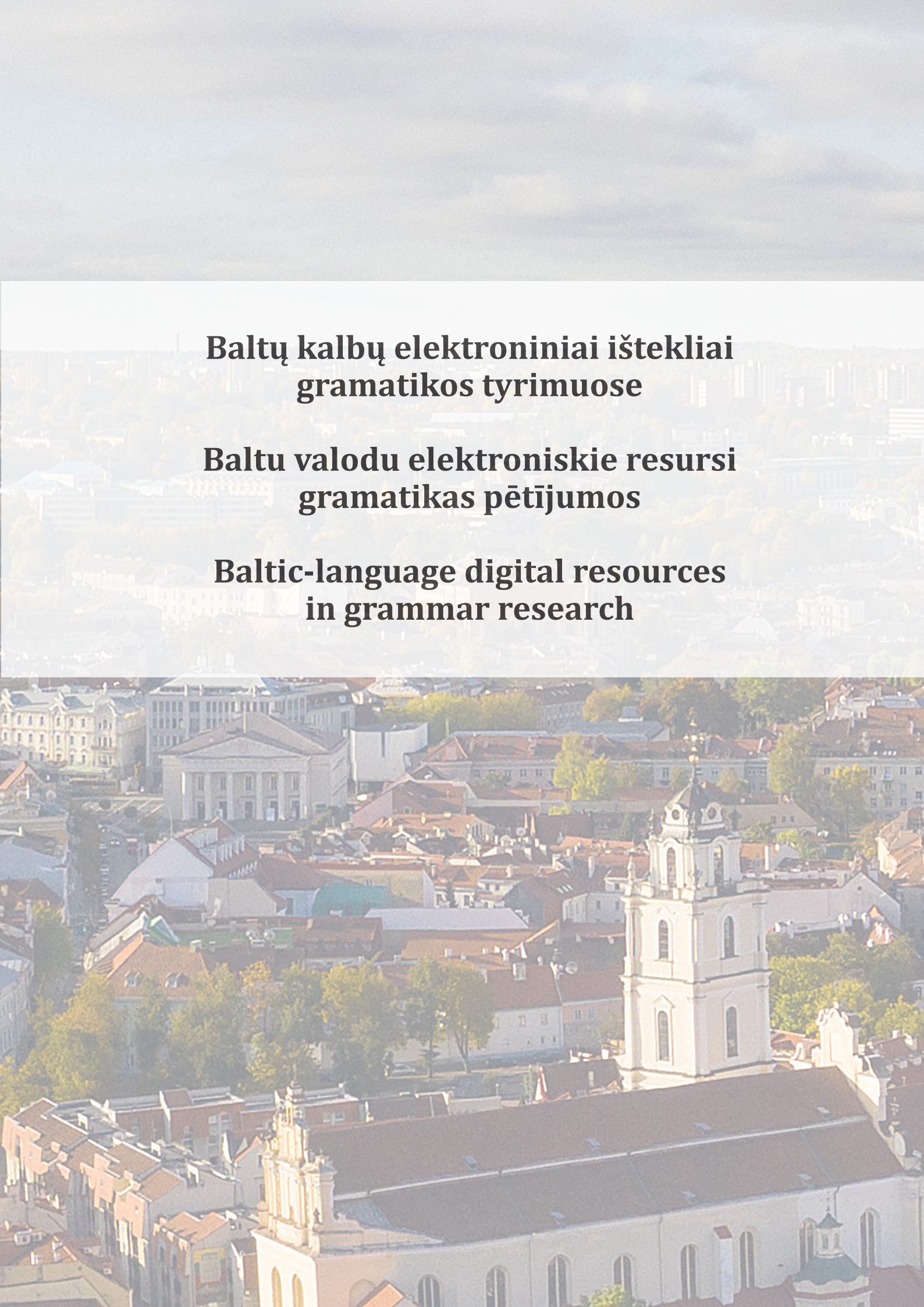


An aerial photograph of a city, likely Vilnius, Lithuania, showing a dense urban landscape with numerous buildings and a prominent church spire in the foreground. The image is overlaid with a semi-transparent white rectangle containing text.

**Baltų kalbų elektroniniai ištekliai
gramatikos tyrimuose**

**Baltu valodu elektroniskie resursi
gramatikas pētījumos**

**Baltic-language digital resources
in grammar research**

VANESA BALMANE

Latvijas Universitāte
balmane3@inbox.lv

Determinatīvie salikteni *lietvārds_{GEN}* + *lietvārds* “Latviešu valodas morfēmu un vārddarināšanas modeļu datubāzes” materiālā

Saliktenis ir vārds, viena leksiska vienība, kuru darina, apvienojot vismaz divus vārdus, resp., vārdformas vai to celmus (Matthews 2014; Kalnača, Lokmane 2021, 68). Pēc sintaktiskās struktūras izdala kopulatīvus un determinatīvus salikteņus (ten Hacken 2021, 15; Kalnača, Lokmane 2021, 69) – kopulatīvus salikteņus darina no vārdrindām, bet determinatīvu salikteņu pamatā ir vārdkopas (Skujiņa 2007, 86, 195). Atkarībā no tā, vai kāds no salikteņa komponentiem attieksmē pret visu salikteni ir hiperonīms, izdala endocentriskus un eksocentriskus salikteņus (Kastovsky 2011, 524; Kalnača, Lokmane 2016, 177).

Referāts balstās uz autores maģistra darbu “Determinatīvie lietvārdu salikteni “Latviešu valodas morfēmu un vārddarināšanas modeļu datubāzes” materiālā”, kas izstrādāts Latvijas Zinātnes padomes fundamentālo un lietišķo pētījumu projektā “Latviešu valodas morfēmu un vārddarināšanas modeļu datubāze (LVMVMD)” (Nr.: lzp-2022/1-0013). Pētījuma materiāls aptver lemmas, kas izgūtas no “Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa LVK2018” un kuru sākotnējais automatizētais segmentējums projekta gaitā ir pārskatīts un labots. Modelis *lietvārds* + *lietvārds* pētāmajā materiālā ir visplašāk sastopamais tieši salikteņu *lietvārds_{GEN}* + *lietvārds* dēļ, jo ar ģenitīvu latviešu valodā izsaka dažādas attieksmes, piemēram, piederības (*mežsargmāja*), īpašības (*brīnumzirgs*), noderības jeb satura (*rotaļlaukums*) (Nītiņa, Grigorjevs 2013, 345–347). Referāta mērķis ir raksturot salikteņus *lietvārds_{GEN}* + *lietvārds* formālā un semantiskā aspektā, uzrādot skaitliskus datus un sagrupējot ekscerpētos piemērus pēc ģenitīva izteiktajām attieksmēm, nosauktajām reālijām u.c. principiem.

Literatūra

- ten Hacken, Pius. 2021. The Nature of Compounding. *Cadernos de Linguística e Teoria da Literatura* 2(1), 1–21.
- Kalnača, Andra, Lokmane, Ilze. 2021. *Latvian Grammar*. Rīga: University of Latvia. <https://doi.org/10.22364/latgram.2021>
- Kastovsky, Dieter. 2011. Diachronic perspectives. Lieber, Rochelle, Štekauer, Pavol (eds.). *The Oxford Handbook of Compounding*. Oxford: Oxford University Press, 511–538. <https://doi.org/10.1093/oxfordhb/9780199695720.001.0001>
- Matthews, Peter (ed.). 2014. *The Concise Dictionary of Oxford Linguistics*. 3rd ed. Oxford: Oxford University Press.
- Pieejams: <https://www-oxfordreference-com.datubazes.lanet.lv/view/10.1093/acref/9780199675128.001.0001/acref-9780199675128>
- Nītiņa, Daina, Grigorjevs, Juris (red.). 2013. *Latviešu valodas gramatika*. Rīga: Latvijas Universitātes Latviešu valodas institūts.
- Skujiņa, Valentīna (red.). 2007. *Valodniecības pamattermiņu skaidrojošā vārdnīca*. Rīga: Latvijas Universitātes Latviešu valodas institūts.

INETA BALODE

Latvijas Universitāte
ineta.balode@lu.lv

DZINTRA LELE-ROZENTĀLE

Latvijas Universitāte
RTU Liepājas akadēmija
dzintra.lele-rozentale@lu.lv

Leksēmas gramatizācija un gramatiskā raksturojuma atspoguļojums leksikogrāfijā

Jebkuras valodas kodifikācijā būtiska nozīme ir vārdnīcām, kuras visbiežāk (preskriptīvi) fiksē un atspoguļo leksēmas lietojumu. Leksikogrāfiskajā praksē tas iekļauj gramatizācijas izvērtējumu, pirmām kārtām, saistībā ar jaunām, visbiežāk aizgūtām, vārdu krājuma vienībām, kā arī jau esošo leksēmu atspoguļojumu atbilstoši gramatiskajai paradigmai un raksturojumam, kas atklājas semantiskā un gramatiskā sasaistē ar citām valodas vienībām.

Šādas norādes par vārdu un tā lietojumu ir leksikogrāfiski svarīgs aspekts, jo vārdnīcas – pirmām kārtām, skaidrojošās vārdnīcas – ir visuzticamākais uzziņas avots, kuru izmanto ikviens, kas profesionālajā darbībā, (sveš) valodas apguvē vai ikdienas saziņā vēlas pārliecināties par pareizu valodas lietojumu. Meklēto atbilžu klāsts var būt visplašākais – gan saistībā ar leksēmu locīšanas paradigmu (piemēram, lietvārda iekļāvumu noteiktā deklinācijā vai verba formu darināšanu), gan ar pareizrakstību, bet jo īpaši ar vārda lietojumu dažādās konstrukcijās, vispirmāk, kolokācijās, kurās gramatikas aspekti ir cieši saistīti ar vārda un frāzes semantiku.

Tādēļ referātā galvenā uzmanība ir pievērsta jautājumam, kā valodas vienību gramatiskais raksturojums ir atspoguļots skaidrojošajā tīmekļa vārdnīcā “Tēzaurs”, kas, šķiet, ir kļuvusi par visbiežāk izmantoto leksikogrāfisko uzziņas avotu, runājot par mūsdienu latviešu valodu. Balstoties piemēros, tiek analizēts vārdnīcas “Tēzaurs” sniegums salīdzinājumā ar vairākām citu valodu, pirmām kārtām, vācu valodas vārdnīcām, parādot, kā “Tēzaurs” ar saviem risinājumiem, sniedzot šķirkļa vārda gramatisko raksturojumu, iekļaujas starptautiskajā leksikogrāfijas kontekstā.

AGNĖ BIELINSKIENĖ

Vytauto Didžiojo universitetas
agne.bielinskiene@vdu.lt

Lietuvių kalbos automatinės sintaksinės analizės plėtojimas ir problematika

Vystantis lietuvių kalbos technologijoms matomas ypač didelis poreikis plėtoti lietuvių kalbos automatinę sintaksinę analizę. 2024–2026 m. Vytauto Didžiojo universitete Skaitmeninių išteklių ir tarpdisciplininių tyrimų institute (<https://sitti.vdu.lt/>) vykdamas Europos Sąjungos NextGenerationEU projektą „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ (Nr. 02-098-K-0001; <https://sitti.vdu.lt/morfologiskai-ir-sintaksiskai-anotuotu-tekstynu-modeliai-dirbtinio-intelektu-apmokymui/>), užsibrėžtas tikslas – parengti didelės apimties (10 mln. žodžių) automatiškai sintaksiškai anotuotą tekstyną. Panašaus pobūdžio tekstyno anotavimo pradžia padaryta 2015 metais kuriant nedidelį sintaksiškai anotuotą tekstyną ALKSNIS (60 tūkst. žodžių) projekte „Lithuania’s Membership in the International Research Infrastructure – CLARIN ERIC“ (<https://clarin.vdu.lt/xmlui/handle/20.500.11821/10>) ir 2019 metais projekte „Information System of Syntactic-Semantic Analysis of Lithuanian Language: Development of Public Services (SEMANTIKA-2)“ (<https://clarin.vdu.lt/xmlui/handle/20.500.11821/21>). Paieška ALKSNIJE prieinama internete (<https://kalbu.vdu.lt/mokymosi-priemones/morfologiskai-anotuotas-tekstynas/>).

