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

© Springer-Verlag Berlin Heidelberg 2012
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

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-

tuvinta programinė įranga, sukurtas originalus lietuviškas kompiuterinis šriftas *Palemonas*, pritaikytas mokslo reikmėms. Tačiau daugelį sukurtų išteklių, produktų ir sistemų reikia nuolat atnaujinti ir plėtoti, kad jie atitiktų kintančius vartotojų poreikius. Menkai išplėtoti semantikos tyrimai lėmė mažesnę kalbos generavimo, teksto interpretavimo ir teksto analizės pažangą. Nors sukauptos gana gausios ir išsamios leksikos duomenų bazės, tačiau nėra *WordNet*, tezauro ir pan. Taip pat trūksta reikiamo lygio kalbos technologijoms pritaikytos lietuvių kalbos gramatikos, sintaksiškai anotuotų tekstynų.

Kuriant išmanesnes ir sudėtingesnes priemones, tokias kaip automatinis vertimas, reikia išteklių ir technologijų, kurie apimtų daugiau lingvistinių aspektų ir leistų semantiškai nuodugniau analizuoti įvedamą tekstą. Gerindami kai kurių bazinių išteklių kokybę ir aprėptį, turėtume gebėti atverti naujas galimybes įsiveržti į pažangesnių technologijų taikymo sritis.

Kol kas lietuvių kalbai pritaikytų technologijų erdvė kokybiškai gana fragmentuota ir menkai sąveiki. Esami kalbos ištekliai, kurie galėtų būti pritaikyti kalbos technologijoms, sukurti atskirų institucijų, mokslininkų grupių ar verslo įmonių nesilaikant bendrų standartų, todėl jų pritaikomumas kalbos technologijoms yra ribotas arba ekonomiškai neefektyvus, turint galvoje išteklių pertvarkymą pagal naujus standartus. Šiuo metu Lietuvoje vykdoma keletas projektų, pagal kuriuos tarptautiniai standartai diegiami senesniuose ištekliuose (pvz., Dabartinės lietuvių kalbos tekstyne) ar kuriami nauji produktai. Didesnis sąveikumas leistų lengviau kurti bendrai Europos kalbinei erdvei būtinus integruotus produktus, tokius kaip daugiakalbiai automatinio vertimo įrankiai, žodynai, semantinės informacijos paieškos priemonės, mažintų lietuviškai kalbančios visuomenės atskirtį, didintų lietuvių kalbos tarptautinį prestižą ir prieinamumą.

Lietuvoje spartesnė informacinės visuomenės plėtra, susidomėjimas kalbos technologijomis ir išteklių kaupimas prasidėjo vos prieš keletą dešimtmečių, todėl norint sukurti tikrai veiksmingų, kasdieniam vartojimui skirtų kalbos technologijų, reikės atlikti nemažai mokslinių tyrimų. Nedidelė kalbos technologijų ir įrankių naudotojų rinka, neišplėtota ir susiskaidžiusi mokslo tyrimų ir studijų infrastruktūra, aiškių prioritetų ir koordinavimo trūkumas neskatina privataus verslo iniciatyvų. Šiuo metu kalbos technologijų srityje dirba keletas įmonių, verslas mažai užsako mokslinių tyrimų.

Kalbos technologijų būklė Lietuvoje teikia pagrindo nuosaikiam optimizmui. Lietuvos Respublikos Vyriausybė pabrėžia siekį užtikrinti kalbos technologijų plėtrą – tai rodo įvairių vyriausybinių institucijų ir Europos Sąjungos struktūrinių fondų finansuojamos programos, pagal kurias kuriamos ir tobulinamos kalbos technologijos. Proveržiai kalbos technologijų srityje skatintų jų diegimą pramonėje, padėtų plėsti ir gerinti viešąsias paslaugas ir pan., taip pat suteiktų galimybę lietuvių kalbą vartoti visose gyvenimo srityse ir komunikacijos terpėse.

Kalbos technologijos padeda Europai vienytis.

Europos kalboms sukurtų ir pritaikytų technologijų skaičius ir jų pritaikymo lygmuo yra gana skirtingas. Žinoma, skiriasi ir veiksmai, kurių reikėtų imtis norint paskatinti konkrečios kalbos technologijų mokslinius tyrimus ir plėtrą, pasirengimas taikyti kalbinius sprendinius ir mokslinių tyrimų lygis. Norint sukurti tikrai veiksmingų, kasdieniam vartojimui skirtų kalbos technologijų, Lietuvai reikės atlikti nemažai tyrimų ir sukurti papildomų išteklių, įrankių, integruoti juos, užtikrinant kuo didesnę sąveiką. Šie tikslai numatyti 2011 metais prasidėjusioje nacionalinėje programoje „Lietuvių kalba informacinėje visuomenėje“.

META-NET tinklo ilgalaikis uždavinys – pristatyti kokybiškas kalbos technologijas, taikomas visose Europos

kalbose, siekiant kultūrine įvairove pagrįstos politinės ir ekonominės vienybės. Šios technologijos padės sugriauti dabartinius barjerus ir nutiesti tiltus tarp Europos kalbų. Visos suinteresuotosios šalys – politikai, tyrėjai, verslo ir visuomenės atstovai – turi suvienyti savo pastangas, kurdami bendrą ateitį.

Ši Baltųjų knygų serija papildo kitus META-NET tinklo strateginius veiksmus, apžvelgiamus šio dokumento priede. Aktualios informacijos, pavyzdžiui, META-NET tinklo vizijos dokumento naujausią versiją arba strateginių mokslinių tyrimų darbotvarkę galima rasti META-NET tinklalapyje: http://www.meta-net.eu.

2

GRĖSMĖS KALBAI: IŠŠŪKIS KALBOS TECHNOLOGIJOMS

Gyvename skaitmeninės revoliucijos, turinčios didžiulį poveikį bendravimui ir visuomenės raidai, metu. Naujausi skaitmeninės informacijos pateikimo ir bendravimo technologijų išradimai kartais prilyginami Johanneso Gutenbergo spausdinimo mašinos išradimui. Ką ši analogija gali byloti apie Europos informacinės visuomenės ir ypač apie mūsų kalbų ateitį?

Esame skaitmeninės revoliucijos liudininkai. Ši revoliucija prilygsta Gutenbergo spausdinimo mašinos išradimui.

Gutenbergo išradimas lėmė svarbius proveržius komunikacijos ir žinių mainų srityje, tokius kaip Martino Lutherio atliktą Biblijos vertimą į gimtąją kalbą ir kt. Vėlesniais amžiais buvo sukurta kultūrinių metodologijų, lengvinančių kalbos apdorojimą ir žinių mainus:

- Ortografinis ir gramatinis labiausiai paplitusių kalbų standartizavimas suteikė galimybių sparčiau plisti naujoms mokslinėms ir intelektinėms idėjoms.
- Bendrinių kalbų susiformavimas suteikė piliečiams galimybių bendrauti tam tikruose (dažniausiai politinių valstybių sienų apibrėžtuose) plotuose.
- Kalbų mokymas ir vertimai iš vienos kalbos į kitą suteikė galimybių keistis informacija skirtingomis kalbomis.
- Redagavimo ir bibliografinės gairės užtikrino spausdintos informacijos kokybę bei prieinamumą.
- Skirtingų žiniasklaidos priemonių – laikraščių, radijo, televizijos, knygų ir kitokių – atsiradimas tenkino skirtingus komunikavimo poreikius.

Per pastarąjį dvidešimtmetį informacinės technologijos padėjo automatizuoti ir palengvinti daugybę procesų:

- Kompiuterinė programinė įranga pakeitė spausdinimą mašinėle ir rankinį tekstų rinkimą.
- Programa *Microsoft PowerPoint* pakeitė projektoriumi rodomas skaidres.
- Elektroniniu paštu siųsti ir gauti dokumentus galime sparčiau nei faksu.
- Programa *Skype* teikia galimybę pigiai skambinti internetu ir rengti virtualius susitikimus.
- Garso ir vaizdo kodavimo formatai palengvina daugialypės terpės turinio keitimą.
- Paieškos sistemos teikia galimybių naudojant esminius žodžius pasiekti reikiamus tinklalapius.
- Internetinės paslaugos, tokios kaip *Google Translate*, teikia galimybių sparčiai, nors ir ne visada tiksliai, versti tekstus.
- Socialinės terpės, tokios kaip *Facebook*, *Twitter* ir *Google+*, lengvina bendravimą, bendradarbiavimą ir keitimąsi informacija.

Nors tokios priemonės ir programos yra naudingos, kol kas jos nepajėgia išlaikyti visapusiškai tvarios daugiakalbės Europos visuomenės, kurioje galėtų vykti netrikdoma informacijos ir prekių sklaida.

2.1 KALBŲ BARJERAI – KLIUVINYS EUROPOS INFORMACINEI VISUOMENEI

Negalime tiksliai numatyti, kokia bus ateities informacinė visuomenė. Tačiau tikėtina, kad komunikacinių technologijų revoliucija skirtingomis kalbomis kalbantiems žmonėms teikia naujų būdų suartėti. Dėl to žmonės patiria spaudimą mokytis naujų kalbų, o programinės įrangos kūrėjai – kurti naujus technologinius sprendinius, kurie užtikrintų susikalbėjimą ir prieigą prie bendrų žinių. Dėl naujų medijų rūšių pasaulinėje ekonominėje ir informacinėje erdvėje sąveikauja vis daugiau kalbų ir jomis kalbančių žmonių informacijos.

> Pasaulinėje ekonominėje ir informacinėje erdvėje sąveikauja vis daugiau kalbų ir jomis kalbančių žmonių informacijos.

Šiuo metu išpopuliarėjusios bendrauti skirtos priemonės (*Wikipedia*, *Facebook*, *Twitter*, *YouTube* ir visai neseniai – *Google+*) yra vos ledkalnio viršūnė.
Šiandien galime akimirksniu iš kito pasaulio krašto parsisiųsti keletą gigabaitų teksto ir tik po to pamatyti, kad jis parašytas kalba, kurios nesuprantame. Anot vienos iš Europos Komisijos pastarojo laikotarpio ataskaitų, 57 proc. interneto naudotojų Europoje perka prekes ir paslaugas ne savo gimtąja kalba (anglų kalba yra labiausiai paplitusi užsienio kalba, po jos – prancūzų, vokiečių ir ispanų kalbos). 55 proc. naudotojų skaito internete pateikiamą informaciją užsienio kalbomis ir vos 35 proc. užsienio kalbomis rašo elektroninius laiškus arba komentarus interneto tinklalapiuose [2]. Prieš kelerius metus anglų kalba gal ir buvo interneto *lingua franca* – didžioji dalis internete pateikiamos informacijos buvo ja parašyta – tačiau dabar padėtis radikaliai pasikeitė. Internetą užplūdo informacija kitomis Europos, taip pat Azijos ir Vidurio Rytų kalbomis.

Neįtikėtina, bet atrodo, kad ši skaitmeninė takoskyra dėl kalbų barjerų nesulaukė labai daug visuomenės dėmesio. Tačiau ji kelia itin aktualų klausimą: kurios Europos kalbos klestės į tinklą sujungtoje informacinėje ir žinių visuomenėje, o kurioms lemta išnykti?

2.2 GRĖSMĖ KALBOMS

Nors spausdinimo mašina ir palengvino keitimąsi informacija, tai lėmė kai kurių Europos kalbų išnykimą. Spaudinių regioninėmis ir mažumos kalbomis buvo mažai, o tokios kalbos kaip kornų ir dalmatų išvis neturėjo raštijos ir tai labai apribojo jų vartojimo sritis. Ar ir interneto poveikis mūsų kalboms bus toks pats?
Europoje kalbama apie 80 kalbų. Jos yra vienas iš gausiausių ir svarbiausių regiono kultūros turtų bei esminė Europos unikalaus socialinio modelio dalis [3]. Nors tokios kalbos kaip anglų ir ispanų greičiausiai išliks besivystančioje skaitmeninėje rinkoje, daugelis Europos kalbų gali tapti nereikalingos į tinklą susietoje visuomenėje. Tai susilpnintų Europos padėtį pasaulyje ir pakenktų strateginiam tikslui užtikrinti kiekvieno Europos piliečio neatsižvelgiant į vartojamą kalbą galimybes būti visateisiu Sąjungos nariu.

> Europos kalbų įvairovė – svarbi jos kultūros turto dalis.

Anot UNESCO ataskaitos apie daugiakalbystę, kalbos sudaro esminę terpę džiaugtis pamatinėmis teisėmis, tokiomis kaip teisė į politinę raišką, švietimą ir dalyvavimą visuomenės gyvenime [4].

2.3 KALBOS TECHNOLOGIJOS – NAUJŲ GALIMYBIŲ KŪRĖJOS

Anksčiau investicijos ir pastangos išsaugoti kalbų įvairovę buvo skiriamos kalbų mokymui ir vertimams. Apy-

tikriais skaičiavimais, 2008 m. Europos vertimų raštu ir žodžiu, programinės įrangos lokalizavimo ir internetinių tinklalapių vertimo rinkos vertė sudarė 8,4 mlrd. eurų, tikėtinas jos metinis augimas yra 10 proc. [5]. Deja, šie skaičiai atspindi tik nedidelę dalį dabartinio ir būsimo poreikio bendrauti skirtingomis kalbomis. Pagrindinis būdas užtikrinti kalbų vartojimo plėtrą ir sklaidą rytdienos Europoje – pasitelkti tinkamas technologijas. Juk būtent technologijos padeda mums tenkinti transporto, energetinius ir kitokius poreikius.

Europai reikia stabilių, lengvai prieinamų kalbos technologijų, pritaikytų visoms Europos kalboms.

Skaitmeninės KT (skirtos visoms rašomosios ir šnekamosios kalbos formoms) padeda žmonėms bendradarbiauti, dirbti, dalytis žiniomis ir dalyvauti socialinėse bei politinėse diskusijose, nepaisant kalbų barjerų ar darbo kompiuteriu įgūdžių. Šios technologijos neretai būna nepastebimos sudėtingų programinės įrangos sistemų dalys, padedančios:

- rasti informacijos naudojantis internetinės paieškos sistema;
- patikrinti, ar tekste nėra rašybos, skyrybos ir gramatinių klaidų;
- peržiūrėti produktų rekomendacijas internetinėje parduotuvėje;
- suprasti automobilio navigacinės sistemos žodines instrukcijas;
- versti tinklalapius naudojant internetines programas.

KT sudaro aibė pagrindinių programų, veikiančių didesnėse programinės įrangos sistemose. META-NET Baltųjų knygų tikslas – apžvelgti, kaip šios pagrindinės programos yra išplėtotos ir pritaikytos kiekvienai iš Europos kalbų.

Jei norės neprarasti pozicijos tarp pasaulio inovacijų lyderių, Europai reikės stabilių, lengvai prieinamų ir į svarbiausias programinės įrangos terpes integruotų KT, pritaikytų visoms Europos kalboms. Neturėdami reikiamo lygio KT, artimiausioje ateityje nesugebėsime užtikrinti kalbos vartotojams interaktyvaus daugiakalbio bendravimo daugialypėje terpėje.

2.4 KALBOS TECHNOLOGIJŲ GALIMYBĖS

Spausdinto žodžio pasaulyje didysis technologinis proveržis buvo galimybė greitai kopijuoti vaizdą ar tekstą tinkamo galingumo spausdinimo mašina. Žmonėms teko sunkus darbas – ieškoti žinių, jas skaityti, versti ir apibendrinti. Mes turėjome laukti Thomo Edisono išradimų, padėjusių įrašyti šnekamąją kalbą. Ši technologija suteikė galimybių gaminti analogines kopijas.

