

LIETUVOS RESPUBLIKOS SEIMAS

NUTARIMAS

**DĖL LIETUVIŲ KALBOS PLĖTROS SKAITMENINĖJE TERPĖJE IR KALBOS
TECHNOLOGIJŲ PAŽANGOS 2021–2027 M. GAIRIŲ PATVIRTINIMO**

2020 m. d. Nr.
Vilnius

Lietuvos Respublikos Seimas, įgyvendindamas Lietuvos Respublikos Konstitucijos 14 straipsnio nuostatą ir Valstybinės kalbos įstatymą, n u t a r i a:

1 straipsnis.

Patvirtinti Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021-2027 m. gaires (pridedama).

2 straipsnis.

1. Pasiūlyti Lietuvos Respublikos Vyriausybei atsižvelgti į Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021-2027 m. gaires rengiant Lietuvos skaitmeninimo plėtros programą 2021–2030 metams ir atitinkamų metų valstybės biudžeto ir savivaldybių biudžetų finansinių rodiklių patvirtinimo įstatymo projektus.

2. Pavesti Valstybinei lietuvių kalbos komisijai atlikti Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021-2027 m. gairių įgyvendinimo stebėseną.

Seimo Pirmininkas

Teikia
Seimo nariai

Eugenijus Jovaiša

Liudas Jonaitis

Vytautas Juozapaitis

Edmundas Pupinis

LIETUVIŲ KALBOS PLĖTROS SKAITMENINĖJE TERPĖJE IR KALBOS TECHNOLOGIJŲ PAŽANGOS

2021–2027 M.

GAIRĖS

I SKYRIUS

BENDROSIOS NUOSTATOS

Pastaraisiais metais žinių visuomenė pereina į kokybiškai naują etapą, kurį žymi sparti inovatyvių informacinių technologijų (toliau – ir IT) plėtra, pirmiausia – didžiųjų duomenų kaupimas bei apdorojimas ir dirbtiniu intelektu grįstų technologijų kūrimas. IT vis plačiau diegiamos visose pagrindinėse visuomenės veiklos srityse, tokiose kaip valstybės administravimas ir teismų sistema, švietimas, mokslas, kultūra ir jos paveldo saugojimas, žiniasklaida, elektroninė bankininkystė, sveikatos apsauga, energetika, viešasis transportas, gamtosauaga, krašto apsauga, verslas ir t. t. Kartu šie pokyčiai kelia ir naujų uždavinių, iš kurių svarbiausi – glaudinti šių sričių sąveiką, užtikrinti kibernetinį saugumą ir apsaugą nuo dezinformacijos sklaidos, plėtoti nuotolinį mokymą, gerinti gyvenimo kokybę, mažinant kalbų barjerus, sprendžiant galėjimo įsidarbinti bei visuomenės senėjimo problemas, mažinant socialinę bei regionų atskirtį. Kalbos technologijos yra svarbi informacinių technologijų dalis ir vienas iš būtinų įrankių šiems uždaviniams spręsti.

1. Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 m. gairės (toliau – Gairės) parengtos siekiant užtikrinti visavertį lietuvių kalbos vartojimą skaitmeninėje terpėje, įtvirtinti ir palaikyti lietuvių kalbos statusą informacinėje visuomenėje, apsaugoti lietuvių kalbą nuo vadinamojo skaitmeninio išnykimo, sudaryti galimybes kitakalbiams integruotis į Lietuvos visuomenę ir mažinti lietuviškai kalbančios bendruomenės atskirtį globalioje žinių visuomenėje. Šie siekiai numatomi įgyvendinti gausinant skaitmeninius kalbos išteklius, plėtojant kalbos technologijas ir viešąsias paslaugas, atitinkančias informacinės visuomenės lūkesčius ir poreikius.

2. Gairės yra valstybinės kalbos politikos strateginis dokumentas, kuriame numatomos veiklos kryptys, uždaviniai ir priemonės, kaip daugiakalbėje skaitmeninėje aplinkoje išsaugoti kalbinę ir kultūrinę tapatybę kaip pagrindinę demokratinės visuomenės raidos ir lygiateisio lietuvių kalbos vartojimo sąlygą, užtikrinančią visavertį Lietuvos piliečių dalyvavimą Lietuvos ir Europos Sąjungos (toliau – ir ES) socialiniame, politiniame ir kultūriniame gyvenime.

3. Gairės parengtos atsižvelgus į Lietuvių kalbos informacinėje visuomenėje 2011–2019 metų programos įgyvendinimo rezultatus ir lietuvių kalbai skirtų technologijų būklės vertinimus. Rengiant Gaires remtasi šiais dokumentais:

- Lietuvių kalbos plėtros informacinėse technologijose 2014–2020 m. gairės, kurioms pritarta Valstybinės lietuvių kalbos komisijos 2013 m. spalio 24 d. posėdyje;
- Valstybės pažangos strategija „Lietuvos pažangos strategija „Lietuva 2030“, patvirtinta Lietuvos Respublikos Seimo 2012 m. gegužės 15 d. nutarimu Nr. XI-2015

„Dėl Valstybės pažangos strategijos „Lietuvos pažangos strategija „Lietuva 2030“ patvirtinimo“;

- Valstybinės kalbos įstatymas;
- Valstybinės kalbos politikos 2018–2022 metų gairės, patvirtintos Lietuvos Respublikos Seimo 2018 m. birželio 27 d. nutarimu Nr. XII-1318 „Dėl Valstybinės kalbos politikos 2018–2022 metų gairių“;
- Valstybinės kalbos politikos 2019–2022 metų gairių įgyvendinimo priemonių planas, pavirtintas Lietuvos Respublikos Vyriausybės 2019 m. vasario 13 d. nutarimu Nr. 147 „Dėl Valstybinės kalbos politikos 2019–2022 metų gairių įgyvendinimo plano patvirtinimo“.

4. Gairėse vartojamos toliau apibrėžtos sąvokos (kitos sąvokos suprantamos taip, kaip jos apibrėžtos 3-iaame šių Gairių punkte nurodytuose ir kituose Lietuvos Respublikos teisės aktuose):

4.1. Atviri duomenys (angl. *open data*) – laisvai prieinami institucijos veikloje ar dokumentuose užfiksuoti duomenys, informacija ar jos dalis, nepaisant jų pateikimo būdo, formos ir laikmenos, įskaitant registro duomenis, registro informaciją, registrai pateiktų dokumentų ir (arba) jų kopijų duomenis, valstybės informacinės sistemos duomenis, kuriuos visi asmenys gali pakartotinai naudoti ir platinti bet kokių tikslų, nurodydami jų šaltinį ir tik tomis pačiomis sąlygomis, kuriomis buvo gauti.

4.2. Didieji duomenys (angl. *big data*) – labai didelė duomenų sandauga, kuriai analizuoti ir apdoroti reikia specialių duomenų bazių valdymo įrankių.

4.3. Dirbtinis intelektas (DI, angl. *artificial intelligence*) – programine įranga grindžiamos virtualiosios technologijos (pavyzdžiui, šnekos sintezatoriai, mašininis vertimas, virtualūs asistentai, paieškos sistemos, teksto, garso ir vaizdo analizės sistemos) arba į techninę įrangą (pavyzdžiui, į pažangius robotus, autonominius automobilius, bepiločius orlaivius ar daiktų interneto objektus) integruojamos išmaniosios technologijos, analizuojančios savo aplinką ir darančios savarankiškus sprendimus nustatytam tikslui pasiekti.

4.4. Elektroninė paslauga (angl. *e-service*) – įvairiais informacinių ir ryšių technologijų kanalais bei priemonėmis nuotoliniu būdu gyventojams ir (arba) verslui teikiama viešoji arba administracinė paslauga.

4.5. Informacinių technologijų sprendinys – informaciją apdorojančių techninių ir programinių priemonių, skirtų tam tikram institucijų, gyventojų, verslo įmonių poreikiui tenkinti, visuma.

4.6. Lokalizavimas (angl. *localization*) – programinės įrangos, svetainės, kitų elektroninių išteklių arba elektroninės paslaugos pritaikymas prie tam tikros kalbinės ir kultūrinės aplinkos.

4.7. Mašininis vertimas (angl. *machine translation*) – automatinis vertimas iš vienos kalbos į kitą vertimo programa.

4.8. Mišrieji duomenys (angl. *multimodal data*) – duomenys, apimantys trijų kategorijų duomenis: sustruktūrintus, pusiau sustruktūrintus ir nesustruktūrintus.

4.9. Neuroninis tinklas (angl. *neural network*) – pagal neuronų veiklos biologinėje nervų sistemoje (smegenyse vykstančių procesų, gaunant informaciją, mokantis bei prisimenant) analogiją sumodeliuota dirbtinio intelekto sistema, kurios elementai geba mokytis gaudami numatytus įvedinius ir iš jų geba gauti išvedinius.

4.10. Natūralioji kalba (angl. *natural language*) – kalba, kuria šneka ir rašo žmonės.

4.11. Natūraliosios kalbos apdorojimas (angl. *natural-language processing*) – kompiuterinės lingvistikos šaka, tirianti kompiuterines sistemas, galinčias atpažinti sakytinę ir rašytinę natūraliąją kalbą ir reaguoti į ją.

4.12. **Ontologija** (angl. *ontology*) – bendrai naudojamas formalus tam tikros srities sąvokų (konceptų), tipų, jų tarpusavio priklausomybės, ryšių, aksiomų, dėsnų ir kt. aprašas.

4.13. **Skaitmeniniai kalbos ištekliai** (angl. *digital language resources*) – susisteminti skaitmeniniai sakytinės ir rašytinės kalbos duomenys (žodynai, tekstynai, terminynai, garsynai ir pan. duomenynai), naudojami švietimo bei mokslo tikslams, kalbos technologijoms ir jomis grindžiamoms paslaugoms kurti.

4.14. **Šnekos atpažinimas** (angl. *speech recognition*) – automatinis sakytinės kalbos garsų atpažinimas, žodžių sintezavimas ir jų užrašymas skaitmeniniu tekstu.

4.15. **Šnekos sintezatorius** – programa ar įrenginys, atliekantis šnekos sintezę.

4.16. **Šnekos sintezė** (angl. *speech synthesis*) – automatinis procesas, kurį atliekant skaitmeninis tekstas paverčiamas garsiniu ir yra perskaitomas.

II SKYRIUS

NACIONALINĖS, EUROPOS SĄJUNGOS IR PRIVATAUS VERSLO INICIATYVOS

Šiame skyriuje apžvelgiamos tarptautinės, nacionalinės ir privataus verslo iniciatyvos, kurios daro ir darys didelę įtaką lietuvių kalbos plėtrai skaitmeninėje terpėje ir kalbos technologijų pažangai.

5. 2003 m. vykusioje UNESCO Generalinės konferencijos 32-oje sesijoje priimta „Rekomendacija dėl daugiakalbystės skatinimo ir naudojimo bei visuotinės prieigos prie kibernetinės erdvės“ (*UNESCO Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace*)¹. Šių Gairių siekiams aktualios rekomendacijos, skirtos daugiakalbio turinio užtikrinimui (1, 2, 3, 4 punktai) ir viešojo turinio prieinamumo užtikrinimui (15, 18, 19 punktai). Tai pat svarbus UNESCO Generalinės konferencijos 37-oje sesijoje patvirtintos „Vidutinės trukmės strategijos 2014–2021“² 9 strateginis tikslas – skatinti saviraiškos laisvę, medijų vystymąsi ir prieigą prie informacijos ir žinių – ir jame suformuluotos nuostatos (80–93 punktai).