Šiuo metu sintaksiškai anotuotas tekstynas rengiamas remiantis *Universal Dependencies* (UD) standartu (<https://universaldependencies.org/>). Tai tarptautinė gramatinio anotavimo sistema, apimanti kalbos dalių, morfologinių ypatybių ir sintaksinių priklausomybių anotavimą ir pritaikyta skirtingoms natūraliosioms kalboms. Tekstyną sudarys 10 mln. žodžių, tekstai apims įvairius žanrus, bus daugiausia visateksčiai. Kadangi projektas susijęs su dirbtinio intelekto plėtra, tai šioje srityje, be visų kitų technologinių sprendimų, didelė svarba teikiama duomenims. Parengti duomenys bus kaip aukso standartas įvairių įrankių (tiek pagrįstų įprastomis technologijomis, tiek panaudojant dirbtinį intelektą) apmokymui.

Pranešime, be pagrindinės informacijos, pristatomi automatinės sintaksinės analizės etapai, kurie apima morfologinį tekstų anotavimą, įtraukiant morfologinio anotavimo tobulinimą, taip pat sukurto automatinio analizatoriaus tobulinimą ir naujų metodų, skirtų anotavimo apsimokymui, taikymą, sintaksinių medžių generavimą, automatinį sintaksinių funkcijų nurodymą ir rankinį sintaksinių medžių tikrinimą, kurį atlieka lingvistų anotuotojų grupė. Priklausomybių medžiai (angl. *dependency trees*) generuojami priklausomybių gramatikos principu pagal minėtą UD standartą. Laikoma, kad sakiny yra hierarchinė struktūra, kurios pagrindinis dėmuo – priklausomybių medžio šaknis (angl. *root*), o kiti dėmenys prijungti remiantis priklausomybe, yra valdomi aukštesnio dėmens.

Projekte numatyta ne tik parengti tekstyną, bet ir pasiekti tam tikrą morfologinio ir sintaksinio anotavimo tikslumą, todėl rankinio anotavimo tikrinimo etapas labai svarbus, nes remiantis teisingai suanotuotais teksta is vyksta tolesnis automatinis tekstų anotavimas, kuris vis gerinamas atsižvelgiant į anotuotojų patikrintą ir ištaisytą anotavimą.

Projekte yra nemažai iššūkių. Vienas iš iššūkių – sukaupti 10 mln. žodžių tekstyną. Tačiau pranešime labiau koncentruojamasi į sintaksinio anotavimo problemas. Pirmiausia daug iššūkių kyla pritaikant UD standartą lietuvių kalbai, nes Lietuvoje vyraujanti tradicinė gramatikos samprata ne visada atitinka UD sampratą. Tenka pritaikyti universalius dalykus ir, pavyzdžiui, tokias pažymas kaip *oblique* (taikoma daiktavardinei frazei kaip nepagrindiniam argumentui arba vardažodinei aplinkybei), *determinant* (būdvardiškesiems įvardžiams). Kadangi tekstai sudėtingi, ypač administraciniai, moksliniai, sudėtinga ir pati sakinių sintaksė, tad kyla nemažai sunkumų dėl sintaksės interpretavimo, konkrečių pažymų parinkimo ir kt.

DIANA BURBIENĖ

Kauno technologijos universitetas
diana.burbienne@ktu.lt

Naujų lietuvių kalbos žodžių eksperimentas žmogiškajam ir dirbtiniam intelektui

Pranešime pristatomas lietuvių kalbos naujažodžių supratimo eksperimentas, kuris yra *Europinio leksinių inovacijų tinklo* (ENEOLI) vykdomų tarpkalbinių tyrimų, analizuojančių naujų žodžių įtraukimą į mentalinį žodyną ir integravimą į kalbinę sistemą, dalis. Pagrindinis klausimas, keliamas pristatomame tyrime – kaip kalbos vartotojai supranta ir interpretuoja naujus ir dėl to nežinomos gimtosios kalbos daiktavardžius.

Eksperimento testas sudarytas ir atliktas siekiant kiekybiškai palyginti atsakymų sutaptį keliais pjūviais: kontaminacinių darinių, sudurtinių žodžių ir vedinių grupėse; skirtingų kalbų grupėse; skirtingų didžiųjų kalbos modelių grupėse; žmonių ir didžiųjų kalbos modelių grupėse.

Tyrime panaudotas sąrašas daiktavardžių, sukurtų tikslingai eksperimento sąlygoms, panaudojant lietuvių kalbos žodžius ir pritaikant darybos principus. Eksperimento rezultatai surinkti dviem būdais. Tyrimo dalyviai *SoSci* platformos apklausoje buvo paprašyti kiekvienam daiktavardžiui parinkti vieną iš siūlomų apibrėžimų. Analogiška užklausa su tokiu pat daiktavardžių ir galimų apibrėžimų sąrašu pateikta generatyvinio dirbtinio intelekto įrankiams.

Tyrimo rezultatai rodo, kad atsakymų sutaptis lietuvių kalbos daiktavardžių darybos būdų grupėse yra panaši tiek žmonių grupėje, tiek GDI įrankių grupėje, tačiau didžiausi sutapties rodikliai yra fiksuojami skirtingiems žodžiams. Analizuojant konkrečius pavyzdžius, matoma, kad lietuvių kalbos vartotojų bendras sutarimas dėl apibrėžimų ryškėja tuomet, kai atitinkamas konceptas kultūriškai aktualesnis ir ryškesnis palyginus su kitais. Tikėtina, kad GDI įrankių atsakymų sutapimus ir skirtumus gali lemti didžiųjų kalbos modelių apmokymui panaudoti lietuvių kalbos tekstų rinkinių sutapimai ir skirtumai.

ANITA BUTĀNE

Latvijas Universitāte
anita.butane@lu.lv

Terminelementi morfēmu un terminu datubāzēs

Terminzinātnē izvirzītās sistēmiskuma prasības izpildi parasti nodrošina ar attiecīgu terminelementu – morfēmu, kā arī salikteņu un vārdkopu komponentu – izmantošanu virsjēdzienu, apakšjēdzienu un blakusjēdzienu apzīmēšanā. Morfēmu datubāzes dod iespēju skaidrāk pamanīt morfēmas, kas vienlīdz raksturīgas dažādiem leksikas slāņiem, un arī tādas, kas biežāk sastopamas tieši terminoloģijā. Par konkrētu morfēmu, kā arī salikteņu un vārdkopu komponentu produktivitāti terminradē savukārt liecina terminu datubāzes, kurās paredzēta iespēja meklēt gan vārdformas, gan arī vārddāļas un kuras vienlaikus rāda arī attiecīgo elementu ekvivalentus citās valodās.

Datubāzēs izgūti saraksti ar interesējošo elementu īpaši noderīgi var būt problēmgadījumu izpētē. Kā viens no tādiem minams, piemēram, Diasporas likumā ietvertais termins *remigrācija*, kas nostājies līdzās agrāk lietotajam terminam *reemigrācija*. Variantu *remigrācija* un *reemigrācija* analizē jāskaatā prefikss *re-* un motivētārvārdi *migrācija* un *emigrācija* ar kopīgo saknes morfēmu *-migr-*. Prefikss *re-* nenodrošina terminoloģijā nepieciešamo viennozīmīgumu, jo tam var būt vairākas nozīmes, tostarp gan nozīme ‘atpakaļ’ (līdzīgi kā terminā *repatriācija*), gan arī nozīme ‘atkal, no jauna’ (līdzīgi kā terminā *reimunizācija*). Attiecīgi *remigrācija* varētu būt gan atkārtota migrācija (pārvietošanās), gan migrācija atpakaļ, savukārt *reemigrāciju* varētu skaidrot arī kā atkārtotu emigrāciju (izceļošanu), lai gan aktuālais apzīmējama jēdziens ir ‘izceļotāju atgriešanās’.

Referātā tiks aplūkoti vairāki terminelementi, kurus morfēmu un terminu datubāzes uzrāda kā īpaši produktīvus vai kuri zināmu apsvērumu dēļ var būt problemātiski.

ANNA FRĪDENBERGA

LU HZF Latviešu valodas institūts
 anna.fridenberga@lu.lv

Salikteņu un vārdu savienojumu šķīrums – viens no problemātiskiem jautājumiem, veidojot “Latviešu valodas vēsturisko vārdnīcu”

“Latviešu valodas vēsturiskā vārdnīca” (LVVV) ir elektroniska vārdnīca, kas katru gadu tiek papildināta, tās autoru mērķis ir vārdnīcas šķirklis apvienot visus latviešu valodas Seno tekstu korpusa 16. un 17. gs. avotos esošos vārdus, to formas un rakstības variantus. Rakstot LVVV, rodas dažādi problēmjautājumi, kas darba gaitā tiek apspriesti vārdnīcas darba grupā, meklējot risinājumus. Viens no šādiem problēmjautājumiem ir salikteņu un vārdu savienojumu šķīrums.

Senajos rakstos bieži vērojamas svārstības starp kopraksti un savrupraksti, ja abi salikteņa resp. vārdu savienojuma komponenti ir nominālas dabas. Piemēram, tādi darinājumi kā *karoganesējs*, *asinanauda* vai *dusmupūķis* senajos tekstos var būt lietoti gan koprakstē, gan savruprakstē un traktējami vai nu kā salikteņi, vai kā vārdu savienojumi. Turklāt dažu autoru, piemēram, G. Manceļa, K. Fīrekera tekstos, šādos veidojumos nereti tiek izmantota biedruzīme. Piemēram, *raksta=mācētāji*, *sirds=sāpes*. Dažkārt nav iespējams novilkt stingru robežu starp īstu (stabilu) salikteni un vārdkopu. Līdz ar to ir iespējami varianti – šādu darinājumu var rakstīt atsevišķā šķirklī (šķirkļa vārds ir saliktenis) vai arī kā vārdu savienojumu šķirklī pie viena no tā komponentiem. Daļai šādu vārdkopu arī vācu valodā atbilst saliktenis (Bergmane). Šādi darinājumi mūsdienu latviešu valodniecībā klasificēti dažādi (piem., sk. Skujiņa, Elksnīte), tomēr būtiski izpētīt vārdu koprakstes vai savruprakstes lietojuma iemeslus no 17. gs. autoru viedokļa, kā arī to lietojuma mainību šo autoru tekstos. Koprakstījumā un savruprakstījumā līdz pat 20. gs. vērojama liela dažādība. Jāpiezīmē, ka salikteņu darināšana kā sintaktisks vārddarināšanas paņēmiens latviešu valodā ir raksturīga – gan runātajā valodā, gan arī salikteņi sastopami vecākajos zināmajos latviešu tekstos (Draviņš), tie raksturīgi arī lietuviešu u.c. ide. valodās (Larsson, sk. arī Endzelīns, Skardžius).

LVVV autori ir vienojušies par vairākiem kritērijiem, lai nolemtu par salikteņa/vārdu savienojuma iekļaušanu atsevišķā šķirklī vai pie kāda no komponentiem. Pamatā tie balstās uz salikteņu noteikšanas pazīmēm (Ahero). Pirmais kritērijs ir rakstība – ja abi komponenti rakstīti kopā, vārds lielākoties tiek klasificēts kā saliktenis un rakstīts atsevišķā šķirklī (*padomadevējs*, *algapuisis*). Salikteņa stabilitātes noteikšanā nozīmīga ir arī semantiskā pazīme – par labu saliktenim liecina tas, ja vārds ir precīzs zināma priekšmeta, jēdziena nosaukums (vai termins). Piemēram, atsevišķā šķirklī rakstīti vārdi *bada=dzeguze* ‘pupuķis’, *kumeļa=pēdi* ‘kumeljpēda’, *mieles=tapa*, *mieļa=tapa* ‘dzērājs’. Šādiem vārdiem arī mūsdienās nereti atbilst saliktenis.

Savukārt, ja tiek pieņemts lēmums par labu vārdu savienojumam un to raksta šķirklī pie viena no komponentiem (vai abiem, ja tie ir vārdnīcā), tiek izveidots tukšais šķirklis ar norādi uz šo šķirklī. Piemēram, vārdam *karoganesējs* LVVV ir norāde uz šķirklī *nesējs*, savukārt šķirklī *nesējs* ir stablais vārdu savienojums Δ *karoga nesējs* ‘karognesējs’ (*karoganesējs*, *karoga nesējs*). Viens no lielākajiem šķirklī vārdnīcā ir *asens*, *asin(i)s*, šiem vārdiem ir ļoti daudz salikteņu/vārdu savienojumu, kas rakstīti gan kopā, gan atsevišķi, gan ar biedruzīmi. Līdz ar to arī salikteņi ar *asens*, *asin(i)s* lielākoties tiek iekļauti šajā šķirklī. Piemēram, vārdam *asinanauda* ir norāde uz abu komponentu šķirklī $\rightarrow$ *asens*, *asin(i)s*; *nauda*. Savukārt gan šķirklī *asens*, *asin(i)s*, gan šķirklī *nauda* ir stablais vārdu savienojums Δ *asenu/asina/asinu nauda* ‘par jaunu darbu saņemta nauda’.