Dabar turint skaitmeninių technologijų galima automatizuoti vertimą, informacijos kūrimą ir žinių tvarkymą visomis Europos kalbomis. Be to, jos gali būti pritaikytos kuriant intuityvias kalbos / šnekos pagrindu veikiančias buities elektronikos, mechanizmų, transporto priemonių, kompiuterių ir robotų sąsajas. Verslo ir pramonės programos vis dar yra ankstyvosios plėtros lygmens, tačiau mokslinių tyrimų ir taikomosios veiklos pažanga atveria naujų galimybių. Pavyzdžiui, konkrečių sričių tekstų automatinis vertimas jau dabar yra gana tikslus, o eksperimentinės programos teikia galimybių tvarkyti informaciją ir žinias bei kurti naują turinį daugeliu Europos kalbų.

Kaip ir daugumos technologijų atveju, pirmosios kalbinės programos, pavyzdžiui, balsu valdomos naudotojų sąsajos ir dialogų sistemos, buvo skirtos itin specializuotoms sritims, jų veikimas gana ribotas. Tačiau švietimo ir pramogų rinkoje esama didžiulių galimybių diegti KT žaidimams, kultūrinio paveldo sąvokoms, mokomosioms ir pramoginėms priemonėms, bibliote-

koms, imitacinėms aplinkoms ir mokomosioms programoms. Mobiliosios informavimo paslaugos, kompiuterinė programinė kalbų mokymo įranga, e.mokymo aplinka, vertinimo priemonės ir programos plagiatui aptikti, tai vos kelios sritys, kuriose KT gali būti ypač vertingos. Tokių daugialypės terpės bendravimo platformų kaip *Twitter* ir *Facebook* populiarumas leidžia manyti, kad ateityje prireiks ir pažangių KT, skirtų stebėti dalyvių žinutes, apibendrinti diskusijas, rodyti nuomonių tendencijas, aptikti emocingus atsakymus, nustatyti autorių teisių pažeidimus arba susekti netinkamo naudojimo atvejus.

Kalbos technologijos padeda įveikti kalbų įvairovės „negalią".

KT teikia daug galimybių Europos Sąjungai. Jos gali padėti išspręsti sudėtingą Europos daugiakalbystės problemą – skirtingos kalbos gali kartu gyvuoti Europos įmonėse, organizacijose ir mokyklose. Gyventojai turi bendrauti be Europos bendrąją rinką skaidančių kalbų barjerų, o KT gali padėti jiems įveikti šią kliūtį bei teikti daugiau galimybių laisvai ir viešai kalbėti savomis kalbomis. Žvelgiant dar toliau į priekį, Europos novatoriškos daugiakalbės technologijos pažymės gaires mūsų tarptautiniams partneriams, pradėsiantiems kurti technologijas savoms daugiakalbėms bendruomenėms. KT gali būti vertinamos kaip „pagalbinės", padedančios visiems įveikti kalbų įvairovės „negalią" ir suteikiančios skirtingoms kalbos vartotojų bendruomenėms daugiau galimybių bendrauti. Galiausiai viena iš aktualių mokslinių tyrimų sričių yra KT pritaikymas vykdant gelbėjimo operacijas nelaimių zonose, kur kiekvienas veiksmas gali lemti gyvybę ir mirtį: ateityje sumanieji robotai, galintys bendrauti įvairiomis kalbomis, turės daug daugiau galimybių gelbėti gyvybes.

2.5 IŠŠŪKIAI, KURIUOS TURI ĮVEIKTI KALBOS TECHNOLOGIJOS

Nors KT pastaraisiais metais padarė nemažą pažangą, dabartiniai technologinės pažangos ir produkcijos naujovių taikymo tempai yra pernelyg lėti. Plačiau taikomos tekstų rašymo sistemų rašybos ir gramatikos tikrintuvės paprastai yra vienakalbės, o ir įdiegtos toli gražu ne visoms kalboms.

Šiuo metu technologinė pažanga yra pernelyg lėta.

Internetinės automatinio vertimo paslaugos yra naudingos, kai norima greitai gauti apytikslį dokumento turinio vertimą, tačiau susiduriama su sunkumais, kai reikia itin tikslaus ir išsamaus vertimo. Dėl natūraliosios kalbos sudėtingumo kalbų modeliavimas programinėje įrangoje ir jų testavimas realioje aplinkoje yra ilgas, brangus procesas, kurį reikia nuolat finansuoti. Vis dėlto Europa privalo išsaugoti lyderės pozicijas spręsdama daugiakalbės bendruomenės technologines problemas ir rasti naujų būdų spartinti plėtrą visose srityse. Metodai galėtų būti susiję tiek su pažanga skaitmeninėje (kompiuterių) srityje, tiek su tokiomis metodikomis kaip individualių užduočių perdavimas žmonių grupėms ar bendruomenei (angl. *crowdsourcing*).

2.6 KALBOS ĮVALDYMAS: ŽMONĖS IR MAŠINOS

Norėdami suprasti, kaip kompiuteriai „išmoksta" kalbą ir kodėl juos taip sunku užprogramuoti ją vartoti, trumpai pažvelkime į tai, kaip pirmąją ir antrąją kalbą išmoksta žmonės, o po to – kaip veikia KT sistemos.
Žmonės įgyja kalbos įgūdžių dviem skirtingais būdais. Kūdikiai išmoksta kalbą girdėdami, kaip bendrauja jų

tėvai, broliai, seserys ir kiti šeimos nariai. Dvejų metų amžiaus vaikai pradeda patys kalbėti – iš pradžių tai būna atskiri žodžiai ir trumpos frazės. Tai įmanoma todėl, kad žmonės genetiškai gali mėgdžioti ir praktiškai taikyti, ką išgirdo.

Vyresnio amžiaus vaikams išmokti antrąją kalbą yra kiek sunkiau, ypač jeigu vaikas auga bendruomenėje, kuriai ta antroji kalba nėra gimtoji. Mokyklose paprastai mokoma užsienio kalbos gramatinės struktūros, žodyno ir rašybos taisyklių, t. y. atliekami pratimai, kuriais kalbos žinios įtvirtinamos pagal abstrakčias taisykles, lenteles ir pavyzdžius.

Dviejų pagrindinių tipų KT sistemos kalbos išmoksta panašiais būdais. Statistiniais (arba pagrįstais duomenimis) metodais duomenys surenkami iš gausybės konkrečių tekstų pavyzdžių. Vienakalbiai tekstai gali būti naudojami tos pačios kalbos mokymo tikslais, pavyzdžiui, tikrinti, ar nėra rašybos klaidų, tačiau automatinio vertimo sistemai „išmokyti“ būtini lygiagretūs tekstai dviem ar daugiau kalbų. Iš tokių tekstų automatinio vertimo algoritmas „išmoksta“ žodžių, trumpų frazių ir visų sakinių vertimo modelius.

Žmonės įgyja kalbos įgūdžių dvejopai: mokydamiesi iš pavyzdžių ir mokydamiesi pagrindinių kalbos taisyklių.

Tokiam statistiniais metodais pagrįstam vertimui gali prireikti milijonų sakinių, o vertimo kokybė gerėja didėjant išanalizuotų tekstų kiekiui. Tai viena iš priežasčių, kodėl paieškos sistemų teikėjai pageidauja sukaupti kaip galima daugiau rašytinės informacijos. Tekstų rašymo ir tokių paslaugų kaip *Google Search* ir *Google Translate* rašybos klaidų taisymo funkcijos pagrįstos statistiniais metodais. Didžiausias statistikos pranašumas yra tas, kad mašina „mokosi“ greitai, įveikdama nesibaigiančius mokomuosius ciklus, nors kartais kokybė gali būti labai įvairi.

Antrasis KT ir ypač automatinio vertimo metodas yra kurti taisyklėmis pagrįstas sistemas. Lingvistikos, kompiuterinės lingvistikos ir kompiuterių mokslo ekspertai visų pirma turi užkoduoti gramatinę analizę (vertimo taisykles) ir sudaryti žodyno sąrašus (leksikonus). Tam reikia daug pastangų, intensyvaus darbo ir laiko. Kai kurios pagrindinės taisyklėmis pagrįstos automatinio vertimo sistemos buvo tobulinamos daugiau nei du dešimtmečius. Pagrindinis taisyklėmis pagrįstų sistemų pranašumas yra tas, kad ekspertai gali geriau kontroliuoti kalbos apdorojimą. Tai leidžia nuosekliai taisyti programinės įrangos klaidas ir palaikyti grįžtamąjį ryšį su naudotoju, ypač tuo atveju, kai taisyklėmis pagrįstos sistemos taikomos kalbų mokymui. Tačiau dėl didelių finansinių sąnaudų taisyklėmis pagrįstų KT turi susikūrusios tik didžiosios kalbos.

Kadangi paprastai statistinių ir taisyklėmis pagrįstų sistemų pranašumai ir trūkumai kompensuoja vieni kitus, šiuo metu atliekami tyrimai siekiant sukurti hibridinius būdus, sujungsiančius šias dvi metodologijas. Tačiau kol kas didesnės sėkmės pasiekta ne taikant šiuos metodus verslo poreikiams, bet tyrimų laboratorijose.

Taigi galima konstatuoti, kad didžioji dalis programų, kurias šiuo metu naudoja informacinė visuomenė, labai priklauso nuo KT. Tai ypač pasakytina apie Europos ekonominę ir informacinę erdvę su jos daugiakalbe bendruomene. Nors pastaraisiais metais KT srityje padaryta reikšminga pažanga, vis dar egzistuoja daugybė galimybių gerinti KT sistemų kokybę. Kituose skyriuose aptarsime lietuvių kalbos vietą Europos informacinėje visuomenėje ir įvertinsime dabartinę lietuvių KT būklę.

3

LIETUVIŲ KALBA EUROPOS INFORMACINĖJE VISUOMENĖJE

3.1 BENDRIEJI DUOMENYS

Lietuvių kalba yra viena iš mažiau vartojamų Europos kalbų – ja kalba apie 4 mln. žmonių, dauguma jų gyvena Lietuvos Respublikoje. Valstybinė lietuvių kalba yra bendra rašomoji ir šnekamoji visiems Lietuvos Respublikos piliečiams, kurių, 2011 m. duomenimis, yra apie 3,2 mln., iš jų lietuvių tautybės – apie 2,7 mln.

84 proc. Lietuvos gyventojų yra lietuviai, 6,1 proc. – lenkai, 4,9 proc. – rusai, 1,1 proc. – baltarusiai, 0,6 proc. – ukrainiečiai, dar po 0,1 proc. sudaro žydų, vokiečių, latvių, totorių, karaimų kilmės piliečiai. Be to, Lietuvoje gyvena apie 3 tūkst. romų bendruomenė, kurios didžiausia koncentracija yra Vilniaus regione (2001 m. surašymo duomenimis). Deja, nuo 2007 m. Lietuvos gyventojų skaičius kasmet mažėja. Šiuos pokyčius lemia mažėjantis gimstamumas bei emigracija, mažinanti Lietuvoje gyvenančių kalbos vartotojų skaičių.

Lietuvių kalba yra viena iš mažiau vartojamų Europos kalbų. Ja kalba vos apie 4 mln. žmonių, dauguma jų gyvena Lietuvos Respublikoje.

Kiek lietuviškai kalbančiųjų yra pasaulyje, gana sudėtinga nustatyti. Spėjama, kad užsienyje gali gyventi per 500 tūkst. lietuviškai kalbančiųjų, kituose šaltiniuose nurodoma, kad ne mažiau nei 15 proc. kalbėtojų. Lietuvių kalba šneka lietuvių tautinės mažumos, gyvenančios Baltarusijoje, Lenkijoje, Latvijoje, bei didelės emigrantų bendruomenės JAV, Kanadoje, Jungtinėje Karalystėje, Airijoje, Ispanijoje, Pietų Amerikoje ir kitur. Pagal kalbančiųjų skaičių lietuvių kalba užima 144 vietą pasaulyje.

Lietuvių kalba priklauso indoeuropiečių kalbų šeimos baltų šakai. Jos artimiausia giminaitė yra latvių kalba, kuria kalbama kaimyninėje Latvijoje.

Pagal socialinės Europos kalbų raidos istoriją, kalbas skirstant į dominuojančiąsias ir dominuojamąsias, lietuvių kalba priskirtina prie antrųjų. Dominuojančiosios kalbos vieną tarmę bendrinėms kalboms formuoti paprastai buvo pasirinkusios ne vėliau kaip Renesanso laikotarpiu (anglų, prancūzų, italų, portugalų), o dominuojamosios susiformavo XIX amžiuje, Tautų pavasario metu (bulgarų, kroatų, lietuvių, slovakų). Bendrinė lietuvių kalba susiformavo XIX amžiaus pabaigoje – XX amžiaus pradžioje.

Lietuvių kalbai būdinga didelė regioninių kalbos atmainų įvairovė. Dvi pagrindinės tarmės – aukštaičių ir žemaičių – skiriasi ne tik fonetinėmis ypatybėmis, bet ir gramatika bei leksika. Šios tarmės skirstomos į keturiolika stambesnių regioninių patarmių, o šios – į smulkesnių teritorinių vienetų šnektas. Patarmės viena nuo kitos skiriasi garsais, žodžių formomis ir kitokiomis savybėmis.

Bendrinė lietuvių kalba susiformavo XX amžiaus pradžioje vienos iš aukštaičių patarmių pagrindu, tačiau re-

gioninis tapatumas ir tarminiai skirtumai vis dar labai ryškūs.

Nuo 1995 metų gestų kalba oficialiai pripažinta Lietuvos Respublikos kurčiųjų gimtąja kalba. Nuo to laiko lietuvių gestų kalba vystosi kaip nepriklausoma kalba.

3.2 LIETUVIŲ KALBOS YPATYBĖS

XIX amžiuje indoeuropeistai išgarsino stebėtiną lietuvių kalbos panašumą į sanskritą, ja imta didžiuotis kaip mažiausiai pakitusios struktūros gyvąja indoeuropiečių kalba. Dėl tos priežasties besimokantys klasikines Europos kalbas (lotynų, senąją graikų) lengviau supranta lietuvių kalbos gramatiką. Lietuvoje didžiuojamasi prancūzų lingvisto Antoine'o Meillet posakiu, jog kiekvienas, norintis išgirsti, kaip kalbėjo indoeuropiečių protėviai, turi važiuoti pasiklausyti lietuvio valstiečio.

Lietuvių kalba – pati konservatyviausia iš indoeuropiečių gyvųjų kalbų, jai pavyko geriausiai išsaugoti daugelį savo archajiškų savybių. Tipologijos požiūriu, lietuvių kalba yra svarbi dėl daugybės unikalių savybių, įskaitant gausias kaitybos formas, charakteringą toninio ir dinaminio kirčio sintezę bei itin įvairialypę žodžių tvarką, atspindinčią sudėtingus diskurso bendravimo ir sintaksinio lygmens santykius.

Rašomojoje lietuvių kalboje milijonus žodžių sudaro net 32 raidės. Būtent tiek jų ir yra norminės lietuvių kalbos abėcėlėje. Šį skaičių nustatė Jonas Jablonskis savo „Lietuviškos kalbos gramatikoje" (1901 m.). Taigi dabartinei abėcėlei yra daugiau nei šimtas metų, tačiau jos atsiradimo istorija apima dar ilgesnį laiką. Lietuvių kalbos abėcėlė yra paremta lotynų kalbos rašmenimis, nors joje yra ir unikalių ženklų – kai kurie iš jų yra originalūs (pavyzdžiui, raidė „ė"). Kiti yra pasiskolinti iš užsienio kalbų (pavyzdžiui, „š", „ž" – iš čekų, ar „ą", „ę" – iš lenkų kalbos). Tačiau kol kas sunku išspręsti sudėtingą lietuvių kalbos kirčiuotų balsių problemą, kuri tampa itin aktuali norint perkelti įvairius žodynus ar kirčiuotus tekstynus į skaitmeninę erdvę – kirtis lietuvių kalboje yra skiriamasis (distinktyvinis), jo vieta nėra fiksuota, taigi jis gali lemti leksinę ar gramatinę žodžio reikšmę (pvz., *nãmo* vns. kilm. – *namõ* adv.). Be to, visi ilgieji skiemenys turi vieną iš dviejų priegaidžių, kurios taip pat skiria žodžių reikšmes (pvz., *áukštas* – *aũkštas* ir pan.). Į šiuos dalykus turi atsižvelgti ir garso technologijų kūrėjai.

