PRITARTA
Valstybinės lietuvių kalbos komisijos 2013 m.
spalio 24 d. posėdyje, protokolo Nr. P-7

LIETUVIŲ KALBOS PLĖTROS INFORMACINĖSE TECHNOLOGIJOSE 2014–2020 M. GAIRĖS

I. BENDROSIOS NUOSTATOS

1. Lietuvių kalbos plėtros informacinėse technologijose gairės (toliau – Gairės) parengtos siekiant užtikrinti visavertį lietuvių kalbos vartojimą skaitmeninėje terpėje, įtvirtinti lietuvių kalbos statusą informacinėje visuomenėje, apsaugoti lietuvių kalbą nuo vadinamojo skaitmeninio išnykimo ir mažinti lietuviškai kalbančios bendruomenės atskirtį globalioje žinių visuomenėje. Šie siekiai numatomi įgyvendinti gausinant išteklius, plėtojant kalbos technologijas ir viešąsias paslaugas, atitinkančias informacinės visuomenės lūkesčius ir poreikius.

2. Gairėse 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 socialiniame, politiniame ir kultūriniame gyvenime.

3. Gairės parengtos atsižvelgus į Lietuvių kalbos informacinėje visuomenėje 2009–2013 metų programos įgyvendinimo rezultatus ir lietuvių kalbai skirtų technologijų būklės vertinimus. Rengiant Gaires remtasi šiais dokumentais: Valstybės pažangos strategijos „Lietuva 2030“ nuostatomis, Valstybinės kalbos įstatymu, Lietuvos informacinės visuomenės plėtros 2011–2019 metų programa, taip pat atsižvelgta į Europos Sąjungos Naujos daugiakalbystės pagrindų strategijos ir Europos skaitmeninės darbotvarkės „2015.eu“ nuostatas.

II. NACIONALINĖS IR EUROPOS SĄJUNGOS INICIATYVOS

4. Valstybės pažangos strategijoje „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 elektroninėje 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.

5. Didžioji dalis iniciatyvų ir įsipareigojimų dėl lietuvių kalbos funkcionavimo informacinėje visuomenėje ir kalbos technologijų kūrimo buvo keliama ir įgyvendinama nacionaliniu lygmeniu. Informacinės visuomenės plėtros 2011–2019 m. programoje, patvirtintoje Lietuvos Respublikos Vyriausybės 2011 m. kovo 16 d. nutarimu Nr. 301 (Žin., 2011, Nr. 33-1547), lietuvių kalbos ir kultūros puoselėjimas pasitelkiant IRT įvardijamas kaip informacinės

visuomenės plėtros prioritetas. Nuo 2000 ųjų Lietuvoje buvo vykdytos dvi programos: Lietuvių kalbos informacinėje visuomenėje 2000–2006 m. programa (patvirtinta Lietuvos Respublikos Vyriausybės 2000 m. balandžio 26 d. nutarimu Nr. 471 (Žin., 2000, Nr. 36-1002), kurią koordinavo Valstybinė lietuvių kalbos komisija, ir Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programa (patvirtinta Lietuvos Respublikos Vyriausybės 2007 m. kovo 21 d. nutarimu Nr. 319 (Žin., 2007, Nr. 40-1487; 2009, Nr. 125-5388), kurios įgyvendinimą koordinuoja Švietimo ir mokslo ministerija kartu su Susisiekimo ministerija. Įgyvendinant pirmąją nacionalinę Lietuvių kalbos informacinėje visuomenėje 2000–2006 m. programą buvo vykdyti atvirųjų programų lokalizavimo, išteklių kūrimo, automatinio kalbos atpažinimo projektai, gerinama kalbos sintezės kokybė, sukurti kompiuterinis šriftas *Palemonas* bei morfologinės analizės ir generavimo įrankiai, pradėti lietuviškų tekstų sintaksinės ir semantinės analizės darbai.

Lietuvių kalbos informacinėje visuomenėje 2009–2013 metų programoje numatyta tobulinti esamus ir kurti naujus kalbinius išteklius, gerinti automatinio kalbos atpažinimo ir kalbos sintezės technologijas, kurti naujus automatinio vertimo įrankius, gerinti ir kurti semantinės analizės ir informacijos paieškos priemones, sukurti interneto portalą, kuriame būtų galima nemokamai naudotis kalbos ištekliais ir technologijomis.

Pagal kitas nacionalines ir tarptautines programas įgyvendinami projektai taip pat prisideda prie lietuvių kalbos išsaugojimo skaitmeninėje erdvėje. 2004–2006 m. Bendrojo programavimo dokumento programiniu laikotarpiu Vytauto Didžiojo universitete sukurtas automatinio anglų–lietuvių kalbų vertimo įrankis (projektas „Internetinės informacijos vertimo priemonės“), sukaupti duomenys, reikalingi įsitraukiant į tarptautinės mokslinės infrastruktūros CLARIN tinklą, pagal Lietuvos mokslų tarybos įgyvendinamą Nacionalinės lituanistikos plėtros 2009–2015 metų programą remiamas lituanistikos mokslinių tyrimų išteklių skaitmeninimas, naujų lituanistikos duomenų bazių kūrimas, palaikymas ir plėtra, internetinių prieigų kūrimas, taip pat skatinama lituanistikos darbų elektroninė leidyba.

