META NET

White Paper Series

Baltųjų knygų serija

THE LITHUANIAN LANGUAGE IN THE DIGITAL AGE

LIETUVIŲ KALBA SKAITMENINIAME AMŽIUJE

Daiva Vaišnienė
Jolanta Zabarskaitė

Springer

White Paper Series Baltųjų knygų serija

THE LITHUANIAN LANGUAGE IN THE DIGITAL AGE

LIETUVIŲ KALBA SKAITMENINIAME AMŽIUJE

Daiva Vaišnienė Lietuvių kalbos institutas
Jolanta Zabarskaitė Lietuvių kalbos institutas

Georg Rehm, Hans Uszkoreit
(redaktoriai, editors)

Editors
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-30757-7 ISBN 978-3-642-30758-4 (eBook)
DOI 10.1007/978-3-642-30758-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012942727

ĮŽANGA PREFACE

Ši Baltoji knyga yra viena iš knygų serijos, skleidžiančios žinias apie kalbos technologijas (toliau – KT) ir jų galimybes. Ji skirta pedagogams, žurnalistams, politikams, kalbos vartotojų bendruomenėms ir pan.

Europos kalboms sukurtų bei pritaikytų technologijų skaičius ir jų pritaikymo lygmuo yra gana skirtingas. Žinoma, skiriasi ir veiksmai, kurių reikėtų imtis norint paskatinti konkrečios KT mokslinius tyrimus ir plėtrą. Šie veiksmai priklauso nuo daugelio veiksnių, tokių kaip kalbos sudėtingumas ir jos vartotojų skaičius.

META-NET, Europos Komisijos finansuojamas kompetencijos tinklas, šioje Baltųjų knygų serijoje atliko turimų kalbos išteklių ir technologijų analizę, į kurią įtrauktos visos 23 oficialiosios bei kitos svarbios nacionalinės ir regioninės Europos kalbos (p. 85). Remiantis analizės rezultatais, konstatuotina, kad kiekvienos kalbos moksliniai tyrimai turi rimtų spragų. Ekspertų atlikta išsamesnė esamos padėties analizė ir įvertinimas galėtų padėti padidinti papildomų tyrimų poveikį ir sumažinti galimą riziką.

2011 metų lapkričio mėnesio duomenimis, META-NET tinklą sudaro 33 šalyse veikiantys 54 mokslinių tyrimų centrai (p. 81), bendradarbiaujantys su suinteresuotomis šalimis – verslo įmonių (programinės įrangos gamintojų, technologijų tiekėjų ir vartotojų), vyriausybės įstaigų, pramonės, tyrimų organizacijų, nevyriausybinių organizacijų, kalbos vartotojų bendruomenių ir Europos universitetų atstovais. Dirbdamas kartu su šiomis bendruomenėmis, META-NET kuria bendrą technologijų viziją ir rengia strateginę mokslinių tyrimų darbotvarkę 2020 metų daugiakalbei Europai.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 85). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of November 2011, META-NET consists of 54 research centres from 33 European countries (p. 81). META-NET is working with stakeholders from economy (software companies, technology providers, users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Šio dokumento autorės nuoširdžiai dėkoja vokiečių Baltosios knygos [1] autoriams, suteikusiems galimybę pasinaudoti medžiaga, kurioje aptariami bendrieji kalbos technologijų dalykai.

Šios Baltosios knygos sudarymas buvo finansuotas pagal Europos Komisijos septintąją bendrąją programą ir IKT politikos paramos programą: T4ME (subsidijų sutartis Nr. 249 119), CESAR (subsidijų sutartis Nr. 271 022), METANET4U (subsidijų sutartis Nr. 270 893) ir META-NORD (subsidijų sutartis Nr. 270 899).

The authors of this document are grateful to the authors of the White Paper on German [1] for permission to reuse selected language-independent materials from their document.

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

TURINYS CONTENTS

LIETUVIŲ KALBA SKAITMENINIAME AMŽIUJE

THE LITHUANIAN LANGUAGE IN THE DIGITAL AGE

SANTRAUKA

Per pastaruosius 60 metų Europa įgijo aiškią politinę ir ekonominę struktūrą, tačiau kultūros ir kalbų požiūriu ji vis dar labai skirtinga. Taigi nuo portugalų iki lenkų, nuo italų iki islandų – kiekvieną dieną bendraudami visuomenės, verslo ir politikos srityse Europos piliečiai neišvengiamai susiduria su kalbų barjerais. Europos Sąjungos institucijos per metus išleidžia apie milijardą eurų daugiakalbystės politikai įgyvendinti, t. y. versti rašytinius tekstus ir žodinę komunikaciją. Tačiau ar ši našta turėtų būti tokia milžiniška? Šiuolaikinės kalbos technologijos ir moksliniai kalbų tyrimai gali labai padėti griaunant tokius kalbų barjerus. Kalbos technologijos, įdiegtos išmaniuosiuose prietaisuose ir programose, ateityje galės padėti europiečiams lengvai susikalbėti ir bendradarbiauti, net jei jie kalba skirtingomis kalbomis.

