās dators. Taču, turpinot darbu

pie šā jautājuma, itin drīz kļuva skaidrs, ka „mašīna" tulkotāju aizstāt nevar. Protams, attīstoties modernajām tehnoloģijām, mašīntulkošanas rīkus iespējams uzlabot, jo īpaši tas attiecināms uz t. s. lielajām valodām (angļu, franču, vācu), kur mašīntulkošanas programmas ātrai informācijas ieguvei izmanto samērā plaši un to kvalitāte salīdzinājumā ar mazajām valodām ir visai pieņemama. Protams, šajā gadījumā nav runa par kvalitatīvu tekstu, tomēr vispārīgu priekšstatu par tulkotajā tekstā pieejamo informāciju iegūt ir iespējams.

Speciālisti intensīvi strādā arī pie mazo valodu, t. sk. latviešu valodas, mašīntulkošanas rīku pilnveidošanas. Pēdējo gadu laikā mašīntulkošanas programmatūru kvalitāte ir ievērojami uzlabojusies, jo tiek realizēti vairāki Eiropas Savienības finansēti projekti. Viens no tādiem ir projekts *META-NET*[1], kas ir uzskatāms par ļoti veiksmīgu sadarbības tīklu mašīntulkošanas jomā. To veido 47 izpētes centri 31 valstī, un to līdzfinansē Eiropas Komisija. *META-NET* veido Daudzvalodu Eiropas tehnoloģisko savienību (*META – Multilingual Europe Technology Alliance*), kurā iesaistījušās vairāk nekā 280 valodas tehnoloģiju pētniecības un izstrādes organizācijas no 40 valstīm. *META-NET* aktivitātes Ziemeļvalstīs un Baltijas valstīs nodrošina projekts *META-NORD*, kurā sadarbību starp vadošajiem šo valstu pētniecības centriem koordinē Latvijas uzņēmums *Tilde*. No Latvijas *META-NET* tīklā piedalās arī Latvijas Universitātes Matemātikas un informātikas institūts.[2]

Runājot par mašīntulkošanas programmu izmantošanu tulkošanas praksē, nevar pilnībā noliegt to piemērotību. Ir vairāki teksta veidi, kuru tulkošanā mašīntulkošanas programmas ir veiksmīgi izmantojamas, piemēram, tehniskajos tekstos (produktu tehniskā dokumentācija, lietošanas pamācības), kur vārdus parasti lieto tikai vienā nozīmē. Piemēram, starptautiskais koncerns *SIEMENS* 20. gadsimta 80. gados savām vajadzībām izstrādāja mašīntulkošanas programmu *METAL (Machine evaluation and translation of natural language)*[3]. Jau 70. gados Monreālas Universitātē izstrādāta programma *TAUM Meteo*, kuru veiksmīgi izmantoja Kanādā laika ziņu tulkošanā no angļu valodas franču valodā (Snell-Hornby et al. 2006, 139). Zināmu popularitāti šobrīd ir ieguvis internetā pieejamais *Google* tulkotājs, kuru daudzi kļūdas pēc uzskata pat par universālu tulkotāja aizvietotāju. Sociālajos tīklos risinās diskusijas par šā rīka kvalitāti, tiek minēti dažādi kuriozi piemēri.

Intensīvi strādājot pie mašīntulkošanas programmatūrām, jau 60. gados kļuva skaidrs, ka datorzinātnes pārstāvji un mašīntulkošanas entuziasti līdz galam nav novērtējuši valodas komplekso raksturu. Skaidrs, ka neviena programma nespēs

[1] META-NET tīmekļa vietne: http://www.meta-net.eu/?set_language=lv
[2] Sīkāku informāciju sk.: http://www.meta-net.eu/mission-lv
[3] META-FORUM tīmekļa vietne: http://www.meta-net.eu/events/meta-forum-2011/

atrisināt sinonīmijas vai homonīmijas jautājumu (*Man šī maizes šķēle šķiet par plānu; Mēs esam izstrādājuši lielisku plānu* – īpašības vārds vai lietvārds akuzatīvā?). Pamazām tika attīstīta ideja, ka kvalitatīvs tulkojums iespējams tikai tad, ja šajā procesā iesaistās arī profesionāls tulkotājs. Tā tika radīta jauna tulkošanas programma, ko nosauca par datorizētās tulkošanas sistēmu.

Datorizētās tulkošanas rīki

Ko nozīmē datorizētā tulkošana? Kāda ir datora loma šajā procesā, un kāda loma tulkotājam? Jau minētajā *Valodniecības pamatterminu vārdnīcā* sniegtā datorizētās tulkošanas definīcija konkrētu priekšstatu par šo jēdzienu nesniedz, proti, tur ir teikts, ka datorizētā tulkošana ir avotteksta tulkošana mērķvalodā, tulkotājam izmantojot datora programmatūras iespējas, piemēram, dators piedāvā variantus, no kuriem tulkotājs izvēlas atbilstošāko (angļu val. *computer-aided translation, computer assisted translation,* saīsināti *CAT,* krievu val. *автоматизированный перевод).* Šķiet, ka atšķirībā no angļu valodas, latviešu un krievu valodā apzīmējums minētajam tulkošanas veidam gluži precīzi neatspoguļo jēdziena būtību, kaut gan, ja lasītājam ir zināms, ka īpašības vārds „datorizēts" nozīmē „ar datora palīdzību", tad pārmetumi par vārdkoptermina „datorizētā tulkošana" precizitāti varētu būt lieki. Taču krievu valodas *автоматизированный перевод* gan var radīt pārpratumus.

Sākotnēji datorizētie tulkošanas rīki tika izstrādāti tikai vārda līmenī. 60. gadu beigās izveidoja plašas terminoloģijas datubāzes – lielākā un pazīstamākā no tām bija Eiropas Komisijas daudzvalodu terminoloģijas datubāze *Eurodicautom*, kura šobrīd ir aizstāta ar *IATE* (*Inter-Active Terminology for Europe*), Eiropas Savienības valodu terminoloģijas datubāzi (ECOT 2006).

80. gadu vidū tika attīstītas terminoloģijas pārvaldības sistēmas, kuras sniedza iespēju izmantot izveidotās terminoloģijas datubāzes tulkošanas procesā. Galvenā šo sistēmu priekšrocība ir tā, ka, savietojot terminoloģijas sistēmu ar tulkojamā teksta formātu, noklikšķinot tajā uz nepazīstamā vārda, šis termins un tā tulkojums automātiski parādās virs tulkojamā teksta. Tātad notiek automātiska terminu atlase no divvalodu terminu datubāzēm, un tekstā tiek piedāvāts iespējamais termina tulkojums. Protams, tulkojums virs teksta parādās vienīgi tad, ja tas iepriekš saglabāts izveidotajā vārdnīcā. Tātad viens no programmatūras elementiem – divvalodīga terminoloģijas vārdnīca.

Nākamais solis datorizētās tulkošanas pilnveidē un tālākattīstībā bija tādas sistēmas izveide, kurā iespējams izveidot pārtulkoto teikumu glabātuvi jeb „teikumnīcu" (neoloģisms izveidots pēc analoģijas ar vārdu „vārdnīca"), kurā tiek uzkrāti pārtulkotie

teikumi. Tādējādi tulkotājam nekad nav jātulko vienreiz jau pārtulkotais teikums vēlreiz. Kā jau minēts iepriekš, tulkojot tekstu ar attiecīgās programmatūras palīdzību, no terminoloģijas datubāzes var „izsaukt" vajadzīgo terminu. Tieši tāpat darbojas arī tulkošanas atmiņa. Tajā ir ievadīti nevis atsevišķi vārdi vai izteicieni, bet veseli teikumi, ko sauc par tulkošanas vienībām jeb segmentiem. Tulkotājs var „pieprasīt" atmiņai nevis viena termina, bet vesela teikuma tulkojumu. Dators piedāvā tulkotājam tādu teikumu, kas, piemēram, atbilst par 70 % tulkojamam teikumam, ja tāds ir tā „krājumos". Tulkotājam pēc tam jāiztulko atlikušie 30 %.

Vislielāko ieguvumu šādas datorizētās tulkošanas programmas nodrošina tiem, kuri ikdienā strādā ar standartizētiem tekstiem, kuros informācija bieži atkārtojas vai mainās tikai daļēji, piemēram, līgumi, lietošanas pamācības, gada pārskati, dažādi reklāmas katalogi, kuros informācija ik pa brīdim tikai jāatjaunina u. tml. Tātad tulkošanas programmas nodrošina iespēju atkārtoti izmantot jau reiz pārtulkotu materiālu, tā samazinot tulkojamā materiāla apjomu, taupot laiku, cilvēkresursus un līdzekļus.

Tulkošanas procesa laikā datorizētās tulkošanas atmiņas rīks *(Translation Memory)* saglabā katru pārtulkoto teikumu un frāzi savā divvalodu datubāzē. Katrā nākamajā teikuma atkārtojumā, visbiežāk teikuma fragmentā, tulkotājs tiek apgādāts ar tulkojumu, kas saglabāts atmiņā. Tad tulkotājs izvēlas pieņemt šo tulkojumu, ja tas simtprocentīgi sakrīt, vai rediģēt to tā, lai padarītu pieņemamu. Šie rīki būtiski uzlabo tulkojuma kvalitāti un teksta vienveidību. Šī tehnoloģija sniedz divas galvenās priekšrocības: 1) laika un finanšu ekonomiju, ja jātulko teksti, kas regulāri atkārtojas, 2) tiek nodrošināta tulkojumu konsistence un augstāka kvalitāte, jo vienādi formulējumi netiek tulkoti dažādi, kas lasītāju nereti var samulsināt.

Tulkošanas atmiņa „palīdz" tulkot tikai tad, ja tajā ir uzkrāts pietiekami daudz tulkošanas vienību. Tulkošanas atmiņas tāpat kā nozaru vārdnīcas veido atbilstoši nozarēm. Iedomājieties, ka jums ir sagatavota tulkošanas atmiņa, kurā ir ievadītas tūkstošiem tulkošanas vienību jeb teikumu pāru angļu un latviešu valodā no visu veidu līgumu tekstiem. Tulkojot kārtējo līgumu, tulkošanas atmiņa jums noteikti sniegs tulkošanas piedāvājumu, jo līguma teksti ir samērā standartizēti. Tas pats ir attiecināms uz jebkura cita veida standartizētu tekstu.

Datorizētā tulkošana Ventspils Augstskolā

VeA Tulkošanas studiju fakultāte, gatavojot jaunos tulkošanas speciālistus, lielu uzmanību pievērš moderno tulkošanas programmatūras izmantošanas prasmju attīstīšanai, jo tulkotāja profesija nav iespējama bez minētajām prasmēm.

Mūsu fakultātes studentu rīcībā šobrīd ir vairākas datorklases, kas ir aprīkotas ar programmas *SDL Trados 2007 Professional* tulkošanas rīkiem: tulkošanas atmiņu (*Translators Workbench – TM*), terminoloģijas datubāzi (*Multiterm*), tulkošanas rīku *Tag Editor*, ar kura palīdzību atšķirībā no tulkošanas atmiņas rīka, kas savietojams tikai ar *Microsoft Word* (*DOC, DOCX, RTF*), iespējams tulkot tekstus vēl citos formātos, piemēram, *Microsoft PowerPoint, Microsoft Excel*, *HTML* u. c. Šis rīks ir praktiski neaizstājams tādu tekstu tulkošanai, kur ir daudz attēlu, shēmu, grafiku, tātad t. s. tehnisko tekstu tulkošanai, piemēram, dažādu iekārtu tehniskās dokumentācijas, lietošanas pamācību tulkošanai.

Tulkošanas specialitātes studenti galvenokārt maģistra studiju līmenī apgūst arī jaunāku *TRADOS* versiju, proti, *SDL TRADOS STUDIO 2009*, kura pēdējos gados aizvien aktīvāk ienāk tulkošanas tirgū, jo patiesībā ir vienkāršāka un ērtāka par iepriekšējo. Viens no tulkošanas programmatūras rīkiem, ar kuru studējošie mācās strādāt, ir t. s. pārtulkoto tekstu sastatītājs *WignAlign*. Ar šā rīka palīdzību iespējams izveidot datorizēto atmiņu no jau esošajiem tulkojumiem. Taču, veidojot atmiņu no tulkojumiem, jo īpaši no tādiem, kurus tulkotājs nav veicis pats, ir jāņem vērā, ka tajos ir iespējamas kļūdas. Ja tulkojums nav veikts teikumu pa teikumam, proti, ja mērķvalodas izteiksmes līdzekļu īpatnību dēļ no diviem avotvalodas teikumiem tulkojumā ir veidots viens teikums, tad tulkotājam attiecīgajā darba vidē ir jāveic daudz labojumu, jo šis rīks „sasien" tulkošanas vienības jeb pārtulkoto teikumu pārus (par teikuma beigu pazīmi tiek uzskatīts punkts un tam sekojoša tukšumzīme), pēc tam šīs vienības ar speciālas funkcijas palīdzību tiek eksportētas uz tulkošanas atmiņu. Šā iemesla dēļ praksē *WignAlign* rīka izveidotās atmiņas pārāk bieži izmantotas netiek, taču atsevišķu teksta veidu tulkošanā tā izmantošana var ievērojami optimizēt tulkotāja darbu. Šajā gadījumā ir runa par Eiropas Savienības (ES) institūciju dokumentiem. Tie ir pieejami ES tīmekļa vietnē *Eur-Lex* visās ES valodās. Starp citu – ļoti noderīgs tulkotājam ir *Eur-Lex* datubāzē pieejamais teksta atveidojums divās valodās, ar kura palīdzību tulkotājs var atrast vissarežģītākos tulkošanas problēmu risinājumus. Izmantojot valodu izvēlni, iespējams izvēlētā ES teksta atveidojums jebkurā ES valodu kombinācijā. Taču jāņem vērā, ka tie ir ES dokumenti un atbilstoši tajos aplūkota ES tematika, kas atspoguļo šīs starptautiskās organizācijas tiesības, pienākumus, funkcijas u. tml.

Docējot datorizētās tulkošanas kursus, studējošajiem tiek uzsvērti arī atsevišķi šīs programmatūras nepilnības. Pats galvenais no tiem – programmatūra ir orientēta nevis uz teksta, bet uz atsevišķu teikumu tulkošanu. Tādējādi programmatūra izmantojama vairāk lietišķajos tekstos, kur pārāk brīva teksta interpretācija nemaz nav pieļaujama. Atsevišķos gadījumos, ņemot vērā konteksta īpatnības, atmiņā pārtulkotais teikums no viena konteksta var izrādīties nepiemērots kādā

citā kontekstā. Tādēļ ļoti svarīgi topošajam tulkotājam ir iemācīties kritisku attieksmi pret jebkuru tekstu.

Datorizētās tulkošanas kursos maģistrantūras studiju programmā apgūtās zināšanas un prasmes tiek pārbaudītas rakstiski. Mūsu augstskolai, sadarbojoties ar uzņēmumu *SDL* augstskolu sertifikācijas programmā, ir iespēja kārtot testu internetā un iegūt *TRADOS* tulkošanas tehnoloģijas lietotāja sertifikātu.

Jāuzsver, ka mūsdienu tulkošanas process ir darbs komandā. Saņemot tulkošanas pasūtījumu, vispirms tiek veikta tā sauktā priekštulkošana (angļu val. *pre-editing*), kurā izmanto tulkošanas atmiņas rīka analīzes funkciju, lai konstatētu, cik tulkošanas atmiņā ir tādu vienību, kas pilnībā sakrīt ar tulkojamā teksta vienībām (teikumiem). Ja šādas vienības atmiņā ir, programmatūrā ir funkcija, ar kuru iespējams automātiski pārtulkot tos teikumus, kuri jau ir tulkotāja datorizētajā atmiņā. Lielos tulkošanas birojos parasti tiek nodarbināts speciālists, kas ir atbildīgs par tekstu sagatavošanu. Šo teksta iepriekšējo analīzi un atbilstošo tulkošanas vienību automātisku tulkošanu sauc par priekštulkošanu. Tādējādi tulkotājs parasti saņem jau daļēji pārtulkotu tekstu, turklāt šādā tulkošanas projektā parasti strādā arī terminologs, kas ir atbildīgs par vienveidīgas terminoloģijas lietojumu pasūtītajā tulkojumā, kā arī teksta redaktors, kaut gan jāuzsver, ka terminologa un redaktora funkcijas nereti veic pats tulkotājs. Par visu tulkojumu un tā kvalitāti ir atbildīgs tulkošanas projekta vadītājs.

Topošajam tulkotājam ir jāapgūst visi šie projekta aspekti. Šīs prasmes tiek apgūtas mūsu maģistra studiju programmas kursā „Tulkošanas darba organizācija un vadība". Atbildīgā pasniedzēja vadībā tiek simulēti apjomīgi tulkošanas projekti, kurā katram studējošajam ir noteikts uzdevums. Katrā projektā neatņemama sastāvdaļa ir attiecīgo tulkošanas rīku izmantošana.

Bez mašīntulkošanas un datorizētās tulkošanas programmām topošie tulkotāji tiek iepazīstināti arī ar citām specializētām tulkošanas programmām, kas paātrina tulkošanu un nodrošina nepieciešamo kvalitāti. Liela nozīme darba atvieglošanai tulkošanas birojā, kurā tiek īstenoti daudzi un apjomīgi tulkošanas projekti, ir programmatūrai, kas nodrošina tulkošanas projektu vadību, veicot automatizētu piedāvājumu un rēķinu sagatavošanu, tādējādi veicinot ciešāku, ātrāku un produktīvāku saikni starp tulkojumu pasūtītājiem un tulkošanas projektu vadītājiem. Šādas programmatūras atvieglo arī projektu peļņas aprēķinu un uzskaiti gan projektu vadītājiem, gan pasūtītājiem, kā arī tulkotāja darba uzskaiti. Projektu vadības sistēmas ir izstrādātas, lai uzlabotu un paātrinātu projektu vadītāju darbu. To iespējas ļauj tulkošanas projektu vadītājam efektīvāk plānot darbu, sekot līdzi projekta izpildei un sazināties ar izpildītājiem un pasūtītājiem, kā arī nodrošināt viņa prasību izpildi.

Sagatavojot tulkošanas speciālistus, praktiskajās nodarbībās vienmēr tiek uzsvērts, ka datorizētie tulkošanas rīki var palīdzēt uzlabot tulkošanas ātrumu, tomēr interneta vārdnīcas, tulkošanas programmas vai terminoloģijas datubāzes nepalīdz, kad tulkojot rodas problēma teksta satura izpratnē, proti, jārod tulkojums kādai īpaši sarežģīti formulētai teksta daļai avotvalodā. Tādā gadījumā datorizētā tulkošanas atmiņa no saviem krājumiem, visticamāk, neko nepiedāvās. Ņemot vērā, ka teksta formulējumu iespējas praktiski ir bezgalīgas, ir skaidrs, ka tulkotājam tulkošanas procesā vienmēr būs jāsniedz savi risinājumi. Nereti specifiskos nozaru valodas tekstos sastopami arī termini, kuru atbilsmes mērķvalodā nav atrodamas, – arī šādā situācijā tulkotājam jāpārslēdzas no rutīnas darba uz problēmas risināšanu. Tādējādi tulkotāja prasme risināt problēmas ir tikpat nozīmīga, cik prasme izmantot tulkošanas tehnoloģijas.

Literatūra

1. Cocci Lucia. CAT Tools for Beginners. 2009. Elektroniskais resurss: *http://translationjournal.net/journal/50caten.htm*
2. ECOT – European Commission Online Translation tool for European Languages, 2006. Pieejams: *http://www.eugris.info/displayresource.asp?ResourceID=5703&*
3. Hutchins, John. The first public demonstration of machine translation: the Georgetown – IBM System, 7th January 1954, 2004. Pieejams: *http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf*
4. Snell-Hornby, Mary, Hönig, Hans G., Kußmaul, Paul. *Handbuch Translation*. Tübingen, 2006.
5. META-FORUM tīmekļa vietne: *http://www.meta-net.eu/events/meta-forum-2011/*
6. VPSV – Valodniecības pamatterminu skaidrojošā vārdnīca. Atb. red. V. Skujiņa. Rīga : LU Latviešu valodas institūts, Valsts valodas aģentūra, 2007.

Normatīvie akti

MK noteikumi Nr. 994 „Kārtība, kādā augstskolas un koledžas tiek finansētas no valsts budžeta līdzekļiem". 12.12.2006.

Jānis Sīlis

Latvijas nacionālo reāliju standartizēta tulkojuma tīmekļa vietnes izveides lingvistiskās semantikas un datortehnoloģiskās problēmas

Projekta „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās" izveides nepieciešamība, mērķi un uzdevumi

Kopš 2004. gada 1. maija, kad Latvija kļuva par Eiropas Savienības sastāvdaļu, īpaši aktuāls ir jautājums par saziņas kvalitāti ar citām ES dalībvalstīm. Piedaloties valsts pētījumu programmas „Letonika: pētījumi par vēsturi, valodu un kultūru" projektā „Valoda un vide" (projekta vadītāja *Dr. habil. philol.* Brigita Bušmane, LU Latviešu valodas institūts), Ventspils Augstskolas (VeA) Tulkošanas studiju fakultātes mācībspēku (vienlaikus arī Ventspils Augstskolas Lietišķās valodniecības centra (LVC) pētnieku un vadošo pētnieku) grupa – profesori *Dr. philol.* Jānis Sīlis un *Dr. philol.* Juris Baldunčiks, asociētās profesores *Dr. Philol.* Maija Baltiņa un *Dr. philol.* Valda Rudziša, docente *Dr. philol.* Astra Skrābane, lektore un VeA doktorante Diāna Pavlovska, kā arī VeA LVC pētniece *Dr. philol.* Tatjana Stoikova – sadarbībā ar Latvijas Universitātes profesoru, Valsts valodas centra (toreiz – Tulkošanas un terminoloģijas centra) direktoru Māri Baltiņu no 2006. gada pavasara prof. J. Sīļa vadībā sāka strādāt pie Latvijas nacionālo reāliju (valsts pārvaldes un pašvaldību institūciju struktūrvienību, dienestu, amatu un dienesta pakāpju, reliģisko, politisko, finanšu, izglītības, kultūras, sporta u. c. institūciju, organizāciju un objektu, administratīvi teritoriālo vienību, naudas, valsts svētku, svinamo dienu, etnogrāfijas un folkloras, vēstures, ģeogrāfisko nosaukumu) standartizēta tulkojuma nodrošināšanas nozīmīgākajās Eiropas valodās.

Projekta „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās" mājaslapā *http://realijas.venta.lv* izvietotais apmēram 1400 Latvijas nacionālo reāliju un to atbilsmju korpuss svešvalodās visplašākajam lietotāju lokam ir pieejams kopš 2009. gada nogales. Mājaslapas datubāzi kā uzziņu un rekomendāciju avotu jau divus gadus var izmantot visi Latvijas un arī ārvalstu interesenti.

Pēc Latvijas un vairāku citu Eiropas valstu uzņemšanas Eiropas Savienībā īpaši aktuāls kļuvis jautājums par to, kādā valodā un cik precīzā izteiksmes formā mūsu valsts sazināsies ar citām ES dalībvalstīm (sk. arī Sīlis 2004, 249–250). Lai arī ES dalībvalstu iedzīvotājiem ir tiesības vērsties ES institūcijās savā dzimtajā valodā, kas

vienlaikus ir viena no ES oficiālajām valodām, tomēr terminoloģijas jomā valodu lietojuma precizitāte skar jebkuru no mums. Dažādu terminoloģijas interpretācijas variantu problēma jebkura lietišķā teksta tulkojumā visbiežāk rodas terminoloģijas nepietiekamas standartizācijas dēļ. Vairāki sociolingvistiski faktori ir viens no iemesliem, kas latviešu valodā izraisījis šādu standartizācijas trūkumu (Sīlis 1999, 62–63). Viens no galvenajiem šīs negatīvās parādības cēloņiem laikā no 1945. gada līdz Latvijas valstiskās neatkarības atjaunošanai bija motivācijas trūkums attīstīt latviešu terminoloģiju un specifisko frazeoloģiju daudzās teorijas un praktiskās darbības jomās, jo mutiskajā un rakstiskajā saziņā (dokumentu apritē) dominēja krievu valoda – šāds stāvoklis, piemēram, bija vērojams tautsaimniecībā un PSRS sociālismam raksturīgās uzņēmējdarbības vadībā.

Pēc pirmo atgūtās neatkarības gadu politiskajām un ekonomiskajām pārmaiņām parādījās vēlme un nepieciešamība straujāk attīstīt latviešu valodas terminoloģiju un frazeoloģiju. Pārmaiņu temps šajā valodas jomā bija tik dinamisks, ka terminu un specializētās frazeoloģijas veidošanu laika trūkuma dēļ nereti nācās veikt nespeciālistiem. Tādējādi bieži vien radās un paralēli tika lietoti sinonīmi ar atšķirīgu kvalitāti.

Iekļaujoties ES kopējā administratīvajā struktūrā valsts, pašvaldības un nevalstisko organizāciju līmenī, radusies akūta nepieciešamība pēc Latvijas organizāciju un tajās strādājošo cilvēku amatu nosaukumu standartizētas interpretācijas lielākajās Eiropas valodās. Galvenā problēma, ar ko šajā jomā sastopas tulki un tulkotāji, ir dažādu administratīvo kultūru nesaderība: Rietumeiropas valstis jeb vecās ES dalībvalstis gadsimtu gaitā izstrādājušas savus administrēšanas principus ar tiem atbilstošu terminoloģiju, kamēr Latvija mantojumā saņēmusi padomju stila administratīvo sistēmu ar attiecīgo terminu kopumu.

Institūciju un administratīvo amatu nosaukumus nevar uzskatīt par terminiem šī jēdziena klasiskajā izpratnē. Administratīvie termini jebkurā valodā – un latviešu valoda šajā ziņā nav izņēmums – ir vai nu kultūrspecifiski vārdi (*Saeima, Satversme, pagasts, tautas nams*), vai arī aprakstošas struktūras, kas arī bieži ir kultūrspecifiskas. Latviešu valodā pēdējās bieži vien atspoguļo padomju laika administratīvās tradīcijas, kad latviskie institūciju nosaukumi dažkārt bija gandrīz burtisks tulkojums no krievu valodas. Šādi postpadomju „transplanti" neiedzīvojas Rietumu tradīcijās (Sīlis 2009, 96–104).

Projekta „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās" mērķa – standartizēta tulkojuma nodrošināšanas nozīmīgākajās Eiropas valodās – īstenošanai tika formulēti šādi uzdevumi:

- Latvijas nacionālo reāliju kopuma apzināšana, standartizācija un kataloģizācija latviešu valodā;

- standartizētā un kataloģizētā Latvijas nacionālo reāliju kopuma atbilsmju atrašana angļu, vācu, franču un krievu valodā (izmantojot arī jau pastāvošos labākos tulkojuma variantus);
- Latvijas nacionālo reāliju alfabētiskā un tematiskā sakārtojuma, kā arī attiecīgo svešvalodu atbilsmju ievietošana projekta mājaslapā.

Latvijas reāliju apzināšanu, standartizāciju un kataloģizāciju latviešu valodā projekta sākumposmā (2006.–2007. gadā) veica Jānis Sīlis, Juris Baldunčiks, Maija Baltiņa, Māris Baltiņš, Diāna Pavlovska, Valda Rudziša un Astra Skrābane. Katra projekta dalībnieka pārziņā bija darbs ar konkrētas jomas reālijām: Valda Rudziša apzināja, standartizēja un kataloģizēja valsts lēmējvaras, izpildvaras un tiesu varas institūciju nosaukumus, svinamo un atceres dienu nosaukumus, Juris Baldunčiks – amatu nosaukumus, komercuzņēmumu, administratīvi teritoriālo vienību un ģeogrāfisko reāliju nosaukumus, Astra Skrābane – ģimenes dzīves un tradīciju, kā arī naudas vienību nosaukumus, Diāna Pavlovska – pašvaldību lēmējvaras un izpildvaras institūciju nosaukumus, Maija Baltiņa – vēsturiskās un folkloras reālijas, Māris Baltiņš – medicīnas iestāžu un to darbinieku amatu nosaukumus, Jānis Sīlis – nevalstisko organizāciju nosaukumus.

2008. un 2009. gadā darbu ar Latvijas reāliju angļu valodas atbilsmēm veica Jānis Sīlis un Juris Baldunčiks, darbu ar vācu atbilsmēm – Valda Rudziša, darbu ar franču atbilsmēm – Astra Skrābane, darbu ar krievu atbilsmēm – Diāna Pavlovska un Tatjana Stoikova.

Projekta mājaslapas izstrāde, dizains un mājaslapas uzturēšanas nodrošināšana bija SIA „Molips" valdes priekšsēdētāja Aleksa Fišmeistera un viņa kolēģu pārziņā. Kopš 2008. gada rudens līdz projekta darbības beigām Ventspils Augstskolas Tulkošanas studiju fakultātes (TSF) sekretāre, VeA maģistrante Dace Rozenberga bija projekta administratīvā un tehniskā asistente.

Pētījuma galvenais uzdevums bija pabeigt standartizēto un kataloģizēto reāliju klasifikācijas optimizēšanu, reāliju precizēšanu latviešu valodā, pievienojot atbilsmes angļu, vācu, franču un krievu valodā, kā arī ievietot visu šo materiālu datubāzē projekta mājaslapā.

Latvijas reāliju svešvalodu atbilsmju izvēlē tika izmantoti principi, kas aizgūti no citu ES dalībvalstu pieredzes.

- Svešvalodas atbilsmes variantā nav pieļaujamas tādu valstu (Lielbritānijas, ASV, Vācijas, Austrijas, Šveices, Francijas, Beļģijas u. c.) kultūras konotācijas, kurās šī svešvaloda ir dzimtā valoda.
- Nosaukumā svešvalodas variantā nav pieļaujama lasītāja neprecīza informēšana par attiecīgās institūcijas vai amata funkcijām vai statusu.

- Nosaukums svešvalodā nedrīkst skanēt absurdi, un tam jāatbilst konkrētās svešvalodas gramatiskajām u. c. normām.
- Svešvalodas variantā, kas atbilst konkrētam latviešu terminam vai nosaukumam, jāievēro konsekvence, pretējā gadījumā rodas nosaukumu paralēlvarianti.
- Nākotnē būtu jāveic pasākumi, kuri noteiktu, ka iestādēm nav tiesību pašām piešķirt sev nosaukumus angļu valodā, nesaskaņojot tos ar institūcijām, kas atbildīgas par standartizāciju šajā jomā.

Projekta „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās" lingvistiskās semantikas un datortehnoloģiskās problēmas, to risinājumi

Reāliju atlases sākumposmā projekta dalībnieku domas par to, kādas reālijas un kāds to apjoms būtu iekļaujams reāliju korpusā, visai būtiski atšķīrās. Izvēle svārstījās no vairākiem desmitiem līdz vairākiem tūkstošiem vienību. Bija jāatrod kāds kvantitatīva rakstura kompromiss, jo korpuss ar dažiem desmitiem vienību, nebūdams reprezentatīvs, zaudētu jēgu, bet 7000 līdz 10 000 vienību liels korpuss atvēlētajā laikā neļautu sasniegt projekta mērķi un mājaslapas lietotājs rezultātā saņemtu nepārskatāmu informācijas „aisbergu".

Korpusa lieluma noteikšanai bija jādefinē arī jēdziens „reālija", lai šī definīcija būtu piemērota tieši konkrētā projekta vajadzībām. Ielūkojoties termina „reālija" lietojumā, tika konstatēts, ka tas satopams tekstos ar visdažādāko tematiku:

„Daudzi vārdi valodas apritē ir tikai īsu brīdi, jo *reālijas*, ko tie apzīmē, ātri kļuvušas neaktuālas..." (Ernstsone 2004)

„avis, kas pazīst Kristus balsi, ir tikpat bibliska *reālija* kā Kristus miesa." (*www.lelb.lv/forums*)

Projekta dalībnieki kritiski izanalizēja šādus jēdziena „reālija" skaidrojumus:

- „Īstenības priekšmets, parādība, dzīva būtne, abstrakts jēdziens, ko nosauc valodas vienība." (VPSV 2007, 321);
- „Reāli pastāvoša lieta, objekts." (Baldunčiks, Pokrotniece 1999, 658);
- „Tie ir vārdi vai izteicieni, kas ikdienas valodā apzīmē priekšmetus, jēdzienus, konkrētai ģeogrāfiskai vietai tipiskas parādības, dažādu tautu, nāciju, valstu un cilšu materiālās dzīves vai sociāli vēsturiskās īpatnības, kam piemīt nacionāla, vietēja vai vēsturiska nokrāsa. Šiem vārdiem nav precīza analoga citās valodās." (*KudoZ* tīkls, *ProZ.com*).

Kas tad ir reālija? Vai tikai īstenības priekšmets, parādība, dzīva būtne, abstrakts jēdziens? Vai tikai vārds, kas nosauc šo īstenības priekšmetu, parādību, dzīvu būtni, abstraktu jēdzienu? Vai tomēr abi aspekti kopā?

Ilustrācijai tiek piedāvāta no konkrēta jēdziena izveidoto terminu „rīkstnieks" un „rīkstniecība" atveide latviešu, angļu un vācu valodā:

- *rīkstnieks = ūdens āderu noteicējs (darba instruments – ne vien rīkste, bet jebkurš āderu atrašanai noderīgas priekšmets)*;
- *water diviner* (burtiskā tulk. *ūdents pareģis), dowser; water divining, dowsing is a practice that attempts to locate hidden water wells*;
- *Wünshelrute = rīkstniecība,* burtiskā tulk. *brīnumrīkste* (pasakās).

Reālija mūsu projektā: īstenības priekšmets, parādība, dzīva būtne, abstrakts jēdziens, kas katras konkrētas valodas gadījumā ir nesaraujami saistīts (veido vienu veselumu) ar šo reāliju nosaucošo valodas vienību (vārdu, vārdkopu u. c.). Tieši šīs saites unikalitāte katrā konkrētā valodā ir tā, kas padara reāliju par nacionālu parādību.