6. Rinkų susiskaidymo problemai spręsti skirta Bendrosios skaitmeninės Europos Sąjungos rinkos (angl. *Digital Single Market*) iniciatyva, kurioje kalbos technologijos vaidina svarbų vaidmenį, kurdamos naują erdvę ekonomikos augimui. Skaitmeninių kalbos technologijų, susijusių su rašytine ir sakytine kalba, ir semantinių internetinių paslaugų naudojimo plėtra turi tapti horizontaliąja politikos strategija, apimančia visus ekonomikos sektorius ir viešąjį sektorių. Mokslinių tyrimų (nekomerciniais ir komerciniais tikslais) inovacijas, pagrįstas automatinę duomenų gavybą iš elektroninių nesustruktūrintos informacijos šaltinių (tekstų) ir jų analize, gali stabdyti tik tai, kad teisinė sistema neaiški, o nacionaliniu lygmeniu laikomasi skirtingų požiūrių. Paminėtina, kad nesustruktūrinta informacija sudaro per 80 proc. visos elektroninės informacijos, esančios pasauliniame kompiuterių tinkle ir lokaliuose kompiuterių tinkluose.

7. 2015 m. Jungtinėse Tautose priimta „Darnaus vystymosi darbotvarkė 2030“, kurioje nustatyta 17 universalių, pasaulinių ir tarpusavyje susijusių darnaus vystymosi tikslų, apimančių daugelį politikos sričių. Šių Gairių siekiams ypač aktualus 16 tikslas „Skatinti taikias visuomenės darniam vystymuisi, užtikrinti prieigą prie teisingumo visiems ir sukurti efektyvias, atskaitingas visų lygių institucijas“ ir jame iškeltas uždavinys užtikrinti viešąją prieigą prie informacijos ir

¹ https://en.unesco.org/sites/default/files/eng_-_recommendation_concerning_the_promotion_and_use_of_multilingualism_and_universal_access_to_cyberspace.pdf

² <https://unesdoc.unesco.org/ark:/48223/pf0000227860>

saugoti pagrindines laisves, vadovaujantis nacionaliniais įstatymais ir tarptautinėmis sutartimis (16.10 papunktis); 4 tikslas „Užtikrinti visa apimančią ir lygiavertę kokybišką švietimą ir skatinti visą gyvenimą trunkantį mokymąsi“; 9 tikslas „Kurti atsparią infrastruktūrą, skatinti visa apimančią industrializaciją ir naujovių diegimą“ ir jame išskirti uždaviniai: plėsti mokslinius tyrimus, modernizuoti pramonės sektorių technologinius pajėgumus visose šalyse, visų pirma besivystančiose šalyse, įskaitant ir siekį iki 2030 metų skatinti naujoves ir labai padidinti mokslinių tyrimų ir plėtros srities darbuotojų skaičių 1 milijonui gyventojų bei valstybines ir privačias lėšas, skiriamas moksliniams tyrimams ir plėtrai (9.5 papunktis); labai padidinti galimybes naudotis informacinėmis ir ryšių technologijomis bei siekti užtikrinti visuotinę ir prieinamą prieigą prie interneto mažiausiai išsivysčiusiose šalyse iki 2020 metų (9c papunktis).

8. Valstybės pažangos strategijoje „Lietuvos pažangos strategija „Lietuva 2030““ įtvirtinti pagrindiniai vertybiniai principai, kuriais siekiama visuomenės ir kiekvieno jos nario inovatyvumo, kūrybiškumo, skaitmeninės įtraukties ir kurie padėtų Lietuvai tapti modernia, veržlia, atvira pasauliui, puoselėjančia savo nacionalinį tapatumą šalimi. Strategijoje numatyta efektyviai taikyti informacinių ir ryšių technologijų (toliau – IRT) priemones, užtikrinančias dinamiškai visuomenei būtinų žinių bei gebėjimų įgijimą ir tobulinimą, kurti moderniausias informacines technologijas ir skaitmeninę infrastruktūrą, taip pat naudoti naujausias technologijas teikiant viešąsias paslaugas skaitmeninėje erdvėje. Veiksmingas viešųjų paslaugų teikimas skaitmeninėje terpėje ir žinių visuomenės plėtra neatsiejami nuo visaverčio valstybinės kalbos funkcionavimo informacinėse technologijose, užtikrinančio lygias piliečių dalyvavimo politiniame, socialiniame ir kultūriniame gyvenime galimybes.

9. Valstybinės kalbos politikos 2018–2022 metų gairėse pažymėta, kad pastaraisiais metais pasiekta pastebima lietuvių kalbos pritaikymo skaitmeninei terpei pažanga: parengta nemažai skaitmeninių kalbos išteklių ir pagrindinių kalbos analizės priemonių (morfologinių požymių nustatymo ir generavimo, rašybos tikrinimo įrankių), sukurta sudėtingų internetinės kalbos paslaugų (mašininis vertimas, kirčiuoklė, teksto anotavimas, įvairios paieškos tekstynuose, šnekos atpažintuvas bei sintezatorius ir kt.), sukurta lietuvių kalbos ontologija, lokalizuota nemažai kompiuterinių programų ir įrankių. Kartu sparti informacinių technologijų plėtra visaverčiam lietuvių kalbos funkcionavimui skaitmeninėje terpėje kelia naujų uždavinių (didžiųjų duomenų analizė, mašininio mokymosi ir neuroninių tinklų pritaikymas kalbos analizei, DI technologijų kūrimas su kalba susijusioms paslaugoms, tobulesnis mašininis vertimas ir kt.). Pažymėta, kad reikia užtikrinti galimybę naudotis lietuviška arba sulietuvinta programine ir technine įranga valstybės institucijose ir įstaigose, mokymo ir studijų įstaigose, viešosios prieigos vietose. Taip pat iškeltas lietuviškų skaitmeninių mokymo priemonių poreikis švietimo sistemoje, nes kitakalbės mokymo priemonės silpnina lietuvių kalbos mokėjimo įgūdžius. Sprendžiant šiuos uždavinius reikia ne tik kokybiškai koordinuoti šios srities veiklą, užtikrinti nuolatinį valstybinį finansavimą bei kryptingas investicijas, bet ir rengti kalbos specifiką bei informacines technologijas išmanančius specialistus, finansuoti fundamentinius ir taikomuosius tyrimus, palaikyti mokslines bei technines infrastruktūras.

10. Nuo 2000-ųjų Lietuvoje buvo vykdytos trys programos, skirtos lietuvių kalbos plėtrai informacinėje visuomenėje: 1) Lietuvių kalbos informacinėje visuomenėje 2000–2006 m. programa (patvirtinta Lietuvos Respublikos Vyriausybės 2000 m. balandžio 26 d. nutarimu Nr. 471 „Dėl Lietuvių kalbos informacinėje visuomenėje 2000–2006 metų programos patvirtinimo“), kurią koordinavo Valstybinė lietuvių kalbos komisija, 2) Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programa (patvirtinta Lietuvos Respublikos Vyriausybės 2007 m. kovo 21 d. nutarimu Nr. 319 „Dėl Lietuvių kalbos informacinėje

visuomenėje 2007–2010 metų programos patvirtinimo“), kurios įgyvendinimą koordinavo Švietimo ir mokslo ministerija kartu su Susisiekimo ministerija, 3) Lietuvių kalbos informacinėse technologijose 2014–2020 m. programa (įtraukta į Skaitmeninės darbotvarkės programą), kurią koordinuoja Susisiekimo ministerija.

Įgyvendinant pirmąją nacionalinę Lietuvių kalbos informacinėje visuomenėje 2000–2006 m. programą buvo vykdyti atvirųjų programų lokalizavimo, išteklių kūrimo, automatinio šnekos atpažinimo projektai, gerinama šnekos sintezės kokybė, sukurti kompiuterinis šriftas *Palemonas* bei morfologinės analizės ir generavimo įrankiai, pradėti lietuviškų tekstų sintaksinės ir semantinės analizės darbai.

Lietuvai įstojus į ES, lietuvių kalbos plėtrai informacinėse technologijose didelį postūmį suteikė valstybės sprendimas finansuoti skaitmenines lietuvių kalbos paslaugas ES struktūrinių fondų lėšomis.

Įgyvendinant antrąją Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą, buvo tobulinami esami ir kurti nauji kalbos ištekliai, tobulinamos automatinio šnekos atpažinimo ir sintezės technologijos, kuriami nauji mašininio vertimo įrankiai, gerintos ir kurtos semantinės analizės ir informacijos paieškos priemonės, sukurtas interneto portalas „Raštija.lt“, kuriame būtų galima nemokamai naudotis kalbos ištekliais ir technologijomis.

Šiose dviejose programose aktyviausiai dalyvavo keturios mokslo ir studijų institucijos: Vilniaus universitetas, Vytauto Didžiojo universitetas, Kauno technologijos universitetas ir Lietuvių kalbos institutas. Mokslo ir studijų institucijos projektus vykdė kartu su verslo įmonėmis ir kitais partneriais.

11. Informacinės visuomenės plėtros 2014–2020 m. programoje „Lietuvos Respublikos skaitmeninė darbotvarkė“, patvirtintoje Lietuvos Respublikos Vyriausybės 2014 m. kovo 12 d. nutarimu Nr. 244 „Dėl Informacinės visuomenės plėtros 2014–2020 metų programos „Lietuvos Respublikos skaitmeninė darbotvarkė“ patvirtinimo“ (toliau – Programa), yra numatytas trečiasis tikslas „puoselėti IRT priemonėmis Lietuvos kultūrą ir lietuvių kalbą – kurti visuomenės poreikius atitinkantį kultūrinį ir lietuvių rašytinės ir sakytinės kalbos sąsajomis pagrįstą skaitmeninį turinį, plėtoti skaitmeninius produktus ir elektronines paslaugas“. Antrasis šio tikslo uždavinys yra „kurti ir plėtoti viešai prieinamus lietuvių kalbos ir rašties skaitmeninius išteklius, diegti juos į IRT ir elektronines paslaugas“. Per 2014–2020 m. laikotarpį iš ES fondų investicijų veiksmų programos 2 prioriteto „Informacinės visuomenės skatinimas“ 02.3.1-CPVA-V-527 priemonės „Lietuvių kalba informacinėse technologijose“ Nr. 02.3.1-CPVA-V-527 buvo skirtas finansavimas lietuvių kalbos sprendiniams skaitmeninėje erdvėje įgyvendinti. 2018 m. pradėti vykdyti 5 projektai: *Lietuvių šneka valdomų paslaugų plėtra („LIEPA-2“)*, *Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinės sistemos viešųjų paslaugų vystymas („Semantika-2“)*, *Mašininio vertimo sistemų ir lokalizavimo paslaugų tobulinimas ir plėtra*, *Integruotų lietuvių kalbos ir rašties išteklių informacinės sistemos plėtra („Raštija 2“)* ir *Lietuvių kalbos išteklių informacinės sistemos plėtra („E. kalba“)*. Iki 2020 m. pabaigos bus sukurta 21 vieša elektroninė paslauga. Šiai priemonei finansuoti iš viso buvo skirta 14 310 635,00 eurai.

