AURELIJA TAMULIONIENĖ
Lietuvių kalbos institutas

Mokslinių tyrimų kryptys: dabartinės lietuvių kalbos vartosenos ir jos variantiškumo tyrimai.

LIETUVIŲ KALBOS PADĖTIS SKAITMENINIAME AMŽIUJE

Apie 24-ąją Jono Jablonskio konferenciją *Skaitmeninių kalbos išteklių plėtros kryptys ir panaudos galimybės*

Jonas Jablonskis (1860–1930) – žymiausias lietuvių kalbos normintojas, reikšmingų lietuvių kalbos mokslui ir praktikai darbų autorius. Pasak Zigmo Zinkevičiaus, „[V]visa Jablonskio veikla – tai ištisa epocha lietuvių bendrinės kalbos istorijoje. Jis mūsų rašomąją kalbą įstatė į tinkamas vėžes [...], jos raidą pasuko sveika linkme“ (Zinkevičius 1992: 155).

Jono Jablonskio vardu pavadinta mokslinė konferencija pirmą kartą surengta 1993 m. Lietuvių kalbos instituto Kalbos kultūros skyriaus ir Vilniaus universiteto Lietuvių kalbos katedros. Nuo to laiko konferencijos tapo tradicinės ir vyksta kiekvieną rudenį (nuo 2003 m. – pakaitomis Lietuvių kalbos institute arba Vilniaus universitete).

2017 m. rugsėjo 29 d. vyko 24-oji mokslinė Jono Jablonksio konferencija *Skaitmeninių kalbos išteklių plėtros kryptys ir panaudos galimybės* (an. Digital Language Resources, Directions of Their Development and Possibilities to Harness Them) surengta Lietuvių kalbos institute. Konferenciją organizavo Lietuvių kalbos instituto Bendrinės kalbos tyrimų centras kartu su Vilniaus universiteto Taikomosios kalbotyros institutu. Savo tematika ši konferencija skyrėsi nuo ankstesnių, jos tikslas – aptarti kalbų ir skaitmeninių technologijų sandūroje kylančias problemas, su kuriomis susiduriama plėtojant skaitmeninius išteklius ir kuriant bendrąją skaitmeninę rinką. Konferencijos akiratyje – semantinis skaitmeninių išteklių lygmuo (žodžių tinklų kūrimas), taip pat garsinio pavidalo suteikimas skaitmeniniams ištekliams, naujų galimybių kūrimas skleidžiant skaitmeninius išteklius įvairiakalbėse bendruomenėse ir kt.

Konferencijoje dalyvavo pranešėjai iš užsienio: Europos Parlamento, Ispanijos, taip pat tyrėjai iš skirtingų Lietuvos institucijų: Lietuvių kalbos instituto, Vilniaus

universiteto, Vytauto Didžiojo universiteto, Baltijos pažangių technologijų instituto, UAB „Netcode“. Konferencijoje perskaityta 17 pranešimų. Pranešėjai savo pranešimus skaitė ir dalijosi savo tyrimais ir patirtimi keturiuose posėdžiuose.

Konferencija prasidėjo įžanginėmis kalbomis. Susirinkusius konferencijos dalyvius ir pranešėjus pasveikino Europos Parlamento narys Algirdas Saudargas ir Lietuvių kalbos instituto direktorė prof. dr. Jolanta Elena Zabarskaitė. Įžanginį žodį tarė Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro vyriausioji mokslo darbuotoja dr. Rita Miliūnaitė. Kalbėta apie tai, kad lietuvių kalbos ir informacinių technologijų sąsajos pastaraisiais metais labai sustiprėjo, turime apčiuopiamų rezultatų tiek skaitmenindami kalbos išteklius, tiek kurdami kalbos analizės, atpažinimo ir sintezavimo įrankius, bet, palyginti su tuo, kaip sparčiai šioje srityje keičiasi pasaulis, lietuvių kalbai dar daug reikia vytis. Kadangi lietuvių kalba yra fleksinė, jai negalime tiesiogiai pritaikyti anglų kalbai sukurtų informacinių technologijų, todėl dabartiniai mūsų mašininio vertimo ir kiti kompiuteriniai įrankiai dar labai netobuli, o pasaulis jau tvirtai eina toliau, dirbtinio intelekto kūrimo kryptimi, įkalbina daiktus ir skverbiasi į vis gilesnius kalbos klodus.

