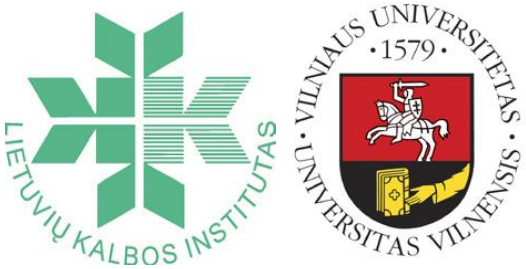

LIETUVIŲ KALBOS INSTITUTO BENDRINĖS KALBOS TYRIMŲ CENTRAS

VILNIAUS UNIVERSITETO TAIKOMOSIOS KALBOTYROS INSTITUTO
LIETUVIŲ KALBOS KATEDRA

24-oji mokslinė Jono Jablonskio konferencija

SKAITMENINIAI KALBOS IŠTEKLIAI, JŲ PLĖTROS KRYPTYS IR PANAUDOS GALIMYBĖS

PRANEŠIMŲ TEZĖS

2017 m. rugsėjo 29 d.

Lietuvių kalbos institutas

Sudarė

AGNĖ ALEKSAITĖ

Spaudai rengė

RITA MILIŪNAITĖ

AURELIJA TAMULIONIENĖ

Mokslo žurnalas „Bendrinė kalba“ (http://www.bendrinekalba.lt/)

Mokslo žurnalas „Lietuvių kalba“ (http://www.lietuviukalba.lt/)

TURINYS

LAIMUTIS BILKIS

Lietuvių kalbos institutas

LIETUVOS VIETOVARDŽIŲ GEOINFORMACINĖS DUOMENŲ BAZĖS STRUKTŪRA, RYŠIAI SU KITOMIS VARDYNO BAZĖMIS IR PLĖTROS KRYPTYS

Lietuvos vietovardžių geoinformacinė duomenų bazė šiuo metu integruota į *Lietuvių kalbos išteklių informacinę sistemą*, jos adresas http://lkiis.lki.lt/lietuvos-vietovardziu-geoinformacine-duomenu-baze. Bazė kuriama tiek vietovardžių mokslinių tyrimų, tiek jų išsaugojimo, sklaidos, taisyklingos vartosenos mokymo tikslais. Jos idėja – sujungti lingvistinius ir geografinius duomenis, pateikti kalbinę informaciją (analizę) apie vietovardžius ir geografinę informaciją apie jais įvardijamus geografinius objektus.

Pagrindiniai vietovardžių bei jais įvardijamų objektų šaltiniai yra: Nacionalinės žemės tarnybos prie ŽŪM ir Valstybės įmonės Registrų centro kuriamos ir administruojamos duomenų bazės (skaitmeniniai žemėlapiai), Lietuvos žemės vardyno anketos (1935–1939 m.) (saugomos Lietuvių kalbos institute, toliau – LKI), Pavardžių ir vietovardžių komisijos vietovardžių kartoteka ir gyvenamųjų vietų vardų bylos (1935–1937 m.) (saugomos LKI), Vilniaus srities gyvenamųjų vietų vardų sąrašas (1940–1945 m.) (saugomas LKI), pokariu sudarytos vietovardžių anketos (kai nėra ar trūksta tarpukario duomenų).

Duomenų bazės viešos prieigos puslapyje sukurtos tókios informacijos apie įvardijamuosius objektus paieškõs sritys: objekto tipas, objekto statusas, dabartinė administracinė teritorinė priklausomybė (savivaldybė, seniūnija, gyvenvietė), tarpukario administracinė teritorinė priklausomybė (apskritis, valsčius, gyvenvietė), upių intakai.

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.

Vietovardžių analizės srityje dar nurodomi objektų fiziniai dydžiai, gyventojų skaičius (gyvenamųjų vietų), objektų nuotraukos (esant galimybei), pateikiamas vietovardžių visų

linksnių kirčiavimas (gyvenamųjų vietų, upių ir ežerų vardų pridedamas ir garsinis linksnių tarimas), kilmės, darybos aiškinimai, pirmojo paminėjimo data ir informacija apie vietovardžių užrašytojus bei pateikėjus.

Bazę integruojant į *Lietuvių kalbos išteklių informacinę 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, priklausančių 11-os savivaldybių (Kalvarijos, Kazlų Rūdos, Marijampolės, Pagėgių, Druskininkų, Birštono, Rietavo, Švenčionių, Ignalinos, Lazdijų, dalies Alytaus) teritorijoms, be to, išanalizuoti visų miestų, miestelių bei seniūnijų centrų vardai ir didesnių nei 50 ha ežerų bei ilgesnių nei 50 km upių onimai.

Duomenų bazės kūrimo sunkumai susiję su turimų vietovardžių gausa ir finansavimo fragmentiškumu.

Ateityje galimos vietovardžių sąsajos su bendrine (tarmine) leksika, ypač semantiniu (motyvaciniu), darybiniu ir lokaliniu požiūriu.

VIRGINIJUS DADURKEVIČIUS
Vilniaus universiteto Taikomųjų mokslų institutas

LIETUVIŲ KALBOS GRAMATIKA SKAITMENINIAME ATVIROJO KODO PASAULYJE

Nenuilstamomis Jono Jablonskio pastangomis jau maždaug 100 metų, kaip lietuvių kalba yra normalizuota ir pritaikyta būti pagrindiniu valstybę vienijančiu instrumentu. Atsiradus civilizaciją keičiančiai naujovei – kompiuteriams – klasikinei gramatikai ir leksikai iškilo būtinybė „apsivilkti naują rūbą“ ir tapti visaverte naujųjų technologijų dalimi. Kalbos dalykai privalėjo būti smulkmeniškai formalizuoti ir užrašyti kitokiu, kompiuteriams suvokiamu, būdu. Ir visų pirma – morfologija, nes tai yra pagrindas bet kokiems gilesniems kompiuterinės lingvistikos darbams (sintaksė, semantinė analizė ir pan.).

Pirmą kartą tokio tipo darbą praeito amžiaus pabaigoje atliko eksperimentinės gamyklos „Bitas“ programuotojas Vytautas Zinkevičius. Beveik du dešimtmečius tai buvo nepamainoma ir vienintelė reali priemonė moderniam lietuvių kalbos tyrinėjimui ir kompiuteriniams taikymams. Tačiau, plečiantis kompiuterinės lingvistikos poreikiams, pradėjo ryškėti šios sistemos trūkumai: nestandartinės, neaprašytos duomenų struktūros, uždaras kodas, nepilnavertis tikrinių vardų, iliatyvo, dviskaitos, sutrumpėjusių bei retų formų realizavimas. Ypač šios sistemos tolesnei plėtrai trukdė uždaras kodas – duomenis ir algoritmą praktiškai galėjo keisti tik pats autorius.

Siekiant išvengti minėtų trūkumų ir sukurti naujos kartos kompiuterinę lietuvių kalbos morfologiją, buvo 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. Visus šiuos reikalavimus gerai atitiko *Hunspell* (https://en.wikipedia.org/wiki/Hunspell) platforma. Be to, šios platformos pagrindu nesunkiai galima turėti rašybos tikrinimą daugelyje (virš 50) taikomųjų programų (*OpenOffice*, *Firefox*, *Chrome*, *Safari*, *InDesign* ir t. t.), atlikinėti žodžių morfologinę analizę bei sintezę, vykdyti intelektualią tekstinės informacijos paiešką ir pan.

Kalbos morfologiją *Hunspell* platformoje aprašo du tekstiniai failai. Viename iš jų (plėtiniu .aff) yra užrašomos kaitybos taisyklės, o kitame (plėtiniu .dic) – žodžių formos su pradine morfologine informacija ir nuorodomis į vieną ar kelias kaitybos taisykles, kurios papildomai gali būti taikomos. Abiejų šių failų sudarymo pagrindas – *Dabartinės lietuvių kalbos gramatika* (Vilnius, 2006) ir VU, VDU bei LRS tekstynai (iš viso apie 1,5 mlrd. žodžių). Buvo užrašyta maždaug 5 000 adresuojamų taisyklių grupių ir atrinkta virš 170 000 lemų. Vykdant lemų atranką buvo apsiribojama tiktai taisyklinga šiuolaikine lietuvių kalba ir vengiama klaidingų, nelietuviškų, neteiktinų, įžeidžiančių, nebevartojamų žodžių. Neaiškiais atvejais buvo tariamasi su VLKK specialistais.

