SKAITMENINIŲ IŠTEKLIŲ IR TARPDISCIPLININIŲ TYRIMŲ INSTITUTO PRIEŠISTORĖ

ANDRIUS UTKA, AGNĖ BIELINSKIENĖ, DARIUS AMILEVIČIUS, ERIKA RIMKUTĖ, INETA DABAŠINSKIENĖ, JOLANTA KOVALEVSKAITĖ, JURGITA VAIČENONIENĖ, LAURA KAMANDULYTĖ-MERFELDIENĖ, VIRGINIJUS DADURKEVIČIUS

Vytauto Didžiojo universitetas

ISSN 1392-0588 (spausdintas)
ISSN 2335-8769 (internetinis)
https://doi.org/10.7220/2335-8769.78
2022. 78

2021 metų lapkričio 24 dieną Vytauto Didžiojo universitete (VDU) įkurtas Skaitmeninių išteklių ir tarpdisciplininių tyrimų institutas (SITTI). Šis faktas žymi nemenką nueitą kelią nuo kelių gerų idėjų iki tarpdisciplininės mokslinės institucijos atsiradimo. SITTI yra trijų VDU institucijų susiliejimas: Kompiuterinės lingvistikos centro, Tarpkultūrinės komunikacijos ir daugiakalbystės tyrimų centro bei Intelektualiųjų technologijų laboratorijos. Šiame straipsnyje apžvelgsime svarbiausius šių institucijų nueito kelio ženklus, vingius ir posūkius.

KOMPIUTERINĖS LINGVISTIKOS CENTRAS

Kompiuterinės lingvistikos centras (KLC) kūrėsi 1992–1995 metais: Stokholmo universitetas 1992 metais padovanojo pirmą kompiuterį, buvo pradėti rinkti lietuvių kalbos tekstai; 1994 metais VDU senatas patvirtino KLC nuostatus ir KLC gavo patalpas (Donelaičio g. 52). Visi šie įvykiai buvo paremti Rūtos Petrauskaitės (ankstesnė pavardė Marcinkevičienė) idėja – sukurti dabartinės lietuvių kalbos tekstyną. Rūta Petrauskaitė ir tapo pirmąja centro vadove, jam vadovavo 16 metų – iki 2011-ųjų. Nuo tų metų centrui pradėjo vadovauti Andrius Utka.

Ilgą laiką KLC vizitinė kortelė buvo (ir vis dar yra) *Dabartinės lietuvių kalbos tekstynas*. Tekstai šiam ištekliui pradėti rinkti nuo 1992 metų. Valstybinis mokslo ir studijų fondas 1997 metais skyrė paramą Dabartinės lietuvių kalbos tekstynui rengti. 2002 metais pasirodė pirmoji tekstyno sąsajos <http://donelaitis.vdu.lt> versija. Ji veikė ilgiau nei 10 metų, o 2011-ųjų pradžioje pristatyta antroji iki šiol naudojama tekstyno sąsajos versija internete <http://tekstynas.vdu.lt>.

Visiems interneto vartotojams prieinama Dabartinės lietuvių kalbos tekstyno sąsaja buvo svarbus įvykis, todėl šis ir kiti KLC sukurti ištekliai, kompiuterinės

analizės programos 2001–2003 metais pristatytos parodose „InfoBalt“ (2001, 2002, 2003, 2005, 2006), „InfoKaunas“ (2002), „Mokslas ir universitetinės studijos“ (2002). 2002 metų parodoje „InfoBalt“ pristatytas projektas Dabartinės lietuvių kalbos tekstynas internete buvo pripažintas geriausiu mokslo darbu ir laimėjo geriausio lietuviško informacinių technologijų, telekomunikacijų ir elektronikos produkto vardą.

KLC parengti ir kiti svarbūs tekstynai: lygiagretieji čekų–lietuvių, lietuvių–čekų, anglų–lietuvių, lietuvių–anglų kalbų porų tekstynai. 2005 metais sukurta internetinė šių lygiagrečiųjų tekstynų sąsaja. 2011–2012 metais parengtas lietuvių–latvių ir latvių–lietuvių kalbų lygiagretusis tekstynas LILA. 2021–2022 metais sukurti anglų–lietuvių kalbų lygiagretusis ir palyginamasis kibernetinio saugumo tekstynai.