Plenariniame posėdyje pirmąjį pranešimą „Gyvoji kalba dirbtiniame prote“ skaitė Europos Parlamento narys Algirdas Saudargas. Pranešėjas kėlė esminį klausimą, į kurį turi atsakyti kiekviena kalbinė bendruomenė: kokiu mastu ji turi pati parengti savo gimtosios kalbos išteklius ir technologijas, o ką galima įsigyti jau pagaminta kitų kalbų pagrindu. Pranešime išsakytos probleminės mintys, kad lietuvių kalba kartu su kai kurių kitų mažų šalių kalbomis „turinti menką technologinę paspirtį arba jos visai neturinti“. Kalbos technologijos nesulaukia tinkamo dėmesio ne tik Lietuvos, bet ir Europos politikų darbotvarkėje. Pranešime pateiktos įžvalgos, rodančios, kad dirbtinio intelekto raida konverguoja į hibridinį pavidalą, kuriame neuroniniai tinklai atitinka nesąmoningus smegenų mechanizmus, o sąmoningą proto veiklą modeliuoja tradicinis, vadinamasis simbolinis dirbtinis intelektas. Kiekviena kalbinė bendruomenė privalo rūpintis, kad kalbos technologijos, atitinkančios simbolinį dirbtinį intelektą išsamiai ir tiksliai atspindėtų visą gimtosios kalbos sandarą, o neuroninių tinklų savimokai būtų sukurta turininga gimtosios kalbos (ir gimtosios kultūros) aplinka.

Antrąjį plenarinį pranešimą „Language equality in the digital age: towards a human language project“ anglų kalba skaitė Rafael Rivera, konsultacinės įmonės „Iclaves“ direktorius. Pranešime išsamiai pristatytas Europos Parlamento Mokslinio perspektyvų tyrimo skyriaus tyrimas „Kalbų lygybė skaitmeniniame amžiuje. Gimtosios kalbos projektas“ (su tyrimu anglų kalba galima susipažinti internete: http://www.europarl.europa.eu/RegData/etudes/STUD/2017/598621/EPRS_STU(2017)598621_EN.pdf). Pranešime apžvelgta dabartinė gimtosios kalbos technologijų padėtis, kiekybiškai įvertinant

ekonominius, socialinius ir lingvistinius kalbos padarinius skaitmeniniame amžiuje, aptarta Europos Sąjungos informacinių ir ryšių technologijų politika. Skaitmeniniame amžiuje kalbos technologijos yra didžiulis iššūkis mažai kalbėtojų turinčioms šalims. Prastai išvystytos kalbų technologijos sudaro atskirtį ir kliūtis. Tai daro įtaką tarpvalstybinėms paslaugoms, darbuotojų judumui, prekybai. Tokių kliūčių priežastis – nesuderinti moksliniai tyrimai ir nepakankamas finansavimas. Pranešėjas pabrėžė, kad kalbų technologijos nesulaukia atitinkamo Europos politikų dėmesio. Remiantis dabartinės padėties analize, kuriama iniciatyva, kuri vienys institucijų, mokslinių tyrimų, pramonės, rinkos ir viešųjų paslaugų politiką.