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 VDU Sintaksinės-semantinės analizės sistemoje (https://semantika.lt/SyntaticAndSemanticAnalysis/Analysis) bei LRS Teisės aktų registre (www.e-tar.lt). Ypač sėkmingas taikymas buvo *Solr*/*Lucene* dokumentų indeksavimo ir paieškos sistemoje, kur pavyko ne tik tiksliai indeksuoti bei surasti visas norimas žodžių formas, bet ir pagreitinti indeksavimo procesą šimtus kartų. Tokių rezultatų pavyko pasiekti pakeitus morfologijos.aff/.dic aprašą baigtiniu būsenų keitikliu (FST – *Finite State Transducer*).

Toliau plėtojant šiuos darbus, reikėtų parengti bent dvi žodyno versijas: norminę, tinkamą rašybai tikrinti, ir pilną (norminė ir nenorminė kalba), tinkamą taikomiesiems gyvosios kalbos analizės uždaviniams spręsti.

GINTARĖ JUDŽENTYTĖ, VILMA ZUBAITIENĖ
Vilniaus universiteto Filologijos fakultetas

TEKSTYNAIS PAREMTI AKADEMINIŲ FRAZIŲ TYRIMAI: FORMALIOJI STRUKTŪRA IR SEMANTIKA

Akademinė, arba mokslo, kalba Lietuvoje tiriama įvairiais aspektais, pvz., Zita Alaunienė (Alaunienė, 2005) yra aptarusi akademinių tekstų struktūrą ir raišką, Audronės Bitinienės darbuose atkreipiamas dėmesys į sintaksines mokslo tekstų ypatybes (Bitinienė 2000, 2009) bei intertekstualumą (Bitinienė 2005), Saulius Damošius tyrė vertinimo raišką mokslinio stiliaus tekstuose (Damošius 2007). Paremtų tekstynais mokslo kalbos tyrimų Lietuvoje nėra gausu. Galima paminėti Jolantos Šinkūnienės studiją „Lietuviškojo humanitarinių ir socialinių mokslo diskurso ypatybės“ (2014), taip pat gretinamuosius autoriaus pozicijos raiškos ir švelninimo, modalumo, savicitavimo (Šinkūnienė 2010, 2011, 2015; Linkevičienė, Šinkūnienė 2012, Šinkūnienė, Van Olmen 2012, Mur Dueñas, Šinkūnienė 2016) bei adverbializacijos lietuvių mokslo kalboje (Smetona, Usonienė, 2012) tyrimus, kurie atlikti remiantis pačių autorių sudarytais lyginamaisiais tekstynais arba Lietuvių mokslo kalbos tekstynu (http://coralit.lt/). Akademinio žodyno (tiek atskirų žodžių, tiek įprastų tekstą kuriančių frazių) analizės iki šiol nėra atlikta. Labai trūksta tyrimų, gretinančių mokslo straipsnių ir akademinių rašto darbų (esė, kursinių, bakalauro, magistro ir kt.) raiškos ypatumus. Tokių tyrimų užsienyje gausu, plg. Aktas, Cortes 2004, 2008, Biber, Conrad, Cortes 2003, 2004, Gledhill 2000, Granger, Paquot 2009, Hyland 2008, Hyland, Tse 2007, Oakey 2002, Paquot 2010, Salazar 2014.

Šio pranešimo tikslas – aptarti 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. Siekiama nustatyti akademiniam rašymui svarbių žodžių sąrašą (Plg. Coxhead 2000, Nation, Coxhead 2001), išskirti pagrindines kolokacijas ir jų plėtinius, taip pat pasikartojančias žodžių sekas įvairiose (plg. https://pearsonpte.com/wp-content/uploads/2014/07/AcademicCollocationList.pdf, Ackermann, Chen 2013, Durrant 2009, Simpson-Vlach, Ellis 2010) akademinio teksto dalyse, aptarti jų sąsajas su retorinėmis teksto dalių funkcijomis, kitaip retoriniais ėjimais ir žingsniais

(Cortes 2013, Swales 1990, 2004, Šinkūnienė 2014, 53) ir aprašyti reikšminių akademinių žodžių, tokių kaip: *tikslas*, *uždavinys*, *metodai*, *išvados*, *rezultatai*; *atlikti*, *analizuoti*, *nustatyti*, *remtis* ir kt.) vartoseną ir semantiką.

PIJUS KASPARAITIS, GINTARAS SKERSYS

Vilniaus universiteto Matematikos ir informatikos fakultetas

LIETUVIŠKO BALSO SINTEZĖS DABARTIS IR PERSPEKTYVOS

Kalba yra natūraliausias žmogaus ir kompiuterio bendravimo būdas, todėl kompiuterių galimybės kalbėti nuolat tobulinamos. Lietuviško balso sintezėje esminis proveržis įvyko 2015 m., kai vykdant projektą „Lietuvių šneka valdomos paslaugos“ (LIEPA) buvo sukurtas lietuviško balso sintezatorius.

LIEPA sintezatorius yra laisvai platinamas kartu su pradiniais tekstais, tai sudarė sąlygas ir kitiems žmonėms rasti naujus jo pritaikymus:

1. Sintezatoriaus debiutas – tarptautinio teatrų festivalio *Sirenos 2014* spektaklis *Remote Vilnius*, kuriame „vaidino“ sintetiniai *Aistės* ar *Edvardo* balsai.

2. Balso įrašus šalia straipsnelių jau pateikia interneto svetainės: *lzinios.lt*, *ukininkopatarejas.lt*, *vilnius.lt*, *m.delfi.lt*, *vle.lt*.

3. Teksto įgarsinimo paslauga *RoboBraille* leidžia automatiškai paversti bet kokius tekstinius dokumentus į garso failus.

4. Pranešimai sintetiniu balsu keleiviams jau skaitomi Vilkaviškio autobusų stotyje.

5. Lietuviškai prašnekęs humanoidinis robotas *NAO* geba pakviesti užsukti į Lietuvos stendą renginių metu, moko tarti lietuviškus žodžius, deklamuoja K. Donelaičio *Metų* ištrauką.

Potencialūs balso sintezės taikymo pavyzdžiai, kuriems įgyvendinti jau buvo žengti pirmi žingsniai:

1. Straipsnelių įgarsinimas *Linux* pagrindu realizuotose svetainėse.

2. Vystantis dirbtiniam intelektui atsiranda virtualių asistentų, kurie geba atsakyti į žmogaus laisva forma užrašytus klausimus. Asistentai jau veikia *Migracijos departamento* svetainėje, deja, nekalba.

3. Autonominis mobilus pardavimų automatas *MIO mobile kiosk*, skirtas vežioti ir pardavinėti užkandžius bei gėrimus parkuose, skveruose, aikštėse. Naudotojų patogumui jis galėtų prabilti lietuviškai.

4. Kai kurios automatinio vertimo programos moka perskaityti balsu verčiamą arba jau išverstą tekstą, deja, lietuviškai nemoka.

5. Kompiuterius vis dažniau pakeičia išmanieji telefonai, turintys savyje balso sintezatorių ir akliesiems skirtą ekrano skaitytuvą. Deja, prakalbinti lietuviškai įmanoma tik *Android* išmaniuosius telefonus. Per standartinę sąsają sintezatoriumi galėtų naudotis ir kitos programos, pvz., navigacija.

RITA MILIŪNAITĖ
Lietuvių kalbos institutas

PAIEŠKOS GALIMYBĖS INTERNETINIAME *LIETUVIŲ KALBOS NAUJAŽODŽIŲ DUOMENYNE*

Pranešime apibūdinamas dabartinis nuo 2011 m. Lietuvių kalbos institute kuriamo internetinio *Lietuvių kalbos naujažodžių duomenyno* (http://naujazodziai.lki.lt/; toliau – ND) būvis ir plėtros kryptys. ND yra internetinis žinynas, skirtas fiksuoti ir leksikografiškai aprašyti XXI a. pradžios lietuvių kalbos naujažodžius – naujai skolintus arba pasidarytus žodžius bei naujas esamų žodžių reikšmes. 2017 m. rugsėjį ND sudaro daugiau kaip 4200 naujažodžių aprašų, juose pateikiama daugiau kaip 16 400 vartosenos pavyzdžių iš įvairių rašytinių ir sakytinių šaltinių.