6. Kalbos technologijų kūrimo ir išteklių kaupimo darbus atlieka mokslo ir studijų institucijos.

Daugiausia sakytinės kalbos technologijų mokslinių tyrimų atliekama Kauno technologijos universitete, Vilniaus universiteto Matematikos ir informatikos institute ir Vytauto Didžiojo universitete.

Kauno technologijos universiteto Kalbos tyrimų laboratorijoje automatinio kalbos atpažinimo tyrimai vyksta nuo 1980 m. Laboratorija yra sukūrusi komandų ir skaitmeninių sekų garsyną. Kuriami lietuviški kompiuteriniai dialogai, sukauptas ir tobulinamas lietuvių sakytinės kalbos garsynas LTDIGITS.

Vilniaus universiteto Matematikos ir informatikos institute sukauptas Lietuvos radijo žinių garsynas LRNO. Institutas su partneriais pagal Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą kuria balsu valdomas paslaugas (pvz.: lietuvių kalbos naujadarų tartuvą, naršytuvą ir valdytuvą, leidžiančius balsu ieškoti informacijos internete ir valdyti kompiuterį), taip pat tobulina sakytinės kalbos technologijas bei įrankius (elektroninio teksto skaitytuvą, komandų ir frazių atpažinimo variklį, lietuvių šnekos atpažinimo variklį ir kt.).

Vytauto Didžiojo universitete sukauptas universalus sakytinės lietuvių kalbos garsynas, vyksta sakytinės lietuvių kalbos automatinio skaidymo tyrimai, kuriama sakytinės lietuvių kalbos automatinė transkripcija (čia kaupiami ir mažesnės apimties specialieji tekstynai, skirti kalboms mokytis, pavyzdžiui, jaunuolių sakytinės kalbos tekstynas SACODEYL).

Vilniaus universitete atlikti sakytinės kalbos sintezės ir jos sistemų pritaikymo akliesiems ir silpnaregiams tyrimai. Lietuviška balso sintezės programa „Aistis“ apima automatinį lietuviškų žodžių skaidymą skiemenimis, žodžių lietuvių kalba parašytame tekste automatinį

kirčiavimą, automatinį lietuviškų tekstų transkribavimą, fonetinių vienetų bazę, lietuviškų tekstų pavertimo sakytine kalba kokybės įvertinimą. Balso sintezatorius MBROLA pagrįstas Vilniaus universitete sukurta fonetinių vienetų baze.

2008 metais sukurta pirmoji automatinio anglų-lietuvių kalbų vertimo sistema, kurios pagrindas yra taisyklėmis pagrįsta technologija. Tačiau projektas, kurį įgyvendino Vytauto Didžiojo universitetas kartu su Rusijos bendrove „ProMT“, daugiau neplėtojamas, jam trūksta tikslinių kalbos išteklių ir įrankių: gausesnių žodynų, anotuotų tekstynų ir pan. Pagal Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą Vilniaus universitetas su partneriais kuria naujas anglų–lietuvių, lietuvių–anglų, prancūzų–lietuvių ir lietuvių–prancūzų vertimo sistemas.

Daugiausia automatiniam vertimui reikalingų išteklių sukaupta ir atvirai internetu prieinama Vytauto Didžiojo universitete: 1) anotuotas Dabartinės rašytinės lietuvių kalbos tekstynas, turintis apie 140 mln. žodžių (pagal Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą Dabartinės lietuvių kalbos tekstyną numatoma papildyti 50 mln. žodžių, taip pat sukurti lietuviško interneto tekstyną (800 mln. žodžių); 2) lygiagretieji lietuvių kalbos ir kitų (anglų, vokiečių, čekų, latvių) kalbų tekstynai. Minėtinas Vilniaus universitete sukauptas lietuvių mokslo kalbos tekstynas *CorALit*. Vis dėlto dabartinių lietuvių kalbos tekstynų modernioms lietuvių kalbos technologijoms (informacijos paieškos, automatinio vertimo ir kitoms sistemoms) nepakanka. Esamiems ir būsimiems tekstynams reikia bendros lietuvių kalbai pritaikytos programinės įrangos, kuri leistų kuo geriau išnaudoti turimus kalbos išteklius ir iš jų gaunamus skaitmeninius aprašus.