Paminėtini ir kiti ištekliai bei tyrimai, į kuriuos įsitraukė KLC mokslininkai: 1993–1995 metais KLC dalyvavo rengiant *Lietuvių kalbos dažninį žodyną*. 1996–2003 metais į lietuvių kalbą išverstas COBUILD žodynas. 1997 metais paruošti lietuviški tekstai (George'o Orwello „1984“ ir Platono „Valstybė“) TELRI kompaktinei plokštelei „Rytai susitinka su Vakarais: daugiakalbių resursų kompendiumas“. 1997 metais KLC parengė tekstynų lingvistikos terminų žodyną anglų kalba. 1997–1999 metais buvo parengta Šv. Rašto vertimo į lietuvių kalbą konkordancija. 1995 metais Vytautas Zinkevičius sukūrė automatinės morfologinės analizės programos „Lemuoklis“ prototipą, o ilgainiui buvo sukurta ir internetinė jo versija. Morfologinės analizės įrankiai buvo toliau tobulinami: 2012–2015 metais sukurtas *Semantika.lt* morfologinis analizatorius.

2007 metais buvo sėkmingai sukurta pirmoji vieša automatinio anglų–lietuvių kalbų mašininio vertimo paslauga. Be KLC darbuotojų, šiame projekte dar dalyvavo Sankt Peterburgo PROMT, UAB „Alna“ ir UAB „Fotonija“ lingvistai bei IT inžinieriai. To meto spaudoje šis darbas buvo įvardytas kaip „didžiausias humanitarinis projektas“. Mašininio vertimo sistema buvo paremta šimtais taisyklių ir kruopščiai sudarytu daugiau kaip 50 000 lemų anglų–lietuvių kalbų žodynu. Neįtikėtina, bet paslauga vis dar veikia ir yra naudojama (<http://vertimas.vdu.lt>).

„Lemuoklis“ panaudotas rengiant morfologiškai anotuotą lietuvių kalbos tekstyną MATAS. Jis rengtas 2000–2005 metais. Iš pradžių šio tekstyno dydis buvo 1 milijonas žodžių, vėliau papildytas, parengtos dvi vartotojams prieinamos versijos (pirmoji <https://clarin.vdu.lt/xmlui/handle/20.500.11821/9>, versija <https://clarin.vdu.lt/xmlui/handle/20.500.11821/33>). Šie tekstynai sudarė galimybes analizuoti lietuvių kalbos morfologinį daugiareikšmiškumą, gramatinių formų ir gramatinių kategorijų pasiskirstymą, rengti dažninius sąrašus, kitais kalbos lygmenimis anotuotus tekstynus.

2016 metais atverta automatiškai morfologiškai anotuota *Dabartinės lietuvių kalbos tekstyno* versija <http://corpus.vdu.lt/lt/>.

2015 ir 2019 metais dviem etapais kurtas automatiškai sintaksiškai anotuotas tekstynas ALKSNIS. Tai vienas pirmųjų automatinės sintaksinės analizės pavyzdžių Lietuvoje, kuris naudojamas kaip aukso standartas tolesnei sintaksinei analizei, pritaikomai kalbos technologijoms kurti. Naujausią tekstyno versiją sudaro 3643 sakiniai (60 196 žodžiai) iš įvairių žanrų tekstų. Tekstynas prieinamas CLARIN-LT saugykloje (<https://clarin.vdu.lt/xmlui/handle/20.500.11821/10 ir https://clarin.vdu.lt/xmlui/handle/20.500.11821/21>), taip pat ir portale <https://kalbu.vdu.lt/>, kur galima atlikti paiešką. Šiam tekstynui parengti reikalingą lietuvių kalbos sintaksinį analizatorių 2012–2015 metais parengė Loïc Boizou ir Francesco Zamblera.