Par salikteņiem tiek uzskatīti un atsevišķā šķirklī rakstīti t.s. celmu salikteņi – arī tad, ja abus komponentus saista biedruzīme, piem., *asa=malka*, *āda=minis*, *māja=vieta*. Parasti patskanis, kas savieno abus darinājuma komponentus, tiek traktēts kā savienotājelements (Bukelskytė-Čepelė). Jāņem vērā arī pētījumi dialektoloģijā par

sēliskajās izloksnēs raksturīgu iezīmi, ka salikteņa pirmais komponents var būt nominatīvā, piem., *pļavazāle, liepaziedi* (Poiša). Iespējams, šie piemēri būtu salīdzināmi ar senajos tekstos lietotajiem salikteņiem, kas īpaši raksturīgi G. Manceļa tekstos (*kāja=pēdi, cauna=cepure*).

Tā kā darbs pie LVVV notiek ilgākā laika periodā, arī pieredze, kā rīkoties, rodas līdz ar ekscerpētajiem piemēriem. Rezultātā – ir pieļaujami varianti, kā tiek klasificēts katrs konkrētais darinājums.

Literatūra

- Ahero A. Salikteņu pazīmes un rakstība. *Latviešu valodas kultūras jautājumi*, 15. laidziens. Rīga: Liesma, 1979, 164–179.
- Bergmane A. Vārdu koprakste un savruprakste. Grām.: A. Bergmane, A. Blinkena. *Latviešu rakstības attīstība*. Rīga: Zinātne, 1986, 98–103.
- Bukelskytė-Čepelė K. Nominal Compounds in Old Latvian Texts in the 16th and 17th Centuries. Dissertation. Stockholm, 2017.
- Draviņš K. Dažas piezīmes par salikteņiem latviešu valodā. *Valodas jautājumi. Rakstu krājums*. Rīga: Avots, 1992, 73–91.
- Larsson J. H. Nominal Compounds in the Baltic Languages. *Transactions of the Philological Society*. 2002, 203–231.
- Poiša M. Vidzemes sēliskās izloksnes, I sēj. – Rīga: Zinātne, 1985; II sēj. – Rīga: Latviešu valodas institūts, 1999.

DAIKI HORIGUCHI

Kyoto University
sirdspuksti@gmail.com

Saliktie divdabji ar pirmajiem internacionālajiem elementiem: korpusu datu analīze

Referātā tiks analizēti aptuveni 350 saliktie divdabji, piemēram, kur pirmie elementi ir internacionālas vārddabjas (piemēram, bionoārdošs, kriminālsodāms, biznesorientēts). Analīze balstīta uz diviem elektroniskajiem resursiem: Līdzsvarotais Latviešu valodas korpus 2022 (tālāk LVK2022) un plašsaziņas līdzekļu tekstu datubāze “Laikrakstu Bibliotēka” (tālāk LB), kas nav izstrādāta valodniecības pētījumiem.

Salīdzinot ar LVK2022, LB ir tehniski mazāk aprīkots un nav izstrādāts kā valodniecības korpus. Uz 2025. g. maija beigām LB satur aptuveni 7.5 miljonus tekstvienību, un tā dati tiek papildināti katru dienu. LB galvenais žanrs ir periodika, bet tā aptver arī dažādus tekstu žanrus, piemēram, tiesību aktus, zinātniskos un tehniskos tekstus. Turklāt paši periodikas teksti iekļauj dažādus funkcionālus paveidus, tostarp daiļliteratūras un zinātniskās valodas. LVK2022 sastāvā periodikas teksti veido 61% no kopējā materiāla, kas zināmā mērā pietuvina to ar LB.

Lai gan pieņemts uzskatīt, ka divdabji tiek veidoti darbības vārda paradigmas ietvaros, saliktie divdabji ir jāaplūko atsevišķi no darbības vārda sistēmas (Vulāne 2013, 271–272). Daļa šo divdabju konversijas ceļā adjektivizējas, kļūstot par patstāvīgām leksiskām vienībām. Tomēr lielākā daļa analizēto salikto divdabju nav reģistrēta leksikogrāfiskajos resursos, kas var liecināt par to ierobežotu izplatību vai okazonālu raksturu.

Aptuveni 55% konstatēto salikto divdabju ar internacionālajiem elementiem ir ciešāmās kārtas pagātnes divdabji (Horiguchi 2025). Šajos darinājumos pirmie elementi norāda uz aģentu vai līdzekli (datorģenerēts), jomu (kultūrorientēts) vai veidu (kriminālsodīts). Daļa darinājumu liecina par iespējamu angļu valodas ietekmi. Angļu valodā tendence veidot saliktus pagātnes divdabjus ir pastiprinājusies kopš 1990. gadiem, sevišķi zinātniskajos tekstos (Maniez 2017). Pārējie salikto divdabju veidi pēc to proporcijas ir šādi: darāmās kārtas tagadnes divdabji (26%), ciešāmās kārtas tagadnes divdabji (13%), vismazāk darāmās kārtas pagātnes divdabji (5%).

Divu resursu datu analīze rāda, ka abi piedāvā salīdzināmu daudzumu valodas fakti, tomēr LB ar nelielu pārsvaru nodrošina vairāk darinājuma piemēru. Piem., LVK2022 vienīgā lemma ar *eiointegr** ir *eiointegrācija*, bet LB atrodama virkne darinājumu, kas ietilpst vienā vārddarināšanas ligzdā: *eiointegrēts*, *eiointegrējusies* un *eiointegrētājs*. Savukārt saliktie divdabji, kas sastopami tikai LVK2022, bieži vien pieder zinātniska stila tekstiem. Referātā tiks aplūkoti arī salikto divdabju lietojuma tendence dažādos tekstu žanros.

Literatūra

- Horiguchi, Daiki. 2025. Deverbālie salikteņi ar internacionālajiem elementiem. Starptautiskā zinātniskā konference “Valodas sistēma, morfēmika un derivatīvā morfoloģija”. Latvijas Universitātē (referāts).
Maniez, François. 2017. An appraisal of recent breakthroughs in machine translation: the case of past participle-based compound adjectives in ESP, ASp, 72, 29–48.
Vulāne, Anna. 2013. Vārddarināšana. Latviešu valodas gramatika. Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: Latvijas Universitātes Latviešu valodas institūts, 190–299.

DALIA JAKULYTĖ

Klaipėdos universitetas
dalia.jakulyte@ku.lt

ASTA BALČIŪNIENĖ

Klaipėdos universitetas
asta.balciuniene@ku.lt

Senujų raštų leksikos ir morfologijos sistemos prototipo – „Knygos nobažnystės“ duomenų bazės – plėtra morfemikos ir žodžių darybos tyrimams

Pranešime pristatoma Senujų raštų leksikos ir morfologijos informacinė sistema (SERMIS), paremta 1653 m. leidinio *Knyga nobažnystės krikščioniškos* reliacine duomenų baze (KNDB). Ši bazė leidžia analizuoti senųjų tekstų morfologinę struktūrą, atlikti semantines paieškas ir leksikos statistinę analizę. Pranešime aptariamos galimybės išplėsti KNDB funkcionalumą, įdiegiant papildomus morfemikos ir žodžių darybos atributus. Tai leistų tyrėjams ne tik žymėti žodžių darybos ypatybes, bet ir peržiūrėti bei tikslinti kitų įrašus, taip kuriant dinamišką, bendradarbiavimu grįstą platformą. Išplėstas modelis galėtų būti integruotas ir į SERMIS sistemą, dar labiau praturtinant senųjų lietuvių kalbos tekstų tyrimų galimybes.

ERIKA JASIONYTĖ-MIKUČIONIENĖ

Vilniaus universitetas
 erika.jasionyte@flf.vu.lt

Lietuvių kalbos klausiamųjų dalelyčių raidos aspektai

Lingvistikos darbuose klausiamųjų dalelyčių klasė nagrinėjama ir iš diachroninės, ir iš sinchroninės perspektyvos. Analizuojamas dalelyčių atsiradimas bei raida, aptariamas klausiamųjų dalelyčių santykis su kitų semantinių grupių dalelytėmis (pvz., modalinėmis, pabrėžiamosiomis ir kt.), taip pat su informacine sakinio ir diskurso, kuriame dalelytės vartojamos, struktūra (Siemund 2001; Bencini 2003; Aijmer 2007; Beijering 2012; Metslang et al. 2017). Lietuvių kalbotyros darbuose analizuota klausiamųjų dalelyčių sandara bei vieta sakinyje (Labutis 1964; Ambrazas 2006), etimologija (Ambrazas 2006; Nau, Ostrowski 2010; Ostrowski 2012), funkcijos ir vartosenos polinkiai rašytiniame diskurse (LKG II 1971, 558–560). Vis dėlto klausiamųjų dalelyčių funkcijų kaita skirtingais kalbos raidos etapais dar nebuvo tirta, o tyrimų, paremtų išsamia autentiškų empirinių duomenų analize, vis dar trūksta.

Pranešimo tikslas – aptarti lietuvių kalbos klausiamųjų dalelyčių funkcinę ir struktūrinę raidą remiantis senųjų XVI–XVII a. raštų, XIX a. ir dabartinės lietuvių kalbos tekstų analize. Senųjų raštų medžiaga rinkta iš *Senųjų lietuvių kalbos raštų duomenų bazės*, sudarytos Lietuvių kalbos instituto (<http://seniejirastai.lki.lt/accept.php?reason=expire>), XIX a. tekstai – iš naujai sudaryto XIX a. lietuvių grožinės literatūros tekstyno (DOI 10.18279/MIDAS.LT-LIT-19.243978), o dabartinės kalbos tekstai – iš *Dabartinės lietuvių kalbos tekstyno* grožinės literatūros patekstynio.

Atlikta analizė rodo, kad sakinio pradžioje (inicialėje) vartojamos klausiamosios dalelytės koduoja ryšį su ankstesniu sakiniu arba tekstu (resp. diskursu), ilgainiui ima žymėti subjektyvią autoriaus poziciją (pvz., abejonę, tam tikrą netikrumo laipsnį) ir, galiausiai, intersubjektyvų kreipimąsi į adresatą (resp. klausytoją). Taigi, galima kalbėti apie klausiamųjų žymiklių semantinę funkcinę plėtrą, siejamą su intersubjektyvizacijos procesais kalboje. Be to, išsiaiškinta, kad šios semantinės klasės dalelyčių kilmę ir funkcijų raidą galima sieti su universaliomis klausiamųjų dalelyčių raidos tendencijomis ir išskirti tokius raidos modelius: (a) klausiamoji dalelytė < koordinuojamieji žymikliai; (b) klausiamoji dalelytė < subordinuojamieji žymikliai; (c) klausiamoji dalelytė < disjunktyviniai žymikliai; (d) klausiamoji dalelytė < pabrėžiamosios dalelytės. Pranešime dalelyčių raidos modeliai aiškinami ir grindžiami skirtingų laikotarpių kalbos medžiaga.

Literatūra

- Aijmer, Karin. 2007. The meaning and functions of the Swedish discourse marker *alltså* – Evidence from translation corpora. *Catalan Journal of Linguistics* 6, 31–59.
- Ambrazas, V. 2006. *Lietuvių kalbos istorinė sintaksė*. Vilnius: Lietuvių kalbos institutas.
- Beijering, K. 2012. *Expressions of epistemic modality in Mainland Scandinavian. A study into the lexicalization-grammaticalization-pragmaticalization interface*. Groningen: University of Groningen Dissertations in Linguistics 106.
- Bencini, G. 2003. Toward a diachronic typology of yes/no question constructions with particles. *Proceedings from the annual meeting of the Chicago Linguistics Society* 39(1), 604–621.
- Labutis, V. 1964. *Lietuvių kalbos sudėtinės dalelytės. Lietuvių kalbos morfologinė sandara ir jos raida*. Vilnius: Mintis, 119–129.
- LKG II 1971. *Lietuvių kalbos gramatika (Morfologija)*, Kazys Ulvydas (vyr. red.), Vilnius: „Mintis“.
- Metslang, H., K. Habicht & K. Pajusalu. 2017. Where do polar question markers come from? *Language Typology and Universals* 70(3), 489–521.
- Ostrowski, N. 2012. Old Lithuanian interrogative particle *biau*. *Baltistica* XLVII (2), 261–265.
- Siemund, P. 2001. Interrogative constructions. In M. Haspelmath, E. König, W. Oesterreicher & W. Raible (eds.), *Language Typology and Language Universals*. Berlin: Walter de Gruyter. 1010–1028.