Didžiausi lietuvių kalbos leksikos skaitmeniniai ištekliai sukaupti ir tvarkomi Lietuvių kalbos institute – skaitmeninis „Lietuvių kalbos žodyno“ leidimas ir duomenų bazė, „Dabartinės lietuvių kalbos žodynas“, Lietuvos vietovardžių geoinformacinė duomenų bazė ir kt. Ištekliai gausinami pagal Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą: numatoma suskaitmeninti dalį lietuvių kalbos ir tautosakos paveldo kartotekų, 6 vienakalbiai (sinonimų, antonimų, frazeologizmų ir kt.) ir 5 dvikalbiai (lietuvių ir latvių, vokiečių, lenkų) žodynai.

Pastaruosius keletą metų lietuvių kalbos semantinės analizės eksperimentai buvo atliekami fragmentiškai, kol kas nepavyko sukurti stabilaus semantinio analizatoriaus ir išsamesnės bendrosios lietuvių kalbos ontologijos. Sparčiau kuriamos specialiosios įvairių sričių ontologijos, sėkmingiau atliekami kai kurių semantinių grupių leksikos tyrimai. Sintaksiškai anotuotą tekstyną kuria Vytauto Didžiojo universiteto ir Kauno technologijos universiteto specialistai, toks tekstynas pradedamas kaupti Lietuvių kalbos institute.

Su informacijos paieška ir valdymu susijusius tyrimus vykdo Vytauto Didžiojo universitetas (projektą „Informacijos valdymo semantinė sistema“ pagal Ekonomikos augimo veiksmų programą remia Europos Sąjungos struktūriniai fondai), Vilniaus universiteto Matematikos ir informatikos institutas, Kauno technologijos universitetas.

Kalbos technologijų, susijusių su turinio semantine ir sintaksine analize bei informacijos paieškos galimybėmis, proveržis numatomas įgyvendinus projektą, kurį pagal Lietuvių kalbos informacinėje visuomenėje 2009–2013 m. programą vykdo Vytauto Didžiojo universitetas kartu su Kauno technologijos universitetu. Numatoma sukurti viešai interneto vartotojams prieinamos lietuvių rašytinės kalbos sintaksinės ir semantinės analizės ir paieškos paslaugas (sintaksinės ir semantinės analizės, lietuviškų svetainių turinio analizės ir paieškos paslaugas), lietuvių kalbos sintaksinės ir semantinės analizės branduolio paslaugas (morfologinės ir sintaksinės analizės, lingvistinės semantinės analizės, specialių sričių semantinės analizės ir paieškos paslaugas).

7. Lietuvių kalbos technologijų industrija nėra išplėtota. Rinkoje veikia keletas bendrovių, kurios dažniausiai yra specializuotos ir atlieka nedidelės aprėpties tikslinius užsakymus (pavyzdžiui, lokalizavimo, morfologinės ir sintaksinės analizės, dokumentų valdymo).

Didesnio masto nacionaliniai ir tarptautiniai verslo projektai įgyvendinami kartu su mokslo ir studijų institucijomis arba su užsienio partneriais. UAB „Tilde informacinės technologijos“ („Tilde IT“) sėkmingai įgyvendino integruojamojo gramatikos tikrintuvo prototipo projektą, kuriam buvo suteikta finansinė parama iš Europos regioninės plėtros fondo pagal Ekonomikos augimo veiksmų programos 1 prioriteto „Ūkio konkurencingumui ir ekonomikos augimui skirti moksliniai tyrimai ir technologinė plėtra“ priemonę „Intelektas LT“. Taip pat „Tilde IT“ prisijungė prie programos *Eurostars* projekto SOLIM (angl. *Spatial Ontology Language for multimedia Information Modeling* – „Daugialypės terpės informacijos modeliavimo erdvinės ontologijos kalba“) ir sukūrė lietuvių kalbos žodžių ryšių, kitaip – semantinio tinklo, duomenų bazės, prototipą (projektas „SemTi“). Šiuo metu „Tilde IT“ dirba su naujomis automatinio vertimo kalbomis (anglų, lenkų, vokiečių, rusų, prancūzų ir kt.), tobulina vertimo technologijas, plečia jų panaudojimą teikiant e. valdžios paslaugas ir didinant vertėjų darbo efektyvumą.

UAB „Fotonija“ kartu su Vytauto Didžiojo universitetu 2006–2008 m. kūrė pirmąją anglų-lietuvių automatinę vertyklę „Internetinė informacijos vertimo priemonė“ (http://vertimas.vdu.lt), taip pat aktyviai dalyvauja kartu su mokslo institucijomis pagal Ekonomikos augimo veiksmų 3 prioriteto „Informacinė visuomenė visiems“ programą įgyvendinamuose projektuose, kuria įrankius, palaikančius lietuvių kalbą skaitmeninėje terpėje, rengia vienakalbius ir daugiakalbius skaitmeninius žodynus.

Lokalizavimo, ontologijų kūrimo ir kitose kalbos technologijų srityse dirba ir kitos verslo įmonės, pavyzdžiui, „Microsoft Lietuva“, „Synergium“, „Sintagma“, „TokenMill“, „HLTech“ ir kt.