Paskutinį plenarinio posėdžio pranešimą „Rašytinė ir sakytinė lietuvių kalba įprastoje ir elektroninėje terpėse“ skaitė Vilniaus universiteto Matematikos ir informatikos instituto profesorius Laimutis Telksnys. Pranešime kalbėta, kad būtina parengti metodus, įrankius – techninę ir programinę įrangą – užtikrinančią taisyklingą lietuvių rašytinės ir sakytinės kalbos vartojimą popieriniuose dokumentuose ir elektroninėje terpėje. Pranešėjas kėlė svarbų klausimą – kaip elgtis keliems milijonams lietuvių, kad neišnyktų daugiau kaip 7 milijardų rašančiųjų ir šnekančiųjų žmonių bei ateinančių išmaniųjų mašinų okeane. Profesorius pateikė ir atsakymą – kad taip nenutiktų, būtina parengti metodus, įrankius – techninę ir programinę įrangą – užtikrinančią taisyklingą lietuvių rašytinės ir sakytinės kalbos vartojimą popieriniuose dokumentuose ir elektroninėje terpėje, harmonizuojančią lietuvių ir kitų kalbų rašytinės ir sakytinės kalbų vartojimą popieriniuose dokumentuose ir elektroninėje terpėje. Reikia sukurti transkribatorius – nelietuvių rašytinės kalbos ženklams pavaizduoti lietuviškais rašytinės kalbos ženklais ir nelietuvių sakytinės kalbos garsams pateikti lietuvių sakytinės kalbos ženklais, tinkančiais automatinei lietuvių šnekos žodžių garsų sintezei.

Pirmajame posėdyje „Šnekos atpažinimas ir sintezė“ perskaityti trys pranešimai.

Audrius Valotka (VU) ir Gediminas Navickas (VU) pranešime „Tolyn nuo Gutenbergo spaudos preso: lietuvių šneka naujausiose technologijose“ kalbėjo apie lietuvių šneką naujausiose technologijose, pristatė Struktūrinių fondų finansuojamą projektą *Lietuvių šneka valdomos paslaugos – LIEPA* (prieiga internete: https://www.raštija.lt/liepa), kurio metu buvo sukurtas, be viso kito, lietuvių šnekos sintezatorius ir atpažintuvas. Šio projekto vykdymo metu lietuvių šnekai atverti vartai į skaitmeninę erdvę, dėl šios priežasties planuojamas projektas *Lietuvių šneka valdomų paslaugų plėtra – LIEPA 2*. Pranešėjai kalbėjo apie projekto LIEPA 2 rezultatus, kurie bus atvirai ir nemokamai prieinami visiems norintiems, skatins lietuvių šnekos naudojimą informacinių technologijų produktuose. Svarbiausias naujojo projekto rezultatas bus galimybė natūralia

šneka bendrauti su daiktais (mobiliaisiais telefonais, planšetėmis, išmaniaisiais laikrodžiais, robotais), duoti komandas žmogaus balsu ir suprasti jų atsakymus. LIEPA 2 numatomos sukurti tipinės paslaugos, rodančios naujas lietuvių šneka valdomų paslaugų panaudojimo galimybes: ugdančio roboto valdytuvą, skambintuvą, taksi iškviestuvą, mobilųjį sintezatorių akliesiems, interneto naujienų skaitytuvą, tarpkalbinį komunikatorių.

Laimutis Telksnys (VU) ir Gediminas Navickas (VU) pranešime „Lietuvių šneka valdomos paslaugos. Padėtis. Perspektyvos“ kalbėjo apie lietuvių šneka valdomų paslaugų padėtį ir perspektyvas, sistemingai vykdomus lietuvių šneka valdomų paslaugų kūrimo ir jų panaudojimo plėtros darbus. Darbai vykdomi sutelkus informacinių technologijų specialistų, lietuvių kalbos filologų žinias ir inžinerines pajėgas. Pirmajame darbų etape sukurtos septynios lietuvių šneka valdomos paslaugos bei padaryta infrastruktūra, užtikrinanti sukurtųjų paslaugų funkcionavimą. Ateityje numatoma sukurti naujas lietuvių šneka valdomas paslaugas mobiliajai elektroninei terpei – išmaniesiems mobiliesiems telefonams, planšetėms, išmaniesiems laikrodžiams, robotams.

Apie lietuviško balso sintezės dabartį ir perspektyvas kalbėjo Pijus Kasparaitis (VU), Gintaras Skersys (VU). Pranešime „Lietuviško balso sintezės dabartis ir perspektyvos“ pristatytos įvairios teksto įgarsinimo paslaugos – LIEPA sintezatorius, – kuris yra laisvai platinamas kartu su pradiniais tekstais, tai sudaro sąlygas ir kitiems žmonėms rasti naujus jo pritaikymus, pavyzdžiui, balso įrašus šalia straipsnelių jau pateikia interneto svetainės: *lzinios.lt*, *ukininkopatarejas.lt*, *vilnius.lt*, *m.delfi.lt*, *vle.lt*., teksto įgarsinimo paslauga *RoboBraille* leidžia automatiškai paversti bet kokius tekstinius dokumentus į garso failus, pranešimai sintetiniu balsu keleiviams jau skaitomi Vilkaviškio autobusų stotyje, virtualūs asistentai jau veikia Migracijos departamento svetainėje ir t. t.