GIEDRĖ JUNČYTĖ

Vilniaus universitetas
giedre.juncyte@flf.vu.lt

The emotive, perception and cognition middles in Lithuanian: a corpus-based study

In Lithuanian, middle marked verbs fall into a number of groups based on their semantics and the function of the marker *-si-*: body action (naturally reflexive), reciprocal, anticausative, cognition middles, facilitative and passive-like (Geniušienė 1987, Holvoet 2020). Middle marked verbs meaning emotive state (e.g. *bijotis* ‘be afraid’ : *bijoti* ‘idem’), change of emotive state (e.g. *išsigąsti* ‘get scared’ : *išgąsdinti* ‘scare’, *įsižeisti* ‘get insulted’ : *įžeisti* ‘insult’), verbs of perception (e.g. *klausytis* ‘listen’ : *klausyti* ‘idem’, *girdėtis* ‘be heard’ : *girdėti* ‘hear’) and cognition (e.g. *prisiminti* ‘remember’ : *priminti* ‘remind’) are considered to form a semantically coherent group, typologically characterized by middle marking (cf. Kemmer 1993: 127-142). However, in Lithuanian, the function of the middle marker in these examples spans from anticausative (cf. *Ada įžeidė Romą*. ‘Ada insulted Romas.’ – *Romas įsižeidė*. ‘Romas got insulted.’) and passive-like (cf. *Aš girdžiu muziką*. ‘I hear music.’ – *Čia girdisi muzika*. ‘The music is heard here.’) to middle marked intransitives (e.g. *bijotis* : *bijoti*) and media tantum (e.g. *džiaugtis* ‘rejoice’, *juoktis* ‘laugh’).

This paper aims at representing and explaining the variety of the function of the middle marker in the verbs of emotion, perception and cognition based on the frequency of usage and syntactic behavior of the verbs in the Morphologically Annotated Corpus of the Contemporary Lithuanian Language (corpus.vdu.lt).

References

- Geniušienė, Emma. 1987. *The Typology of Reflexives*. Berlin: Mouton de Gruyter.
- Holvoet, Axel. 2020. *The Middle Voice in Baltic*. Amsterdam; Philadelphia: John Benjamins Publishing Company.
- Kemmer, Suzanne. 1993. *The Middle Voice*. Amsterdam; Philadelphia: John Benjamins.

ANDRA KALNAČA

Latvijas Universitāte
kalnaca@latnet.lv

ILZE LOKMANE

Latvijas Universitāte
ilokmane@latnet.lv

Latviešu valodas konstrukcija *kas tur ko + nenoteiksme* un miratīvs: korpusa datu analīze

Miratīvs ir semantiska kategorija, kas izsaka teksta autora pārsteigumu par informācijas saturu (sk., piem., Aikhenvald 2004, 2012; Simeonova 2015; Lau, Rooryck 2017). Latviešu valodā miratīvs netiek izteikts morfoloģiski, taču ir dažādas konstrukcijas, kurām pārsteiguma izteikšana var būt gan primārā, gan sekundārā nozīme (tipoloģiskā aspektā sk., piem., DeLancey 2001, 2012, Mocini 2014). Šīs konstrukcijas ir daļēji leksikalizējušās, jo ietver īpašas darbības vārda formas, kā arī tādus vietniekvārdus un apstākļa vārdus kā, piem.:

- 1) norādāmie vietniekvārdi *tas, tāds*;
- 2) nenoteiktais / jautājamais vietniekvārds *kas*;
- 3) deiktiski apstākļa vārdi *tur, kur* (vēsturiski saistīti ar vietniekvārdiem *tas, kas*).

Lai gan šīs konstrukcijas latviešu gramatikas pētījumos aplūkotas jau iepriekš (piem., Nītiņa, Grigorjevs 2013; Kalnača, Lokmane, Rūmniece 2019; Kalnača, Lokmane 2020, 2021), tās nav sasaistītas ar pārsteiguma izteikšanu un tā semantiku plašākā tipoloģiskā skatījumā. Tāpēc šis referāts veltīts konstrukcijas *kas tur ko + nenoteiksme* semantikas analīzei, balstoties uz vairāku latviešu valodas korpusu kvalitatīviem datiem (piem., *Līdzsvarotais mūsdienu latviešu valodas tekstu korpus LVK2018* un *LVK2022, Karogs, Tīmeklis 2020*). Viens no datu izguvei izmantotajiem vaicājumiem ir: [lemma="kas"][lemma="tur"][lemma="kas"&tag="p....a.*"] [tag="vm.n.*"], iegūtie dati ir šādi:

- 1) LVK2028 – 19
- 2) LVK2022 – 47
- 3) *Karogs* – 239
- 4) *Tīmeklis 2020* – 341

Izgūto konstrukciju semantikas analīze liecina, ka visvairāk konstrukcija *kas tur ko + nenoteiksme* tiek lietota, izsakot tieši negatīvu pārsteigumu:

- (1) *Un kas tur ko sarežģīt?* (LVK2018)

Turklāt konstrukcijās ietvertie darbības vārdi nenoteiksmē ir saistāmi ar vismaz trim semantiskām grupām:

- 1) izziņas darbības vārdi (piem., *domāt, prātot, saprast*);
- 2) runāšanas darbības vārdi (piem., *stāstīt, strīdēties, ziņot*);
- 3) emociju darbības vārdi (piem., *uztraukties, kautrēties, brīnīties*).

Minētā konstrukcija ir veidojusies no īpaša latviešu valodas salikta teikuma tipa ar nenoteiksmes palīgteikumu:

- (2) *Tur bija ko **padomāt***. (LVK2018)

Konstrukcija leksikalizējas, ja tiek lietota jautājuma teikumā kopā ar jautājamo vietniekvārdu *kas*, turklāt tiek atmests virsteikuma izteicējs (par nenoteiksmes teikumu leksikalizēšanās tendencēm plašāk sk. Holvoet 2000). Faktiski šī procesa sekas ir jautājuma teikuma pārtapšana par retorisku jautājumu, ko ievada desemantizējies deiktisks vārds, latviešu valodas gadījumā vietniekvārds *kas* nominatīvā, kam seko tāds pats vietniekvārds akuzatīvā un apstākļa vārds *tur*. Savukārt darbības vārda leksēma var būt mainīga gan apgalvojumā (1), gan noliegumā (3).

Tomēr retoriskie jautājumi pēc būtības ir izsaukuma teikumi, jo tie emocionāli uzsver kādu apgalvojumu (vai tā noliegumu):

(3) *Kādi vēl padomi, kas tur ko nezīnāt!* [Par matu griezumumu.] (www.apollo.lv)

Vienlaikus šādi retoriski jautājumi izsaka pārsteigumu un norāda, ka bijis gaidīts kaut kas cits (sk. (1) un (3) piemēru). Tas attiecas arī uz konstrukcijām ar darbības vārdu noliegumā.

Tādējādi referātā aplūkojamās konstrukcijas *kas tur ko + nenoteiksme* iespējams uzskatīt par leksiski gramatisku miratīva izteikšanas līdzekli. Miratīva nozīme veidojas no pretējām gaidām, pārsteiguma un negatīva vērtējuma pretstatā vispārīgam viedoklim par izteicējā ietvertās darbības iespējamību. Konstrukcijā ietilpstošais apstākļa vārds *tur*, lai gan ir desamentizējis, norāda uz kaut kur esošām varbūtējām vispārīgām zināšanām.

Literatūra

- Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: Oxford University Press.
- Aikhenvald, A. Y. (2012). The essence of mirativity. *Linguistic Typology*, 16, 435–485.
- DeLancey, S. (2001). The mirative and evidentiality. *Journal of Pragmatics*, 33, 369–382.
- DeLancey, S. (2012). Still mirative after all these years. *Linguistic Typology*, 16, 529–564.
- Holvoet, A. (2000). Infinitival relative clauses in Latvian: Their structure, development and tendency toward lexicalisation. *Linguistica Lettica*, 7, 99–116.
- Kalnača, A., Lokmane, I., Rūmniece, I. (2019). Nenoteiksmes predikāts latviešu valodā, tā latīņu un sengrieķu analogi un Kārļa Mīlenbaha “abzolutais infinitīvs”. *Valoda: nozīme un forma*. 10, 76–91. <https://doi.org/10.22364/vnf.10.08>
- Kalnača, A., Lokmane, I. (2020). Pragmatic aspects of Latvian Predicative Infinitive Constructions. *Studies about Languages*, 37, 74–91. <https://doi.org/10.5755/j01.sal.1.37.25881>
- Kalnača, A., Lokmane, I. (2021). *Latvian Grammar*. Riga: University of Latvia Press. <https://doi.org/10.22364/latgram.2021>
- Karogs. CLARIN-LV digitālā bibliotēka. 2022. <http://hdl.handle.net/20.500.12574/83>
- Latviešu valodas timekļa korpus (Timeklis 2020). Pieejams: <https://korpuss.lv/id/T%C4%ABmeklis2020>
- Lau, M. & Rooryck, J. (2017). Aspect, evidentiality, and mirativity. In M. Lau & W. Adelaar (Eds.), *Lingua Special Issue Essays on Evidentiality*. *Lingua*, 186–187, 110–119.
- Līdzsvarotais mūsdienu latviešu valodas tekstu korpus (LVK2018). CLARIN-LV digitālā bibliotēka. 2018. Pieejams: <http://hdl.handle.net/20.500.12574/11>
- Līdzsvarotais mūsdienu latviešu valodas tekstu korpus (LVK2022). CLARIN-LV digitālā bibliotēka. 2022. Pieejams: <http://hdl.handle.net/20.500.12574/84>
- Mocini, R. (2014). Expressing Surprise. A Cross-Linguistic Description of Mirativity. *Altre Modernità*, 11, 136–156.
- Nītiņa, D. & Grigorjevs, J. (2013). *Latviešu valodas gramatika*. Rīga: Latvijas Universitātes Latviešu valodas institūts.
- Simeonova, V. (2015). On the semantics of mirativity. *Proceedings of the 2015 annual conference of the Canadian Linguistic Association*. Ottawa: University of Ottawa. <http://cla-acl.ca/wp-content/uploads/Simeonova-2015.pdf>

KRISTĪNE LEVĀNE-PETROVA

LU Matemātikas un informātikas institūts
kristine.levanepetrova@lumii.lv

MIKUS GRASMANIS

LU Matemātikas un informātikas institūts
mikus.grasmanis@lumii.lv

BAIBA SAULĪTE

LU Matemātikas un informātikas institūts
baiba.saulite@gmail.com

“Nacionālās korpusu kolekcijas” izmantošana latviešu valodas biežuma saraksta izveidē

Vārdu biežuma saraksti ir nepieciešami gan valodnieciskos pētījumos un dažādu valodas resursu izveidē (t. sk. salīdzināmajā valodniecībā, valodas apguvē, gramatiku, vārdnīcu un citu leksisko datubāzu veidošanā), gan arī dažādās dabiskās valodas apstrādes jomās. “Latviešu valodas biežuma vārdnīca” (Jakubaite et al. 1966–1976), kas atspoguļo literārās valodas funkcionālo stilu leksiku, izstrādāta 20. gs. 70. gados un darbs pie pirmā sējuma veikts rokām (Jakubaite et al. 1966, 8–9). Mūsdienās latviešu valodas biežuma sarakstu izveidei pieejama “Nacionālā korpusu kolekcija” (NKK; Saulīte et al., 2022), un katram ir iespēja iegūt kāda konkrēta korpusa vārdu biežuma sarakstu. Lai censtos atspoguļot valodas lietojumu kopumā, parasti šādus biežuma sarakstus izstrādā uz līdzsvarota korpusa pamata.