Šioje programoje aktyviausiai dalyvavo trys mokslo ir studijų institucijos: Vilniaus universitetas, Vytauto Didžiojo universitetas ir Lietuvių kalbos institutas. Mokslo ir studijų institucijos projektus vykdė kartu su verslo įmonėmis.

12. Lietuvos mokslo taryba (toliau – LMT) yra pagrindinė mokslinius tyrimus finansuojanti institucija Lietuvoje, vykdanči valstybės mokslo konkursinį finansavimą. 2014–2020 m. laikotarpio moksliniai tyrimai, susiję su lietuvių kalbos plėtra skaitmeninėje terpėje,

galėjo būti finansuojami įvairiose nacionalinėse ir tarptautinėse LMT programose. Atsižvelgiant į tematiką, lietuvių kalbos projektai skaitmeninėje terpėje turėjo daugiau galimybių būti finansuojami pagal Lituanistikos 2016–2024 m. programą, mokslininkų grupių projektų finansavimo priemonę bei tarptautinio bendradarbiavimo programą COST.

Lituanistikos 2016–2024 m. programoje iki 2020 m. įvyko septyni kvietimai, kurių metu 268 lituanistinių tyrimų projektams buvo skirtas 9,7 mln. eurų finansavimas. Kalbos technologijų projektų dalis šioje programoje yra nedidelė: buvo finansuoti tik du Vytauto Didžiojo universiteto kalbos technologijų mokslo projektai, kurių bendras biudžetas 186 tūkst. eurų (tai 1,90 proc. viso programos biudžeto).

Mokslininkų grupių projektų konkursai vyksta nuo 2010 m. Vertinant 2014–2020 m. laikotarpį, humanitarinių ir socialinių mokslų sričiai V–IX kvietimuose 154 projektams buvo skirtas 20,2 mln. eurų finansavimas, iš jų trims lietuvių kalbos plėtros skaitmeninėje erdvėje projektams skirta 390 tūkst. eurų (1,92 proc.). Finansavimą gavusios institucijos: Vilniaus universitetas, Vytauto Didžiojo universitetas ir Mykolo Romerio universitetas.

Lietuvos mokslininkai aktyviai dalyvauja COST (*European Cooperation in Science & Technology*) asociacijos bendradarbiavimo ir mobilumo veiklose. 2014–2020 m. laikotarpiu kalbos technologijoms skirtose COST³ veiklose dalyvavo Vilniaus universitetas (IC1406), Vytauto Didžiojo universitetas (IC1408, IC1207, CA18209), Kauno technologijų universitetas (IC1002) ir Mykolo Romerio universitetas (CA18209).

13. 2015 m. parengtas naujas mokslinių tyrimų infrastruktūrų (MTI) kelrodis. Jame pateikiami Lietuvai aktualių Europos mokslinių tyrimų infrastruktūrų, įtrauktų į Europos mokslinių tyrimų infrastruktūrų strategijos forumo (ESFRI) parengtą Europos MTI kelrodį, taip pat kitų tarptautinių mokslinių tyrimų infrastruktūrų atitikmenys, kuriuose tikslinga siekti Lietuvos narystės. Į 2015 m. kelrodį įtrauktas CLARIN-LT konsorciumas, kuris atstovauja Lietuvai europinėje CLARIN ERIC infrastruktūroje. Lietuvių kalbos plėtrai skaitmeninėje erdvėje aktualios ir kitos dvi humanitarinių ir socialinių mokslų MTI – LIDA (humanitarinių ir socialinių mokslų duomenų archyvas) bei PITI *Aruodai*.

14. Lietuva turi visas prielaidas tapti skaitmeninių technologijų inovacijų, įskaitant inovacijas kalbos technologijų srityje, lydere. Todėl Lietuva su kitomis ES narėmis pasirašė DI bendro vystymo Deklaraciją⁴. 2019 m. Ekonomikos ir inovacijų ministerija parengė ir deklaravo Lietuvos DI vystymo strategiją⁵, kurioje pabrėžta kalbos technologijų, kaip vienos svarbiausių DI technologijų, plėtros svarba šalies ūkiui ir verslui bei mokslui. Ši iniciatyva bus įgyvendinama 2021–2027 m.

Europos Komisija 2020 m. vasario 20 d. paskelbė baltąją knygą „Dirbtinis intelektas. Europos požiūris į kompetenciją ir pasitikėjimą“⁶, kurioje pateikiamos Dirbtinio intelekto technologijų, Europos duomenų ir susijusių technologijų raidos ir reguliavimo strateginės gairės. Taip pat Europos Komisija 2020 m. vasario 19 d. paskelbė Europos duomenų strategiją⁷.

15. ES struktūrinių fondų lėšomis finansuojamos priemonės, skirtos paskatinti įmones investuoti į inovaciniams gaminiams, paslaugoms ar procesams kurti reikalingus mokslinius tyrimus bei eksperimentinę plėtrą (toliau – MTEP), taip pat paskatinti įmonių plėtrą ir naujų

³ <https://www.cost.eu/>

⁴ <https://ec.europa.eu/jrc/communities/en/node/1286/document/eu-declaration-cooperation-artificial-intelligence>

⁵ https://ec.europa.eu/knowledge4policy/publication/lithuanian-artificial-intelligence-strategy_en

⁶ <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence>

⁷ <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy>

inovacinių įmonių steigimąsi investuojant į MTEP ir inovacijų infrastruktūros kūrimą ir plėtrą (priemonės „Inočekiai“, „Intelektas“, „Eksperimentas“). Šiose priemonėse svarbius kalbos technologijų projektus vykdo Baltijos pažangių technologijų institutas, Vilniaus universitetas, Vytauto Didžiojo universitetas, bendradarbiaudami su UAB „Amberlo“, UAB „Tilde informacinės technologijos“ ir kitomis įmonėmis.

16. Rengiamas 2021–2030 metų nacionalinis pažangos planas, kuriuo siekiama nustatyti pagrindinius ateinantį dešimtmetį valstybėje siekiamus pokyčius, užtikrinančius pažangą socialinėje, ekonominėje, aplinkos ir saugumo srityse. 2021–2030 metų nacionalinis pažangos planas pakeičia 2014–2020 metų nacionalinės pažangos programą. Jis užtikrina ne tik strateginio planavimo tęstinumą, bet ir kokybinį strateginio valdymo Lietuvoje pokytį. Siekiant įgyvendinti strateginio valdymo reformą, 2021–2030 metų nacionalinis pažangos planas tampa pagrindiniu valstybės pokyčių planavimo dokumentu, kuriame įvertinamos valstybės finansinės galimybės šiuos pokyčius įgyvendinti, integruojant ES, kitų tarptautinių šaltinių bei papildomas nacionalinio biudžeto lėšas. Tęstinė veikla, nepatenkanti į šį planą, bus įgyvendinama ir finansuojama tęstinės veiklos lėšomis. Plane numatytiems pokyčiams įgyvendinti rengiamos nacionalinės plėtros programos. Viena iš tokių programų – Ekonomikos ir inovacijų ministerijos rengiama Skaitmeninio plėtros 2021–2030 m. programa, kuri pakeis Skaitmeninės darbotvarkės programą. Rengiant Lietuvos skaitmeninio plėtros 2021–2030 m. programą, daugiausia dėmesio bus skiriama inovatyvių ir vartotojams patrauklių viešųjų skaitmeninių sprendinių kūrimui ir plėtrai, pasitelkiant DI technologijas, kurios leistų piliečiams ir verslui bendrauti valstybine kalba (rašytine ir sakytine), sukurti lietuvių kalbai specifinius DI technologijų sprendinius, pritaikytus lietuvių kalbos gramatikos ir leksinės semantikos ypatumams.

17. 2021–2027 m. Skaitmeninės Europos programos projekte numatyta, kad viešojo administravimo institucijų ir paslaugų modernizavimas skaitmeninėmis priemonėmis yra itin svarbus, siekiant sumažinti administracinę naštą verslui ir piliečiams, todėl būtina pagreitinti jų sąveiką su valdžios institucijomis, padaryti ją patogesnę ir pigesnę, taip pat padidinti piliečiams ir verslo subjektams teikiamų paslaugų efektyvumą ir kokybę, sudaryti galimybę šias paslaugas gauti gimtąja kalba. Kadangi tam tikros viešosios paslaugos jau dabar teikiamos ES lygiu, ir tokių paslaugų tik daugės, turėtų būti užtikrinta, kad piliečiai ir verslo subjektai galėtų gauti aukštos kokybės skaitmenines paslaugas visoje Europoje.

18. Skaitmeninė erdvė turi milžinišką potencialą palengvinti informacijos ir viešųjų paslaugų pasiekiamumą, tačiau jeigu neprisitaikoma prie visų vartotojų galimybių ir poreikių, užprogramuojama socialinė atskirtis, kurios padarinius jaus visa visuomenė. Europos Komisija yra įsipareigojusi užtikrinti vienodas galimybes visiems gyventojams, o ypač neįgaliesiems, ir siekia, kad skaitmeninė erdvė jiems taptų labiau pritaikyta ir įtrauki. Lietuva taip pat siekia, kad visi gyventojai turėtų galimybę visavertiškai naudotis skaitmeniniais sprendiniais, todėl itin svarbi jų pritaikymo neįgaliesiems politika.

Atsižvelgtina ir į tai, kad 2016 m. spalio 26 d. Europos Parlamento ir Tarybos direktyva (ES) 2016/2102 dėl viešojo sektoriaus institucijų interneto svetainių ir mobiliųjų programų prieinamumo. Ši direktyva įpareigoja valstybes nares, taip pat ir Lietuvą, pritaikyti visas viešojo sektoriaus interneto svetaines neįgaliesiems ir specialiųjų poreikių turintiems piliečiams. Taigi lietuvių kalba valdomi sprendiniai turi būti integruoti ir į Lietuvos valstybės viešojo sektoriaus institucijų interneto svetaines. Itin aktualūs tampa lietuvių šneka valdomi skaitmeniniai sprendiniai, pavyzdžiui, neįgaliesiems būtinas lietuvių kalbos sintezatorius (skaitytuvas), leidžiantis kompiuteriu skaityti lietuviškus tekstus ir naršyti internete.

19. EFNIL (angl. *European Federation of National Institutions for Language*)⁸ – 2003 m. įsteigta Europos nacionalinių kalbos institucijų federacija, vienijanti ES valstybių narių svarbiausių kalbos organizacijų ir kitų nacionalinių kalbos institucijų atstovus (atstovai Lietuvoje – Lietuvių kalbos institutas ir Valstybinė lietuvių kalbos komisija). Federacija ypač daug dėmesio skiria nacionalinėms ES kalboms ir Europos kalbų įvairovei. Šios federacijos tikslai itin susiję su kalbos technologijų (ir DI technologijų) kūrimu bei plėtra.