Lietuvos ūkis turi didžiulės naudos iš Europos bendrosios rinkos: 2010 metais prekyba su Europos Sąjunga sudarė 61 proc., o su kitomis Europos šalimis – dar 3 proc. viso Lietuvos eksporto. Tačiau kalbų barjerai gali stabdyti verslą, ypač jei kalbame apie mažas ir vidutinio dydžio įmones, neturinčias lėšų pakeisti situaciją.

Alternatyva tokiai daugiakalbei Europai būtų leisti įsigalėti vienai kalbai, kuri ilgainiui pakeistų visas kitas kalbas. Tačiau tai sukeltų sunkumų įvairiakalbiams Europos piliečiams.

Klasikinis būdas įveikti kalbų barjerus – mokytis užsienio kalbų. Tačiau išmokti 23 oficialiąsias ir dar beveik 60 kitų Europos kalbų nesinaudojant technologijomis europiečiams būtų neįveikiama užduotis ir kliūtis siekiant Europos ekonominės, politinės ir mokslinės pažangos. Geriausia išeitis – kurti plačių galimybių technologijas (angl. *key enabling technology*). Jos suteikia Europos rinkos dalyviams didžiulio pranašumo ne tik Europos bendrojoje rinkoje, bet ir palaikant prekybinius ryšius su trečiųjų šalių besivystančiomis rinkomis. Pagaliau kalbos technologijų sprendiniai turėtų tapti tiltais, jungiančiais įvairias Europos kalbas.

Kalbos technologijos – ateities raktas.

Informacinės technologijos keičia mūsų kasdienį gyvenimą. Paprastai kompiuterius naudojame tekstams rašyti, redaguoti, skaičiuoti, ieškoti informacijos ir vis dažniau – klausytis muzikos, peržiūrinėti nuotraukas ir filmus. Su savimi nešiojamės kišeninius kompiuterius, kuriais galime skambinti, rašyti elektroninius laiškus, gauti informacijos ir susirasti pramogų, kur bebūtume. Kokia yra šio plataus informacijos, žinių ir kasdienio bendravimo skaitmeninimo įtaka mūsų kalbai? Ar mūsų kalba pasikeis, o gal iš viso išnyks?

Dauguma šiuo metu pasaulyje egzistuojančių 6 000 kalbų globalizuotoje skaitmeninėje informacinėje visuomenėje neišgyvens. Manoma, kad ne mažiau nei 2 tūkst. kalbų artimiausiais dešimtmečiais lemta išnykti. Kitos bus vartojamos šeimose ir miestų rajonuose, tačiau tikrai ne platesniame verslo ir mokslo pasaulyje. Kokios yra lietuvių kalbos galimybės išlikti? Kalbos statusas priklauso ne vien nuo ja kalbančių žmonių ar ja parašytų knygų, sukurtų kino filmų ir ja transliuojančių televizijos stočių skaičiaus, bet ir nuo kalbos vartojimo skaitmeninėje informacinėje erdvėje bei programinei įrangai kurti. Tai aktualu lietuvių kalbai, kuri yra viena iš ma-