Lietuvių kalboje, kuriai būdingos linksniuotės, dauguma žodžių formų yra sudaromos naudojant afiksus, t. y. galūnes. Galūnės yra svarbiausia priemonė pažymėti žodžių sintagminius santykius sakinyje ir (arba) žodžių formų santykius paradigmoje. Galūnės dažniausiai yra nevienareikšmės, t. y. galūnė apima dvi ar daugiau gramatinių funkcijų, ir dėl to žodžio forma priskiriama tokiam pačiam morfologinių kategorijų kiekiui.

Priesagos taip pat plačiai vartojamos lietuvių kalbos žodžių formoms sudaryti. Jos dažniausiai reiškia žodžių formų paradigminius, o ne sintagminius ryšius. Kaitybinės priesagos naudojamos pažymėti būdvardžių ir daugelio prieveiksmių laipsnius, veiksmažodžių laikus ir nuosakas bei neasmenines veiksmažodžių formas: bendratį, dalyvius ir veiksmažodinius prieveiksmius (būdinius).

Sudarant naujas žodžių formas, afiksai (ypač veiksmažodžių paradigmoje) dažnai derinami su šaknies balsių kaita.

Be paprastųjų (sintetinių) žodžių formų, sudaromų su afiksais, paradigmoje galima aptikti ir aprašomųjų (analitinių) žodžių formų, kurias sudaro pagrindinis ir pagalbinis žodis.

Pagal bendrąsias morfologines, sintaksines ir semantines savybes žodžiai skirstomi į gramatines klases, tradiciškai vadinamas kalbos dalimis. Lietuvių kalboje skiriama 11 kalbos dalių: daiktavardis, būdvardis, skaitvardis, įvardis, veiksmažodis, prieveiksmis, dalelytė, prielinksnis, jungtukas, jaustukas ir ištiktukas.

Sintaksinis ryšys apibrėžia betarpišką santykį tarp sakinio žodžių formų, žodžių grupių ir dėmenų. Lietuvių kalboje sintaksinius ryšius reiškia galūnės ir, kiek rečiau,

kaitybinės priesagos, kurias dažnai papildo struktūriniai žodžiai – prielinksniai, jungtukai ir dalelytės. Žodžių eilės tvarka nėra tokia svarbi gramatiniams ryšiams parodyti. Pavyzdžiui, žodžių eiliškumas parodo būdvardžio sintaksinę funkciją tokiuose žodžių junginiuose kaip *Gražios gėlės* (pažyminys) ir *Gėlės gražios* (predikatyvas). Sakinyje žodžių formas į grupes susieja intonacija, sustiprinanti jų sintaksinį ryšį (betarpiškai susijusios žodžių formos paprastai sudaro intonacinį vienetą); be to, ji parodo ir sakinių tipus. Skiriami trijų tipų sintaksiniai ryšiai: tarpusavio sąsaja, derinimas ir šliejimas. Lietuvių kalbos žodžių tvarka yra laisva, taigi tą pačią mintį įmanoma pasakyti įvairiais būdais (nors kai kurios struktūros gali būti vartojamos tik stilistiniais sumetimais). Ką jau kalbėti apie eliptinius sakinius, kuriuose praleisti žodžiai gali būti tik numanomi iš konteksto. Be to, sakiniai gali būti labai ilgi ir sudėtingos struktūros, tai taip pat sunkina automatinį apdorojimą.

Gausu daugiareikšmių žodžių, todėl besimokančiam lietuvių kalbos gali būti sudėtinga atpažinti vieno ar kito žodžio reikšmę ir formą.

Nemažai gramatinių formų, pavyzdžiui, lietuvių kalbos vardažodžiai turi linksnio, giminės ir skaičiaus gramatines kategorijas. Be to, būdvardžiai gali būti įvardžiuotiniai / neįvardžiuotiniai, kaitomi laipsniais, būti bevardės giminės ir pan.

Dar sudėtingesnis lietuvių kalbos veiksmažodis, nes yra asmenuojamųjų ir neasmenuojamųjų formų, kurios turi ir vardažodžių, ir veiksmažodžių savybių, t. y. jos kaitomos skaičiais, laikais, rūšimis, linksniais, giminėmis.

Kai kurios lietuvių kalbos ypatybės sunkina skaitmeninį kalbos apdorojimą.

3.3 DABARTINĖ RAIDA

Nors turinti raštijos tradiciją nuo XVI amžiaus, lietuvių kalba buvo sunorminta tik XX amžiaus pradžioje – tuo metu parašyta norminamoji lietuvių kalbos gramatika, pradėtas leisti žodynas (tezauras), kurio paskutinis 20 tomas pasirodė 2002 metais.

Tik pradėjusi įsitvirtinti bendrinė lietuvių kalba patyrė nemažai iššūkių. Nuo pačios raštijos pradžios jai didelę įtaką darė slavų kalbos. Sovietiniu laikotarpiu buvo remiamas ir skatinamas rusų kalbos mokymasis ir vartojimas, o lietuvių kalbos vaidmuo kai kuriose srityse, pavyzdžiui, mokslo ir valstybės administravimo, buvo ribojamas. Dabartinė vyresnioji ir vidurinioji gyventojų karta išaugo apsupta rusų kalbos ir kultūros. Kadangi svetimos kilmės žodžiai yra ne vien kalbos, bet ir visuomenės gyvenimo atspindys, tuo metu į lietuvių kalbą pateko nemažai skolinių, ypač terminų, administracinės kalbos konstrukcijų ir pan. Rusų kalbos įtaka vis dar stipriai juntama kai kuriose periferinėse kalbos atmainose: žargone, nenorminėje leksikoje ir pan.

Pastarųjų keliolikos metų Lietuvos politiniai, ekonominiai, socialiniai ir kultūriniai procesai lėmė itin staigius lietuvių kalbos žodyno pokyčius. Per 1993–1997 m. spaudoje užfiksuota daugiau nei 700 naujų svetimų žodžių šaknų [6]. Dažniausiai tai skoliniai iš anglų kalbos arba žodžiai, į lietuvių kalbą patekę per šią kalbą. Tai lėmė po nepriklausomybės atkūrimo prasidėjusi sparti informacinių technologijų plėtra bei naujų kultūrinių, socialinių ir ekonominių galimybių atsiradimas. Dabartinio lietuvių kalbos tekstyno duomenimis, vien per 1991–1996 m. lietuvių kalbos žodyną papildė per 10 tūkst. naujažodžių, tikėtina, kad dabar šie procesai dar spartesni. Nuo 1990 m. informacinę erdvę užplūdo anglakalbė, ypač amerikietiškoji, populiarioji kultūra: serialai, laidos, muzika ir pan. Nors užsienio filmai, serialai ar televizijos laidos verčiami į lietuvių kalbą, toks kultūrinis pasikeitimas turėjo nemažos įtakos lietuvių kalbai ir kultūrai. Anglų kalba laikoma svarbiausia pasaulinės materialinės ir intelektinės rinkos tarpininke, tad jos vaidmuo didėja ir Lietuvos ekonominiame, socialiniame ir kultūriniame gyvenime: stiprėja ne tik anglų

kalbos mokymosi, bet ir specialybės įgijimo, darbo, intelektinės kūrybos šia kalba motyvacija. Šiuo metu jaunajai kartai artimesnė ir didesnį prestižą turi anglų kalba kaip *lingua franca*, su kuria siejama ne tik kultūrinė integracija, bet ir studijos, karjeros perspektyvos ir pan. Kol kas nėra atlikta pakankamai tyrimų, bet tikėtina, kad daugiausia anglų kalbos skolinių ar ištisų konstrukcijų vartojama jaunimo, ypač priklausančio tam tikroms subkultūroms, kalboje.

Lietuvoje anglų kalba veikia tas pačias kalbos vartojimo sritis, kaip ir kitur. Bene dažniausiai anglų kalba vartojama ten, kur kreipiamasi į jaunimo auditoriją, pavyzdžiui, 75 proc. kino anonsų pateikiama angliškai arba yra mišrūs. Kitur padėtis yra geresnė, pavyzdžiui, Lietuvos televizijos transliuoja daugiau nei 60 proc. lietuviškos reklamos, o mišriose reklamose angliškas dažnai būna tik produkto pavadinimas [6].

Nemažą rūpestį kelia mokslo kalbos raida. Siekiant tarptautinio pripažinimo ir sklaidos, kai kurios mokslo sritys beveik nepublikuoja savo tyrimų rezultatų Lietuvoje lietuvių kalba. Mokslo tarptautiškumas ypač skatinamas, nepaisant nuogąstavimų, kad taip skurdinama lietuviška mokslo terminija, lietuvių kalba išstumiama iš specifinių vartojimo sričių, menkėja motyvacija tobulinti specialybės kalbą aukštosiose mokyklose.

Diskutuojant, kokia turėtų būti tolesnė lietuvių kalbos raida ir perspektyvos, lietuvių kalbos įsitvirtinimas informacinėje visuomenėje būtų geras argumentas, kad ji yra moderni ir funkcionali komunikacijos priemonė.

3.4 KALBOS PADĖTIS IR VARTOJIMAS LIETUVOJE

Lietuvių kalba turi valstybinės kalbos statusą, įtvirtintą Lietuvos Respublikos Konstitucijoje. Šio statuso įgyvendinimą, t. y. valstybinės kalbos vartojimą viešajame gyvenime, jos apsaugą ir kontrolę, taip pat atsakomybę už pažeidimus reglamentuoja Valstybinės lietuvių kalbos įstatymas (1995 m.). Už šio įstatymo nuostatų vykdymą yra atsakinga Valstybinė lietuvių kalbos komisija, kuri teikia pasiūlymus dėl juridinio reguliavimo ir svarsto kalbos norminimo ir vartojimo klausimus.

Lietuvių kalba yra valstybinė – toks jos statusas įtvirtintas Lietuvos Respublikos Konstitucijoje.

Valstybės ir savivaldos institucijos, įmonės ir organizacijos privalo tarpusavyje bendrauti valstybine kalba. Komunikacijos, transporto, sveikatos priežiūros ir socialinės apsaugos, policijos ir teisėtvarkos tarnybų bei kitokių įstaigų, teikiančių paslaugas gyventojams, vadovai turi užtikrinti, kad atitinkamos paslaugos gyventojams būtų teikiamos valstybine kalba.

Lietuviškai transliuojamos 34 nacionalinės ir vietinės TV programos, 52 radijo stotys [7]. Vaizdo ir garso programos ir kino filmai, viešai demonstruojami Lietuvoje, turi būti išversti į valstybinę kalbą arba rodomi su lietuviškais subtitrais. Taigi vertimai yra labai aktuali ir svarbi sritis, atsižvelgiant ir į tai, kad verstinės knygos sudaro apie trečdalį visų lietuviškai išleistų knygų (2010 m. duomenimis, iš 2962 lietuvių kalba išleistų knygų 982 buvo vertimai [8]). Beje, Lietuvos žiniasklaida (spauda, televizija, radijas ir pan.), visi knygų ir kitokie leidėjai privalo laikytis taisyklingos lietuvių kalbos normų. Kaip laikomasi valstybinės kalbos vartojimo ir taisyklingumo reikalavimų, kontroliuoja Valstybinė lietuvių kalbos inspekcija.

Svarbiausios lietuvių kalbos politikos nuostatos yra šios:

- Lietuvių kalba yra valstybės ir jos gyventojų bendravimo priemonė visose viešojo gyvenimo srityse, vienas svarbiausių valstybės suverenumo ir vientisumo požymių.
- Lietuvių kalbos politika turi tenkinti visuomenės, įskaitant ir užsienyje gyvenančius tautiečius, socialinės, nacionalinės ir kultūrinės vienybės poreikį.

- Lietuvių kalbos politika turi derėti su Europos Sąjungos kalbų politika, skatinančia išlaikyti daugiakultūrės Europos kalbų įvairovę, laikomą viena didžiausių Europos vertybių.
- Lietuvių kalbos politika turi ugdyti sąmoningą ir kūrybišką visuomenės požiūrį į lietuvių kalbos vartojimą, lietuvių kalbos vertės ir savitumo suvokimą.
- Lietuvių kalbos tarmės yra lietuvių kalbos ir kultūros turtas, todėl yra saugotinos bei palaikytinos.
- Valstybinės kalbos politika turi apimti komunikacijos sistemas, skirtas specialiųjų poreikių turintiems žmonėms.
- Lietuvių kalbos plėtros prioritetai – skaitmeninės kalbos sistemos ir ištekliai internete. Lietuvių kalba turi plėtotis kaip Europos Sąjungos daugiakalbių terminologinių bei vartojimo išteklių sudedamoji dalis. Automatinis vertimas iš / į lietuvių kalbą – svarbi kalbos vartojimo Europos Sąjungos erdvėje sudedamoji dalis.

Politinių pastangų apsaugoti ir remti lietuvių kalbą taip pat pakanka: valstybinės kalbos vartojimą ir statuso apsaugą reglamentuoja Valstybinės lietuvių kalbos įstatymas (1995 m.), vartojimo ir taisyklingumo kontrolę – Valstybinės lietuvių kalbos inspekcijos įstatymas (2001 m.), terminologijos išteklių plėtrą – Terminų banko įstatymas (2003 m.). Be to, kalba, kaip kultūrinės tapatybės dalis, įtraukta į kultūrinio ir etninio paveldo apsaugos teisės aktus.

Lietuvių kalbos politika turi derėti su Europos Sąjungos kalbų politika, skatinančia išlaikyti daugiakultūrės Europos kalbų įvairovę, laikomą viena didžiausių Europos vertybių.

Lietuvai tapus Europos Sąjungos nare, prasidėjo naujas lietuvių kalbos raidos etapas – įgytas oficialios Europos Sąjungos kalbos statusas užtikrino lietuvių kalbos vartojimą ir sklaidą daugiakalbėje Europos Sąjungos erdvėje, paspartėjo kalbos išteklių, reikalingų visaverčiam kalbos funkcionavimui daugiakalbėje aplinkoje, kaupimas ir pan.

Lietuvių kalbos vartojimą bei plėtrą remia įvairios valstybinės institucijos bei visuomeninės organizacijos: Valstybinė lietuvių kalbos komisija, Valstybinė lietuvių kalbos inspekcija, Lietuvių kalbos draugija ir kt. Įgyvendinama nemažai valstybės remiamų programų, skirtų kalbos tyrimams ir sklaidai skatinti. Vienas iš svarbių lietuvių kalbos tyrimų ir sklaidos centrų yra Lietuvių kalbos institutas, kuriame veikia visuomenei atviras Kalbos muziejus. Įvairiais aspektais lietuvių kalba tiriama Lietuvos universitetuose – nuo tradicinių empirinių tyrimų iki KT taikymo galimybių.

Be to, visuomenė, ypač moksleiviai ir jaunimas, taip pat įtraukiama į lietuvių kalbos vartojimo ir sklaidos iniciatyvas, kurias organizuoja valstybės institucijos, mokslo įstaigos ir verslo įmonės, pavyzdžiui, rašomas Nacionalinis diktantas, rengiami įvairūs konkursai: dailyraščio, svetimžodžių keitimo lietuviškais žodžiais konkursas „Kalbą kuriu AŠ“, taisyklingos lietuvių kalbos vartojimo informacinėse technologijose konkursas „Švari kalba – švari galva“, kuriuo siekiama skatinti moksleivius elektroninėje terpėje vartoti lietuviškus rašmenis bei taisyklingą lietuvių kalbą. Kasmet renkamos taisyklingiausios metų knygos, gražiausias lietuviškas įmonės pavadinimas, gražiausias lietuviškas žodis ir pan.