20. Kuriant kalbos technologijas ir su kalba susijusias paslaugas, labai svarbios yra privataus verslo iniciatyvos. Didžiulę įtaką kalbos technologijų plėtrai pirmiausia daro stambiaus tarptautinio verslo (pavyzdžiui, *Google, Microsoft, Facebook, Amazon, IBM* ir kt.) iniciatyvos, taip pat duomenų rinkimo iniciatyvos, tokios kaip *Mozilla Common Voice* ir *Glosbe*. Didžiausios pasaulio kalbos technologijų bendrovės buriasi į *LT-Innovate*⁹ kalbos technologijų pramonės asociaciją, kurioje dalijamasi idėjomis bei kuriami strateginiai sprendiniai.

Iš lietuviško verslo iniciatyvų paminėtini UAB „TokenMill“ atvirai prieinami lietuviški įrankiai bei sulietuvinta natūraliosios kalbos apdorojimo *SpaCy* platforma, taip pat UAB „Tilde informacinės technologijos“ sukurtos lietuvių šnekos atpažinimo ir mašininio vertimo demonstracijos internete.

21. ELRC (angl. *European Language Resource Coordination*)¹⁰ – Europos kalbų išteklių koordinavimo iniciatyva, sukurta siekiant surinkti kalbos išteklius mašininio vertimo sistemoms, kurias naudos viešųjų paslaugų teikėjai visose ES valstybėse narėse, taip pat Islandijoje bei Norvegijoje. Ji padės aprūpinti mašininio vertimo platformą CEF.AT kalbos ir vertimo duomenimis (vienkalbiais ir dvikalbiais), susijusiais su kasdiene Europos šalių institucijų veikla.

22. CLARIN ERIC (angl. *Common Language Resources and Technology Infrastructure*)¹¹ yra šalių ir tarpvyriausybinių organizacijų konsorciumas, kuris rūpinasi atvirąja prieiga prie skaitmeninių kalbos duomenų ir technologijų visoje Europoje ir už jos ribų. CLARIN-LT yra Lietuvos nacionalinis konsorciumas, koordinuojamas Vytauto Didžiojo universiteto, kuris yra CLARIN ERIC narys nuo 2014 m. spalio 25 d.

23. ELG (angl. *European Language Grid*)¹² – ES finansuojamas projektas, pagal kurį 2019–2021 m. kuriamas Europos kalbų tinklas – platforma (angl. *scalable cloud platform*), suteiksianti prieigą prie visų Europos kalbų komercinių ir nekomercinių kalbų technologijų paslaugų ir duomenų rinkinių, įskaitant šimtus veikiančių įrankių ir paslaugų, taip pat tūkstančius duomenų rinkinių ir išteklių. Palaikant technologijomis pagrįstą daugiakalbystę, daugiausia dėmesio skiriama Europai, kurioje yra 24 oficialios ES valstybių narių kalbos, 60 neoficialių ar mažų kalbų, taip pat vartojamos imigrantų ir svarbių prekybos partnerių kalbos. Europos kalbų tinklą sudaro ir uždavinius įgyvendina 32 nacionaliniai kompetencijos centrai (nuo 2019 m. Lietuvoje toks centras – Lietuvių kalbos institutas).

24. Europos Komisijos CEF (angl. *Connecting Europe Facility*)¹³ programa finansuoja bendrųjų ir daugkartinio naudojimo skaitmeninių paslaugų infrastruktūrą (DSI), dar vadinamą statybiniais blokais, rinkinį. CEF programoje kuriamus pagrindinius struktūrinius blokus galima

⁸ <http://www.efnil.org/>

⁹ <http://lt-innovate.org>

¹⁰ <http://www.lr-coordination.eu/>

¹¹ <http://clarin.eu>

¹² <https://www.european-language-grid.eu/>

¹³ <https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL>

pakartotinai naudoti įgyvendinant bet kurį Europos projektą, siekiant palengvinti skaitmeninių viešųjų paslaugų teikimą nepaisant sienų ir įvairiuose sektoriuose. Šiuo metu yra aštuoni pagrindiniai blokai: „Big Data Test Infrastructure“, „Context Broker“, „eArchiving“, „eDelivery“, „eID“, „eInvoicing“, „eSignature“ ir „eTranslation“. Pastaroji paslauga – moderniausia iš visų mašininio vertimo paslaugų internete; ji yra nemokama, skirta viešojo sektoriaus institucijoms ir leidžia versti dokumentus iš bet kurios oficialios ES kalbos į bet kurią kitą ES oficialią kalbą. „eTranslation“ garantuoja visų išverstų duomenų konfidencialumą ir saugumą.

25. „Gimtosios kalbos projektas“ (angl. *Human Language Project (HLP)*)¹⁴ yra didelio masto ilgalaikė mokslinių tyrimų (taikomųjų ir fundamentinių), plėtros ir inovacijų programa, kurioje inovacijos ir komercializavimas glaudžiai sąveikauja, siekiant maksimaliai padidinti kalbų technologijų poveikį Europos ekonomikai ir visuomenei. 2017 m. buvo atlikta išsami studija, kurioje, remiantis dabartinės padėties analize, pateikiama argumentų, kodėl reikėtų imtis šios daugiadalykės, plataus masto koordinuojamos iniciatyvos¹⁵.

26. „Europos atvirojo mokslo debesija“ (angl. *European Open Science Cloud (EOSC)*). Šią iniciatyvą, kaip Europos debesijos iniciatyvos dalį, 2016 m. pasiūlė Europos Komisija, siekdama sukurti konkurencingą duomenų ir žinių ekonomiką Europoje. 2016 ir 2017 m. vyko išsamos konsultacijos su suinteresuotomis mokslo ir studijų institucijomis, 2017 m. birželio mėn. įvykusiame pirmajame EOSC aukščiausiojo lygio susitikime EOSC deklaracijai pritarė daugiau nei 70 institucijų.

27. FAIR (angl. *Findable, Accessible, Interoperable, Re-usable*; liet. „atrandami, pasiekiami, sąveikūs, pakartotinai naudojami“) – duomenų rinkimo iniciatyva, kurią 2016 m. paskelbė mokslininkų ir organizacijų konsorciumas¹⁶. Šiuo metu daugelis duomenų platformų deklaruoja savo duomenų valdymo plano atitiktį FAIR principams.

III SKYRIUS

LIETUVIŲ KALBOS TECHNOLOGIJŲ BŪKLĖ IR PLĖTROS GALIMYBĖS

28. Vertinant lietuvių kalbos technologijų būklę ir jų plėtros galimybes, svarbu paminėti pagrindines šioje srityje dirbančias mokslo ir studijų institucijas, taip pat verslo subjektus, apžvelgti jų atliktus darbus ir įvardyti tolesnių darbų poreikį, atsiremiant į Lietuvių kalbos plėtros informacinėse technologijose 2014–2020 m. gaires.

28.1. Svarbiausios mokslo ir studijų institucijos, prisidedančios prie lietuvių kalbos plėtros skaitmeninėje erdvėje, yra šios¹⁷:

28.1.1. Baltijos pažangiųjų technologijų institutas (BPTI) – dalyvavo mokslo projektuose „Intelektas LT“, „Inočekiai“ ir LMT finansuojamuose projektuose, bendradarbiauja su mokslo ir studijų institucijomis, taiko kalbos technologijas gynybos ir saugumo srityje.

¹⁴ <http://human-language-project.eu/>

¹⁵ *Kalbų lygybė skaitmeniniame amžiuje. Gimtosios kalbos projektas*. Europos Parlamento tyrimų tarnyba. Mokslinių perspektyvų tyrimo skyrius (STOA). Briuselis. 2017, 166 p.

¹⁶ Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*. 3: 160018. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175>

¹⁷ Pateikiamos abėcėlės tvarka.

28.1.2. Kauno technologijos universitetas (KTU) – dalyvauja ES struktūrinių fondų, „Inočekių“ bei COST projektuose.

28.1.3. Lietuvių kalbos institutas (LKI) – dalyvauja ES struktūrinių fondų, Europos Komisijos ir LMT mokslo projektuose.

28.1.4. Vilniaus universitetas (VU) – dalyvauja ES struktūrinių fondų ir LMT mokslo projektuose, „Inočekių“ bei COST projektuose.

28.1.5. Vytauto Didžiojo universitetas (VDU) – dalyvauja ES struktūrinių fondų, „Inočekių“, COST ir LMT mokslo projektuose.

28.2. Svarbiausios privataus verslo bendrovės, prisidedančios prie lietuvių kalbos plėtros skaitmeninėje erdvėje, yra šios¹⁸:

28.2.1. UAB „Algoritmų sistemos“ – informacinių sistemų kūrimo ir diegimo įmonė, turinti patirties plėtojant šnekos technologijų sprendinius. Bendradarbiaudama su VU, projekte „LIEPA“ yra sukūrusi lietuvių kalbos šnekos sintezės ir įvairių su šneka susijusių paslaugų sprendinius.

28.2.2. UAB „Amberlo“ – specializuojasi kurti daugiakalbių, DI ir kalbos technologijomis grįstų teisinių paslaugų sprendinius.

28.2.3. UAB „ATEA“ – lygmenyse „verslas–verslui“, „verslas–viešajam sektoriui“, „viešasis sektorius–verslui“ ir „mokslas–verslui“, „verslas–mokslui“ specializuojasi kurti ir įgyvendinti kalbos technologijų infrastruktūrinius sprendinius. Bendradarbiaudama su VDU vykdo projektus „Semantika-1“ ir „Semantika-2“.

28.2.4. UAITB „Fotonija“ – lygmenyse „verslas–verslui“, „verslas–viešajam sektoriui“, „viešasis sektorius–verslui“ ir „mokslas–verslui“, „verslas–mokslui“ teikia skaitmeninių tekstų automatinės analizės sprendinių kūrimo ir pritaikymo bei kalbos technologijų duomenų rinkinių formavimo paslaugas.

28.2.5. UAB „Leksinova“ ir UAB „Lexnet“ – specializuojasi elektroninių teisinių dokumentų automatinės analizės ir DI bei kalbos technologijomis grįstų teisinių paslaugų teikimo srityse.

28.2.6. UAB „NETCODE“ – programinius sprendinius kurianti įmonė, turinti didelę patirtį dirbti su lietuvių kalbos skaitmeniniais ištekliais, juos apdoroti ir atverti visuomenei.

28.2.7. UAB „Proit“ – yra sukaupusi didelę patirtį kalbos technologijų plėtros projektuose, tokiuose kaip skaitmeninių kalbos išteklių sistemų bei elektroninių kalbos paslaugų, skirtų verslui, kūrimas.

28.2.8. MB „Tetragrama“ – lygmenyse „verslas–verslui“, „verslas–viešajam sektoriui“, „viešasis sektorius–verslui“ ir „mokslas–verslui“, „verslas–mokslui“ teikia skaitmeninių tekstų automatinės analizės sprendinių kūrimo ir pritaikymo bei kalbos technologijų duomenų rinkinių formavimo paslaugas.

28.2.9. UAB „Tilde informacinės technologijos“ – aktyviai dalyvauja ES struktūrinių fondų projektuose. Lygmenyse „verslas–verslui“, „verslas–viešajam sektoriui“, „viešasis sektorius–verslui“ ir „verslas – mokslui“, „mokslas–verslui“ teikia DI grįstas kalbos technologijų (šnekos atpažinimo, sintezės, mašininio vertimo, virtualių asistentų ir kt.) plėtros ir lokalizavimo paslaugas.