Dabartinės lietuvių kalbos tekstynas suteikė galimybių analizuoti ir pastoviuosius žodžių junginius. Jiems KLC tyrėjai skyrė nemažą dėmesį nuo pat centro įsteigimo pradžios. Kaip svarbiausi paminėtini šie pastoviesiems žodžių junginiams skirti darbai: *Lietuvių kalbos daiktavardinių frazių žodynas* (sudarytas 2012 metais, jo duomenų bazė randama <https://klc.vdu.lt/fraziu-zodynas/>); *Švietimo ir mokslo terminų žodynas* (sudarytas 2010–2012 metais, <http://daukantas.vdu.lt/moksliniai-terminai/>); *Lietuvių kalbos pastoviųjų žodžių junginių duomenų bazė* (sudaryta 2016–2022 metais, <https://resursai.pastovu.vdu.lt/paieska/paprastoji>); *Lietuvių kalbos kolokacijų žodynas* (sudarytas 2019 metais, <https://pastovu.vdu.lt/wp-content/uploads/2021/11/zodynas.pdf>); *Lietuvių kalbos arbitraliųjų kolokacijų žodynas* (sudarytas 2022 metais, <https://arka.pastovu.vdu.lt/zodynas/>); *Lietuvių–anglų kibernetinio saugumo terminų bazė* (sudaryta 2022 metais, <https://klc.vdu.lt/dvitas/lt>).

Maždaug nuo 2017 metų KLC tyrėjai įsitraukė į lietuvių kalbos kaip svetimosios tyrimus. Įgyvendinant projektą „Užsienio baltistikos centrų ir Lietuvos mokslo ir studijų institucijų bendradarbiavimo skatinimas“, sudarytas *Mokomasis lietuvių kalbos tekstynas* (<https://kalbu.vdu.lt/mokymosi-priemones/mokomasis-tekstynas/>), *Mokomasis lietuvių kalbos vartosenos leksikonas* (<https://kalbu.vdu.lt/mokymosi-priemones/leksikonas/>).

KLC mokslininkai nuo 2004 metų apgynė 9 disertacijas tekstynų ir kompiuterinės lingvistikos tematika. 2002 metais Rūtai Petrauskaitei suteiktas habilituoto mokslų daktaro laipsnis už tyrimus „Tekstynų lingvistika ir lietuvių kalbos vartosena“.

KLC (su rėmėjais ir partneriais) aktyviai dalyvavo organizuojant tarptautinius renginius: 1997 metų pavasarį KLC drauge su TELRI projektu organizavo II tarptautinį seminarą „Kalbų technologijos daugiakalbėje Europoje“; 2007, 2014 ir 2020 metais surengė tarptautines konferencijas „The Baltic Conference on Human Language Technologies“. 2017 metais surengtos CLARIN-PLUS tarptautinės dirbtuvės – „Creation and Use of Social Media Resources“.

Be to, KLC rengė ir įvairius seminarus, skirtus kalbos technologijoms pristatyti ir populiarinti: 2001 metais surengė seminarą „Vytauto Didžiojo universitetas – informacinei visuomenei“; 2006 metais – seminarą „Kompiuterinės lingvistikos centro darbai: lygiagretusis tekstynas“; 2009 metais kartu su VDU Lietuvių kalbos katedra surengė konferenciją „Lietuvių kalba ir naujosios technologijos: galimybės ir problemos“. Nuo 2013 metų kartą per mėnesį pradėti rengti „KLC penktadieniai“. Šiais neformaliais susitikimais siekiama diskutuoti aktualiomis tekstynų ir kompiuterinės lingvistikos temomis ir rasti tyrėjų, atliekančių šios srities tyrimus. Seminaruose siekiama populiarinti šias mokslo sritis pritraukiant bendraminčių, diskutuoti aktualiomis temomis neformaliai.

KLC tyrėjai per visą centro gyvavimo laikotarpį parengė daugiau nei 400 mokslinių publikacijų, kelias mokymosi priemones, 2 teminius „Darbų ir dienų“ numerius (skirtus tekstynų lingvistikai (2000, nr. 24) ir vertimo tyrimams (2006, nr. 45)). Paminėtinos dvi monografijos: Rūta Petrauskaitė „Lietuvių kalbos kolokacijos“ (2010); Agnė Bielinskienė, Loïc Boizou, Gintarė Grigonytė, Jolanta Kovalevskaitė, Erika Rimkutė ir Andrius Utka „Lietuvių kalbos terminų automatinis atpažinimas ir apibrėžimas“ (2015).