3.5 KALBA ŠVIETIMO SRITYJE

Pagal Valstybinės lietuvių kalbos įstatymą valstybė garantuoja visų pakopų išsilavinimą gimtąja lietuvių kalba. Valstybinis baigiamasis lietuvių kalbos egzaminas yra vienintelis privalomas visoms vidurinio lavinimo mokykloms, kurių dėstomoji kalba yra lietuvių. Tokį pat egzaminą laikys ir tautinių mažumų mokyklų moksleiviai. Šalyje veikia mokyklos tautinių mažumų mokiniams,

kuriose mokoma nevalstybine – rusų, lenkų, baltarusių – kalba bei mišrios mokyklos. Yra mokyklų anglų, vokiečių, prancūzų, hebrajų kalbomis [9].

Nepaisant, regis, nemažo dėmesio lietuvių kalbos mokymui, kiekvienais metais prastėja lietuvių kalbos mokėjimo rezultatai visuose mokymosi koncentruose. 2007 m. nacionalinio tyrimo duomenimis, tik 39 proc. aštuntokų tepasiekė pagrindinį lietuvių kalbos mokėjimo lygmenį [9].

Lietuvių kalbos pasiekimai yra labai išsibarstę, o lyginant su 2006 m. duomenimis, gerokai sumažėję [10]. Pagal OECD PISA tyrimą, penkiolikmečių Lietuvos mokinių skaitymo gebėjimų vidutiniai rezultatai 2009 m. taip pat buvo gerokai žemesni nei tarptautinis vidurkis [11]. Beje, ypač suprastėjo berniukų skaitymo gebėjimai. Kiek geresni pradinio mokymo rezultatai: net 99 proc. IV klasės Lietuvos mokinių pasiekė žemiausią 2006 m. tarptautinio skaitymo gebėjimų tyrimo PIRLS (angl. *Progress in International Reading Literacy Study*) lygmenį, tačiau aukščiausią lygį pasiekia tik 5 proc. (tarptautinis vidurkis – 9 proc.) [12].

Nemaža dalis mokinių mano, kad nėra gabūs lietuvių kalbai: 2008 m. nacionalinio mokinių lietuvių kalbos mokėjimo tyrimo duomenimis, 52 proc. dešimtokų teigė manantys, kad nėra gabūs lietuvių kalbai, tik pusė mokinių teigė, kad jiems patinka lietuvių kalbos pamokos [11].

> Lietuvių kalba strategijoje suvokiama kaip moderni visuomenės bendravimo priemonė, vartojama ir vartotina visose gyvenimo srityse ir terpėse.

Proveržio tikimasi iš besikeičiančio požiūrio į lietuvių kalbos ir literatūros mokymą. 2010 m. patvirtinta „Lietuvių kalbos ugdymo bendrojo lavinimo programas vykdančiose mokyklose 2010–2014 metų strategija“ [13], kurioje numatyta, kad mokykla turi plėtoti tokią lietuvybę, kuri suteikia viską, ko XXI amžiaus žmogui reikia, kad jis augtų laisvas, pasitikintis savimi, kritiškai mąstantis, kūrybingas, atsakingas. Lietuvių kalba strategijoje suvokiama kaip moderni visuomenės bendravimo priemonė, vartojama ir vartotina visose gyvenimo srityse ir terpėse. Kalbos kitimas, pritaikymas šiuolaikinės informacinės visuomenės reikmėms, atvirumas ir gebėjimas atsinaujinti yra jos išlikimo sąlyga.

Daug dėmesio skiriama lietuvių kalbos mokymui ir sklaidai Lietuvoje bei pasaulyje. Intensyvėjant migracijai, stengiamasi sudaryti sąlygas iš užsienio grįžtantiems vaikams mokytis lietuvių kalbos ir toliau tęsti mokslą Lietuvoje, o ketinantieji išvykti gali pasirinkti savarankiško ar nuotolinio mokymosi būdą [13]. Tačiau pripažįstama, kad nesukūrus šiuolaikinių nuotolinio mokymo sistemų nemaža dalis išvykusių vaikų praras ryšį su lietuvių kalba ir kultūra.

Plečiamas lietuvių kalbos mokymas tautinių mažumų mokyklose, taip suteikiant daugiau galimybių kitakalbiams integruotis į Lietuvos darbo rinką, gauti informaciją ir bendrauti viešojoje erdvėje.

3.6 TARPTAUTINIAI ASPEKTAI

Nors nemaža dalis lietuvių, paklausti, kuo garsi Lietuva pasaulyje, ilgai nemąstydami atsakys – krepšiniu (iš tikrųjų garsių krepšininkų ar trenerių lietuviški vardai ar pavardės yra nebloga lietuviškos tarties treniruotė ir pirmoji pažintis su lietuvių kalba krepšinio mėgėjams visame pasaulyje), vis dėlto tradiciškai viena iš pagrindinių nacionalinių vertybių ir Lietuvos prisistatymo pasauliui objektų laikoma pati lietuvių kalba. Kai kuriems Lietuvos gyventojams jų tautiškumo supratimas ligi šiol asocijuojasi su kalbine tapatybe.

> Lietuvių kalbą, jos tarmes, tautosaką yra tyrę ir aprašę įvairių tautybių mokslininkai – šių tyrimų geografija apima nuo kaimyninių valstybių iki Japonijos, Australijos, JAV.

Ne tik kalbos archajiškumu, bet ir pagoniško tikėjimo tradicijomis, originalia tautosaka Lietuva traukė garsius menininkus Adomą Mickevičių, Prosperą Mérimée, Johanną Wolfgangą von Goethe ir kt. Bene ilgiausiai Europoje išlaikytas pagoniškas tikėjimas ir papročiai itin domino mitologus, ne veltui iš Lietuvos kilusi viena iš žymiausių Europos proistorės tyrėjų Marija Gimbutienė, mitologija domėjosi ir Paryžiaus semiotikos mokyklos kūrėjas Algirdas Julius Greimas.

Lietuvių kalbą, jos tarmes, tautosaką yra tyrę ir aprašę įvairių tautybių mokslininkai – šių tyrimų geografija apima nuo kaimyninių valstybių iki Japonijos, Australijos, JAV. Tradicija svarbiausiuose indoeuropeistikos centruose studijuoti lietuvių kalbą išlaikyta ir šiandien. Užsienio universitetuose veikia apie 10 lituanistikos centrų, šiuo metu vykdančių savarankiškas lituanistikos ar baltistikos studijų programas. Pasaulyje, daugiausia Europoje, iš viso veikia per 30 įvairaus dydžio centrų, kuriuose tiriama ar dėstoma lietuvių ar baltų kalbos ir kultūra.

Tradicija svarbiausiuose indoeuropeistikos centruose studijuoti lietuvių kalbą išlaikyta ir šiandien.

Lietuvos Respublikos švietimo ir mokslo ministerija remia ir skatina lituanistikos centrus. Kasmet teikiamos Kazimiero Būgos vardo stipendijos užsieniečiams, užsienio šalių aukštosiose mokyklose studijuojantiems lietuvių kalbą.

3.7 LIETUVIŲ KALBA INTERNETE

2011 m. duomenimis, beveik 64,1 proc. Lietuvos gyventojų naudojasi internetu namie ar darbe, o 16–24 metų amžiaus grupėje naudotojų skaičius yra dar didesnis ir siekia 96,6 proc. [14]. Tačiau tikslių duomenų apie tai, kokia kalba jie naudojasi internetu, nėra. Itin gerai išvystyta IRT infrastruktūra – Lietuva pirmauja Europoje pagal šviesolaidinio plačiajuosčio tinklo skverbtį (23 proc.), užima pirmąją vietą pasaulyje pagal mobiliojo ryšio abonentų skaičių, tenkantį 100 gyventojų, užima antrąją vietą pasaulyje pagal interneto ryšio greitį ir turi tankiausią Europoje belaidžio interneto prieigos taškų tinklą Europoje (875) [15].

Lietuva pirmauja Europoje pagal šviesolaidinio plačiajuosčio tinklo skverbtį, užima pirmąją vietą pasaulyje pagal mobiliojo ryšio abonentų skaičių, tenkantį 100 gyventojų, užima antrąją vietą pasaulyje pagal interneto ryšio greitį ir turi tankiausią Europoje belaidžio interneto prieigos taškų tinklą Europoje.

2010 metais registruota 126 tūkstančių *.lt* sričių vardų, iš jų apie 1 400 – su specifiniais lietuviškais rašmenimis (*ė*, *š* ir pan.). Daugumos jų turinys yra lietuviškas. Lietuviško turinio puslapių yra ir *.eu*, *.org*, *.com* sričių varduose.

Internete vis daugiau atsiranda viešųjų paslaugų, prieinamų lietuvių kalba. Pagrindinių viešųjų paslaugų perkėlimo į elektroninę terpę lygis 2005 m. Lietuvoje siekė 64 proc. Dar sparčiau į internetą perkeliamos verslui skirtos paslaugos – jų lygis 2005 m. siekė 76 proc., o gyventojams – tik 56 proc. [16]. Įgyvendinant nacionalines programas siekiama daugiau viešųjų paslaugų perkelti į interneto erdvę, didinti lietuviško turinio apimtį internete skaitmeninant ir skleidžiant Lietuvos kultūros paveldą, sudarant sąlygas Lietuvos gyventojams naudotis IRT, turinčiomis lietuviškas sąsajas. Taip mažinama skaitmeninė atskirtis, užtikrinama, kad technologijomis būtų lengva naudotis, jos būtų pritaikytos žmonėms su negalia.

2010 metais registruota 126 tūkstančių *.lt* sričių vardų, iš jų apie 1 400 – su specifiniais lietuviškais rašmenimis.

Vis populiaresni tampa žinių portalai, internetu pasiekiami pagrindiniai lietuviški spaudos leidiniai, kai kurie mokslo žurnalai ir pan. Iš svarbiausių lietuviško turinio sklaidos projektų minėtinas kultūros paveldo portalas www.epaveldas.lt, virtualioje erdvėje jungiantis bibliotekų, muziejų, kitų paveldo institucijų išteklius, mokomųjų išteklių portalai www.emokykla.lt, www.emokymas.lt ir pan. Planuojama sukurti portalą, kuriame būtų dedami atvirai prieinami kalbos ištekliai ir technologijos, sukurti pagal prasidedančią programą „Lietuvių kalba informacinėje visuomenėje".

Kitame skyriuje pristatomos KT ir pagrindinės jų taikymo sritys bei lietuvių kalbai pritaikytų KT įvertinimas.

4

LIETUVIŲ KALBAI PRITAIKYTOS KALBOS TECHNOLOGIJOS

KT – tai programinės įrangos sistemos, skirtos apdoroti žmonių kalbą, kuri gali būti šnekamoji ir rašomoji. Nors žmogaus evoliucijos požiūriu šnekamoji kalba yra seniausia kalbinio bendravimo forma, sudėtinga informacija ir didžioji dalis žmonėms žinomų faktų yra saugoma bei perduodama rašytiniais tekstais. Šnekamajai kalbai ir tekstams skirtos technologijos apdoroja arba atkuria kalbos formas skirtingai, nors visos jos yra pagrįstos žodynais, gramatikos ir semantikos taisyklėmis. Taigi KT sujungia kalbą su įvairių formų žiniomis, neatsižvelgdamos į raiškos priemones. Dešinėje pusėje pateiktame paveiksle pavaizduota KT aplinka. Bendraudami sujungiame kalbą su kitomis bendravimo terpėmis ir informacijos perdavimo priemonėmis, pavyzdžiui, kalbėdami galime gestikuliuoti, keisti veido išraišką. Skaitmeniniai tekstai susiję su vaizdais ir garsais. Filmuose gali būti naudojama šnekamoji ir rašomoji kalba. Kitaip tariant, šnekos ir teksto technologijos jungiasi ir sąveikauja su kitomis technologijomis, lengvinančiomis įvairiarūšio bendravimo ir daugialypės terpės dokumentų apdorojimą (žr. 1 paveikslą).

Toliau aptarsime pagrindines KT taikymo sritis, pavyzdžiui, kalbos taisyklingumo tikrinimą, paiešką internete, šnekamosios KT ir automatinį vertimą, kurios apima tokius technologijų pritaikymo būdus:

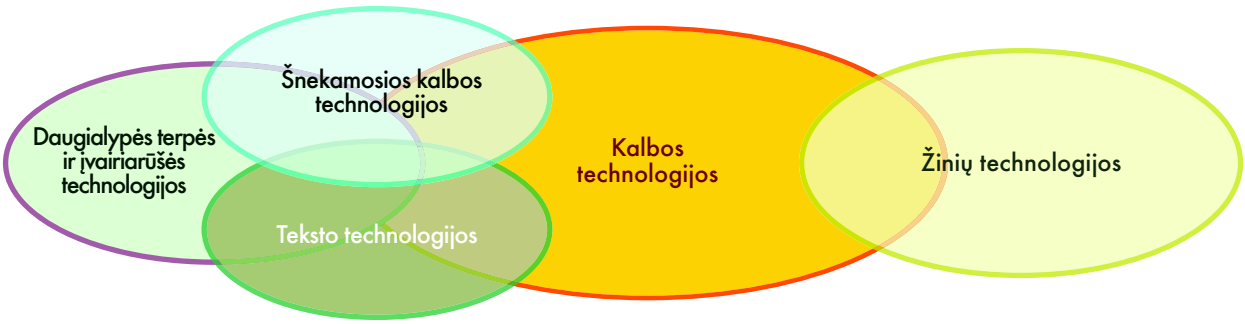
- rašybos ir gramatikos tikrinimą;
- pagalbą kuriant dokumentus;
- kalbų mokymąsi;
- informacijos paiešką;
- informacijos išgavimą;
- tekstų santraukų kūrimą;
- atsakymus į klausimus;
- šnekos atpažinimą;
- šnekos sintezę.

KT tyrimo sritis yra gerai įsitvirtinusi ir įvadinės literatūros galima rasti vis daugiau. Susidomėjusiems skaitytojams siūlomos šios publikacijos: [17, 18, 19, 20, 21]. Prieš aptardami išvardytas pritaikymo sritis, trumpai pakalbėkime apie tipinės KT sistemos architektūrą.

4.1 KALBOS TECHNOLOGIJŲ TAIKYMO ARCHITEKTŪRA

Kalbos apdorojimo programinę įrangą paprastai sudaro keletas komponentų, atspindinčių skirtingus kalbos aspektus. 2 paveiksle parodyta itin supaprastinta tipinės teksto apdorojimo sistemos struktūra. Pirmieji trys moduliai apdoroja įvedamo teksto struktūrą ir nustato pradinius semantikos duomenis:

1. Pirminis duomenų apdorojimas (angl. *preprocessing*): duomenys „išvalomi“, išanalizuojamas arba pašalinamas formatavimas, nustatoma įvesties kalba ir pan.
2. Gramatinė analizė: atliekama žodžių morfologinė analizė, nustatomos kalbos dalys ir pagrindinės žodžių formos, surandamas veiksmažodis, jo papildi-



1: Kalbos technologijos

niai, aplinkybės ir kitos kalbos dalys, nustatoma sakinio struktūra.