Antrajame posėdyje „Skaitmeniniai lietuvių kalbos ištekliai“ perskaityti keturi pranešimai.

Pirmąjį pranešimą „E. kalba – skaitmeninių kalbos išteklių naudojimo(si) inovacija“ skaitė Elena Jolanta Zabarskaitė (LKI), Deimantė Budriūnaitė (LKI), Skirmantas Šermukšnis (NetCode). Pranešime išsamiai pristatytas naujasis Lietuvių kalbos instituto projektas, skirtas Lietuvių kalbos išteklių informacinei infrastruktūrai plėtoti (LKIIS) (prieiga internete: www.lkiis.lt). Pranešime daugiausia kalbėta apie naujojo Lietuvių kalbos instituto projekto *E. kalba* infrastruktūros generuojamas paslaugas – „Paieška žodžių tinkle“, „E. rinkodara“, „E. sąvokos“ ir „E. patarimai“. Pranešėjai pristatė šių paslaugų funkcionalumus, paslaugų technologinius aspektus. Projekto funkcionalumai apims inovatyvias dirbtinio intelekto technologijas, skirtas vartotojų nuomonių analizei, paslauga „E. sąvokos“ užtikrins išplėstinę paiešką ir sąvokų praturtinimą, integruojant papildomus kalbinius išteklius, tokius kaip dvikalbiai žodynai,

kitų kalbų žodžių tinklai, bus sudarytos galimybės naudojantis interaktyviu žodžių darybos vedliu greitai ir patogiai pasirinkti tinkamus žodžio darybos būdus pagal norimą kategorinę ir grupinę darybos reikšmę arba pagal žodžio darybos priemones. Pranešime pristatyti ir pagrindiniai projekto *E. kalba* rezultatai bei planuojama nauda.

Rita Miliūnaitė (LKI) savo pranešime „Paieškos galimybės internetiniame *Lietuvių kalbos naujažodžių duomenyne*“ pristatė, kaip kalbos vartotojams atskleisti kuo įvairesnių jų poreikius atitinkančių *Lietuvių kalbos naujažodžių duomenyno* (prieiga internete: http://naujazodziai.lki.lt/) naujažodžių ypatybių. Šiuo metu galima paieška pagal tokius parametrus: antraštinį žodį, naujažodžio kilmę, originalo formą, rašybos variantus, vartojimo sritį ir t. t. Duomenų bazės tvarkytojams prieinama ir platesnė paieška, reikalinga duomenims įvairiais pjūviais redaguoti. Projekto LKIIS metu bus kuriama išplėstinė informacijos paieškos sistema. Pranešėja pabrėžė, kad *Lietuvių kalbos naujažodžių duomenynas* yra lankstus ir atviras plėtrai skaitmeninis išteklius. Tokį jo pobūdį pirmiausia diktuoja patys duomenys – nuolat kintantis ir atsinaujinantis lietuvių kalbos leksikos sluoksnis, taip pat naujažodžių tyrėjams atsiveriantys vis nauji duomenų požymiai ir besikeičiantys vartotojų poreikiai. Šį išteklių numatoma integruoti į LKIIS (prieiga internete: http://lkiis.lki.lt/) ir įtraukti ne tik į bendrą paieškos šioje duomenų sankaupoje sistemą, bet ir panaudoti kuriant žodžių tinklus. Pranešime vaizdžiai parodyta, kuo šis duomenynas gali būti naudingas tiek kalbos specialistams, tiek visiems naujažodžiais besidomintiems vartotojams.