Esminės privataus ir viešojo sektorių sąveikos, kalbos išteklių ir technologijų standartizavimo bei sklaidos siekiama kuriant tarptautinius tinklus META NET ir CLARIN.

Įkurtas Europos mokslo tyrimų infrastruktūros konsorciumas ERIC (angl. *European Research Infrastructure Consortium*), kurio plėtrą finansuoja nacionalinės vyriausybės. Iš CLARIN projekto išaugęs konsorciumas siekia sukurti europinį tinklą ir užtikrinti kalbinių išteklių bei kalbos technologijų priemonių sklaidą

Lietuvių kalbos institutas įgyvendino META-NORD projektą, finansuojamą Europos Komisijos lėšomis pagal Informacijos ir ryšių technologijų politikos rėmimo programą, jo tikslas – sukurti: 1) Baltijos ir Šiaurės šalių (Danijos, Estijos, Suomijos, Islandijos, Latvijos, Lietuvos, Norvegijos ir Švedijos) skaitmeninių išteklių paieškos ir metaduomenų infrastruktūrą, prieinamą sistemoje META-SHARE (http://www.meta-share.eu/); 2) lietuvių kalbos išteklių, suvienodintų pagal META-NET skaitmeninius formatus ir aprašytų pagal bendrą metaduomenų sistemą, prieigos tašką.

8. 2008 m. Europos Tarybos rezoliucijoje dėl Europos daugiakalbystės strategijos pabrėžiamas kalbinės įvairovės ir tarpkultūrinio dialogo skatinimas didinant vertimo galimybes, kad būtų galima Europoje ir pasaulyje skleisti idėjas ir žinias įvairiomis kalbomis. Rezoliucijoje skatinama plėtoti daugiakalbes terminijos duomenų bazes ir kalbos technologijas, ypač susijusias su vertimo poreikiais, nustatyti kalbos technologijų sklaidą visose Europos Sąjungos kalbose ir esmines jų taikymo sritis.

Europos Parlamento rezoliucijoje dėl daugiakalbystės – Europos turto ir bendro rūpesčio (2008/2225(INI) – konstatuojama, kad informacijos ir ryšių technologijos visų pirma turėtų būti naudojamos daugiakalbystei skatinti, todėl pabrėžiama tarptautinio standarto ISO 10646 „Informacinės technologijos. Universalusis koduotų ženklų rinkinys“ tinkamo naudojimo

svarba, nes tokiu būdu sudaromos galimybės vartoti visų kalbų abėcėles Europos institucijų ir valstybių narių administracinėse sistemose ir žiniasklaidoje. Taip pat patariama ir skatinama naudoti informacijos bei ryšių technologijas kaip būtinas kalbų mokymo ir mokymosi priemones.

Įgyvendinant Europos Sąjungos daugiakalbystės strategines iniciatyvas 2009–2011 metais pagal mokslinių tyrimų programą FP7-ICT ir inovacijų programą CIP-ICT-PSP pradėti vykdyti 25 kalbos technologijų projektai, kurių bendra vertė apie 56 mln. eurų. Šių projektų tikslas – įveikti kalbų barjerus kuriant bendrą rinką, keičiantis informacija, bendraujant ir kt. Siekiama sukurti skaitmeninę daugiakalbę ekonomiką ir žinių bendruomenę, galinčią nepaisyti kalbų ir valstybių sienų, laisvai keistis informacija ir pritaikyti savo gebėjimus. Daugelio projektų tikslas – kurti technologinius automatinio vertimo pagrindus naudojant socialinių tinklų aplinką ir daugiakalbio interneto turinį, taip pat skatinti viešojo sektoriaus ir bendrosios skaitmeninės rinkos dalyvių bendradarbiavimą efektyviai panaudojant sukauptus kalbos išteklius (projektai META-NET, CLARIN, daugiakalbio tinklo (*Multilingual Web*) iniciatyvos pagal 2009-2010 CIP-ICT-PSP programą ir kt.). Pabrėžtina, kad turėtų būti ypač skatinama lietuvių kalbos įtrauktis į tarptautines programas, pageidautinas aktyvesnis dalyvavimas tarptautinėse mokslinių tyrimų infrastruktūrose.