Vispārlietojama biežuma saraksta iegūšanai, iespējams, pietiktu ar apjomīgu, kvalitatīvu, rūpīgi līdzsvarotu un precīzi nomarkētu korpusu, bet šādu datu latviešu valodai nav. Šī pētījuma mērķis ir noskaidrot, kā latviešu valodas biežuma saraksta izveidē izmantot vairākus NKK korpusus – kurus korpusus izmantot un kāda ir sarakstā iekļaujamo leksēmu atlases metodika. Izmantoto datu apjoms ir ievērojams un darbs notiek galvenokārt automatiski, ar minimālu pārskatīšanu, tāpēc ir svarīgi analizēt galvenās problēmas, kas atklājas, veidojot šādu sarakstu, t. i., automatiskā markētāja neprecizitātes (Paikens et al. 2024), pareizrakstības kļūdas korpusos, homonīmija u. tml.

NKK ietilpst 34 automatiski morfoloģiski markēti latviešu valodas korpusi, no kuriem biežuma saraksta izveides datu kopā ir iekļauti 19 (kopējais apjoms – pusotrs miljards tekstvienību). Šie korpusi izvēlēti pēc vairākiem kritērijiem, piemēram, datu daudzveidība (dažādu valodas paveidu, t. sk. daļlīteratūras un sarunvalodas teksti), datu publicēšanas laiks (priekšroka dota korpusiem, kuros iekļautie teksti publicēti, sākot no 20. gs. 80. gadiem) un datu kvalitāte (nav iekļauti korpusi, kuros ir daudz automatiskās rakstzīmju atpazīšanas kļūdu).

Rēķinot vārdu (leksēma kopā ar vārdšķiru) biežumu, vispirms katrā korpusā tika saskaitīts katra vārda lietošanas reižu skaits. Vārdi, kas sastopami mazāk nekā trīs korpusos, netika iekļauti rezultātos, bet pārējiem tika izrēķināts to biežums (vārda lietojums pret korpusa kopējo apjomu) katrā korpusā atsevišķi. Pēc tam tika iegūts vārda biežums visos korpusos, rēķinot to kā vidējo aritmētisko starp šī vārda biežumiem katrā korpusā. Tad visi vārdi tika sakārtoti dilstošā biežumu secībā un tika atstāti 25 000 biežāk lietotie.

Biežuma saraksta izveidei no korpusiem tika ņemtas visas leksēmas, kurās ir tikai mazie latviešu alfabēta burti. Sarakstam netika pievienotas atsevišķas vārdu grupas, piemēram, reziduāļi, saīsinājumi, īpašvārdi, leksēmas, kas ir tikai vienu rakstzīmi garas, vai tādas, kas pārsniedz 32 rakstzīmes. Lai mazinātu drukas kļūdu ietekmi, galīgā saraksta vārdiem rangā no 10 000 līdz 25 000 tika prasīts, lai tie būtu starp vārdnīcas “Tēzauris” leksēmām.

Automātiski iegūtais biežuma saraksts vairākkārt tika pārskatīts un precizēts – sākotnējos datos konstatētās automātiskā marķētāja kļūdas uz nākamajām versijām lielākoties novērstas (korpusi pārmarķēti ar precizētu automātisko marķētāju). Neprecizitātes, kas saglabājušās saraksta galaversijā, īpaši analizētas, piem., automātiski nepareizi noteikta pamatforma (*solīns, pulciņš* → *solīnš, pulciņš*), pareizrakstības kļūdas (*cilveks*), dažādas citas kļūdas, kas ietekmē biežumu, piem., īpašvārdi, kas var būt arī sugasvārdi (*bels, kalniņš, dons*), saīsinājumi, kas automātiski marķēti kā izsaukmes vārdi (*ha, pag*).

Iegūtais latviešu valodas biežuma saraksts jau šobrīd tiek sastatīts ar lībiešu un latgaliešu valodas vārdiem, tāpat tas tiek ņemts vērā, papildinot “Tēzauru” ar informāciju par leksēmu izrunu. Biežuma rēķināšanas metodiku plānots izmantot, izstrādājot rīku, kas pēc individuāli izvēlētiem parametriem lietotājam pašam ļautu veidot dažādus biežuma sarakstus no NKK datiem.

Literatūra

Jakubaite, Tamāra, Kristovska, Dainuvīte, Ozola, Vera, Prūse, Rudīte, Sika, Nira. *Latviešu valodas biežuma vārdnīca*. Rīga: Zinātne, 1966–1976.

Saulīte, Baiba, Dargis, Roberts, Grūzītis, Normunds, Auziņa, Ilze, Levāne-Petrova, Kristīne, Pretkalniņa, Lauma, Rituma, Laura, Paikens, Pēteris, Znotiņš, Artūrs, Strankale, Laine, Pokratniece, Kristīne, Poikāns, Ilmārs, Bārzdiņš, Guntis,

Skadiņa, Inguna, Baklāne, Anda, Saulespurēns, Valdis, Ziediņš, Jānis. 2022. Latvian National Corpora Collection – Korpus.lv. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. 5123–5129. Pieejams: <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.548.pdf>.

GUNTA LOČMELE

Latvijas Universitāte
gunta.locmele@lu.lv

Pirmās latviešu reklāmas kā valodu kontaktu un tulkošanas attīstības pētījumu avots

Latviešu valodas vēsturiskās attīstības pētījums balstās uz oriģinālteksti un tulkojumiem digitālajos resursos un korpusā un skata valodu kontaktus, semantiski pragmatiskās izmaiņas un tulkošanas stratēģijas 19. gs. reklāmās.

Agrīnā latviešu preses reklāma ir maz pētīts žanrs, kas atklāj valodu kontaktus un tulkošanas praksi 19. gadsimtā. Tā ir nozīmīga semantikas, pragmatikas un translatoģijas pētījumos Baltu valodu kontekstā.

Pirmo latviešu sludinājumu uzbūve bieži ir līdzīga vācu sludinājumiem, arī vārdformas un sintakse ir aizgūtas no tās – plaši ienāk kalki un aizguvumi. Otra tiešā kontaktvaloda ir krievu valoda. Ir vērojama ietekme no netiešiem kontaktiem ar citām koloniālajām valodām, piemēram, ar angļu valodu. Latvijā ienāk preces arī no Britu impērijas, un to angļiskie nosaukumi kā termini iedzīvojas latviešu valodā.

Pētījumā izmantots korpuss „Latviešu Avīzes“, kurā ietverti šī laikraksta teksti no 1822. līdz 1915. gadam. Korpusa veidošanā izmantota automātiskā rakstzīmju atpazīšana (OCR), tāpēc, kā atzīmē arī paši tā veidotāji, korpusā bieži tiek vērotas atpazīšanas kļūdas (<https://korpuss.lv/id/LAv%C4%ABzes>). Otrs pētījuma pamatresurss ir digitalizēts preses krājums *periodika.lv*. Dažādās vārdformas, aizguvumi, kalki un sintaktiskās konstrukcijas paralēli ir meklētas un pārbaudītas abos šajos digitālajos resursos. Visumā teicamajā resursā *periodika.lv* dažkārt arī vērojamas teksta atpazīšanas kļūdas, arī 18. gs. vācu prese, kurā tiek meklētas pētāmo preču un uzņēmēju reklāmas svešvalodās, krājumā ne vienmēr ir pilnībā piemērotas digitālai izmantošanai. Turklāt daudziem vārdiem ir rakstības varianti.

Ārzemēs pētīta tulkotā reklāma ar uzsvāru uz tulkošanas stratēģijām (Jettmarová 2004) un kultūru mijiedarbību (Torresi 2021), latviešu preses pirmsākumi pētīti nācijas veidošanās aspektā (Zelče 2009; 2023). Šķiet piedāvātais tulkošanas un valodu mijiedarbības aspekts šos pētījumus papildina. Reklāmas ir kultūrteksti, kas atspoguļo varas struktūras un ideoloģiju. Analizējot to tulkoto saturu, var atklāt, vai un kā koloniālās, patriarhālās un Rietumu normas tika pieņemtas vai arī pret tām tika izrādīta pretestība, un kā šajos procesos tika iesaistīta valoda.

Literatūra

- Jettmarová, Zuzana. 2004. "Linguistic Aspects of the Translation of Advertisements." In *Übersetzung – Translation – Traduction. An International Encyclopedia of Translation Studies* 1, edited by Harald Kittel, Armin Paul Frank, Norbert Greiner, Theo Hermanns, Werner Koller, José Lambert, and Fritz Paul, 655–661. Berlin and New York: Walter de Gruyter.
- Torresi, Ira. 2021. *Translating Promotional and Advertising Texts*. 2nd ed. London and New York: Routledge. <https://doi.org/10.4324/9781003131595>.
- Zelče, Vita. 2009. *Latviešu avīžniecība: Laikraksti savā laikā un sabiedrībā (1822-1865)*. Rīga: Zinātne.

PAULA MIKELSONE

Latvijas Universitāte
paula.mikelsone@lu.lv

Vokāļu kontrakcija priedēkļa un saknes sadūrā runas korpusu datos

Vokāļu kontrakcija latviešu valodā ir fonētiski fonoloģiska parādība, kad runā divi atšķirīgām zilbēm piederoši vokāļi tiek savilkti kopā vienā garā zilbē, piemēram, “vienā garā patskanī: *neesmu* > [næ:smũ] vai divskanī: *aerodroms* > [aj̥rodroms]” (Skujiņa, 2007; Auziņa u.c., 2013). Zūdot zilbes robežai, var saplūst arī morfēmu robeža, ja kontrakcija notikusi to sadūrā (Auziņa, 2010). Aplūkotas iespējamās kontrakcijas pozīcijas starp priedēkļa vokāli un saknes vai otrā priedēkļa vokāli.

Līdz šim uzskatīts, ka vokāļu kontrakcija biežāk sastopama izloksnēs (Laua, 1997), kur tā ir arī vairāk aprakstīta (Markus, 1993). Pateicoties runas korpusiem iespējams pētīt vokāļu kontrakciju arī standartvalodā. Pētījumā izmantoti dati no trīs latviešu valodas saistītās runas korpusiem: LATE fonētiski marķēts runas korpuss (fonLATE), LATE sarunu korpuss (LATE-sarunas) un LATE plašsaziņas līdzekļu korpuss (LATE-mediji).

Parādība korpusos ir vērojama biežāk nekā iepriekš aprakstīts. Tā ir fakultatīva pozicionāla skaņu pārmaiņa, kuras realizēšanās nav regulāra vai obligāta, to ietekmē ekstralingvistiski faktori: runas temps, situācija un valodas lietotājs. Izrunā vērojams variants. Palielinoties runas tempam, pieaug kontrahēšanās un citu pozicionālo pārmaiņu biežums, jo runa tiek optimizēta ātrākai saziņai, kurā vārdu nozīme ir saprotama kontekstā. Izolēti atskaņojot runas ierakstu fragmentus (vārdus), kas lietoti saistītā runā, to uztveršana var būt apgrūtināta un pat neiespējama vairāku pārmaiņu dēļ, piemēram, fonēmu zuduma, saplūšanas un kvalitātes maiņas.

Kontrahēšanās divu patskaņu sadūrā vērojama biežāk nekā gadījumos, kur saduras pozīcijā ir patskanis un divskanis vai patskanis un diftongiskais savienojums.

Visproduktīvākā vokāļu sadura kontrakcijai ir starp priedēkļa *ne-* īso patskani [e] un tam sekojošu saknes pirmo patskani [e] vai [æ]. Pēc artikulācijas pozīcijas abi šie patskaņi ir palatāli. Ja saknes patskanis ir plats [æ], tad arī priedēkļa patskaņa kvalitāte tam pielāgojas un patskanis kļūst plats, rezultējoties garā patskanī [æ:], savukārt ja patskaņu kvalitātes ir vienādas, tie tiecas saplūst vienā garā monoftongā [e:]. Šie priekšējās rindas patskaņi [e], [e:], [æ], [æ:] kontrahējas aptuveni 50% aplūkoto piemēru. “Latviešu valodas pareizrakstības un pareizrūnas vārdnīcā” norādīts, ka literārajā valodā kontrakcija pieļaujama tikai verba *nebūt* personu formās (Ceplītis, 1995), bet norma būtu jāpaplašina, balstoties uz valodas datiem.