KLC įvykdė apie 30 tarptautinių ir nacionalinių projektų, iš jų paminėtini trys didžiausi projektai, finansuoti iš Europos Sąjungos struktūrinių fondų: „Internetinė informacijos vertimo priemonė“ (2005–2007), „Semantinis informacijos valdymo variklis“ (Semantika-1, 2010–2012) ir „Lietuvių kalbos sintaksinės-semantinės analizės sistema tekstynui, lietuviškam internetui ir viešojo sektoriaus taikymams“ (Semantika-2, 2018–2020).

Nuo 2008 iki 2022 metų KLC ir Informatikos fakulteto panašius tyrimus vykdantys mokslininkai buvo susibūrę į mokslinių tyrimų klasterį „Teksto ir balso skaitmeniniai tyrimai, išteklių ir technologijų kūrimas bei taikymas“ (vėliau pervadintas į „Skaitmeniniai kalbos duomenys ir intelektinės technologijos“).

KLC buvo ir aktyvus tarptautinių organizacijų narys. 1997 metais KLC tapo tarptautinės TELRI asociacijos nariu. 2015 metais Lietuva įsitraukė į Europos mokslinių tyrimų infrastruktūrą CLARIN ERIC (angl. *Common Language Resources and Technology Infrastructure*), Europos ir kitų pasaulio šalių socialinių ir humanitarinių mokslų tyrėjams užtikrinančią atvirą ir patikimą skaitmeninių kalbos išteklių ir jų analizės įrankių sklaidą bei prieigą vienoje platformoje, bendrus kalbos išteklių kūrimo standartus ir ilgalaikį duomenų saugojimą. Įkurtą CLARIN-LT konsorciumą sudarė trys institucijos partnerės: Vytauto Didžiojo universitetas (koordinuojanti institucija), Vilniaus universitetas ir Kauno technologijos universitetas. 2015–2016 metais konsorciumas gavo Švietimo, mokslo ir sporto ministerijos finansavimą; tai padėjo sukurti reikiamą infrastruktūrą ir vykdyti pagrindines veiklas: išlaikyti ir administruoti CLARIN-LT centrą, plėtoti ir atnaujinti lietuvių

kalbos išteklius bei analizės įrankius, atvirai prieinamus sertifikuotoje saugykloje, teikti konsultacijų ir kompetencijų tobulinimo paslaugas tyrėjams. 2016 metais prof. habil. dr. Rūta Petrauskaitė išrinkta CLARIN ERIC generalinės asamblėjos pirmininke dvejų metų laikotarpiui, o nacionalinės koordinatorės pareigas perėmė doc. dr. Jurgita Vaičenonienė. 2020 metais prie konsorciumo prisijungė dvi naujos institucijos: Mykolo Romerio universitetas ir Baltijos pažangių technologijų institutas, o 2021 metais – dar ir Klaipėdos universiteto Baltijos regiono archeologijos ir istorijos institutas. Kitas svarbus pasiekimas – 2020 metais CLARIN-LT tapo Helsinkio universiteto koordinuojamo CLARIN žinių centro morfologiškai turtingų kalbų sistemų ir sistemų (SAFMORIL) nare. Infrastruktūros brandą pripažino Lietuvos mokslo taryba, ji rekomendavo CLARIN-LT įtraukti į tuo metu atnaujinamą Lietuvos mokslinių tyrimų infrastruktūrų 2020–2023 metų planą. Šiuo metu CLARIN-LT funkcionuoja kaip viena brandžiausių humanitarinių mokslų tarptautinį statusą turinčių infrastruktūrų Lietuvoje, siekiama įgyti paslaugų teikimo centro (*CLARIN B centre*) sertifikatą, jis ypač svarbus norint visiškai integruotis į šią tarptautinę infrastruktūrą.