Darba gaitā viena no laikietilpīgākajām problēmām izrādījās reāliju klasifikācijas sistēmas izveide. Šādu sistēmu bija vēlams radīt vēl pirms konkrēto reāliju apzināšanas, lai šis process noritētu pēc iespējas efektīvāk un organizētāk – tātad arī ātrāk. V. Rudziša (ne vien filoloģijas doktore, bet arī diplomēta juriste), sākot apzināt valsts institūciju nosaukumus, nonāca pie atziņas, ka loģisks klasifikācijas princips varētu būt Valsts pārvaldes iekārtas likumā izmantotais iestāžu un to darbinieku amatu klasifikācijas princips. Sākotnēji šo klasifikācijas principu projekta dalībnieki mēģināja piemērot visām apzināmo reāliju grupām. Tomēr šis princips pilnīgi neder ģimenes dzīves, vēstures, folkloras reāliju un ģeogrāfisko veidojumu klasifikācijai, kā arī ir tikai daļēji piemērojams nevalstisko organizāciju un komercstruktūru klasifikācijai. Šāds klasifikācijas princips būtu grūti uztverams plašākam reāliju datubāzes lietotāju lokam, kas nav juristi vai valsts pārvaldes speciālisti (Sīlis 2009, 76–103).

Tika meklēta cita pieeja klasifikācijai, un pēc A. Skrābanes ieteikuma tika pieņemts tēzaura princips (tikai šim projektam piemērojams darba termins). Klasifikācijas centrā implicīti ir kategorija „cilvēks" (eksplicīti šī kategorija neparādās, bet tiek izmantota kā atskaites punkts). Tika izveidoti četri kategoriju līmeņi: virskategorijas „Sabiedrība", „Sadzīve" un „Teritorija", katra no šīm virskategorijām ietvēra trīs zemākus kategoriju līmeņus. Piemēram, virskategorijas „Sabiedrība" 1. līmenis ir kategorijas „Institūciju nosaukumi" un „Amatu nosaukumi"; kategorijā „Institūciju nosaukumi" ietilpst pakārtotas kategorijas „Valsts institūcijas" un „Pašvaldību institūcijas", „Nevalstiskās organizācijas" un „Uzņēmējsabiedrības"; kategorijā „Valsts institūcijas" ietilpst apakškategorijas „Lēmējvaras institūcijas",

„Izpildvaras institūcijas" un „Tiesu varas institūcijas" utt. Kategorijā „Nevalstiskās organizācijas" netiek izmantotas tematiskās apakškategorijas, bet gan septiņas alfabētiski sakārtotas NVO tipu (*aģentūra, apvienība, asociācija, centrs, dienests, draudze, fonds, klubs, partija, sabiedrība, serviss* utt.) grupas.

Problēma bija vienoties arī par klasifikācijas temporālo aspektu (Sīlis 2009, 76–103). Tika nolemts, ka reālijas tiks apzinātas sinhroniskā griezumā, jo to svešvalodu atbilsmes jāizmanto atbilstoši mūsdienu situācijas prasībām, bet pagātnē lietotiem institūciju, amatu u. c. nosaukumiem varētu pievērsties kādā citā projektā. Diahroniskā pieeja izmantota vienīgi folkloras reāliju jomā, kur citu principu izmantot nav iespējams.

Dažās reāliju kategorijās apzināto vienību skaits sasniedza vairāk nekā tūkstoti. Skaita samazināšanai risinājums tika atrasts, atsakoties no īpašvārdiem nosaukumos: [Rīgas] Amatniecības vidusskola, [Rīgas] Būvamatniecības vidusskola, [Rīgas] Mākslas un dizaina vidusskola, [Rīgas Pārdaugavas] profesionālā vidusskola u. c.

Tālāk parādīts reāliju klasifikācijas procesa rezultāts:

Klasifikācijas kategoriālo līmeņu hierarhiska shēma

1. līmenis
Sabiedrība – Sadzīve – Teritorija

2. līmenis (1. līmeņa kategorijas ir kvadrātiekavās)
[Sabiedrība]
↓
Institūciju nosaukumi – Amatu nosaukumi
[Sadzīve]
↓
Darbs – Tradīcijas – Ģimenes dzīves notikumi
[Teritorija]
↓
Administrat. teritoriālā dalījuma nosaukumi – Ģeogrāf. veidojumu nosauk.

3. līmenis (abu augstāko līmeņa kategorijas ir kvadrātiekavās)
[Sabiedrība: Institūciju nosaukumi]
↓
Valsts instit. – Pašvaldību instit. – NVO – Komercsabiedrības
[Sabiedrība: Amatu nosaukumi]
↓
Amati – Pakāpes un nosaukumi
[Sadzīve: Darbs]
↓

3. un 4. līmeņa kategoriju nav
[Sadzīve: Tradīcijas]
↓
Svētku un atceres dienas
[Sadzīve: Ģimenes dzīves notikumi]
↓
3. līmeņa kategoriju nav
[Teritorija: Administratīvi teritoriālā dalījuma nosaukumi]
↓
3. līmeņa kategoriju nav
[Teritorija: Ģeogrāfisko veidojumu nosaukumi]
↓
3. līmeņa kategoriju nav

4. līmenis (visu augstāko līmeņa kategorijas ir kvadrātiekavās)

katrā šā līmeņa kategorijā ir konkrētu reāliju/reāliju grupas nosaukumi

[Sabiedrība: Institūciju nosaukumi: Valsts institūcijas]
↓
Likumdošanas varas inst. – Izpildvaras inst. – Tiesu varas inst.
[Sabiedrība: Institūciju nosaukumi: Pašvaldības institūcijas)
↓
Lēmējstruktūras – Izpildvaras institūcijas
[Sabiedrība: NVO]
↓
konkrētu NVO nosaukumi
[Sabiedrība: Komercsabiedrības]
↓
Valsts komercsab. – Pašvaldību komercsab. – Privātās komercsab.
[Sadzīve: Tradīcijas]
↓
Svētku dienas – Atceres un piemiņas dienas – Folkl. un reliģ. svētki
[Sadzīve: Ģimenes dzīves notikumi]
↓
konkrētu reāliju nosaukumi
[Teritorija: Administr. terit. dalījuma un ģeogr. veidojumu nosaukumi]
↓
Republikas pilsēta – Pilsēta ar lauku teritoriju – Novads, pagasts, rajons
Ģeogrāfiskie veidojumi
↓
konkrētu reāliju nosaukumi

Projekta „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās" mērķis ir izskaust lingvistiski un kultūrspecifiski kļūdainus variantus, rekomendējot vienu nosaukumu, ko varētu uzskatīt par pareizu. Tas gan ne vienmēr ir iespējams, jo vairākām iestādēm, organizācijām un uzņēmumiem to dibināšanas dokumentos minēti arī šo struktūru nosaukumi vairākās svešvalodās; šie svešvalodu varianti tiek aktīvi lietoti lietišķajā sarakstē, ja tā notiek attiecīgajā svešvalodā. Ar līdzīgu problēmu jāsastopas arī, piemēram, Latvijas augstākās izglītības iestāžu nosaukumos. Problēmas parasti rodas gadījumos, ja augstākā mācību iestāde nav universitāte, akadēmija vai institūts, bet gan augstskola. Vārds „augstskola" šāda tipa valsts un privātajās institūcijās angļu nosaukumos kļūst par *Institution of Higher Education, Higher School, School, University College*. Visi šie nosaukumi formāli nav nepareizi, bet ieteikt vienu variantu visām augstskolām nav iespējams, jo esošie nosaukumi apstiprināti šo augstskolu satversmēs un tādējādi ieguvuši likumīgu spēku (Sīlis 2009, 76–103).

Projekta gaitā bija radušies arī tīri tehniska rakstura sarežģījumi, kas saistīti ar apstākli, ka ne visi iesaistītie pieredzējušie valodniecības un tulkojumzinātnes speciālisti bija spējīgi efektīvi strādāt elektroniskajā vidē, ievadot projekta mājaslapas datubāzē latviešu reālijas un to atbilsmes svešvalodās.

Šīs un citu datortehnoloģiska rakstura problēmu risināšanā nenovērtējamu palīdzību projekta valodniekiem sniedza projekta partneris SIA „Molips" (*http://www.vatp.lv/molips-sia*), kas nodarbojas ar jaunu, inovatīvu un klientam pievilcīgu risinājumu izstrādi IT jomas attīstībai – no mājaslapu izveides līdz informāciju sistēmu projektēšanai un ieviešanai. SIA „Molips" specializācija ir interneta mājaslapu izstrāde, balstoties uz *PHP, HTML, Java Script, CSS, AJAX, Actionscript/Flash* un *MySQL* tehnoloģijām.

SIA „Molips" dalība „Letonikas" programmas projektā „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās":

- 2007. gads – izveidots un izmēģināts projekta mājaslapas pirmvariants;
- 2008. gads – pilnveidota un administrēta projekta mājaslapa (*http://realijas.venta.lv*), izveidots administrēšanas darba panelis (valodniekiem – projekta dalībniekiem) un uzturēta vietne;
- 2009. gads – turpinās projekta mājaslapas un tajā ievietoto reāliju datubāzes uzturēšana;
- 2010. gads – mājaslapā tiek veikti pēdējie uzlabojumi, un tā tiek padarīta pieejama lietotājiem.

Projekta valodnieku un datorspeciālistu grupu sadarbības modelis darba gaitā bija šāds:

- projekta valodnieki savstarpējās diskusijās formulē savus darba uzdevumus (arī reāliju klasifikācijas pamatprincipus);
- valodnieki kopējās darba sēdēs ar „Molipa" datortehnologiem formulē darba uzdevumus „Molipam";
- „Molips" savstarpējās diskusijās pārformulē valodnieku vēlmes sev saprotamos darba uzdevumu formulējumos.

Tika konstatēts, ka svarīgi ievērot tieši šādu secību: valodnieku vajadzības → datortehnologu risinājums (un ne otrādi).

SIA „Molips" problēmuzdevumi projektā „Latvijas nacionālo reāliju standartizēts tulkojums Eiropas valodās":

- izveidot valodniekiem saprotamu, ērti lietojamu un „muļķudrošu" (*foolproof*) administrēšanas darba paneli – projekta nobeiguma posmā šajā vidē strādāja ne vien profesionāli valodnieki/tulkojumzinātnieki, bet arī VeA TSF bakalaura programmu studenti (zinātniskā darba prakse 7. semestrī);
- izprast valodnieku/tulkojumzinātnieku ideju par četriem reāliju klasifikācijas līmeņiem, lai šo principu varētu atspoguļot mājaslapā (savdabīgs reāliju sistemātiskais katalogs).

Sākotnēji projekta uzdevumi ietvēra arī priekšlikumu formulēšanu par īpašu normatīvo aktu izstrādi, kas fiksētu standartizētos Latvijas nacionālo reāliju (administratīvās terminoloģijas, amatu un iestāžu nosaukumu) tulkojumus svešvalodās, ņemot vērā citu Eiropas Savienības dalībvalstu pieredzi šajā jomā. Šie priekšlikumi tiktu izstrādāti sadarbībā ar juristiem un citiem speciālistiem. Diemžēl darbs pie latviešu nacionālo reāliju apzināšanas, kataloģizācijas un standartizācijas, kā arī piemērotāko angļu, vācu, franču un krievu valodas atbilsmju atrašana prasīja daudz vairāk laika, nekā iepriekš bija paredzēts, tāpēc no šā uzdevuma projekta dalībnieki bija spiesti atteikties.

Jau tika minēts, ka visplašākie Latvijas sabiedrības slāņi un arī ārvalstu interesenti mājaslapas datubāzi var izmantot kā uzziņu un rekomendāciju avotu. Reāliju meklēšana pēc alfabētiskā un tematiskā principa ir iespējama jebkurā no piecām mājaslapā atrodamajām valodām, tāpēc ārvalstu lietotāji, ievadot angļu, vācu, franču vai krievu nosaukumu kā meklēšanas sākumpunktu, var iegūt atbilsmes arī pārējās četrās valodās. Projekta pētniecības grupa cer, ka reāliju paveidu dažādība padarījusi datubāzi noderīgu gan jebkuras valsts pārvaldes, tautsaimniecības, izglītības, kultūras utt. nozares iestādēm, gan arī jebkura sociālā slāņa, vecuma grupas un dzimuma privātpersonai.

No valodnieku viedokļa bija svarīgi, ka datorspeciālisti spēja izprast lingvistisko vajadzību specifiku, iedzīvinot reāliju klasifikācijas principus un līmeņus konkrētas programmatūras formā, izstrādāt valodniekiem ērtu un drošu darba paneli, kā arī izveidot lietotājiem parocīgu un viegli izprotamu konkrētas reālijas meklēšanas sistēmu. Projekta īstenošana un iegūtie rezultāti uzskatāmi demonstrē, ka valodnieku un datortehnologu sadarbība ir ne tikai interesanta pašiem sadarbības partneriem, bet arī neapšaubāmi auglīga Latvijas datorlingvistikas tālākā attīstībā.

Literatūra

1. Baldunčiks, J., Pokrotniece, K. *Svešvārdu vārdnīca.* Rīga : Jumava, 1999, 880 lpp.
2. Ernstsone, V. *Deviņdesmito gadu sarunvaloda Austrumu un Rietumu ietekmju krustpunktā.* [Skatīts 2011. gada 8. septembrī.] Pieejams: *http://www.liis.lv/latval/stilistika/ernstsone.htm.*
3. Diskusija par kristoloģiju Emanuela Svēdenborga sakarā. *LELB*, 2006. [Skatīts 2011. gada 29. augustā.] Pieejams: *http://www.lelb.lv/forums/?fu=searchfor&srch=kristolo%C4%A3ija*.
4. KudoZ tīkls *ProZ.com The translation workplace*. [Skatīts 2011. gada 12. septembrī.] Pieejams: *http://lav.proz.com/*.
5. Sīlis, J. Sociolinguistic Aspects of Translation and Interpreting. In: *Interpreting and Translation as Intercultural Communication: Theory, Practice, Instruction Methods.* Ventspils, 1999, 57.– 66. lpp.
6. Sīlis, J. Interpretation Problems in Translations of Names of Institutions and Administrative Posts. In: *Information Society and Modern Business*. Rīga : Jumava, 2004, 249.–254. lpp.
7. Sīlis, J. *Tulkojumzinātnes jautājumi. Teorija un prakse*. Ventspils Augstskola, 2009, 264 lpp.
8. Sīlis, J. Valoda un vide. No: *Letonikas 3. kongresa zinātniskie raksti*. Rīga : Latvijas Zinātņu akadēmija, 2009, 76.–103. lpp.
9. VPSV – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Atb. red. V. Skujiņa. Rīga : LU Latviešu valodas institūts, Valsts valodas aģentūra, 2007, 623 lpp.

Anita Helviga

Ieskats datorlingvistikas terminoloģijas iezīmēs un attīstības tendencēs

Ievads

Terminoloģijas izstrāde ikvienā zinātnes nozarē ir kopīgs darbs gan attiecīgās nozares speciālistiem, gan valodniekiem. Terminrades procesam ir vairākas daļas: sabiedrības terminoloģisko vajadzību izzināšana, terminu izstrāde, terminu saskaņošana, terminu apstiprināšana, terminu publiskošana, terminu grozīšana, terminu definīciju izstrāde (Baltiņš 2007, 401–402). Šajā rakstā netiks nošķirti atsevišķi terminrades procesa soļi, bet sniegts ieskats terminoloģijas iezīmēs kopumā. Iespējams, ne visi publikācijā aplūkotie vārdi un jēdzieni ir apstiprināti termini, bet tie ir lietoti datorlingvistikas aprakstīšanai, tāpēc šajā pārskatā tiks dēvēti par datorlingvistikas terminoloģiju, kaut arī vairāki no tiem ir attiecināmi uz datorzinātni kopumā. Zināmu ieguldījumu informācijas un komunikācijas tehnoloģijas (IKT) terminoloģijas izstrādes pieredzes apkopojumā devusi Ilze Irēna Ilziņa (Ilziņa 2010).

Datorlingvistika kā salīdzinoši jauna zinātnes nozare attīstās ļoti strauji, nepieciešamība pēc jauniem jēdzieniem un terminiem ir aktuāla. Klasiskais terminrades process ir pārāk laikietilpīgs, lai nodrošinātu mūsdienīgu nacionālo terminu datubāzi. Tā kā ne vārdnīcas, ne citas datubāzes nevar pilnībā apmierināt nepieciešamību pēc datorlingvistikas terminiem, tad nereti var veidoties nekonsekvence terminu veidošanā, lietojumā un skaidrojumā.

Juris Borzovs, Ilze Irēna Ilziņa, Valentīna Skujiņa un Ilze Vancāne ir izstrādājuši teorētisko pamatojumu sistēmiskai datorterminoloģijas izstrādei latviešu valodā, nosaucot desmit principus:

1) vienam terminam oriģinālvalodā atbilst viens noteikts termins tulkojuma valodā;
2) atšķirīgiem terminiem oriģinālvalodā arī tulkojuma valodā sniedzami atbilstoši atšķirīgi termini;
3) daudznozīmīgam vārdam termina funkcijā oriģinālvalodā jācenšas atrast vārdu ar līdzīgu nozīmju diapazonu arī tulkojuma valodā;
4) jāizvēlas tāds ekvivalents, lai, to attulkojot, nepārprotami tiktu izmantots tas pats oriģinālvalodas vārds;

5) veidojot jaundarinājumu, jāievēro tā iederīgums attiecīgajā terminu sistēmā un līdzība ar tam tuviem un analoģiskiem terminiem, vēlams, lai jaundarinājums būtu ērti izmantojams par tālāku atvasinājumu bāzi;
6) vārdu aizgūstot, jāraugās, cik tas iederīgs tulkojuma valodā gan semantiskā, gan fonētiskā un morfoloģiskā ziņā;
7) ja ir iespējami internacionālu vārdu un pašcilmes vārdu sinonīmi, priekšroka dodama pašcilmes vārdiem;
8) praksē jau ieviesušos vārdus bez pietiekama pamatojuma mainīt nevajadzētu;
9) ikdienā lietojamiem terminiem pievēršama lielāka uzmanība, tiem jābūt īsiem, precīziem, labskanīgiem, viegli uztveramiem; reti lietojamiem terminiem prasības var nebūt tik stingras;
10) neviens no iepriekšminētajiem principiem nav absolutizējams. (Borzovs u. c. 2001)

Terminoloģijas speciāliste Valentīna Skujiņa savā pamatnostādnē ir strikta: „Lai cik specifiski ir jēdzieni un tos apzīmējošie termini, tomēr katram terminam ir jāatbilst latviešu valodas vārddarināšanas likumībām un latviešu literārās valodas normām un attīstības tendencēm." (Skujiņa 2002, 192) Praksē redzams, ka ne vienmēr izdodas ievērot šo prasību.

Publikācijas mērķis – dot ieskatu datorlingvistikas terminoloģijas galvenajās iezīmēs un attīstības tendencēs, raksturojot tās latviešu valodas gramatikas un latviešu valodas terminoloģijas aspektā. Mērķa realizēšanai izpētīta Ilzes Auziņas, Maijas Baltiņas, Gunta Bārzdiņa, Normunda Grūzīša, Renāra Kudiņa, Valērija Krugļevska, Everitas Milčonokas (Andronovas), Guntas Nešpores, Anitas Ozoliņas, Baibas Saulītes, Ingunas Skadiņas, Andreja Spektora, Māra Šteinberga, Jurģa Šķiltera, Ilzes Vancānes, Andreja Vasiļjeva, Andreja Veisberga rakstos lietotā leksika. Aplūkotas vairāk nekā divdesmit publikācijas un prezentācijas. Terminoloģijas jautājumu teorētiskais pamatojums ņemts no Valentīnas Skujiņas, Māra Baltiņa, Jura Baldunčika u. c. valodnieku pētījumiem. Piemēriem izmantoti teikumi no aplūkotās literatūras, tāpēc nobeigumā dots kopīgs avotu un literatūras saraksts.

Salikteņdarināšana un vārdkopterminu veidošana

Latviešu valodā salikteņdarināšana ir īpaši produktīva terminu veidošanā, jo saliktenī visuzskatāmāk ir iespējams izteikt virsjēdziena un apakšjēdziena pazīmi (Skujiņa 2002, 88). Tādējādi pašsaprotami, ka salīdzinoši jaunajā zinātnes nozarē veidoti vairāki salikteņtermini, kuru pirmo komponentu veido celms *dator-*, piemēram, *datorlingvistika*, *datorleksikons*, *datorfonds*, *datoranalīze*, *datorversija*, *datorresursi*, vai *mašīn-*, piemēram, *mašīntulkošana*, *mašīnapmācība*,

mašīnmācīšanās. Ilze Irēna Ilziņa apgalvo, ka ir vairāk nekā 90 salikteņterminu, kuru veidošanā izmantota leksēma *dator-* (Ilziņa 2010, 75).

Ir izveidojusies jauna zinātnes nozare – datorlingvistika[1], kuras uzdevums ir lingvistisko pētījumu automatizācija. (Ozoliņa 1997, 219)

Ir salīdzinoši vienkārši datorleksikonā dot vārda morfoloģisko informāciju, jo šī informācija ir konkrēta un galīgi aprakstāma. (Milčonoka 2000, 209)

Trešais veids ir tulkotāja programmrīki un mašīntulkošanas sistēmas. (Spektors 2000, 296)

Uzkrājot šādu precedentu korpusu (ontoloģiju), tā vēlāk var tikt izmantota jaunu tekstu automātiskai semantiskai analīzei, lietojot mašīnmācīšanās metodes. (Bārzdiņš, Grūzītis, Nešpore 2008, 13)

Sastopami līdzīgas izcelsmes salikteņtermini, kuru otrais komponents ir *-ontoloģija*, piemēram, *mikroontoloģija*, *paraugontoloģija.*

Pašlaik šī pieeja ir tikai teorētiskas izstrādes stadijā, ir izveidots neliels skaits paraugontoloģiju, tiek attīstīta μ-ontoloģiju rakstīšanas metodika. (Saulīte u. c. 2008, 134)

Bet tikpat produktīvs šis vārds *ontoloģija* (arī *ontoloģisks)* ir vārdkopterminu veidošanas procesā, atrodoties gan atkarīgā komponenta lomā, piemēram, *ontoloģijas vienības*, *ontoloģiskā semantika*, gan neatkarīgā komponenta lomā, piemēram, *OntoSem ontoloģija*, *universālās ontoloģijas*, *terminu ontoloģijas.*

Tajā pašā laikā šīs sastatīšanas rezultāts ir ļoti vērtīgs lingvistisks resurss, kas padara iespējamu gan dabiskās valodas tekstu automatizētu translēšanu atbilstoši OntoSem ontoloģijai. (Bārzdiņš u. c. 2006, 15)

Ļoti aktīvs elements vārdkopterminos ir vārds *korpuss*, kas izmantots kā neatkarīgais komponents, piemēram, *tekstu korpuss*, *valodas korpuss*, *nacionālais korpuss*, *anotēts korpuss*, *marķēts korpuss*, *nemarķēts korpuss*, *vienvalodas korpuss*, *paralēlais korpuss*, *divvalodu korpuss*, *daudzvalodu korpuss*, *specializēts korpuss*, *sinhronisks korpuss*, *diahronisks korpuss.*

Precīzs termins veidots, izmantojot vārda *korpuss* sakni celmu salikteņterminā *korpuslingvistika* (Andronovs, Andronova 2011), bet paralēli tiek izmantota arī šī termina vārdkopas forma *korpusa lingvistika* (Baltiņa 2006; VPSV 2007, 196). Dažādos laika periodos un arī atsevišķos izdevumos vērojama iezīme lietot šo terminu gan kā vārdkopterminu, t. i., atsevišķi kā divus nesaplūdušus vārdus, gan kā salikteņterminu. Lai atšķirtu salikteni no vārdu savienojuma, jāņem vērā

[1] Pasvītrojums teikumā norāda uz terminu, kurš tiek pētīts.

vairākas pazīmes (Ahero 1979), bet vārdkopas *korpusa lingvistika* un salikteņa *korpuslingvistika* semantiskajā lietojumā atšķirības nav būtiskas. Vērojot tendenci – strauji palielināties saliktu vārdu īpatsvaram latviešu valodā –, ir pieļaujams, ka šī un citas līdzīgas vārdkopas tiks pilnīgi aizstāta ar sintaktiski veidotu salikteni.

Vārda semantikas maiņa terminā

Dažkārt datorlingvistikā jaunu terminu veidošanai izmantoti jau zināmi vārdi, visbiežāk svešvārdi, kuriem tiek dots jauns nozīmes skaidrojums, piemēram, *korpuss*, *anotācija*, *konkordance* u. c. Visiem šiem vārdiem var atrast skaidrojumu svešvārdu vārdnīcās, un tas ir pilnīgi vai daļēji atšķirīgs no tā, kādu jēgu un saturu konkrētajam vārdam piedēvējuši datorlingvistikas speciālisti. Piemēram, vārds *anotācija*, vispārlietojamā leksikā nozīmē – *īss grāmatas, sacerējuma vai cita informācijas avota raksturojums, dažkārt ar novērtējumu* (SvV 1999, 58). Datorlingvistikā ar *anotētu tekstu* saprot marķētu tekstu, kas sagatavots tālākai apstrādei. Līdzīgi var vērot jēdziena *korpuss* semantikas maiņu. Sākotnēji datorlingvistikā izmantojamo tekstu kopumu dēvēja gan par *datorfondu*, gan par *tekstu masīvu*, gan par *elektronisko tekstu krājumu* (Spektors 2000). Desmit gadu laikā ir nostabilizējies termins *korpuss*, lai gan ne visu *tekstu krājumu* dēvē par *korpusu*. Terminam *korpuss* datorlingvistikā ir atšķirīga semantika no svešvārdu vārdnīcās skaidrotajām 3–4 nozīmēm ('pamatelements ierīcei', 'viena no vairākām celtnēm', 'karaspēka vienība', 'personu kopums'). Terminu *korpuss* arvien biežāk izmanto valodniecības speciālisti ārpus datorlingvistikas, runājot par apjomīgu izpētei sagatavotu valodas materiālu kopumu. Tas nozīmē, ka jēdziena semantika turpina attīstīties ārpus vienas zinātnes nozares robežām.

Ontoloģija – termins no citas zinātnes

Kā starpdisciplināra zinātņu nozare, datorlingvistika nereti pārņem terminus no radniecīgām zinātņu nozarēm. Visvairāk kopīgo terminu var saskatīt ar datorzinātni un valodniecību (gramatiku), bet nereti – ar citām zinātņu nozarēm (filozofiju, matemātiku, fiziku, loģiku u. c.)

Vārds *ontoloģija* „Svešvārdu vārdnīcā" (SvV 1999, 535) skaidrots ar divām nozīmēm: 1) mācība par esamību, kurā skaidroti esamības vispārīgie pamati, principi, struktūra un likumsakarības; 2) pieņēmumi par to, kādas lietas eksistē vai var eksistēt attiecīgajā realitātes jomā, kādi varētu būt to eksistences apstākļi, no kā un kādā veidā tās varētu būt atkarīgas utt.; katrai zinātnes nozarei un katrai teorijai ir sava ontoloģija. Rakstā, kurā atklāta ontoloģijas jēdziena būtība datorlingvistikā, ir dots plašāks skaidrojums, ko šis termins nozīmē datorzinātnē:

„Datorzinātnē termins *ontoloģija* ir aizgūts no filozofijas; ar to apzīmē virzienu, kas apraksta dažādus pasaules objektus un to savstarpējās attieksmes. Ar ontoloģijām tiek saprasti dažādi formāli pasaules modeļi, kas veidoti, balstoties uz universālu jēdzienu sistēmu. Tā ir hierarhiska sistēma, kur katrs jēdziens ir nosaukts kādā nosacītā vārdā un kur ir definētas attieksmes starp jēdzieniem – gan hierarhiskas attieksmes ar citiem jēdzieniem, gan jēdzieniem atbilstošo vārdu iespējamās funkcijas teikumā u. tml. Informācija par pasaules zināšanām tiek pierakstīta loģikas formulās, tāpēc to var izmantot, apstrādāt, pārbaudīt (meklēt modelī pretrunas u. tml.) ar datoru." (Saulīte u. c. 2008, 132)

Šāda pieeja, kad autori sākumā izklāsta darbā lietoto terminu būtību un saturu, ir absolūti nepieciešama, ja jēdzieni tiek aizgūti (pārņemti) no citām zinātņu nozarēm, kā šajā gadījumā – no filozofijas.

Pamatojoties uz terminu *ontoloģija*, tiek veidoti jauni termini *mikroontoloģija*, *nozares ontoloģija*, *terminu ontoloģija*, *vispārīgā ontoloģija*, *universālā ontoloģija*, *konkrētās valodas ontoloģija*. Kā redzams, lielākoties veidoti vārdkoptermini ar diviem (vienā gadījumā – trijiem) komponentiem. Saskaņā ar latviešu valodas vārdkopu struktūras likumībām (Skujiņa 2002, 111) prepozitīvi apzīmējošajā funkcijā lietoti substantīvi (*nozares*, *terminu*, *valodas*) un adjektīvi (*vispārīgā*, *universālā*). Vienā gadījumā jaunais termins veidots kā saliktenis – *mikroontoloģija*. Salikteni visuzskatāmāk ir iespējams izteikt virsjēdziena un apakšjēdziena pazīmi un tādējādi īstenot terminu sistēmiskuma prasību (Skujiņa 2002, 88). Salikteņdarinājumi zinātniskajā terminoloģijā ir ļoti izplatīti. Šajā gadījumā izmantots jau zināms termins *ontoloģija*, kuram pievienots prepozitīvais elements *mikro-* ar nozīmi 'ļoti mazs, sīks' (SvV 1999, 481), līdzīgi kā veidots termins *mikroorganisms*.

Neparasti ir tas, ka datorlingvisti rakstu valodā šo terminu pieraksta, aizstājot salikteņa pirmo komponentu ar simbolu μ (mikro) un izmantojot defisi. Tādējādi *μ-ontoloģija* termina funkcijā tiek lietots kā defissavienojums. Šādā pozīcijā, kad simbolisku apzīmējumu ar defisi piesaista sekojošajam vārdam, latviešu valodnieki (Blinkena 1969, 385) iesaka defisi atmest, paturot adefisālu konstrukciju, piemēram, *μ ontoloģija*. Arī pašā jaunākajā izdevumā „Latviešu interpunkcija" šāds defises lietojums nav paredzēts (Blinkena 2009, 407–411).

Termini ar defisi un vienotājdomuzīmi

Datorlingvistikā tiek izmantoti termini, kas veidoti kā defissavienojumi. Latviešu valodā šādas vienības tiek uzskatītas par salikteņiem (Skujiņa 2002, 115–120). Komponentu saistīšanai sakārtojuma attieksmē latviešu valodā mēdz izmantot vienotājdomuzīmi. Ar vienotājdomuzīmi saistīto vārdu savienojumus pēc

analoģijas ar defisavienojumiem nosacīti var saukt par vienotājdomuzīmes savienojumiem. Taču terminoloģijā vienotājdomuzīmes savienojumi netiek izmantoti patstāvīgā termina funkcijā. Tos lieto tikai terminelementa funkcijā – vārdkopterminu atkarīgajos komponentos, piemēram, *Džoula–Lenca likums*. Datorlingvistikā vērojami tādi termini, kuros iekļauti vienotājdomuzīmes savienojumi *grafēmas–fonēmas* (*likumi, transkripcija*), *teksta–runas* (*sintēze, sistēma*), *jautājumu–atbilžu* (*sistēma*).

Vispirms notiek automatizēta ieejas teksta transkribēšana, izmantojot izstrādātos grafēmas–fonēmas atbilsmju likumus. (Auziņa 2000, 17)

Veidojot teksta–runas sintēzes sistēmu latviešu valodai, tiek izmantota konkatenācijas metode. (Auziņa 2005, 21)

Dažkārt vērojama nekonsekvence un vienotājdomuzīmes savienojumi tiek pierakstīti ar defisi, piemēram, *teksta-runas sistēma*.

Pašlaik pasaulē nav nevienas teksta-runas sistēmas, kas pilnībā balstītos uz vārdiem vai zilbēm. (Auziņa 2003, 65)

Iespējams, tas notiek tikai tehnisku iemeslu dēļ vai pēc redaktora ieskatiem, bet nav izslēdzama arī iespēja, ka nepietiekamas kompetences dēļ netiek nošķirta atšķirība starp defisi un vienotājdomuzīmi. Praksē atrodami arī gadījumi, kad vienotājdomuzīmes vietā (domuzīme bez atstarpēm abās pusēs) lietota nevis defise, bet gan parastā domuzīme, piemēram, *jautājumu – atbilžu sistēma*.

Šādam korpusam sasniedzot jau minētos 150 miljonus vārdlietojumu, vajadzētu kļūt par pamatu kā mākslīgā intelekta sistēmu izstrādei, tā arī turpmākiem latviešu valodas pētījumiem. Tad varētu ķerties arī pie grūtākiem uzdevumiem, kā jautājumu – atbilžu un dialogu sistēmu izstrādes latviešu valodā, kas tālākā nākotnē kalpos par nodrošinājumu valodas izdzīvošanai datorizētajā pasaulē. (Spektors 2000, 298)

Abreviatūras (iniciāļu un zilbju saīsinājumi)

Viens no visintensīvāk izmantotajiem vārddarināšanas paņēmieniem, veidojot terminus datorzinātnē, t. sk. datorlingvistikā, ir saīsinājumu salikteņi jeb abreviatūras. Tām veltīta raksta apjomīgākā daļa.

Pirms detalizētākas abreviatūru lietojuma analīzes datorlingvistikā un šai zinātnes nozarei veltītajos zinātniskajos tekstos, sniegts koncentrēts to teorētiskais raksturojums.

Andrejs Veisbergs, aprakstot mūsdienu vārddarināšanas procesu, uzsver trīs iezīmes: abreviāciju, strupināšanu un saplūdeņu veidošanu. „Valodas demokratizācija visumā veicina tieši īsināšanas procesus. Zināmu iespaidu uz tiem atstāj arī kontaktvalodu ietekme (sevišķi angļu valodas pieaugošā ietekme) uz latviešu valodu." (Veisbergs 1997, 271)

Abreviatūru lietojums strauji pieaug mūsdienu valodas dažādās lietojuma sfērās. Tās lieto ne tikai atsevišķu nozaru speciālisti, bet arvien vairāk šie saīsinājumi ieplūst arī vispārlietojamā rakstu un runas valodā. Abreviatūras tiek iekļautas gan saīsinājumu vārdnīcās, gan skaidrojošajās vārdnīcās, gan tulkojošajās vārdnīcās, gan terminu vārdnīcās. Andrejs Bankavs rakstā „Abreviatūras 90. gadu leksikā" konstatē: „Abreviatūras kļuvušas par sava veida XX gadsimta lingvistisku universāliju. Tās plaši tiek lietotas tehnikā, reklāmtekstos, publicistikā. [..] Reizē saīsinājumi ir leksikas visnepastāvīgākā daļa, kas mainās, mainoties politiskajiem, sociālekonomiskajiem un citiem apstākļiem. Nemainīga paliek šo vienību pieauguma tendence." (Bankavs 1997, 7)

Kā norādīts „Mūsdienu latviešu literārās valodas gramatikā", abreviatūras intensīvi latviešu valodā veidojušās, sākot ar 20. gadsimta otro pusi (MLLVG 1959, 221). Tolaik tas noticis spēcīgā krievu valodas ietekmē. 20. gadsimta beigās un 21. gadsimta sākumā dominē no angļu valodas aizgūtie saīsinājumi. Sākot ar 20. gadsimta 90. gadiem, ir izteikta angļu valodas kā dominējošās kontaktvalodas ietekme uz latviešu terminoloģijas veidošanos (Baltiņš 2008).