9. Bendrõsios skaitmeninės riñkos, grindžiamos sparčiuoju ir itin sparčiu internetu, sąveikiomis taikomosiomis programomis, taip pat suteikiančios tvarios ekonominės ir socialinės naudos, siekis užsibrėžtas Europos skaitmeninėje darbotvarkėje, kuri yra viena iš septynių pagrindinių 2020 m. Europos strategijos iniciatyvų. Šios darbotvarkės pagrindinės sritys: 1) dinamiška skaitmeninė bendroji rinka; 2) veiksminga IT produktų ir paslaugų sąveika ir standartų diegimas; 3) pasitikėjimo ir saugumo užtikrinimas; 4) sparčiojo ir itin spartaus interneto prieiga; 5) investicijos į mokslo tyrimus ir inovacijas; 6) skaitmeninio raštingumo didinimas, įgūdžių gerinimas ir įtraukties stiprinimas; 7) IRT teikiama nauda ES visuomenei. Naujoji 2014–2020 m. finansinė perspektyva *Connecting Europe Facility* tarp prioritetų numato skaitmeninės infrastruktūros iniciatyvas, mažinančias vidinės ES rinkos susiskaidymą ir skaitmeninę atskirtį, skatinančias pažangą ir kuriančias naujas darbo vietas.

10. Didžioji dalis programų, kurias šiuo metu naudoja informacinė visuomenė, labai priklauso nuo kalbos technologijų, todėl svarbu, kad valstybinės lietuvių kalbos statusas būtų užtikrinamas teikiant elektronines viešąsias paslaugas, teikiant prieigą prie reikiamos informacijos. Nors pastaraisiais metais lietuvių kalbos technologijų srityje padaryta reikšminga pažanga, ji nėra pakankama, kad kalba būtų visavertiškai vartojama elektroninėje erdvėje.

III. LIETUVIŲ KALBOS TECHNOLOGIJŲ BŪKLĖ IR PLĖTROS GALIMYBĖS

11. Lietuvoje spartesnė informacinės visuomenės plėtra, susidomėjimas kalbos technologijomis ir išteklių kaupimas prasidėjo prieš keletą dešimtmečių, tačiau 2011 m. Informacinės visuomenės plėtros komiteto užsakymu atliktas tyrimas parodė, kad 18 proc. Lietuvos gyventojų jau naudojasi skaitmeninėmis lietuvių kalbos paslaugomis skaitmeninėje erdvėje.

Lietuvių kalba yra viena iš vadinamųjų nekomercinių Europos kalbų, todėl plėtojant kalbos technologijas susiduriama su sunkumais ir problemomis, būdingomis mažiau vartojamų kalbų raidai. Šių technologijų plėtra labai priklauso nuo kitų šalių patirties ir paramos bei tarptautinio bendradarbiavimo. Kita vertus, kalbos technologijų plėtojimas yra svarbiausia lietuvių kalbos funkcionalumo, žinomumo ir lietuviškos kultūros sklaidos daugiakalbėje Europoje stiprinimo proceso sudedamoji dalis. Be to, tik visavertis lietuvių kalbos gyvavimas informacinėje

erdvėje gali padėti įveikti skaitmeninę atskirtį dėl kalbos barjero, mažinti socialinę atskirtį ir didinti kitomis kalbomis pateikiamos informacijos prieinamumą.

Reikia daugiau pastangų kaupiant skaitmeninius lietuvių kalbos išteklius, atliekant kalbos technologijų tyrimus ir diegiant naujoves. Be to, dėl būtinybės sukaupti didelį kiekį duomenų ir dėl kalbos technologijų sistemų sudėtingumo būtina sukurti naujų informacijos mainų ir tarptautinio bei tarpinstitucinio bendradarbiavimo infrastruktūrų.

12. Labai svarbu kurti sąlygas, skatinančias visavertį lietuvių kalbos vartojimą elektroninėje erdvėje, reikiamų specialistų rengimą ir prieinamų IRT sprendinių bei lietuvių kalbos bazinių išteklių kūrimą.

Turi būti sudaromos galimybės lietuvių kalbą vartoti visuose skaitmeniniuose įrenginiuose. Iki šiol Lietuvoje platinama įrenginių (pavyzdžiui, išmaniųjų telefonų, planšetinių kompiuterių), nepritaikytų Lietuvos rinkai, tai yra nesudarančių galimybės vartoti lietuviškus rašmenis, rinktis meniu lietuvių kalba. Nepakankamai dėmesio skiriama programinei įrangai lietuvių kalba. Lietuvos kompiuterininkų sąjungos 2011 m. atliktos apklausos duomenimis, iš universitetuose ir kolegijose naudojamų kompiuterių operacinę sistemą lietuvių kalba turi 25 procentai, sulietuvintą raštinės paketą – 33 procentai.

Valstybinės lietuvių kalbos komisijos 2013 m. surinktais duomenimis, ministerijose ir joms pavaldžiose institucijose naudojama 39,5 proc. sulietuvintų programų, teatruose ir koncertinėse organizacijose naudojama tik 27,4 proc., muziejuose – 26,3 proc. Geriau sulietuvintos programinės įrangos naudojimu rūpinamasi bibliotekose (62,8 proc.), taip pat savivaldybių administracijose (60 proc.).

Šiai sričiai trūksta teisinio reguliavimo. Lietuvių kalbos vartojimo skaitmeniniuose įrenginiuose ir lietuviškos ar sulietuvintos programinės įrangos pasirinkimo galimybės numatytos konstitucinio Valstybinės kalbos įstatymo projekte, kuris iki šiol nėra priimtas.