Novērota garā patskaņa rašanās (*neesat* > [næ:sat]), tomēr valodā iespējama divskaņa rašanās atsevišķos gadījumos atzīta par iemeslu zilbju nošķiršanai, piemēram, divu blakus esošu priedēkļu gadījumā (*neizdarīja, neuztver*). Garo zilbju saplūšanai nepieciešamas vēl kādas vokālisma pārmaiņas, kas visbiežāk saistītas ar fonēmu redukciju. Vērotas saplūdušas formas gan garo patskaņu pārveidojumos (*neēdu* > [næ:du], *jāuzpasē* > [j̥a:spasē]), gan divskaņu (*saaicināti* > [saj̥:cinat̪i], *neaug* > [næ:ug]), savukārt diftongiskie savienojumi valodas runas plūsmā ar priedēkļa patskani nekontrahējas (*neilgs, painteresējies*). Nav fiksētas tādas pārmaiņas, kas šādās pozīcijās pieļautu kontrahētas formas veidošanos. Šī tendence ļauj secināt, ka diftongiskā savienojuma komponenti ir noturīgāki par divskaņa komponentiem.

Valodas lietojumā vērojama tendence no vokāļu kontrahēšanas izvairīties uztveramības dēļ, tādēļ reizēm runā tiek iesprausts puspatskanis [i̯] vai retāk troksnenis [j̥], kas iezīmē zilbju robežu. Tas vērots piemēros pirms divskaņa /ie/ (*paies* > [paj̥ies], *saiet* > [saj̥iet]), bet arī vārdos, kur morfēmu sadūrā ir tikai zema mēles pacēluma patskaņi, piemēram, *neatgriezies* > [nej̥adgriezi:sies].

Literatūra

- Aužiņa, I. (2010) Zilbju robeža vārdos, kur priedēkļa un saknes sadūrā ir divi vokāļi. *XI starptautiskā baltistu kongresa "baltu valodu pagātne, tagadne un nākotne" referātu tēzes*. Rīga: Latvijas Universitāte, 12–13.
- Aužiņa, I. u.c. (2013) *Latviešu valodas gramatika*. Rīga: LU Latviešu valodas institūts.
- Ceplītis, L. u.c. (1995) *Latviešu valodas pareizrakstības un pareizrūnas vārdnīca*. Rīga: Avots.
- Laua, A. (1997) *Latviešu literārās valodas fonētika*. Rīga: Zvaigzne ABC.
- Markus (Markusa), D. (1993) *Zilbes intonācija latviešu valodas skaņu mācībā*. Rīga: Latvijas Universitāte.
- Skujiņa, V. (red.) (2007). *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga: LU Latviešu valodas institūts.

GUNTA NEŠPORE-BĒRZKALNE

LU Matemātikas un informātikas institūts
gunta.nespore@lumii.lv

MADARA STĀDE

LU Matemātikas un informātikas institūts
stade.madara@gmail.com

The variety of semantic links in the electronic dictionary “Tēzaurs”

“Tēzaurs” (publicly available for viewing at <https://tezaurs.lv> and for downloading at <http://hdl.handle.net/20.500.12574/107>) is the largest electronic Latvian explanatory dictionary (Grasmanis et al. 2023) with more than 400,000 entries. It is regularly updated and edited. Our aim is to make the dictionary entries as informative as possible. In the last years, “Tēzaurs” has become a multifunctional resource for structured lexical data – a collection of objects (entries, lexemes, senses, examples etc.), which can be linked in various ways (Klints et al. 2024). It has served as a basis for several research initiatives increasing its scope and connection to other resources.

The first step in dictionary development is word entry revision and editing according to consistent criteria (see Lokmane, Rituma, 2021; Lokmane et al. 2025). Most of the revised entries have been processed in order to add them to the Latvian WordNet or supplement them with derivational links. In addition, revisions of entries and senses are carried out after receiving user suggestions or comments. The Latvian National Corpora Collection (Saulite et al. 2022) is used to determine whether word senses are valid and still in use nowadays, or if a new sense has appeared and should be added to the word entry.

After entry revision, word senses are linked with other senses to expand the Latvian WordNet. A wordnet is a lexico-semantic resource that links word senses based on their semantic relations (Fellbaum 1998). The basic unit of the Latvian WordNet is a synonym set (or synset). Synsets are then interlinked based on partial synonymy, antonymy, hyponymy, meronymy and other relations; sometimes synsets are also grouped in gradation sets (Klints et al. 2024). After interlinking within the Latvian WordNet, Latvian synsets are also linked with corresponding units of the Princeton English WordNet. Thus, in addition to its explanatory properties, “Tēzaurs” fulfills the functions of a synonym and translation dictionary as well. The Latvian WordNet currently constitutes 9483 synsets interlinked with 6396 links of various types; in addition, 3636 interlingual links to the Princeton WordNet have been formed so far.

In addition, dictionary entries are supplemented with derivational links: semantic links between word senses and morphological derivational links between lexemes. Semantic labels are attached to both ends of the link. The novelty of this approach is that it analyzes the relations between separate senses of polysemous derivatives and their base words, not just the polysemy of derivational affixes (Lokmane et al. 2025). So far the semantic labels are defined for lexemes of four basic word classes – nouns, verbs, adjectives, and adverbs. The first processed data consist of ~1220 morphological derivational links and 1760 semantic links, available in the latest release of “Tēzaurs”. Parallel to the semantic and morphological linking, new entries are created for the most frequent regular derivatives that were previously not included in “Tēzaurs”. This is done both to make it easier for the language user or learner to find the target word, and also because the sense system of the derivatives mostly does not fully correspond to the sense system of the base word.

The variety of semantic links in “Tēzaurs” allows the dictionary users to obtain a wide scope of information on a certain word sense in one place in the section “Saistītās nozīmes” (Related senses) under each respective sense.

The processing of such entries is a time-consuming process, and methods for its partial automation are currently being researched. However, manual verification will remain a crucial part of the annotation process to maintain the overall quality of the resource.

References

- Fellbaum, C. (ed.). 1998. WordNet: An electronic lexical database. MIT Press.
- Grasmanis, M., Paikens, P., Pretkalnina, L., Rituma, L., Strankale, L., Znotins, A., Gruzitis, N. 2023. Tēzaurs.lv – The Experience of Building a Multifunctional Lexical Resource. Electronic lexicography in the 21st century (eLex): Invisible Lexicography.
- Klints, A., Grasmanis, M., Nespore-Berzkalne, G., Pretkalnina, L., Stade, M., Gruzitis, N., Lokmane, I., Paikens, P., Rituma, L., Spektors, A. 2024. Tēzaurs as a Digital Multifunctional Lexical Resource. *Baltic Journal of Modern Computing*, 12(4), 513–525.
- Lokmane, I., Rituma, L. 2021. Verba nozīmju nošķiršana: teorija un prakse. *Valoda: nozīme un forma*, 12, Rīga: LU Akadēmiskais apgāds, 142–162.
- Lokmane, I., Grasmanis, M., Klints, A., Nespore-Berzkalne, G., Paikens, P., Pretkalnina, L., L. Rituma, L., Stade, M., Taurina, E. 2025. Towards a Derivational Semantics Resource for Latvian. *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, 359–366.
- Saulite, B., Dargis, R., Gruzitis, N., Auzina, I., Levane-Petrova, K., Pretkalnina, L., Rituma, L., Paikens, P., Znotins, A., Strankale, L. 2022. Latvian National Corpora Collection – Korpuss.lv. *Proceedings of 13th Language Resources and Evaluation Conference (LREC)*.

JURGIS PAKERYS

Vilniaus universitetas
jurgis.pakerys@flf.vu.lt

AGNĖ NAVICKAITĖ-KLIŠAUSKIENĖ

Vilniaus universitetas
agne.navickaite@flf.vu.lt

VIRGINIJUS DADURKEVIČIUS

Vytauto Didžiojo universitetas
virginijus.dadurkevicius@vdu.lt

Mišriuoju būdu sudaryti daiktavardžiai Jungtinio lietuvių kalbos tekstyno duomenimis

Šalia pagrindinių (grynųjų) darybos būdų – sufiksacijos, prefiksacijos, paradigmatizacijos ir kompozicijos – lietuvių kalbotyroje skiriami ir rečiau pasitaikantys mišrieji: prefiksacija-sufiksacija ir kompozicija-sufiksacija (Urbutis 1978: 283, 289–290; Keinys 1999: 22–23, 105). Mišriaisiais būdais pasidaryti daiktavardžiai daug dėmesio kol kas nėra susilaukę: apie juos neužsimenama pagrindinėse gramatikose, trumpai jie tepaminėti ir žodžių darybos studijose, pvz.: *į-pėd-in-is* ← *pėd-a*, *prieš-dė-l-is* ← *dė-ti* (prefiksacija-sufiksacija), *žmog-ėd-r-a* ← *žmog-us*, *ės-ti*, *ketur-vėj-inink-as* ← *Ketur-i vėj-ai* (kompozicija-sufiksacija) (Urbutis: 1978: 289–290; Urbutis 1999; Keinys 1999: 23, 105). Pastaruoju metu daugėjant įvairių elektroninių lietuvių kalbos leksikos išteklių – duomenynų ir tekstynų – pavyksta rasti vis naujų pavyzdžių, tad galima daugiau pasakyti apie mišriosios daiktavardžių darybos polinkius, pavyzdžiui, kad dažnesnė yra kompozicija-sufiksacija, pvz.: *saul-ė-deng-t-ės* ← *saul-ė*, *deng-ti laisv-a-rink-inink-as* ← *laisv-oji rink-a* (Murmulaitytė 2021: 91–92; Aleksaitė 2022: 54, 141).

Paminėtina, kad mišriosios daiktavardžių darybos atvejų yra ir lietuvių kalbai artimiausioje latvių kalboje, žr., pvz., Auziņa et al. (2013: 197, 201, 248–249, 261–262). Kaip tipiškiausi čia nurodomi modeliai *bez- + N + -niek-s* (*bez-pajumt-niek-s* ‘benamis’ ← *bez pajumt-es* ‘be pastogės’) ir *bez- + N + -īb-a* (*bez-bail-īb-a* ‘bebaimiškumas’ ← *bez bail-ēm* ‘be baimės’), šalia jų minimi ir kiti, tokie kaip *Adj + N + -īb-a* (*šaur-pier-īb-a* ‘siaurakaktiškumas’ ← *šaur-a pier-e* ‘siaura kakta’), *N + V + -īb-a* (*graud-kop-īb-a* ‘grūdininkystė’ ← *graud-us kop-t* ‘grūdais rūpintis’) ir kt.

Mūsų atlikto tyrimo šaltinis buvo 1,3 mlrd. žodžių apimties Jungtinis lietuvių kalbos tekstynas (Dadurkevičius 2020a, 2020b). Jame vertinome svarbiausių daiktavardžių priesagų produktyvumą, bet kartu pasižymėjome ir mišriosios darybos atvejus. Pastebėjome, kad daugiausia tokių darinių yra vardažodinės kilmės, ypač su ypatingu turėtojų priesaga *-ininkas*, *-ė*. Čia dominuoja du modeliai: (1) *N + N + -inink-as*, *-ė* (*kaim-o-turizm-inink-as* ← *kaim-o turizm-as*) ir (2) *Adj + N + -inink-as*, *-ė* (*žali-a-kort-inink-as* ← *žali-oji kort-a*). Pasitaikė ir darinių, besiremiančių skaitvardžiu ir būdvardžiu (*Num + N + -inink-as*, *-ė*: *tri-kėd-inink-as* ← *trys kėdės*) arba tam tikrais posakiais, kur išlaikoma kaitybinė pirmojo dėmens forma, pvz., *tėve-mūs-inink-as* ← *tėve mūs-ų*. Kai pamatu eina prielinksnis ir daiktavardis (*be-pens-inink-as* ← *be pens-ij-os*), tokį darinį galima laikyti pasidarytą ir prefiksacijos-sufiksacijos (kur *be-* funkcionuoja kaip priešdėlis), ir kompozicijos-sufiksacijos būdu, plg. Paulauskienė (1994: 49, 88, 92–96) apie prielinksnius kaip dūrinių pirmuosius sandus. Kitos mūsų tirtos priesagos mišriojoje daryboje pasitaikė rečiau – buvo rasta po keletą darinių su vardažodžių abstraktų priesagomis *-yst-ė* ir *-izm-as*, pvz.: *vien-a-sakin-ystė* ← *vien-as sakin-ys*, *sen-bambekl-izm-as* ← *sen-as bambekl-is*. Kaip ir anksčiau minėtais atvejais, matyti, kad dariniai remiasi žodžių junginiais, kurių kamienai suduriami išlaikant pamatinio junginio dėmenų tvarką.