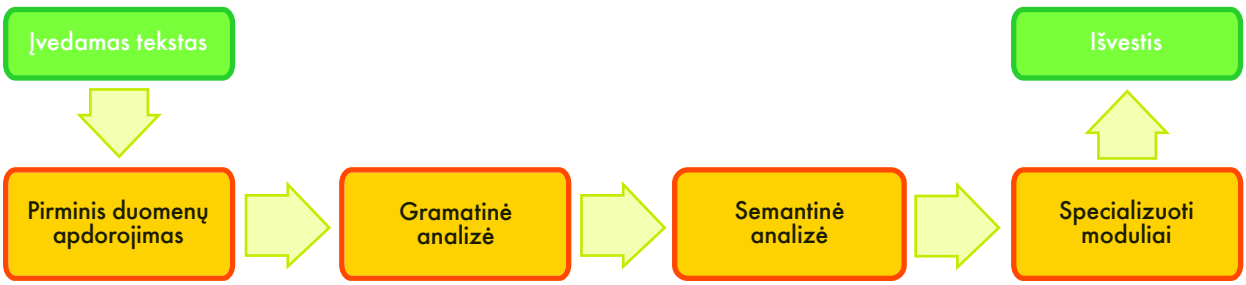
3. Semantinė analizė: nustatoma žodžio reikšmė (t. y. išanalizuojama apytikslė žodžių reikšmė tiriamame kontekste); išsprendžiamos anaforos (t. y. nustatoma, kurie įvardžiai sakinyje atitinka kuriuos daiktavardžius) ir posakių pakeitimo problemos; sakinio prasmė pateikiama kompiuteriui suvokiamu būdu.

Išanalizavus tekstą, specialioms užduotims sukurti moduliai gali atlikti ir kitokius veiksmus, pavyzdžiui, sukurti teksto santrauką ar ieškoti informacijos duomenų bazėse.

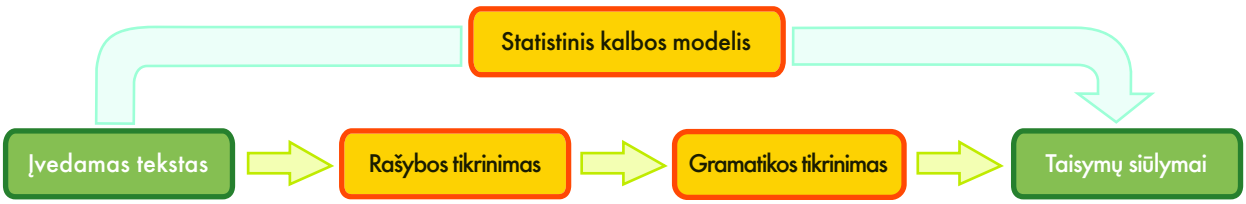
Pristatę pagrindines sritis, kuriose taikomos KT, trumpai apžvelgsime dabartinę KT tyrimų ir švietimo būklę, jau įgyvendintas ir šiuo metu vykdomas mokslinių tyrimų programas. Po to pateiksime profesionalų atliktą pagrindinių KT įrankių ir išteklių įvertinimą įvairiais aspektais – prieinamumo, išbaigtumo, kokybės ir pan. Bendroji lietuvių kalbos KT būklė pateikiama 8 lentelėje.

4.2 PAGRINDINĖS TAIKYMO SRITYS

Šiame skyriuje daugiausia dėmesio skirsime svarbiausiems KT įrankiams ir ištekliams, apžvelgsime KT padėtį Lietuvoje. Įrankius ir išteklius, kurių pavadinimai tekste paryškinti, galima rasti ir šio skyriaus pabaigoje pateiktoje 8 lentelėje.



2: Tipinė teksto apdorojimo programos architektūra



3: Kalbos taisyklingumo tikrinimas (viršuje – pagrįstas statistiniais metodais, apačioje – taisyklėmis)

4.2.1 Rašybos ir gramatikos tikrinimas

Kam yra tekę naudotis teksto rašymo programa, pavyzdžiui, *Microsoft Word*, žino, kad joje yra funkcija tikrinti, ar nėra rašybos klaidų. Ji klaidas randa, parodo ir siūlo taisymo variantus. Pirmosios rašybos klaidų taisymo programos palygindavo atrinktus žodžius su taisyklingos rašybos žodynu. Šiandien tokios programos yra kur kas sudėtingesnės. Naudodamos su konkrečia kalba susietus teksto analizės algoritmus, jos aptinka morfologines (pvz., daugiskaitos darybos klaidas) ir sintaksines (pvz., praleistus veiksmažodžius arba nesuderintus veiksnius ir tarinius, pvz., *ji *rašyti laišką*) klaidas. Tačiau dauguma rašybos klaidų taisymo programų neras jokių klaidų tokiame tekste anglų kalba [22]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Tokiai analizei atlikti reikalinga arba ekspertų į programinę įrangą kruopščiai perkelta konkrečios kalbos **gramatika**, arba statistinis kalbos modelis. Šiuo atveju modelis apskaičiuoja tam tikro žodžio vartojimo kitų žodžių apsuptyje tikimybę. Statistinį kalbos modelį galima sukurti automatiškai, panaudojant didelį skaičių taisyklingų kalbos duomenų (**tekstyną**). Šie du metodai daugiausia buvo išplėtoti naudojant anglų kalbos duomenis. Taigi nė vieno metodo negalima lengvai pritaikyti lietuvių kalbai, kadangi lietuvių kalbos žodžių tvarka nėra fiksuota, o kaitybos sistema yra kur kas sudėtingesnė (nuo 2002 m. VDU mokslininkai nemažai dirbo su statistiniu lietuvių kalbos modeliu, įrankius galima nemokamai atsisiųsti [23]).

Kalbos taisyklingumo tikrinimo programos naudojamos ne tik rašant tekstus, jas galima pritaikyti ir dokumentų kūrimo pagalbinėse sistemose.

Kalbos taisyklingumo tikrinimo programos naudojamos ne tik rašant tekstus, jas galima pritaikyti ir dokumentų kūrimo pagalbinėse sistemose (angl. *authoring support systems*), t. y. programinėje terpėje, kurioje laikantis ypatingų standartų kuriami sudėtingų informacinių technologijų sistemų, sveikatos priežiūros, inžinerijos ir kitokių sričių vadovai bei kitokie dokumentai. Nuogąstaudamos, kad klientai pradės skųstis dėl to, jog produktas bus naudojamas ne taip, kaip reikia, o dėl nesuprantamų instrukcijų bus pateikta reikalavimų kompensuoti žalą, bendrovės vis daugiau dėmesio skiria techninės dokumentacijos kokybės gerinimui, tuo pat metu orientuodamosios į tarptautines rinkas (versdamos tekstus arba juos pritaikydamos konkrečiai kalbai). Pažanga, pasiekta apdorojant natūralią kalbą, paskatino plėtoti dokumentų kūrimo pagalbines sistemas. Šios sistemos padeda techninių dokumentų autoriui naudoti tos srities žodyną ir sakinių struktūras, atitinkančias srities taisykles, verslo terminiją.

Šioje srityje ką pasiūlyti turi vos keletas Lietuvos bendrovių. 1992–1994 m. UAB „Fotonija“ sukūrė rašybos tikrinimo programą *Juodos Avys*, kuri buvo nuolat tobulinama. Automatinio rašybos tikrinimo metu veikia algoritmas, padedantis išvengti savaiminio pavardės ar pavadinimo pakeitimo į kitą. Atpažįstami neteiktini žodžiai ir siūloma juos keisti į tinkamus. Teikiami pasiūlymai dėl trūkstamų specifinių lietuviškų rašmenų, tokių kaip *š*, *ž*, *ū*, *ė* taisymo. Į tikrintuvę integruotas lietuviškų skiemenų rašymo su brūkšneliu įrankis.

UAB „Tilde IT“ lietuvių kalbos rašybos tikrinimo programą sukūrė 2001 m. Ši bendrovė tobulina rašybos tikrinimo programą ir kuria naują gramatikos tikrinimo programą, kuri analizuos sakinio struktūrą, atpažins praleistus arba nereikalingus kablelius ir kitokius skyrybos ženklus, tikrins, ar nėra sintaksės ir leksikos klaidų. Gramatikos tikrintuvė veiks ne tik su platforma *Microsoft Office*, bet ir su platforma *Open Office* ar internetinėmis programomis. Ji bus nesunkiai suderinama su kitomis programomis, kuriose naudojamos su kalba susijusios funkcijos (pvz., įmonės išteklių tvarkymo programomis, verslo sprendimais ir pan.). Naudotojai turės galimybę išbandyti naująją gramatikos klaidų taisymo programą 2012 metais.

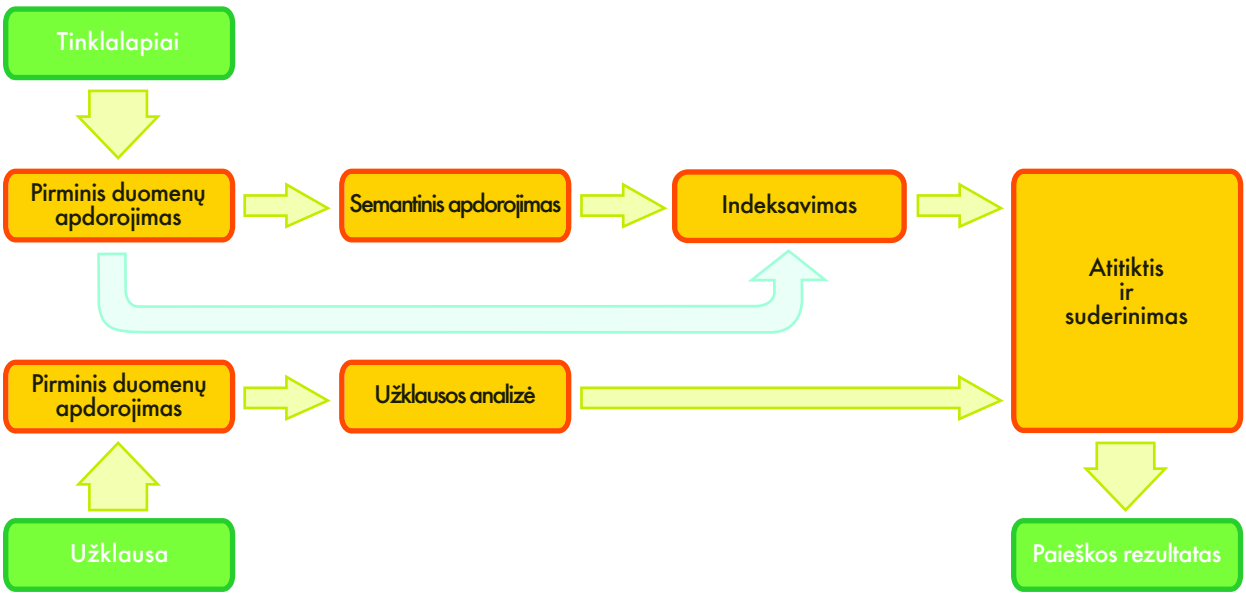
Kalbos tikrinimo funkcija svarbi ne tik rašybos tikrinimo ir dokumentų kūrimo programose – ji reikalinga ir skaitmeninėje terpėje mokantis kalbų. Be to, kalbos tikrinimo įrankiai automatiškai taiso į paieškos sistemas įvedamas užklausas, teikia teisingų užklausų pasiūlymus – čia kaip pavyzdį galime paminėti sistemos *Google* įrankį *Galbūt jūs norėjote ieškoti...* (angl. *Did you mean...*).

4.2.2 Paieška internete

Paieška internete, vidaus tinkluose ar skaitmeninėse bibliotekose šiandien turbūt yra plačiausia, tačiau mažiausiai išplėtota KT pritaikymo sritis. Paieškos sistema *Google*, kuri buvo pristatyta 1998 metais, šiuo metu apdoroja apie 80 proc. visų užklausų [24]. Sistemos *Google* paieškos sąsaja ir langas, kuriame pateikiami rezultatai, ne itin pasikeitė palyginti su pirmąja sistemos versija. Tačiau dabartinė *Google* versija turi rašybos klaidų taisymo ir esminių semantinės paieškos funkcijų, analizuojančių užklausos terminų kontekstinę reikšmę ir galinčių padidinti paieškos tikslumą [25]. Sistemos *Google* sėkmė rodo, kad gausybė sukauptų duomenų ir veiksmingi rodyklių sudarymo įrankiai gali padėti pasiekti neblogų rezultatų taikant statistinį metodą.

Norint apdoroti įmantresnes informacijos užklausas, būtinos nuodugnesnės lingvistinės tekstų interpretavimo žinios. Eksperimentai su **leksikos ištekliais**, pavyzdžiui, kompiuteriui suprantamais tezaurais ar ontologiniais kalbų ištekliais (pvz., *WordNet* anglų kalba ar *GermaNet* vokiečių kalba), parodė, kad vartojant pirminių terminų sinonimus, tokius kaip *Atomkraft* [atominė energija], *Kernenergie* [atominė galia] ir *Nuklearenergie* [branduolinė energija], ar net ne taip glaudžiai tarpusavyje susijusius terminus, randama vis daugiau internetinių puslapių.

Naujosios kartos paieškos sistemos turės būti grindžiamos kur kas pažangesnėmis KT, ypač turint galvoje užklausas, kurias sudaro klausimas ar kitoks sakinys, o ne keli esminiai žodžiai. Apdorodama užklausą „Pateikite man sąrašą bendrovių, kurias per pastaruosius penkerius metus įsigijo kitos bendrovės“, KT sistema turi išanalizuoti sakinį sintaksės ir semantikos požiūriu bei pateikti rodyklę, kuri leistų operatyviai rasti reikiamus dokumentus. Norint pateikti patenkinamą atsakymą, prireiks išnagrinėti sakinio sintaksę nustatant, kad naudotojas prašo pateikti sąrašą bendrovių, kurios buvo įsigytos, o ne kurios įsigijo kitų bendrovių. Apdorodama posakį *pastaruosius penkerius metus*, sistema turi nustatyti reikiamus metus. Be to, užklausą reikia palyginti su gausybe nesusistemintų duomenų, ieškant naudotojo pageidaujamos informacijos. Šis procesas vadinamas *informacijos paieška*, jį sudaro tam tikrų dokumentų paieška



4: Paieškos internete architektūra

ir vertinimas. Rengdama bendrovių sąrašą sistema dar turi atpažinti ir konkrečią žodžių seką dokumente, pavyzdžiui, bendrovės pavadinimą – toks procesas vadinamas *įvardytų subjektų atpažinimu*.

Naujosios kartos paieškos sistemos turės būti grindžiamos kur kas pažangesnėmis kalbos technologijomis.

Dar sunkesnis darbas – derinti užklausą, pateiktą viena kalba, ir dokumentus, skelbiamus kita kalba. Informacijos skirtingomis kalbomis paieška reiškia, kad reikės automatiškai išversti užklausą į visas įmanomas informacijos šaltinių kalbas, o po to rezultatus vėl išversti į užklausos kalbą.

Dabar duomenys vis dažniau pateikiami ne tekstiniais formatais, tad didėja daugialypėje terpėje – paveikslėliuose, garso ir vaizdo bylose – esančios informacijos paieškos poreikis. Garso ir vaizdo bylų informaciją kalbos atpažinimo modulis turi paversti tekstu (arba fonetine transkripcija), kuris galės būti palygintas su naudotojo užklausa.

Lietuvių kalbos pritaikymui šios technologijos tik pradedamos intensyviau plėtoti. Su šia sritimi susijusius tyrimus ir projektus 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. Semantinių tinklų, ontologijų kūrimo, žinių ir dokumentų tvarkymo srityje pradeda dirbti kai kurios KT srityje dirbančios verslo kompanijos, pavyzdžiui, UAB „Sintagma“, sukūrusi dokumentų tvarkymo sistemą *Avilys*. UAB „Tilde IT“ nuo 2008 metų plėtoja semantinių sistemų srities projektus ir šiuo metu įgyvendina lingvistinio semantinio tinklo kūrimo projektą *SemTi*, taip pat dalyvauja tarptautiniame SOLIM projekte (angl. *Spatial Ontology Language for Multimedia Information Modeling*). Atskirų sričių, pavyzdžiui, bibliotekų, mokslo, ontologijos pradedamos intensyviau diegti tik pastaruoju metu.