Naujažodžių temą pranešime „Naujažodžių darybos tyrimų perspektyvos (*Lietuvių kalbos naujažodžių duomenyno* atvejis) tęsė Daiva Murmulaitytė (LKI), ji kalbėjo apie naujažodžių darybos tyrimus, perspektyvas, kaip ryškėjančius naujažodžių darybos reiškinius tirti pasitelkiant *Lietuvių kalbos naujažodžių duomenyną*. Pranešėja pristatė, kad jau pradiniu šios duomenų bazės kūrimo etapu preliminariai pasirengta ir naujažodžių darybos analizei – tam skirtuose laukuose numatyta nurodyti darinių rūšis, jų darybos formantus, darybos kategorijas (reikšmes), giminiškus žodžius ir kt.; sukaupta dalis duomenų. Ypatingųjų požymių lauke kai kurie naujažodžiai pažymėti kaip pavieniai (pavyzdžiui, *varškėfobija* ir kt.), autoriniai (pavyzdžiui, *demaguogija* ir kt.), kontaminaciniai (pavyzdžiui, *murmanas* ir kt.). Pranešėja pabrėžė, kad išsamiai ir kokybiškai žodžių darybos analizei taip pat svarbu atsižvelgti į kai kuriuos pradžioje nenumatytus fiksuoti dalykus – darinio pamatinio žodžio kalbos dalį, darybos pamato pakitimus, analogijos, vertimo vaidmenį darantis naujažodį, netipinės darybos apraiškas ir kt. Svarbu sukurti gerą – išsamią, lanksčią, patogią naudoti, pildyti bei koreguoti – indeksų sistemą.

Laimutis Bilkis (LKI) pranešime „Lietuvos vietovardžių geoinformacinės duomenų bazės struktūra, ryšiai su kitomis vardyno bazėmis ir plėtros kryptys“

pristatė *Lietuvos vietovardžių geoinformacinę duomenų bazę* (prieiga internete: http://lkiis.lki.lt/lietuvos-vietovardziu-geoinformacine-duomenu-baze). Pranešime kalbėta apie šios bazės struktūrą, ryšius su kitomis vardyno bazėmis. Pranešėjas pristatė duomenų bazės viešos prieigos puslapyje sukurtas paieškos sritis: objekto tipą, objekto statusą, dabartinę administracinę teritorinę priklausomybę (savivaldybė, seniūnija, gyvenvietė), tarpukario administracinė teritorinė priklausomybė (apskritis, valsčius, gyvenvietė), upių intakus. Bazėje informacijos apie vietovardžius galima ieškoti pagal šiuos požymius: vardas, giminė, skaičius, kirčiuotė, darybos būdas, kilmė pagal pamatinio žodžio priklausymą kalbai, kilmė pagal pamatinio žodžio priklausymą leksinei grupei. Bazėje galima susirasti reikiamos gyvenvietės (kaimo, bažnytkaimio, dvaro, miestelio, vienkiemio) teritorijoje tarpukariu užrašytus autentiškus vietovardžius, matyti tikslią ar apytikslę jų vietą žemėlapyje. Pranešime akcentuota, kad bazę integruojant į LKIIS sistemą sukurtos trejopos nuorodos į kitas vardyno bazes: į kitą vietovardį, iš kurio vietovardis kilęs, į *Lietuvių pavardžių duomenų bazėje* esantį asmenvardį, iš kurio vietovardis kilęs, į istorinius to vietovardžio užrašymus, esančius *Istorinių vietovardžių duomenų bazėje*. Šiuo metu į bazę įkelta ir aprašyta (aprašoma) apie 25 000 vietovardžių.

Po pietų pranešėjai ir dalyviai rinkosi į trečiąjį posėdį „Tekstynų lingvistika“.