Mišrioji daryba, kai antruoju kompozicijos sandu eina veiksmazodis, mūsų medžiagoje pasitaikė retai ir jos interpretacija kėlė abejonių. Iš veiksmo atlikėjo priesagą turinčių vedinių, pavyzdžiui, minėtinas *bendr-a-įkūr-ėj-as* (← *bendr-ai įkur-ti*; plg. angl. *co-found-er*). Galbūt mišriosios darybos rezultatais tikėtų laikyti ir priesagą *-yb-a*, ir daiktavardžio bei veiksmazodžio kamienus turinčius daiktavardžius, tokius kaip *kaln-a-kas-yb-a*, *knyg-riš-yb-a*, jei juos sietume ne tik su *kalnakas-ys*, *knygriš-ys*, bet ir tiesiai su pastarųjų pamatais – *kaln-as* ir *kas-ti*, *knyg-a* ir *riš-ti*, plg. anksčiau minėtą latvių N + V + *-īb-a* modelį.

Literatūra ir šaltiniai

- Aleksaitė, Agnė 2022, *Lietuvių kalbos naujažodžių daryba (2011–2019 m. Naujažodžių duomenyno pagrindu)*, humanitarinių mokslų daktaro disertacija, Vilnius: Lietuvių kalbos institutas.
- Auziņa, Ilze, Ieva Brenķe, Juris Grigorjevs, Inese Indričāne, Baiba Ivulāne, Andra Kalnača, Linda Lauze, Ilze Lokmane, Dace Markus, Daina Nītiņa, Gunta Smiltņiece, Baiba Valkovska, Anna Vulāne 2013, *Latviešu valodas gramatika*, Dainas Nītiņas un Jura Grigorjeva vadībā, Rīga: LU Latviešu valodas institūts.
- Dadurkevičius, Virginijus 2020a, Wordlist of Lemmas from the Joint Corpus of Lithuanian. CLARIN-LT digital library in the Republic of Lithuania (<http://hdl.handle.net/20.500.11821/41>).
- Dadurkevičius, Virginijus 2020b, *Assessment Data of the Dictionary of Modern Lithuanian versus Joint Corpora*, CLARIN-LT digital library in the Republic of Lithuania (<https://clarin.vdu.lt/xmlui/handle/20.500.11821/36>).
- Keinys, Stasys 1999, *Bendrinės lietuvių kalbos žodžių daryba*. Šiauliai: Šiaulių universiteto leidykla.
- Murmulaitytė, Daiva 2021, *Naujažodžių darybos ir morfemikos tyrimų perspektyvos (Lietuvių kalbos naujažodžių duomenyno atvejis)*, Vilnius: Lietuvių kalbos institutas (<https://doi.org/10.35321/e-pub.16.naujadaros-tyrimu-perspektyvos>).
- Paulauskienė, Aldona 1994, *Lietuvių kalbos morfologija: paskaitos lituanistams*, Vilnius: Mokslo ir enciklopedijų leidykla.
- Urbutis, Vincas 1978, *Žodžių darybos teorija*. Vilnius: Mokslo.
- Urbutis, Vincas, 1999, Darybos būdas, *Lietuvių kalbos enciklopedija*, parengė Kazys Morkūnas, redagavo Vytautas Ambrazas, Vilnius: Mokslo ir enciklopedijų leidybos institutas, 140.

ERIKA RIMKUTĖ

Vytauto Didžiojo universitetas
erika.rimkute@vdu.lt

Nauji lietuvių kalbos gramatiškai anotuoti tekstynai: morfologiškai anotuoto tekstyno rengimas

2024–2026 m. vykdant Europos Sąjungos *NextGenerationEU* projektą „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ (Nr. 02-098-K-0001; <https://sitti.vdu.lt/morfologiskai-ir-sintaksiskai-anotuotu-tekstynu-modeliai-dirbtinio-intelektu-apmokymui/>), rengiami morfologiškai ir sintaksiškai anotuoti tekstynai (jų dydis – 10 mln. žodžių). Vytauto Didžiojo universiteto Skaitmeninių išteklių ir tarpdisciplininių tyrimų institute (SITTI, anksčiau – Kompiuterinės lingvistikos centras) jau yra parengti anotuoti tekstynai: morfologiškai anotuotas tekstynas MATAS (1,7 mln. žodžių) ir sintaksiškai anotuotas tekstynas ALKSNIŠ (apie 60 tūkst. žodžių). Nors jau turima anotuotų tekstynų sudarymo patirties, vis dėlto rengiant naujus gramatiškai anotuotus tekstynus susiduriama su daugybe iššūkių: 1) gauti autorių leidimus naudoti jų tekstus (anotuotų tekstynų modeliai bus naudojami įvairių įrankių (tiek pagrįstų įprastomis technologijomis, tiek panaudojant dirbtinį intelektą) apmokymui, o toks tekstyno panaudojimas kelia daug nerimo ir baimių autoriams, leidykloms); 2) apsispręsti dėl tekstynų sudarančių tekstų proporcijų (bus stengiamasi sudaryti iš apylygių mokslinės, administracinės, grožinės literatūros ir žiniasklaidos tekstų dalių); 3) išsiaiškinti tekstyno vieneto (angl. *token*) sampratą (ar sudaro tik žodiniai ar ir nežodiniai vienetai (skaitmenys, skyrybos ženklai, simboliai) – tai ypač aktualu norint per trumpą laiko tarpą parengti didelius anotuotus tekstynus); 4) pasirinkti anotavimo ir anotuotų tekstų redagavimo įrankius; 5) pritaikyti tarptautinius anotavimo standartus. Sintaksiškai anotuotas tekstynas bus anotuotas tarptautiniu *Universal Dependencies* (toliau – UD) standartu (<https://universaldependencies.org/>). UD yra atviras tarptautinis projektas, kuriame dalyvauja tyrėjai, rengiantys morfologiškai ir sintaksiškai anotuotus tekstynus. UD kaip anotavimo standartas šiuo metu jau yra pritaikytas daugiau nei šimtui kalbų, UD svetainėje prieinama apie 200 tekstynų, anotuotų UD standartu. Šis standartas susiformavo iš tipologinių gramatinių teorijų (dėl to galima nuosekliai anotuoti tipologiškai skirtingas kalbas, taip plečiant automatinio natūraliosios kalbos apdorojimo ir lingvistinių tyrimų galimybes).

Šiame pranešime norima pristatyti morfologiškai anotuoto tekstyno rengimo etapus ir kilusias problemas:

- 1) pirmiausia tekstai anotuojami automatiškai morfologiškai (naudojant įrankį „Morfuoklis“ (<https://sitti.vdu.lt/morfuoklis/lt/>), pagrindiniame anotavimo standarte „Jablonskis“ (<https://sitti.vdu.lt/jablonskis-lt/>) naudojamos lietuvių lingvistiniuose darbuose įprastos kalbos dalių ir gramatinių kategorijų pažymos. Bet tam, kad morfologiškai anotuotus tekstus būtų galima anotuoti sintaksiškai tarptautiniu UD standartu, iš karto reikėjo pertvarkyti kai kurių kalbos dalių anotacijas, pvz., kelintinius skaitvardžius laikyti būdvardžiais; nurodyti trumpinių kalbos dalį; kai kuriuos įvardžius pažymėti kaip determinantus (pažyma *det*);
- 2) vienodai anotuoti lietuvių kalbos žodynuose, gramatikose neaiškiai aprašytus, nevienodai traktuojamus žodžius (ypač nekaitomuosius, pvz., tik – dalelytę irrieveiksmį, būdvardžiais ir dalyviais laikytinus žodžius, pvz., galimas, papildomas);
- 3) remiantis naujausia „Bendrinės lietuvių kalbos žodyno“ informacija, pakoreguoti kalbos dalių pažymas „Morfuoklyje“ (pvz., „Dabartinės lietuvių kalbos žodyne“ dar laikomarieveiksmiu ir ištiktuku, o „Bendrinės lietuvių kalbos žodyne“ – dalelyte ir ištiktuku);
- 4) spręsti techninius iššūkius: gerinti automatinį sakinių, žodžių segmentavimą (angl. *tokenization*), vienareikšminimo kokybę, „Morfuoklyje“ išlaikyti lietuvių kalbai įprastas gramatines pažymas, bet kartu jas tinkamai konvertuoti į tarptautinius standartus; išanalizuoti ir į „Morfuoklio“ leksinę bei gramatinę duomenų bazę įtraukti neatpažintus vienetus.

BAIBA SAULĪTE

LU Matemātikas un informātikas institūts
baiba.saulite@gmail.com

ILZE AUZIŅA

LU Matemātikas un informātikas institūts
ilze.auzina@lumii.lv

Spontānas runas marķējuma līmeņi un gramatiskā analīze

Spontānā runa ir cilvēka dabiskās komunikācijas forma, kas bieži vien ir nestrukturēta un neplānota. Ja neskaita rūpīgi iestudētas runas, runāto valodu parasti nevar segmentēt tādās vienībās, kas būtu ekvivalentas rakstītam teikumam (Kahane et al. 2021), jo ierakstītai runai nav nekāda dalījuma teikumos. Šī referāta mērķis ir konstatēt spontānās runas gramatiskās analīzes galvenos aspektus, lai labāk izprastu un interpretētu tās struktūru un saturu.

“LATE sarunu korpusā” ir iekļauta t.s. spontānā runa, arī sagatavots vai lasīts teksts, un šo datu analīze atklāj galvenās runātās valodas gramatikas iezīmes, piemēram:

- 1) tajā kā analīzes vienību nevar izmantot teikumu, jo runā teikumu nav, bet izteikumu robežas mēdz būt neskaidras,
- 2) izteikumu sintaktiskā struktūra var būt vienkāršota,
- 3) liela daļa runas vienību ir izsacījumi, resp., to pamatā nav teikuma struktūras shēmas,
- 4) izteikuma gramatisko struktūru ietekmē tas, ka runātā valoda veidojas spontāni, runas brīdī, ir interaktīva un neprognozējama.

Runas korpusos ir vairāk marķējuma līmeņu nekā teksta korpusos, un tajos iekļauta daudzveidīga informācija. Par pirmo marķējuma līmeni uzskatāma runas transkripcija pati par sevi, tāpēc ka runa ir sadalīta segmentos, sastatīta ar audioierakstu, tai ir pievienoti dažādi metadati, kā arī norādīta izruna, leksiskas un gramatiskas parādības, kas raksturīgas runai (piem., pauzes, atkāpes no normas, pārteikšanās). Tāpat “Nacionālajā korpusu kolekcijā” iekļautie runas korpusi ir automātiski morfoloģiski marķēti. Lai runas korpusus marķētu sintaktiski, ir jāizveido sintaktiskās marķēšanas metodika, un viens no galvenajiem uzdevumiem ir noskaidrot, kāda vienība runas korpusā būtu uzskatāma par sintaktiskās analīzes vienību.

Runas korpusos audiosignāls tiek segmentēts jeb dalīts „minimālās intonatīvi pabeigtās saziņas vienībās, kam var būt noteikts strukturāls paraugs, bet tas nav obligāts” (Skujiņa 2007, 173). Šie izteikumi tiek saukti arī par klauzālām (*clausal units*) vai ilokutīvām vienībām (IU) – runas segmentiem, kas atbilst vienam runas aktam, piemēram, jautājumam vai pavēlei. Segmentējot spontānas vai sagatavotas (parasti – lasīts teksts) runas ierakstus, tiek ievērots intonatīvais princips – segmenta robežas tiek noteiktas, ņemot vērā runas intonācijas elementus: pauzi, intensitāti, melodiju (skaņu frekvences maiņu), loģisko akcentu, runas tempu. Tātad dalīšana segmentos var būt balstīta uz prosodiju, bet prosodiskās vienības no sintaktiskajām vienībām var ievērojami atšķirties (Kahane, Lacheret 2019) un nav tādu vispārzināmu zīmju, kas norobežotu izteikumus runas plūsmā (Biber et al. 2021), turklāt dalījums segmentos var būt arī subjektīvs. “LATE sarunu korpusā” iekļautās transkripcijas nav pierakstītas, ievērojot dalījumu teikumos un interpunkcijas principus.