13. Kompiuterinė lingvistika ir kalbos technologijos, kaip atskira disciplina, kol kas nėra įtvirtinta Lietuvos aukštojo mokslo sistemoje. Nė vienas universitetas nesiūlo visų lygmenų kalbos technologijų studijų, dėl to šioje srityje dažniausiai dirba mokslininkai, baigę lingvistikos ir (arba) informatikos studijas. Su šia sritimi susijusių studijų programų (pavyzdžiui, Kompiuterinės lingvistikos bakalauro studijų programa Kauno technologijos universitete, Taikomosios lietuvių kalbotyros magistro studijų programa Vytauto Didžiojo universitete) absolventų skaičius nedidelis ir negali patenkinti kvalifikuotų kalbos technologijų srities darbuotojų paklausos. Be to, reikėtų stiprinti šioje srityje dirbančių verslo ir mokslo bei studijų institucijų ryšius, numatant tokius bendradarbiavimo būdus, kaip dalyvavimą bendruose projektuose, stažuotes ir pan. Taigi efektyviausi reikiamų specialistų rengimo būdai labiau sietini su kvalifikacijos tobulinimo veiklomis (stovyklomis, konferencijomis ir pan.).

14. Vieni iš pagrindinių sėkmingos kalbos technologijų plėtros veiksnių yra kalbos išteklių ir bazinių įrankių kokybė, prieinamumas bei laisvo jų naudojimo galimybė. Pastaruoju metu skiriama daug dėmesio kalbos technologijų ir skaitmeninių išteklių standartizavimui, prieigos galimybių plėtimui, šie dalykai siejami su kultūrinio paveldo išsaugojimu ir tarptautiškumo plėtra. Siekiama kalbos išteklius ir technologijas Europos mokslinių tyrimų erdvėje jungti į tarptautines infrastruktūras (CLARIN ERIC konsorciumas), skatinti verslo ir mokslo sričių bendradarbiavimą dalijantis patirtimi, standartizuojant išteklius, siekiant didesnių panaudojimo galimybių bei sklaidos (META-NET tinklas). Lietuvoje šie darbai pradėti įgyvendinant META-NORD projektą, o esamų išteklių analizė atlikta siekiant įsijungti į CLARIN tinklą. Vis dėlto dar

turėtų būti nustatyti pagrindiniai išteklių ir kalbos technologijų prieinamumo ir sklaidos principai, numatyta licencijavimo tvarka, skatinama taikyti tarptautinius standartus.

15. Pagal META-NORD projektą buvo parengta lietuvių kalbos skaitmeninių išteklių baltoji knyga. Joje pateikti 2012 m. atlikto lietuvių kalbos technologijų būklės tyrimo apibendrinti rezultatai:

– moksliniai tyrimai leido sėkmingai sukurti gana kokybišką programinę bazinės teksto analizės įrangą, pavyzdžiui, morfologinės ir sintaksinės analizės įrankius, tačiau pažangesnių technologijų, pavyzdžiui, informacijos paieškos sistemos, garso automatinės vertyklės, semantinio teksto anotavimo įrankių, kuriems reikia nuodugnesnio lingvistinio apdorojimo ir semantinių žinių, kol kas tėra tik užuomazgos;

– kuo daugiau lingvistinių ar semantinių žinių reikia sričiai plėtoti, tuo daugiau esama spragų (pvz., informacijos paieškos, teksto semantikos sritys);

– nors sukaupta neblogos kokybės specializuotų tekstynų ar garsynų, jie nepakankamai pritaikytini ir standartizuoti, jų tvarumas nėra efektyviai užtikrinamas;

– trūksta automatiniam vertimui skirtų daugiakalbių lygiagrečiųjų tekstynų;

– labai trūksta daugialypės terpės duomenų tyrimų.

16. Galima teigti, kad daugelyje lietuvių kalbos technologijų specifinių sričių šiandien turime tik riboto funkcionalumo programinę įrangą. Sudėtingesni įrankiai, pavyzdžiui, sintaksiškai anotuoti tekstynai, leksinės semantinės žinių bazės ar sąvokų taksonomijos, tokios kaip *WordNet* ir pan., lietuvių kalbai dar nesukurti. Pažangiausi įrankiai, susiję su semantiniais ir sintaksiniais tyrimais, tik dabar pradedami plėtoti.

Labai netolygi lietuvių kalbos išteklių ir technologijų plėtra trukdo sėkmingai kurti kalbos modelius. Semantikos tyrimų trūkumas lemia mažesnę kalbos generavimo ir teksto analizės pažangą.

Kai kurios sakytinės lietuvių kalbos apdorojimo technologijos veikia gana gerai ir sėkmingai integruojamos industrinėse srityse. Sparčiau plėtojami šnekos sintezės tyrimai ir taikymas, bet atpažinimo kokybė ir pritaikymo galimybės prastesnės.