Pēc izcelsmes iedala nacionālās abreviatūras (*GNP – Gaujas Nacionālais parks*) un aizgūtās abreviatūras. Aizgūtās abreviatūras visbiežāk ir no angļu valodas (*BBC – British Broadcasting Corporation*, *NATO – North Atlantic Treaty Organization*), bet ir arī tulkotās abreviatūras (*ASV – Amerikas Savienotās Valstis*, *ES – Eiropas Savienība*).

Pēc struktūras izdalītas divu veidu abreviatūras: zilbju salikteņi (*proforgs*, *kolhozs*) un iniciāļu salikteņi (*ASV*, *ANO*, *HES*). (MLLVG 1959, 221–223) Iniciāļi, kas veidoti nevis no burtiem, bet no skaņām (akronīmi), nereti pārvēršas sugas vārdos, kurus raksta ar mazajiem burtiem, piemēram, *TEC – tecs*, *VEF – vefs* vai *vefiņš*, *NEP – neps*, *KAMAZ – kamazs*.

Savukārt pēc lietojuma būtu izdalāmas vispārzināmās abreviatūras, kuras lieto jebkurš valodas lietotājs, (*SOS, NATO, HES*) un speciālās, kuras lieto kādas konkrētas zinātņu nozares valodā un no kurām veidojas arī termini. Jau 20. gadsimta 70. gados Rita Kalnbērziņa rakstīja, ka abreviatūras bieži vien izmanto tikai attiecīgās zinātnes nozares speciālisti (Kalnbērziņa 1975).

Pašlaik datorzinātnē un arī datorlingvistikā dominē angļu valodas abreviatūru nepastarpināta un bieži arī netulkota aizgūšana.

Datorlingvistikas (arī datorzinātnes) pētījumiem veltītajos rakstos atrodamas ļoti daudzas iniciāļu abreviatūras – *HTML*, *TMR*, *URI*, *RDF*, *XML*, *UML*, *W3C*, *OWL*, *PDDL*, *LPC*, *OCR*, *URL*, *XHTML*, *HTTP*, *SGML*, *GVS*, *SPARQL*, *RDF/XML*, *N3*, *RDF/XML-ABBREV* u. c. Ielūkojoties „Latviešu valodas saīsinājumu vārdnīcā" (Bankavs 2003), redzams, ka saīsinājumu datorzinātnē ir 2–3 reizes vairāk nekā tādās senās zinātņu nozarēs kā fizikā un ķīmijā. Valodas zinātniskais stils pieprasa jebkurā tekstā vai prezentācijā, minot kādu saīsinājumu pirmo reizi, sniegt informāciju par tā atšifrējumu. Šķiet, tas nav nepieciešams, izmantojot tādas vispārzināmas abreviatūras kā *WWW* vai arī *HTTP*. Piemēram:

Lietvedības terminu vārdnīcu paredzēts ievietot Internetā WWW *lapās un tā būs brīvi pieejama.* (Spektors 1999, 56)

Biežāk izplatīts paņēmiens ir norādīt abreviatūras burtus, neatšifrējot tos angļu valodas vārdus, no kuriem saīsinājums veidojies. Piemēram:

Resursu aprakstīšanas ietvars (turpmāk RDF*) ir elastīga tehnoloģija, kas saistīta ar mēģinājumiem radīt semantisko tīmekli.* (Šteinbergs 2008)

Informācijas precizitātei šādā gadījumā pietrūkst norādes, ka *RDF* ir veidots no vārdiem *Resource Description Framework – resursu aprakstīšanas ietvars*.

Stipri apgrūtināta zinātniskā teksta uztvere ir tad, ja aiz abreviatūras neseko nekāda paskaidrojoša norāde.

Pagaidām tikai divu grāmatu, Platona „Valsts" (aptuveni 60 000 vārdlietojumu) un Dž. Orvela „1984" (aptuveni 100 000 vārdlietojumu) teksti ir marķēti atbilstoši SGML *prasībām.* (Spektors 1999, 54)

Lietojums: nozarspecifisku ontoloģiju aprakstīšana un verbalizācija (alternatīva UML *diagrammām un loģikas formulām).* (Saulīte, Grūzītis 2009)

Vērojama vēl viena tendence, lietojot abreviatūras – uzrādīt angļu valodas terminu, no kura veidots saīsinājums, bet nenorādīt tā latvisko skaidrojumu.

Sporta leksikons tika izveidots UNL *(Universal Networking Language) projekta ietvaros, dalībniekiem izvēloties tekstus un aprakstot tos ar starpniekvalodas palīdzību.* (Milčonoka 2000, 209)

Dažādo terminoloģijas resursu metadatu aprakstīšanai EuroTermBank projekta ietvaros tiek izmantots TeDIF *(Termilology Documentation Interchange Format) standarts, kas iedibina kopēju terminoloģijas bibliogrāfisko un faktoloģisko datu formātu.* (Vasiļjevs, Rirdance 2008, 114)

Ilze Auziņa, pētot latviešu valodas runas datormodelēšanu, atzinusi, ka dažādu saīsinājumu un abreviatūru lietošana apgrūtina teksta transkribēšanu, nepieciešams veikt teksta priekšapstrādi, saīsinājumu izvēršanu, atšifrēšanu (Auziņa 2005, 25). Līdzīga problēma ir valodniekam, kurš nav datorzinātnes speciālists, viņam vispirms ir jāapgūst datorzinātnes un datorlingvistikas specifiskā valoda, lai spētu uztvert informāciju šai zinātnes apakšnozarē. Tas var atgrūst, atbaidīt, mulsināt šādu tekstu lasītāju. Situāciju sarežģī daudzo abreviatūru lietojums tekstos, kam skaidrojums meklējams ārpus konkrētā teksta citos informācijas avotos – vārdnīcās, rokasgrāmatās, enciklopēdijās.

Pētījumam veikts eksperiments – ņemti 13 nejauši izvēlēti datorzinātnes un datorlingvistikas terminu saīsinājumi jeb iniciāļu abreviatūras, kurām tekstā nebija dots skaidrojums (atšifrējums) vai arī tas bija dots tikai daļēji. Ar 7 dažādu vārdnīcu palīdzību mēģināts noskaidrot katras abreviatūras atšifrējumu un līdz ar to arī terminu latviešu valodā. Izmantoti dažādi uzziņas avoti, no kuriem 6 – elektroniski un tikai viens grāmatas izdevums:

1) *Lielā terminu vārdnīca. Datortermini* (*www.termini.lv*);
2) *Letonikas* enciklopēdijas un vārdnīcas (*www.letonika.lv*);
3) LU MII AILab vārdnīcu serveris (*www.tezaurs.lv*);
4) Brīvā enciklopēdija *Vikipēdija* (*lv.wikipedia.org*);
5) Eiropas terminu banka *EuroTermBank* (*www.eurotermbank.com*);
6) Valsts valodas centra mājaslapas sadaļa *Terminu datubāze* (*www.vvc.gov.lv*);
7) Datorterminu vārdnīca *Personālie datori. Angļu-latviešu-krievu skaidrojošā vārdnīca*. Izdevējs Rīga : Dati, 1998. 256 lpp.

Eksperiments tika veikts 2010. gada maijā, ievadot meklētājā konkrēto abreviatūru un pierakstot saņemto rezultātu. Ja rezultāts tika saņemts (nesalīdzinot un neapspriežot tā precizitāti), tad tabulā atzīmēts X. Tukšās šūnas tabulā norāda, ka rezultāts par attiecīgo abreviatūru netika saņemts. Rezultāti atspoguļoti 1. tabulā.

Redzams, ka neviena no izvēlētajām vārdnīcām un enciklopēdijām nevarēja dot atšifrējumu visiem izvēlētajiem piemēriem, kaut arī visi šie uzziņu avoti uzskatāmi par ļoti populāriem. Visvairāk noderīga izrādījās enciklopēdija *Vikipēdija* (noderīguma koeficients 0,769). Salīdzinoši noderīgi informācijas avoti eksperimentā bija arī Eiropas terminu banka *EuroTermBank*, (noderīguma koeficients 0,538) un Valsts valodas centra mājaslapas terminu datubāze (noderīguma koeficients 0,462). Visās izvēlētajās vārdnīcās varēja atrast tikai divu abreviatūru skaidrojumu (*HTTP* un *HTML*), kam vērojams ļoti plašs lietojums datorzinātnes literatūrā, un līdz ar to tie būtu uzskatāmi par vispārlietojamajiem saīsinājumiem.

1. tabula. **Abreviatūru nozīmes skaidrojumu pieejamība dažādos uzziņu avotos**

Abreviatūra un avots, no kura tā ņemta	Atšifrējums angļu valodā	Jēdziena tulkojums latviešu valodā	www.termini.lv	www.letonika.lv	www.tezaurs.lv	lv.wikipedia.org	www.eurotermbank.com	www.vvc.gov.lv	Personālie datori. Angļu-latviešu-krievu skaidrojošā vārdnīca. Rīga : Dati, 1998
HTML Vairākos avotos	Hypertext Markup Language	Hiperteksta iezīmēšanas valoda	X	X	X	X	X	X	X
HTTP Vairākos avotos	Hyper Text Transport Protocol	Hiperteksta transporta protokols	X	X	X	X	X	X	X
LPC Skadiņa, Vasiļjevs 2008	Linear Predictive Coding	Lineārā prognozēšanas kodēšana				X			
OCR Šķilters 2010	Optical Character Recognition	Rakstzīmju optiskā pazīšana	X				X	X	X
OWL SemTi-Kamols	Web Ontology Language	Tīmekļa ontoloģijas valoda				X			
PDDL SemTi-kamols	Planning Domain Definition Language	Plānošanas domēna definēšanas valoda				X			
RDF Bārzdiņš, Grūzītis u. c. 2006	Resources Description Framework	Strukturāls resursu apraksts			X		X		
SPARQL Šteinbergs 2008	Protocol and RDF Query Language	RDF vaicājumu valoda				X			
TMR Bārzdiņš, Grūzītis u. c. 2006	Text Meaning Representation	Teksta nozīmes reprezentēšana							
UML Saulīte, Grūzītis 2009	Unified Modeling Language	Vienotā modelēšanas valoda				X			

URL Šteinbergs 2008	Uniform Resource Locator	Vienotais resursu vietrādis	X	X		X	X	X	X
W3C Šteinbergs 2008	World Wide Web Consortium	Globālā tīmekļa konsorcijs				X	X	X	
XML Bārzdiņš, Grūzītis u. c. 2006	eXtensible Markup Language	Paplašināmās iezīmēšanas valoda	X	X	X	X	X	X	
	Uzziņas avota noderīguma koeficients (konkrētajā eksperimentā)	Koeficients = iegūto rezultātu skaits/ meklējamo atbilžu skaits	0,385	0,308	0,308	0,769	0,538	0,462	0,308

Atsevišķiem terminiem abreviatūras formātā ne tikvien neizdevās atrast skaidrojumu, bet vārdnīcas piedāvāja pavisam citas nozīmes jēdzienu, ko apzīmē ar identisku abreviatūru. Piemēram, enciklopēdija *Vikipēdija* abreviatūru *TMR* skaidro kā '*Tunnel Magnetoresistance* – tuneļa magnētiskā pretestība', nevis '*Text Meaning Representation* – teksta nozīmes reprezentēšana', kā tas ir lietots datorzinātnē.

> *Ontoloģiskā semantika ir teorija par dabīgās valodas pierakstā izteiktu jēgu un valodas apstrādes tehnikām, kas patvaļīga teksta nozīmes izgūšanai, tās formālai reprezentēšanai (TMR – Text Meaning Representation) un spriešanai par zināšanām, izteiktām vai atvasināmām šajā tekstā, kā centrālo resursu izmanto formalizētu pasaules modeli.* (Bārzdiņš u. c. 2006)

Savukārt abreviatūru *LPC* vārdnīca *Tēzaurs* skaidro kā 'Latvijas Pārtikas centrs', bet enciklopēdija *Vikipēdija* kā '*Linear Predictive Coding* – lineārā prognozēšanas kodēšana'. Tieši otrais skaidrojums ir nepieciešams, lai saprastu izteikumu:

> *Ar dinamiskās programmēšanas algoritmu tiek noteikts konkrētam kontekstam vispiemērotākais fragments, salīdzinot skaņas viļņu spektrālos raksturlielumus. Visbeidzot no izvēlētajiem fragmentiem ar LPC sintēzes metodi tiek sintezēts runas signāls.* (Skadiņa, Vasiļjevs 2008, 126)

Visas aplūkotās abreviatūras veidotas no angļu valodas terminu iniciāļiem. Tas saprotams, jo datorzinātnē dominē termini, kas aizgūti no angļu valodas, bet pētījumā apzināti nav izmantotas vārdnīcas un enciklopēdijas angļu valodā. Izmantojot „Latviešu valodas saīsinājumu vārdnīcu" (Bankavs 2003), kurā ir ietverti 330 saīsinājumi informātikā, skaidrojumu atrast izdevās tikai divām no eksperimentam pakļautajām 13 abreviatūrām. Tas nozīmē, ka datorzinātnes un datorlingvistikas termini (tāpat kā

zinātne kopumā) attīstās ļoti strauji un ir absolūti nepieciešams katrā apjomīgākā pētījumā precizēt lietoto terminu un to saīsinājumu nozīmes, jo neviens uzziņu avots nespēj nodrošināt visaptverošu informāciju terminoloģijā.

Struktūras aspektā ļoti savdabīgas ir abreviatūras, kurās izmanto burtus kopā ar cipariem, piemēram, *N3*, *W3C*. Termins '*World Wide Web Consortium* – globālā tīmekļa konsorcijs' abreviatūrā tiek lietots kā *W3C*. Latviešu valodā nav tāda precedenta, kad tā vietā, lai rakstītu līdzās trīs vienādus iniciāļu burtus, piemēram, *TTT* vai *RRR*, izvēlas šādu pieraksta veidu – *T3* vai *R3,* respektīvi, *3T* vai *3R*. Ja šāda vārddarināšanas tendence kļūs izplatīta latviešu valodā, tad tā būs uzskatāma par ļoti savdabīgu vārdu strupināšanu, kas lielā mērā veidojusies datorzinātnes valodas ietekmē.

Angļu valodas spiediena rezultātā abreviatūrās tiek lietoti burti, kuru latviešu valodas alfabētā nemaz nav, piemēram, *X*, *Y*, *W*, *Q*.

Jāatzīst, ka datorzinātnē strauji attīstās tāds vārddarināšanas paņēmiens kā zilbju salikteņi. Savienojot kopā vienā vārdā atsevišķu vārdu pirmās zilbes, iegūst salikteņus, kurus lieto kā terminus, piemēram, *OntoSem*, *WordNet*, *FrameNet*, *EuroTermBank*, *MultiTerm* u. c. No latviskas cilmes vārdiem tradicionāli šādi salikteņi netiek veidoti (MLLVG 1959, 221). Latviešu valodai šādi veidojumi gan izklausās sveši, gan izskatās nepareizi, jo katra zilbe rakstīta ar lielo sākumburtu avotvalodas ietekmē.

Pirmais ievērojamais resurss – WordNet *—, kura arhitektūras iedvesmas avots ir psiholingvistiskas teorijas par cilvēka leksisko atmiņu, tika izstrādāts Prinstonas universitātē.* (Bārzdiņš u c. 2006)

Dokumenta datnē tika aizzīmēti termini un ievadīti atsevišķā MultiTerm *datubāzē.* (Vancāne, Krugļevskis 2003, 160)

FrameNet *datu bāzē pašlaik ir vairāk nekā 10 000 vārdu nozīmju, kas veido apmēram 825 situācijas.* (Saulīte u. c. 2008, 129)

Ir labi saprotama nepieciešamība ļoti ātri piemeklēt jaunus vārdus un terminus latviešu valodā, lai aprakstītu pētījumus. Vieglāk ir sākotnēji lietot no angļu valodas pārņemtu jēdzienu tulkojumus, nereti – kalkus, pamazām nonākot pie precīza nosaukuma latviešu valodā. Ieskatam – iepriekš citētā raksta autoru skaidrojums terminu izvēles sakarā: „Tā kā līdz šim latviešu valodā tikpat kā nav pētījumu situāciju semantikā, latviski nav stabilu terminu ar šo metodi saistīto jēdzienu nosaukšanai. Burtiski tulkojot angļu valodas vārdus *frame semantics* (piemēram, ietvaru, rāmju, sistēmas, uzbūves, struktūras semantika), pēc raksta autoru domām, nerodas priekšstats par aprakstāmo jēdzienu. Angļu valodas vārds *frame* (vai *framework*) latviešu valodā bieži tiek tulkots kā *ietvars*, tomēr nereti tas rada liekvārdību un

nepalīdz izprast nosauktos jēdzienus. Tāpēc raksta autori ir izvēlējušies terminu *situācija*, jo tas precīzāk raksturo šīs metodes būtību (vārdu nozīmes tiek grupētas pēc situācijas, kurā tās parasti tiek lietotas) un to var veiksmīgi izmantot, nosaucot arī pārējos šīs metodes pamatjēdzienus (situācijas semantika, situācija, situācijas elements)." (Saulīte u. c. 2008, 129) Šis raksta fragments precīzi parāda jaunu terminu meklējumu un veidošanas ceļu. Ne vienmēr termins, kurš ir piemeklēts kā atbilstošākais konkrētajai vajadzībai, tiek pieņemts un lietots zinātnes valodā. Bieži vien par terminu dzīvotspējīgāks ir profesionālisms vai žargonisms (Skujiņa 2002, 39–40). Publikācijas autore klātienē pārliecinājās, kā datorlingvistikas speciālists lekcijā doktorantiem lieto vārda *frame* tiešu pārcēlumu latviešu valodā, runājot par *fremiem* un *fremēšanu*. Aizguvums atveidots pēc izrunas un iekļauts latviešu valodas gramatiskajā sistēmā. Šeit un līdzīgos gadījumos jārunā par identiskajiem aizguvumiem (identismiem) terminu un tehnisku nosaukumu veidošanā. Tos aplūko Juris Baldunčiks rakstā „Angļu valodas ietekmes desmit izpausmes veidi mūsdienu latviešu valodā" (Baldunčiks 2010). Savukārt Ilze Irēna Ilziņa labi ir aprakstījusi kalkus, metaforas un žargonvārdus IKT terminoloģijā (Ilziņa 2010), tāpēc šai rakstā tie detalizētāk nav aplūkoti.

Dažkārt vienā rakstā līdzās tiek lietots latviešu valodai gramatiski tik neiederīgais zilbju saīsinājums (*OntoSem*) un pilnais termins (*ontoloģiskā semantika*).

OntoSem *modelis ar plašu sintaktisko un semantisko struktūru un nozīmju procedūru palīdzību ļauj aprakstīt un izskaitļot valodas (leksikona) dinamiskos aspektus, kas ir ļoti būtiski reālā, efektīvā tekstu analīzē un daudznozīmības risināšanā.* (Bārzdiņš u. c. 2006)

*Ontoloģiskā semantika ir teorija par dabīgās valodas pierakstā izteiktu jēgu un valodas apstrādes tehnikām, kas patvaļīga teksta nozīmes izgūšanai, tās formālai reprezentēšanai (*TMR – *Text Meaning Representation) un spriešanai par zināšanām, izteiktām vai atvasināmām šajā tekstā, kā centrālo resursu izmanto formalizētu pasaules modeli.* (Bārzdiņš u. c. 2006)

Iespējams, ka ilgākā laika posmā šie zilbju saīsinājumi iekļausies latviešu valodā, izmantojot latviskas vārddarināšanas iezīmes, piemēram, *ontosems* vai *ontosema*. Valodas praksē līdzīgi ir noticis ar dažādiem no krievu valodas pārņemtiem zilbju saīsinājumiem, piemēram, *kolhozs* 'kolektīvā saimniecība', *sovhozs* 'padomju saimniecība'.

Ja šāda skaņu abreviatūra valodā iegūst pilnnozīmes vārda tiesības, tā kļūst lokāma un no tās var atvasināt jaunus vārdus (Kalnbērziņa 1975, 77), piemēram, varētu veidoties atvasinājumi *ontosemēt*, *ontosemēšana*, *ontosemējums* u. tml. Iespējams, ka šādā virzienā notiks produktīva vārddarināšana jaunu terminu nepieciešamības dēļ. Vēl viens iespējamais terminu attīstības ceļš ir veidot tā sauktās jauktās abreviatūras, kad reducētā vārda daļa (zilbe) savienota ar pilnu

vārdu, respektīvi, *ontosemantika*. Arī šādi precedenti latviešu valodā jau ir pazīstami, piemēram, *komjaunatne*, *ģeofizika* u. c.

2007. gada 21. februārī Latvijā notika Eiropas terminu bankas *EuroTermBank* atklāšana (Kautiņa 2007). Šī projekta uzdevums ir nodrošināt vienotu, centralizētu piekļuvi daudzvalodu terminoloģijas resursiem internetā. Sekojot solījumam, ka *EuroTermBank* noderēs ne tikai terminu tulkošanā un pareizas to lietošanas nostiprināšanā (Linde 2007), tas tika izmantots pētījuma eksperimentā. Kā jau iepriekš teikts, šī terminu banka tikai viduvēji spēja skaidrot datorzinātnes terminus, kas izteikti ar abreviatūrām. Bet, aplūkojot šīs datubāzes nosaukumu *EuroTermBank*, jāsecina, ka nosaukumā izmantota zilbju abreviatūra ar lielo burtu lietojumu katras zilbes sākumā. Turklāt pirmajā komponentā, saglabājot angļu valodas rakstību *Euro-*, bet izrunājot latviski *Eiro-*, veidojas savdabīgs hibrīdtermins. Varētu atļauties izteikt prognozes, kādas varētu būt pārmaiņas šī nosaukuma lietojumā latviešu valodā. Pirmkārt, šis datubāzes nosaukums piesaistīs galotni un no īpašvārda kļūs par sugasvārdu, veidosies apelatīvs – *eirotermbanka*. Otrkārt, iespējams, notiks šī vārda strupinājums, atmetot pirmo komponentu – *termbanka*. Treškārt, pieļaujama iespēja, ka latviešu valodas lietotāji dos priekšroku salikteņdarināšanas ieteikumiem (Skujiņa 2002, 88–97) un izvēlēsies lietot nevis zilbju saīsinājumu, bet salikteni – *terminbanka*. Vārds *terminbanka* samērā precīzi atspoguļo jēdzienu, ko šis veidojums ietver, labi iekļaujas latviešu valodā kā 4. deklinācijas lietvārds, un detalizētāki paskaidrojumi par tā nozīmi nav nepieciešami. Valsts valodas komisijas priekšsēdētājs Andrejs Veisbergs, komentējot jaunās terminu bankas izveidi 2007. gada augustā laikrakstā „Izglītība un Kultūra", nelieto šo savdabīgo saīsinājumu *EuroTermBank*, bet gan pilnu nosaukumu *Eiropas terminu banka*.

„Ar jauno projektu Latvijas terminoloģiskie resursi iekuģo plašajos starptautiskajos ūdeņos un būs pieejami visiem. Eiropas terminu banka demonstrē datorlingvistikas iespējas un nāks par labu gan latviešu terminoloģijai un tās lietotājiem, gan valodniekiem un terminologiem, kam tagad ir iespēja sastatīt daudzu valodu terminoloģiju. Šāds sastatījums atvieglo terminradi, jo ļauj redzēt citu valodu risinājumus," atzīmē profesors Andrejs Veisbergs, Valsts valodas komisijas priekšsēdētājs. (Linde 2007)

Līdzīgs pēc struktūras ir semantiskā tīmekļa projekta nosaukums *SemTi-Kamols*, kas veidots pēc hibrīdterminu principa. Vārds ir saliktenis, ko veido trīs komponenti. Pirmie divi komponenti ir zilbju abreviatūras: *sem-* no vārda *semantisks* un *-ti-* no vārda *tīmeklis*. Salikteņa trešais komponents ir pilns latviešu valodas vārds *kamols*. Projekta nosaukums retāk tiek lietots pēdiņās, biežāk – bez tām.

Ceļā uz latviešu valodas teksta semantiskās analīzes līdzekļu izveidi SemTi-Kamola ietvaros ir izstrādātas metodes un rīki arī zemāko līmeņu padziļinātai analīzei un anotēšanai [..] (Bārzdiņš, Grūzītis, Nešpore 2008)

Nobeigums

Datorlingvistikas (arī datorzinātnes) terminoloģijā vērojamas tādas iezīmes, kas kopīgas visai latviešu terminoloģijas attīstībai, bet ir arī atšķirīgas tendences. Aktīvi tiek izmantota salikteņdarināšana un vārdkopterminu veidošana. Vērojama vārdu semantikas maiņa terminā, kas īpaši raksturīga jaunas zinātnes nozares straujas attīstības posmā. Nereti sastopami kalki, hibrīdtermini, apelatīvi. Kā starpdisciplināra zinātņu nozare, datorlingvistika bieži pārņem terminus no radniecīgām zinātņu nozarēm (valodniecības, datorzinātnes, filozofijas, loģikas u. c.). Īpaši plaši datorlingvistikā tiek izmantotas abreviatūras. Tās ir angļu valodas cilmes, netulkotas, biežāk iniciāļu, bet ir arī zilbju abreviatūras. Gramatiskajā aspektā vērojamas neatbilsmes latviešu valodas normām lielo sākumburtu lietojumā zilbju abreviatūrās, kā arī defises un vienotājdomuzīmes izmantojumā. Leksiskajā aspektā ne vienmēr tiek nošķirta precīzu terminu lietošana no profesionālismu un žargonismu lietojuma. Jūtams, ka datorlingvistikas terminoloģija, tāpat kā šī zinātnes nozare kopumā, ir straujas attīstības stadijā, tādēļ nepieciešams regulāri atjaunināt informāciju vārdnīcās. Jauno terminu nozīmes precizēšana vēlama katra raksta vai plašāka pētījuma sākumā, lai nodrošinātu semantisko precizitāti pareizas informācijas sniegšanā un uztverē.

Literatūra

1. Ahero, Antonija. Salikteņu pazīmes un rakstība. No: *Latviešu valodas kultūras jautājumi,* 15. laidiens. Rīga : Liesma, 1979, 164.–179. lpp.
2. Andronova, E., Andronovs, A. Latviešu valodas korpuss un tā izmantošana. No: *Valodas prakse: vērojumi un ieteikumi*. Populārzinātnisku rakstu krājums Nr. 6. Rīga : Latviešu valodas aģentūra, 2011, 41.–57. lpp.
3. Auziņa, Ilze. Fonēmu bibliotēkas izstrāde, tās nozīme runas sintēzē. No: *Baltistika IX, 2000. Starptautiskais baltistu kongress „Baltu valodas laikmetu griežos"03.10.2000.–06.10.2000.* Referātu tēzes. Rīga : LU Latviešu valodas institūts, 2000, 16.–19. lpp.
4. Auziņa, Ilze. Segmentu izvēle runas sintēzei. No: *Linguistica Lettica*, Nr. 12. Rīga : Latviešu valodas institūts, 2003, 61.–72. lpp.
5. Auziņa, Ilze. *Latviešu valodas izrunas datormodelēšana*. Promocijas darba kopsavilkums filoloģijas doktora grāda iegūšanai valodniecības zinātņu nozares latviešu sinhronās valodniecības apakšnozarē. Rīga : Latvijas Universitāte, 2005.
6. Baldunčiks, Juris. Angļu valodas ietekmes desmit izpausmes veidi mūsdienu latviešu valodā. No: *Linguistica Lettica*, Nr. 19. Rīga : Latviešu valodas institūts, 2010, 62.–73. lpp.
7. Baltiņa, Maija. No vārdu sarakstiem līdz korpusa lingvistikai. No: *Linguistica Lettica*, Nr. 15. Rīga : Latviešu valodas institūts, 2006, 77.–84. lpp.
8. Baltiņš, Māris. German, Russian, English – Their Impact on Latvian Terminology: Similarities and Discrepancies. No: *VALODA – 2008. Valoda dažādu kultūru kontekstā.* Zinātnisko rakstu krājums XVIII. Daugavpils : Daugavpils Universitātes Akadēmiskais apgāds „Saule", 2008, 325.–330. lpp.
9. Bankavs, Andrejs. Abreviatūras 90. gadu leksikā. No: *Linguistica Lettica*, Nr. 1. Rīga : Latviešu valodas institūts, 1997, 7.–13. lpp.

10. Bankavs, Andrejs. *Latviešu valodas saīsinājumu vārdnīca*. Rīga : Avots, 2003.

11. Bārzdiņš, Guntis, Grūzītis, Normunds, Kudiņš, Renārs, Nešpore, Gunta, Spektors, Andrejs. Latviešu valoda semantiskajā tīmeklī. No: *Latvijas Zinātņu akadēmijas Vēstis*, A, 2006, Nr. 6, 60. sēj., 26.–42. lpp.

12. Bārzdiņš, Guntis, Grūzītis, Normunds, Nešpore, Gunta. Metodes un rīki tekstu korpusa daudzdimensionālai anotēšanai: projekta *SemTi-Kamols* pieredze. No: *Starptautiskās konferences „BALTU DIENAS un BALTĀS NAKTIS – apaļais galds „Baltu filoloģija": pēc desmit gadiem" tēžu krājums*. Sanktpēterburga, 2008, 12.–14. lpp.

13. Blinkena, Aina. *Latviešu interpunkcija*. Rīga : Zinātne, 1969.

14. Blinkena, Aina. *Latviešu interpunkcija*. Rīga : Zvaigzne ABC, 2009.

15. Borzovs u. c. 2001 – Borzovs, Juris, Ilziņa, Ilze Irēna, Skujiņa, Valentīna, Vancāne, Ilze. Sistēmiska latviešu datorterminoloģijas izstrāde. No: *LZA Vēstis*, 2001, 55. sēj. 1./2., 83.–91. lpp.

16. Ilziņa, Ilze Irēna. ISO standarti un informācijas un komunikācijas tehnoloģijas terminoloģijas izstrādes pieredze. No: *Linguistica Lettica*, Nr. 19. Rīga : Latviešu valodas institūts, 2010, 74.–81. lpp.

17. Kalnbērziņa, Rita. Abreviatūru struktūra un to lietošana latviešu valodā. No: *Latviešu valodas kultūras jautājumi*, 11. laidiens. Rīga : Liesma, 1975, 74.–88. lpp.

18. Kautiņa, Lāsma. Latvija realizējusi starptautiskas terminu bankas izveidi. *Tilde*, 21.02.2007 [skatīts 2010. g. 24. martā]. Pieejams: http://termini.lza.lv

19. Linde, Jānis. Jauns terminu resurss. *Izglītība un Kultūra*, 07.08.2007 [skatīts 2010. g. 24. martā]. Pieejams: http://termini.lza.lv

20. Milčonoka, Everita. Vārda raksturojums datorleksikonā. No: *Baltistika IX, 2000. Starptautiskais baltistu kongress „Baltu valodas laikmetu griežos" 03.10.2000.–06.10.2000.* Referātu tēzes. Rīga : LU Latviešu valodas institūts, 2000, 209.–211. lpp.

21. MLLVG 1959 – *Mūsdienu latviešu literārās valodas gramatika*, I. Rīga : Latvijas PSR Zinātņu akadēmijas izdevniecība, 1959, 221.–223. lpp.

22. Ozoliņa, Anita. 17. gs. tekstu datorfonda izveides programmlīdzekļi. *Linguistica Lettica*, Nr. 1. Rīga : Latviešu valodas institūts, 1997, 219.–225. lpp.

23. Saulīte, Baiba, Grūzītis, Normunds. Teksta informatīvās struktūras analīze koreferences noteikšanai kontrolētā latviešu valodā. *14. starptautiskās zinātniskās konferences „Vārds un tā pētīšanas aspekti" 2009. gada 26.–27. novembrī Liepājas Universitātē materiāli*. [skatīts 2010. g. 7. aprīlī]. Pieejams: http://www.semti-kamols.lv/doc_upl/Lieppa09.pdf

24. Saulīte, Baiba, Nešpore, Gunta, Bārzdiņš, Guntis, Grūzītis, Normunds. μ-ontoloģijas – situāciju semantikas un ontoloģiskās semantikas apvienojums. *Letonikas otrais kongress. Valodniecības raksti-2*. Rīga : Latvijas Zinātņu akadēmija, 2008, 128.–135. lpp.

25. *SemTi-Kamols*. [skatīts 2010. g. 6. apr.]. Pieejams: http://www.semti-kamols.lv

26. Senie – *Senie. Latviešu valodas seno tekstu korpuss*. [skatīts 2010. g. 6. apr.]. Pieejams: http://www.korpuss.lv/senie/about.htm

27. Skadiņa, Inguna, Vasiļjevs, Andrejs. Latviešu valoda jaunākajās lietojumprogrammās. No: *Letonikas otrais kongress. Valodniecības raksti-2*. Rīga : Latvijas Zinātņu akadēmija, 2008, 118.–127. lpp.

28. Skujiņa, Valentīna. *Latviešu terminoloģijas izstrādes principi*. Otrais, labotais un papildinātais izdevums. Rīga : Latviešu valodas institūts, 2002.

29. Spektors, Andrejs. Latviešu valoda Internetā un datorlingvistikas resursi. No: *Konferences „Latviešu valoda – esamība, vide, konteksti" materiāli 1997. gada 14. martā Rīgā*. Rīga : PBLA, 1997, 46.–53. lpp.

30. Spektors, Andrejs. Latviešu valodas datorlingvistikas resursi. *Baltu filoloģija*, VIII. Zinātniskie raksti, 619. sējums. Rīga : Latvijas Universitāte, 1999, 53.–59. lpp.

31. Spektors, Andrejs. Datorlingvistika un tās resursi. No: *Baltistika IX, 2000. Starptautiskais baltistu kongress „Baltu valodas laikmetu griežos" 03.10.2000.–06.10.2000*. Referātu tēzes. Rīga : LU Latviešu valodas institūts, 2000, 296.–298. lpp.