Pirmąjį pranešimą „Lietuvių kalbos morfologiškai ir sintaksiškai anotuoti tekstynai“ skaitė Erika Rimkutė (VDU), Agnė Bielinskienė (VDU), Loïc Boizou (VDU), Andrius Utka (VDU). Kalbėta apie du anotuotus lietuvių kalbos tekstynus, parengtus Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre (tekstynų prieiga internete: http://nl.ijs.si/ME/V4/msd/html/index.html; https://ufal.mff.cuni.cz/tred/). Anotuoti tekstynai – pagrindiniai ištekliai, be kurių neapsieinama plėtojant kalbos technologijas. Jie paprastai naudojami kitiems natūraliosios kalbos ištekliams ir įrankiams kurti tokiose srityse, kaip automatinio kalbos atpažinimo sistemos, automatizuotas vertimas ir pan. Morfologiškai anotuotas tekstynas MATAS rengtas 2002–2014 metais. Jį sudaro 1,6 mln. žodžių iš įvairių stilių tekstų. Tekstynas parengtas 1 mln. žodžių tekstyno, sudaryto 2006 m., pagrindu pritaikant statistinius modelius. Sintaksiškai anotuotas tekstynas ALKSNIS, kaip aukso standartas tolesniems tyrimams ir ištekliams, parengtas 2016 m. Šį tekstyną sudaro 2 355 sakiniai (apie 30 tūkst. žodžių), imti iš įvairių stilių tekstų. Tekstyno anotavimas paremtas automatinio morfologinio ir sintaksinio anotavimo principais, pritaikytas sintaksinių priklausomybių modelis.

Tekstynų temą savo pranešime „Duomenų bazė lietuvių kalbos pastoviesiems junginiams“ toliau tęsė gausiausias konferencijos pranešėjų kolektyvas Erika Rimkutė (VDU), Agnė Bielinskienė (VDU), Loïc Boizou (VDU), Ieva Bumbulienė (Baltijos pažangių technologijų institutas), Jolanta Kovalevskaitė

(VDU), Tomas Krilavičius (Baltijos pažangių technologijų institutas), Justina Mandravickaitė (Baltijos pažangių technologijų institutas), Laura Vilkaitė (Baltijos pažangių technologijų institutas). Pranešimą konferencijoje pristatė Erika Rimkutė (VDU), ji daugiausia kalbėjo apie dabartinės rašytinės lietuvių kalbos pastoviųjų žodžių junginių tyrimo metodiką, rengiamą tekstynu paremtą lietuvių kalbos kolokacijų žodyną. Vykdant projektą *Lietuvių kalbos pastoviųjų žodžių junginių automatinis atpažinimas (PASTOVU)* (nr. LIP-027/2016) (žr. http://mwe.lt/), siekiama sukurti dabartinės rašytinės lietuvių kalbos pastoviųjų žodžių junginių tyrimo metodiką, parengti tekstynu paremtą lietuvių kalbos kolokacijų žodyną. Projekte sudarytas ir naudojamas 2014–2016 m. *Delfi.lt* tekstynas, kurio apimtis – 72 mln. žodžių. Kolokacijų žodynas bus rengiamas remiantis duomenų baze. Joje bus pateikta įvairialypė informacija apie pastoviuosius junginius: gramatinė, leksinė informacija, vartosenos dažnumas, teksto rubrika, konkordanso pavyzdžiai ir pan.

Tekstynų tyrimus pristatė ir Vilniaus universiteto atstovės Gintarė Judžentytė (VU), Vilma Zubaitienė (VU). Pranešime „Tekstynais paremti akademinių frazių tyrimai: formalioji struktūra ir semantika“ kalbėta apie akademinės kalbos frazių struktūrą ir semantiką, remiantis šiuo metu Vilniaus universitete kuriamo Studentų rašto darbų tekstynu, kuris yra vienas iš Valstybinės lietuvių kalbos komisijos remiamo projekto *Studentų darbų fraziškumo tyrimai ir interaktyvusis frazemų sąvadas* numatomų rezultatų. Projektu siekiama nustatyti akademiniam rašymui svarbių žodžių sąrašą, išskirti pagrindines kolokacijas ir jų plėtinius, taip pat pasikartojančias žodžių sekas įvairiose akademinio teksto dalyse, aptarti jų sąsajas su retorinėmis teksto dalių funkcijomis ir aprašyti reikšminių akademinių žodžių, tokių kaip: *tikslas*, *uždavinys*, *metodai*, *išvados*, *rezultatai*; *atlikti*, *analizuoti*, *nustatyti*, *remtis* ir kt. vartoseną ir semantiką.

Paskutiniame posėdyje „Automatizuota kalbos sandaros ir tekstų analizė“ perskaityti keturi pranešimai.