17. Tolesni darbai turėtų užpildyti išsamesnės semantinės tekstų analizės spragą ir sukaupti trūkstamų išteklių, skatinti kurti taikomuosius prototipus, būtinus viešųjų paslaugų teikimui skaitmeninėje erdvėje. Kuriant išmanesnes ir sudėtingesnes priemones, tokias kaip automatinės vertyklės, reikia išteklių ir technologijų, kurie apimtų daugiau lingvistinių aspektų ir leistų semantiškai nuodugniau analizuoti įvedamą tekstą.

Prioritetu laikytinas įsiliejimas į daugiakalbes mokslines tyrimų ir išteklių infrastruktūras (išnaudojant kompiuterių debesijos teikiamas galimybes, kuriant daugiakalbius technologinius sprendinius ir pan.), kurios leidžia susieti mažiau išplėtotus kalbų išteklius, skatina tarptautinius mokslo tyrimus ir pažangesnių technologijų ir įrankių kūrimo patirties perėmimą ir jų diegimą į viešąsias elektronines paslaugas.

IV. LIETUVIŲ KALBOS PLĖTROS INFORMACINĖSE TECHNOLOGIJOSE TIKSLAI

18. Siekiant sėkmingo lietuvių kalbos gyvavimo informacinėse technologijose svarbiausi yra šie vienas nuo kito neatsiejami kalbos technologijų plėtros tikslai:

18.1. užtikrinti visavertį lietuvių kalbos vartojimą skaitmeninėje terpėje, gerinti mokslinių tyrimų kokybę;

18.2. plėtoti rašytinės ir sakytinės kalbos technologijų ir išteklių infrastruktūrą, kurti ir tobulinti viešai prieinamus IT sprendinius bei išteklius;

18.3. diegti lietuvių kalbos skaitmeninius produktus viešosiose elektroninėse paslaugose.

V. LIETUVIŲ KALBOS PLĖTROS INFORMACINĖSE TECHNOLOGIJOSE UŽDAVINIAI IR PRIEMONĖS

19. Pirmasis tikslas apima lietuviškos skaitmeninės aplinkos kūrimą ir plėtrą, esamų kalbos technologijų ir išteklių prieinamumo ir sklaidos užtikrinimą, žmogiškųjų išteklių tobulinimą.

19.1. Numatomi šie pirmojo tikslo uždaviniai:

19.1.1. sukurti lietuvišką visavertę skaitmeninę aplinką ir skatinti ja naudotis;

19.1.2. tobulinti specialistų, dirbančių kalbos technologijų srityje, kvalifikaciją;

19.1.3. užtikrinti kalbos technologijų ir išteklių, sukurtų valstybės ir Europos struktūrinių fondų lėšomis, sklaidą ir nemokamą prieigą, taip pat jų panaudojimą kuriant naujus ar iš esmės patobulintus produktus ar paslaugas.

19.2. Siūlomos šios priemonės:

19.2.1. lietuvinti vartotojams reikalingą programinę įrangą, ją atnaujinti ir didinti prieinamumą;

19.2.2. kurti ir tvarkyti lokalizavimo terminijos bazes, vertimo atmintis, kompiuterinės leksikos sąvokų taksonomijas ir ontologijas;

19.2.3. užtikrinti galimybę Lietuvoje platinamuose skaitmeniniuose įrenginiuose naudotis visavertėmis klaviatūromis su lietuviškais rašmenimis;

19.2.4. kurti ir diegti įrankius, ugdančius vartotojų įpročius naudoti lietuviškus rašto ženklus elektroninėje komunikacijoje;

19.2.5. plėsti ir tobulinti informacinę sistemą, užtikrinančią prieigą prie valstybės ir ES struktūrinių fondų lėšomis sukurtų išteklių ir technologijų;

19.2.6. standartizuoti esamus ar valstybės ir ES struktūrinių fondų lėšomis kuriamus kalbos išteklius ir technologijas, numatyti jų panaudos ir licencijavimo teisinę bazę;

19.2.7. skatinti kalbos technologijų, kompiuterinės lingvistikos, programinės įrangos lokalizavimo specialistų ugdymą ir kvalifikacijos tobulinimą (pavyzdžiui, trijų pakopų studijų programų, taip pat ir jungtinių atsiradimą, aukštos kvalifikacijos specialistų rengimą, įsijungimą į tarptautinius kompetencijos tinklus, projektus);

19.2.8. diegti kalbos technologijomis pagrįstas paslaugas viešuosiuose interneto prieigos taškuose, skatinti jomis naudotis ir parengti visuomenei skirtų metodinių priemonių, suteikiančių būtinų žinių ir įgūdžių;

19.2.9. ugdyti naudojimosi kalbos technologijomis ir kalbos skaitmeniniais ištekliais įgūdžius, integruojant šias temas per lietuvių kalbos ir informacinių technologijų pamokas.