Kol kas šioje srityje pastangos fragmentiškos ir didesnio proveržio tikimasi iš Lietuvos Respublikos Vyriausybės inicijuotos „Lietuvių kalbos informacinėje visuomenėje“ 2009–2013 metų programos, pagal kurią numatoma sukurti priemones, pritaikytas teikti sintaksinės-semantinės analizės paslaugą, analizuoti lietuviškų interneto svetainių turinį, atlikti pagal jį paiešką ir pan.

4.2.3 Šnekamosios kalbos technologijos

Šnekamosios KT taikomos norint sukurti sąsajas, leisiančias naudotojams bendrauti su kompiuteriu balsu, o ne naudojantis grafiniu ekranu, klaviatūra ir pele. Šiandien naudotojo balso sąsajos (angl. *voice user interfaces*, VUI) paprastai naudojamos teikiant iš dalies arba visiškai automatizuotas telefono paslaugas klientams, darbuotojams ar partneriams. Naudotojo balso sąsajos yra itin aktualios bankininkystėje, tiekimo sistemose, viešajame transporte ir telekomunikacijose. Kitos sritys, kuriose taikomos šnekamosios KT, yra automobilių navigacijos sistemų sąsajos ir šnekamosios kalbos sąsaja vietoje išmaniųjų telefonų grafinių arba liečiamųjų ekranų sąsajų.

Šnekamosios kalbos technologijos taikomos norint sukurti sąsajas, leisiančias naudotojams bendrauti su kompiuteriu balsu, o ne naudojantis ekranu, klaviatūra ir pele.

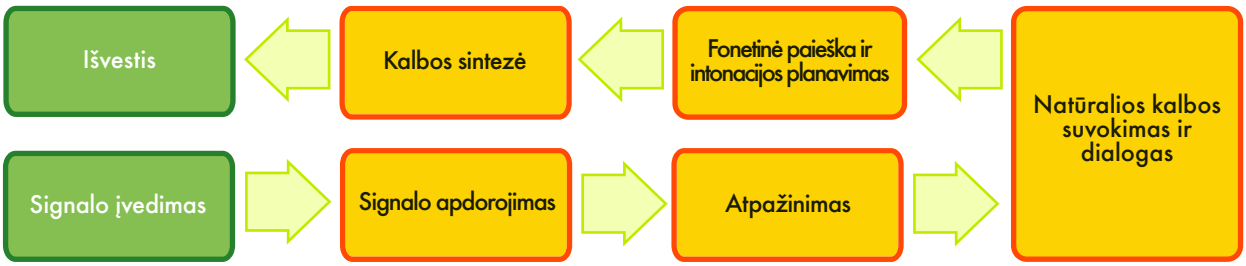
Šnekamajai kalbai skirtos keturios technologijos:

1. Automatinis **kalbos atpažinimas** (angl. *automatic speech recognition*, ASR) nustato, kokius žodžius sudaro naudotojo ištarta garsų seka.
2. Natūralios kalbos suvokimo technologija analizuoja naudotojo ištartos frazės morfologinę ir sintaksinę struktūrą ir ją interpretuoja pagal atitinkamos sistemos taisykles.
3. Dialogo valdymo technologija nustato, kokių veiksmų imtis atsižvelgiant į naudotojo įvestus duomenis ir sistemos funkcijas.
4. **Kalbos sintezė** (rašytinio teksto pavertimas šnekamąja kalba, angl. *text-to-speech*, TTS) sistemos atsakymą naudotojui transformuoja į garsą.

Vienas iš didžiausių keblumų, su kuriais susiduria automatinio kalbos atpažinimo sistemos, yra tikslus naudotojo ištartų žodžių atpažinimas. Reikia apriboti galimų naudotojo pasakymų skaičių iki tam tikro esminių žodžių sąrašo arba rankiniu būdu sukurti kalbos modelius, apimančius daugybę natūralios kalbos pasakymų. „Mokant“ kompiuterines programas, kalbos modelius galima kurti ir automatiškai, pasitelkiant šnekamosios kalbos garsynus – gausias kalbos garsinių bylų ir tekstų transkripcijų rinkinius. Apribojus ištariamų pasakymų skaičių, žmonės verčiami naudotis balso sąsajomis itin tiksliai ir dėl tos priežasties tokios sistemos gali tapti nepatogios naudotojams. Kita vertus, sukurti, priderinti ir prižiūrėti turiningus bei išsamius kalbų modelius galima tik labai padidinus išlaidas. Kalbų modeliais pagrįstos ir naudotojui galimybę lanksčiau išreikšti pageidavimus suteikiančios (vartojant posakį *Kuo galiu jums padėti?*) naudotojo balso sąsajos dažniausiai būna automatinės ir gerokai priimtinesnės.

Bendrovės, kurdamos naudotojo balso sąsajų išvestis, linkusios naudoti profesionalių diktorių iš anksto įrašytas frazes. Statinių pasakymų atveju, kai žodžiai nepriklauso nuo konkretaus konteksto ar naudotojo asmeninių duomenų, tokia sistema gali pasirodyti itin patraukli. Tačiau dinamiškesnis turinys gali nukentėti dėl nenatūralios intonacijos, kadangi atskiri garsinių bylų segmentai būna tiesiog susieti vienas su kitu. Dabartinės kalbos sintezės sistemos vis tobulėja (nors jas vis dar reikia gerinti), jos sugeba atkurti natūraliai skambančią dinamišką kalbą.

Per pastarąjį dešimtmetį buvo gerokai standartizuoti rinkoje turimų šnekamosios KT sąsajų įvairūs techno-



5: Nesudėtingo balsinio dialogo architektūra

loginiai komponentai. Be to, kalbos atpažinimo bei kalbos sintezės rinkos ir toliau intensyviai konsolidavosi. Dvidešimties didžiųjų šalių G20 (ekonomiškai stabilių šalių, pasižyminčių dideliu gyventojų skaičiumi) rinkose vyravo vos penki pasauliniai dalyviai, iš kurių Europoje aktyviausios buvo bendrovės „Nuance“ (JAV) ir „Loquendo“ (Italija). 2011 metais bendrovė „Nuance“ paskelbė įsigijusi bendrovę „Loquendo“ – tai yra dar vienas rinkos konsolidacijos žingsnis.

Lietuvoje šnekamosios KT moksliniai tyrimai nuo 1980 metų atliekami Kauno technologijos universitete (KTU). Jau daugelį metų šie darbai dirbami Vilniaus universiteto Matematikos ir informatikos institute, taip pat tyrimai atliekami Vytauto Didžiojo universitete.

Kauno technologijos universiteto Kalbos tyrimų laboratorijoje automatinio kalbos atpažinimo tyrimai tęsiami nuo 1980 m. Laboratorija yra sukūrusi komandų ir skaitmeninių sekų garsyną. Kuriami lietuviški kompiuteriniai dialogai, sukauptas ir tobulinamas lietuvių šnekamosios kalbos garsynas LTDIGITS. Jį sudaro ištisinės skaičių sekos ir lietuviški žodžiai kompiuteriui valdyti. Lietuvių kalbos ženklų tyrimai atliekami ir Vilniaus universiteto Matematikos ir informatikos institute, sukaupsiame Lietuvos radijo žinių garsyną LRNO. Vytauto Didžiojo universitete sukauptas universalus šnekamosios lietuvių kalbos garsynas (čia kaupiama ir mažesnės apimties specialiųjų tekstynų, skirtų kalboms mokytis, pavyzdžiui, jaunuolių sakytinės kalbos tekstynas SACODEYL ir pan.). Vyksta ir šnekamosios lietuvių kalbos automatinio skaidymo tyrimai, kuriama šnekamosios lietuvių kalbos automatinė transkripcija.

Nors tyrimai ir toliau vyksta siekiant gerinti kokybę, automatinio kalbos atpažinimo programinė įranga šiuo metu sėkmingai taikoma teisėsaugoje, telefonijoje, švietime, transporte, internete ir kitur.

Vilniaus universitete atlikti šnekamosios kalbos sintezės ir tokių sistemų pritaikymo akliesiems ir silpnaregiams tyrimai. Lietuviškos balso sintezės programos *Aistis* svarbiausi komponentai yra šie: a) automatinis lietuviškų žodžių skaidymas skiemenimis; b) žodžių lietuvių kalba parašytame tekste automatinis kirčiavimas; c) automatinis lietuviškų tekstų transkribavimas; d) fonetinių vienetų bazė; e) lietuviškų tekstų pavertimo šnekamąja kalba kokybės įvertinimas. Šia programa lengva naudotis, ji skirta specialiųjų poreikių, pavyzdžiui, fizinę negalią turintiems naudotojams ar senyvo amžiaus žmonėms. Balso sintezatorių MBROLA galima lengvai rasti internete, jis yra pagrįstas Vilniaus universitete Alekso Girdenio ir Pijaus Kasparaičio sukurta fonetinių vienetų baze.

Programos *Aistis* komponentai buvo pritaikyti kuriant lietuvių kalbos sintezatorių *WinTalker Voice*, kuriame buvo įdiegti du balsai: *Gintaras* ir *Aistis2*. Šią programą Lietuvos aklųjų ir silpnaregių draugijos užsakymu išleido čekų bendrovė „Rosasoft“. Derėtų paminėti, kad Lietuvoje yra apie 7 tūkst. žmonių, turinčių specialiųjų

poreikių, o 2010 m. kovo 1 d. duomenimis, visoje Lietuvoje kompiuteriu naudojosi 258 aklieji ir silpnaregiai. Dar vieną nemokamą balso sintezatorių sukūrė UAB „Etalinkas“. Šis sintezatorius yra pritaikytas dirbti su operacinėmis sistemomis *Windows* ir *Linux*.

Vytauto Didžiojo universitete šnekos atpažinimui tekstyno pagrindu sukurti statistiniai lietuvių kalbos modeliai, ištisinės šnekos atpažintuvo prototipas, apimantis daugiau nei 1 mln. žodžių formų, automatinio kirčiavimo programa su homografų vienareikšminimu, prieinama internete, bei lietuvių kalbos garsų trukmių modeliai.

Vis labiau populiarėjantys išmanieji telefonai, kaip nauja bendravimo su naudotojais priemonė greta fiksuotųjų telefonų, interneto ir elektroninio pašto, ateityje lems didžiulius pokyčius. Tai turės įtakos ir šnekamosios KT taikymui. Ilgainiui telefoninės naudotojo balso sąsajos sustiprės, o šnekamoji kalba taps naudotojams dar svarbesne ir patogesne priemone duomenims į išmaniuosius telefonus įvesti. Šiuos pokyčius skatins vis tikslesnis nuo kalbėtojo nepriklausantis kalbos atpažinimas, panaudojant šnekos diktavimo paslaugas. Tokios centralizuotos paslaugos jau dabar yra siūlomos išmaniųjų telefonų naudotojams.

4.2.4 Automatinis vertimas

Sumanymas pritaikyti kompiuterius versti iš vienos kalbos į kitą kilo 1946 m., o šeštajame ir vėliau devintajame praėjusio amžiaus dešimtmetyje šiems tyrimams buvo skirta nemažai lėšų. Tačiau ir šiuo metu **automatinis vertimas** vis dar negali tenkinti visuotinio vertimo poreikių.

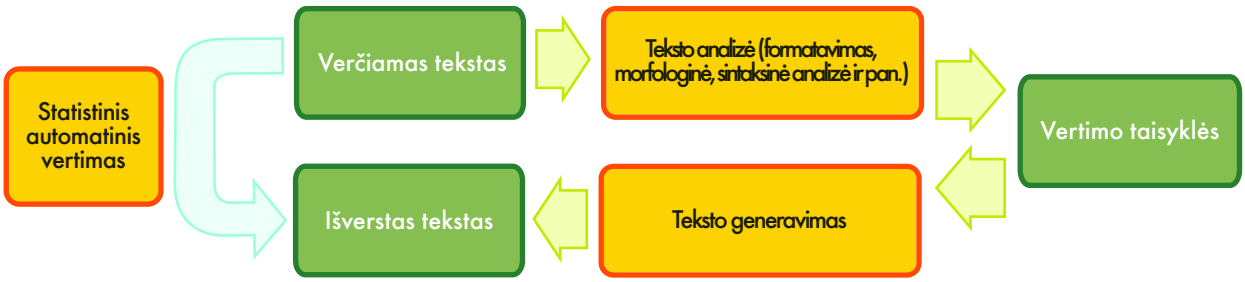
Paprasčiausias automatinio vertimo būdas – automatiškai pakeisti vienos kalbos žodžius kitos kalbos žodžiais. Šis būdas gali būti naudingas tose srityse, kurių kalba yra labai standartizuota ir šabloniška, pavyzdžiui, rengiant orų prognozes. Tačiau norint gauti kokybišką ne tokių standartinių tekstų vertimą, didesni teksto vienetai (frazės, sakiniai ar net visos pastraipos) turi būti sugretinti su jų artimiausiais atitikmenimis ta kalba, į kurią verčiama. Sunkiausia įveikti žmonių kalbos polisemiją, kuris kelia problemų skirtingiems kalbos lygmenims, pavyzdžiui, nustatant daugiareikšmių žodžių leksinę reikšmę (*jaguaras* gali būti ir automobilio markė, ir gyvūnas) arba linksnį ar reikiamą formą sintaksinėse konstrukcijose, pavyzdžiui:

- *I was happy to read a book.*
- *Aš buvau laimingas:*
  1. *perskaitęs knygą.*
  2. *skaitydamas knygą.*
  3. *galėdamas perskaityti knygą.*

Vienas iš būdų sukurti automatinio vertimo sistemą – pasinaudoti kalbos taisyklėmis. Verčiant iš vienos kalbos į jai giminingą kalbą, kaip jau minėta, galima būtų tiesiog pakeisti žodžius, tačiau taisyklėmis pagrįstos (arba į lingvistines žinias orientuotos) sistemos dažniausiai analizuoja įvedamą tekstą ir sukuria jo tarpinį simbolinį pavidalą, pagal kurį gali būti sukurtas išverstasis tekstas. Tokių metodų sėkmė itin priklauso nuo galimybės turėti išsamius žodynus, kuriuose pateikiama morfologinė, sintaksinė ir semantinė informacija, bei gausius gramatikos taisyklių rinkinius, parengtus profesionalių kalbininkų. Tai labai ilgas ir, žinoma, labai daug kainuojantis darbas.

Paprasčiausios automatinio vertimo programos tiesiog pakeičia vienos kalbos žodžius kitos kalbos žodžiais.

Praėjusio amžiaus devintojo dešimtmečio pabaigoje, pradėjus kurti galingesnius kompiuterius ir jiems atpigus, imta labiau domėtis automatinio vertimo statistiniais modeliais. Statistiniai modeliai išplėtoti analizuojant dvikalbius tekstynus, pavyzdžiui, **lygiagretųjį tekstyną** *Europarl*, kuriame pateikiama Europos Parlamento



6: Automatinis vertimas (kairėje – pagrįstas statistiniais metodais, dešinėje – taisyklėmis)

posėdžių medžiaga vienuolika Europos kalbų. Turėdamos pakankamai duomenų, statistinės automatinio vertimo sistemos apdoroja lygiagrečias tekstų skirtingomis kalbomis versijas ir randa galimus žodžių modelius – dėl to jos tinka apytiksliams vertimams iš vienos kalbos į kitą. Tačiau, kitaip nei taisyklėmis pagrįstos sistemos, statistinės (arba duomenimis pagrįstos) automatinio vertimo sistemos dažnai pateikia gramatiškai netaisyklingų tekstų. Duomenimis pagrįstos automatinio vertimo sistemos yra pranašesnės, kadangi reikia mažiau žmogaus indėlio ir pastangų, be to, tokios sistemos gali atsižvelgti į tam tikras kalbos ypatybes (pvz., idiomas), kurių kalbos žiniomis pagrįstos sistemos gali ir nepastebėti.