Referātā tiks analizēti kritēriji, pēc kā noteikt minimālo sintaktiskās analīzes vienību (intonācija, gramatiskā struktūra, pauzes utt.) spontānā runā, tāpat tiks pētīti starptautiski pieņemtie apzīmējumi intonatīvo vienību atspoguļošanā un tiks iztirzāts, ko, dalot runas ierakstus izteikumos, var secināt no līdzšinējā dalījuma segmentos un vai izteikumu nošķiršanā var palīdzēt prosodija.

Avots

- Auziņa, I., Dargis, R., Rābante-Bušā, G., Ļaksa-Timinska, I., Gailīte, E., Auziņa, A. (2024) LATE sarunu korpuss (LATE-sarunas), *CLARIN-LV digitālā bibliotēka*, <http://hdl.handle.net/20.500.12574/113>. Pieejams: <https://korpuss.lv/id/LATE-sarunas>

Literatūra

- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., Quirk, R. (2021) *Grammar of Spoken and Written English*. doi: [10.1075/z.232](https://doi.org/10.1075/z.232)
- Kahane S., Lacheret A. (2019) *Syntax and prosody mapping: What and how? The case of intonational periods and illocutionary units*. In Lacheret-Dujour et al. (2019), 339–363.
- Kahane, S., Caron, B., Strickland, E., Gerdes, K. (2021) Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*. Sofia, Bulgaria. Association for Computational Linguistics, 35–47.
- Skujiņa, Valentīna (red.). (2007) *Valodniecības pamatterminu skaidrojošā vārdnīca*. Rīga: Latvijas Universitātes Latviešu valodas institūts.

INTA URBANOVIČA

Latvijas Universitāte
inta.urbanovica@lu.lv

Čehu valodas elektroniskās lietojumprogrammas “Morfio” izmantošana latviešu valodas paronīmu izpētē

Pašreiz latviešu valodā tiek izstrādāta paronīmu – pēc skaniskās un morfoloģiskās uzbūves līdzīgu, bet semantiski atšķirīgu vienas vārdšķiras vārdu (Skujiņa 2007, 287) – vārdnīca, un tāpēc ir nepieciešams iegūt pēc iespējas apjomīgāku un pilnīgāku paronīmu sarakstu. Referāta mērķis ir demonstrēt lietojumprogrammas “Morfio” iespējas latviešu valodas paronīmu kopas izstrādē.

Čehu valodas nacionālajā korpusu kolekcijā (Institute of the Czech National Corpus, 2025) ir pieejami dažādi elektroniskie rīki un resursi, kas izstrādāti korpusu materiālu specifiskai apstrādei. Viens no šādiem rīkiem ir “Morfio” – lietojumprogramma, kas paredzēta vārdu derivatīvo saikņu identificēšanai un morfēmu produktivitātes izpētei (Institute of the Czech National Corpus, 2019). “Morfio” programmu ir iespēja izmantot četrās valodās – čehu, poļu, slovāku un latviešu.

“Morfio” latviešu valodas versija ir piesaistīta latviešu valodas korpusiem, un ar tā palīdzību var identificēt un izgūt no korpusu tekstiem vārdu pārus, kam ir viena sakne vai celms un atšķirīgi afiksi (priedēkļi, piedēkļi) vai morfēmu kopas (izskaņas). Latviešu valodā līdz šim nav izstrādāta šāda veida programma, un, lai iegūtu līdzīgus datus, nacionālo korpusu kolekcijas rīkos ir jāizmanto sarežģīti vaicājumu veidi. Savukārt “Morfio” ir samērā vienkārša un ērti izmantojama lietojumprogramma, ar kuras palīdzību var iegūt diezgan apjomīgu skaitu derivatīvi līdzīgu vārdu pāru. Katram darinājumam pārīm ir pievienoti arī kvantitatīvie rādītāji par lietojuma biežumu.

Paronīmu vārdnīcas vajadzībām datu apstrāde īstenota trīs posmos:

- vaicājumu izvēle un datu kvantitatīva analīze;
- iegūto vārdu pāru filtrēšana un kvalitatīva izvērtēšana paronīmijas aspektā;
- iegūto paronīmu pāru nozīmes differences izpēte.

Referātā tiks prezentēti “Morfio” lietojumprogrammas izmantošanas rezultāti paronīmu identificēšanā, kā arī tiks veikta šo rezultātu kvantitatīva un kvalitatīva analīze un izvērtēta to lietderība latviešu valodas paronīmu kopas izstrādē.

Literatūra

- Institute of the Czech National Corpus. 2019. Morfio (2013–2019). Pieejams: <https://morfio.korpus.cz>
- Institute of the Czech National Corpus. 2025. The Czech National Corpus (2025). Pieejams: <https://www.korpus.cz>
- Skujiņa, Valentīna. 2007. Valodniecības pamatterminu skaidrojošā vārdnīca. Rīga: LU Latviešu valodas institūts.

EVELĪNA ZILGALVE

Latvijas Universitāte
evelina.zilgalve@lu.lv

Latviešu valodas elektronisko resursu izmantojums valodas konsultācijās

Valodas konsultāciju pakalpojumi lielā daļā Eiropas valstu ir pētniecības institūtu uzdevums, piemēram, Čehijā – Čehu valodas institūts, Igaunijā – Igaņu valodas institūts vai Slovēnijā – Frana Ramovša Slovēņu valodas institūts. Daļā valstu konsultācijas sniedz arī valsts iestādes (arī valsts pētniecības iestādes), piemēram, Dānijā – Dāņu valodas padome, Lietuvā – Valsts lietuviešu valodas komisija, Norvēģijā – Norvēģijas Valodas padome, Latvijā – Latviešu valodas aģentūra (sk. Language consultations 2023; par vēl citām institūcijām sk. arī Ludānyi 2020).

Dažādās iestādēs atšķiras konsultāciju sniegšanas laiks un formas, piemēram, telefonkonsultāciju pieejamība, iespēja uzdot jautājumus katru dienu vai tikai noteiktās dienās un stundās u. tml., taču jautājumu tematika ir līdzīga: neskaidrības gramatikas vai pareizrakstības normās, leksikas, sevišķi svešvārdu, skaidrojums, iederīgāko vārdu meklējumi (Telmer 2024, 2).

Vairums latviešu valodas konsultāciju skar personvārdu atveidi (sk. MK 114), atlikusī daļa, kas prasa lielāko izpēti, skar tieši gramatikas, leksikas, pareizrakstības, stila neskaidrības (procentuālo sadalījumu sk. Zilgalve [b. g.]). Konsultācijas tiek sniegtas gan e-pastā, gan telefoniski, un abos gadījumos tiek prasīta ātra un precīza atbilde, kas pamatota vārdnīcās, gramatikās vai citās normatīvās publikācijās, turklāt nereti klients pats jau ir tās apskatījis un vēršas pie valodas konsultantiem, jo nav radis atbildi vai uzgājis avotos pretrunas.

21. gadsimtā tiek pārskatīta vārdnīcu un gramatiku loma valodas normēšanā (McLelland 2021) un uzsvērta digitālo mediju nozīme valodas standartizācijā (Bellamy 2021), un arī mūsdienīgas valodas konsultācijas vairs nav iedomājamas bez elektronisko resursu izmantojuma (par korpusu lietojumu valodas konsultācijās sk. Ludānyi 2024; Ludānyi, Domonkosi 2023).

Referātā aplūkots elektronisko resursu izmantojums latviešu valodas konsultāciju sniegšanā:

- vārdnīcu platformas tezaurs.lv vārdnīcas “Tēzaurs”, “Mūsdienu latviešu valodas vārdnīca” (MLVV) un “Latviešu literārās valodas vārdnīca” (LLVV);
- Nacionālā korpusu kolekcija (NKK);
- periodika.lv.

Apskatītas elektronisko resursu iespējas, sniedzot atbildes uz sabiedrības jautājumiem:

- kādas normatīvās norādes atrodamas MLVV, piemēram, ieteikums vārda “ofiss” vietā lietot “birojs”,
- kā, izmantojot vārdnīcu un korpusu datus, iespējams izsecināt vārdnīcās neaprašītu nozīmes, piemēram, vārdiem “shēmot” vai “tāpatot”,
- kopā vai šķirti rakstāmu jēdzienu meklējumi kā vārdnīcās, tā korpusos, piemēram, “gulašzupa”/“gulaša zupa”, “ķirštomāts”/“ķiršu tomāts”, “veļasmazgātava”/“veļas mazgātava”,
- vārda lietojuma dati no NKK un periodika.lv, piemēram, vārda “forši” lietojuma pirmsākumi 19. gadsimta beigās vai vārdnīcās nefiksētais vārds “savpatns” Imanta Ziedoņa un Vairas Viķes-Freibergas leksikā.

Secināms, ka elektronisko resursu pieejamība nemazina valodas konsultāciju pieprasījumu skaitu, taču nereti ir skaidrojams, kā darboties ar tiešsaistes vārdnīcām, piemēram, daļai vārdnīcu lietotāju nepieciešams skaidrot “Tēzaurā” atrodamās uzziņas, lai gan tajā atrodamas norādes uz citām vārdnīcām un enciklopēdijām. Ir vērojama arī gluži pretēja tendence: lietotāji ir labi iepazinušies ar dažādiem elektroniskajiem resursiem, atraduši tajos trūkumus vai pretrunas, un valodas konsultantu uzdevums ir, balstoties uz pieejamajiem datiem, sniegt savu viedokli.

Literatūra

- Bellamy 2021 – Bellamy, John. 2021. Contemporary Perspectives on Language Standardization: The Role of Digital and On-line Technologies. Language Standardization. Cambridge: Cambridge University Press, 691–712.
- Language consultations 2023 – Language consultations: traditions and current trends. Workshop programme. https://valoda.lv/wp-content/uploads/2023/09/Workshop_programme_ENG-2.pdf.
- LLVV – Ceplītis, Laimdota un Spektors, Andrejs Latviešu literārās valodas vārdnīca (LLVV), 2025, Pavasara versija. CLARIN-LV digitālā bibliotēka. <http://hdl.handle.net/20.500.12574/126>.
- Ludányi 2020 – Ludányi, Zsófia. 2020. Language consulting: a brief European overview. DOI: 10.36007/eruedu.2020.3.025-047.
- Ludányi 2024 – Ludányi, Zsófia. 2024. The role of corpora in the work of the Language Consulting Service of the HUN-REN Hungarian Research Centre for Linguistics. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 15(2), 47–73. <https://doi.org/10.12697/jeful.2024.15.2.03>.
- Ludányi, Domonkosi 2023 – Ludányi, Zsófia, Domonkosi Ágnes. 2023. Language consulting and language management from the perspective of the Hungarian language Consulting Service. *Taiko-moji kalbotyra* 20. 74–88. <https://doi.org/10.15388/Taikalbot.2023.20.6>.
- McLelland 2021 – McLelland, Nicola. 2021. Grammars, Dictionaries and Other Metalinguistic Texts in the Context of Language Standardization. Language Standardization. Cambridge: Cambridge University Press, 263–293.
- MK 114 – Ministru kabineta 2004. gada 2. marta noteikumi Nr. 114 „Noteikumi par personvārdu rakstību un lietošanu latviešu valodā, kā arī to identifikāciju”. <https://likumi.lv/ta/id/85209>.
- MLVV – Zuicena, Ieva et al. Mūsdienu latviešu valodas vārdnīca (MLVV), 2025, Pavasara versija. CLARIN-LV digitālā bibliotēka. <http://hdl.handle.net/20.500.12574/128>.
- NKK – Nacionālā korpusu kolekcija. <https://korpuss.lv/.periodika.lv> – <https://periodika.lndb.lv/>.
- Telmer 2024 – Standard of language consulting. https://ujc.cas.cz/wp-content/uploads/2024/09/Standard-of-language-consulting_FV.pdf.
- “Tēzaurs” – Spektors, Andrejs et al. Tēzaurs.lv, 2025, Pavasara versija. CLARIN-LV digitālā bibliotēka. <http://hdl.handle.net/20.500.12574/127>.
- Zilgalve [b. g.] – Zilgalve, Evelīna [b. g.] Latvian Language Standardization: The perspective of language consultants (iesniegts publicēšanai).