28.2.10. UAB „Tokenmill“ – lygmenyse „verslas–verslui“, „verslas–viešajam sektoriui“, „viešasis sektorius–verslui“ ir „mokslas–verslui“, „verslas – mokslui“ teikia skaitmeninių tekstų automatinės analizės sprendinių kūrimo ir pritaikymo bei kalbos

¹⁸ Pateikiamos abėcėlės tvarka.

technologijų duomenų rinkinių formavimo paslaugas. Specializuojasi DI technologijų, medijų stebėsenos, natūraliosios kalbos apdorojimo ir supratimo, kalbos generavimo srityse.

29. Lietuvių kalbos technologinės erdvės būklę galima įvertinti pagal septynias technologijų sritis: nacionalinės kalbos technologijų ir duomenų infrastruktūros, kalbos duomenys ir duomenų rinkiniai, mašininis vertimas ir lokalizavimas, šnekos technologijos, natūraliosios kalbos apdorojimas, natūraliosios kalbos supratimas ir natūraliosios kalbos generavimas.

29.1. Nacionalinės kalbos technologijų ir duomenų infrastruktūros.

29.1.1. „Raštija LT“ – Vilniaus universiteto Matematikos ir informatikos instituto sukurta „Integruotų lietuvių kalbos ir raštijos išteklių informacinė sistema“ (<http://www.raštija.lt>): žinių bazė, paieškos įrankiai, integruoti antrosios bei trečiosios nacionalinių programų projektai.

29.1.2. CLARIN-LT – 2015 m. kovo 31 d. įkurtas Lietuvos nacionalinis konsorciumas (CLARIN ERIC narys), kurį šiuo metu sudaro 5 mokslo institucijos: Vytauto Didžiojo universitetas (koordinadorius), Kauno technologijos universitetas, Vilniaus universitetas, Mykolo Romerio universitetas ir Baltijos pažangiųjų technologijų institutas.

29.1.3. Valstybinės kalbos technologijų informacinės sistemos:

29.1.3.1. LKSSAIS („Semantika“) (www.semantika.lt) (vykdytojas Vytauto Didžiojo universitetas, partneris Kauno technologijos universitetas). Projekto „Semantika-1“ metu sukurta informacinė sistema, kuri teikia šešias kalbos technologijų paslaugas (lygmenys „žmogus–mašina“ ir „mašina–mašina“): naršymas tekstynuose, tekstų analizė, automatinis rašybos klaidų taisymas. Lankomumas – per 300 000 panaudos atvejų per metus. Projekto „Semantika-2“ metu sistema modernizuota: visa apimtimi veikia debesijos technologijų pagrindu (mikroservisai, tinklinės paslaugos, paskirstytosios sistemos principas ir t. t.), teikia modernizuotas kalbos technologijų paslaugas (fonogramų transkribavimas tekstu, automatinis santraukų sudarymas, rašybos klaidų taisymas, tekstų analizė, socialinės medijos tekstų analizė, išplėstinė paieška tekstuose), įtraukia *Docker* repozitoriumą, iš kurio vartotojai gali patogiai parsisiųsti projekto dalyvių sukurtus atvirojo kodo IT sprendinius, saugomus *Docker* kontaineriuose.

29.1.3.2. Lietuvių kalbos išteklių informacinė sistema (Lietuvių kalbos institutas). Pagal programos „Lietuvių kalba informacinėje visuomenėje“ projektą „IRT sprendimų bei turinio, padedančių išsaugoti lietuvių kalbą viešojoje erdvėje, kūrimas bei galimybių naudotis jais sudarymas“ Lietuvių kalbos institutas kartu su Lietuvių literatūros ir tautosakos institutu, Vilniaus universitetu ir tuomečiu Lietuvos edukologijos universitetu sukūrė ir visuomenei 2015 m. pristatė Lietuvių kalbos išteklių informacinę sistemą (<http://lkiis.lki.lt/>). Pastaraisiais metais Lietuvių kalbos institutas šią sistemą modernizuoja ir pildo naujais sprendiniais (2018–2020 m. vykdomas projektas „Lietuvių kalbos išteklių informacinės sistemos plėtra (E. kalba)“, finansuojamas iš ES struktūrinių fondų). Lietuvių kalbos išteklių informacinė sistema „E. kalba“ (<http://lkiis.lki.lt/> >> <http://ekalba.lt>) apima 19 vienakalbių bei dvikalbių žodynų, 10 įvairių kalbos bei tautosakos duomenų bazių ir kartotekų išteklių („Lietuvių kalbos žodyno“ kartotekos, Mįslių kartoteka, Pokario partizanų dainų kartoteka, Liaudies tikėjimų kartoteka, Lietuvos vietovardžių geoinformacinė duomenų bazė, Istorinių vietovardžių duomenų bazė, Pavardžių duomenų bazė, Tarmių archyvas), 9 elektronines paslaugas (pavyzdžiui, „Paieška žodžių prasmių tinkle“, „E. sąvokos“, „E. pavadinimas“, „Nuomonių analizė“, „Žodžių darybos vedlys“, „Kalbos patarimai“, „E. mokymai ir kalbos žaidimai“ ir kt.).

29.1.4. Vertinant nacionalinių kalbos technologijų ir duomenų infrastruktūrų aktualumą ir perspektyvas, būtina užtikrinti nuolatinę jų plėtrą šiomis kryptimis: 1) pildyti infrastruktūras

reikalingais skaitmeniniais ištekliais (pavyzdžiui, regioninių atmainų bei senųjų raštų kalbos duomenimis, kalbos lauką išplėsti istorijos, geografijos, kultūros paveldo ištekliais, kaupti trūkstamus kalbos duomenų išteklius (žr. 29.2.7)); 2) atnaujinti infrastruktūrų techninę įrangą bei užtikrinti jos palaikymą; 3) integruoti infrastruktūras į stambesnes nacionalines, Europos ir tarptautines kalbų išteklių sistemas; 4) užtikrinti infrastruktūrose saugomų technologijų ir duomenų atvirumą. Be to, svarbu skatinti nacionalinių duomenų infrastruktūrų kūrėjų bendradarbiavimą tarpusavyje ir specializavimąsi tam tikroje srityje.

29.2. Kalbos duomenys ir duomenų rinkiniai.

29.2.1. Tekstynai.

29.2.1.1. Vytauto Didžiojo universitete vykdant keletą projektų toliau plėtoti šie tekstynai: Dabartinės lietuvių kalbos tekstynas; Lietuviško interneto socialinių tekstų tekstynas LITIS; Bendrasis interneto (žiniasklaidos) tekstynas (BIT); Sakytinės lietuvių kalbos tekstynas, vadovėlių tekstų rinkinys KLASIUS; ekspertų rankiniu būdu morfologiškai anotuotas tekstynas („auksinis standartas“) MATAS; ekspertų rankiniu būdu sintaksiškai anotuotas tekstynas („auksinis standartas“) ALKSNIS. Daugumos tekstynų duomenims yra užtikrinta atviroji prieiga.

29.2.1.2. Lietuvių kalbos institute kuriami ir pildomi šie tekstynai¹⁹: Tarmių tekstynas, Senosios lietuvių kalbos tekstynas, Senųjų raštų duomenų bazė, Moderniosios tapatybės ideologinio naratyvo tekstynas ir kt. Vis dar yra daugybė nesuskaitmenintų tarminių išteklių, kurie turėtų būti pritaikyti naudotis tarpregioniniu ir tarpvalstybiniu lygiu, pavyzdžiui, turizmo tikslais. Būtina užtikrinti išteklių gausinimą ir integraciją į bendrą Lietuvių kalbos išteklių informacinę sistemą „E. kalba“ bei kitas pasaulines (ar europines) kalbų išteklių sistemas.

29.2.2. Garsynai.

29.2.2.1. Vilniaus universiteto kuriamas lietuvių šnekos garsynas „LIEPA“ – fonetiškai reprezentatyvi lietuvių šnekos duomenų bazė, pritaikyta šnekos technologijų moksliniams tyrimams ir konstravimo darbams, elektroninėms paslaugoms teikti²⁰. Šiuo metu vykdomas projektas „LIEPA-2“, kuriuo planuojama esamą garsyną papildyti iki 1000 valandų.

29.2.2.2. Vytauto Didžiojo universitetas ir Kauno technologijos universitetas (atskiri juose dirbantys tyrėjai), vykdydami įvairius taikomuosius projektus, turi sukaupę garsynų, kurių apimtys svyruoja nuo dešimties iki šimtų valandų. Šie duomenys dažniausiai surinkti be pateikėjų licencijų (sutikimų), todėl yra uždari. Šie garsynai yra naudojami įvairioms šnekos atpažinimo priemonėms kurti, pavyzdžiui, lietuviškų fonogramų (šnekos failų) transkripcijos tekstu sprendiniui („Semantika-2“ projekte).

29.2.3. Skaitmeniniai žodynai ir vertimo atmintys.

29.2.3.1. Lietuvių kalbos instituto interneto svetainėje (<http://lkiis.lki.lt/> ir <http://lki.lt/skaitmeniniai-lietuviu-kalbos-istekliai/>) visuomenei prieinami šie skaitmeniniai žodynai: 9 vienakalbiai („Bendrinės lietuvių kalbos žodynas“, „Dabartinės lietuvių kalbos žodynas“, „Lietuvių kalbos žodynas“, Lietuvių kalbos naujažodžių duomenynas, „Sinonimų žodynas“, „Antonimų žodynas“, „Frazeologijos žodynas“, „Palyginimų žodynas“, „Sisteminis lietuvių kalbos žodynas“), 12 dvikalbių žodynų (šių kalbų: lietuvių–latvių, latvių–lietuvių, lietuvių–vokiečių, vokiečių–lietuvių, lietuvių–lenkų, lenkų–lietuvių, lietuvių–anglų, anglų–lietuvių, lotynų–lietuvių, senovės graikų–lietuvių, lietuvių–vengrų, baltarusių–lietuvių), iš kurių didžioji dalis – suskaitmeninti popierinių knygų variantai. Nuolat pildomi ir rašomi, neturintys popierinio formato, yra šie vienakalbiai internetiniai žodynai: „Bendrinės lietuvių kalbos

¹⁹ <http://lki.lt/skaitmeniniai-lietuviu-kalbos-istekliai/>

²⁰ <https://www.xn--ratijs-ckb.lt/liepa/infrastrukt%C5%ABrin%C4%97s-paslaugos/garsynas-liepa/7569>

žodynas“ ir Lietuvių kalbos naujažodžių duomenynas. Atsižvelgiant į visuomenės poreikius, būtina atnaujinti sinonimų, antonimų, frazeologijos žodynus, taip pat suskaitmeninti tarties, junglumo, enciklopedinius ir kt. žodynus ir užtikrinti, kad vieno langelio principu būtų prieinama kuo daugiau skaitmeninių išteklių.

29.2.3.2. Vilniaus universiteto projekte „Visuomenei aktualios programinės įrangos lokalizavimas, programoms reikalingų priemonių sukūrimas“ sukurti žodynai ir vertimo atmintys viešai prieinami per informacinę sistemą „Raštija“²¹.

29.2.3.3. Mokslo ir enciklopedijų leidybos centras sukūrė ir nuolat naujina „Visuotinės lietuvių enciklopedijos“ (VLE) suskaitmenintą versiją²².