Daugiausia dėmesio pranešime skiriama klausimui, kaip kalbos vartotojams atskleisti kuo įvairesnių jų poreikius atitinkančių ND naujažodžių ypatybių. Šis klausimas atrodo paprastas tik iš pirmo žvilgsnio. Kuriant įvairius skaitmeninius išteklius, skaitmeninant žodynus (didįjį *Lietuvių kalbos žodyną*, *Dabartinės lietuvių kalbos žodyną*, *Sinonimų žodyną*, *Frazeologijos žodyną*, rengiamą naują *Bendrinės lietuvių kalbos žodyną* ir kt.), jau esama nemažai patirties. Tačiau ND turi esminių skirtumų.

Minėtuose žodynuose pirminė kalbos vartotojų (toliau – vartotojų) prieiga prie duomenų yra antraštinis žodis. Vartotojams dažniausiai reikia informacijos apie kokį nors jiems reikalingą žinomą žodį: jo rašybą, kirčiavimą, gramatikos ir vartojimo ypatumus. ND atveju vartotojai patenka į mažai pažįstamą leksinę aplinką ir, turėdami kokios nors informacijos apie reikiamą naują žodį (pavyzdžiui, žinodami originalią jo formą arba apytikslę apibrėžtį, arba vien lietuvišką atitikmenį), gali iš anksto nežinoti paties naujažodžio arba tikslaus jo pavidalo. Tai kelia klausimus: kas duomenyne svarbu vartotojams? Nuo ko pradėti paiešką? Kaip pateikti kuo daugiau ND panaudos galimybių?

ND, kaip naujų reiškinių sankaupa, dėsningai sukuria puikią erdvę, traukiančią tiek paprastus, dažnai smalsumo vedamus kalbos vartotojus, tiek švietimo sistemos dalyvius (mokinius ir studentus, mokytojus ir dėstytojus), tiek kitus tekstų kūrėjus ar redaguotojus, taip pat kalbos naujovių tyrėjus. Šiems poreikiams iki 2017 m. rugsėjo ND buvo galima paieška

pagal tokius parametrus: antraštinį žodį, naujažodžio kilmę, originalo formą (jei naujažodis skolintas arba išsiverstas), rašybos variantus, 51-ą vartojimo sritį, kai kuriuos ypatingesnius požymius (pvz., kontaminacinius darinius, išeivių vartojamus naujažodžius ir pan.), taip pat pagal pavartojusio žymesnio žmogaus ar naujadaro autoriaus pavardę. ND duomenų bazės tvarkytojams prieinama ir platesnė paieška, reikalinga duomenims įvairiais pjūviais redaguoti.

ND 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. ND numatoma integruoti į Lietuvių kalbos išteklių informacinę sistemą (LKIIS, http://lkiis.lki.lt/) ir įtraukti ne tik į bendrą paieškos šioje duomenų sankaupoje sistemą, bet ir panaudoti kuriant žodžių tinklus. Todėl esama paieška jau yra per siaura.

Atsižvelgiant į visus šiuos veiksnius ir turint galvoje atskiras tikslines vartotojų grupes, kuriama išplėstinė ND informacijos paieškos sistema. Pranešime parodoma, kuo ND gali būti naudingas tiek kalbos specialistams, tiek visiems naujažodžiais besidomintiems vartotojams.

DAIVA MURMULAITYTĖ
Lietuvių kalbos institutas

NAUJAŽODŽIŲ DARYBOS TYRIMŲ PERSPEKTYVOS (*LIETUVIŲ KALBOS NAUJAŽODŽIŲ DUOMENYNO* ATVEJIS)

Lietuvių kalbos naujažodžių duomenyne (toliau – ND) kaupiami 3 formų ir 5 tipų neologizmai, iš viso – 14 rūšių naujieji leksikos vienetai. Žodžių darybos požiūriu nagrinėtinos 4 rūšys – naujai pasidaryti, skolinti, atgiję ir naujas reikšmes įgiję žodžiai. ND naujažodžių klasifikacija – dėl *naujai skolintos šaknies* tipo – ne visai atitinka įprastą jų skirstymą į skolinius ir darinius, tokią šaknį gali turėti ir skoliniai, ir hibridiniai dariniai. Tai neatsispindi ND duomenų klasifikacijoje, bet yra svarbu tiriant naujažodžių darybą.

Pradiniu ND 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 (tokių tyrimo laikotarpiu rasta 426: *karštė, įsikatinti, lovininkas, -ė, varškėfobija* ir kt.), autoriniai (145: *alitas, demaguogija, fotosofija, išvirkštėnai* ir kt.), kontaminaciniai (143: *irba, irklentė, murmanas*, *siaurovizija*; *glokalus, -i, hakatonas, kospėjus*, *kryminalizacija* ir kt.). Pastaroji informacija svarbi nagrinėjant okazinius darinius, kurie dažnai nepaklūsta žodžių darybos dėsniams.

Išsamiai ir kokybiškai žodžių darybos analizei taip pat svarbu atsižvelgti į kai kuriuos pradžioje nenumatytus fiksuoti dalykus – darinio pamatinio žodžio kalbos dalį (esminę, nustatant žodžio darybos kategoriją), darybos pamato pakitimus, analogijos, vertimo vaidmenį darantis naujažodį, netipinės darybos (*užužužužužpernykštis, -ė, gražulesnis, -ė*) apraiškas ir kt. Šiais aspektais naujažodžiai gali būti žymimi įvairiais indeksais, pagal kuriuos paskui gali būti grupuojami, lyginami ir analizuojami. Naudojant indeksus ND duomenis galima surūšiuoti kitais pjūviais negu buvo numatyta kuriant pirminę duomenų bazės struktūrą. Galima derinti paieškos kryptis, siaurinti paiešką ir pan. Svarbu sukurti gerą – išsamią, lanksčią, patogią naudoti, pildyti bei koreguoti – indeksų sistemą.

Turint galvoje, kad naujažodžiai tirtini ir morfemiškai, bei siekiant smulkiau suklasifikuoti minėto *naujai skolintos šaknies* tipo žodžius, ND duomenys pirmiausia suskirstyti į turinčius tik savas, tik skolintas (išskyrus galūnę, kuri skolinantis yra adaptacinė), mišrias morfemas ir morfemų kilmės atžvilgiu neaiškius ar abejotinus atvejus. Tam naudoti indeksai L (*karulis, krivė, plaukuosena, protmūšis*), S (*felinologija, kobido, okupendumas*), H (*paguglinti, paneuropietiškumas, žalfanis, -ė*), L? (*subumčikinti, babatukas*), S? (*eksas, -ė, delfinizmas*), H? (*alitas*, *baudžiakas, klikt*). Jau indeksuojant pastebėta, kad yra tokių tik skolintas morfemas turinčių žodžių, kurie akivaizdžiai pasidaryti lietuvių kalboje (*sakurozė* 'sakurų psichozė – masiškas fotografavimasis prie žydinčių sakurų (ypač Lietuvos didmiesčiuose)', *kalafiorgeitas* 'skandalas, kurį Lietuvoje žmonės sukėlė dėl itin išaugusių kainų; jo pradžia buvo vieno piliečio pasipiktinimas žiedinių kopūstų kaina [...]'). Atskiro indekso prireikė norint atskirti lietuviškos darybos kontaminacinius darinius nuo skolintų. Pagal jį išsiaiškinta, kad ND tokie žodžiai sudaro apie 35% visų kontaminacinių darinių, dar apie 5% greičiausiai yra vertiniai. Taip pat akivaizdu, kad vertiniai apskritai sudaro nemažą dalį ND fiksuotų naujažodžių, todėl juos (kaip ir netipinės darybos žodžius, analoginius darinius ir kt.) irgi reikia suindeksuoti, atskirai išsirinkti ir ištirti.