32. SvV 1999 – *Svešvārdu vārdnīca*. Jura Baldunčika redakcijā. Rīga : Jumava, 1999.

33. Šķilters, Jurģis. Korpuss un nacionālā identitāte: LNB vieta korpusa izveidē. *CLARIN projektam veltītā semināra materiāli 26.02.2010.* [skatīts 2010. g. 24. martā]. Pieejams: www.clarin.lv

34. Šteinbergs, Māris. *Datorlingvistika. Autoreferāts*. [skatīts 2010. g. 25. martā]. Pieejams: http://www.ante.lv/xwiki/bin

35. Baltiņš, Māris. Terminrades process pēdējo piecpadsmit gadu laikā: pagātnes mantojums un nākotnes perspektīvas. Valsts valodas komisija. No: *Latviešu valoda 15 neatkarības gados: Lingvistiskā situācija, attieksme, procesi, tendences*. Rīga : Zinātne, 2007, 401.–461. lpp.

36. VPSV – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Atb. red. V. Skujiņa. Rīga : LU Latviešu valodas institūts, Valsts valodas aģentūra, 2007.

37. Vancāne, Ilze, Krugļevskis, Valērijs. Vārdkopterminu struktūra un datorizēta meklēšana tekstos. No: *Linguistica Lettica*, Nr. 12. Rīga : Latviešu valodas institūts, 2003, 154.–162. lpp.

38. Vasiļjevs, Andrejs, Rirdance, Signe. Latviešu valodas terminoloģijas konsolidēšana vienotā terminu bankā. No: *Letonikas otrais kongress. Valodniecības raksti-2*. Rīga : Latvijas Zinātņu akadēmija, 2008, 106.–117. lpp.

39. Veisbergs, Andrejs. Īsinātās vārddarināšanas formas latviešu valodā. No: *Linguistica Lettica*, Nr. 1. Rīga : Latviešu valodas institūts, 1997, 271.–282. lpp.

Korpuslingvistika un korpusu analīzes rīki

Anna Briška

Ieskats projekta „HipiLatLit" speciālā korpusa izveidē un izmantošanas iespējās

Cilvēkresursi, kultūra, sociālekonomiskais kapitāls un vides kvalitāte ir reģionālo konkurētspēju raksturojošie faktori. Tie ir svarīgi, lai piesaistītu augsti kvalificētu darbaspēku un uzlabotu inovāciju attīstību. Šajā kontekstā būtiski ir saglabāt teritoriālo identitāti, vietējo, unikālo reģiona kultūru un teritoriju dažādību. Līdz ar to svarīgi ir apzināties, kas veicinātu Latgales un Lietuvas austrumu reģiona unikālā potenciāla izmantošanu reģiona konkurētspējas paaugstināšanā, respektējot globalizācijas tendences, kā arī ievērojot Eiropas Savienības, sevišķi Baltijas jūras reģiona valstu, kontekstu.

2011. gada janvārī Rēzeknes Augstskolas (RA) vadībā tika sākts projekts "Development of Research Infrastructure for Education in the Humanities in Eastern Latvia, Lithuania" („Humanitārās izglītības pētniecības infrastruktūras izveide Austrumlatvijā, Lietuvā", saīsinātais nosaukums – „HipiLatLit"). Tā pamatmērķis ir divu gadu laikā modernizēt humanitāro zinātņu jomu augstākajā izglītībā Lietuvas un Latvijas austrumu pierobežā, izveidojot kopīgu pētniecības infrastruktūru. Projekta realizēšanā piedalās arī Vītauta Dižā Universitāte (Kauņa, Lietuva) un Latvijas Universitātes Matemātikas un informātikas institūts (Rīga, Latvija).

Projekta laikā paredzēta paralēlā valodas korpusa (latviešu-lietuviešu un lietuviešu-latviešu) un speciālā latgaliešu valodas korpusa izstrāde, kā arī „Lietuviešu-latviešu-latgaliešu valodas leksikona" izveide. Tāpat arī plānota kopīga

doktorantūras programmas baltistikā izstrāde, kas ir viens no kopīgas humanitārās jomas pētniecības infrastruktūras nosacījumiem. Plānotajiem projekta rezultātiem šobrīd nav analogu Latvijā un Lietuvā. Projekta inovatīvais raksturs ir saistīts ar Latvijā maz aprobētu pieeju izmantošanu – divu valodu vai valodas paveida (latgaliešu rakstu valodas) tekstu datubāzu veidošanu, datorlingvistikas rīku izmantošanu studiju procesā, pētniecībā, biežāk lietoto vārdu noteikšanu.

Projekts tiek realizēts ar Latvijas–Lietuvas pārrobežu sadarbības programmas 2007.–2013. gadam atbalstu. Projekta laikā ir plānots sasniegt četrus galvenos rezultātus.

Pirmkārt, ir plānots izveidot paralēlo latviešu–lietuviešu un lietuviešu–latviešu tekstu korpusu, kura apjoms ir astoņi miljoni vārdlietojumu.

Otrkārt, ir paredzēts radīt mūsdienu latgaliešu valodas tekstu korpusu, plānotais apjoms – apmēram viens miljons vārdlietojumu. Abi tekstu korpusi būs pieejami pētnieciskiem mērķiem, projekta partneru mājaslapās (*www.ru.lv, donelaitis.vdu.lt, lumii.lv*) izvietojot saiti uz tiem.

Treškārt, lai saglabātu un pētītu baltu valodas un paplašinātu to funkcionalitāti, tiks veikts izstrādāts „Lietuviešu-latviešu-latgaliešu valodas leksikons", kas ietvers aptuveni 10 tūkstošus biežāk rakstos lietoto vārdu.

Ceturtkārt, ir paredzēts organizēt starpdisciplināras doktorantūras līmeņa studiju programmas izveidi baltistikā, kas būs pirmā šāda veida programma Latvijā. Tāpat arī tiks nodrošināta Rēzeknes Augstskolas realizētās maģistra studiju programmas moduļu papildināšana ar īpašiem kursiem datorlingvistikā, kam tiktu nodrošināta pēctecība arī doktorantūras programmā, sagatavojot augsta līmeņa speciālistus ar datorlingvistikas un citām moderno tehnoloģiju prasmēm.

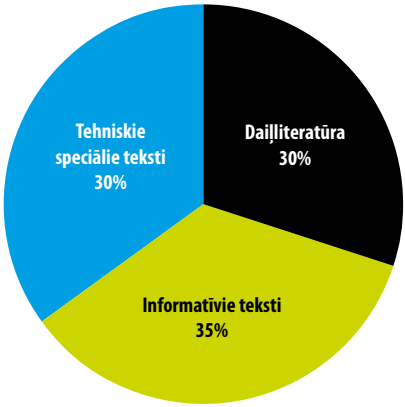
Projekts paredz arī daudzveidīgas aktivitātes, lai sasniegtu iepriekš minētos rezultātus. „HipiLatLit" laikā notiks divas starptautiskas konferences: 2011. gada rudenī Kauņā „Datorlingvistikas iespējas humanitārajā izglītībā un pētniecībā" un 2012. gada rudenī Rēzeknē projekta noslēguma konference „Humanitāro zinātņu pētniecības infrastruktūras izveide Latvijas un Lietuvas austrumu pierobežā". Dažādos pieredzes apmaiņas braucienos iesaistīti studenti, docētāji un pētnieki, piemēram, lai iepazītu speciālo un paralēlo korpusu veidošanas specifiku, projekta darbinieki konsultējās ar Tartu Universitātes un Kārļa Universitātes Prāgā speciālistiem. Savukārt no 2011. gada novembra līdz 2012. gada augustam Kauņā un Rēzeknē plānots apmācīt trīs mērķa grupas – studentus, pētniekus un skolotājus par korpusu lietošanu. Mācības būs atvērtas arī interesentiem (tulkiem, mediju pārstāvjiem u. c.). Tādā veidā tiks nodrošināta arī individuālo pētījumu kvalitāte, pētījumu pieejamība plašam lietotāju lokam, kā arī sekmēta projekta

rezultātu atpazīstamība sabiedrībā. Kā viena no darba metodēm ir ekspedīcijas pētījuma „Lietuviešu-latviešu-latgaliešu valodas leksikons" datu vākšanai. Pirmā ekspedīcija 2011. gada jūlijā notika Lietuvas un Latvijas pierobežā, Biržu rajonā, bet otrā – 2012. gada jūnijā Latgalē, piedaloties gan Rēzeknes Augstskolas, gan Vītauta Dižā Universitātes studentiem un pētniekiem.

Nepieciešamību izstrādāt „HipiLatLit" speciālo (latgaliešu) valodas tekstu korpusu pamato vairāki argumenti. Šāds korpuss ir nozīmīgs avots valodas pētījumiem. Tas noder mācību līdzekļu veidošanā, valodas mācīšanā, informācijas izguvē, tulkošanā, leksikogrāfijā, terminoloģijas veidošanā.

Latgaliešu valodas tekstu korpusa izveidē tiek ņemta vērā arī latviešu valodas tekstu korpusa izveides pieredze, ko veiksmīgi turpina projekta sadarbības partneris LU aģentūra „Latvijas Universitātes Matemātikas un informātikas institūts". Izstrādājot latgaliešu valodas tekstu korpusa proporcijas, pētnieki ir saskārušies ar situāciju, ka valodas lietojuma sfēru dažādība ir ierobežotāka salīdzinājumā ar latviešu valodas tekstiem. Meklējot atbilstošāko risinājumu un ņemot vērā nosacījumu, ka arī latgaliešu valodas tekstu korpusam svarīgi ir pārstāvēt dažādas valodas lietojuma sfēras, tika izdalīti trīs tekstu veidi (sk. 1. attēlu). Pirmkārt, **informatīvie teksti**, kas veidotu 35 % no kopējā tekstu apjoma, otrkārt, **tehniski speciālie teksti** 35 %, treškārt, **daiļliteratūra** 30 %. Pie informatīviem tekstiem pieder raksti laikrakstos, avīzēs, žurnālos, interneta portālos, kā arī ziņas un ceļojumu apraksti, kas ir plaši izplatīts žanrs latgaliešu vidū, it sevišķi pateicoties interneta emuāru popularitātes pieaugumam. Tehniski speciālie teksti ietver zinātniskos rakstus, reliģiskos tekstus un mācību līdzekļus. Daiļliteratūras tekstu kopā tiks ietverta latgaliski publicētā proza un dzeja.

1. attēls. **Latgaliešu valodas tekstu korpusa proporcijas**



Projekta „HipiLatLit" laikā, veidojot latgaliešu valodas tekstu korpusu, tiek ņemti vērā trīs kritēriji:

1) korpusā tiek iekļauti teksti latgaliski, kas izdoti Latvijā, sākot ar 1989. gadu;
2) tekstu datubāzē tiek iekļauti pilni teksti;
3) tekstiem jābūt autentiskiem – tādiem, kādi tie ir publicēti, bez labojumiem, pat ja ir ieviesušās drukas kļūdas.

Strādājot pie latgaliešu tekstu korpusa veidošanas, ir konstatētas vairākas problēmas, pirmkārt, līdz šim nav radīta visu latgaliski izdoto tekstu bibliogrāfija. Tādējādi nav iespējams apzināt visus latgaliski publicētos tekstus. Vienotas bibliogrāfijas izveidi kavē fakts, ka daļa literatūras izdota ārpus Latvijas – trimdā dažādās valstīs. Tas ierobežo arī dažādu izdevumu pieejamību Latvijā, un plašākai sabiedrībai ir liegta iespēja iepazīties ar izciliem latgaliešu rakstnieku darbiem. Līdz ar to viens no risinājumiem ir šo tekstu digitalizācija.

Otrkārt, latgaliešu tekstu tehniskā atpazīšana lielā mērā tiek veikta manuāli. Lai arī Rēzeknes Augstskolas brīvprātīgie studenti izmanto programmu *ABBY Fine Reader* tekstu skenēšanai, tā sniedz tikai daļējas iespējas tekstu atpazīšanai, jo atpazīst latviešu valodu. Visi teksti tiek pārbaudīti, salīdzinot grāmatas tekstu un elektronisko tekstu, un laboti manuāli. Lai atvieglotu šo darbu, ir jārada vārdu un gramatikas pārbaudītājs latgaliski, tomēr arī šis ir sarežģīts uzdevums, jo latgaliešu rakstības tradīcijas krietni atšķiras.

Projektā „HipiLatLit", sekojot līdzi jaunākajām tendencēm tehnoloģiju attīstībā un valodu pētniecībā pasaulē, ir iespējams modernizēt humanitāro zinātņu jomu arī lokālā, reģionālā līmenī, saskatot vērtības, kas mums dotas tepat līdzās. Pat sākot ar literāru tekstu digitalizēšanu un tai sekojošu datubāzu veidošanu, ir iespējams vēl vairāk apzināties un saglabāt savu reģionālo, teritoriālo identitāti un kultūru daudzveidību un nodrošināt nozīmīgu avotu turpmākiem pētījumiem, vārdnīcām, mācību līdzekļiem, kā arī sniegt ierosmi jaunu ideju realizēšanā.

Literatūra

1. *Valodniecības pamatterminu skaidrojošā vārdnīca*. Atbildīgā redaktore V. Skujiņa. Rīga : LU Latviešu valodas institūts, LVA, 2007.
2. Latviešu valodas korpusa koncepcija. [tiešsaiste] Rīga : Latvijas Universitātes Matemātikas un informātikas institūts, 2005. Pieejams: *www.korpuss.lv*

Guna Rābante

Runas korpuss: izveide un izmantošana

Pētot valodu, ļoti palīdz jau iepriekš elektroniskā formā apkopoti dati. Īpaši svarīgi tas ir valodas un ar to saistītu jomu pētījumos. Pētniekam pašam meklēt zinātniskajam darbam nepieciešamo materiālu ir ļoti laikietilpīgi. Nekad iepriekš nav zināms, vai izvēlētajā grāmatā vai periodiskajā izdevumā būs pētniekam vajadzīgie materiāli. Lai atvieglotu meklēšanu, ir izveidots „Līdzsvarots mūsdienu latviešu valodas tekstu korpuss" (sk. *www.korpuss.lv*). Tas ļauj ielūkoties plašā tekstu krājumā, meklēt tajā interesējošo informāciju, redzēt latviešu valodas attīstības tendences.

Lai arī rakstītās valodas pētniekam atrast materiālus pētījumam nav viegli, vēl grūtāk tos iegūt ir runātās valodas pētniekam. Tas ir viens no iemesliem, kāpēc runātā valoda ir pētīta salīdzinoši maz. Runas ierakstu veikšana dabiskos apstākļos ir sarežģīta. Slepus to darīt nav ētiski. Ja personas zina, ka tiek ierakstītas viņu sarunas, runa nav brīva, tādēļ nav derīga pētījumiem. Tikai tad, kad runātāji ir aizmirsuši, ka viņu balsis tiek ierakstītas, un runā brīvi, ieraksts ir derīgs pētījumiem. Turklāt sarunas tēma, tās maiņa nekad iepriekš nav paredzama, tāpat arī izmantotā leksika, frāzes, izteikumu uzbūve utt. Līdz ar to var būt daudz ierakstīta materiāla, bet maz derīgā pētījumam. Šo runas pētniekiem tik nozīmīgo problēmu varētu palīdzēt atrisināt runas korpuss.

Runas korpusa izveide

Runas korpusa pamatā ir liels krājums transkribētu audioierakstu. Tiem ir jābūt ļoti labas kvalitātes. Jo specifiskāks runas korpuss, jo labākas kvalitātes ierakstiem jābūt, piemēram, leksikas pētījumiem fona troksnis netraucē, bet fonētikas pētījumiem tas nav pieļaujams, jo traucē skaņu akustiskajai analīzei.

Metadatu pievienošana

Veicot audioierakstus, jāapkopo arī metadati, t. i., dati par datiem: informācija par runātājiem, ierakstu u. tml. Tas ir svarīgi vēlākajiem pētījumiem. Minimālais metadatu apjoms par runātājiem ietver ziņas par cilvēka dzimumu, vecumu un dzimto valodu. Šo sarakstu iespējams papildināt ar ziņām par dzīvesvietu, nodarbošanos, vaļaspriekiem, izglītību, zināmajām svešvalodām un citiem datiem, kas var palīdzēt saprast cilvēka valodas lietojuma nianses.

Paralēli jāaizpilda arī ieraksta protokols. Jānorāda datums un vieta, kur ieraksts veikts. Jāpieraksta runātājam piešķirtais identifikācijas kods. Svarīgi ir atzīmēt, kurš ierakstu ir veicis un kas to transkribējis.

Datu marķēšana

Runas korpusu var sākt izmantot, ja audiofaili ir transkribēti, t. i., precīzi pierakstīti. Vispiemērotākā ir ortogrāfiskā transkripcija – tieša teiktā atveide, ievērojot ortogrāfijas likumus.

Transkripcijā norāda runātājus pēc to identifikācijas koda.

Ortogrāfiskajai tranksripcijai ir vairāki līmeņi. Vienkāršākajā variantā tiek pierakstīti tikai vārdi, norādot runai raksturīgās pazīmes. Tātad norāda arī nepabeigtos vārdus, nepareizi vai specifiski izrunātos vārdus. Kā pieturzīmes ieteicams lietot tikai punktus, jo runā teikumu uzbūve ievērojami atšķiras no rakstītajiem tekstiem, tāpēc bieži vien nav iespējams noteikt, vai komats ir vai nav nepieciešams.

Sarežģītākajā līmenī apzīmē visas dzirdamās skaņas un precizē vārdu izrunu un nozīmi, ja tie atšķiras no normētās valodas.

Apzīmē ar runātāju nesaistītās skaņas jeb fona troksni. Tas var būt gan vienreizējs, gan ilgstošs. Ilgstošs fona troksnis ir, piemēram, citas sarunas, automašīnas rūkoņa, radio, TV u. c. Vienreizējie fona trokšņi bieži vien nav identificējami vai arī ir pārāk dažādi, lai tos būtu vērts precizēt. Tiem lieto kopēju apzīmējumu: <troksnis> (sk. 2. tabulu).

Norāda ar runātāju saistītās skaņas, piemēram, klepu, šķavas, smieklus, ieelpu vai izelpu.

Nosaka pauzes un to garumu. Pauzes ir divu veidu – aizpildītās un neaizpildītās. Neaizpildītās ir klusums, aizpildītās ir runātāja vilkta skaņa bez semantiskas nozīmes, piemēram, (*āā*), (*ēē*), (*mm*) un citas, kas palīdz iegūt laiku nākamā izteikuma precizēšanai. (Auziņa 2011)

1. tabula. **Ortogrāfiskās transkripcijas līmeņi**

Audio	
Ortogrāfiskā transkripcija	**zin kā pieauguši cilvēki pieauguši cilvēki jau tā vairs nedara.**
Ortogrāfiskā transkripcija ar runai raksturīgām pazīmēm	**zin <zini> kā .hh hh. pieauguši cilvēki (.) @@ <@>pieauguši </@> cilvēki jau tā vairs nedara.**

2. tabula. **Papildu apzīmējumi ortogrāfiskajā transkripcijā**

Apzīmējums	Skaidrojums
<teksts>	Vārda literārā forma vai nepabeigta vārda pilnā forma
.hh	Ieelpa
hh.	Izelpa
(.)	Īsa, neaizpildīta pauze
(āā) (ēē) (mm)	Aizpildīta pauze
@@@	Smiekli
<@>teksts<@>	Smejoties izrunāts viens vai vairāki vārdi
<troksnis>	Neidentificēts troksnis audioierakstā
<klepus>	Klepošana
<šķavas>	Šķaudīšana
<dz>teksts<dz>	Sacītais tiek izdziedāts
<hh.>teksts<hh.>	Teksts sacīts izelpojot
{teksts}	Neskaidra runa
{----}	Nesaprotama runa
[teks]ts	Vienlaicīga runa

Transkripcijā tiek parādīts, piemēram, vai teksts tiek runāts dziedot, smejoties vai izelpojot. Tiek norādīti arī neskaidri izrunāti vārdi vai frāzes, liekot neskaidro tekstu figūriekavās. Iezīmē nepabeigtos vārdus, nepareizi izrunātos vārdus, pēc tiem norādot pareizo variantu. Šajā grupā iekļaujas arī vārdi, kas izrunāti neprecīzi runas defektu dēļ.

Norāda vienlaicīgu runu, ja divi vai vairāki runātāji izsakās reizē.

Morfoloģiskā marķēšana

Nākamais solis ir transkripcijas marķēšana. Morfoloģiskajā marķējumā katrs vārds tiek analizēts, balstoties uz morfoloģisko pazīmju kopu. Sintaktiskajā marķējumā tiek norādīta katra vārda loma teikumā. Semantiskajā marķējumā nosaka vārdu nozīmes. Šie transkripcijas marķējumi daļēji sakrīt ar tekstu marķējumiem valodas korpusā.

Ir daži marķējumi, kas tekstu korpusā nav iespējami. Tie ir fonētiskais un prosodiskais marķējums. Fonētiskajā apzīmē katru skaņu. Prosodiskajā marķējumā norāda intonācijas. Tās ir divu veidu – zilbes un runas jeb teikuma intonācijas (sk. 3. tabulu).

3. tabula. **Runas datu aprakstīšanas līmeņi (Auziņa 2011, 18)**

RUNA				
	5.	Semantiskais marķējums		
	4.	Sintaktiskais marķējums		
	3.	Morfoloģiskais marķējums		
		Fonētiskā transkripcija	Prosodiskais marķējums	
	2.	Metadati	Loģiskā struktūra	Reprezentācijas marķējums
	1.	Ortogrāfiskā transkripcija		
	0.	Audiodatnes		

Runas korpusa izmantošana

No runas korpusa iegūtie dati ir dažādi. Tāds pats ir arī to izmantojums. Runas korpusu pētījumiem var izmantot gan sociālajās, gan humanitārajās zinātnēs. Visplašāk runas korpusu var izmantot valodniecībā: leksikoloģijā var pētīt leksiku, vārdu nozīmes; sociolingvistikā – iegūt informāciju par valodas lietojumu dažādās sociālajās grupās, vecuma grupās; sintaksē un morfoloģijā pētīt valodas sistēmu runā, atšķirības un līdzības ar rakstu valodu teikumu uzbūvē un vārdu darināšanā.

Lai pētītu skaņas ir nepieciešams specializēts runas korpuss. Tajā ir jābūt ļoti labas kvalitātes audioierakstiem bez fona trokšņiem. Vislabāk šos ierakstus veikt kādā ierakstu studijā. Šajā korpusā audiofails ir jātranskribē gan ortogrāfiskajā transkripcijā, gan *SAMPA* (*Speech Assessment Methods Phonetic Alphabet*) mašīnlasāmajā pierakstā, kura simboli atbilst starptautiskā fonētiskā alfabēta (*IPA*, *International Phonetic Alphabet*) simboliem, jāapzīmē katra skaņa.

Runas korpusu var izmantot arī pedagoģijā. Noskaidrojot biežāk lietotās frāzes dažādās situācijās, var veidot mācību līdzekļus cittautiešiem. Izmantojot korpusā iekļautos datus, ir iespējams izveidot vārdu un vārdformu biežuma sarakstu. Tas ir noderīgs, atlasot mācību līdzekļos ietveramo leksiku.

Izveidojot specializētu bērnu runas korpusu, var sekot bērnu valodas attīstībai, noskaidrot, cik gadu vecumā bērns sāk lietot konkrētus jēdzienus, sintaktiskās konstrukcijas, frazeoloģismus u. c.

Runas korpuss ir izmantojams arī reklāmas jomas pārstāvjiem, analizējot, kuras frāzes ir iedarbīgākās cilvēka pārliecināšanai, kuras tiek uztvertas ar nepatiku.

Tā kā runas korpuss atspoguļo visas latviešu valodas izpausmes, tas ietver arī cittautiešu runu. Līdz ar to var analizēt biežāk pieļautās kļūdas, pētīt, kuri vārdi

parādās biežāk, kuri retāk, kādas runātāja dzimtās valodas iezīmes visbiežāk saklausāmas latviešu valodā.

Iegūtos datus var izmantot vārdnīcu un mācību līdzekļu sastādīšanā, lai valodas mācīšanu un mācīšanos padarītu efektīvāku. Tāpat arī runas korpuss nepieciešams runas sintēzes un analīzes programmu izveidei un uzlabošanai, tā nodrošinot iespēju pilnīgāk izmantot latviešu valodu virtuālajā vidē.

Secinājumi

Runas korpusa izveidei ir nepieciešami lieli cilvēkresursi un daudz laika. Ļoti svarīga ir uzmanība un precizitāte. Īpaši tas attiecas uz runas korpusiem, kuros dati ir pierakstīti fonētiskajā transkripcijā. Tomēr ieguvumi ir daudz lielāki nekā ieguldītais darbs.

Izveidojot pilnīgu un sabalansētu runas korpusu, zinātnei pavērtos daudz plašākas iespējas. Sagatavošanās pētījumiem neaizņemtu tik daudz laika, līdz ar to pētījuma rezultāti un galaprodukts pie lietotājiem nonāktu īsākā laika posmā.

Literatūra

1. Auziņa, Ilze. Runas datu transkribēšana un marķēšana: problēmas un risinājumi. No: *Vārds un tā pētīšanas aspekti*-15 (1). Liepāja: LiepU, 2011, 18 lpp.
2. *Līdzsvarots mūsdienu latviešu valodas korpuss* [tiešsaiste]. Pieejams: www.korpuss.lv

Līga Vogina

Atgriezeniskie darbības vārdi un teikuma semantiskās lomas: latviešu valodas korpusa pieredze

Atgriezeniskie jeb refleksīvie darbības vārdi aplūkoti ikvienā latviešu valodas gramatikā. Kaut arī tiem ir viens kopīgs formāls rādītājs (atgriezeniskā galotne), tie veido neviendabīgu vārdu grupu. Emīlija Soida norādījusi, ka atgriezenisko darbības vārdu grupai „veidojas vairākas spilgtas pazīmes, kas var noderēt, izšķirot jautājumu par šās verbu grupas vietu latviešu valodas verbu sistēmā" (Soida 2009, 214). Šīs darbības vārdu grupas neviendabīgumu pamanījuši jau pirmo gramatiku autori. E. Soida analizējusi pirmajās latviešu valodas gramatikās izteiktos viedokļus par atgriezeniskajiem darbības vārdiem un secinājusi, ka „gramatiku autori jautājumam par atgriezeniskajiem verbiem katrs izveidojis savu risinājumu un novērojuši, kā arī uzsvēruši dažādas latviešu valodas atgriezeniskā verba pazīmes" (Soida 2009, 210). Arī Dzintra Paegle norādījusi, ka atgriezeniskums valodā izpaužas neviendabīgi (Paegle 2003, 129). Tas, ka šī darbības vārdu grupa nav tik vienkārši analizējama, atspoguļots arī „Mūsdienu latviešu literārās valodas gramatikā", kur sacīts, ka „latviešu valodā ir daudz tādu refleksīvu verbu, kas paliek ārpus apskatītā kārtas kategorijas jēdziena" (Mllvg 1959, 561). Minētie novērojumi atspoguļo temata problemātiku – latviešu valodniecībā atgriezeniskie darbības vārdi raksturoti katrā gramatikā, taču līdz šim mēģinājumi tos sistematizēt nesuši līdzi atziņu, ka daļa šo darbības vārdu paliek ārpus piedāvātās sistēmas.

Aplūkojot teorētiskajā literatūrā sniegtos atgriezenisko darbības vārdu piemērus, jāsecina, ka tie lielākoties neatklāj plašāku kontekstu, kurā šie verbi tiek lietoti. Visas atgriezenisko darbības vārdu nozīmes apvienotas vienkopus, sīkāk neanalizējot, kādi semantiskie aktanti ir katra atgriezeniskā darbības vārda argumentu struktūrā. Ekscerpējot valodas materiālu klasiskā veidā (izrakstot piemērus no daiļliteratūras vai publicistikas), jārēķinās, ka ne vienmēr iespējams atrast piemērus, kas atšķirtos no gramatikās tradicionāli minētajiem. Tomēr tas nenoliedz šādu piemēru esamību. Lai iegūtu pilnīgāku priekšstatu par valodu, lietderīgi izmantot mūsdienīgus resursus, piemēram, latviešu valodas tekstu korpusu. Tekstu korpusā iespējams pētīt atgriezenisko darbības vārdu nozīmes plašākā kontekstā un dabiskā, nepārveidotā valodas materiālā, tādējādi ir iespējams nošķirt visas nozīmes, kas piemīt atgriezeniskajiem darbības vārdiem. Atgriezenisko darbības vārdu klasifikācijai jāizmanto semantisko lomu teorija. Tieši darbības vārda apkaime ļauj šķirt subjekta, objekta un nepersonas verbus.

Šī raksta mērķis nav pārskatīt atgriezenisko darbības vārdu vietu latviešu valodas sistēmā, bet gan ieskicēt, kuras teorētiskās atziņas nepilnīgi atklāj latviešu valodas lietojumu un būtu papildināmas ar tekstu korpusa materiāliem.

Atgriezeniskie darbības vārdi ir darbības vārdi, kam „ir atgriezeniskais formants un kam vēsturiski, iespējams, bijusi vidējās kārtas nozīme, piemēram, *mazgāties*, *slaucīties*, *celties*" (VPSV 2007, 52). Vidējā kārta ir „darbības vārda kārta, kas izsaka, ka darbības subjekts reizē ir arī darbības objekts" (VPSV 2007, 443). Mūsdienu latviešu valodā vidējās kārtas nav, tomēr tās nozīmi izsaka daļa atgriezenisko darbības vārdu, piemēram, *es mazgājos* 'es mazgāju sevi' vai *tu ģērbies* 'tu ģērb sevi'. „Mūsdienu latviešu literārās valodas gramatikā" šādi atgriezeniskie darbības vārdi saukti par tieši atgriezeniskiem, kas „vienā formā apvieno subjekta un objekta nozīmi: darbība izriet no subjekta un vērsta uz objektu, bet objekts ir pats subjekts. [..] No darbības virziena viedokļa verba formā *mazgājos* apvienojas divi pretēji darbības virzieni: *mazgāju (sevi)* un *tieku mazgāts* [..]" (Mllvg 1959, 556). Atgriezeniskais formants norāda darbības vērsumu no subjekta uz objektu un otrādi. Tas ir „afikss vai vietniekvārds, kas sākotnēji norādījis, ka darbības vēršas atpakaļ uz darītāju" (VPSV 2007, 52). Vidējās kārtas nozīme ir atgriezenisko darbības vārdu prototipiskā nozīme, taču „Mūsdienu latviešu literārās valodas gramatikā" norādīts, ka „šī tipiskā refleksīvo verbu kārtas nozīme tomēr nav produktīva, verbu ar šādu subjekta un objekta identitāti nav daudz" (Mllvg 1959, 556). Gramatikā minēti vien daži konkrētas nozīmes verbi: *mazgāties*, *ģērbties*, *apsegties*, *rotāties* (= *mazgāt sevi*, *ģērbt sevi*, *apsegt sevi*, *rotāt sevi*), kā arī daži abstraktākas nozīmes darbības vārdi: *aizbildināties*, *attaisnoties*, *glābties*, *izlutināties*, *izrādīties*, *izglītoties*, *saukties* (sk. Mllvg 1959, 556). Lai arī teorijā norādīts, ka atgriezenisko darbības vārdu, kuriem subjekts un objekts sakrīt, nav daudz, līdzsvarotā mūsdienu latviešu valodas tekstu korpusa materiālos iespējams bez jau minētajiem darbības vārdiem atrast vēl citus atgriezeniskos darbības vārdus, kuriem ir prototipiska (vidējās kārtas) nozīme:

Pirmajās dienās centušies paši savām rokām ***ieziesties*** *ar aizsargkrēmu.* (timeklis-1.0)

Ikreiz biezi un izšķērdīgi ***ieziedās*** *ar Nivea krēmu.* (timeklis-1.0)

Lote ***ietinās*** *caurspīdīgā lakatā.* (miljons-2.0m)

Mani savāks vīri ķiteļos un palīdzēs ***ieģērbties*** *tai halātā ar tām garajām rokām.* (timeklis-1.0)

Vairums komiksu liek varonim ***ieģērbties*** *zaļās triko biksēs.* (timeklis-1.0)

Minētie piemēri atklāj, ka darbības darītājs un darbības objekts jeb cietējs sakrīt, tātad tie ir koreferenti.

Atsevišķi būtu nošķirami tādi atgriezeniskie darbības vārdi, kuriem darbības objekts jeb cietējs sakrīt ar darītāja ķermeņa daļu:

(Es) ***skujos*** *arī tikai katru otro dienu.* (miljons-2.0m)

Te aizmirstu ***saķemmēties****, te paēst, te pienācīgi apkārtējai temperatūrai saģērbties.* (timeklis-1.0)

Ikdienā Irmgarde ***krāsojas*** *pavisam drusciņ, pavisam atturīgi – pensionārei nepiedienas ne savus gadus, ne savu necilo sociālo stāvokli citiem bāzt acīs.* (miljons-2.0m)

Sievietes vispār ***krāsojas****, lai patiktu sev, apkārtējiem un būtu interesantas pretējam dzimumam.* (miljons-2.0m)

Jāteic, ka šāda atgriezenisko darbības vārdu grupa „Mūsdienu latviešu literārās valodas gramatikā" vispār nav minēta. Savukārt E. Soidas darbā „Vārddarināšana" šāda grupa ir minēta un tajā ierindoti arī darbības vārdi *grimēties*, *pūderēties*, *skrullēties*, *smiņķēties* (Soida 2009, 214). Valodniece norādījusi, ka „refleksīvais elements tikai norāda, ka darbība, kaut arī neaptver visu tās veicēju, tomēr neskar citus objektus, bet paliek darbības veicēja sfērā un risinās viņa labā" (Soida 2009, 214).