20. Antrasis tikslas susijęs su kalbos technologijų ir išteklių plėtra, daugiakalbio skaitmeninio turinio valdymo ir prieigos įrankių kūrimu, automatinio vertimo, balso apdorojimo ir kt. įrankių plėtra.

20.1. Numatomi šie antrojo tikslo uždaviniai:

20.1.1. tobulinti lietuvių šnekos atpažinimo ir sintezės technologijas ir įrankius, kurti ir plėtoti didesnio funkcionalumo kalbos analizės ir sintezės įrankius;

20.1.2. plėtoti daugialypės terpės išteklius, įrankius ir turinio analizės bei valdymo tyrimus;

20.1.3. aktyviai reaguoti į kalbos technologijų raidos poreikius ir operatyviai atlikti reikiamus mokslinius tyrimus, kurti ir plėtoti sudėtingesnius skaitmeninius lietuvių kalbos išteklius, užtikrinti jų tvarumą, prieinamumą ir panaudojimo galimybes.

20.2. Siūlomos šios priemonės:

20.2.1. tobulinti automatinę transkripciją, diktavimo sistemas, specializuotų sričių dialogų sistemas;

20.2.2. plėsti specialiųjų sričių ir bendruosius anotuotus garsynus, atsižvelgiant į regioninius, socialinius ir kt. kalbos vartosenos ypatumus;

20.2.3. kalbos sintezę išplėsti semantine ir sintaksine sakytinės kalbos analize bei diegti prozodinius modelius, gerinti ištisinės lietuvių kalbos šnekos atpažinimo kokybę;

20.2.4. kurti ir tobulinti interneto turinio analizės ir valdymo įrankius (pavyzdžiui, sintaksinius ir semantinius bei naujosios leksikos analizatorius, daugiakalbės informacijos paieškos ir santraukų kūrimo įrankius);

20.2.5. kurti viešai prieinamą integruotą automatinio teksto ir lietuvių šnekos vertimo infrastruktūrą;

20.2.6. plėsti sintaksiškai anotuotą tekstyną ir kurti kalbos technologijoms pritaikytą lietuvių kalbos gramatiką;

20.2.7. kurti ir plėsti laisvai prieinamus vienakalbius ir daugiakalbius išteklius: žodynus, duomenų bazes, anotuotus vienakalbius ir lygiagrečiuosius tekstynus (pavyzdžiui, bendruosius ir specialiųjų sričių, ypač mokslo, teisės aktų, interneto kalbos);

20.2.8. kurti bendrąsias ir specialiąsias leksines semantines žinių bazes ar sąvokų taksonomijas ir ontologijas;

20.2.9. parengti skaitmeninius įrankius, skirtus daugiapakopiam nacionaliniam semantiniam tinklui, kaip lietuviško dirbtinio intelekto pamatui, kurti;

20.2.10. kurti tikslines ir laisvai prieinamas išteklių ir technologijų infrastruktūras (pavyzdžiui, atskirų sričių lingvistams, kalbos technologijų kūrėjams, vertėjams, mokykloms, verslui).

21. Trečiasis tikslas numato socialiojo išmaningumo, elektroninio dalyvavimo ir interaktyvios pagalbos plėtrą, siekiant didesnio kalbos technologijų pritaikymo e. valdžios paslaugoms, socialiai prieinamos interaktyvios skaitmeninės pagalbos, paslaugų pritaikymo specialiesiems poreikiams ir pan.

21.1. Numatomi šie antrojo tikslo uždaviniai:

21.1.1. diegti rašytinės ir sakytinės kalbos technologijas viešosiose ir specialiems poreikiams pritaikytose paslaugose;

21.1.2. kurti ir pritaikyti naujoms skaitmeninėms terpėms esamas e. mokymo sistemas ir priemones.

21.2. Siūlomos šios priemonės:

21.2.1. kurti ir diegti tam tikroms sritims (pvz., sveikatos apsaugos, švietimo, viešojo administravimo, teisėsaugos) ir specialiųjų poreikių turintiems žmonėms pritaikytas lietuvių šnekos technologijomis paremtas paslaugas;

21.2.2. kurti semantines žinių tvarkymo sistemas, naudojančias lietuviškų dokumentų sintaksinės-semantinės analizės metodus bei priemones ir skirtas viešojo ir privataus sektorių informacijai valdyti;

21.2.3. kurti lietuvių kalbos technologijas specialiesiems valstybės ir viešojo sektoriaus poreikiams;

21.2.4. diegti specializuotų sričių daugialypės terpės automatinio vertimo sistemas e. valdžios paslaugoms teikti;

21.2.5. kurti ir diegti kalbos technologijomis ir ištekliais paremtą interaktyvų edukacinį turinį (pvz., edukacinius žaidimus, žinynus), pritaikomą įvairiems skaitmeniniams įrenginiams.

_____