Šie ir kiti indeksuojant ryškėjantys naujažodžių darybos reiškiniai yra mažai tirti ar iš viso netirti. ND, kaip įvairią informaciją apie naujažodžius kaupiantis ir pateikiantis šaltinis bei jų tyrimo priemonė, labiau pritaikius jį žodžių darybos ir morfemų analizei, suteikia tam dideles galimybes.

RAFAEL RIVERA
iClaves

LANGUAGE EQUALITY IN THE DIGITAL AGE: TOWARDS A HUMAN LANGUAGE PROJECT

In the digital era, language barriers represent a major challenge preventing European citizens and businesses from fully benefiting from a truly integrated Europe. These barriers particularly affect the less educated and older population, as well as speakers of smaller and minority languages, thus creating a notable language divide. Language barriers have a profound effect on (1) cross-border public services, (2) fostering a common European identity, (3) workers' mobility, and (4) cross-border e-commerce and trade, in the context of a Digital Single Market.

The emergence of new technological approaches such as deep-learning neural networks, based on increased computational power and access to sizeable amounts of data, are making Human Language Technologies (HLT) a real solution to overcoming language barriers. However, several factors, such as market fragmentation, uncoordinated research and insufficient funding, are hindering the European HLT industry, while putting under-resourced languages in danger of digital extinction.

Moreover, language technologies are not properly represented in the agenda of European policy-makers, although they are likely to be crucial for the construction of a fair and truly integrated European Union.

Based on the analysis of the current state of affairs, we argue for setting up a multidisciplinary large-scale coordinated initiative, the European Human Language Project (HLP). Within the HLP, eleven policies are proposed and assessed. These policies are grouped into: institutional policies, research policies, industry policies, market policies, and public service policies.

DANIELIUS ALGIRDAS RALYS

Vilniaus universiteto Taikomųjų mokslų institutas

MAŠININIS VERTIMAS LIETUVIŲ KALBAI

Vertimas iš kitos kalbos visuomet yra tam tikras intelektualinis iššūkis. Gal čia gali padėti kompiuteris? 1949 m. Warren'as Weaver'is pasiūlė panaudoti kompiuterius tekstams versti. Atsiranda terminas – mašininis vertimas (MV). Mašininis vertimas pirmaisiais dešimtmečiais buvo sparčiai vystomas, siekiant įgyti strateginį pranašumą šaltajame kare. Populiariausios verčiamos kalbos – rusų ir anglų. Vyrauja pažodinis vertimas, sukuriami dideli kompiuteriniai dvikalbiai žodynai, apimantys virš 200 000 žodžių.

1950–1960 metais atsiranda mašininio (kompiuterinio) vertimo sistemos, kurias galima pavadinti taisyklinėmis (*rule-based*). Jos kuriamos laikantis požiūrio, jog kalbą galima aprašyti naudojant tam tikrų taisyklių (taip pat ir gramatinių) sistemą. Tai buvo optimistinis laikotarpis – tikėtasi per keletą metų sukurti tobulą mašininį vertimą. Tačiau kompiuteris sunkiai „supranta" gramatiką. Verčiant fleksines kalbas reikia keliasdešimt tūkstančių taisyklių, kurios turi būti tarpusavyje suderintos. To niekas dar nėra tinkamai padaręs. Toliausiai pažengusios sistemos buvo SYSTRAN MV sistema, pradėta naudoti Europos Komisijoje, bei rusiška PROMT vertimo sistema. Vėliau taisyklinio MV pažanga sulėtėjo. Europinis EUROTRA projektas (1982–1992 m.), kainavęs apie 50 000 000 ECU, baigiasi nesėkme – šimtai specialistų taip ir nesukūrė veikiančios MV sistemos. Taip prasidėjo rimta taisyklinio MV krizė. Jau daug metų trypčiojama vietoje.

Kilo klausimas – jei negalime parašyti tiek daug taisyklių, tai gal galima versti be gramatikos? 1990 m. įvyksta naujas proveržis – *IBM Thomas J. Watson Research Center* tyrėjų grupė suformuluoja statistinio mašininio vertimo pagrindus. Vertimo procesas prilyginamas tam tikro pranešimo perdavimui triukšmingu kanalu. Dekoduojama remiantis Bajeso teorema. Vertimas remiasi tekstynais, ypač svarbūs yra dvikalbiai tekstynai. Vyko greitas statistinio MV tobulinimo procesas. Europos Komisijos remto projekto *EuroMatrix* metu sukurtas universalus atvirojo kodo statistinio mašininio vertimo programinės įrangos paketas MOSES, kurio pagrindu sukurtos pramoninio lygio MV sistemos. Buvo gauti geri rezultatai – pasirodo, galima

versti neturint nei žodyno, nei jokio supratimo apie gramatiką! Šis metodas labai palengvino fleksinių kalbų vertimą.

Mašininio vertimo pasiekimai 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 (http://vertimas.vdu.lt/twsas/). Taisyklinio vertimo variklį pateikė rusų kompanija PROMT, o kiti lingvistiniai resursai buvo paruošti Lietuvoje. Bendro pobūdžio tekstų vertimo kokybės įverčiai BLEU metrikoje (procentais) – apie 10. Praktikoje tai reiškia, kad adekvačiai suprantamas tik kas trečias sakinys. Ši vertimo priemonė dar turi nemažas galimybes pagerinti vertimo kokybę, pvz., plečiant frazių žodyną. Nuo 2008 m. rugsėjo 25 d. *Google Translate* palaiko ir lietuvių kalbą. 2014 m. atliktų testų duomenimis, bendro pobūdžio tekstų vertimo kokybės BLEU įverčiai buvo apie 17.

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 (https://www.versti.eu/). 2014 m. atliktų testų duomenimis bendro pobūdžio tekstų vertimo kokybės BLEU įverčiai daugiau kaip dvigubai viršijo taisyklinio vertimo rezultatus ir buvo praktiškai tolygūs *Google* vertimo sistemai. Verčiant tam tikrų sričių dokumentus (pvz., teisės) BLEU įverčiai maždaug dvigubai aukštesni, nei verčiant bendrojo pobūdžio tekstus, ir žymiai viršijo *Google* (2014-09-19) rezultatus. Vis dėlto tokie, net geriausi mašininiai vertimai dažnai prašyte prašosi geresnio vertėjo įsikišimo. Tad ar mašinos gali versti išties gerai?

Pastarieji keleri metai žada naujus proveržius naudojant neuroninį mašininį vertimą. Neuroniniai tinklai patys konstruoja transformavimo taisykles. Naujausios neuroninio MV sistemos jau kartais verčia geriau už vidutinišką vertėją! Vilniaus universitetas ruošiasi 2018 m. pradėti įgyvendinti naujos kartos neuroninį mašininį vertimą anglų, lietuvių, lenkų, prancūzų, rusų ir vokiečių kalboms.

Taigi, naujausi mašininio vertimo pasiekimai neaplenkia ir lietuvių kalbos.

ERIKA RIMKUTĖ, AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, ANDRIUS UTKA
Vytauto Didžiojo universitetas

LIETUVIŲ KALBOS MORFOLOGIŠKAI IR SINTAKSIŠKAI ANOTUOTI TEKSTYNAI

Pranešime pristatomi du anotuoti lietuvių kalbos tekstynai, parengti Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre (KLC). 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. Tekstynui anotuoti naudotas KLC parengtas morfologinis anotatorius. Tekstynas yra sužymėtas dviem formatais: KLC sukurtu formatu ir tarptautiniu TEI P5. Morfologinės pažymos, sudarytos remiantis MULTEXT-East formato (http://nl.ijs.si/ME/V4/msd/html/index.html) pavyzdžiu, kur kiekviena santrumpa atitinka konkrečią morfologinę kategoriją (nuo 2 iki 14).