Nākamā atgriezenisko darbības vārdu nozīme, kas teorētiskajā literatūrā tiek minēta, ir savstarpējas darbības jeb reciprokas darbības nozīme. Savstarpējas darbības vārdi ir darbības vārdi, kas „izsaka divu vai vairāku subjektu savstarpēju darbību, piemēram, *apspriesties*, *cīnīties*, *sarunāties*. Latviešu valodā savstarpējas darbības vārdi parasti ir atgriezeniski" (VPSV 2007, 344). „Mūsdienu latviešu literārās valodas gramatikā" savstarpējas nozīmes darbības vārdi raksturoti kā darbības vārdi, kas „izsaka, ka viena un tā pati darbība, iziedama no diviem vai vairākiem darītājiem, pamīšus pāriet uz pašiem darītājiem. Katrs darbības subjekts pamīšus ir arī darbības objekts" (Mllvg 1959, 558). Savstarpējas darbības nozīme īsi pieminēta Brigitas un Laimdota Ceplīšu „Latviešu valodas praktiskajā gramatikā": „atgriezeniskie darbības vārdi norāda, ka darbība vēršas uz pašu darītāju vai arī ka tā ir savstarpēja" (Ceplīte, Ceplītis 1997, 64). Dz. Paegle norāda, ka savstarpējas darbības vārdiem subjekta un objekta nozīme netiek diferencēta, piemēram, *Kaimiņi sarunājās arvien skaļāk* vai *Viņi apkampjas un aiziet tumsā*, izņēmums ir gadījumi, kad atgriezeniskais darbības vārds savstarpējas darbības nozīmē piesaista substantīvu instrumentālī. Tādā gadījumā subjekta un objekta nozīme tiek šķirta (Paegle 2003, 130), piemēram, *Es negribu sacensties* ***ar draudzeni***. Šādu piemēru sakarā E. Soida uzsvērusi, ka „subjekts pie šiem verbiem lietojams daudzskaitļa formā [..]" (Soida 2009, 215). Mūsdienu latviešu valodas tekstu korpusa

materiālos tik tiešām ir atrodami vairāki piemēri, kuros redzami „īstie" savstarpējas darbības vārdi (tātad teorētiskās atziņas šajā ziņā ir pamatotas):

*Četri brāļi(.) joprojām cenšas **satiekties** pie mammas.* (miljons-2.0m)

*Pēc vakariņām dzeram jasmīnu tēju un **sarunājamies**.* (miljons-2.0m)

*Mēs **sarunājamies** par pilsētu, kurā esam uzauguši.* (miljons-2.0m)

*Kad Berluskoni tika iecelts par Itālijas premjeru, daudzi politiķi nevēlējās ar viņu **sarokoties**.* (miljons-2.0m)

*Mēs ar Leopoldu **apkampjamies**, tad viņš nolec no prāmja un kāpj stāvajā krastā.* (timeklis-1.0)

Minētajos piemēros arī redzams, ka subjekts ir daudzskaitļa formā. Taču līdzās iepriekšējiem piemēriem noteikti jāmin tādi piemēri, kuros subjekts ir vienskaitlī:

*Man nepatīk pukstēt un **rāties**.* (timeklis-1.0)

*Bērns **spārdās** un kviec kā sivēns.* (timeklis-1.0)

*Liepājas domes sabiedrisko attiecību speciālists Aigars Štāls **rājas**, ka pilsēta neesot paredzēta tikai šoferiem.* (miljons-2.0m)

*Gavēnī jādzīvo klusu un godīgi, nedrīkst kliegt, svilpot, **mēdīties** un citādi palaidņoties.* (timeklis-1.0)

Redzams, ka, pretēji norādēm teorētiskajā literatūrā par vairākiem darītājiem, kas vienlaikus ir arī otra darītāja veiktās darbības cietēji, minētajos teikumos ir tikai viens darītājs. Tieši korpusa materiālā iespējams nošķirt „īstas" savstarpējas darbības situācijas no tādām, kurās ir vispārināts otrs darbības darītājs, kas vienlaikus ir arī vispārināts cietējs.

Teorētiskajā literatūrā norādīts uz to, ka atgriezeniskie darbības vārdi lietojami tikai darāmajā kārtā. Tā, piemēram, Dz. Paegle raksta: „Latviešu valodai ir raksturīgi, ka vairums refleksīvo verbu lietojami tikai aktīva nozīmē, t. i., darbības veicējs ir dzīva būtne." (Paegle 2003, 129). „Mūsdienu latviešu literārās valodas gramatikā" uzsvērts, ka „ar verbu refleksīvajā formā retumis izsaka darbību ar pasīvu nozīmi, piemēram, *nauda glabājas [tiek glabāta] bankā*, *muzejā krājas [tiek krātas] vērtības*, *aparāts nolietojas [tiek nolietots]*. Šīs formas uztveramas arī kā norise bez subjekta aktivitātes. [..] Refleksīvās formas tīra pasīva nozīmē latviešu valodā nav parastas. Neliterāri ir vienkāršrunā retumis sastopamie darinājumi *kaut kas dzirdas (dzirdams)*, *redzas*

(redzams), *romāns viegli lasās (lasāms)*, *vārds lokās (lokāms)* u. tml." (Mllvg 1959, 560–561). Tāpat B. un L. Ceplīšu gramatikā norādīts, ka teikumi *biļetes pārdodas kasē*, *izrāde skatās patīkami* vai *vienošanās izpildās veiksmīgi* latviešu valodā nav pieļaujami (sk. Ceplīte, Ceplītis 1997, 65). Gramatikas autori šādu apgalvojumu pamato, norādot, ka atgriezenisko darbības vārdu nevar lietot tiešā darbības vārda vietā situācijās, kurās darītāja aktivitāti nevar samazināt (tātad pareizi būtu sacīt *biļetes tiek pārdotas kasē* vai *biļetes pārdod kasē*, bet ne *biļetes pārdodas kasē*) (Ceplīte, Ceplītis 1997, 65). Arī skolu gramatikā uzsvērts, ka atgriezeniskos darbības vārdus attiecina uz dzīvu, rīkoties spējīgu darītāju, par piemēru minot, ka mēs varam rakstīt domrakstu, bet domraksts pats nerakstās (Latviešu valoda 1998, 191).

Tomēr, aplūkojot piemērus korpusā, jāsecina, ka ciešamās kārtas nozīmē lietotu atgriezenisko darbības vārdu skaits pieaug. Spilgtākie no piemēriem ir šādi:

*Vienkārši tikai pasaciņa **stāstās**.* (miljons-2.0m)

*Cilvēka vārds **rakstās** ar lielo burtu.* (timeklis-1.0)

*Šie diski **pārdodas** iepakoti.* (timeklis-1.0)

*No valsts budžeta piešķirtie līdzekļi **glabājas** Valsts kasē.* (miljons-2.0m)

Minētajos piemēros var domāt par teikuma sintaktiskajā struktūrā nesvarīgu darītāju, tādēļ runātājs to neizsaka gramatiski.

*Cienījamie deputāti! Kamēr **drukājas** rezultāti… Māris Rudzītis vēlas sniegt arī paziņojumu.* (saeima-2.0)

*Man tagad sapņos **redzas** vairs tikai tavas kurpju šņores.* (timeklis-1.0)

Savukārt šajos teikumos runātājs gribējis norādīt, ka darbība norit pati par sevi, ka to nav iespējams ietekmēt – drukāšanas ātrums nav atkarīgs no cilvēka, kurš vēlas izdrukāt rezultātus, tāpat arī sapņus nav iespējams ietekmēt.

Raksta noslēgumā jāsecina, ka teorētiskajā literatūrā visbiežāk aplūkotas tās atgriezenisko darbības vārdu nozīmes, kuras ir nepārprotamas un samērā viegli nošķiramas. Pārējās nozīmes tiek noklusētas vai raksturotas ļoti skopi un nekonsekventi. Visbiežāk minēta atgriezenisko darbības vārdu prototipiskā nozīme un savstarpējas darbības nozīme, reizēm arī darbības nejaušības nozīme. Nav minēta, piemēram, pārmērīgi veiktas darbības nozīme, ko iespējams konstatēt korpusa materiālos:

*Latvieši **sadzērās** tarhūnu.* (timeklis-1.0)

*Viņš **saēdās** čūskogas.* (timeklis-1.0)

Turpretī ciešamās kārtas nozīme tiek nostumta perifērijā, tai parasti tiek atvēlēts viens teikums vai īsa rindkopa. Visbiežāk tiek norādīts, ka šāds darbības vārdu lietojums latviešu valodā nav visai raksturīgs vai nav vēlams. Tomēr dzīvajā valodā šādu piemēru skaits palielinās, tādēļ atgriezeniskie darbības vārdi analizējami semantisko lomu teorijas aspektā. Tā ļauj sasaistīt teikumā nosauktās situācijas dalībniekus un to lomas ar teikuma sintaktisko struktūru. Savukārt mūsdienu latviešu valodas korpuss sniedz iespēju ne vien ātri un ērti iegūt datus, bet arī pētīt dzīvu, nesamākslotu un nepārveidotu valodu, tādējādi nodrošinot pilnīgāku valodas izpēti un pārskatot teorētiskās atziņas.

Avoti

timeklis-1.0 – *Latviešu valodas tīmekļa korpuss*. Pieejams: www.korpuss.lv

miljons-2.0m – *Līdzsvarots mūsdienu latviešu tekstu korpuss.* Piejams: www.korpuss.lv

saeima-2.0 – *Latvijas Republikas 5.–9. Saeimas sēžu stenogrammmas*. Pieejams: www.korpuss.lv

Literatūra

1. Ceplīte, B., Ceplītis, L. *Latviešu valodas praktiskā gramatika*. Rīga : Zvaigzne ABC, 1997.
2. *Latviešu valoda 10.–12. klasei.* Autoru kolektīvs. Rīga : Zvaigzne ABC, 1998.
3. MIlvg – *Mūsdienu latviešu literārās valodas gramatika. I. Fonētika un morfoloģija.* Autoru kolektīvs. Rīga : LPSR ZA izdevniecība, 1959.
4. Paegle, Dz. *Mūsdienu latviešu literārās valodas morfoloģija.* 1. daļa. Rīga : Zinātne, 2003.
5. Soida, E. *Vārddarināšana*. Rīga : LU Akadēmiskais apgāds, 2009.
6. VPSV – *Valodniecības pamatterminu skaidrojošā vārdnīca.* Autoru kolektīvs. Rīga, 2007.

Gita Elksnīte

„Latviešu valodas seno tekstu korpusa" izmantojums vārdkopu pētniecībā

Raksta pamatā izmantoti promocijas darbā „Nominālās vārdkopas Georga Manceļa tekstos" (Elksnīte 2011) aprakstītie pētījumi, kuriem materiāls iegūts, frontāli ekscerpējot G. Manceļa vācu-latviešu vārdnīcu „Lettus" (Lettus) un tās pielikumus „Phraseologia Lettica" (PhL) un „Zehen Gespräche Deutsch und Lettisch" (Zehen G). Piemēri no spredikļu grāmatas „Lettische Langgewünschte Postill" (LP1, LP2, LP3) atlasīti „Latviešu valodas seno tekstu korpusā" (turpmāk – „Seno tekstu korpuss"), kas izstrādāts Latvijas Universitātes Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijā.

Promocijas darbā tika analizētas nominālās vārdkopas no vārdnīcas „Lettus" un tās pielikumiem, bet spredikļu grāmatas teksts tika lietots kā salīdzinājums vārdnīcā iegūtajām vienībām. Salīdzinājums bija nepieciešams, jo vārdnīcā nominālās vārdkopas ievietotas pārsvarā bez plašāka konteksta, tādēļ ir lietderīgi redzēt, kā tās funkcionē saistītā tekstā kā teikuma veidotājelementi, kā tās ietekmē teikuma struktūru, kādi teikuma locekļi tās ir. Spredikļu grāmatā var gūt plašāku ieskatu par tiem vārdkopu tipiem, kuri vārdnīcas tekstā lietoti reti, piemēram, adjektīviskās vārdkopas un substantīviskās vārdkopas ar atributīvu divdabi atkarīgajā komponentā (plašāk par to Elksnīte 2011).

Raksta mērķis ir parādīt, kā tika iegūti un apstrādāti piemēri no Georga Manceļa spredikļu grāmatas „Lettische Langgewünschte Postill", izmantojot „Seno tekstu korpusu".

Meklēšanas iespējas „Seno tekstu korpusā"

Lai atrastu nepieciešamo informāciju, „Seno tekstu korpusā" meklēšanai tika izmantoti vairāki lietojumrīki:

1) navigācija korpusa saturā;
2) meklēšana vārdformu indeksā;
3) konkordanču rīks.

Sākumā piemēru atlasei tika izmantots konkordanču rīks, ar kura palīdzību, ierakstot meklēšanas logā vārdformu vai tās daļu un izvēloties nepieciešamos avotus, tika atrastas nominālās vārdkopas (sīkāk par to raksta apakšnodaļās „Adjektīvisko vārdkopu meklēšana" un „Vārdkopu ar atributīvo divdabi meklēšana").

Lietojumrīks „Navigācija korpusa saturā" tika lietots, ja bija nepieciešams pārbaudīt teksta pareizību vai arī noskaidrot plašāku kontekstu. Šo rīku var izmantot tikai tajos gadījumos, ja ir skaidri zināma lappuse un avota nosaukums. Lietojumrīkā „Navigācija korpusa saturā" jāizvēlas vajadzīgais avots, jāatrod nepieciešamā lappuse, jāatver tā un jāmeklē piemērs, atverot meklētāju ar taustiņu kombināciju *ctrl+F* un ierakstot tajā kādu no vārdkopas komponentiem.

Bieži vien plašāks konteksts nepieciešams adjektīviskajām vārdkopām, jo to galvenais komponents pilda izteicēja funkciju teikumā un ir saistīts ar saitiņu, kas ne vienmēr atrodas tam blakus:

Vnd tudeļ ***kļua*** *Wings no śawas Spittalibas* ***schkiests*** .. (LP1 161, 157);

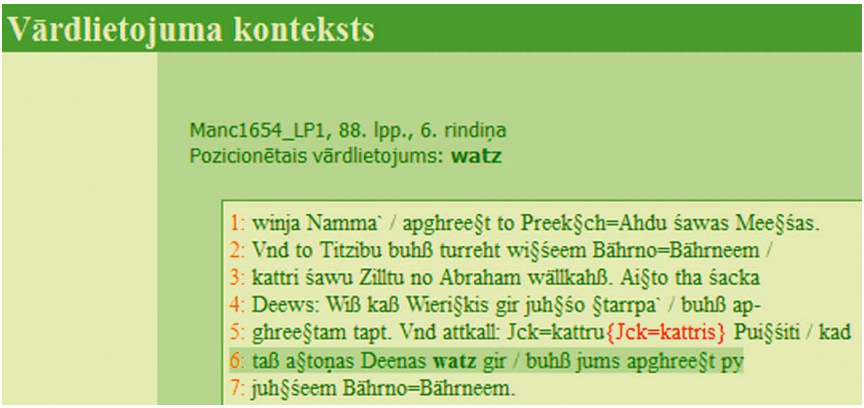
.. *tomähr* ***gir*** *teem ta Śirrds* ***pilla*** *no Blehdibas* .. (LP2 149);

Mehß ***tohpam*** *beß Nopällnu* ***taißni*** *no Deewa Schälastibas* .. (LP2 206).

Lai atrastu lappusi vai avotu konkrētam piemēram, tika izvēlēts lietojumrīks „Meklēšana vārdformu indeksā". Piemēram, lai noskaidrotu lappusi vai avotu teksta fragmentam ***taß astoņas Deenas watz gir***, meklēšanas logā tika ierakstīts vārds ***watz.*** Lai nebūtu jāpārskata liels daudzums lappušu, izraudzīts vārds, kurš varētu būt lietots retāk. Meklēšanas rezultātos parādās adjektīva dažādie rakstības varianti (***watz, wätz, Wätz***), no kuriem tad izvēlas vajadzīgo, spiežot uz krustiņu un pēc tam uz norādīto lappusi un rindkopu (sk. 1. attēlu).

1. attēls. **Vārdformas *watz* meklēšanas rezultāts**

2. attēls. **Vārdformas *watz* lietojuma konteksts**



Šajā gadījumā spredikņu grāmatā konstatēts tikai viens piemērs ar vārdformu ***watz***, taču bieži vien ir jāatver un jāpārbauda vairākas lappuses, lai atrastu meklēto piemēru (sk. 2. attēlu).

Tāpat lietojumrīku „Meklēšana vārdlietojumu indeksā" var izmantot, lai pārbaudītu tekstu vai iegūtu plašāku kontekstu.

Burtu atveide „Seno tekstu korpusā"

Lai pareizi ierakstītu vajadzīgo vārdu *meklētājā*, tika ņemti vērā gan „Seno tekstu korpusa" burtu apzīmējumi, gan T. G. Fennela izdotie G. Manceļa teksti mūsdienām atbilstošā rakstībā (Fennell 1988, 1989).

Iegūtais materiāls bija jāpārveido atbilstoši seno tekstu ortogrāfijas principiem, jo apzīmējumi atšķīrās no oriģināla rakstības:

Śchiß Kunngs gir tuwe py teem / kattrreem śalau§u§chas Śirrdes gir / und palieds teem / kattreem śaśi§ts Prahts gir .. (LP2 274)

*.. taß by śataiśam**a**` Deen**a**` Leeldeen**a**` / leel**a**` Peek=Deen**a**` / §tarrp**a**` tre§§chu unnd śä§tu Stundu / ka taß Śwähts Marcus unnd Śwähtz Jahnis to Laiku währ**a**` jämuśchi ..* (LP3 103)

*Ka aß Na§is muh§śam Pe§titajam Śirrdy ghree§e / ka winjam d§irrdeht by / ka taß Caiwas ar pillu Mutt Winju / to Auxteteizamu Deewa Dählu / kuŗŗa Mutteh muh§cham nhe kahds Willtz attra§ts / nei zittas Blehņas muh§ch**a~** panahktas / wi§śeem d§irrdoht / par Deewa=Saimotaju und Apśmehjehju lammaja ..* (LP3 50)

1. tabula. **Burtu apzīmējumi**

Apzīmējumi „Seno tekstu korpusā”	Apzīmējumi promocijas darbā
a`	à
a~	ā
m~	m
n~	n
o^	ô
u^	û
§	s

*.. tahda Zillwäzigha Mahziba muh**§**śu no**§**ku**m~**uśchu Śirrdi nhe warr ee=preezenaht ..* (LP1 477)

*Es usraughu to Nabbaghu / und kam śalau**§**ieta Śirrds gir / und kaß bie**§**tahs prek**§**ch ma**n~**u Wahrdu ..* (LP1 5)

*Juhß / **o^** Chri**§**titi Deewa Bährni / turreeteeß tahdu Śirrdi / kahda tam Koninjam Dawid by ..* (LP1 66)

„Seno tekstu korpusa” veidotāji izmantojuši paragrāfa zīmi §, lai atveidotu burtu ʃ, ar kuru G. Mancelis apzīmējis fonēmu [s] (Bergmane, Blinkena 1986, 70). Sintakses vēstures pētniecībā „Seno tekstu korpusa” materiālu izmanto dažādi. Kornēlija Pokrotniece (Pokrotniece 2006, 272–280; Pokrotniece 2008, 27–38) un Elga Skrūzmane (Skrūzmane 2009, 307–316) nepārveido piemērus no „Seno tekstu korpusa”. Raksta autore savā promocijas darbā „Nominālās vārdkopas Georga Manceļa tekstos” piemērus no „Seno tekstu korpusa” pārveidojusi atbilstoši seno tekstu ortogrāfijas principiem (sk. 1. tabulu). Visvairāk bija jālabo paragrāfa zīmes, kuras lietotas biežāk nekā citi tehnisku iemeslu dēļ lietotie burtu apzīmējumi.

Adjektīvisko vārdkopu meklēšana

Vispirms, izmantojot T. G. Fennela izdoto G. Manceļa vārdnīcu „Lettus” un tās pielikumu „Phraseologia Lettica” mūsdienām atbilstošā rakstībā (Fennell 1988; Fennell 1989), tika sastādīts adjektīvu reģistrs. Pavisam tika konstatēts aptuveni 300 leksēmu. Tad, izmantojot „Seno tekstu korpusu”, šie adjektīvi tika meklēti spredi̦ku grāmatā „Lettische Langgewünschte Postill”.

Lai atrastu nepieciešamo vārdu, bija jāņem vērā:

1) seno tekstu rakstība;
2) vārda iespējamās formas.

Lai noskaidrotu vārdu rakstību, tika izmantots gan T. G. Fennela darbs, gan paša G. Manceļa teksts vecajā drukā.

Piemēru atlasei tika izmantots konkordanču rīks, ar kura palīdzību, ierakstot meklēšanas logā vārda nemainīgo daļu, liekot aiz tās % zīmi un izvēloties vajadzīgos avotus (Manc1654_LP1, Manc1654_LP2, Manc1654_LP3), tika atrastas visas iespējamās konkordanču rindas ar meklēto vārdformu centrā. Atverot pasvītroto vārdformu, tiek piedāvāts plašāks konteksts, no kura var izkopēt nepieciešamo teksta fragmentu, kā arī lappusi un avota saīsinājumu. Pēc tam izkopētais teksts tiek apstrādāts, nodzēšot rindu numerāciju, izlabojot seno tekstu ortogrāfijas principiem neatbilstošās zīmes, liekot iekavās saīsinājumu un lappuses numuru, pārveidojot tekstu *Times New Roman* fontā, ja nepieciešams, un slīprakstā, trekninot vārdus, kas veido nominālo vārdkopu (sk. 2. tabulu).

Raksta autore kartotēku veidojusi *Microsoft Word* programmā, izveidojot tabulu, kurā ievieto konstatētos piemērus, taču ir iespējams izvēlēties arī citas programmas, piemēram, *Microsoft Excel* vai kādu no datubāzu apstrādes programmām (sk. 3. tabulu).

2. tabula. **Meklētā vārdforma, vaicājums un atrastie piemēri**

meklējamā vārdforma	vaicājums korpusā	Korpusā atrastā konkordance	Pārveidotais teksts
gräßns	*gräßn%*	23: Taß gir §pehziegs ar śawu Ällkonu / unnd ißkai§śa 24: tohß kattri **gräßni** gir śawas Śirrds prahta`. LP2, 84. lpp	*Taß gir spehziegs ar śawu Ällkonu / unnd ißkaiśa tohß kattri **gräßni gir** śawas Śirrds **prahtà**.* (LP2, 84)
pills	*pill%*	Śataiśi tawu Śirrdi / laid tai buht 9: **pilla** Titzibas eek§chan Chri§tum JE§um / **pilla** Mieläſtibas prett 10: Deewu und tawu Tuwaku / **pilla** Deewa Peeluhkśchanas / **pil** 11: **la** Ghaidieśchanas und Pazeeśchanas / **pilla** Dohmaśchanas us 12: to muh§chighu D§iewośchanu. LP1, 505. lpp.	*Śataiśi tawu Śirrdi / laid tai **buht pilla Titzibas** eekschan Christum JEsum / **pilla Mielästibas** prett Deewu und tawu Tuwaku / **pilla** Deewa **Peeluhkśchanas / pilla Ghaidieśchanas** und **Pazeeśchanas / pilla Dohmaśchanas** us to muhschighu Dsiewośchanu.* (LP1 505)
stippris	*§tippr%*	Bett tas Bährns augha / und 18: kļua **§tippris** Gharra` / pills Ghuddribas / und Dee- 19: wa Schäla§tiba by py to. LP1, 73. lpp	Bett tas Bährns augha / und **kļua stippris Gharrà** / pills Ghuddribas / und Deewa Schälastiba by py to. (LP1 73)
nheschehlighs	*%§chehlig%*	Ai§to / kad teem Juddeem tick nhe§chehligha Śirrds by prett 26: to Kunghu JE§um .. LP3, 157. lpp	*Aisto / kad teem Juddeem **tick nheschehligha** Śirrds **by prett to Kunghu JEsum** ..* (LP3 157)

3. tabula. **Kartotēkas fragments**

Wehśch tahjahß / und **irr itt dischans rahms** Ghais. (LP1 188)	Daśch skattahß us zitto Ļauscho Dsiewośchanu / und **gir tschacklis redseht** ko ziti nhepareise darra (LP1 392)
Śchiß Autz gir krahßnis / jaux / dischans Tebbikis bijis / no dsälltänahm Siedehm / **śarrko= śarrkanahm** / und **deegha=balltahm** Siedehm (LP3 198)	Apdohma nu wehl / ka schinnieß Kahsahß wiß **lieds pillam ghattaws** und **śataisśietz**. (LP2 351)
Turr buhśśim mehs **dauds taißnaki nhe ka Adams** / pirrms wings ghräkoja / **dauds stippraki nhe ka Simsons** / **dauds krahßnaki / nhe ka Absolons** / **dauds ghuddraki / nhe ka Salomons** (LP1 21)	Auxta gir śchie Leeta / und leela / **tick leela** / ka **nheneeka auxtaka** / **leelaka** / **briniggaka** buht warr. (LP2 9)

Līdzīgi tika meklētas substantīviskas vārdkopas ar saskaņotu adjektīvu, numerāli, pronomenu. Nav grūti atrast vārdkopas ar piekļautu adverbu, nelokāmo numerāli, jo to formas ir nemainīgas, piemēram, *deßmit, deßmits, dauds, mas, zeek, tick, magkeniet*.

Vārdkopu ar atributīvo divdabi meklēšana

Lai sāktu meklēt divdabjus „Seno tekstu korpusā", vispirms tika noskaidrotas iespējamās divdabju formas morfoloģijas vēsturē (LLVMSA 2002, 425–467). Substantīviskās vārdkopas ar atributīvu divdabi tika atrastas, ierakstot korpusa meklēšanas logā divdabju izskaņas, priekšā liekot % zīmi.

Darāmās kārtas tagadnes divdabji tika meklēti, ierakstot vaicājumu *%ośch%* un iezīmējot vajadzīgos avotus. Piemēram:

%ośch% → Deews jaw ween reis tha gir raddijis / ka 7: nheneeka nhe warr augt / ja taß nhe diegs no kahdas Śähklas / 8: laid buht ta Labbiba / jeb **§eedoścha** Sahle / jeb Pugki§chi .. LP1, 255. lpp. → *Deews jaw ween reis tha gir raddijis / ka nheneeka nhe warr augt / ja taß nhe diegs no kahdas Śähklas / laid buht ta Labbiba / jeb **seedoścha Sahle** / jeb Pugkischi* .. (LP1 255)

%ośch% → Tapehtz 24: tahdu Zillwäku räd§ädammi / buhß mums ad§iet to leelu Schälä- 25: §tibu tha wi§śewallditaja Deewa / katters mums muh§śas d§irrdo- 26: śchas Au§śis / und runnajo§chu Mehl dehwis gir / und tam par ta- 27: du leelu Labb=darri§chanu Ghodu / Ślawu unnd Patteitzibu śa- 28: tziet. LP2, 210. lpp. → *Tapehtz tahdu Zillwäku rädsädammi / buhß mums adsiet to leelu Schälästibu tha wisśewallditaja Deewa / katters mums muhśśas **dsirrdośchas Ausśis** / und **runnajoschu Mehl** dehwis gir / und tam par tadu leelu Labb=darrischanu Ghodu / Ślawu unnd Patteitzibu śatziet*. (LP2 210)

%ośch% → .. tad Winjo Sem- 13: me par **däggoſchu** Picki taps .. LP2, 155. lpp. → *.. tad Winjo Semme **par däggoſchu Picki** taps ..* (LP2 155)

Izmantojot šāda veida meklēšanu, tiek atrasti pilnīgi visi vārdi, kuros ir burtu salikums *-ośch-*, piemēram, *addoſchanu, dsiewoſchana, Deewa=Saimoſchanu, deßmitts=tuhxtoſchus, droſcha, dußmoſchanas, mälośchana, peedośchana* u. c.

Līdzīgi tika meklēti atributīvi divdabji ar izskaņām *-ams*, *-ohts*, *-dams* un verbāladjektīvs ar izskaņu *-tins*:

-ams → *%am%* → Kad Deews teem illghe usluh- 25: kojiß / und pehtz tohß śohda / dohd Wings teem trie§śamu und 26: kluxtamu Śirrdi / tad teem Beßdeewigheem Meers nhe 27: gir. LP3, 200. lpp. → *Kad Deews teem illghe usluhkojiß / und pehtz tohß śohda / dohd Wings teem **triesśamu** und **kluxtamu Śirrdi** / tad teem Beßdeewigheem Meers nhe gir*. (LP3 200)

-ohts → *%oht%* → Pils / śa§paidiets / śakrattietz / unnd 17: pahr=eijohts Mährs / taps jums juh§śa` Klehpy ee=dohts. LP2, 304{284}. lpp. → *Pils / śaspaidiets / śakrattietz / unnd **pahr=eijohts Mährs** / taps jums juhśśà Klehpy ee=dohts*. (LP2 304)

-dams → *%dam%* → .. kaß ißniegdamas 8: Leetas meckleh / taß lied§e ißniex. LP2, 92. lpp. → *.. kaß **ißniegdamas Leetas** meckleh / taß liedse ißniex*. (LP2 92)

-tins → *%tin%* → Behdigha / **§chälotina** 7: Leeta gir wi§śo Zillwäko D§iewiba / no Mahtes Mee§śas / 8: teekam tee Semmeh aprakti tohp / kattra muh§śo wi§śo 9: Mahte gir. LP2, 269{249}. lpp. → *Behdigha / **schälotina Leeta** gir wiśśo Zillwäko Dsiewiba / no Mahtes Meeśśas / teekam tee Semmeh aprakti tohp / kattra muhśśo wiśśo Mahte gir*. (LP2 269)

Lai atrastu ciešamās kārtas pagātnes divdabjus, bija jāparedz visas iespējamās izskaņu formas. Piemēram:

%ts → *Śchiß Kunngs gir tuwe py teem / kattrreem śalaususchas Śirrdes gir / und palieds teem / kattreem **śaśists Prahts** gir / śacka taß Konings Dawids*. (LP2 274)

%ta → *.. taß Kungs JEsus Christus tohß nabbaghus Ghrezeneekus / kattri gir ka ißklieduśchi Awis / und **pasusta apruhśśäta Nauda** / meckleh / tick illghe / kamähr Wings tohß attrohd ..* (LP2 61)

%tas → *Bett kad tee pee=ehdehß / śatzija Wings us śaweems Mahzekļeems / śackrajeta tahß **attlicktas Drußkas** / ka no teems nheneeka ißbahrstahß*. (LP1 327)

%tam → *Redsi / tu / itt tu patz eśśi taß Kohx / katters nhe kahdu labbu Auglu neß / töw buhß nozirrstam / und ka kahdam ißkalltuścham **śatruhdätam Praulam** Elles däggußchà zepply mäśtam kluht*. (LP3 75)

%teems → Eß nhe äßmu śuhtietz / ka ween py teems ***pasusteems Ahweems*** .. (LP1 304)

%tu → Nhe buhtu tad tam Kungham JEsu winja Śirrds sprahghusśi / rädsädams ka winja Ļaudeems teems Judeems / kattrus wings śawu pirrmu ***peedsimbtu Dählu*** */ uñ śawu mieļu Dwehsśel dehwe / by ißposteems tapt*. (LP2 193)

%to → Wings ghribbohtz irr tahdo ***nhedsimmto Bährno*** */ kattri wehl śawas Mahtes Meesśahß ghull / Pestitais buht*. (LP1 366)

Pēc tāda paša principa tika meklēti arī darāmās kārtas pagātnes divdabji.

Problēmas

1. Biežāk lietotajiem vārdiem jāizskata daudz lappušu, lai atlasītu tos piemērus, kuros ir nominālās vārdkopas. Līdz ar to var rasties neprecizitātes vārdkopu biežuma noteikšanā, var kādu izlaist, taču tas netraucē nonākt pie vispārīgām atziņām. Piemēram, ja jāatrod substantīviskās vārdkopas netiešajā pārvaldījumā, tad *meklētājā* ieraksta atbilstošos prievārdus, kuri lietoti pietiekami bieži.
2. Teksts jāpārveido atbilstoši seno tekstu ortogrāfijas principiem. Apzīmējumi atšķiras no oriģinālrakstības.
3. Nav visiem tekstiem pieejams faksimils. Diemžēl G. Manceļa spredi̇ķu grāmatai ir tikai ar „Seno tekstu korpusa" pieņemtajiem apzīmējumiem pārveidotais teksts. Meklējot nepieciešamo materiālu, tika konstatēti piemēri, par kuru atbilstību oriģināla tekstam radās šaubas, piemēram:

MJeļi *Draughi / Jßghaju§cha` Śwehdena` e§śeeta juhß d§irrdejuśchi* (LP2 186)

Zitz attkal śatza / Mehß ka Deewa Ļaudis ä§śam ***bry*** */ mums nhe buhß doht / bett to Tee§śu Ba§nizas=Schkirr§ta` eelickt.* (LP2 382)

Nu gir taß Kungs JE§us taß ***ļohteauxte*** *teitzams Deews muh§chighe*. (LP1 93)

Oriģināla faksimili ir svarīgi tāpēc, ka teksti „Seno tekstu korpusā" tika ierakstīti ar roku (*www.korpuss.lv/senie/about.htm#sagatavosana*).

Jāņem, protams, vērā projekta dalībnieku atzinumi, ka „Seno tekstu korpuss" tiek un tiks papildināts un nākotnē paredzēts tajā ievietot citus oriģināla faksimilu attēlus (*www.korpuss.lv/senie/about.htm#plani*).

4. Nepieciešamo piemēru var atrast tikai pēc viena vārda. Nevar, piemēram, ierakstīt vārdu savienojumu, vārdkopu, frazeoloģismu.

5. „Seno tekstu korpusā" nav pirmās tulkojošās vārdnīcas, kurā viena no valodām ir latviešu. Tajā pašā laikā ir pieejama K. Fīrekera vārdnīca.

Korpusa veidotāji norādījuši, ka, izskatot Latvijas Nacionālās bibliotēkas sagatavoto kopkatalogu „Seniespiedumi latviešu valodā 1525–1855", „tika nolemts sākumā elektroniskā formā sagatavot svarīgākos 17. gs. pirmizdevumus, atkārtotos izdevumus apstrādājot nākotnē" (*www.korpuss.lv/senie/about.htm#plani*). Tomēr rodas jautājums, vai G. Manceļa vārdnīca nav „svarīga"?

„Seno tekstu korpusa" pozitīvās puses

1. Nevajag pārrakstīt piemērus vēlreiz, jo tos var viegli iekopēt vajadzīgajā dokumentā. G. Manceļa vārdnīca un tās pielikumi nav „Seno tekstu korpusā", tāpēc piemērus vajadzēja ievadīt datorā ar roku, kas bija laikietilpīgs process, jo bieži vien bija jāmeklē vajadzīgie burti pie simboliem, kā arī vajadzēja datorā ielādēt citus burtu veidus (*Fontra* un *IndoBalt*).
2. Vajadzīgos piemērus var atrast, nelasot visu spredi̦ku grāmatas tekstu.
3. Piemērus var uzreiz grupēt, izveidojot katram vārdkopu tipam savu atsevišķu mapīti.
4. Konkordancē var atrast vajadzīgās leksēmas uzreiz vairākos avotos, jo ir atļauts iezīmēt un atvērt vairākus tekstus vienlaikus. Izmantojot taustiņu *ctrl*, var iezīmēt arī avotus, kas neatrodas blakus.