Automatinis vertimas iš lietuvių kalbos – itin didelis iššūkis.

Ir taisyklėmis, ir duomenimis pagrįstų automatinio vertimo sistemų pranašumai ir trūkumai paprastai kompensuoja vieni kitus, todėl šiuo metu tyrėjai daugiausia dėmesio skiria hibridiniams metodams, susiejantiems šias dvi technologijas. Vienas toks metodas yra pagrįstas ir į taisykles, ir į duomenis orientuotomis sistemomis, be to, turi modulį, sugebantį pasirinkti, kuris būdas geriausiai tinka konkrečiam sakiniui išversti. Tačiau sakinių, ilgesnių nei, tarkim, 12 žodžių, vertimo rezultatai tikrai nebus tobuli. Tokiu atveju geriausiai pasirinkti pagal prasmę tinkamiausias kiekvieno vertimo varianto dalis. Šis procesas gali būti gana sudėtingas, kadangi alternatyvių variantų sutampančios dalys ne visuomet yra aiškios, jas reikia papildomai sugretinti.

Automatinis vertimas iš lietuvių kalbos – itin didelis iššūkis. Laisva žodžių tvarka ir veiksmažodinės konstrukcijos sunkina analizę, o dėl linksniuočių įvairovės sunku parinkti reikiamos giminės ir linksnio žodžius.

Mažiau vartojamų kalbų, tokių kaip baltų kalbos, automatinio vertimo tyrimų įrankiai, kaip ir apskritai pačios KT, nėra labai gerai išplėtoti. Lietuvoje atlikta keletas su automatiniu vertimu susijusių darbų. Internete šiuo metu galima rasti tris vertimo įrankius: projektą WIFTA [26], sistemą *Google translator* ir *Vertimo vedlį* [27]. Pirmoji sistema buvo sukurta 2008 metais drauge su Rusijos bendrove „ProMT“, jos pagrindu tapo taisyklėmis pagrįsta technologija. Ši sistema verčia atsižvelgdama į teksto morfologines, sintaksines ir semantines savybes. Projektas sėkmingai baigtas ir nuo 2008 m. prieinamas internetu: http://vertimas.vdu.lt. Į ją kreipiamasi 127 mln. kartų per metus ir ja naudojasi maždaug 1 mln. unikalių vartotojų per metus. Registruotiems vartotojams suteikta galimybė naudotis kompiuterinių terminų ir verslo žodynais.

Atvirai internetu prieinamas Vytauto Didžiojo universitete sukauptas dabartinės rašomosios lietuvių kalbos tekstynas, turintis apie 140 mln. žodžių [28]. Be to, sudaromas lygiagretusis lietuvių kalbos ir kitų (anglų, vo-

kiečių, čekų) kalbų tekstynas, rengiama ir daugiau specialioms sritims skirtų tekstynų (pavyzdžiui, Vilniaus universitete sukauptas lietuvių mokslo kalbos tekstynas *CorALit* [29]). Deja, dabartinio lietuvių kalbos tekstyno modernioms lietuvių KT (informacijos paieškos, automatinio vertimo ir kitoms sistemoms) nebepakanka. Esamiesiems ir būsimiesiems tekstynams reikia bendros lietuvių kalbai pritaikytos programinės įrangos, kuri leistų kuo geriau išnaudoti turimus kalbos išteklius ir iš jų gaunamus skaitmeninius aprašus. Viena iš privalomų sąlygų klasikinėms automatinio vertimo sistemoms yra galimybė naudotis didžiuliu lygiagrečiuoju tekstynu, iš kurio kompiuteris gali mokytis. Didelių lygiagrečiųjų tekstynų trūkumas yra pagrindinė priežastis, dėl kurios Baltijos šalys tik dabar pradeda eksperimentuoti su automatinio vertimo sistemomis.

Sistemos *Google* automatinio vertimo programinė įranga yra pagrįsta statistiniu metodu, o jos galimybės apima apie 30 kalbų, taip pat ir lietuvių.

Vertimo vedlys – tai eksperimentinis automatinio vertimo įrankis, kurį sukūrė Lietuvių kalbos institutas kartu su UAB „Tilde IT“. Bandomoji versija verčia iš lietuvių kalbos į anglų kalbą. Automatizuota vertimo priemonė analizuoja sakinių struktūrą ir automatiškai siūlo sakinio, jo dalies ar atskirų žodžių vertimą. Ji yra pagrįsta statistine vertimo technologija ir sistemomis *Giza++* ir *Moses*. Bendrovė „Tilde IT“ kuria lietuvių–anglų kalbų automatinio vertimo sistemą, integruodama statistinius ir taisyklėmis pagrįstus automatinio vertimo metodus, diegdama novatoriškas iš keleto žodžių sudarytų frazių apdorojimo funkcijas. Vertimo tikslumas – 30 proc., o sistema yra nuolat tobulinama. Bendrovės „Tilde IT“ automatinio vertimo sistema naudojama ne tik tekstams versti, bet ir ieškoti informacijos skirtingų kalbų šaltiniuose.

Tikimasi, kad automatinio vertimo sistemų kokybę dar galima labai pagerinti. Paprastai šioje srityje susiduriama su tokiomis problemomis, kaip galimybės pritaikyti kalbos išteklius tam tikrai sričiai ar naudotojo poreikiams, sukurti jų darbinę sąveiką su terminijos bazėmis ir vertimų atmintimi. Be to, dauguma šiuo metu egzistuojančių sistemų yra pritaikytos anglų kalbai, galimybės versti iš lietuvių kalbos į kitokias kalbas ir iš kitokių kalbų į lietuvių kalbą yra ribotos, dėl to stringa bendras vertimų srautas, taip pat naudotojai privalo mokytis skirtingose sistemose taikomų skirtingų priemonių žodynams koduoti.

Vertinimas padeda palyginti automatinio vertimo sistemų kokybę, skirtingus metodus ir šių sistemų būklę skirtingose kalbų porose. Pateiktoje 7 lentelėje, kuri sudaryta Europos Komisijos projekto „Euromatrix+“ metu, parodyti dvidešimt dviejų iš dvidešimt trijų oficialiųjų Europos Sąjungos kalbų (airių kalba į lyginimą nebuvo įtraukta) lyginimo rezultatai. Rezultatai išdėstyti eilės tvarka, remiantis įvertinimo skale BLEU, pagal kurią už geresnį vertimą skiriamas aukštesnis balas [30]. Vertėjas žmogus gautų apie 80 balų įvertinimą.

Geriausi rezultatai (pažymėti žalia ir mėlyna spalva) buvo pasiekti verčiant kalbas, kurios yra išsamiai ištirtos pagal koordinuotas programas ir turinčios daugybę lygiagrečiųjų tekstynų (pvz., anglų, prancūzų, olandų, ispanų ir vokiečių kalbos). Kalbos, kurių įvertinimo rezultatai prastesni, pažymėtos raudona spalva. Šios kalbos arba nepakankamai ištirtos, arba jų struktūra smarkiai skiriasi nuo kitų kalbų (pvz., vengrų, maltiečių ir suomių kalbos).

4.3 KITOS TAIKYMO SRITYS

KT taikomųjų programų kūrimas yra susijęs su aibe papildomų užduočių, kurios ne visuomet matomos sistemos naudotojui, tačiau suteikia sistemai daugiau funkcionalumo. Visos šios užduotys yra svarbūs mokslinių tyrimų objektai, pastaruoju metu tapę kompiuterinės lingvistikos atšakomis.

Pavyzdžiui, atsakymai į klausimus šiuo metu yra aktyvi mokslinių tyrimų sritis, kuriai parengta anotuotų teks-

	Kalba, į kurią verčiama – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

7: 22 Europos Sąjungos kalbų automatinio vertimo rezultatai – Machine translation for 22 EU-languages [31]

tynų ir paskelbta mokslinių konkursų. Atsakymų į klausimus sąvoka apima daugiau nei vien paiešką pagal esminius žodžius (kurios metu paieškos sistema pateikia galimai tinkamų dokumentų rinkinį), ji suteikia naudotojams galimybę užduoti konkretų klausimą, į kurį sistema pateikia vienintelį atsakymą. Pavyzdžiui:

> *Klausimas: Kiek metų turėjo Neilas Armstrongas, kai jis išsilaipino Mėnulyje?*
>
> *Atsakymas: 38.*

Akivaizdu, kad atsakymai į klausimus yra susiję su internetine paieška, tačiau šiuo metu šis terminas apima ir mokslinių tyrimų problemas – kokie gali būti skirtingi klausimų tipai ir būdai į juos atsakyti; kaip turi būti analizuojami ir palyginami dokumentai, kuriuose gali slypėti atsakymas (ar tokiuose dokumentuose pateikiami prieštaringi atsakymai); kaip galima konkrečią informaciją (atsakymo) patikimai išgauti iš dokumento, neignoruojant konteksto.

Kalbos technologijų taikomosios programos, įdiegtos didesnėse programinėse sistemose, gali atlikti labai svarbias funkcijas.

Savo ruožtu tai yra susiję su informacijos išgavimu (angl. *information extraction*), sritimi, kuri buvo itin populiari ir daranti poveikį, kai praėjusio amžiaus dešimtojo dešimtmečio pradžioje kompiuterinė lingvistika pasuko statistikos linkme. Informacijos išgavimo tikslas – nustatyti tam tikruose dokumentuose glūdinčią konkrečią informaciją, pavyzdžiui, nustatyti pagrindinius bendrovių įsigijimo proceso dalyvius, apie kuriuos buvo rašoma laikraščių straipsniuose. Kita įprasta tyrimų erdvė – teroristų išpuolių ataskaitos. Čia svarbiausia tekstą užpildyti pagal šabloną, kuriame nurodomas kaltininkas, taikinys, išpuolio laikas, vieta ir rezultatai. Konkrečioms sritims pritaikytų šablonų pildymas yra svarbiausias bruožas, dėl kurio informacijos

išgavimas priskiriamas „užkulisinėms“ technologijoms, formuojančioms dar vieną tiksliai apibrėžtą tyrimų sritį, kuri praktiškai turi būti įtvirtinta tinkamoje taikomojoje terpėje.

Santraukų kūrimas ir **tekstų generavimas** – dvi tarpinės technologijos, kurios gali veikti ir savarankiškai, ir atlikti pagalbinę funkciją didesnėse sistemose. Santraukos glaustai pateikia ilgo teksto esmę, tai yra viena iš programos *Microsoft Word* funkcijų. Paprastai pagal šį metodą statistiniu būdu nustatomi „svarbiausi“ teksto žodžiai (t. y. žodžiai, kurie apdorojamame tekste pasitaiko dažniausiai, nors šiaip kalboje jie gerokai retesni) ir sakiniai, kuriuose tokių „svarbių“ žodžių yra daugiausia. Tokie sakiniai paimami iš teksto ir sudedami, suformuojant santrauką. Itin paplitusiame komerciniame modelyje santraukos kūrimas yra tiesiog būdas sakiniams išrinkti, suskaidant tekstą į jo sakinių poaibį. Alternatyvus, kol kas nedaug ištirtas metodas yra generuoti visiškai naujus sakinius, kurių pirminiame tekste nėra. Tuo atveju reikia gerai išanalizuoti ir suvokti tekstą, taigi kol kas šis metodas nėra labai patrauklus ir visuotinai taikomas. Paprastai tekstų generatorius retai kada funkcionuoja kaip savarankiška programa, jis būna įdiegtas į didesnių sistemų terpę. Pavyzdžiui, teksto generatorių galima rasti klinikinės informacijos sistemose, kaupiančiose, saugančiose ir apdorojančiose ligonių duomenis. Ataskaitų rengimas yra tik dar viena sritis, kurioje gali būti pritaikoma santraukų kūrimo technologija.

Lietuvių kalbai šios technologijos nėra pakankamai išplėtotos, palyginti su anglų kalba, ir kol kas yra tik eksperimentinio lygmens: Vytauto Didžiojo universitete atliekami pavieniai lietuviškų tekstų santraukų rengimo, švietimo ir mokslo terminų automatinio identifikavimo tyrimai ir pan.

Lietuvių kalba įtraukta į tarptautinius projektus. Latvijos bendrovė SIA „Tilde“ įgyvendina FP5 projektą CLARITY: „Pasiūlymas dėl informacijos paieškos skirtingomis kalbomis ir tekstų bei garsinių dokumentų organizavimo“. Informacijos paieškos skirtingomis kalbomis sistema CLARITY buvo sukurta anglų–latvių, latvių–anglų, vokiečių–latvių, latvių–vokiečių, rusų–latvių, latvių–rusų, lietuvių–anglų, anglų–lietuvių, vokiečių–lietuvių, lietuvių–vokiečių, lietuvių–rusų ir rusų–vokiečių kalbų poroms. Kalbant apie baltų kalbas, dokumentų paieškos naudojant tiesioginio užklausos vertimo funkciją rezultatai rodo, kad tikslumo vidurkis gali siekti daugiau nei 70 proc., palyginti su vienakalbės paieškos rezultatais.

4.4 ŠVIETIMO PROGRAMOS

KT – tarpdalykinė sritis, jungianti kalbos mokslą, informatiką, matematiką, filosofiją, psicholingvistiką ir kitus susijusius mokslus. Kaip atskira disciplina ji kol kas nėra įtvirtinta Lietuvos aukštojo mokslo sistemoje. Keletas universitetų yra įsteigę atskirus kompiuterinės lingvistikos centrus (pvz., Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras) arba laboratorijas (pvz., Kauno technologijos universiteto Kalbos tyrimų laboratorija). Šiuo metu Kauno technologijos universiteto Humanitarinių mokslų fakultete dėstoma vienintelė kompiuterinės lingvistikos bakalauro studijų programa. Ši programa buvo pradėta 2003 metais, o iki 2010 metų ją baigė 73 studentai. Toks nedidelis absolventų skaičius negali patenkinti nuolat didėjančios kvalifikuotų KT srities darbuotojų paklausos.

Vilniaus universitete ir Vytauto Didžiojo universitete artimos srities studijų programose dėstoma kompiuterinės lingvistikos ir KT kursų. Nuo 2011 metų Vilniaus universiteto Kauno humanitarinių mokslų fakultete pradėta audiovizualinių vertimų magistro studijų programa. Vytauto Didžiojo universitete nuo 2006 m. veikia skaitmeninės lingvistikos magistro studijų programa. Kol kas nė vienas universitetas nesiūlo nuoseklių visų lygmenų studijų, dėl to KT srityje dažniausiai dirba mokslininkai, baigę lingvistines ir (arba) informatikos studijas.

Mokslinė tyrimų bazė formuojama ir ištekliai kaupiami Matematikos ir informatikos bei Lietuvių kalbos institutuose (pastarajame 2010 m. įkurta Skaitmeninių kalbos išteklių laboratorija).

4.5 NACIONALINIAI PROJEKTAI IR INICIATYVOS

Lietuvoje spartesnė informacinės visuomenės plėtra, susidomėjimas KT ir išteklių kaupimas prasidėjo vos prieš keletą dešimtmečių. Kadangi kalbančiųjų lietuvių kalba nėra daug, KT komercinė rinka nėra labai didelė, be to, Lietuvoje nėra tokių šiuolaikinių technologijos gigantų kaip BMW ar NOKIA, o KT srityje dirba vos keletas verslo įmonių.