Sintaksiškai anotuotas tekstynas ALKSNIS, kaip aukso standartas tolesniems tyrimams ir ištekliams, parengtas 2016 m. Šį tekstyną sudaro 2355 sakiniai (apie 30 tūkst. žodžių), imti iš įvairių stilių tekstų. Tekstyno anotavimas paremtas automatinio morfologinio ir sintaksinio anotavimo principais, pritaikytas sintaksinių priklausomybių (angl. *dependencies*) modelis. Sintaksiniu analizatoriumi, kuris sukurtas KLC *Haskell* kalba, automatiškai sugeneruoti priklausomybių medžiai (angl. *dependency trees*) pateikiami grafiškai medžio principu, kur kiekviena medžio viršūnė atitinka sakinio žodį, skyrybos ženklą ar kitą sakinio vienetą. Priklausomybių ryšiai tarp žodžių yra nurodomi briaunomis, o prie kiekvieno žodžio sutrumpintai pateikiama morfologinė ir sintaksinė informacija. Sintaksinei informacijai nurodyti naudojama 18 sintaksinių pažymų ir jų variantų, pvz., *Pred* – predikatas, *Sub* – subjektas ir t. t. Sintaksiniai medžiai pavaizduoti *TrEd* (https://ufal.mff.cuni.cz/tred/) redaktoriumi, kuriuo medžius galima redaguoti, atlikti morfologinių arba sintaksinių funkcijų paiešką. Visas automatiškai suanotuotas tekstynas peržiūrėtas ir pataisytas lingvistų.

Abu anotuoti tekstynai prieinami tarptautinėje mokslinių tyrimų infrastruktūroje *Clarin* (http://clarin-lt.lt/).

ERIKA RIMKUTĖ, AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, IEVA BUMBULIENĖ, JOLANTA KOVALEVSKAITĖ, TOMAS KRILAVIČIUS, JUSTINA MANDRAVICKAITĖ, LAURA VILKAITĖ

Vytauto Didžiojo universitetas

DUOMENŲ BAZĖ LIETUVIŲ KALBOS PASTOVIESIEMS JUNGINIAMS

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, į jį įeina 12 tekstų kategorijų: DELFI veidas, projektai, DELFI mokslas, DELFI auto, sportas, DELFI gyvenimas, DELFI žmonės, DELFI pilietis, verslas ir kt. Tekstyno 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. Iš duomenų bazės bus galima pasirinkti reikalingą informaciją kolokacijų žodynui ar kitam leksikografiniam darbui.

Šiuo metu duomenų bazėje sukelti visi dvižodžiai pastovieji junginiai, kurie buvo pažymėti bandomajame tekstyne (tai minėto tekstyno dalis – 72 tūkst. žodžių). Iš viso yra apie 2700 lemų (antraštinių formų) ir daugiau nei 35 000 kaitybinių formų. Prieš tai tekstynas buvo automatiškai morfologiškai anotuotas, tik taip buvo gãlima skirtingas to paties junginio kaitybines formas suvesti į vieną lemą. Naudojamos *Universal Dependency* gramatinės pažymos (žr. http://universaldependencies.org/u/pos/).

Visas anotuotas tekstynas bus pasiekiamas per *BlackLab* (http://inl.github.io/BlackLab/) – tekstyno administravimo programą, kuri suteikia plačias paieškos galimybes.

Duomenims saugoti naudojama *MongoDB* duomenų bazė, joje įrašai saugomi JSON formatu (žr. https://www.mongodb.com/what-is-mongodb). Įrašams peržiūrėti ir redaguoti naudojama *Mongo-express* administravimo sistema (žr. https://github.com/mongo-express/mongo-express), suteikianti prieigą prie duomenų bazės per interneto naršyklę.

Mongo-express suteikia galimybę ieškoti norimų įrašų pagal esančius laukus ir lengvai redaguoti, įrašyti reikalingą informaciją naudojant grafinę vartotojo sąsają.

ALGIRDAS SAUDARGAS
Europos Parlamentas

GYVOJI KALBA DIRBTINIAME PROTE

Pranešimu siekiama paaiškinti motyvus, paskatinusius inicijuoti Europos Parlamento Mokslinio perspektyvų tyrimo skyriaus tyrimą „Kalbų lygybė skaitmeniniame amžiuje. Gimtosios kalbos projektas“. Nerimą sukėlė 2010–2012 metais META-NET surinkti ir pateikti duomenys apie Europos kalbų technologinę paspirtį. Lietuvių kalba kartu su kai kurių kitų mažų šalių kalbomis pagal visus keturis kriterijus buvo įvertinta žemiausiu balu, kaip „turinti menką technologinę paspirtį arba jos visai neturinti“. Per pastaruosius metus padėtis mažai tepasitaisė. Tapo akivaizdu, kad kalbos technologijos nesulaukia tinkamo dėmesio ne tik Lietuvos, bet ir Europos politikų darbotvarkėje. Todėl Europos Parlamente buvo surengta konferencija ir prieš trejus metus inicijuotas tyrimas.

Esminis klausimas, 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. Pastaruoju metu pasiekta ženkli pažanga įvairiose srityse. Ypač daug žadantys rezultatai gauti naudojant dirbtinį intelektą, pavyzdžiui, dirbtinius daugiasluoksnės savimokos neuroninius tinklus. Formuojasi įspūdis, kad tereikia kaupti labai didelius duomenų kiekius, o dirbtinis intelektas visą informaciją apie gimtąją kalbą pasiims pats.

Pranešime pateikiamos įžvalgos, rodančios, kad yra ne visiškai taip. Pateikiamas palyginimas tarp požiūrio į žmogaus protavimo mechanizmus, pagrįsto šiuolaikinės kognityvinės neuropsichologijos tyrimų rezultatais, ir dirbtinio intelekto metodų panoramos. Iš šio palyginimo išplaukia, 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. Tas pats tinka ir kalbos technologijoms, jei jas suprasime kaip kalbos mechanizmų žmogaus smegenyse (gyvosios kalbos) modelius. Todėl kiekviena kalbinė bendruomenė privalo rūpintis, kad kalbos technologijos, atitinkančios simbolinį dirbtinį intelektą (vadinamosios „taisyklėmis grįstos“ programos) išsamiai ir tiksliai atspindėtų visą gimtosios kalbos sandarą, o neuroninių tinklų savimokai būtų sukurta turininga gimtosios kalbos (ir gimtosios kultūros) aplinka.

MINDAUGAS ŠINKŪNAS
Lietuvių kalbos institutas

KAS NUO KO NUSIRAŠĖ? BIBLIJOS VERTIMŲ ISTORIJOS TYRIMO AUTOMATIZAVIMAS IR VIZUALIZACIJA

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. Algoritmo veikimą ir apskaičiuojamą procentinę išraišką daugiausia lemia eilučių ilgis, trumpiausia leidžiama lyginti seka ir transliteracijos tikslumas.

Skirtingas eilučių ilgis (biblinė eilutė gali būti cituojama kaip fragmentas, kuris lygintinas su pilna eilute) reikalauja atsižvelgti į šaltinių chronologiją ir pagal ją rinktis atitinkamus panašumo skaičiavimo parametrus. Optimaliausia trumpiausia leidžiama lyginti ženklų seka nustatyta eksperimentiškai.

Palyginimą automatizuoti kliudo nevienareikšmė senųjų raštų ortografija. Ji niveliuojama transliteracijos taisyklėmis, kurių taikymas diferencijuojamas atsižvelgiant į svarbiausias kiekvieno autoriaus dėsningas rašybos ypatybes. 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. Duomenys juose gali būti rikiuojami pagal šaltinių chronologiją, pagal panašumo vidurkį, pagal lygintų eilučių skaičių, pagal eilučių eilės tvarką šaltinyje, pagal atskirų eilučių panašumą, pagal eilučių panašumo skirtumus dviejuose šaltiniuose ir t. t.

Programinės analizės metu gauti Bretkūno *Postilės* (1591) biblinių eilučių palyginimo rezultatai atitinka ankstesnių tyrėjų kitais metodais prieitas išvadas.

Lyginimas neapima ir negali įvertinti kitakalbių Biblijos vertimų įtakos. Automatiniai skaičiavimai patys savaime neįrodo eilučių perėmimo. Procentinės panašumo išraiškos nėra absoliučios, jos kinta priklausomai nuo pasirenkamų skaičiavimo parametrų. Svarbiausiu darbo rezultatu laikytina sukurta priemonė, parodanti orientacinį šaltinių ar jų dalių panašumo santykį, tai atveria galimybes tolesniems, detalesniems tyrimams.