Posėdis prasidėjo Danieliaus Ralio (VU) pranešimu „Mašininis vertimas lietuvių kalbai“ apie naujausius mašininio vertimo istoriją, pasiekimus, įvairių institucijų vykdomus projektus, taip pat naujus proveržius naudojant neuroninį mašininį vertimą. Pranešėjas kalbėjo apie mašininio vertimo raidą ir pasiekimus, kurie šiandien efektingai taikomi ir lietuvių kalbai. 2005–2007 m. Vytauto Didžiojo universitetas vykdė ES Struktūrinių fondų finansuojamą projektą *Internetinė informacijos vertimo priemonė*. Rezultatas – vieša internetinė vertimo iš anglų į lietuvių k. paslauga (prieiga internete: http://vertimas.vdu.lt/twsas/). 2012–2014 m. Vilniaus universitetas vykdė ES finansuojamą projektą *Anglų–lietuvių–anglų ir prancūzų–lietuvių–prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas*. Rezultatas – vieša internetinė vertimo paslauga (prieiga internete: https://www.versti.eu/). Vis dėlto

net geriausiems mašininiams vertimams reikalingas vertėjo įsikišimas. Tad ar mašinos gali versti išties gerai? Pastarieji keleri metai žada naujus proveržius naudojant neuroninį mašininį vertimą. Vilniaus universitetas ruošiasi 2018 m. pradėti įgyvendinti naujos kartos neuroninį mašininį vertimą anglų, lietuvių, lenkų, prancūzų, rusų ir vokiečių kalboms. Pasidžiaugta, kad naujausi mašininio vertimo pasiekimai neaplenkia ir lietuvių kalbos.

Virginijus Dadurkevičius (VU) pranešime „Lietuvių kalbos gramatika skaitmeniniame atvirojo kodo pasaulyje“ kalbėjo apie naujos kartos kompiuterinės lietuvių kalbos morfologijos kūrimą, žodžių morfologinę analizę bei sintezę, intelektualią tekstinės informacijos paiešką ir pan. Atsiradus kompiuteriams klasikinei gramatikai ir leksikai iškilo būtinybė „apsivilkti naują rūbą“ ir tapti pilnaverte naujųjų technologijų dalimi. Tam sukurta naujos kartos kompiuterinė lietuvių kalbos morfologija, iškelti tokie tikslai: naudoti ir kurti tik atvirąjį kodą; tik patys duomenys, bet ne jų forma ir interpretavimas, gali turėti lietuvių kalbai specifinių savybių, visas programinis kodas turi būti universalus, tikti bet kuriai kitai kalbai; maksimaliai pasinaudoti kitoms pasaulio kalboms sėkmingai pritaikytais sprendimais. Sudarymo pagrindas – *Dabartinės lietuvių kalbos gramatika* (Vilnius, 2006) ir tekstynai (iš viso apie 1,5 mlrd. žodžių). Vidutinio šiuo metu internete atsirandančio teksto „atpažįstamumas“ yra apie 99 %, t. y. vidutiniškai 99-is iš 100-o žodžių morfologinis analizatorius interpretuoja teisingai. Šis naujai sukurtas morfologinės analizės būdas buvo sėkmingai pritaikytas Vytauto Didžiojo universiteto Sintaksinės-semantinės analizės sistemoje (prieiga internete: https://semantika.lt/SyntaticAndSemanticAnalysis/Analysis) ir Lietuvos Respublikos Seimo Teisės aktų registre (prieiga internete: www.e-tar.lt).