29.2.4. Geoinformaciniai kalbos duomenys.

Lietuvių kalbos institute kuriama Lietuvos vietovardžių geoinformacinė duomenų bazė. Ji integruota į Lietuvių kalbos išteklių informacinę sistemą „E. kalba“ ir nuolat plečiama. Numatyta, kad ši bazė apims Lietuvos Respublikos savivaldybių teritorijų lingvistinius-geografinius duomenis (įvairių geografinių objektų – gyvenamųjų vietų, hidrografinių objektų, susisiekimo objektų, istorinių-kultūrinių objektų, žemės dangos objektų ir kitus pavadinimus, susietus su tiksliais jų koordinatėmis žemėlapyje), taip pat šių objektų fiksavimo istoriniuose šaltiniuose, kilmės ir kt. ypatybes, papildant šiuos duomenis garsine bei vaizdine informacija. Būtina nuolatinė duomenų bazės plėtra, tiek didinant geografinę aprėptį, tiek įtraukiant daugiau geografinių objektų, tačiau šis kalbininkų ir kartografijos specialistų pradėtas itin didelės apimties ir nacionalinės svarbos darbas vyksta su pertrūkiais, nes neužtikrinamas nuolatinis finansavimas.

29.2.5. Ontologijos.

29.2.5.1. Vytauto Didžiojo universiteto vykdomo projekto „Semantika-1“ metu sukurta *Bendroji Lietuvių kalbos ontologija*, ji keliuose kituose projektuose išplėtota ir transformuota į *LitWordNet*. Šiuo metu vyksta *LitWordNet*o susiejimo su anglų kalbos *WordNet* darbai.

29.2.5.2. Sukurta lietuviškų medicinos terminų ontologija „Snomed CT“²³ – atvirosios prieigos, atviras duomenų rinkinys.

29.2.5.3. Lietuvių kalbos institute baigiama kurti elektroninė paslauga „E. sąvokos“, kuri apima šių sričių ontologijas: medicinos (žmogaus anatomija), finansų (ekonomikos terminai) ir informacinių technologijų (kompiuterinė technika ir jos dalys). Numatoma sudaryti galimybę kurti ir tvarkyti šių sričių ontologijas panaudojant lietuvių kalbos prasminio žodžių tinklo duomenis bei integruojant kitų kalbų žodžių tinklus, taip praturtinant sąvokos semantinę aplinką. Būtina užtikrinti ontologijų gausinimą ir kitų sričių ontologijų integravimą į Lietuvių kalbos išteklių informacinę sistemą.

29.2.6. Įterptinių vektorių kalbos modeliai.

29.2.6.1. UAB „Tokenmill“ atvirojoje prieigoje pateikė w2v algoritmu apdorotą socialinės žiniasklaidos įterptinių vektorių modelį²⁴.

29.2.6.2. Vytauto Didžiojo universitetas atvirojoje prieigoje pateikė *Fasttext* algoritmu parengtą socialinės medijos tekstų įterptinių vektorių modelį²⁵, o šiuo metu rengia BERT ir

²¹ <https://www.xn--ratija-ckb.lt/lokalizavimas/%C5%BEodynai/336>

²² <https://www.vle.lt/>

²³ <https://www.snomed.lt/>

²⁴ <http://word2vec.tokenmill.lt/>

²⁵ <http://fasttext.vdu.lt/>

ELMO algoritmais apdorotus socialinės žiniasklaidos tekstų įterptinių vektorių modelius, kurie bus pateikti atvirojoje prieigoje kaip atviri duomenų rinkiniai.

29.2.7. Įvertinant kalbos duomenų ir jų rinkinių poreikį, pažymėtini du aspektai. Pirma, visų rūšių duomenys turi būti nuolat gausinami ir atnaujinami, kad atspindėtų kuo įvairesnes kalbos vartojimo sritis bei kalbos pokyčius ir atlieptų įvairių tikslinių visuomenės grupių reikmes. Antra, šie duomenys yra pagrindinis DI technologijų ir sprendinių šaltinis. Todėl palengvėjus nesustruktūrintų duomenų tekstinio ir garsinio turinio prieigai ypač išaugo sudėtingų anotuotų ir sustruktūrintos informacijos duomenų – garsynų, įterptinių vektorių tekstynų, geoinformacinių bazių (be vietovardžių, įtraukiant ir kitus kalbos duomenis), sintaksiškai anotuotų tekstynų ir ontologijų – poreikis. Kaupiant šiuos duomenis, svarbu atsižvelgti ne tik į jų apimtį, bet ir į kokybę, privačių duomenų apsaugą bei į visuotinį prieigos atvirumą priimtinais formatais ir jų prieinamumo didinimą (pavyzdžiui, per nacionalines infrastruktūras, MTI portalus bei per kuriamą nacionalinį atvirų duomenų portalą²⁶).

29.3. Mašininis vertimas ir lokalizavimas.

29.3.1. Vilniaus universitete, vykdant projektą „Anglų–lietuvių–anglų ir prancūzų–lietuvių–prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas“, sukurta mašininio vertimo sistema ALPMAVIS ir visuomenei prieinama vieša internetinė statistinio mašininio vertimo paslauga²⁷, pasiekama taip pat per informacinę sistemą „Raštija.lt“²⁸. Nuo 2018 m. pabaigos vykdant naują projektą „Mašininio vertimo sistemų ir lokalizavimo paslaugų tobulinimas ir plėtra“ kuriama naujos kokybės neuroniniais tinklais paremta atvira ir nemokama vertimo aplinka. Esamoje infrastruktūroje tobulinamos jau sukurtos mašininio vertimo sistemos, diegiamos papildomos mašininio vertimo kalbų poros, diegiami šnekos atpažinimo ir sintezės sprendiniai, infrastruktūra bus pritaikyta elektroninės valdžios paslaugoms teikti.

29.3.2. Vilniaus universitete įgyvendintas projektas „Visuomenei aktualios programinės įrangos lokalizavimas, programoms reikalingų priemonių sukūrimas“: lokalizuotas atvirojo kodo programinės įrangos paketas ir sukurtos vertimo atmintys (rezultatai prieinami per informacinę sistemą „Raštija.lt“).

29.3.3. UAB „Tilde informacinės technologijos“ suteikia galimybę nemokamai naudotis naujausiais neuroniniais tinklais grįstomis daugiakalbėmis mašininio vertimo sistemomis²⁹.

29.3.4. Minėtinos ir ne Lietuvoje sukurtos vertimo sistemos: *Google Translate* ir *Microsoft Bing Translator* ir *eTranslation*. *Google Translate* ir *Microsoft Bing Translator* vertimo sistemų privalumas yra tas, kad nemokama mašininio vertimo paslauga yra siūloma daugumai pasaulio kalbų, įskaitant ir lietuvių. *E-translation* sistema yra nemokama valstybinėms institucijoms, skirta ES oficialių kalbų vertimams ir labiau pritaikyta administracinės ir teisinės kalbos tekstų vertimams.

29.3.5. Tolesniam mašininio vertimo sistemų tobulinimui trūksta daugiakalbių duomenų bazių (dvikalbių lygiagrečiųjų tekstynų), taip pat specializuotų tekstų duomenų, kurie užtikrintų aukštesnę mašininio vertimo kokybę. Reikėtų susitelkti į dvikalbių atvirų duomenų (dvikalbių tekstynų, įvardytų esybių duomenų bazių) kūrimą ir jų atvėrimą visuomenei.

29.4. Šnekos technologijos (šnekos atpažinimas ir šnekos sintezė).

²⁶ <https://data.gov.lt/> (kuris yra kuriamas Ekonomikos ir inovacijų ministerijos kartu su Informacinės visuomenės plėtros komitetu).

²⁷ <https://www.versti.eu>

²⁸ <https://www.raštija.lt>

²⁹ <https://translate.tilde.com>

29.4.1. Šnekos atpažinimas.

29.4.1.1. Projekte „Semantika-2“ Vytauto Didžiojo universiteto mokslininkai sukūrė ir išplėtojo nemokamą atvirojo kodo lietuviškų fonogramų (šnekos failų) transkripcijos tekstu sprendinį, kuris atpažįsta laisvai formuluojamą šneką šiose srityse: bendrinė kalba, teisės ir medicinos kalbinės atmainos. Sprendinys pateikiamas kaip elektroninė paslauga ir gali būti naudojamas tolesniam taikymui bei paslaugoms kurti.

29.4.1.2. Projekte „LIEPA“ Vilniaus universiteto mokslininkai sukūrė penkias paslaugas, kuriose šnekos atpažinimo technologijos pritaikytos valdyti kompiuterį balsu: „Naršytuvas“ (naršymo valdymas balsu), „Pažintuvas“ (mokymosi valdymas balsu), „Valdytuvas“ (kompiuterio valdymas balsu), „Ieškotuvai“ (UNESCO paveldo išteklių ieškojimas balsu), „Pagalbininkas“ (valdymas balsu, skirtas neįgaliesiems). Projekte „LIEPA-2“, taikant šnekos atpažinimo technologijas, kuriamos dar keturios paslaugos: „Ugdančiojo roboto valdytuvas“ (humanoidinio vaikams skirto roboto valdymas), „Skambintuvas“ (skambinimas telefono kontaktams), „Taksi iškviestuvai“ ir „Tarpkalbinis komunikatorius“ (lietuvių–kinų kalbų).

29.4.1.3. UAB „Tilde informacinės technologijos“ suteikia galimybę nemokamai naudotis šnekos atpažinimo programa, kuri paverčia šneką į tekstą iš anksčiau įrašyto garso failo arba diktuojamo teksto³⁰.

29.4.1.4. Google kompanija siūlo lietuvių šnekos atpažinimo paslaugas mobiliesiems įrenginiams ir nemokamą *Rašymo balsu* paslaugą *Google Docs* priemonėje. Nors kompanija teikia patogią paslaugos teikimo „mašina – mašina“ lygmens sąsają, tačiau būtent lietuvių šnekai, ypač jos taikymui specialiose srityse, kompanijos *Google* šnekos atpažinimo kokybė ženkliai atsilieka nuo Lietuvoje kuriamų sprendinių.

29.4.2. Šnekos sintezė.

29.4.2.1. Projekte „LIEPA“ Vilniaus universiteto mokslininkai sukūrė dvi paslaugas, kuriose pritaikytos šnekos sintezės technologijos: 1) „Tartuvas“ (garsinis lietuvių kalbos naujažodžių žodynėlis, sukurtas kartu su Lietuvių kalbos institutu), 2) lietuvių šnekos sintezatorius akliems – su SAPI5 standartu suderinamas lietuvių šnekos sintezatorius, skaitantis balsu tai, kas rodoma kompiuterio ekrane. Projekte „LIEPA-2“ kuriamos dar dvi paslaugos: mobilusis šnekos sintezatorius akliems ir interneto naujienų skaitytuvas.

29.4.2.2. Vytauto Didžiojo universitete sukurtas lietuvių šnekos sintezavimo prototipas, kuris šiuo metu pateiktas testuoti verslo ir socialiniams partneriams. Planuojama plėtoti tinklinę paslaugą (*Google Cloud Natural Language API* principu).