TARPKULTŪRINĖS KOMUNIKACIJOS IR DAUGIAKALBYSTĖS TYRIMŲ CENTRAS

VDU Tarpkultūrinės komunikacijos ir daugiakalbystės tyrimų centro ištakos susijusios su vaikų kalbos tyrimais. 1994 metais, dalyvaujant Vienos universiteto prof. Wolfgango U. Dresslerio vadovaujamame tarptautiniame projekte ,,The Acquisition of Pre- and Protomorphology“, skirtame vaikų kalbos įsisavinimo tyrimams, buvo pradėti kaupti lietuvių vaikų kalbos duomenys. Ineta Dabašinskienė (anksčiau – Savickienė), Pawełas Wójcikas ir Magdalena Smoczyńska sukaupė ir suskaitmenino dviejų vaikų kalbos įrašus bei, naudodami kompiuterinę programą CHILDES (angl. *Child Language Data Excange System* <http://childes.psy.cmu.edu/>, MacWhinney 2000), sukūrė transkribuotą ir gramatiškai anotuotą šių vaikų kalbos duomenų bazę. Minimą duomenų bazę galima laikyti pirmuoju sakytinės lietuvių kalbos tekstynu, kuriuo remiantis atlikti tyrimai leido susiformuoti lietuviškajai psicholingvistikos mokyklai, taip pat atvėrė erdvę sakytinės kalbos tyrimams. Dabašinskienės ir kolegų pradėta kurti lietuvių vaikų kalbos duomenų bazė plečiama iki šiol. Šiuo metu VDU kaupiamą vaikų kalbos tekstyną sudaro tekstine forma transkribuoti ir morfologiškai anotuoti 7 vaikų kalbos įrašai, surinkti ilgalaikio stebėjimo metodu ir apimantys vaikų kalbos duomenis nuo gramatinės sistemos formavimosi pradžios iki jos įsitvirtinimo, vidutiniškai nuo 1;7 metų iki 3;5 metų. Tai vienas didžiausių vaikų kalbos tekstynų Europoje.

Vaikų kalbos tyrimai peraugo į platesnius sakytinės kalbos tyrimus, apimančius spontaninę šnekamąją kalbą ir parengtą viešąjį kalbėjimą. Siekiant sukaupti natūralios sakytinės suaugusiųjų kalbos duomenis, 2006 metais VDU buvo pradėtas kurti *Sakytinės lietuvių kalbos tekstynas* (<http://sakytinistekstynas.vdu.lt/>), kurį šiuo metu sudaro apie 300 000 morfologiškai anotuotų žodžių formų. Vykdant įvairius Lietuvos mokslo tarybos finansuojamus projektus sakytinės kalbos duomenys kaupti įvairiuose Lietuvos regionuose. Transkribavimui ir gramatiniam sakytinės kalbos duomenų anotavimui pritaikyta programa CHILDES, kuria anotuojant tekstus remiamasi specialia forma pateikiamu leksikonu. Sakytinės kalbos anotavimui naudojamą leksikoną parengė Ineta Dabašinskienė, Andrius Utka, Laura Kamandulytė-Merfeldienė (plačiau žr.: Kamandulytė-Merfeldienė, Laura. Sakytinės lietuvių kalbos tekstynas – natūralios vartosenos tyrimų šaltinis. *Taikomoji kalbotyra*, 2017, nr. 9), jis nuolat papildomais naujais žodžiais ir jų formomis.

2008 metais Tarpkultūrinės komunikacijos ir daugiakalbystės centro mokslininkai kartu su kolegomis įsitraukė į klasterį „Daugiakultūriškumo ir kalbos kaitos tyrimai globalizacijos kontekste“. Klasterio mokslininkai analizavo kalbos vartojimo ypatybes, kalbos sistemos pokyčius, susijusius su globalizacija, kalbos raidos ypatumus daugiakultūrėje ar daugiakalbėje visuomenėje, tyrė klausimus, susijusius su kalbos politika. Vykdant įvairius projektus buvo ne tik plečiami anksčiau aprašyti tekstynai, bet ir sukurta kalbos raidos diagnostikos priemonių, eksperimentinių kalbos įsisavinimo tyrimo testų, parengta standartizuojama kalbos raidos vertinimo metodika ir tam skirta priemonė.