Secinājumi

„Seno tekstu korpuss" palīdzēja gūt plašāku ieskatu vārdnīcu tipos, kuri vārdnīcas tekstā lietoti retos gadījumos, piemēram, adjektīviskās vārdkopas, substantīviskās vārdkopas ar atributīvu divdabi atkarīgajā komponentā, substantīviskās vārdkopas netiešajā pārvaldījumā. Tā, piemēram, vārdnīcā tika konstatētas ~100 substantīvisku vārdkopu ar atributīvu divdabi, bet „Seno tekstu korpusā" – ~750.

Literatūra

1. Bergmane Anna, Blinkena Aina. *Latviešu rakstības attīstība.* Latviešu literārās valodas vēstures pētījumi. Rīga : Zinātne, 1986, 435 lpp.
2. Elksnīte Gita. *Nominālās vārdkopas Georga Manceļa tekstos*. Promocijas darbs. Liepāja : 2011, 182 lpp.
3. Fennell Trevor Garth. *A Latvian-German Revision of G. Mancelius' "Lettus"* (1638). Melbourne: Latvian Tertiary Committee, 1988, 110 lpp.
4. Fennell Trevor Garth. *A Latvian-German Revision of G. Mancelius' "Phraseologia Lettica"(1638)*. Melbourne: Latvian Tertiary Committee, 1989, 150 lpp.
5. Lettus – LETTUS, Das ist Wortbuch / Samt angehengtem täglichem Gebrauch der Lettischen Sprache; Allen vnd jeden Außheimischen / die in Churland / Semgallen vnd Lettischem Liefflande bleiben / vnd sich redlich nehren wollen / zu Nutze verfertigt / durch GEORGIVM MANCELIVM Semgall. der H. Schrifft Licentiatum &c, Erster Theil. Cum Grat. & Priv. S. R. M. Svec. Gedruckt vnnd verlegt zu Riga / durch GERHARD. Schröder / Anno M. DC. XXXVIII. *Altlettische Sprachdenkmäler in Faksimiledrucken*. Herausgegeben von August Günther. II. Band. Heidelberg: Carl Winters Universitätsbuchhandlung, 1929, 2.–222. lpp.
6. LP1, LP2, LP3 – Mancelis Georgs. *Lettische Langgewünschte Postill.* Pieejams: *http://www.ailab.lv/senie/*
7. PhL – PHRASEOLOGIA LETTICA, Das ist: Täglicher Gebrauch der Lettischen Sprache. Verfertigt durch GEORGIUM MANCELIUM Semgallum, der H. Schrifft Licentiatum &c. Ander Theil. Diesem ist beygefüget das Spruchbuch Salomonis. Zu Riga Gedruckt vnnd Verlegt durch Gerhard. Schröder/1638. *Altlettische Sprachdenkmäler in Faksimiledrucken*. Herausgegeben von August Günther. II. Band. Heidelberg: Carl Winters Universitätsbuchhandlung, 1929, 223.–414. lpp.
8. Pokrotniece Kornēlija. Nelokāmo divdabju funkcionālais lietojums Manceļa „Postillā". *Letonikas pirmais kongress. Valodniecības raksti.* Rīga : LZA, 2006, 272.–280. lpp.
9. Pokrotniece Kornēlija. Divdabji G. Manceļa sprediķu grāmatā I. *Linguistica Lettica*. 18. krāj. Rīga : LU Latviešu valodas institūts, 2008, 27.–38. lpp.
10. Skrūzmane Elga. Dažas frazeoloģijas īpatnības G. Manceļa „Postillā". *Vārds un tā pētīšanas aspekti* : rakstu krājums, Nr. 13(1). Liepāja : LiePA, 2009, 307.–316. lpp.
11. Zehen G – Mancelis Georgs. *Zehen Gespräche Deutsch und Lettisch*.. Riga, 1685. Ekscerptiem izmantots klišeju novilkums. Rīga, 1942. 55.
12. Latviešu valodas seno tekstu korpuss. Pieejams: *http://www.korpuss.lv/senie/*

Datorprogrammas zinātniski pētniecisko darbu izstrādē

Lienīte Litavniece, Sandra Murinska

Eiropas Sociālā fonda projekta „Teritoriālās identitātes lingvokulturoloģiskie un sociālekonomiskie aspekti Latgales reģiona attīstībā" datubāzu veidošana un datu apstrādes metodoloģija

Pētnieciskais darbs neatkarīgi, vai tas tiek realizēts skolā, augstskolā vai zinātniskajā institūtā, pētniekam rada sava veida pienākumu. Katrs pētījums sākas ar ideju, kura materializējas konkrētā mērķī un veicamās aktivitātēs jeb uzdevumos. Sekmīga mērķa sasniegšana nav iespējama bez konkrētai situācijai precīzi izvēlētām un izmantotām metodēm. Katru zinātnes nozari raksturo tai specifiskas pamatinformācijas un datu ieguves, apstrādes un interpretācijas metodes.

Šī raksta mērķis ir sniegt priekšstatu par datu apstrādes iespējām Statistisko datu apstrādes paketes (*Statistical Package for the Social Sciences – SPSS*) programmā humanitārajās un sociālajās zinātnēs. Autores informē par Rēzeknes Augstskolā īstenotā Eiropas Sociālā fonda (ESF) finansētā projekta „Teritoriālās identitātes lingvokulturoloģiskie un sociālekonomiskie aspekti Latgales reģiona attīstībā" laikā iegūtajiem datiem un to apstrādes iespējām, izmantojot *SPSS* programmu, salīdzinot datu (aptaujas anketu un fotogrāfiju) ievades un apstrādes iespējas datorprogrammās *MS Excel* un *SPSS*.

„Dati (*data* < latīņu *datum* 'dotais' < *dare* 'dot') ir tekstuālas, skaitliskas vai grafiskas ziņas, kas kaut ko raksturo un noder par pamatu secinājumiem, kā arī skaitliska vai citāda informācija, ko apstrādā ar datoru" (Svešvārdu vārdnīca 2007, 132). Dati sniedz informāciju, kas pēc tam var tikt interpretēta dažādi. Tie var tikt analizēti objektīvi, apkopojot kādas parādības statistiku, biežumu. Savukārt, izmantojot metodi, kas balstās uz subjektīviem faktoriem, var tikt panākta atšķirīga datu interpretācija. Datu analīzes pieeju nosaka pētījuma mērķis, izvēlētās metodes un to lietojums.

Pēc pētījuma sagaidāmā rezultāta datus iespējams klasificēt **kvantitatīvajos** un **kvalitatīvajos.** Kvantitatīvie dati sniedz atbildi uz jautājumiem – *cik daudz, cik bieži*. Lielākoties tā ir skaitliska informācija par kādas parādības izplatību. „Kvantitatīvais pētījums parasti ietver eksperimentālo vai vismaz kvazieksperimentālo (daļēji eksperimentālo) dizainu, datu vākšanu, lietojot standartizētas procedūras vai instrumentus laboratorijā vai līdzīgā institūcijā, skaitliskos datus, statistisko analīzi un hipotēzes pārbaudi par sprieduma vispārināšanas līdzekli" (Kroplijs, Raščevska 2010, 18). Savukārt kvalitatīvie dati ļauj skaidrot parādības būtību, aspektus. Tie sniedz atbildi uz jautājumiem *kāpēc* un *kā*. Tiek iegūta informācija par to, kā tiek saprasta realitāte: „Kvalitatīvo datu pētījums tiecas lietot neeksperimentālus dizainus dabīgā vidē, vākt dažādu veidu aprakstošus stāstījuma veida datus (..), analīzes mērķis ir atklāt stāstījuma nozīmi" (Kroplijs, Raščevska 2010, 18). Izmantojot kvantitatīvo un kvalitatīvo datu veidus, iespējams sniegt daudzpusīgu situācijas vai parādības raksturojumu un analīzi.

Rēzeknes Augstskola (turpmāk – RA) kā reģionālā augstākās izglītības iestāde savā pētnieciskajā darbībā, pamatojoties uz kvantitatīvajiem un kvalitatīvajiem datiem, veic reģiona un Latvijas attīstībai nozīmīgus pētījumus (vides, humanitārajā, pedagoģijas, sociālajā u. c. jomās). To īstenošanai tiek izmantotas daudzveidīgas pētnieciskās metodes un datu apstrādes programmas.

RA 2009. gada 1. decembrī sāka realizēt ESF finansētu projektu „Teritoriālās identitātes lingvokulturoloģiskie un sociālekonomiskie aspekti Latgales reģiona attīstībā". Projekts tiks turpināts līdz 2012. gada 30. novembrim, tā laikā:

- tiks izveidota Baltijas reģionā pirmā lingvoteritoriālā vārdnīca;
- tiks veikta Latgales reģiona pilsētu lingvistiskās ainavas datu salīdzināšana ar citām Baltijas valstu pilsētām;
- tiks izstrādāts Latgales reģiona tautsaimniecības un pilsētu piesaistes spēju izvērtējums;
- tiks sagatavoti priekšlikumi pašvaldību un tautsaimniecības speciālistu darba efektivitātes paaugstināšanai (sk. projekta mājas lapā *http://tilra.ru.lv*).

Lai sekmīgi sasniegtu plānotos rezultātus, būtiski ir pareizi izmantot iegūto informāciju. Projektā iesaistītais personāls darbojas trīs aktivitātēs (lingvistiskās ainavas izpēte, pilsētu pievilcības izpēte un lingvoteritoriālās vārdnīcas izveide), koncentrējoties uz noteiktu mērķu sasniegšanu, izmantojot starpdisciplināru pieeju.

Tā pamatā ir aptaujas anketu (iedzīvotāju, uzņēmēju, tūristu), pilsētvidē iegūto fotogrāfiju un statistikas datu apstrāde un analīze.

Datu ievades un apstrādes atšķirības *MS Excel* un *SPSS* programmā

Projekta pilsētas pievilcības izpētes aktivitātes rīcībā ir iedzīvotāju, uzņēmēju un tūristu aptaujas rezultāti un statistiskie dati, lai izvērtētu Latgales pilsētu pievilcību un izstrādātu koncepciju Rēzeknes pilsētas pievilcības uzlabošanai, kā arī izveidotu datubāzi ar rādītājiem par 14 Latgales reģiona pilsētu sociālekonomisko attīstību un piesaistes spējas (pievilcības) aspektiem.

Viens no laikietilpīgākajiem un darbietilpīgākajiem datu iegūšanas veidiem ir anketēšana, ar kuras palīdzību iegūst pētījumam nepieciešamo informāciju. Būtiska nozīme ir aptaujas mērķgrupai, kuru reprezentē pamatota izlase, un aptaujas anketas saturam. Tā veidošanā būtiski atcerēties jautājumus:

- Kāds ir anketēšanas mērķis (jautājumu formulējums)?
- Kā anketu sastādīt, lai vieglāk apstrādātu iegūtos datus?

Otrais jautājums nosaka tālāko pētniecisko rīcību: ir jāizlemj, kāda programma tiks izmantota datu apstrādei un kā pareizi šajā programmā dati tiks ievadīti.

Aptaujas datu apstrādei ir iespējams izmantot, piemēram, *MS Excel*, kas ir pieejama *MS Office* standartpakā, bet daudz plašākas iespējas piedāvā *SPSS* programma. Pamatojoties uz projekta laikā iegūtajiem Balvu pilsētas iedzīvotāju aptaujas anketu datiem, salīdzinātas datu ievades un apstrādes atšķirības iepriekš minētajās programmās.

Kopumā projekta laikā veikta anketēšana 14 Latgales pilsētās (Daugavpilī, Rēzeknē, Ludzā, Balvos, Krāslavā, Dagdā, Subatē, Ilūkstē, Zilupē, Viļānos, Viļakā, Kārsavā, Līvānos un Preiļos). Iedzīvotāju aptaujas anketā ietverti 82 jautājumi, kas sadalīti 18 tematiskajos blokos (informācija par respondentu; apmierinātība ar dzīvi pilsētā; materiālais stāvoklis; darbs un nodarbinātības iespējas; izglītība; drošība; ekoloģiskā kvalitāte; veselība; sociālā aprūpe; kultūra; brīvais laiks; sports; sabiedriskā līdzdalība; tirdzniecības pakalpojumi; sabiedriskais transports; personiskais transports; pakalpojumu kvalitāte un pieejamība; reģionālā piederība).

1. tabula. **Piemērs no iedzīvotāju aptaujas anketas**

1.1. Dzimums:	Vīrietis	1	Sieviete	2
1.2. Vecums:				gadi
1.3. Izglītība:	Augstākā	1		
	Vidējā	2		
	Profesionālā	3		
	Pamatskolas	4		
	Cita atbilde	5		
1.4. Nodarbošanās:	Strādāju algotu darbu			1
	Strādāju, pašnodarbinātais			2
	Nestrādāju, bezdarbnieks (*norādiet, cik sen – gadi, mēneši*) _____			3
	Bērna kopšanas atvaļinājumā			4
	Cita atbilde			5

Tā kā datu apstrādei tika izvēlēta *SPSS* programma, aptaujas anketa veidota tā, lai datus varētu apstrādāt ar šo programmu (sk. 1. tabulu). *SPSS* programma izvēlēta, jo tai ir plašas un daudzpusīgas iespējas darbā ar dažādu apjomu un veidu datiem.

Iespējamajiem atbilžu variantiem tiek piešķirts kods, kurš atbilstoši pēc tam tiek ievadīts programmā. Šāda anketas veidošanas metode atvieglo datu ievadi pētnieka izvēlētajā programmā. *SPSS* programmā (atšķirībā no *MS Excel*) datu ievadi var veikt tikai horizontāli (sk. 1. attēlu).

Ja pētnieks sākotnēji datu ievadei ir izmantojis *MS Excel* programmu (datu horizontālais izvietojums), tos var viegli importēt uz *SPSS*. Tādēļ pētniekam, pirms

1. attēls. **Datu ievades piemērs SPSS programmā**

	kods	dzimums	vecums	izglitiba	nodarb	bezd_ilg	tautiba
1	B1	1	32	4	3	3	1
2	B2	1	32	4	3	1	1
3	B3	1	34	1	3	1	1
4	B4	1	54	3	3	100	3
5	B5	1	37	1	3	24	1
6	B6	1	37	3	3	1	3
7	B7	1	42	3	3	12	1
8	B8	2	30	1	3	12	1
9	B9	2	47	2	3	6	1

sākot darbu ar kritēriju izveidi un ievadi *SPSS* vidē, ir jāpārdomā, kāds ir pētījuma mērķis, lai datu izvietojums netraucētu tā sasniegšanu.

Būtiskas *SPSS* ir izvēlnes, kuras atrodas lapas kreisajā apakšējā stūrī: datu skats (*Data View*) un mainīgo skats (*Variabla View*). Vienā dokumentā ir atrodami gan ievadītie dati, gan arī kodu sistēma, kas padara darbu ar datiem pārredzamu un kompaktu.

Mainīgo skatā tiek ievadīts apraksts par konkrēto anketas jautājumu: kolonnas nosaukums (*Name*), veids (*Type*), ieraksta garums šūnā (*Width*), vietu skaits aiz decimālā atdalītāja (*Decimals*), mainīgā apraksts jeb skaidrojums (*Label*), vērtība (*Values*), trūkstošās vērtības (*Missing*), kolonnas platums (*Columns*), ierakstu izlīdzinājums datu skatā (*Align*) un mērījumu skalas veids (*Measure*) (sk. 2. attēlu).

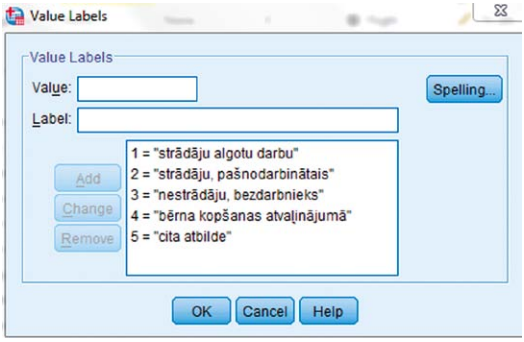
Svarīga ir vērtību (*Values*) kolonna, kurā tiek ievadīta informācija par visiem iespējamajiem atbilžu variantiem.

2. attēls. **Mainīgo skats**

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	kods	String	24	0	Anketas kods	None	None	8	Left	Nominal
2	dzimums	String	24	0	Dzimums	{1, vīrietis}...	None	8	Left	Nominal
3	vecums	Numeric	8	0	Vecums	None	None	8	Right	Scale
4	izglitiba	Numeric	8	0	Izglītība	{1, augstāk...	-999	8	Right	Ordinal
5	nodarb	Numeric	8	0	Nodarbošanās	{1, strādāju ...	-999	8	Right	Nominal
6	bezd_ilg	Numeric	8	0	Perioda ilgums ...	None	None	8	Right	Scale
7	tautiba	Numeric	8	0	Tautība	{1, latvietis}...	-999	8	Right	Nominal
8	gim_stav	Numeric	8	0	Ģimenes stāvo...	{1, precējies...	-999	8	Right	Nominal
9	berni	Numeric	8	0	Bērni	{1, ir}...	-999	8	Right	Nominal

Vērtību ievadīšana (sk. 3. attēlu) nozīmē, ka katram iespējamajam atbilžu variantam uz konkrēto jautājumu tiek piešķirts skaitlis, kas ir kā kods.

3. attēls. **Mainīgo vērtību kodēšana**



Datu apstrādes procesā lietotājs, izmantojot mainīgo vērtības vai vērtības apraksta maiņas rīku, var izvēlēties kodus (sk. 1. attēlu) vai kodu atšifrējumus (sk. 4. attēlu), kas ir vēl viena salīdzinošā priekšrocība SPSS programmā (atšķirībā no *MS Excel*).

4. attēls. **Datu ievades piemērs ar vērtību aprakstu**

kods	dzimums	vecums	izglitiba	nodarb	bezd_ilg	tautiba
B1	vīrietis	32	pamatskolas	nestrādāju, bezdarbnieks	3	latvietis
B2	vīrietis	32	pamatskolas	nestrādāju, bezdarbnieks	1	latvietis
B3	vīrietis	34	augstākā	nestrādāju, bezdarbnieks	1	latvietis
B4	vīrietis	54	profesionālā	nestrādāju, bezdarbnieks	100	krievs
B5	vīrietis	37	augstākā	nestrādāju, bezdarbnieks	24	latvietis
B6	vīrietis	37	profesionālā	nestrādāju, bezdarbnieks	1	krievs
B7	vīrietis	42	profesionālā	nestrādāju, bezdarbnieks	12	latvietis
B8	sieviete	30	augstākā	nestrādāju, bezdarbnieks	12	latvietis
B9	sieviete	47	vidējā	nestrādāju, bezdarbnieks	6	latvietis

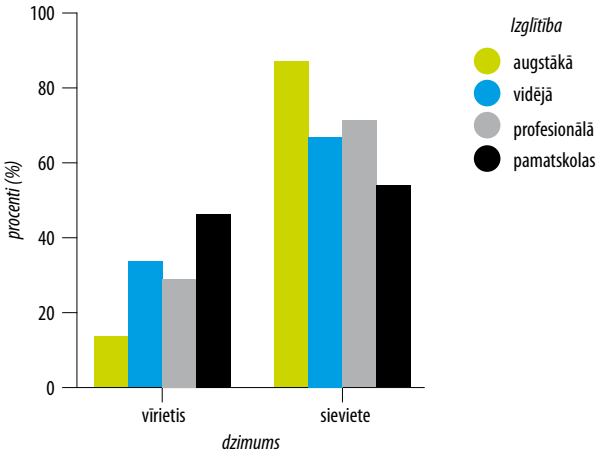
Datu ievade programmā *SPSS*, salīdzinot ar *MS Excel*:

- datu ievade notiek tikai horizontāli;
- salīdzinoši ilgāks datu ievades posms, bet dati un to atšifrējumi pieejami vienā dokumentā un ir ērti lietojami.

Datu apstrādē *SPSS* programma piedāvā plašas iespējas (frekvenču, korelāciju, regresiju aprēķinus, kopsakarību tabulu veidošanu u. c.). Pamatojoties uz Balvu pilsētas iedzīvotāju anketas datiem, raksta autori sniegs priekšstatu par dažām datu apstrādes iespējām.

Attēlu un grafiku veidošana ir nozīmīga datu analīzes sastāvdaļa. Paralēli vienkāršajiem grafiku veidiem, ko piedāvā *MS Excel*, ar *SPSS* ir iespējams izveidot klastera veida grafikus (dati tiek grupēti, pamatojoties uz kādu noteiktu pazīmi). Pētāmajā piemērā (sk. 5. attēlu) ir redzams respondentu sadalījums pēc izglītības līmeņa un dzimuma. Šādā veidā ir iespējams iegūt grafiskus attēlus pētniekam interesējošā datu griezumā.

5. attēls. **Balvu pilsētas iedzīvotāju aptaujas dati pēc respondentu izglītības līmeņa un dzimuma**



Interesējošos datus var attēlot arī tabulā un salīdzināt iegūtos rezultātus skaitliskā izteiksmē (2. tabula).

2. tabula. **Respondentu atbildes kopsakarību tabulās**

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Dzimums* Izglītība	130	99,2%	1	8%	131	100%

Dzimums* Izglītība Crosstabulation

Count

		Izglītība				
		augstākā	vidējā	profesionālā	pamatskolas	Total
Dzimums	vīrietis	7	11	9	6	33
	sieviete	46	22	22	7	97
Total		53	33	31	13	130

Bieži vien ir svarīgi arī veikt aprēķinus, lai zinātu procentuālo respondentu sadalījumu. Līdzīgu tabulu ir iespējams izveidot arī *MS Excel* programmā. Ar *SPSS* iegūtā 3. tabula atspoguļo respondentu dalījumu absolūtos skaitļos un procentuālā izteiksmē pēc noteikta kritērija. Aprēķinātie procenti tiek grupēti divās kolonnās:

- procenti (*Percent*) – dati aprēķināti no kopējā respondentu skaita;
- procenti no gadījumiem (*Valid Percent*) – dati aprēķināti no to respondentu skaita, kuri ir snieguši savu atbildi uz konkrēto jautājumu.

Veidojot anketas, bieži vien rodas nepieciešamība lūgt respondentiem izvēlēties vairākus sev tīkamus atbilžu variantus, savukārt tas sarežģī datu ievadi un apstrādi. *SPSS* programma paredz šādu situāciju rašanos, un, veicot datu ievadi, katra iespējamā atbilde tiek veidota kā atsevišķs jautājums. Atbilstoši noformējot (izmantojot atbilstošās izvēlnes), programma šīs atbildes uztver kā vienu kopīgu jautājumu. 4. tabulā ir redzamas atbildes uz jautājumu *Kā Jūs pavadāt brīvo laiku?*, kurš saistīts ar iepriekš raksturoto situāciju. Kolonnā „procenti" programma ir aprēķinājusi procentus respondentu atbildēm, pamatojoties uz kopējo respondentu skaitu. Savukārt kolonnā „procenti no gadījuma" (katrs respondents varēja izvēlēties ne vairāk kā trīs atbilžu variantus) kā pamats procentu aprēķinam kalpo kopējais atbilžu skaits. Un, veicot datu analīzi, pētniekam ir iespēja izvēlēties sev nepieciešamo kolonnu un veikt iegūto rezultātu interpretāciju.

Salīdzinot *MS Excel* un *SPSS* atšķirības datu apstrādes procesā, jāsecina, ka *SPSS* programmā ir plašākas datu reprezentēšanas iespējas un datu analīzes process ir salīdzinoši vienkāršāks.

3. tabula. **Respondentu atbilžu biežums**

Tautība

		Frequency	Percent	Valid Percent
Valid	latvietis	100	76,3	76,9
	latgalietis	7	5,3	5,4
	krievs	20	15,3	15,4
	polis	1	,8	,8
	baltkrievs	1	,8	,8
	ukrainis	1	,8	,8
	Total	130	99,2	100,0
Missing	-999	1	,8	
Total		131	100,0	

4. tabula. **Respondentu atbilžu biežums jautājumiem ar vairākām iespējamajām atbildēm**

	Atbildes		Procenti no gadījumiem
	Skaits	Procenti	
Brīvo laiku pavada sportojot	18	6,1 %	14,6 %
Brīvo laiku pavada, lasot grāmatas	54	18,2 %	43,9 %
Brīvo laiku pavada, skatoties TV	64	21,5 %	52,0 %
Brīvo laiku pavada, „sērfojot" internetā	33	11,1 %	26,8 %
Brīvo laiku pavada, spēlējot datorspēles	4	1,3 %	3,3 %
Brīvo laiku pavada dejojot	11	3,7 %	8,9 %
Brīvo laiku pavada, pastaigājoties svaigā gaisā	43	14,5 %	35,0 %
Brīvo laiku pavada bāros, kafejnīcās, restorānos	8	2,7 %	6,5 %
Brīvo laiku pavada draugu tusiņos	23	7,7 %	18,7 %
Brīvo laiku pavada, piedaloties pašdarbības kolektīvos	14	4,7 %	11,4 %
Brīvajā laikā nodarbojas ar rokdarbiem, floristiku	10	3,4 %	8,1 %
Brīvo laiku pavada brīvā dabā	15	5,1 %	12,2 %
Kopā	**297**	**100,0 %**	**241,5 %**

SPSS programma piedāvā ne tikai aptaujas datu, bet arī fotogrāfiju kā informācijas avota ievadi un apstrādi.

Lingvistiskās ainavas datu ievade un apstrāde *SPSS*

Lingvistiskā ainava (*Linguistic Landscapes*) ir sociolingvistikas pētniecības metode, ar kuras palīdzību iespējams veikt rakstītās informācijas izpēti publiskajā telpā, konstatējot valodu situāciju konkrētajā teritorijā. Šīs metodes teorētiskais pamatlicējs ir nīderlandiešu valodnieks Dirks Gorters, kurš piedāvā to kā jaunu pieeju multilingvālas vides izpētē. Tā ir iespējama, pateicoties mūsdienu digitālajām tehnoloģijām (vizuālajai un tekstuālajai mijiedarbei un izplatībai publiskajā komunikācijā).

Metode paredz publiskajā telpā (veikalu izkārtnes, nosaukumi, afišas, reklāmas plakāti, vides reklāmas, reklāmas uz sabiedriskā transporta, ielu un vietu nosaukumu plāksnes, valdības iestāžu publiskā informācija, grafiti u. c.) esošo rakstīto valodas zīmju ieguvi noteiktos ģeogrāfiskos apgabalos.

Latvijā un RA šīs metodes aizsācēji ir *Dr. philol*. H. F. Martens un asoc. prof. S. Lazdiņa, kas 2008. gadā īstenoja projektu „Latvijas lingvistiskā ainava Baltijas valstu kontekstā". Tā laikā tika iegūtas valodas zīmes Baltijas valstu pilsētās – Rēzeknē, Ventspilī, Alītā un Pērnavā (izvēlētās pilsētas vieno līdzvērtīgs iedzīvotāju skaits, ģeogrāfiskais izvietojums u. c. faktori). Rezultātā tika veikti vairāki pētījumi (Lazdiņa, Pošeiko, Marten 2008).

6. attēls. **Valodas zīme Druskininkos (Lietuva)**

7. attēls. **Valodas zīme Rēzeknē (Latvija)**

2010. gadā RA realizētā projekta „Teritoriālās identitātes lingvokulturoloģiskie un sociālekonomiskie aspekti Latgales reģiona attīstībā" aktivitāšu laikā datubāze tika papildināta ar kvalitatīvo un kvantitatīvo datu vākšanu vēl divās Baltijas valstu pilsētās – Druskininkos (Lietuvā) un Narvā (Igaunijā), ko veica pētnieki S. Lazdiņa, H. F. Martens, O. Senkāne, I. Matisovs, S. Pošeiko un S. Murinska.

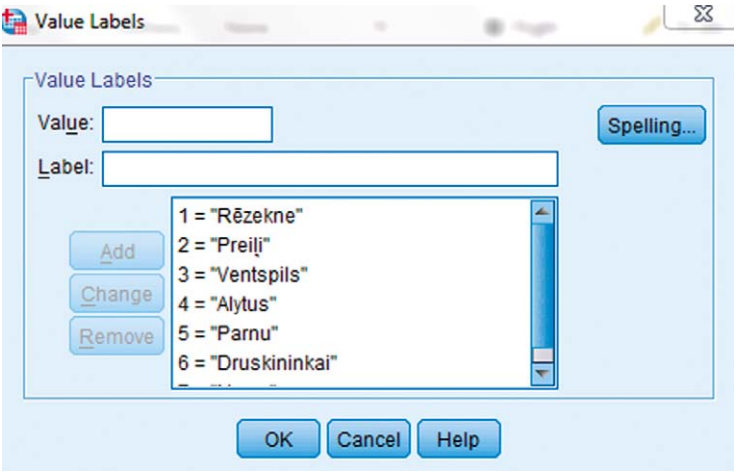
Lingvistiskās ainavas izpētes datu iegūšanas un izlases veidošanas princips ir šāds: kvantitatīvie dati jeb valodas zīmes tiek iegūtas, fotografējot (sk. 6. un 7. attēlu) pilsētas publiskajā telpā ar blīvu uzņēmumu, iestāžu koncentrāciju, savukārt kvalitatīvie dati jeb ielu intervijas pēc nejaušības principa – uzdodot jautājumus pilsētas ielās un uzņēmumos sastaptajiem cilvēkiem (vietējiem iedzīvotājiem, tūristiem, veikalu darbiniekiem u. c.). Kopumā iepriekšminētajās pilsētās tika iegūtas 4884 rakstu valodas zīmes un 53 intervijas.

Apkopotās valodas zīmes ievadītas datorprogrammā *SPSS*, un tad veikta to analīze, interpretējot informāciju par katru fotogrāfiju, skatot to saistībā ar kvalitatīvajiem datiem un noteiktā valsts vēsturiskajā, sociālajā, kultūras un ekonomiskajā kontekstā.

Datu apstrādes procesu iespējams iedalīt vairākos secīgos posmos. Pirmais posms ietver kritēriju izveidi un kodēšanu, pēc kuriem tiks noskaidrota valodas situācija pilsētā. Kopumā zīmju analīzei tika atlasīti 26 kritēriji, kas ir būtiski sociolingvistisku pētījumu veikšanai, piemēram, pilsēta, iela, privātu vai valsts institūciju zīmes, zīmes atrašanās vieta, veikala veids, zīmes veids, valodu skaits, pirmā valoda.

Katram kritērijam ir vairāki apakškritēriji, piemēram, 1. kritērijs – pilsēta. Attiecīgi datorprogrammā katrai pilsētai tiek piešķirts savs kods, piemēram, Rēzekne – 1, Preiļi – 2, Ventspils – 3 utt. (sk. 8. attēlu). Apakškritēriju izveide ļauj fiksēt konkrētu valodas zīmju atrašanās vietu, veidu, kā arī konstatēt valodu skaitu, to kontaktus, līdz ar to – arī izplatību.

8. attēls. **Mainīgo vērtību (kritēriju) kodēšana**



Rezultātā izveidojas kategoriju jeb kritēriju tabula (sk. 9. attēlu), pēc kuras iespējams analizēt kvantitatīvos un kvalitatīvos datus.

9. attēls. **Kritēriju ievade mainīgo skatā *(Variable view)***

	Name	Type	Width	Decimals	Label	Values
9	Placement	Numeric	10	0	Placement of S...	{1, above th...
10	Number_lan...	Numeric	10	0	Number of Lang...	{1, 1 Langu...
11	Presence	Numeric	10	0	Presence of Pr...	{1, Proper N...
12	First_Langu...	Numeric	8	0	First Language,...	{1, Latvian}...
13	Second_La...	Numeric	8	0	Second Langua...	{1, Latvian}...
14	Third_Langu...	Numeric	8	0	Third Language...	{1, Latvian}...
15	Fourth_Lan...	Numeric	8	0	Fourth Languag...	{1, Latvian}...
16	Size	Numeric	8	0	Size on Multilin...	{1, The sam...
17	First_L_inSize	Numeric	8	0	First Language ...	{1, The sam...
18	Sec_L_inSize	Numeric	8	0	Second Langua...	{1, The sam...
19	Third_L_inSi...	Numeric	8	0	Third Language...	{1, The sam...
20	Fourth_L_in...	Numeric	8	0	Fourth Languag...	{1, The sam...
21	Type_of_Font	Numeric	8	0	Type of Font	{1, The sam...
22	Amount_of_...	Numeric	8	0	Amount of infor...	{1, The sam...
23	Translation	Numeric	8	0	Translation	{1, Word by...
24	Language_c...	Numeric	8	0	Language contact	{1, Mixing in...
25	Reaction	Numeric	8	0	Reaction by per...	{1, Negative...
26	Spoken	Numeric	8	0	Spoken with pe...	{1, Jes}...

Otrajā posmā datorprogrammā tiek ievadīti attēli un veikta valodas zīmju analīze pēc iepriekš minētajiem kritērijiem. Tādējādi tiek izveidota plaša fotogrāfiju un to vienību datubāze, kas turpmākajā procesā tiek izmantota, analizējot konkrētās teritorijas lingvistisko situāciju.

Katram attēlam tiek piešķirts numurs, ko veido pilsētas kods, iela un kārtas skaitlis, piemēram, ja zīme atrodas Rēzeknē, tās numurs tiek veidots šādi: 1 – Rēzekne; 2 – Latgales iela, 001 – zīmes kārtas skaitlis. Šādā veidā fotogrāfija tiek analizēta pēc visiem izvirzītajiem kritērijiem, piemēram, piederības (privāta vai valdības institūcija), funkcijas (izkārtne, plakāts, cita informācija u. tml.), valodas vienību skaita (dominējošā valoda pēc secības un lieluma) u. c. būtiska informācija, kas liecina par valodas lietojumu un izvēli.

Par valodas izvēli konkrētajā fotogrāfijā visbiežāk informāciju spēj sniegt kvalitatīvie dati – intervijas ar uzņēmuma darbiniekiem. Tomēr par to informē arī citi objekta darbības principi un faktori, piemēram, uzņēmuma darbības sfēra (viesnīcā paralēli valsts valodai tiek izmantota angļu, vācu vai kāda cita svešvaloda), pilsētvides etnolingvistiskais raksturojums, ģeogrāfiskais novietojums, vēsturiskie apstākļi, kas ietekmē noteiktas teritorijas komerciālo vidi.

Trešais posms piedāvā plašas datu apstrādes iespējas (jau iepriekš minēto biežumu un kopsakarību veidošanu) (sk. 5. tabulu).

Ar *SPSS* programmu ir iespējams valodas zīmju savstarpējās korelācijas, t. i., noskaidrot mainīgo savstarpējas atkarības pakāpi (sk. 6. tabulu).