29.4.2.3. UAB „Tilde informacinės technologijos“ suteikia galimybę nemokamai naudotis šnekos sintezės programa, kuri paverčia tekstą į šneką²³. Programoje naudojama projekte „LIEPA“ sukurta teksto sintezės technologija. Šiuo metu įmonė testuoja naują neuroniniais tinklais pagrįstą šnekos sintezės technologiją.

29.4.3. Šiuolaikinės šnekos atpažinimo ir sintezės pagrindą sudaro anotuoti garsynai, naudojami šnekos atpažintuvams ir sintezatoriams apmokyti. Lietuvių kalbai tokių duomenų labai trūksta, todėl būtina sutelkti pastangas įvairių sričių, dialektų, amžiaus grupių, foninės aplinkos ir kt. požymius turintiems garsynams kurti ir atverti visuomenei.

29.5. Natūraliosios kalbos apdorojimas.

29.5.1. Vytauto Didžiojo universiteto projektuose „Semantika-1“, „Semantika-2“, CLARIN-LT, „Lietuvių kalbos pastoviųjų žodžių junginių automatinis atpažinimas

³⁰ <https://www.tilde.lt/snekos-technologijos>

(PASTOVU)“ ir kt. sukurti visi pagrindiniai skaitmeninio teksto lietuvių kalba bazinės analizės įrankiai, kurie vartotojams pateikiami kaip nemokami atvirojo kodo sprendiniai: segmentatorius, lemuoklis, morfologijos analizatorius, kalbos dalių atpažintuvas, sintaksės analizatorius, rašybos klaidų tikrintuvas, teksto normalizatorius, lietuvių kalbos tekstų indeksavimo išplėstinei paieškai sprendinys, pastoviųjų junginių atpažintuvas ir kiti. Sprendiniai apima norminę (bendrinę) ir socialinės medijos kalbos atmainas, taip pat specialiąsias – žiniasklaidos, teisės, medicinos – kalbos sritis.

29.5.2. Kauno technologijos universitetas projekte „Semantika-2“ sukūrė išsamios tekstų statistinės analizės sprendinį, vertinantį teksto skaitomumo lygį, brandos lygį ir kitus lygius.

29.5.3. UAB „Tokenmill“, pasinaudodama savo ir Vytauto Didžiojo universiteto kurtais kalbos duomenimis, lietuvių kalbos tekstų analizei pritaikė pasaulyje populiarią daugiakalbį kompleksinį atvirosios prieigos sprendinį *Spacy*.

29.5.4. Tolesnei natūraliosios kalbos apdorojimo plėtrai, tobulėjant giliojo mokymosi algoritmams, reikalingi gausūs, patikimi ir įvairių temų duomenys, parengti mašiniam mokymuisi.

29.6. Natūraliosios kalbos supratimas (apimantis semantines technologijas).

29.6.1. Vytauto Didžiojo universitete „Semantika-1“, „Semantika-2“, CLARIN-LT ir kitų projektų metu sukurti kalbos supratimo ir semantinės analizės sprendiniai, kurie vartotojams pateikiami kaip atvirojo kodo sprendiniai: paprastas ir aspektais grįstas sentimentų (nuomonių) analizatorius, neapykantos / įžeidžios kalbos atpažintuvas, automatinis dokumentų santraukų sudarymas, įvardytų esybių atpažintuvas. Vidiniuose universiteto projektuose pradėti rekomendacinių sistemų, automatinio žinių ištraukimo (kasybos) ir skaitmeninės žinių bazės teisės ir medicinos srityje sprendinių kūrimo darbai. Dėl finansavimo stygiaus jie vyksta lėtai.

29.6.2 Tobulėjant giliojo mokymosi algoritmams, natūraliosios kalbos supratimo plėtrai reikalingi gausūs, patikimi ir įvairių temų duomenys, parengti mašiniam mokymuisi.

29.7. Natūraliosios kalbos generavimas. Reikia pripažinti, kad šioje kalbos technologijų srityje Lietuvoje žengiami pirmieji žingsniai. Minėtina UAB „Tokenmill“, kuri sukūrė ne vieną natūraliosios kalbos generavimo sprendinį verslo reikmėms.

30. Kalbos technologijų lygį ir pažangą Lietuvoje galima įvertinti pagal tai, kaip įgyvendinti trys Lietuvių kalbos plėtros informacinės technologijose gairėse (2014–2020) iškelti tikslai: 1) užtikrinti visavertį lietuvių kalbos vartojimą skaitmeninėje terpėje, gerinti mokslinių tyrimų kokybę; 2) plėtoti rašytinės ir sakytinės kalbos technologijų ir išteklių infrastruktūrą, kurti ir tobulinti viešai prieinamus IT sprendinius bei išteklius; 3) diegti lietuvių kalbos skaitmeninius produktus viešosiose elektroninėse paslaugose. Darytina bendra išvada, kad šie tikslai sėkmingai vykdomi, tačiau pabrėžtina, kad patys tikslai yra tęstiniai ir priklauso nuo sparčiai kintančių kalbos technologijų pažangos pasaulinėje rinkoje bei visuomenės poreikių.

30.1. Įgyvendinant pirmąjį praėjusio laikotarpio tikslą, buvo sėkmingai įvykdyta didelė dalis uždavinių. Buvo intensyviai lokalizuojama programinė įranga³¹, kaupiama vertimo atmintis ir kuriamos leksinės bazės ir ontologijos (žr. 29.3), toliau plėtos mokslinių tyrimų infrastruktūros: informacinė sistema „Raštija“ (žr. 29.1.1) bei CLARIN-LT (žr. 29.1.2), LKSSAIS ir LKIIS (žr. 29.1.3). Tai užtikrina kalbos technologijų ir išteklių, sukurtų valstybės ir ES struktūrinių fondų lėšomis, sklaidą ir nemokamą prieigą. Pažymėtina, kad 2019 m., palyginti su 2018 m., padaugėjo gyventojų, kurie naudojami su lietuvių kalba ir Lietuvos kultūros paveldu

³¹ <https://www.xn--ratija-ckb.lt/naujienos/lokalizavimo-naujienos/214>

susijusiomis elektroninėmis paslaugomis. Tokiomis elektroninėmis paslaugomis kaip internetinės vertimo priemonės, automatinės kalbos atpažinimo priemonės naudojosi 15 proc. Lietuvos gyventojų (plg. 2018 m. IV ketvirtį – 10 proc.).³² Taip pat apskritai daugėja internete viešai prieinamų lietuvių kalbos ir raštijos išteklių, priemonių, elektroninių paslaugų naudotojų (plg. 2018 m. 62 proc., 2019 m. – 64 proc.)³³.

Vienas iš lietuvių kalbos privalumų yra tai, kad ji, būdama viena iš oficialiųjų ES kalbų, dalyvauja tokiuose svarbiuose ES procesuose kaip mašininis vertimas, dalykinės terminologijos vertimas ir terminų bazių kūrimas. Tiesa, kalbos technologijų specialistų tobulinimas Lietuvoje iki šiol labai vangus dėl nekomercinės lietuvių kalbos specifikos ir riboto Lietuvos IT bendrovių suvokimo, kaip kalbos technologijas galima panaudoti inovacijoms kurti. Kompiuterinė lingvistika ir kalbos technologijos, kaip atskiras mokomasis dalykas, kol kas nėra įtvirtintos Lietuvos aukštojo mokslo sistemoje. Nė vienas universitetas nesiūlo visų lygmenų kalbos technologijų studijų, dėl to šioje srityje dažniausiai dirba tik patirtį sukaupę universitetų kalbos technologijų srities mokslininkai ir tyrėjai. Kita vertus, nekomercinė lietuvių kalba ir didelės pagrindinių jos technologijų kūrimo bei įgyvendinimo sąnaudos rinkoje nesukuria kvalifikuotų kalbos technologijų srities darbuotojų paklausos. Tiesa, visaverčių studijų trūkumą iš dalies atsveria augantis komercinių nuotolinių mokymų populiarumas ir prieinamumas, todėl motyvuoti specialistai gali įgyti reikiamą kvalifikaciją ir be tradicinių aukštojo mokslo studijų. Dėl visų šių aplinkybių tiek visavertis lietuvių kalbos vartojimas skaitmeninėje terpėje, tiek kalbos technologijų specialistų kvalifikacijos kėlimas ir toliau išlieka svarbiais uždaviniais.

30.2. Praėjusio laikotarpio antrojo tikslo įgyvendinimas apėmė kalbos technologijų ir išteklių plėtrą, daugiakalbio skaitmeninio turinio valdymo ir prieigos įrankių kūrimą, mašininio vertimo, balso atpažinimo ir kitų kalbos technologijų įrankių plėtrą. Sėkmingą šio tikslo įgyvendinimą daug lems iš ES struktūrinių fondų investicijų veiksmų programos 2 prioriteto „Informacinės visuomenės skatinimas“ 02.3.1-CPVA-V-527 priemonės „Lietuvių kalba informacinėse technologijose“ Nr. 02.3.1-CPVA-V-527 skirtas finansavimas penkiems projektams, kuriuos įvykdžius iki 2020 m. pabaigos bus atlikta didelė dalis 2014–2020 m. gairėse numatytų tobulinti priemonių: automatinė transkripcija ir diktavimo sistemos (žr. 29.4), specialieji ir bendrieji anotuoti garsynai (žr. 29.2.2), išplėsta kalbos sintezė (žr. 29.4.2), interneto turinio analizės ir valdymo įrankiai (žr. 29.5), integruota automatinio teksto vertimo ir lietuvių šnekos infrastruktūra (žr. 29.3), padidintas sintaksiškai anotuotas tekstynas (žr. 29.2.1.1), kuriami vienakalbiai ir daugiakalbiai ištekliai (žr. 29.2), semantinės duomenų bazės (žr. 29.2.4), plečiamos išteklių ir technologijų infrastruktūros (žr. 29.1), kuriamos naujos pažangios e. paslaugos (žr. 29.1).

30.3. Trečiojo tikslo įgyvendinimas apėmė lietuvių kalbos skaitmeninių produktų diegimą viešosiose elektroninėse paslaugose, ypač pabrėžiant jų pritaikymą e. valdžios paslaugoms bei pritaikymą specialiesiems poreikiams. Šį tikslą, kaip ir antrąjį, įgyvendinti padeda ES struktūrinės priemonė „Lietuvių kalba informacinėse technologijose“. Įvykdžius jos projektus iki 2020 m. pabaigos, bus sukurta 21 nauja elektroninė paslauga. Dauguma paslaugų, taip pat IT sprendinių informacinėse sistemose „Raštija“ bei „Semantika.lt“ bus pritaikytos neįgaliųjų poreikiams.