Tarpkultūrinės komunikacijos ir daugiakalbystės centro mokslininkai yra koordinavę ir vykdę daugiau kaip 20 tarptautinių ir nacionalinių mokslo projektų, skirtų sakytinės suaugusiųjų kalbos, vaikų kalbos, daugiakalbystės, lingvistinio kraštovaizdžio, kalbos tapatybės tyrimams. Iš jų vertėtų išskirti projektą „Friendly Resources for Playful Speech Therapy – Frepy“, apdovanotą Švietimo mainų paramos fondo Kokybės statulėle, taip pat Europos Komisijos – Europos kalbų ženklu. Paminėtina, kad šio projekto metu sukurta priemonė tebėra viena dažniausiai naudojamų kalbos terapijos priemonių logopedų praktinėje veikloje.

Tarpkultūrinės komunikacijos ir daugiakalbystės centro temomis apgintos 8 mokslo disertacijos, parengta daugiau kaip 200 mokslo publikacijų. Šio centro mokslininkai nuolat teikia ekspertines konsultacijas valstybinėms ir kitoms suinteresuotoms institucijoms (Švietimo, mokslo ir sporto ministerijai, Nacionalinei švietimo agentūrai, Lietuvos logopedų asociacijai, Lietuvos Respublikos specialiųjų pedagogų asociacijai ir t. t.).

INTELEKTUALIŲJŲ SISTEMŲ LABORATORIJA

2018–2020 metais, gavus Europos Sąjungos struktūrinių fondų ir Lietuvos biudžeto finansavimą projektui „Semantika-2“, susidarė sąlygos VDU vykdomuose fundamentiniuose ir taikomuosiuose kalbos technologijų tyrimuose atlikti kokybinį šuolį ir pažangiausias giliojo mokymosi technologijas pradėti taikyti visa apimtimi kuriant sprendimus lietuvių kalbai.

Pasaulyje natūraliosios kalbos technologijose nuo 2010 metų prasidėjo vadinamoji neuroninė revoliucija, kai giliojo mokymosi ir įterptinių žodžių metodų taikymo rezultatai parodė, kad tokiais metodais galima pasiekti puikių rezultatų sprendžiant daugelį natūraliųjų kalbų technologijų uždavinių (kalbos modeliavimo, šnekos atpažinimo bei šnekos sintezavimo, analizės ir daugelį kitų). Nors VDU Informatikos fakultete nuo seno buvo dėstomi mašininio ir giliojo mokymosi dalykai, tik vykdant projektą „Semantika-2“ susidarė tinkama tarpdisciplininė terpė: projektą vykdė Humanitarinių mokslų, Informatikos ir Teisės fakultetų tyrėjai. Jie turimą įdirbį bei patirtį papildė naujomis giliojo mokymosi technologijų patirtimis ir kompetencijomis siekdami projekte numatytų, kalbos technologijomis grįstų tarpdisciplininių tikslų įgyvendinimo (medicinoje, teisėje ir kt.). 2020 metais Lietuvos verslo konfederacija, įvertinusi atliktą proveržį, VDU už sukurtą giliojo mokymosi technologijomis grįstą šnekos atpažinimo sprendimą skyrė prestižinį „Metų mokslo paslauga verslui 2020“ apdovanojimą.

Vykdant „Semantika-2“ projektą, keli Informatikos fakulteto tyrėjai susibūrė ir 2018 metų pabaigoje įkūrė Intelektualiųjų sistemų laboratoriją, jai vadovavo vienas iš Lietuvos dirbtinio intelekto strategijos kūrėjų Darius Amilevičius. 2019 metais laboratorija tapo Europos dirbtinio intelekto tyrimų laboratorijų konfederacijos (CLAIRE) nare. Nuo įkūrimo pradžios iki 2021 metų, kai laboratorija ir jos tyrėjų komanda buvo įtraukti į SITTI sudėtį, laboratorijoje buvo kuriami intelektualiomis technologijomis grįsti šnekos atpažinimo, šnekos sintezės, automatinės tekstų analizės, muzikos signalų analizės ir kiti sprendimai. Buvo sėkmingai įgyvendinti du MTEP projektai: „Astra“ ir „Muzika“.