Gramatikos skaitmeninimo temą tęsė Daiva Šveikauskienė (LKI) ir Vytautas Šveikauskas (LKI). Pranešime „Lietuvių kalbos skaitmeninė gramatika“ kalbėta apie Lietuvių kalbos institute pradėtą kurti lietuvių kalbos skaitmeninę gramatiką, kuri gali būti panaudota ir kitose kalbos kompiuterinio apdorojimo srityse: gramatinei analizei, tiesioginiam duomenų vertimui ir kt. Pristatytas projektas, kurio metu planuojama prisijungti prie tarptautinės sistemos DIGITAL GRAMMARS, apimančios šiuo metu 32 kalbas. Pagrindinis skaitmeninės gramatikos sukūrimo tikslas yra panaudoti ją nestatistiniais metodais veikiančiose automatinio vertimo sistemose. Lietuvių kalbai tai labai aktualu. 2017 metų *Tildės* duomenimis, *Google* vertimų kokybė į lietuvių kalbą ir iš lietuvių kalbos yra prastesnė net nei latvių ar estų kalboms. Todėl Lietuvai labai svarbu kurti alternatyvų automatinio vertimo variantą. Šiuo metu yra parengtas bandomasis skaitmeninės gramatikos pavyzdys, teapimantis keletą žodžių, tačiau ir iš jo galima matyti, kad čia vertimo kokybė daug geresnė ypač tiems sakiniams, kuriuose atsispindi specifiniai lietuvių kalbos bruožai. Kol naudojama anglų kalbos žodžių tvarka, neblogus vertimus pateikia ir statistiniai

metodai, tačiau jei sakinyje nestandartinis žodžių išsidėstymas, *Google* vertimas sakinio prasmės neperduoda. Skaitmeninė gramatika gali būti panaudota ir kitose kalbos kompiuterinio apdorojimo srityse: gramatinei analizei, tiesioginiam duomenų vertimui ir kt. Skaitmeninių gramatikų kūrimui vadovauja Geteborgo (Švedija) universiteto profesorius Aarne Ranta.

Paskutinį konferencijos pranešimą „Kas nuo ko nusirašė? Biblijos vertimų istorijos tyrimo automatizavimas ir vizualizacija“ skaitė MINDAUGAS ŠINKŪNAS (LKI). Pranešime kalbėta apie Biblijos vertimų istorijos tyrimų automatizavimą ir vizualizaciją. Biblinių eilučių palyginimo rezultatus pranešėjas parodė vaizdžiose ir gausiose diagramose. Biblijos citatų gausa senuosiuose lietuvių raštuose apsunkina jų ryšių tyrimus. Reikalinga patogi duomenų platforma, leidžianti biblines eilutes grupuoti pagal tyrėją dominančius požymius. Daugiausia sunkumų kelia citatų panašumo vertinimas. Kompleksinė leksikos, sintaksės ir morfologijos (o tam tikrais atvejais – ir rašybos) analizė tai leidžia padaryti, tačiau šitoks filologinis tyrimas yra lėtas, o jo automatizuoti šiuo metu neįmanoma. Palyginimą pavyko automatizuoti ciklais veikiančiu algoritmu, kurio uždavinys – dviejose ženklų sekose aptìkti identiškus junginius ir apskaičiuoti, kurią lyginamų sekų dalį jie apima. Palyginimą automatizuoti kliudo nevienareikšmė senųjų raštų ortografija. Įvertinus neautomatiškai ir automatiškai transliteruotas eilutes, nustatyta, kad transliteracijos būdas esminės įtakos rezultatams neturi. Rezultatai gali būti ginčytini lyginant skirtingomis tarmėmis parašytus tekstus (dialektologinių ypatybių niveliacija nebuvo išbandyta). Gautì panašumo rezultatai apibendrinami dvimatėse diagramose. Programinės analizės metu gauti Bretkūno *Postilės* (1591) biblinių eilučių palyginimo rezultatai atitinka ankstesnių tyrėjų kitais metodais prieitas išvadas.

Konferencija baigėsi diskusijomis. Konferencijos pranešimų pagrindu parengti straipsniai bus spausdinami mokslo žurnaluose *Bendrinė kalba* (prieiga internete: www.bendrinekalba.lt) ir *Lietuvių kalba* (www.lietuviukalba.lt).

LITERATŪRA

Zinkevičius Zigmas 1992: *Lietuvių kalbos istorija* 5. *Bendrinės kalbos iškilimas*. Vilnius: Mokslas, 144–155.

Įteikta 2017 m. lapkričio 3 d.

AURELIJA TAMULIONIENĖ
Lietuvių kalbos institutas
Petro Vileišio g. 5, LT-10308 Vilnius, Lietuva
aurelija.tamulioniene@lki.lt