³² <https://ivpk.lrv.lt/lt/naujienos/vis-daugiau-gyventoju-naudojasi-su-lietuviu-kalba-ir-lietuvos-kulturos-paveldu-susijusiomis-paslaugomis>

³³ <https://ivpk.lrv.lt/lt/veiklos-sritys-1/informacines-visuomenes-statistika>

IV SKYRIUS

TIKSLAS, UŽDAVINIAI IR JŲ ĮGYVENDINIMAS

31. Pasaulyje keičiasi mokslinė kalbos technologijų paradigma, neįtikėtinu greičiu plėtojamos intelektualiosios technologijos, robotizacija ir atsiranda daiktų internetas. Kaip teigiama DI baltojoje knygoje³⁴, pasaulyje sugeneruojamų duomenų kiekis nuo 33 zetabaitų 2018 m. augs iki 175 zetabaitų 2025 m. Kiekviena nauja duomenų banga yra proga Europai ir Lietuvai įsitvirtinti duomenų ekonomikoje ir tapti pasauline šios srities lydere. Be to, per artimiausius penkerius metus iš esmės keisis duomenų saugojimo ir tvarkymo būdas. Šiuo metu 80 proc. debesijos duomenų tvarkymo ir analizės procesų vykdoma duomenų centruose ir centriniuose kompiuterijos įrenginiuose, o 20 proc. – išmaniuosiuose prie tinklo jungiamuose objektuose, tokiuose kaip automobiliai, buitinės technikos prietaisai arba gamybiniai robotai, ir naudotojai po ranka esančiuose kompiuteriniuose įrenginiuose. Iki 2025 m. šios proporcijos turėtų iš esmės keistis.

Šioje dinamiškoje ir milžiniškoje informacinėje terpėje labai svarbu neatsilikti nuo pažangiausių kalbos technologijų plėtros krypčių pasaulyje, neprarasti tapatybės ir užtikrinti visavertį lietuvių kalbos gyvavimą skaitmeninėje terpėje. Lietuvių kalbos integravimas į naujuosius procesus itin reikšmingas tiek visuomenės, tiek mokslo, tiek ir ekonomikos plėtrai. Tokiame kontekste formuluojamas Gairių tikslas, uždaviniai ir jų įgyvendinimo priemonės.

32. Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 m. gairių tikslas yra užtikrinti visavertį lietuvių kalbos funkcionavimą skaitmeninėje terpėje ir jos lietuvinimo pažangą, skatinti lietuvių kalbai pritaikytų technologijų plėtrą, gerinti jomis grįstų paslaugų visuomenei kokybę. Tikslas formuluojamas atsižvelgiant į Valstybinės kalbos politikos 2018–2022 m. gaires (42.3 papunktis).

33. Šiam tikslui įgyvendinti numatomi trys uždaviniai:

33.1. didinti specialistų, dirbančių kalbos technologijų srityje, kompetenciją ir kelti visuomenės gebėjimo naudotis kalbos technologijų teikiamomis galimybėmis lygį;

33.2. kaupti ir gausinti atvirus, patikimus, kokybiškus, pakartotinai pritaikomus skaitmeninius kalbos išteklius ir kitus skaitmeninius kalbos duomenų rinkinius;

33.3. plėtoti kalbos technologijų infrastruktūrą, kalbos technologijų taikymą viešajame sektoriuje ir viešosiose paslaugose, kurti ir tobulinti viešai prieinamus IT sprendinius ir priemones.

34. Pirmas uždavinys apima aukštųjų mokyklų programų tobulinimą, edukacinių iniciatyvų ir edukacinio turinio plėtrą.

Žmogiškųjų išteklių tobulinimas yra labai svarbi sąlyga, kuri lems sėkmingą lietuvių kalbos plėtrą skaitmeninėje terpėje ir kalbos technologijų pažangą. Todėl visais lygiais, pradedant nuo ankstyvojo amžiaus, turi būti skatinamas kalbos technologijų, kompiuterinės lingvistikos, dirbtinio intelekto technologijų, programinės įrangos lokalizavimo specialistų ugdymas, rengimas ir kvalifikacijos tobulinimas:

- integruojant šias temas į įvairių dalykų (ypač lietuvių kalbos ir informacinių technologijų) pamokas vidurinio ugdymo programose,
- rengiant trijų pakopų studijų aukštosiose mokyklose programas, kur būtų rengiami

³⁴ „Dirbtinis intelektas. Europos požiūris į kompetenciją ir pasitikėjimą“.

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_lt.pdf

- aukštos kvalifikacijos specialistai,
- įsitraukiant į tarptautinius kompetencijos tinklus,
- vykdant mokslinius ir struktūrinius projektus,
- atliekant mokslinius tyrimus,
- skatinant nacionalinį ir tarptautinį bendradarbiavimą visais lygiais,
- kuriant ir diegiant kalbos technologijomis ir ištekliais grįstą interaktyvų edukacinį turinį (pavyzdžiui, skaitmeninius kalbos žinynus, mokymo priemones ir kt.) bei pritaikant jį įvairiems skaitmeniniams įrenginiams.

35. Antras uždavinys apima patikimų, kokybiškų, pakartotinai naudojamų ir atvirųjų skaitmeninių išteklių, jų rinkinių kūrimą, kaupimą ir gausinimą.

Siekiant, kad valstybinės lietuvių kalbos technologijų sprendiniai kuo greičiau prilygtų komercinių kalbų, visų pirma anglų kalbos, technologijų lygiui ir būtų patrauklūs naudoti daugiakalbėse sistemose, kuriamose ne tik Lietuvos, bet ir užsienio gamintojų, ypač daug dėmesio turi būti skiriama skaitmeniniams kalbos duomenims ir ištekliams. Tinkamai kuriami ir plėtojami kalbos ištekliai ir duomenys, atitinkantys tarptautinius standartus, leis įveikti prarają tarp šiuo metu atskirai plėtojamų disciplinų, pavyzdžiui, mašininio mokymosi ir giliojo mokymosi (kuriems būdingas ribotas interpretavimas, poreikis turėti daug duomenų modeliams mokyti ir mokytis iš koreliacijų) ir simbolinių metodų (kurių taisyklės kuria žmogus).

35.1. Todėl šį uždavinį įgyvendinant, būtina kurti ir plėtoti:

35.1.1. bendruosius kalbos duomenis ir išteklius (tekstynus, dažninius sąrašus, žodynus, garsynus, ontologijas, taip pat mišriuosius duomenis (angl. *multimodal data*), apimančius bendrojo pobūdžio ir specialiąsias sritis), kurie reikalingi kuriant kalbos technologijas bei jų taikymą, skaitmenines žinių bazes, semantinius tinklus ir kt.;

35.1.2. mokomuosius kalbos duomenis ir išteklius (specialiai mašininiam mokymuisi parengtus tekstynus, garsynus, įterptinius vektorių modelius, skaitmeninius sakytinės, rašytinės kalbos modelius, auksinius standartus ir kt., taip pat mišriuosius duomenis (angl. *multimodal data*), apimančius bendrojo pobūdžio ir specialiąsias sritis), kurie ypač svarbūs DI technologijų taikymui ir IT sprendiniams.

35.2. Sukurtus ir sukauptus duomenis ir išteklius reikia saugoti tam skirtose viešose, privačiose, nacionalinėse ar tarptautinėse duomenų infrastruktūrose, užtikrinant jų:

35.2.1. atvirąją prieigą,

35.2.2. pakankamą kiekį ir įvairovę, atsižvelgiant į tinkamą ir proporcingą kalbos reiškinių bei visuomenės aspektų įvairovės atspindėjimą,

35.2.3. saugumą, patikimumą, teisingumą, tikslumą, sąveikumą ir pakartotinį, pritaikymą, laikantis gerosios pasaulinės praktikos ir tarptautiniu lygiu pripažintų formatų ir standartų,

35.2.4. asmens duomenų apsaugą, atsižvelgiant į Bendrąjį duomenų apsaugos reglamentą ir kitus galiojančius teisės aktus, reikalavimus bei normas,

35.2.5. autorių ir gretutinių teisių apsaugą,

35.2.6. apsaugą nuo visų rūšių diskriminacijos (lyties, rasės, tautybės, pilietybės ir kt. pagrindu).

36. Trečias uždavinys apima kalbos technologijų taikymą viešajame sektoriuje ir jų plėtoją viešosiose paslaugose, lietuvių kalbos technologijų IT sprendinių ir priemonių kūrimą ir tobulinimą.

Labai svarbu, kad viešojo administravimo institucijos, sveikatos priežiūros įstaigos, komunalinių ir transporto paslaugų įmonės, finansų priežiūros institucijos ir kiti viešieji subjektai savo veikloje pradėtų sparčiau ir plačiau diegti valstybinės kalbos technologijų sprendinius, priemones, produktus ir paslaugas. Tai duos valstybei ketveriopą naudą:

- **viešojo intereso paslaugas teikiantiems subjektams** bus užtikrinama mažesnė paslaugų kaina, didesnis paslaugų tvarumas ir patrauklumas, patogesnis naudojimas ir tinkamesnė piliečių teisių ir laisvių apsauga,
- **piliečiams** bus suteikiamos patogesnės ir patrauklesnės viešosios paslaugos bei jų pritaikymas specialiesiems ir neįgaliųjų poreikiams,
- **verslui** bus atveriamos naujos plėtros galimybės, pavyzdžiui, kuriant naujos kartos inovatyvius produktus ir paslaugas,
- **tyrėjams ir inovacijų raidai** bus sudaromos geresnės galimybės vykdyti nekomercinės lietuvių kalbos technologijų taikomuosius tyrimus, eksperimentinę plėtrą ir kurti naujos kartos IT sprendinius, produktus ir paslaugas, taip užtikrinant, kad lietuvių kalbos technologijų sprendiniai ir taikymas savo kokybe ir efektyvumu prilygtų jų analogams, skirtiems anglų kalbai.

Todėl šį uždavinį įgyvendinant būtina:

36.1. skatinti sektorinį ir tarpsektorinį dialogą, užtikrinant glaudesnę mokslo ir viešojo sektoriaus bei mokslo ir verslo bendradarbiavimą, kad viešųjų ir komercinių paslaugų teikėjams būtų pasiūlytas efektyvus veiksmų planas ir priemonės, siekiant sudaryti palankesnes kalbos technologijų kūrimo, eksperimentavimo ir diegimo sąlygas kuo didesniame viešųjų paslaugų skaičiuje;

36.2. tobulinti lietuvių kalbai pritaiktų hibridinių ir DI technologijų sprendinius šiose srityse: bendravimo robotų virtualiųjų asistentų kūrimo (angl. *chat bots*), kompiuterinės lingvistikos, informacijos paieškos ir gavybos, natūraliosios kalbos apdorojimo, supratimo ir generavimo, mašininio vertimo, dialogų sistemų, sentimentų (nuomonių) analizės, tekstų klasifikavimo, automatinio santraukų sudarymo, duomenų bei informacijos iš nesustruktūrintos informacijos (teksto, garso, mišrių) šaltinių gavybos, visaverčio ir visapusiško šnekos technologijų plėtojimo;

36.3. kurti arba toliau plėtoti kalbos technologijų saugojimo ir dalijimosi infrastruktūras, skatinti atvirojo kodo programų kūrimą ir užtikrinti tinkamą kalbos technologijų įrankių ar sprendinių patentavimą ir licencijavimą;

36.4. užtikrinti, kad konkrečių priemonių ir viešųjų pirkimų organizatoriai tinkamai ir pagrįstai atsižvelgtų į naujausių ir pažangiausių kalbos technologijų sprendinių teikiamą naudą valstybinės lietuvių kalbos stiprinimui ir įtvirtinimui skaitmeninėje erdvėje bei išmaniuosiuose įrenginiuose (įskaitant robotus ir robotizuotas sistemas).

37. Gairių įgyvendinimo stebėseną atlieka Valstybinė lietuvių kalbos komisija.