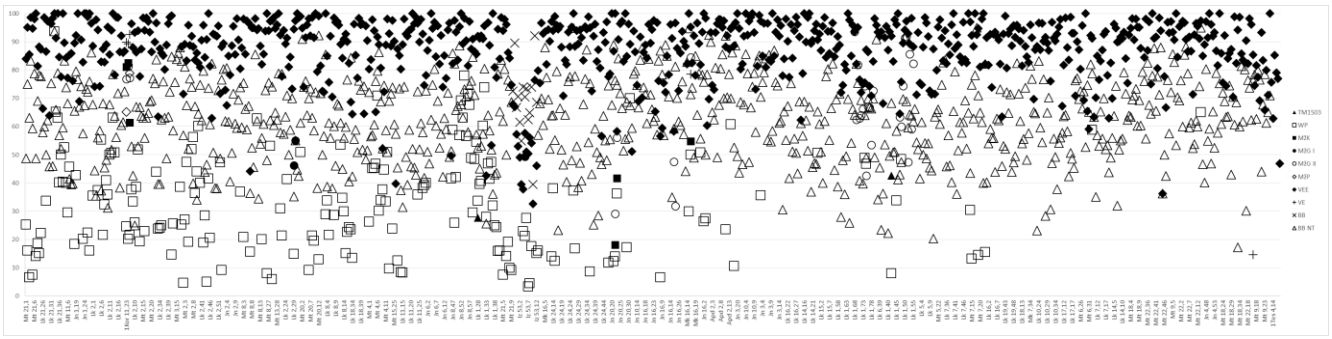

Diagrama: 1591 m. Bretkūno *Postilės* perikopių, kurias sudaro 750 Biblijos eilučių, panašumas į kituose XVI a. šaltiniuose vartojamas. *TM* – XVI a. pr. poteriai, *MžK* – Mažvydo katekizmas (1547), *MžG* I–II – 1566 ir 1570 m. giesmynai, *WP* – 1573 m. postilė, *MžP* – 1589 m. „Parafrazis“, *VE* – Vilento *Enchiridionas* (1579), *VEE* – Vilento *Evangelijos bei Epistolos* (1579), *BB* – Bretkūno *Biblija*, *BB NT* – Bretkūno *Naujasis Testamentas*.

LAIMUTIS TELKSNYS

Vilniaus universiteto Matematikos ir informatikos institutas

RAŠYTINĖ IR SAKYTINĖ LIETUVIŲ KALBA ĮPRASTOJE IR ELEKTRONINĖJE TERPĖSE

Žmonės bendrauja tiesiogiai ar/ir per internetą šnekėdami įvairiomis kalbomis, vartodami rašto ženklus: lotyniškus, arabiškus, kiriliką, hieroglifus.

Reikšminga, kad žmonių bendriją papildo sparčiai plintančios išmaniosios mašinos – robotai, automobiliai, išmanieji daiktai, su kuriais dirbdami šnekėsime.

Klausimas: kaip elgtis mums, keliems milijonams lietuvių, kad neišnyktume daugiau kaip 7 milijardų rašančiųjų ir šnekančiųjų žmonių bei ateinančių išmaniųjų mašinų okeane.

Atsakymas: 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.

Atlikus šiuos darbus bus užtikrinta, kad:

– žmonių dialogas su informacinių technologijų išmaniosiomis priemonėmis (telefonais, planšetėmis, dėvimąja kompiuterine įranga, išmaniaisiais daiktais, robotais) vyktų taisyklinga lietuvių rašytine ir/arba sakytine kalba;

– Lietuvos elektroninėse duomenų saugyklose nelietuviški asmenvardžiai būtų pateikiami lietuviškų rašto ženklų kodais bei originalo rašmenimis, o nelietuviškų asmenvardžių tarimas balsu būtų pateikiamas garso įrašais, padarytais juos tariant lietuvių kalbos fonemų garsais ir garso įrašais originalo kalba.

LAIMUTIS TELKSNYS, GEDIMINAS NAVICKAS

Vilniaus universiteto Matematikos ir informatikos institutas

LIETUVIŲ ŠNEKA VALDOMOS PASLAUGOS. PADĖTIS. PERSPEKTYVOS

Pasaulis keliasi į elektroninę terpę.

Auga joje šnekų – šnekamųjų kalbų – vartojimo reikšmė.

Kalba yra kultūros vertybė, neįkainojamas turtas, pridėtinės vertės šaltinis.

Lietuvių kalba nekomercinė.

Rūpinantis išlaikyti lietuvių šnekos tvirtą padėtį, daugiau kaip 7 milijardų įvairiomis kalbomis šnekančiųjų žmonių ir ateinančių robotų bei išmaniųjų mašinų okeane, sistemingai vykdomi lietuvių šneka valdomų paslaugų kūrimo ir jų panaudojimo plėtros dabai.

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.

Sukurtos priemonės gamybininkams, paslaugų tiekėjams teikia pavyzdžius, rodančius lietuvių šnekos valdomų paslaugų panaudojimo galimybes, padeda taupiau pritaikyti lietuvių šneka valdomas paslaugas įvairioms veikloms tobulinti.

Pasaulis intensyviai keliasi į mobiliąją elektroninę terpę. Todėl numatoma sukurti naujas lietuvių šneka valdomas paslaugas mobiliajai elektroninei terpei – išmaniesiems mobiliesiems telefonams, planšetėms, išmaniesiems laikrodžiams, robotams – užtikrinančias galimybes gauti paslaugas vartojant 3000 lietuvių žodžių žodyną ir tokios apimties žodynui būtiną infrastruktūrą – garsyną.

AUDRIUS VALOTKA, GEDIMINAS NAVICKAS
Vilniaus universiteto Filologijos fakultetas,
Vilniaus universiteto Matematikos ir informatikos institutas

TOLYN NUO GUTENBERGO SPAUDOS PRESO: LIETUVIŲ ŠNEKA NAUJAUSIOSE TECHNOLOGIJOSE

Pranešime apžvelgiami pastarųjų metų darbai lietuvių šnekos sintezės ir atpažinimo srityje, atskleidžiamos galimybės bei planai pritaikyti šnekos sintezės ir atpažinimo sprendinius naujausiose technologijose.

Europai nestinga mokslinių ir technologinių inovacijų kalbų technologijų srityje, tačiau Lietuva šiuo atžvilgiu yra gerokai atsilikusi. Viena iš šio atsilikimo priežasčių – lietuvių kalba yra nekomercinė, todėl, naudodamiesi svetima produkcija, visada būsime technologijų nuošalėje.

Šiuo metu lietuvių šneka informacinėse technologijose yra panašioje situacijoje, kaip prieš 30 metų, kai buvo sprendžiami lietuviškų diakritinių ženklų kompiuterijoje naudojimo klausimai. Dauguma kompiuterių buvo neįgalūs dirbti su lietuviškomis raidėmis, teko jas sukurti, įtraukti į pasaulinius standartus. Gyvename laikais, kai šneka vis plačiau naudojama kaip žmogaus ir kompiuterio sąsaja. Ar galėsime kalbėtis su kompiuteriais lietuviškai? Tam reikalingi tyrimai lietuvių šnekos atpažinimo ir sintezės srityse. Ir ne tik tyrimai, bet ir jų pagrindu atsirandantys informacinių technologijų produktai – leidžiantys lietuviams bendrauti su kompiuteriais gimtąja kalba.

Todėl 2013 metais prasidėjo Struktūrinių fondų finansuojamas projektas „Lietuvių šneka valdomos paslaugos – LIEPA“, kurio metu buvo sukurtas, be viso kito, lietuvių šnekos sintezatorius ir atpažintuvas. Tai pirmasis didelės apimties projektas, skirtas lietuvių šnekos tyrimams ir taikymams skaitmeninėje erdvėje. Šio projekto vykdymo metu lietuvių šnekai atverti vartai į skaitmeninę erdvę, tačiau sustoti negalima, nes padėti tik pamatai. Dėl šios priežasties planuojamas projektas „Lietuvių šneka valdomų paslaugų plėtra – LIEPA 2“. Projekto LIEPA 2 rezultatai bus atvirai ir nemokamai prieinami visiems norintiems ir turėtų paskatinti lietuvių šnekos naudojimą informacinių technologijų produktuose.