žiau vartojamų, ne tokių patrauklių rinkos požiūriu Europos kalbų – ja kalba apie 4 mln. žmonių, dauguma jų gyvena Lietuvos Respublikoje. Lietuvių kalba turi valstybinės kalbos statusą, įtvirtintą Lietuvos Respublikos Konstitucijoje, šio statuso apsaugą ir valstybinės kalbos vartojimą reglamentuoja Valstybinės lietuvių kalbos įstatymas bei kiti teisės aktai. Be to, kalba, kaip kultūrinės tapatybės dalis, įtraukta į kultūrinio ir etninio paveldo apsaugos teisės aktus. „Lietuvos informacinės visuomenės plėtros" 2011–2019 metų programoje yra iškeltas strateginis tikslas – pagerinti Lietuvos gyventojų gyvenimo kokybę ir įmonių veiklos aplinką naudojantis informacinių ir ryšių technologijų (IRT) teikiamomis galimybėmis ir pasiekti, kad iki 2019 metų ne mažiau kaip 85 proc. Lietuvos gyventojų naudotųsi internetu. Šio tikslo prioritetas yra elektroninio turinio ir paslaugų plėtra, jų naudojimo skatinimas. Prioritetui pasiekti Lietuvos Vyriausybė kelia du uždavinius: 1. skaitmeninti Lietuvos kultūros paveldo objektus ir jų pagrindu kurti viešai prieinamus skaitmeninius produktus, taip užtikrinti skaitmeninio turinio išsaugojimą ir sklaidą elektroninėje erdvėje; 2. diegti lietuvių kalbos skaitmeninius produktus į IRT, siekiant užtikrinti visavertį lietuvių kalbos (rašytinės ir šnekamosios) funkcionavimą visose valstybės gyvenimo srityse. Ar šių politinių pastangų pakaks įtvirtinant lietuvių kalbą Europos daugiakalbėje informacinėje erdvėje?

Lietuvai tapus Europos Sąjungos nare, prasidėjo naujas lietuvių kalbos raidos etapas – įgytas oficialios Europos Sąjungos kalbos statusas užtikrino lietuvių kalbos vartojimą ir sklaidą Europos Sąjungos institucijose, paspartėjo kalbos išteklių bei technologijų, reikalingų visaverčiam kalbos funkcionavimui daugiakalbėje aplinkoje, kūrimas ir diegimas. Vis dėlto lietuvių kalba yra viena iš vadinamųjų „nekomercinių" Europos kalbų, todėl plėtojant kalbos technologijas ji susiduria su sunkumais ir problemomis, būdingomis mažiau vartojamų kalbų raidai. Šių technologijų plėtra labai priklauso nuo kitų šalių patirties ir jų paramos bei tarptautinio bendradarbiavimo. Kita vertus, kalbos technologijų plėtojimas yra svarbiausia lietuvių kalbos funkcionalumo, žinomumo ir studijų bei lietuviškos kultūros sklaidos daugiakalbėje Europoje stiprinimo proceso sudedamoji dalis. Lietuvių kalbos visavertis funkcionavimas skaitmeninėje erdvėje tapo ypač svarbiu lietuvių kalbos išlikimo ir sklaidos veiksniu. Informacinėje visuomenėje kalbos gyvybingumą ir patrauklumą lemia galimybės greitai ir patogiai keistis daugiakalbe informacija, gauti paslaugas ir pan. Informacinės technologijos lietuvių kalbai atveria naujus bendravimo, tekstų rengimo, informacijos sklaidos ir paieškos būdus. Šiuolaikinių komunikacijų greitis ir geografinė aprėptis palengvina bendravimą lietuvių kalba, daugėja lietuviško turinio ir paslaugų internete, kuriami įrankiai, padedantys vartoti taisyklingą kalbą, tenkinantys specialiuosius vartotojų poreikius ir pan. Kita vertus, pokyčiai šioje srityje tokie spartūs, kad lietuvių kalbos planavimas ir plėtra nebespėja laiku spręsti visų iššūkių. Vartotojams greičiau ir paprasčiau pasiekiami produktai ir informacija anglų kalba lemia palyginti menką lituanizuotos programinės įrangos populiarumą, lėtą kalbos technologijų ir įrankių diegimą bei sklaidą, nepakankamą skaitmeninių kalbos išteklių ir įrankių plėtrą.

Lietuvoje, kaip ir daugelyje Europos šalių, kalbos technologijų erdvė yra netolygiai plėtojama. Moksliniai tyrimai leido sėkmingai sukurti gana kokybišką programinę įrangą bazinei teksto analizei, pavyzdžiui, įrankius morfologinei ir sintaksinei analizei. Tačiau pažangesnių technologijų, kurioms reikia nuodugnesnio lingvistinio apdorojimo ir semantinių žinių, kol kas tėra tik užuomazgos. Parengta nemažai pirminių skaitmeninių kalbos išteklių (elektroninių žodynų, tekstynų, terminynų) ir pagrindinių kalbos analizės priemonių (morfologinių požymių nustatymo ir generavimo, rašybos tikrinimo įrankių), sukurtas lietuviškas sintezatorius, lietuvių ir anglų kalbų automatinio vertimo sistemos, sulie-