Izmantojot Spīrmena korelācijas koeficienta aprēķinu (ar šo datu apstrādes veidu tiek noskaidrota mainīgo savstarpējā atkarība), ir secināts, ka konkrētajā piemērā

5. tabula. **Fotogrāfijas dati kopsakarību tabulā**

Town* Government or Private Crosstabulation

Count

		Government or Private			**Total**
		government	**private**	**not known**	
Town	Rēzekne	142	1043	3	1188
	Preiļi	23	28	0	51
	Ventspils	58	601	3	662
	Alytus	59	590	5	654
	Pärnu	63	228	0	291
	Druskininkai	83	473	4	560
	Narva	301	1039	138	1478
Total		**729**	**4002**	**153**	**4884**

6. tabula. **Korelācijas**

			Town	Number of Languages	Branch of shop
Spearman's rho	Town	Correlation Coefficient	1.000	.148**	.080**
		Sig. (2-tailed)	.	.000	.000
		N	4884	4883	4882
	Number of Languages	Correlation Coefficient	.148**	1.000	.051**
		Sig. (2-tailed)	.000	.	.000
		N	4883	4883	4881
	Branch of shop	Correlation Coefficient	.080**	.051**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	4882	4881	4882

***. Correlation is significant at the 0.01 level (2-tailed).*

starp pētāmajām pazīmēm pastāv statistiski nozīmīga korelācija ar varbūtību 99 % (par to liecina divas zvaigznītes pie aprēķinātajiem koeficientiem). Šādā veidā var tikt aprēķinātas jebkuru divu mainīgo korelācijas.

Rakstā dots tikai neliels ieskats *SPSS* programmas izmantošanā, apstrādājot datus humanitāro zinātņu pētījumos, piemēram, valodniecībā, veicot empīrisku pētījumu publiskajā telpā.

Veicot valodas situācijas izpēti publiskajā telpā, iespējams noskaidrot, kā valoda funkcionē dažādās sfērās sabiedriskajā vidē, kā tā pielāgojas apkārtējās vides izmaiņām, vai kultūrvide spēj ietekmēt valodas lietojumu konkrētā teritorijā utt. Izmantojot fotogrāfiju kā datu nesēju, iespējams fiksēt valodas izmaiņas, to funkcionalitāti arī diahroniski, tādējādi veidojot efektīvu komunikācijas stratēģiju publiskajā telpā.

Izmantojot *SPPS* programmā veidoto lingvistiskās ainavas datubāzi, projekta laikā izstrādāti vairāki individuālie pētījumi par valodu situāciju Baltijas valstīs, latgaliešu valodu publiskajā telpā, valodu kontaktiem vides reklāmās u. c. Pētnieki H. F. Martens, S. Lazdiņa, S. Pošeiko un S. Murinska ir sagatavojuši grāmatas „Linguistic Landscapes, Multilingualism and Social Change: Diversité des approches" (red. *C. Hélot, M. Barni, R. Janssens & C. Bagna*. Publ.: *Peter Lang*; grāmata ir sagatavošanā) nodaļu par lingvistisko ainavu sešās Baltijas valstu pilsētās.

Lingvistiskās ainavas izpēte sniedz ieguldījumu arī citās zinātņu nozarēs, piemēram, sociālajās, veicot detalizētu lingvistiskās vides izpēti, iespējams sniegt priekšlikumus uzņēmējiem un citiem interesentiem par valodas lietojumu publiskajā vidē, kas ir saistoši gan adresantam, gan arī adresātam.

Secinājumi

Datubāzu veidošana Statistisko datu apstrādes paketē dod plašas iespējas veikt dažādu kvantitatīvo un kvalitatīvo datu apstrādi un iegūt daudzveidīgu informāciju, ko izmantot zinātniski pētnieciskajā procesā. Datu grupēšana un apkopošana (grafiski un tabulu veidā) padara informāciju uzskatāmu, sistematizētu un pārredzamu, tādējādi ekonomējot laiku datu analīzei.

Apkopojot iepriekšminētās atziņas par darbu *SPSS* vidē, jāsecina, ka programma ir nozīmīgs datorrīks statistisko datu apstrādē, vērojot pētāmā objekta attīstības tendences sinhroniskā aspektā (kas ir pamats diahroniskajiem pētījumiem). To piedāvā noteikt un analizēt izveidotās datubāzes.

Būtiskākais trūkums ir programmas iegādes izmaksas salīdzinājumā ar *MS Excel*, bet tai ir svarīga nozīme pētnieciskajā procesā.

Literatūra

1. Gorter, Durk. *Linguistic Landscape: A New Approach to Multilingualism*. Clevedon : Multilingual Matters, 2006.
2. Kroplijs, Artūrs, Raščevska, Malgožata. *Kvalitatīvās pētniecības metodes sociālajās zinātnēs*. Rīga : Raka, 2010, 190 lpp.
3. Lazdiņa, Sanita, Pošeiko, Solvita, Marten, Heiko, F. Lingvistiskās ainavas metode – netradicionāls ceļš multilingvisma jautājumu izpētē un mācīšanā. No: *Tagad*. LVAVA zinātniski metodisks izdevums, 1. Rīga : LVAVA, 2008, 43.–49. lpp.
4. Rēzeknes Augstskolas ESF finansētais projekts *Teritoriālās identitātes lingvokulturoloģiskie un sociālekonomiskie aspekti Latgales reģiona attīstībā*. Pieejams http://tilra.ru.lv
5. *Svešvārdu vārdnīca :* vairāk nekā 16000 citvalodu cilmes vārdu un terminoloģisku vārdkopu. Red. sast. J.Baldunčiks; sast. K. Pokrotniece. Rīga : Jumava. 2007, 132. lpp.

Guna Pūce

Konkordances programmas izmantošana Rucavas izloksnes priedēkļverbu izpētē

Darbā „Priedēkļverbi Rucavas izloksnē" ir pētīti Rucavas izloksnes priedēkļverbi, to gramatiskās, semantiskās un fonētiskās iezīmes, salīdzinot ar latviešu literāro valodu.

Rucavas izloksnes īpatnības ir sāktas reģistrēt jau 19. gs beigās un 20. gs. sākumā, bet darbā izmantoti 20. un 21. gs. mijā vāktie Rucavas izloksnes materiāli.

Pētījumā izmantoti **izlokšņu materiāli**:

1) Liepājas Universitātes Humanitārās fakultātes studentu vākumi folkloras un dialektoloģijas prakses laikā (materiāli glabājas Kurzemes Humanitārajā institūtā), jāatzīmē, ka darbā saglabātas studentu rakstītās diakritiskās zīmes;
2) 2004. gadā izdotajā grāmatā „Mana novada valoda: Lejaskurzeme" ievietotie izlokšņu teksti;
3) 2007. gadā ar programmas „Letonika: pētījumi par vēsturi, valodu un kultūru" atbalstu izdotajā grāmatā „Rucavā, tur Paurupē…" publicētie Rucavas izloksnes teksti.

Raksturīgi, ka teicēji bijuši galvenokārt vecākās paaudzes pārstāvji.

Pētījumā izmantotas vairākas **dialektu pētīšanas metodes**.

Deskriptīvā (aprakstošā) metode – valodas pētīšanas metode, kuras pamatā ir aprakstošs skaidrojums (VPSV 2007, 38–39) – visvecākā un reizē arī mūsdienīga valodniecības metode. Tā ir fakta konstatēšana, neliela fakta analīze, aprakstīšana. Dialektoloģijā raksturīga izloksnes īpatnību reģistrēšana, skaidrošana un salīdzināšana ar atbilstošajām latviešu literārās valodas normām un tradīcijām, kā arī ar citām izloksnēm.

Matemātiskā, statistiskā metode – lietota vārdu biežuma noteikšanai konkrētajā apgabalā. Pētnieks rīkojas ar vidējiem biežumiem un relatīvo biežumu, kas ir novērotā biežuma attiecība pret teksta garumu (Koduhovs 1987, 250). Statistiskā metode izmantota Rucavas izloksnes priedēkļverbu biežuma noteikšanā, kā arī atsevišķu gramatisku un semantisku parādību analīzei. Šo metodi izmantojusi Ieva Ozola, rakstot promocijas darbu par divdabjiem Vitrupes izloksnē. Izlokšņu pētniekiem noderīgi arī iedzīvotāju skaita un nacionālā sastāva statistiskie pētījumi. Kā atzīmē Benita Laumane: „Statistisko datu salīdzināšana ļauj izsekot pagasta iedzīvotāju skaita un nacionālā sastāva izmaiņām gandrīz gadsimta garumā" (Laumane 2008, 364).

Datorlingvistiskas metodes. Datorlingvistika – starpdisciplināra zinātnes nozare, kurā pēta un pilnveido valodas izmantošanu un reproducēšanu elektroniskajos informācijas tehnoloģijas līdzekļos, kā arī pēta dažādus valodas aspektus, izmantojot datortehnoloģiju (VPSV 2007, 77). Datortehnoloģija ļauj valodas pētniecībā izmantot jaunas metodes, piemēram, konkordanču analīzi.

Ar konkordanču analīzes rīka palīdzību darbā tika:

1) ekscerpēti un apstrādāti Rucavas izloksnes priedēkļverbi;
2) noteikts verbu priedēkļu lietojuma biežums.

Aleksanders Krudens (*Alexander Cruden*) 1736. gadā publicēja karaļa Džeimsa Bībeles tulkojuma konkordanci. Tulkojuma priekšvārdā viņš raksta: „Konkordance ir vārdnīca vai Bībeles indekss, kurā visi vārdi, kas sastopami Svētajos Rakstos, ir sakārtoti alfabēta secībā, un blakus pievienotas dažādas teksta vietas, kur tie parādās, lai palīdzētu atrast pantus un lai varētu salīdzināt viena un tā paša vārda vairākas nozīmes" (Andronova 2009).

Izloksnes avoti apstrādāti, izmantojot konkordances programmu *Antconc 3.2.1.w*, ko izveidojis Antonijs Laurencs (*Laurence Anthony*) Japānā – Vasedas Universitātes Zinātnes un inženierijas fakultātē.

Lai programma tekstā spētu atpazīt visas diakritiskās zīmes, Rucavas izloksnes materiālus bija nepieciešams noformēt vienotā fontā. Pašlaik Liepājas Universitātē bez tradicionālajiem pieejami trīs veidu fonti, ko var izmantot dialektoloģijā: *Fontra*, *Indobalt*, *Palemonas* fonti. Darbā izmantots tikai *Palemonas* fonts – oriģināls lietuviešu valodas fonts, kuru lieto zinātniskos nolūkos. Zīmes veidotas uz latīņu alfabēta bāzes. *Palemonas* nosaukumam ir simboliska saikne starp 16. gs. populāru lietuviešu cilmes teoriju un jauno fontu.[1] Jāatzīmē, ka *Palemonas* fonts tiek nepārtraukti papildināts un precizēts.

Tā kā teksti no krājuma „Rucavā, tur Paurupē… Etnogrāfija, folklora, valoda" bija rakstīti *Fontra* un *Indobalt* fontā, tie tika pārveidoti *Palemonas* fontā.

Strādājot ar programmu *AntConc 3.2.1.w,* visi dokumenti tiek saglabāti vienkārša teksta formātā ar kodējumu *UTF-8*.

Priedēkļverbi tekstā tika marķēti ar zīmi <*pr*>, iekavās ierakstot LPA (jo izloksnes materiālus vākuši Liepājas Pedagoģijas augstskolas // Akadēmijas studenti), pierakstītāja iniciāļus (vārda, uzvārda pirmos burtus), gadu, vietu, jo dažkārt valodas īpatnības var izpausties tikai viena teicēja runā.

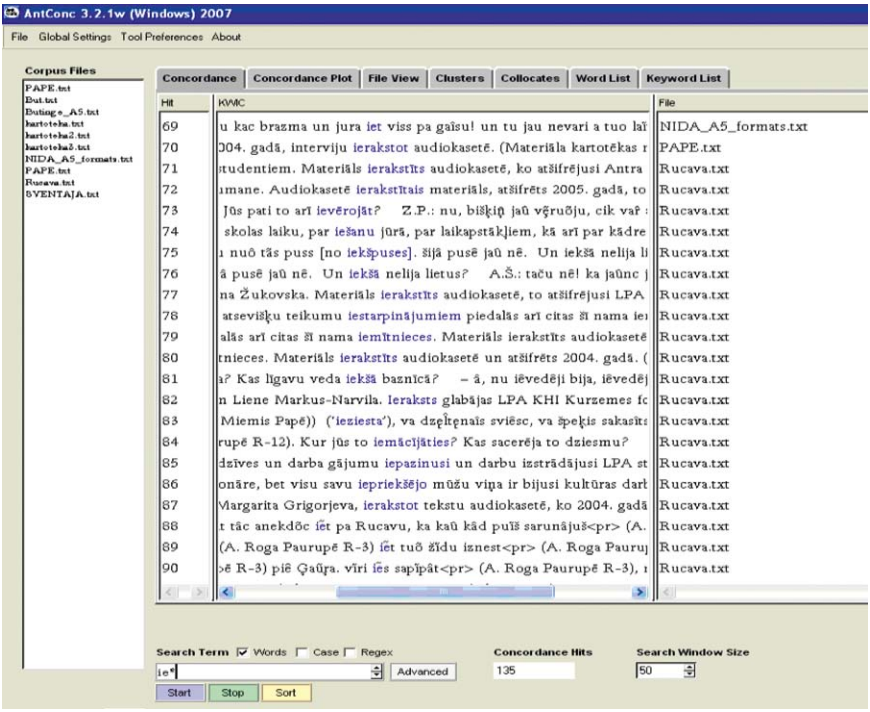
[1] http://www.vlkk.lt/lit/palemonas.html

Piemēram:

ār šitām krīčolēm atsvākļu nesanāks <pr> (LPA I.O. 1999 Rucava); *kâc kuŗ guļ vaî sêž atsarēmis <pr>* (LPA L.K. 1996 Rucava) *kuôku; ciēsu iŗ grūti iznīdēt <pr>* (LPA Dz.M. 2003 Rucava), *ja tā iêaûgusi druvā; Vīliņģis sakamp̃e <pr>* (LPA D.B. 1998 Rucava) *siêna klêpi uñ dacēle <pr>* (LPA D.B. 1998 Rucava); *paņèm <pr>* (LPA I.O. 1999 Rucava) *kurm̃tini ùn kurm̃stines, ārā aūksc; pakaci <pr>* (LPA I.L. 1997 Rucava) *kuñtini; uzsamāvu <pr>* (LPA D.B. 1998 Rucava) *liêliņus uñ gāju dzīt luōpus ganuôs; Mikurm̃ šuôdiên tās kričas galīgi nuôkritušas <pr>* (LPA D.B. 1998 Rucava)*; apsiêt <pr>* (LPA Z.P. 1999 Rucava) *kuñtīni; nuô skrizduōlēm iznāk <pr>* (LPA L.Z. 2003 Rucava) *laba zapte.*

Marķējums <*pr*> nepieciešams, lai ar *AntConc 3.2.1.w* programmas palīdzību atrastu tieši priedēkļverbus, jo, izmantojot, piemēram, regulāro izteiksmi *ie*,* tiks atrasti ne tikai priedēkļverbi, bet arī verbi **bez priedēkļiem** (piemēram, *te pēkšņi veš maīna kau ka virzienu jeb nak kau kac brāzma un jura iet viss pa gâisu* (A. Šteina R-33)), **lietvārdi** (piemēram, *piedalās arī citas šī nama iemītnieces* (M. Stammere R – 16)), un **apstākļa vārdi** (piemēram, *Kas līgavu veda iekšā baznīcā* (M. Stammere R – 16)) u. tml. (Sk. 1. attēlu.)

1. attēls. **Konkordances vaicājumam *ie****



Ar *AntConc 3.2.1.w* programmu iespējams automātiski uzzināt pētāmo priedēkļverbu daudzumu un noskaidrot, ka analizētajos tekstos konstatēti 1026 priedēkļverbi.

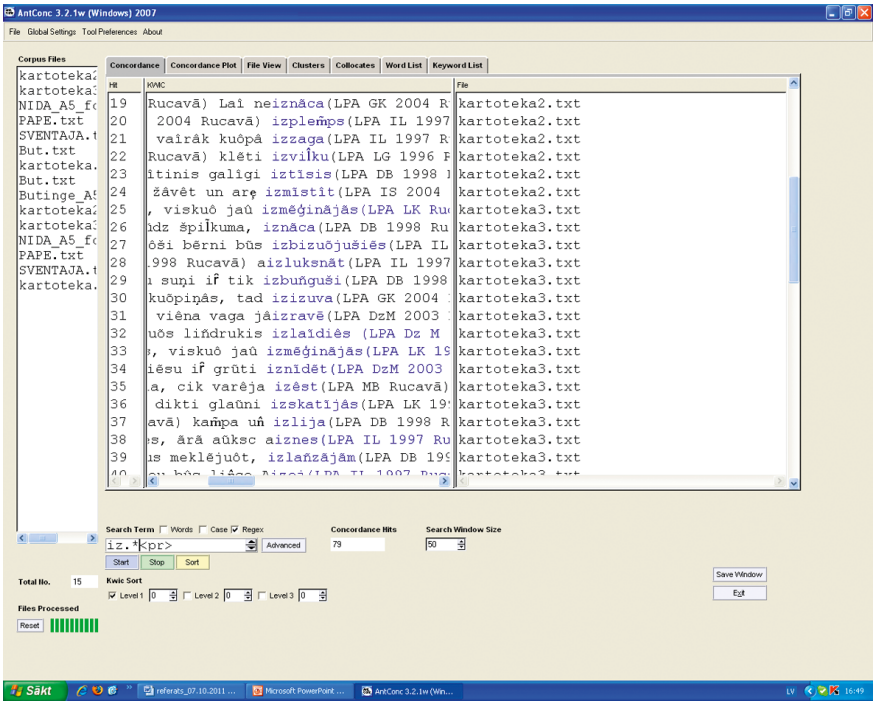
Lai aplūkotu vienu priedēkļverbu, piemēram, verbu ar priedēkli *iz-*, vaicājuma logā ieraksta iz.* <*pr*> un iegūst konkordanču piemērus, kas ļauj secināt, ka ar priedēkli *iz-* analizētajā materiālā atrodami 79 verbi (sk. 2. attēlu).

Mūsdienu latviešu literārās valodas verbu nenoteiksmes morfoloģiskajā sastāvā funkcionē vienpadsmit prefiksu: *aiz-*, *ap-*, *at-*, *ie-*, *iz-*, *no-*, *pa-*, *pār-*, *pie-*, *sa-*, *uz-* (Mllvg 1959, 351; Soida 2009, 227).

Arī Rucavas izloksnē konstatēti visi latviešu valodā sastopamie priedēkļi (4. attēls). Pētījumā secināts, ka Rucavas izloksnes analizētajos tekstos verbi, kas atvasināti ar priedēkli *aiz-*, konstatēti 38 reizes, t. i., 4 % no ekscerpētajiem verbiem; *ap-* 84 reizes, t. i., 8 %, *at-* 93 reizes, t. i., 9 %; Rucavas izloksnē verbs ar piedēkli *da-* lietots vienu reizi: *dacelt*, piemēram:

Vīliņģis sakaŭmpe <*pr*> (LPA D.B. 1998 Rucava) *siêna klêpi uņ dacêle* <*pr*> (LPA D.B. 1998 Rucava);

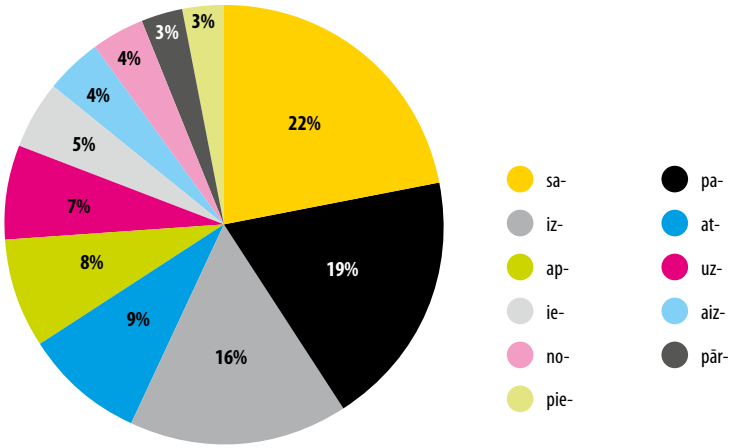
2. attēls. **Konkordances vaicājumam iz.***



Rucavas izloksnes analizētajos tekstos priedēklis *ie-* konstatēts 56 reizes, t. i., 5 % no ekscerpētajiem priedēkļverbiem; verbi ar priedēkli *iz-* (*is-*) konstatēti 162 reizes, t. i., 16 % no priedēkļverbiem; priedēklis *nuo-* konstatēts 41 reizi, t. i., apmēram 4 % no ekscerpētajiem priedēkļverbiem; priedēklis *pa-* konstatēts 193 reizes, t. i., 19 % no ekscerpētajiem priedēkļverbiem; priedēklis *pār-* konstatēts 28 reizes, t. i., 3 % no ekscerpētajiem priedēkļverbiem; verbi ar priedēkli *pie-* konstatēti 29 reizes, t. i., 3 % no priedēkļverbiem; Rucavas izloksnē priedēklis *sa-* ir produktīvs un analizētajos tekstos reģistrēts 226 reizes, t. i., 22 % no ekscerpētajiem priedēkļverbiem; priedēklis *uz-* konstatēts 71 reizi, t. i., 7 % no ekscerpētajiem priedēkļverbiem. (Sk. 3. attēlu.)

Datorlingvistikas pētīšanas metodes **mīnusi**, analizējot LiepU Kurzemes Humanitārā institūta materiālus:

3. attēls. **Priedēkļverbu lietojuma biežums Rucavas izloksnē**



1) studentu prakšu laikā savākto materiālu pārrakstīšana – darbietilpīgs un laikietilpīgs process;
2) datoros, kuros *Palemonas* fonts nav instalēts, dialektālos tekstus nevar izlasīt.

Datorlingvistikas metodes **plusi**, analizējot LiepU Kurzemes Humanitārā institūta materiālus:

1) iespēja ātri un bez kļūdām veikt darbietilpīgus uzdevumus;
2) var ātri sakārtot priedēkļverbus pēc konteksta;
3) var ātri sakārtot priedēkļverbus pēc biežuma;

4) izmantojot plašo *Palemonas* fontu, nav jāmeklē vajadzīgie simboli vairākos fontos;
5) datorizēti apstrādātos materiālus iespējams izmantot pētniecības darbā kompaktāk un ērtāk.

Avoti

1. LiepU Kurzemes Humanitārā institūta apvidvārdu kartotēka (rokrakstā LiepU Kurzemes Humanitārajā institūtā)
2. *Mana novada valoda: Lejaskurzeme*. Liepāja : LiePA, 2004, 262.–289. lpp.
3. *Rucavā, tur Paurupē…Etnogrāfija, folklora, valoda.* Atkārtots izdevums. Liepāja : LiePA, 2008. 401.–468. lpp.

Literatūra

1. Andronova, Everita. *Latviešu valodas seno tekstu korpusa 5 gadi*. Pieejams: http://www.ailab.lv/users/Everita/Latviesu_valodas_korpuss050209.ppt#391,13
2. Andronova, Everita. Mūsdienu latviešu valodas korpuss un tā izmantošana. *CLARIN* projekta seminārs 2009. gada 4.–5. februāris. Slīdrāde. Pieejams: http://www.ailab.lv/users/Everita/Latviesu_valodas_korpuss050209.ppt#391,13,P.S.
3. Koduhovs, Vitālijs. *Vispārīgā valodniecība*. Rīga : Zvaigzne, 1987.
4. Mllvg – *Mūsdienu latviešu literārās valodas gramatika*: 1. d. Rīga : LPSR ZA izd., 1959.
5. Soida, Emīlija. *Vārddarināšana*. Rīga : LU Akadēmiskais apgāds, 2009.
6. VPSV – *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga : LU Latviešu valodas institūts, 2007.

Diāna Bravacka

Datorprogramma „MonoConc Pro", tās izmantošana pētnieciskā darba izstrādē

Informācijas un komunikācijas tehnoloģiju (IKT) attīstība neapturami virzās uz priekšu, un ir tikai pozitīvi, ja datorprogrammas spēj atvieglot ne tikai ikdienas saziņu, bet arī nodrošināt rutinizēta darba optimizāciju pētniekiem dažādās zinātnes jomās.

Lietojumprogramma *MonoConc Pro for Windows* (turpmāk tekstā – *MP2.2*) ir izmantojama pētniecisko darbu izstrādē, ja jāiegūst konkrēta informācija, jāapstrādā plašs datu apjoms vai jāekscerpē no tā vienumi (Barlow 2003). Tā tika praktiski lietota, veidojot raksta autores maģistra darbu, kura izstrādes laikā izveidots pilnīgs I. Ziedoņa epifāniju korpuss.

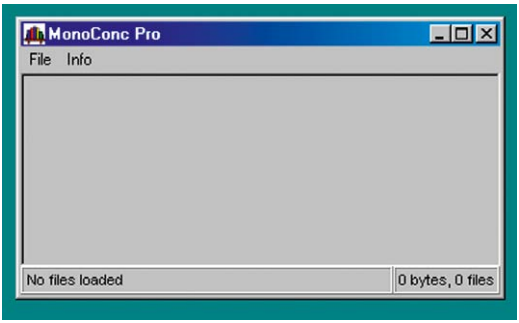
Šī lietojumprogramma datu analīzei un salīdzināšanai ļauj izvēlēties vairākus (tomēr savstarpēji samērojamus) korpusus vienlaikus. Nevar salīdzināt, piemēram, divu dažādu valodu tekstu korpusus. Tāpat var izmantot jau gatavus korpusus (piemēram, *Brown corpus, London–Lund Corpus, COBUILD Bank of English, British National Corpus, American National Corpus, MICASE, Helsinki Corpus, ICLE*), kuru anotācijas ir dotas arī *MP2.2* instrukcijas datnē, vai veidot savus tekstu korpusus, kā tas tika darīts raksta autores pētījumā. To var veikt, izmantojot tekstus, kas saglabāti *.txt* formātā.

Tātad, kā minēts, korpusus var veidot no jebkuriem tekstiem, kas ir programmai saprotamā (*.txt*) elektroniskā formātā, piemēram, skenējot grāmatas tekstu un pārveidojot ar lietojumprogrammas *ABBYY Reader* palīdzību *.doc* datnēs (kā to piedāvā *ABBYY Reader* brīvpiekļuves un bezmaksas programma), kuras savukārt viegli pārveidojamas *MP2.2* izmantojamās datnēs. Tādējādi iegūstami gramatiski nemarķēti korpusi, ar kuriem var strādāt *MP2.2* darba vidē dažādu pētījumu ietvaros. Domājams, ka pētījumu apjoms ir pietiekami plašs, jo programma *MP2.2* izmantojama kā filoloģijā, tā arī vēsturē, socioloģijā u. tml.

Sākot darbu un iekārtojot darbavietu uz datora darba virsmas, ieteicamas vairākas lietotāja paša jaunizveidotas mapes jeb direktoriji: tekstiem (korpusiem), darbvietu uzstādījumu saglabāšanai, rezultātu (atbilstības un biežumi) saglabāšanai, kā arī viena – sarežģītai meklēšanai. Tas ļaus lietotājam labāk orientēties lielajā datu apjomā un to analīzē. Tā, piemēram, jau minētajā pētījumā par I.Ziedoņa epifānijām.

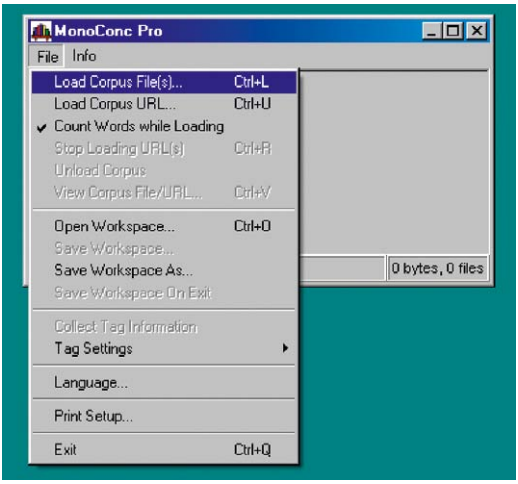
Kad iekārtotas datu glabāšanas mapes, var sākt darbu *MP2.2* programmā. Sākuma logs izskatās diezgan vienkāršs (sk. 1. attēlu). Loga vadības pogas izmantojamas pēc vajadzības, arī darba vides palielināšanai datora monitorā.

1. attēls. **MP2.2 programmā sākuma logs**



Pēc tās atvēršanas obligāti jāuzstāda valoda vai šrifts, kas adekvāti spēj atspoguļot teksta korpusa saturu. Ja nepieciešami statistiskie rādītāji par visiem vārdlietojumiem, ieteicams pēc valodas uzstādīšanas izvēlēties (atzīmēt ar ķeksīti) *File/Count Words while Loading* vārdu uzskaitei un tikai pēc tam sākt korpusu ielādi *(File/Load* CorpusFile*(s))*. Šīs izvēlnes iekārtojums redzams 2. attēlā.

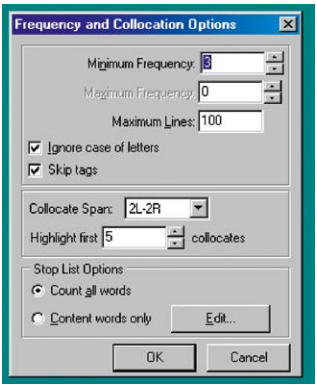
2. attēls. **Vārdlietojumu statistikas iegūšana**



Komandjoslā ***File*** izvēlnē *Load Corpus File(s)* katru reizi tiek piedāvāts ielādēt korpusu(s), kam jāatbilst iepriekš norādītajiem parametriem (proti – *.txt* formātā). Lai izvēlētos vairākas datnes reizē, var vajadzīgo iezīmēt (viena direktorija ietvaros), papildus turot nospiestu taustiņu *CTRL*, un apstiprināt savu izvēli ar *Open*. Pēc korpusu ielādēšanas var sākt meklēšanu un datu izgūšanu. Jāatceras, ka meklējumi notiek **visos** *ielādētajos* korpusos vienlaikus un to skaitu var redzēt programmas loga kreisajā apakšējā malā. Jebkurā brīdī no programmas var izmest liekos korpusus vai papildināt to ar jauniem.

Lai uzstādītu biežuma un atbilstības rādītājus, seko izvēlnei ***Frequency/Frequency Options***, kur dialoglogā uzstāda nepieciešamos parametrus: minimālo vārdu uzskaites soli, rindu skaitu, kolokāciju skaitu pa kreisi un pa labi no pētāmās leksēmas (tas svarīgāk kontekstanalīzes pētījumiem) u. c. nepieciešamos parametrus. Tas uzskatāmi redzams 3. attēlā.

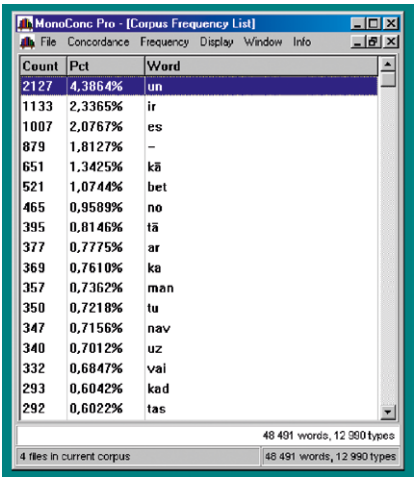
3. attēls. **Biežuma un atbilstības rādītāju uzstādīšanas dialoglogs**



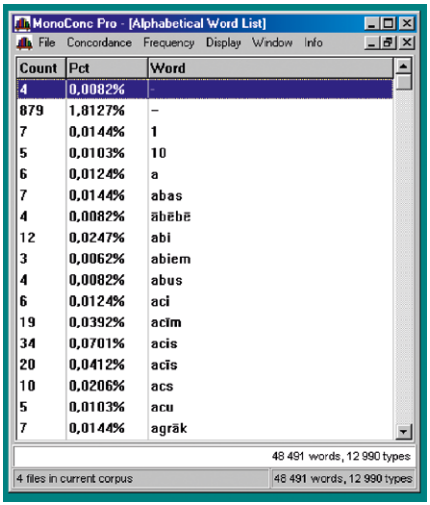
Statistiskos rādītājus par **visiem** vārdiem, kas lietoti korpusos vairāk nekā 44 reizes (šos uzstādījumus pēc nepieciešamības var arī mainīt, bet šoreiz izmantoti programmas noklusētie parametri), ar *MP 2.2* iegūst, izvēloties ***Frequency***/*Corpus Frequency Data*/***Frequency*** *Order* (sk. 4. attēlu), saglabājot un nodēvējot iegūto *.txt* datni pēc vajadzības.

Šādi saglabātu teksta datni pēc tam viegli pārnest *MS Excel* vidē un veikt citas nepieciešamās (piemēram, leksēmu atlases, aprēķinu, diagrammu veidošanas u. c.) darbības, veidot tabulas *.doc* datnē u. tml., jo pašā *MP2.2* vidē šādas darbības veikt nav iespējams. Ja ir jāiegūst **visi** vārdlietojumi alfabētiskā secībā, tad seko komandizvēlnei ***Frequency***/*Corpus Frequency Data*/***Alphabetic*** *Order* (sk. 5. attēlu).

4. attēls. **Korpusa vārdformu biežuma rādītājs**

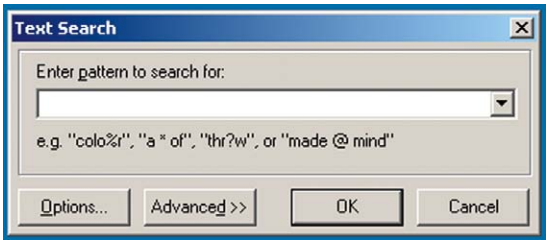


5. attēls. **Korpusa vārdformu biežuma saraksts alfabēta secībā**



Tālāk jau var strādāt ar konkrētu meklēšanu un meklēto datu šķirošanu, ar burtiem marķējot nepieciešamo informāciju. Ja ir jāiegūst dati ar konkrētām leksēmām vai frāzēm, tad komandjoslā ***Concordance***/*Search*.. izvēlas (ieraksta) nepieciešamo vārdformu/vārda daļu. Šis dialoglogs ir pietiekami vienkāršs (sk. 6. attēlu).