Esminiai planuojami LIEPA 2 rezultatai – infrastruktūriniai sprendimai: 1000 valandų lietuvių šnekos garsynas, 3000 žodžių lietuvių šnekos atpažintuvas ir patobulintas lietuvių šnekos sintezatorius. Taip pat labai svarbu, kad šie lietuvių šnekos IT sprendimai kuriami taip, kad galės būti taikomi ne tik stacionariuosiuose kompiuteriuose, bet ir mobiliojoje terpėje: mobiliuosiuose telefonuose, planšetiniuose kompiuteriuose, robotuose, išmaniuosiuose laikrodžiuose.

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.

Projekto rezultatai bus skirti viešajam naudojimui, t. y. „normaliai“ kalbantiems žmonėms, todėl nuo anksčiau sukurto „laboratorinio“ garsyno bus žengiama prie „tikro“ kalbėjimo: su foniniu triukšmu, vyrų ir moterų, jaunų ir senų, įvairaus tembro ir kalbėjimo greičio, emocijų raiškos ir pan., taigi naujasis garsynas remsis gyvo pokalbio įrašais.

Kuriant infrastruktūrines paslaugas bus suvienytos lietuvių kalbos specialistų ir informatikų pajėgos. Garsynui kurti bus pasitelkiami iki šiol sukurti garsynai ir sukurtas didžiausias lietuvių šnekos anotuotas garsynas (ne mažiau kaip 1000 valandų anotuoto teksto).

Lietuvių šnekos sintezatorius ir atpažintuvas – tai lyg „statybiniai blokai“, kurie paskatins programinės įrangos kūrėjus naudoti lietuvių šnekos atpažinimo ir sintezės funkcionalumą, nes jo nebereikės patiems kurti – užteks tiesiog integruoti LIEPA 2 sukurtus ir laisvai platinamus variklius į savo kuriamą programinę įrangą. Tokių pavyzdžių jau turime. Projekto LIEPA sukurtą sintezės variklį jau naudoja ne viena interneto svetainė, tarp jų: laikraštis „Lietuvos žinios“, portalo *DELFI.lt* mobilioji svetainė, Vilniaus savivaldybės svetainė.

Garsynas „LIEPA 2“ galės būti laisvai ir nemokamai naudojamas tolesniems lietuvių šnekos tyrimams bei naujiems atpažintuvams kurti.

LIEPA 2 numatomos sukurti tipinės paslaugos, rodančios naujas lietuvių šneka valdomų paslaugų panaudojimo galimybes:

1 **Ugdančio roboto valdytuvas** – vaikų sprendimų priėmimo gebėjimus ugdantis humanoidinis robotas.

2. **Skambintuvas** – asmens kontaktų mobiliajame telefone valdytuvas balsu, pritaikomas vartotojo poreikiams.

3. **Taksi iškviestuvas** balsu lietuvių kalba.

4. **Mobilusis sintezatorius** akliesiems skaitantis tekstus lietuviškai per mobiliuosius telefonus.

5. **Interneto naujienų skaitytuvas** – balsu skaitantis vartotojo pasirinktas naujienas iš interneto.

6. **Tarpkalbinis komunikatorius** – lietuvių–kinų kalbų.

Kuriamų paslaugų tikslas – parodyti, kaip veikia infrastruktūriniai sprendimai, kaip jie gali būti taikomi, ir paskatinti lietuvių šnekos technologijų plitimą bei taikymą IT produktuose.

DAIVA ŠVEIKAUSKIENĖ, VYTAUTAS ŠVEIKAUSKAS
Lietuvių kalbos institutas

LIETUVIŲ KALBOS SKAITMENINĖ GRAMATIKA

Lietuvių kalbos institute pradėta kurti lietuvių kalbos skaitmeninė gramatika. Šiuo projektu planuojama prisijungti prie tarptautinės sistemos DIGITAL GRAMMARS, apimančios šiuo metu 32 kalbas. Tarp jų jau yra estų ir latvių kalbos. 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 (šioje vertimo sistemoje naudojami statistiniai metodai) 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.

Skaitmeninę gramatiką sudaro du lygmenys: abstraktusis ir konkretusis. Abstrakčiajame lygmenyje kaupiamos sąvokos, tai yra forma, bendra visoms kalboms. Konkretusis lygmuo susijęs su atskira kalba ir atspindi specifinius jos bruožus. Vertimo metu sakinys transformuojamas į abstraktų semantinį pavidalą, iš kurio generuojamas kitos kalbos sakinys remiantis tik tos kalbos, į kurią jis yra verčiamas, savybėmis, ir visiškai neatsižvelgiama į tos kalbos morfologiją, iš kurios jis buvo verstas. Taip galima gauti daug tikslesnius rezultatus, nes visada garantuotai perduodama tiksli sakinio reikšmė.

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.

JOLANTA ZABARSKAITĖ, DEIMANTĖ BUDRIŪNAITĖ, SKIRMANTAS ŠERMUKŠNIS

Lietuvių kalbos institutas, *NetCode*

E. KALBA – SKAITMENINIŲ KALBOS IŠTEKLIŲ NAUDOJIMO(SI) INOVACIJA

XXI amžiuje, atsiradus kūrybos visuomenei, naujosioms medijoms ir skaitmeninėms aplinkoms vis labiau besiskverbiant į visuomenės gyvenimą, atsigręžiama į kalbą kaip fenomeną, kuris užprogramuotas kurti kultūrinę, socialinę ir ekonominę vertę. Globalizacijos epochoje kalbos tampa *takiosiomis* (Z. Baumano terminas): bet kuri nacionalinė kalba, tarp jų ir lietuvių, nebeegzistuoja izoliuotai, atliepdama tik vienos kalbinės bendruomenės poreikius. Išskirtinumas yra svarbus globalizacijos procesų elementas, neleidžiantis bet kuriai, ypač mažai vartojamai (*less-widely used*) kalbai ir kultūrai, nugrimzti informacijos apie pasaulio kalbas ir kultūras gaudesyje. Atsižvelgiant į tai, pasaulio įvairovė ir virtualybėje turi išlikti pasaulio įvairove, o kalbos ištekliai, kylant naujiems poreikiams, iššūkiams ir lūkesčiams, – natūraliai kisti: turi rastis naujų skaitmeninių kalbos išteklių ideologijų ir koncepcijų, kurtis prasmių koridoriai, sąvokų inžinerijos, kalbos – žinių – meninio žinojimo išteklių saitynai, informacijos paieškos galimybių įvairovė.

Pranešime pristatomas naujasis Lietuvių kalbos instituto projektas, skirtas plėtoti Lietuvių kalbos išteklių informacinę infrastruktūrą LKIIS www.lkiis.lt, kuri buvo sukurta įgyvendinant projektą *„IRT sprendimų bei turinio, padedančių išsaugoti lietuvių kalbą viešojoje erdvėje, kūrimas bei galimybių jais naudotis sudarymas“*.

Pirmuoju projekto vykdymo etapu buvo sukurtas susistemintų skaitmeninių lietuvių kalbos išteklių paieškos saitynas: suskaitmeninta 11 vienakalbių ir dvikalbių žodynų ir 5 kartotekos, sukurtos elektroninės šių išteklių duomenų bazės ir jų valdymo įrankiai, sukurta vieša šių išteklių prieiga visuomenei bei integruotos elektroninės paslaugos (e. mokymai, e. kartotekos, e. žodynai). Remiantis LKIIS kaupiama portalo naudojamumo statistika, suskaitmenintus lietuvių kalbos išteklius ir sukurtas el. paslaugas 2015 m. naudojo vidutiniškai po 5000 vartotojų per mėnesį, fiksuojama po 9000 aktyvių sesijų. Vertinant nuo sistemos naudojimo / naudojimosi pradžios šis skaičius didėjo: kasdien prie portalo prisijungdavo

vidutiniškai po 180 naujų aktyvių vartotojų. Naujuoju „E.kalba“ projektu siekiama integruoti į LKIIS naujus struktūrizuotus lietuvių kalbos išteklius ir sukurti technologijas, sprendimus ir įrankius lietuvių kalbos išteklių semantinei paieškai atlikti, jos rezultatams vizualizuoti bei šių technologijų pagrindu sukurti naujas e. paslaugas bei e. sprendimus vartotojams.

Pranešime daugiausia dėmesio skiriama naujojo Lietuvių kalbos instituto projekto „E. kalba“ infrastruktūros generuojamoms paslaugoms – *„Paieška žodžių tinkle“*, *„E. rinkodara“*, *„E. sąvokos“* ir *„E. patarimai“* pristatyti bei funkcionalumams aptarti, paslaugų technologiniams aspektams pateikti. Projektu numatoma, kad kuriant žodžių tinklą bus taikomi tarptautiniu mastu pripažinti žodžių tinklų duomenų modeliai bei praktikos (pvz., *Princeton Wordnet duomenų modelis*), pirminė žodžių tinklo versija apims pasaulinės žodžių tinklų asociacijos (angl. *The Global WordNet Association*) rekomenduojamą bendrų bazinių sąvokų (angl. *Comon Base Concpets*) rinkinį, kuris buvo apibrėžtas *EuroWordNet* ir *BalkaNet* projektuose, taip pat pirminė žodžių tinklo versija bus parengta integruojant esamus leksikografinius išteklius (*Dabartinės lietuvių kalbos žodyną*, *Sinonimų žodyną*, *Antonimų žodyną* ir kt.).

Projekto funkcionalumai apims inovatyvias dirbtinio intelekto (angl. *artificial intelligence*) 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 (*Princeton* žodžių tinklu (angl. *Princeton WordNet*) bus vizualizuojami sąvokos ryšiai, vartotojams), 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 pristatomi bei aptariami ir pagrindiniai projekto „E. kalba“ rezultatai bei planuojama nauda: numatoma galimybė esamus lietuvių kalbos išteklius pritaikyti rinkodaros tikslams, pavyzdžiui, atliekant nuomonių analizę virtualioje erdvėje, aptariamos galimybės esamus lietuvių kalbos išteklius pritaikyti teikiant tiek viešąsias, tiek ir privačias elektronines paslaugas, pavyzdžiui, vykdant vartotojų užklausų analizę ir praturtinimą kalbine informacija, pristatoma galimybė pakartotinai panaudoti lietuvių kalbos žodžių prasminio tinklo duomenis kuriant komercinius produktus ir paslaugas. Projektu taip pat numatoma užtikrinti ir didesnį praktinių žodžių darybos įrankių ir kalbos patarimų prieinamumą internete, nuolatinį praktinių kalbos patarimų atnaujinimą, atitinkantį šių dienų aktualijas, užtikrinti esamų leksikografinių išteklių sąveikumą ir jų susiejimą prasminiais ryšiais, sukuriamos galimybės naudotis

integruotais ir prasminiais ryšiais susietais leksikografiniais lietuvių kalbos ištekliais bei lietuvių kalbos žodžių tinklą susieti su kitų kalbų žodžių tinklais (daugiau kaip 200 kalbų) ir tokiu būdu kurti technologinius sprendimus bendrajai Europos skaitmeninei rinkai (angl. *digital single market*).

APIE AUTORIUS

AGNĖ BIELINSKIENĖ – dr., Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro mokslo darbuotoja; el. p. agne.bielinskiene@vdu.lt

LAIMUTIS BILKIS – dr., Lietuvių kalbos instituto Baltų kalbų ir vardyno tyrimų centro vyresnysis mokslo darbuotojas; el. p. laimutis.bilkis@lki.lt

LOÏC BOIZOU – dr., Vytauto Didžiojo universiteto Užsienio kalbų, literatūros ir vertimo studijų katedros lektorius; el. p. lboizou@gmail.com

DEIMANTĖ BUDRIŪNAITĖ – Lietuvių kalbos instituto Bendrųjų reikalų ir infrastruktūros skyriaus vadovė; el. p. deimante.budriunaite@lki.lt

IEVA BUMBULIENĖ – Baltijos pažangių technologijų instituto jaunesnioji mokslo darbuotoja; el. p. ieva.bumb@gmail.com

VIRGINIJUS DADURKEVIČIUS – Vilniaus universiteto Taikomųjų mokslų instituto inžinierius; el. p. dadurka@gmail.com

GINTARĖ JUDŽENTYTĖ – dr., Vilniaus universiteto Lietuvių kalbos katedros lektorė; el. p. gintare.judzentyte@flf.vu.lt

PIJUS KASPARAITIS – dr., Vilniaus universiteto Matematikos ir informatikos fakulteto docentas; el. p. pkasparaitis@yahoo.com

JOLANTA KOVALEVSKAITĖ – dr., Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro mokslo darbuotoja; el. p. jolanta.kovalevskaite@vdu.lt

TOMAS KRILAVIČIUS – prof. dr., Baltijos pažangių technologijų instituto Technologijų direktorius; el. p. t.krilavicius@gmail.com

JUSTINA MANDRAVICKAITĖ – Baltijos pažangių technologijų instituto jaunesnioji mokslo darbuotoja; el. p. justina@bpti.lt

RITA MILIŪNAITĖ – dr., Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro vyriausioji mokslo darbuotoja; el. p. rita.miliunaite@lki.lt

DAIVA MURMULAITYTĖ – dr., Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro vyriausioji mokslo darbuotoja; el. p. daiva.murmulaityte@lki.lt

GEDIMINAS NAVICKAS – Vilniaus universiteto Matematikos ir informatikos instituto Projektų skyriaus vyresnysis specialistas; el. p. gediminas.navickas@mii.vu.lt

DANIELIUS ALGIRDAS RALYS – Vilniaus universiteto Taikomųjų mokslų instituto inžinierius; el. p. danielius.ralys@gmail.com

ERIKA RIMKUTĖ – doc. dr.; Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro vyresnioji mokslo darbuotoja; el. p. erika.rimkute@vdu.lt

RAFAEL RIVERA – *iClaves* direktorius; el. p. rafaelrivera@iclaves.es

ALGIRDAS SAUDARGAS – Europos Parlamento narys;
el. p. algirdas.saudargas@europarl.europa.eu

GINTARAS SKERSYS – dr., Vilniaus universiteto Matematikos ir informatikos fakulteto asistentas; el. p. gintaras.skersys@mif.vu.lt

SKIRMANTAS ŠERMUKŠNIS – *Netcode* projektų vadovas;
el. p. skirmantas.sermuksnis@gmail.com

MINDAUGAS ŠINKŪNAS – dr., Lietuvių kalbos instituto Raštijos paveldo tyrimų centro mokslo darbuotojas; el. p. mindaugas.sinkunas@lki.lt

VYTAUTAS ŠVEIKAUSKAS – Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro programuotojas; el. p. vytautas.sveikauskas@lki.lt

DAIVA ŠVEIKAUSKIENĖ – dr., Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro mokslo darbuotoja; el. p. daiva.sveikauskiene@lki.lt

LAIMUTIS TELKSNYS – prof. habil. dr., Vilniaus universiteto Matematikos ir informatikos instituto profesorius, el. p. laimutis.telksnys@mii.vu.lt

ANDRIUS UTKA – doc. dr., Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro vadovas ir mokslo darbuotojas; el. p. andrius.utka@vdu.lt

AUDRIUS VALOTKA – dr., Vilniaus universiteto Filologijos fakulteto projektų vadovas, asistentas; el. p. audrius.valotka@flf.vu.lt

LAURA VILKAITĖ – dr., Baltijos pažangių technologijų instituto jaunesnioji mokslo darbuotoja; el. p. vilkaite.laura@gmail.com

JOLANTA ZABARSKAITĖ – prof. dr., Lietuvių kalbos instituto direktorė, Bendrinės kalbos tyrimų centro vyriausioji mokslo darbuotoja; el. p. jolanta.zabarskaite@lki.lt

VILMA ZUBAITIENĖ – doc. dr., Vilniaus universiteto Lietuvių kalbos katedros docentė; el. p. vilma.zubaitiene@flf.vu.lt