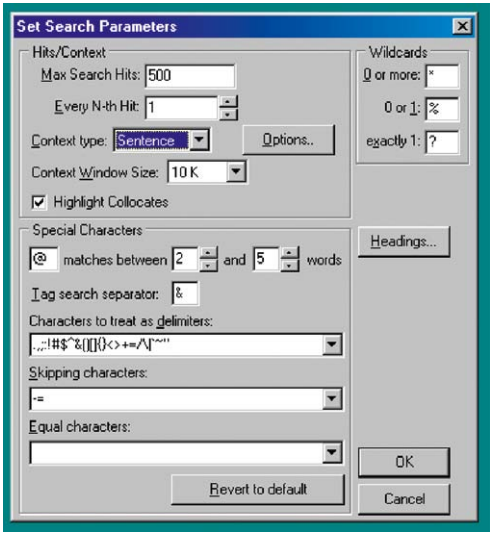
6. attēls. **Vaicājuma dialoglogs**



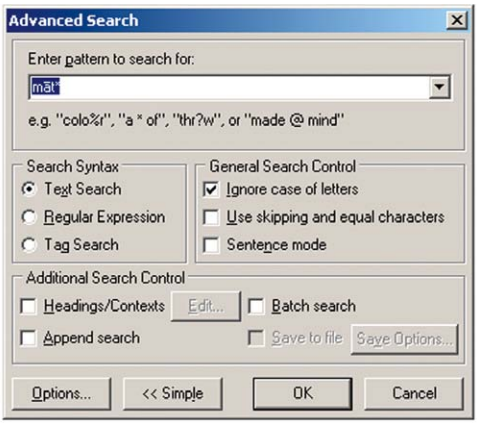
Programma piedāvā arī paplašinātu meklēšanu, kuras parametrus uzstāda ar ***Concordance****/Advanced Search…* (sk. 8. attēlu). vai izmanto ***Concordance****/Search Options* (sk. 7. attēlu). Te svarīga apakšizvēlne *Wildcards*, kas atrodas loga augšējā stūrī. Tā ļauj saprast, ka meklēšanā var tikt izmantoti t. s. rakstzīmju aizstājējsimboli:

- nevienu vai daudzas rakstzīmes virknes sākumā, vidū vai beigās – * (resp., zvaigznīte),
- nevienu vai tikai vienu rakstzīmi virknes sākumā, vidū vai beigās – % (resp., procenti),
- tikai (tieši) vienu rakstzīmi virknes sākumā, vidū vai beigās – ? (resp., jautājuma zīme).

7. attēls. **Meklēšanas parametri**



8. attēls. **Paplašinātas meklēšanas dialoglogs**



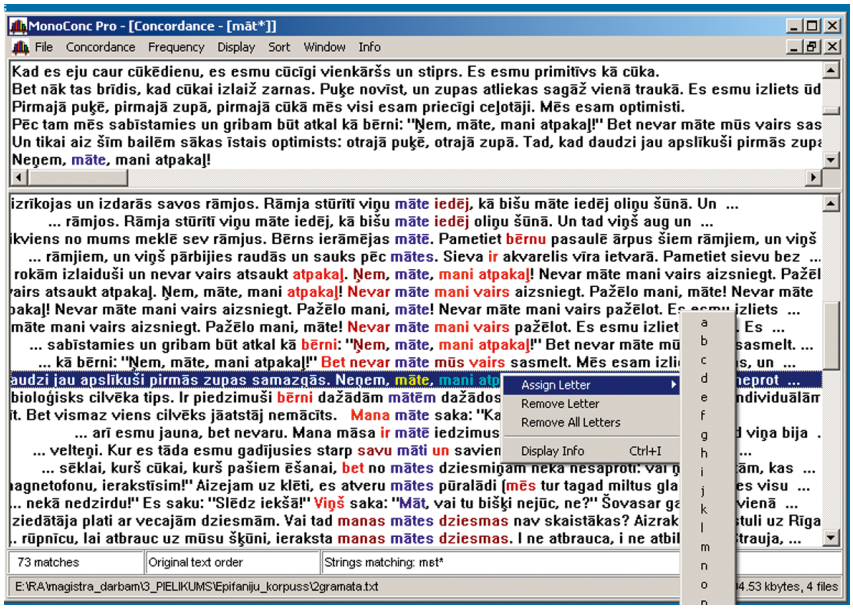
Meklēšanas rezultātā izgūst konkordanču rindas. Lai aplūkotu plašāku kontekstu, ir jāiezīmē konkrēta rindiņa. Plašāks konteksts ir redzams saskarnes augšējā logā. Ieskatu šajā darbībā var gūt 9. attēlā, kur parādīta arī iespēja „iezīmēt" (nosacīti – marķēt) nepieciešamo informāciju, ko pēc tam var atlasīt pēc marķējuma parametru atbilstības un izmantot atbilstoši pētījuma mērķim un nolūkam. To iegūst, ar peles labo pogu uzklikšķinot uz nepieciešamās rindiņas. Tepat arī var pēc vajadzības noņemt iezīmes – gan visas reizē (*Remove All Letters*), gan arī atsevišķas (*Remove Letter*), ja gadījies kļūdīties. Datus ar iezīmēm tāpat var turpināt apstrādāt un kārtot komandjoslā atrodamajā izvēlnē *Sort*.

Komandjoslā *Display* tiek piedāvātas dažas ļoti nepieciešamas iespējas:

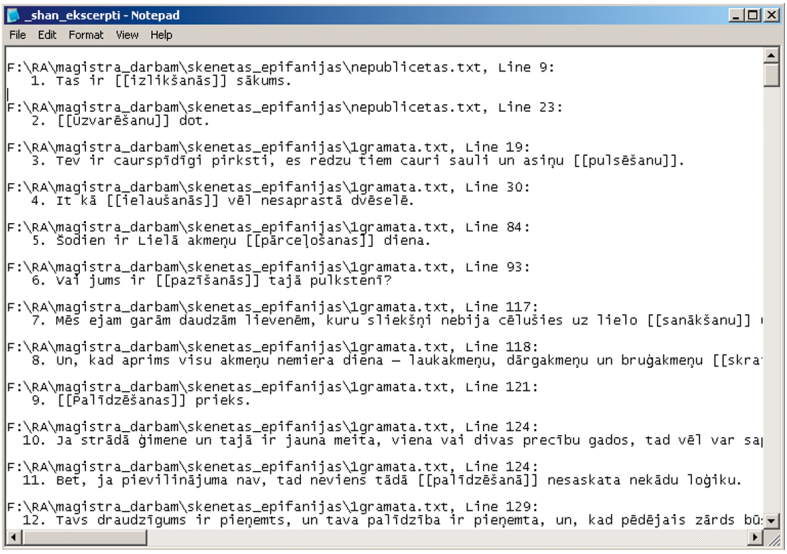
- *Highlight Collocates* – iezīmēt kolokāciju vārdus (noklusētajos parametros – sarkanos toņos);
- *Conceal Hits* – paslēpt meklēto leksēmu, aizstājot to ar noteiktas (noklusējumā – zilas) krāsas svītru;
- *Conceal Collocates* – paslēpt blakusesošās (biežāk lietotās, iepriekš uzrādītās) leksēmas, aizstājot to ar noteiktas (noklusējumā – sarkanas) krāsas svītru, atstājot pašu kontekstu;
- *Delete Item(s) (CTRL+D)* – izdzēst nevajadzīgo (iezīmēto) rindiņu, kā arī *Distribution,* kura tiks aplūkota nedaudz tālāk.

Atrastos piemērus ir iespējams saglabāt *.txt* datnē. 10. attēlā redzams, ka programma piedāvā saglabātu datni jau ar numurētiem ekscerptiem (leksēmām teikuma kontekstā), kas ir izcelti dubultās kvadrātiekavās. Turklāt pēc pieprasījuma uzstādījumos redzama arī meklējuma (vaicājuma) atrašanās vieta un korpuss, no kura tas izgūts.

9. attēls. **Vaicājuma *māt** konkordances**



10. attēls. **Saglabātu piemēru datnes fragments**



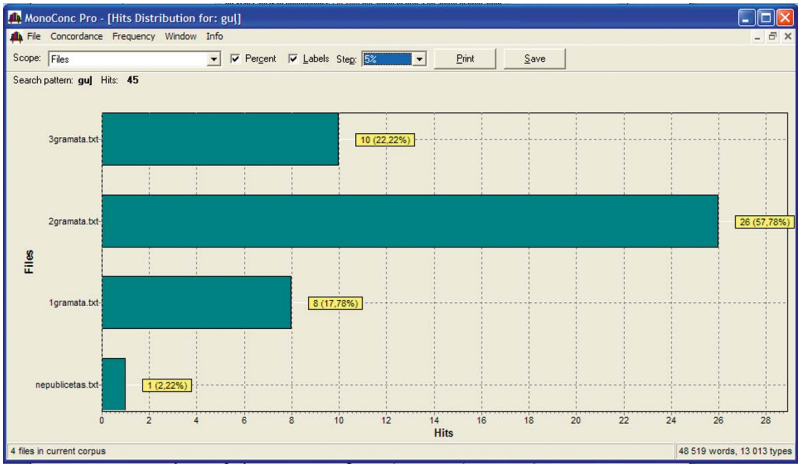
11. attēls. **Datu apstrāde *MS Excel* programmā**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Word: guļ															
2																
3	4-Left		3-Left		2-Left		1-Left		1-Right		2-Right		3-Right		4-Right	
4					5 un		4 viņa		3 un		3 guļ		4 un		3 un	
5					3 guļ		3 ceļi						3 uz			
6					3 lai		3 vēl						3 vēl			
7																

Lai iegūtu statistisku informāciju par vārdlietojumu savienojumiem ar konkrēto vārdu (piemēram, *guļ*), var izmantot *MP2.2* komandkarti *Frequency/Collocate Frequency*, kurā iegūtos datus saglabā teksta datnē un no tās pārnes uz *MS Excel* (ja nepieciešams ar šo izgūto informāciju veikt aprēķinus vai veidot diagrammas) vai *MS Word* darba vidi (skat. 11. attēlu):

Pētnieciskos nolūkos ļoti noderīga var būt informācija, ko var izgūt komandkartē *Display/Distribution*… Te tiek uzskatāmi parādīts, kurā korpusā, ja tiek lietoti vairāki, atrastā vārdforma lietota skaitliski un procentuāli biežāk. Programma *MP2.2* piedāvā jau gatavu skatu, kur redzams vārdformas biežums pa konkrētiem korpusiem (datnēm), turklāt jau salīdzinošā aspektā (sk. 12. attēlu).

12. attēls. **Vārdformas *guļ* biežuma izplatība dažādās korpusa datnēs**



Secinājumi

Lai gan *MP2.2* vidē ir iespējams veidot sarežģītus un daudzveidīgus vaicājumus, programma nedos gaidītos rezultātus, ja korpusā būs daudz grafiskas informācijas vai arī dzeja, kas līdzinās verlibram. Līdzīgas problēmas varētu rasties ar drāmas tekstiem, ja tie iepriekš nav apstrādāti, marķējot runātājus un viņu teikto. Paliek neatbildēts jautājums arī par remarku funkciju šādā tekstu apstrādē. Taču, ja visus iepriekšminētos šķēršļus novērš *.txt* datnē, tad pētījumu iespējas paliek tās pašas. Vislabāk šī lietojumprogramma ir izmantojama prozas tekstu analīzē.

Tātad programmai *MP2.2* ir raksturīga viegla un ātra meklēšanas iespēja plašā tekstu korpusā (ap 180 miljoniem vārdu – D.B.).

Uzskatāmu diagrammu veidošanas neiespējamība pašā *MP2.2* programmas vidē ir lielākais trūkums, bet pētniekam, kam ir pietiekama datorprasme, nav neiespējami operēt ar šajā lietojumprogrammā iegūto informāciju.

Literatūra

1. Barlow, Michael. *Concordancing and Corpus Analysis Using MP 2.2*. Athelstan, 2003 [programmas *MP 2.2* instrukcija *.pdf* datnē].

Latviešu valodas resursi un rīki e-vidē

Elektroniskās vārdnīcas

Latviešu literārās valodas vārdnīca

Pieejamība tiešsaistē: *http://www.tezaurs.lv/llvv/*

„Latviešu literārās valodas vārdnīcas" (8 sējumi) mašīnlasāmā versija izveidota LU MII Mākslīgā intelekta laboratorijā ar Valsts pētījumu programmas „Letonika" finansiālu atbalstu. Vārdnīcā ir ap 64 000 šķirkļu. Papildiespējas: iespējams izlocīt šķirkļa vārdu un noklausīties šķirkļa vārda izrunu (tikai daļai šķirkļu).
Mobilā versija: *www.tezaurs.lv/wap*

Skaidrojošā vārdnīca (pavasara versija (21.03.2012.))

Pieejamība tiešsaistē: *http://termini.lza.lv/term.php*

Vārdnīca tiek veidota, izmantojot līdz šim publicētās latviešu valodas skaidrojošās vārdnīcas. Tajā ir vairāk nekā 199 014 šķirkļu. Sastādījis A. Spektors.
Mobilā versija: tezaurs.lv/wap (paredzēta izmantošanai mobilajā telefonā)

Latviešu valodas vārdnīca

Pieejamība tiešsaistē: *http://www.tezaurs.lv/lvv/*

Vārdnīcā ir ap 30 000 šķirkļu. Izmantoti Latviešu valodas vārdnīcas (Rīga : Avots, 1987) materiāli. Projektu finansēja Jaunmeksikas Valsts universitāte (*New Mexico State University*) un Laskrusas Rotarī klubs (*Las Cruces Rotary Club*). Mašīnlasāmā versija izstrādāta LU MII.

K. Mīlenbaha un J. Endzelīna „Latviešu valodas vārdnīca"

Pieejamība tiešsaistē: *http://www.tezaurs.lv/mev/*

Vārdnīcā ir aptuveni 75 000 šķirkļu (4 pamatsējumi un 2 papildu sējumi). Vārdnīcas elektroniskā versija izstrādāta LU MII. 2000. gadā ar Latviešu Fonda finansiālu atbalstu tika ieskenēti vārdnīcas pamatsējumi un sākts darbs pie MEV elektroniskās versijas izstrādes. 2003. un 2004. gadā ar Latvijas Republikas Izglītības un zinātnes ministrijas finansiālu atbalstu datorā tika ievadīti K. Mīlenbaha un J. Endzelīna „Latviešu valodas vārdnīcas" papildsējumi, kā arī pilnveidota vārdnīcas pamatsējumu elektroniskā versija.

Ieskenējot un manuāli ievadot vārdnīcu, pilnībā ir saglabāta oriģinālā rakstība; daudzveidīgas meklēšanas iespējas, piemēram, iespējams meklēt šķirkļa vārdu gan mūsdienu ortogrāfijā, gan oriģinālrakstībā, gan meklēt pēc avota u.tml.
Jebkuram ir brīvi pieejama MEV izmēģinājuma versija (*A* un *Ā* burts).
Noslēdzot MEV konsultanta līgumu un saņemot lietotāja vārdu un paroli, vārdnīcas tekstu var brīvi izmantot izglītības un pētniecības mērķiem.

Mūsdienu latviešu valodas vārdnīca (saīsinājums – MLVV)

Pieejamība tiešsaistē: *http://www.tezaurs.lv/mlvv/*

Interneta lietotājiem tiek piedāvāta jaunas, mūsdienīgas latviešu valodas skaidrojošās vārdnīcas pirmā daļa no *A* līdz *M* burtam. Vārdnīcu LU Latviešu valodas institūtā veido neliels autoru kolektīvs Ievas Zuicenas vadībā. Darbs pie vārdnīcas tiek turpināts, ir paredzēts pakāpeniski publicēt internetā arī pārējo vārdnīcas tekstu, kamēr lasītājiem būs pieejama visa vārdnīca pilnā apjomā (~70 000 šķirkļu). Vārdnīca paredzēta plašam lietotāju lokam.

Latviešu valodas sinonīmu vārdnīca

Pieejamība tiešsaistē: *http://www.letonika.lv/groups/default.aspx?g=5&r=1108*

Vārdnīcā apkopoti vairāk nekā 84 500 latviešu valodas vārdu, katram no tiem ir dots viens vai vairāki sinonīmi. „Latviešu valodas sinonīmu vārdnīca" ir lielisks palīgs vēstuļu un sacerējumu rakstīšanā, kā arī noder, lai mācītos, pētītu un bagātinātu latviešu valodu. Izstrādātājs: SIA *Tilde*

Latviešu-lietuviešu vārdnīca

Pieejamība tiešsaistē: *http://www.letonika.lv/groups/default.aspx?g=2&r=10621063*

„Latviešu-lietuviešu vārdnīcā" ir ap 43 000 šķirkļu. Vārdnīcas saturu izstrādājis Vītauta Dižā Kauņas Universitātes Letonikas centra vadītājs *doc. dr.* Alvīds Butkus (*Alvydas Butkus*), elektroniskā versija veidota SIA *Tilde*. Vārdnīca ir brīvi pieejama tiešsaistē.

Lietuviešu-latviešu vārdnīca

Pieejamība tiešsaistē: *http://www.letonika.lv/groups/default.aspx?g=2&r=10631062*

„Lietuviešu-latviešu vārdnīcā" ir ap 60 000 vārdu un izteicienu. Tā ir 1995. gadā publicētās, pašlaik plašākās un modernākās lietuviešu-latviešu vārdnīcas (autori: Jons Balkevičs, Laimute Balode, Apolonija Bojāte, Valters Subatnieks, redaktors Alberts Sarkanis) elektroniskā versija. Vārdnīcas lietošanu ērtu padara īpašā meklēšanas un šķirkļu pārlūkošanas vide. Elektroniskā versija veidota SIA *Tilde*.

Igauņu-latviešu vārdnīca

Pieejamība tiešsaistē: *http://www.letonika.lv/groups/default.aspx?g=2&r=10611062*

„Igauņu-latviešu vārdnīcas" pamatā ir igauņu un latviešu valodnieka, tulkotāja un pasniedzēja Kārļa Abena (1896–1976) sastādītā un 1967. gadā izdotā vārdnīca. 2007. gadā ar Latvijas Valsts valodas aģentūras atbalstu vārdnīcu pārskatīja un papildināja igauņu un latviešu valodas speciālistes, tulkotājas un redaktores – Andra Kalnača, Ērika Krautmane, Jana Šteinberga-Ranki un Urve Aivare. No vārdnīcas tika izņemta liela daļa padomju propagandas šķirkļu, kā arī tā tika papildināta ar aptuveni 2 000 jauniem šķirkļiem, kas aptver biežāk lietotos mūsdienu igauņu valodas vārdus un terminus. Vārdnīcā ir aptuveni 26 000 vārdu tulkojumu un izteicienu.

Citas noderīgas saites

LU MII: Mākslīgā intelekta laboratorijas vārdnīcu serveris: *http://www.tezaurs.lv/*

Vārdnīcas internetā: *http://www.vvk.lv/?sadala=180*

Vārdnīcas un terminu datu bāzes: *http://www.hzf.lu.lv/saites/*

Letonika.lv (SIA *Tilde* vārdnīcas): *http://www.letonika.lv/groups/default.aspx?g=2*

Terminu datubāzes

Valsts valodas centra Terminu datubāze

Pieejamība tiešsaistē: *www.vvc.gov.lv/advantagecms/LV/terminologija/terminudatubaze.html*

SIA *Tilde* terminoloģijas portāls

Pieejamība tiešsaistē: *http://termini.tilde.lv/DesktopDefault.aspx*

Lielā terminu vārdnīca

Pieejamība tiešsaistē: *http://termini.lv*

Latvijas nacionālo reāliju standartizēts tulkojums

Pieejamība tiešsaistē: *http://realijas.venta.lv/*

Akadēmiskā terminu datubāze *AkadTerm*

Pieejamība tiešsaistē: *http://termini.lza.lv/term.php*

Elektroniskie latviešu valodas mācību līdzekļi

Latviešu valoda pieaugušajiem (A1 valodas prasmes līmenis)

Pieejamība tiešsaistē: *http://www.valoda.lv/Papildus_Materiali/eapmaciba2/default.htm*

Mācību līdzeklis paredzēts latviešu valodas apguvei e-vidē. Mācību līdzeklī ir 7 tēmas: *Iepazīšanās, Adrese, Ģimene, Dzīvesvieta, Pilsētā, Ikdiena, Ēdienreizes.* Katras tēmas beigās ir vārdnīca un interaktīvi uzdevumi. Mācību līdzeklī ir neliela latviešu valodas gramatika, papildus iespējams arī apgūt alfabētu, skaitļus, dienu un mēnešu nosaukumus. Mācību līdzeklis ir izstrādāts Latviešu valodas aģentūrā.

Latviešu valoda sākumskolai

Pieejamība tiešsaistē: *http://valoda.ailab.lv/latval/sakumskolai/galvena.swf*

Paredzēts vispārizglītojošo skolu skolēniem. Teorija izklāstīta, izmantojot animāciju un skaņu, ir 180 interaktīvu uzdevumu. Mācību materiāls izstrādāts LU Matemātikas un informātikas institūtā ar Latviešu valodas aģentūras atbalstu.

Latviešu valoda pamatskolai

Pieejamība tiešsaistē: *http://valoda.ailab.lv/latval/pamatskolai/galvena.swf*

Mācību līdzeklis paredzēts vispārizglītojošo skolu skolēniem. Teorija izklāstīta, izmantojot animāciju un skaņu; ir vairāk nekā 240 interaktīvu uzdevumu. Mācību materiāls izstrādāts LU Matemātikas un informātikas institūtā LIIS projekta laikā.

Citi elektroniskie mācību līdzekļi

Pieejamība tiešsaistē: *http://valoda.ailab.lv/latval/*

Latviešu tautas pasakas un teikas

Pieejamība tiešsaistē: *http://valoda.ailab.lv/folklora/pasakas/*

Prof. P. Šmits ir apkopojis latviešu pasakas un teikas 15 sējumos, kuri pirmoreiz izdoti 1925.–1937. g. Pirmā grāmata satur „Ievadu" 130 lpp. apjomā, kur sniegti visi nepieciešamie paskaidrojumi. Pasakas klasificētas pēc pasaku tipiem. „Sorosa Fonds – Latvija" finansētā projekta laikā datorā tika ievadītas visas 15 grāmatas, saglabājot oriģinālrakstību. Projekts īstenots LU MII Mākslīgā intelekta laboratorijā.

Latviešu tautas ticējumi

Pieejamība tiešsaistē: *http://valoda.ailab.lv/folklora/ticejumi/*

Prof. P. Šmita apkopotie latviešu tautas ticējumi. Vairāk nekā 36 000 vienību. Ticējumus iespējams meklēt pēc atslēgas vārda. Projekts īstenots LU MII Mākslīgā intelekta laboratorijā.

Latviešu sakāmvārdu datorfonds

Pieejamība tiešsaistē: *http://valoda.ailab.lv/folklora/sakamvardi/*

Izveidots LU MII sadarbībā ar Latviešu folkloras krātuvi. Datorfondā ietvertas vairāk nekā 20 000 vienības.

Folkloristikas elektroniskā bibliotēka (FEB)

Pieejamība tiešsaistē: *http://www.korpuss.lv/feb/*

Folkloristikas elektroniskā bibliotēka ir folkloras pētniecisko materiālu glabātuve, kas veidota ar nolūku padarīt pieejamus dažādu laika perioda materiālus. Šobrīd FEB ir ievietoti ap 200 rakstu un to turpinājumi, kas publicēti galvenokārt 19. gs. beigu un 20. gs. sākuma latviešu periodikā. Projekts īstenots ar Valsts Kultūrkapitāla fonda finansiālo atbalstu.

Krišjāņa Barona Dainu skapis

Pieejamība tiešsaistē: *http://www.dainuskapis.lv*

Virtuālais Krišjāņa Barona Dainu skapis ļauj ikvienam interneta lietotājam ielūkoties Dainutēva vākumā un meklēt savu tautasdziesmu. Turklāt virtuālā Dainu skapja veidotāji ļauj mums ņemt līdzi latvju dainas arī savā mobilajā telefonā
Mobilais Dainuskapis: *http://www.dainuskapis.lv/mobile*
Dainuskapis WAP: *http://www.dainuskapis.lv/wap/*

Folklora, mitoloģija un pseidomitoloģija literatūrā

Pieejamība tiešsaistē: *http://www.liis.lv/folklora/folkloristika/liter/*

Tiešsaistē lasāma Ausekļa darbu izlase, Garlība Merķeļa „Vidzemes senatne", Andreja Pumpura eposs „Lāčplēsis" un dzejoļu krājumā „Tēvijā un svešumā" ietvertie dzejoļi, kā arī Anša Lerha-Puškaiša sacerējums „Kurbads – senlatvju varonis".

Latviešu valodas folkloras krātuves interneta vietne

Pieejamība tiešsaistē: *http://www.lfk.lv*

LU Literatūras, folkloras un mākslas institūta Latviešu valodas folkloras krātuvē atrodama dažāda informācija gan par latviešu folkloru, gan Latviešu valodas folkloras krātuvi (LFK), piemēram, par LFK nodibināšanu un vēsturi, tās ieskaņojumu un attēlu kolekcijām. Vietnē ir apskatāmi un dzirdami folkloras materiāli, ir iespējams noskaidrot folklorā lietotu, bet mūsdienu cilvēkam nepazīstamu vārdu nozīmi.

Latviešu literatūras klasika

Pieejamība tiešsaistē: *http://www.korpuss.lv/klasika*

Vietnē apkopoti latviešu literatūras vecmeistaru darbi, kurus neaizsargā autortiesību likums: Apsīšu Jēkaba, Brāļu Kaudzīšu, Māteru Jura, R. Blaumaņa, A. Deglava, J. Purapuķes, Zeiboltu Jēkaba, Raiņa, A. Pumpura darbi; arī T. Zeiferta „Latviešu rakstniecības vēsture". Projekts realizēts LU MII ar Kultūrkapitāla fonda atbalstu.

Latviešu literatūras interneta bibliotēka (Letonika.lv)

Pieejamība tiešsaistē: *http://letonika.lv/literatura/default.aspx?*

Šobrīd bibliotēkā atrodami vairāk nekā 30 latviešu autoru darbi: dzeja, lugas, stāsti, noveles un citu žanru sacerējumi – kopumā ap 200 pilnteksta darbu kolekcijas (vairāk nekā 22 000 lappušu). Šeit atrodama arī informācija par 22 latviešu klasiķiem. Bibliotēka tiek papildināta.

Latviešu valodas seno tekstu korpuss

Pieejamība tiešsaistē: *http://www.korpuss.lv/senie/*

Korpusā iekļauti gan 16. gs., gan 17. gs., gan 18. gs. sākuma latviešu rakstu pieminekļi un to indeksi. Senie teksti elektroniskā veidā ir pieejami un brīvi izmantojami pētnieciskiem mērķiem, pateicoties 90. gadu vidū saņemtajam Sorosa fonda – Latvija atbalstam, 2002. gadā piešķirtajam Kultūrkapitāla fonda un LU finansējumam, kā arī Izglītības un zinātnes ministrijas finansējumam 2003. gadā. 2004. gadā, izmantojot VKKF piešķirto radošo stipendiju, seno tekstu korpuss tika papildināts ar vēl diviem avotiem. Latviešu valodas seno tekstu uzkrāšanu un apstrādi nodrošina LU Humanitāro zinātņu fakultātes (agrāk Filoloģijas un mākslas zinātņu fakultātes) Baltu valodniecības katedra un LU Matemātikas un informātikas institūta Mākslīgā intelekta laboratorija.

Līdzsvarots mūsdienu latviešu valodas tekstu korpuss

Pieejamība tiešsaistē: *http://www.korpuss.lv*

Līdzsvarots 3,5 miljonus vārdlietojumu liels mūsdienu latviešu valodas tekstu korpuss, ko veido publicistikas, daiļliteratūras, zinātniskie un populārzinātniskie teksti, normatīvie akti, Saeimas stenogrammas un citi teksti. Teksti ir morfoloģiski marķēti. Vaicājumiem korpusā jāizmanto pārlūkprogramma *Bonito*. Korpuss izstrādāts LU Matemātikas un informātikas institūtā. Tas tapis sadarbībā ar Latviešu valodas aģentūru.

Latviešu valodas tīmekļa korpuss

Pieejamība tiešsaistē: *http://www.korpuss.lv*

Ar *SemTi-Kamols* gramatisko analizatoru automātiski, eksperimentāli marķēti tekstu fragmenti no Latvijas meklētāja savāktajām tīmekļa lapām. Apjomīgais analīzes process tika īstenots *BalticGrid* infrastruktūrā.

Latvijas Republikas 5.–9. Saeimas sēžu stenogrammu korpuss

Pieejamība tiešsaistē: *http://www.korpuss.lv*

Publiski pieejams LR Saeimas stenogrammu korpuss. Apjoms: aptuveni 4 milj. vārdlietojumu. Marķēti runātāji.

Latvijas Nacionālā digitālā bibliotēka

Pieejamība tiešsaistē: *http://www.lnb.lv/lv/digitala-biblioteka*

Digitālās bibliotēkas mērķis ir nodrošināt Latvijas Nacionālās bibliotēkas un citu atmiņas institūciju krājumu digitalizāciju, padarot tos pieejamus internetā. Šobrīd digitālajā bibliotēkā atrodamas digitalizētu laikrakstu, attēlu, karšu, grāmatu, nošu un skaņu ierakstu kolekcijas. Resursi tiek papildināti.

Dzīvesstāsts. Nacionālās mutvārdu vēstures projekts (NMV)

Pieejamība tiešsaistē: *http://www.dzivesstasts.lv/lv/default.htm*

Nacionālās mutvārdu vēstures krājumā apkopoti atmiņu stāstu audio ieraksti un rakstītie teksti. NMV krājuma veidošana sākta 1992. gadā. Tajā galvenokārt ir vecākās paaudzes dzīvesstāstu ieraksti – latviešu literārajā valodā, dialektos un dažādās izloksnēs, ierakstīti Latvijā un ārzemēs – Norvēģijā, Zviedrijā, Anglijā, ASV. NMV krājumā ir apmēram 3 000 autoru audio ierakstu.

Morfoloģiskais analizators

Pieejamība tiešsaistē: *http://eksperimenti.ailab.lv/kamols/*

Projekta *SemTi-Kamols* laikā izstrādātā latviešu valodas morfoloģiskā analizatora vienkāršots tīmekļa serviss, kurš dotajām vārdformām izveido atbilstošu pamatformu sarakstu, norādot arī vārdšķiras. Ja vārdforma ir daudznozīmīga, tiek piedāvātas visas iespējamās pamatformas. Analizatora leksikons (~50 000 leksēmu) un vārddarināšanas likumi vēl ir samērā nepilnīgi – ne visām vārdformām tiks noteiktas (visas) pamatformas. Analizators izstrādāts LU MII.

Sintaktiskais analizators

Pieejamība tiešsaistē: *http://www.semti-kamols.lv/?sadala=218*

Analizētājs veic teikumu gramatisko (morfoloģisko un sintaktisko) analīzi – nosaka katra vārda morfoloģisko formu un, balstoties uz tām, nosaka iespējamo teikuma sintaktisko struktūru. Sintaktiskais analizators LU MII.

Marķētājs

Pieejamība tiešsaistē: *http://www.semti-kamols.lv/?sadala=217*

Marķētājs ir pusautomātisks programmrīks, kas paredzēts latviešu valodas tekstu morfoloģiskai un sintaktiskai analīzei. Ar tā palīdzību katrai vārdformai tiek dots morfoloģiskais raksturojums un, ja iespējams, noteikta arī vārda sintaktiskā loma teikumā. Sintaktiskais analizators LU MII.

Vārda analīze

Pieejamība tiešsaistē: *http://www.letonika.lv/groups/default.aspx?g=5&r=1100*

Lai varētu ātri noskaidrot, kā pareizi rakstāmi latviešu vārdi dažādos locījumos, *Tilde* ir izveidojusi pareizrakstības uzziņu sistēmu. Tajā ietverta vairāk nekā 10 gadu laikā *Tildē* uzkrātā latviešu valodas vārdu datu bāze, kas ir lielākā un pilnīgākā šāda veida krātuve. Tās mērķis ir sekmēt latviešu literārās valodas normu ievērošanu un lietošanu un visiem interesentiem palīdzēt apgūt latviešu valodu. *Tildes* tehnoloģija ļauj ar automātisko līdzekļu palīdzību noskaidrot vairumam latviešu valodas vārdu gramatisko informāciju un vārdformas.

Materiāls sagatavots, izmantojot internetā pieejamo informāciju un *CLARIN* projekta laikā sagatavoto pārskatu par Latvijā veidotajiem valodas resursiem elektroniskā formā un valodas apstrādes rīkiem (sk. *www.clarin.lv*).

Ziņas par autoriem

Juris Baldunčiks	*Dr. philol.*, Ventspils Augstskola
Uldis Bojārs	*Dr. sc. comp.*, Latvijas Universitātes Sociālo un politisko pētījumu institūts
Diāna Bravacka	*Mg. philol.*, Rēzeknes Augstskola
Līva Brice	*Mg. sc. soc.*, Latvijas Universitātes Sociālo zinātņu fakultāte
Anna Briška	Rēzeknes Augstskola, projekta „Humanitārās izglītības pētniecības infrastruktūras izveide Austrumlatvijā, Lietuvā" koordinatore
Eduards Cauna	*Mg. phys.*, Ventspils Augstskola
Gita Elksnīte	*Dr. philol.*, Liepājas Universitāte
Normunds Grūzītis	*Dr. sc. comp.*, LU Matemātikas un informātikas institūts, Mākslīgā intelekta laboratorija
Anita Helviga	*Mg. philol.*, Liepājas Universitāte
Lienīte Litavniece	*Dr. oec.*, Rēzeknes Augstskola
Sandra Murinska	*Mg. philol.*, Rēzeknes Augstskola
Jānis Naglis	*Mg. sc. comp.*, Ventspils Augstskola
Gunta Nešpore	*Mg. philol.*, LU Matemātikas un informātikas institūts, Mākslīgā intelekta laboratorija
Jānis Pencis	*Mg. sc. soc.*, LU Sociālo zinātņu fakultāte, komunikācijas zinātnes doktorants
Guna Pūce	*Mg. philol.*, Liepājas Universitāte
Guna Rābante	*Bc.*, LU Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijas asistente
Valda Rudziša	*Dr. philol.*, Ventspils Augstskola
Jānis Sīlis	*Dr. philol.*, Ventspils Augstskola
Līga Vogina	*Mg. philol.*, Latvijas Universitātes Humanitāro zinātņu fakultāte

Latviešu valoda digitālajā vidē: datorlingvistika
Informatīvi izglītojoša semināru cikla materiāli. Rakstu krājums

Valsts aģentūra „Latviešu valodas aģentūra"
Lāčplēša iela 3-5, Rīga, LV-1011
www.valoda.lv