Didžioji dalis iniciatyvų ir įsipareigojimų dėl lietuvių kalbos funkcionavimo informacinėje visuomenėje ir KT kūrimo atsiranda nacionaliniu lygmeniu. 2000 metais pradėta įgyvendinti pirmoji nacionalinė programa „Lietuvių kalba informacinėje visuomenėje“, apimanti 2000–2006 m. laikotarpį. Šią programą koordinavo Valstybinė lietuvių kalbos komisija, ją įgyvendinant buvo sprendžiamos lokalizavimo, išteklių kūrimo ir kitos problemos:

- Automatinio kalbos atpažinimo plėtra, įskaitant lietuvių kalbos ypatybių tyrimą, atskirų šnekamosios kalbos žodžių atpažinimo priemonės prototipo kūrimą, Lietuvos radijo naujienų transliacijų garsyno LRNO tobulinimą, kompiuterinių dialogų lietuvių kalba tyrimus, kalbos sintezės kokybės gerinimą, lietuvių kalbos balso technologijų bandomųjų programų kūrimą, automatinį lietuvių šnekamosios kalbos skaidymą ir lietuvių kalbos automatinės transkripcijos kūrimą.
- Lietuvių kalbos elementų standartizavimas informacinėse technologijose, įskaitant kompiuterinio šrifto *Palemonas* sukūrimą, lokalizaciją ir pan.
- Reikalingų išteklių vertimas ir priemonių kūrimas, įskaitant specializuotų tekstų pažodinio vertimo kompiuterinę sistemą, lietuvių ir čekų kalbų lygiagrečiojo tekstyno sudarymą ir atnaujinimą, morfologinės analizės ir generavimo įrankio sukūrimą.
- Pradėti lietuviškų tekstų sintaksinės ir semantinės analizės darbai.

Už programos „Lietuvių kalba informacinėje visuomenėje“ antrąjį etapą (2009–2013 metų) atsakingas Informacinės visuomenės plėtros komitetas prie Susisiekimo ministerijos. Programoje numatyta sukurti interneto portalą, kuriame bus galima nemokamai naudotis visais turimais kalbos ištekliais ir technologijomis, plėtoti turimus ir kurti naujus kalbos išteklius, gerinti automatinio kalbos atpažinimo ir kalbos sintezės technologijas, kurti naujus automatinio vertimo įrankius, gerinti ir kurti semantinės analizės ir informacijos paieškos priemones.

Skatinami šios srities moksliniai tyrimai bei išteklių kūrimas. Lietuvos mokslo taryba pradėjo pirmąją nacionalinę programą „Valstybė ir tauta: paveldas ir tapatumas“, apimančią ir paveldo skaitmeninimą, lituanistinio paveldo ir tapatumo vieningos informacinės infrastruktūros koncepcijos parengimą (šioje programoje įgyvendintas projektas „Lituanistinių skaitmeninių išteklių metaduomenų sistemos sukūrimas ir suderinimas su CLARIN“). Lietuvos mokslo taryba taip pat finansuoja „Nacionalinės lituanistikos plėtros“ 2009–2015 metų programą, kurios paskirtis – plėtoti ir skatinti lituanistikos mokslinius tyrimus, padėti įgyvendinti lituanistikos mokslinių tyrimų prioritetą, sustiprinti lituanistikos mokslinių tyrimų rezultatų indėlį į valstybės humanistikos plėtrą, suteikti mokslinį pagrindą tautinės savimonės ugdymui ir lituanistinio paveldo apsaugai.

Verslo įmonių, veikiančių KT srityje, nėra gausu. Galima būtų paminėti šias įmones: „Tilde IT“ , „Fotonija“, „Microsoft Lietuva“, „CID Baltic“, „Synergium“, „Sintagma“, „TokenMill“, „HLTech“.

Neabejotinas lyderis KT srityje yra UAB „Tilde IT“, kuri Lietuvos rinkoje dirba jau 12 metų. Bendrovė daug dėmesio skiria programinės įrangos lokalizavimui, techninių dokumentų vertimams, lietuvių kalbai pritaikytos programinės įrangos kūrimui. „Tilde IT“ yra viena didžiausių lokalizacijos paslaugų teikėjų Lietuvoje. Bendrovė nuolatos bendradarbiauja su tarptautinėmis lokalizacijos ir vertimų įmonėmis.

Šiuo metu „Tilde IT“ gerina automatinio vertimo kokybę, kuria ir tobulina rašybos bei gramatikos tikrinimo sistemas. Kartu su Lietuvių kalbos institutu, Matematikos ir informatikos institutu bei Vilniaus universiteto Filologijos fakultetu bendrovė inicijuoja ir įgyvendina mokslinių tyrimų ir plėtros projektus, kurių tikslas – sukurti programinės įrangos prototipų.

Semantinių sistemų projektus bendrovė „Tilde IT“ įgyvendina nuo 2008 m. Kadangi „Tilde IT“ Europos rinkai tiekia automatinio vertimo technologijas, naujosios technologijos bus naudojamos kartu su metodika, skirta gerinti automatinio vertimo rezultatus. „Tilde IT“ siekia sukurti lietuvių kalbos žodžių sąsajų duomenų bazę – lingvistinę semantikos bazę. Lietuvių kalbos semantikos informacinis tinklas galėtų padėti rinkodaros profesionalams prognozuoti visuomenės reakcijas į siūlomas produktų reklamos akcijas, pakuotes ar pavadinimus. Tokį mąstymo modelį galima pritaikyti kuriant naujus produktus ir generuojant naujas ar nestandartines idėjas. Vienas žodis gali turėti daugiau nei 15 sinonimų, nors paprastai žmonės žino vos 5 ar 6 iš jų. Išsamus semantikos tinklas padėtų išsaugoti lietuvių kalbos sinonimų įvairovę.

„Tilde IT“ prisijungė prie programos *Eurostars* projekto SOLIM (angl. *Spatial Ontology Language for multimedia Information Modeling* – „Daugialypės terpės informacijos modeliavimo erdvinės ontologijos kalba“). Projektas skirtas tobulinti informacijos analizę atsižvelgiant į kontekstą ir pasitelkus erdvės ir pokyčių sąvokas peržengti statiško pasaulio ribas. Šio projekto tikslas – internetinę ontologijos kalbą (angl. *Web Ontology Language*) pritaikyti veiksmingam erdvinės informacijos saugojimui ir aiškinimui bei pademonstruoti tokio pritaikymo naudą automatiškai apdorojant tekstinę ir grafinę informaciją.

Nuo 1991 metų UAB „Fotonija“ diegė lietuvių kalbą skaitmeninėje terpėje, kurdama ir tobulindama tvarkykles *WinLika*, *Lika*, projektuodama lietuvių kalbos šriftą *Aistika*, kurdama teksto tvarkymo programą *Mainukai*, dokumentų konvertavimo priemonę *Korektorius*, tekstų kūrimo, redagavimo ir korektūros programą *Redaktorius*, rašybos klaidų taisymo programą *Juodos Avys*. Svarbi bendrovės „Fotonija“ darbo sritis – vienakalbių ir daugiakalbių žodynų sudarymas. Tai tarptautinių žodžių žodynas *Interleksis*, anglų–lietuvių kalbų žodynas *Anglonas* ir jo atitikmuo prancūzų kalba *Frankonas*.

Lokalizavimo, ontologijų kūrimo ir kitose KT srityse projektus įgyvendina ir kitos verslo įmonės, pavyzdžiui, „Microsoft Lietuva“, „CID Baltic“, „Synergium“, „Sintagma“, „TokenMill“, „HLTech“ ir kt.

Įgyvendinus ankstesnes programas ir projektus buvo sukurta ir išplėtota keletas svarbių lietuvių kalbai pritaikytų priemonių ir išteklių. Kitame skyriuje apibendrinama dabartinė lietuvių KT būklė.

4.6 TURIMI KALBOS IŠTEKLIAI IR ĮRANKIAI

8 lentelėje apibendrinama dabartinė KT pritaikymo lietuvių kalbai būklė. Turimų įrankių ir išteklių įvertinimo balai pagrįsti konkrečios srities ekspertų nuomone, kurie pateikė vertinimus pagal septynis parametrus nuo 0 (labai žemas) iki 6 (labai aukštas).

Lietuvių kalbos būklės rezultatus apibendrintai galima pateikti taip:

- Moksliniai tyrimai leido sėkmingai sukurti gana kokybišką programinę įrangą bazinei teksto analizei,

	Kiekybė	Prieinamumas	Kokybė	Aprėptis	Išbaigtumas	Tvarumas	Pritaikomumas
Kalbos technologijos (įrankiai, technologijos ir pritaikymo sprendiniai)							
Kalbos atpažinimas	2	0	2	1	1	0	2
Kalbos sintezė	3	2	2,5	2,5	1,5	1	2
Gramatinė analizė	2	1,5	2,5	2	1,5	1	2
Semantinė analizė	1,3	1	1,3	1	0	0	0,3
Teksto generavimas	0	0	0	0	0	0	0
Automatinis vertimas	2	3	2,5	2,5	2	2	2
Kalbos ištekliai (ištekliai, duomenys ir žinių bazės)							
Tekstynai	1,5	1,5	2,5	2,5	2	2,5	2,5
Garsynai	2	1	2	2	1	1	2
Lygiagretieji tekstynai	2	2	1,5	1,5	2	2	4
Leksikos ištekliai	2,5	2	2,5	2	2	0,5	2,5
Gramatikos	0	0	0	0	0	0	0

8: Kalbos technologijų pritaikymo lietuvių kalbai būklė

pavyzdžiui, įrankius morfologinei ir sintaksinei analizei. Tačiau pažangesnių technologijų, kurioms 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 ir pan.), daugiau dėmesio reikia skirti nuodugnesniam lingvistiniam apdorojimui.
- Nors sukaupta neblogos kokybės specializuotų tekstynų ar garsynų, jie nepakankamai parengti, kai kurie iš jų yra pasiekiami tik naudojantis specializuotomis, individualiomis prieigos priemonėmis, kai kuriais naudotis galimybių iš viso nėra.
- Nemaža dalis išteklių ir įrankių nestandartizuoti, jų tvarumas nėra efektyviai užtikrinamas. Norint standartizuoti duomenis ir jų perdavimo formatus būtinos koordinacinės programos ir iniciatyvos.
- Trūksta automatiniam vertimui skirtų lygiagrečiųjų tekstynų. Kol kas labiau išplėtotas vertimas iš lietuvių į anglų kalbą, kadangi šiai kalbų porai yra sukaupta daugiausia duomenų.
- Labai trūksta daugialypės terpės duomenų.

Apibendrinant galima teigti, kad daugelyje lietuvių kalbos tyrimų specifinių sričių šiandien turime tik riboto funkcionalumo programinę įrangą. Sudėtingesni įrankiai (angl. *advanced tools*), pavyzdžiui, sintaksiškai anotuoti tekstynai (angl. *treebanks*), leksinės semantinės žinių bazės ar sąvokų taksonomijos, tokios kaip *WordNet* ir pan., lietuvių kalbai dar nesukurti arba tik kuriami. Nors neseniai sukurti automatinio vertimo įrankiai, pažangiausi ištekliai, vadinamieji bendrieji taiky-

mai (angl. *general applications*), tik dabar pradedami plėtoti [32]. Akivaizdu, tolesni tyrimai turėtų užpildyti išsamesnės semantinės tekstų analizės spragą ir pasirūpinti trūkstamų išteklių, tokių kaip lygiagretieji tekstai automatiniam vertimui, *WordNet* ir pan., kaupimu.

4.7 KALBŲ PALYGINIMAS

Įvairiose kalbos vartotojų bendruomenėse KT taikymo lygmuo ir būklė yra skirtingi. Šiame skirsnyje Europos kalbos bus lyginamos pagal šias taikymo kategorijas: automatinį vertimą ir šnekamosios kalbos apdorojimą, teksto analizę, taip pat bus vertinami pamatiniai ištekliai, būtini KT plėtoti. Kalbos suskirstytos į penkias sankaupas:

1. Puikus palaikymas
2. Geras palaikymas
3. Vidutinis palaikymas
4. Fragmentiškas palaikymas
5. Menkas palaikymas arba jo visai nėra

KT palaikymo lygmuo nustatomas remiantis šiais kriterijais:

Šnekamosios kalbos apdorojimas: kalbos atpažinimo technologijų kokybė, kalbos sintezės technologijų kokybė, sričių aprėptis, garsynų skaičius ir dydis, šnekamąja kalba paremtų technologijų pritaikymo mastas ir įvairovė.

Automatinis vertimas: automatinio vertimo kokybė, kalbų porų kiekis, kalbos sričių ir reiškinių aprėptis (*linguistic phenomena and domains*), lygiagrečiųjų tekstynų dydis ir kokybė, automatinio vertimo pritaikymų kiekis ir įvairovė.

Teksto analizė: teksto analizės technologijų (morfologijos, sintaksės, semantikos) kokybė ir aprėptis, lingvistinių reiškinių ir sričių aprėptis, pritaikymų kiekis ir įvairovė, anotuotų tekstynų kokybė ir dydis, leksinių išteklių (pvz., *WordNet*) ir gramatikų kokybė ir aprėptis.

Ištekliai: tekstynų, garsynų ir lygiagrečiųjų tekstynų kokybė ir dydis, leksinių išteklių ir gramatikų kokybė ir aprėptis.

9–12 lentelėse matyti, kad lietuvių kalba gerokai atsilieka nuo KT lyderių, pavyzdžiui, anglų kalbos, kuri pirmauja bemaž visose sankaupose. Lietuvių kalba dažniausiai atsiduria toje pačioje sankaupoje, kaip ir kitos mažiau vartotojų turinčios, taigi ne taip komerciškai patrauklios Europos kalbos, tokios kaip latvių, slovakų, slovėnų. Kita vertus, lietuvių kalbos ištekliai ir technologijos labai netolygiai išplėtoti, pavyzdžiui, sukauptos gana gausios ir išsamios terminų duomenų bazės, tačiau nėra *WordNet*, tezauro ir pan. Visai nėra KT pritaikytos lietuvių kalbos gramatikos. Tai trukdo sėkmingai kurti kalbos modelius.

Itin silpnai išplėtoti semantikos tyrimai lėmė lėtesnę kalbos generavimo, teksto interpretavimo ir teksto analizės pažangą. Tuo tarpu šnekamosios kalbos apdorojimo kai kurios technologijos veikia pakankamai gerai ir sėkmingai integruojamos versle. Sparčiau plėtojami šnekos sintezės tyrimai ir taikymas, o kalbos atpažinimas gerokai sudėtingesnis.

Vis dėlto kuriant išmanesnes ir sudėtingesnes priemones, tokias kaip automatinis vertimas, reikia išteklių ir technologijų, kurie apimtų daugiau lingvistinių aspektų ir leistų semantiškai nuodugniau analizuoti įvedamą tekstą. Gerindami kai kurių bazinių išteklių kokybę ir aprėptį, turėtume gebėti atverti naujas galimybes įsiveržti į pažangesnių technologijų taikymo sritis, tarp jų ir itin geros kokybės automatinį vertimą.

4.8 IŠVADOS

Šioje Baltųjų knygų serijoje pirmą kartą mėginome įvertinti, kokiu lygiu KT yra pritaikomos 30 Europos kalbų ir atlikti lyginamąją analizę. Žinodamos spragas, poreikius ir trūkumus, Europos bendruomenė, plėtojanti KT, susijusios verslo įmonės dabar turi galimybių inicijuoti platesnio masto mokslinius tyrimus ir plėtros programas, kad

būtų sukurta iš tikrųjų daugiakalbė ir technologiškai pažangi Europa.

Atskleidėme didžiulius skirtumus tarp Europos kalbų. Kai kurios kalbos turi gana geros kokybės programinę įrangą ir išteklius, tačiau kitoms (dažniausiai „mažesnėms“ kalboms) to trūksta. Daugelis kalbų neturi svarbiausių teksto analizės technologijų ir būtiniausių išteklių toms technologijoms plėtoti. Kitos kalbos turi pagrindinių įrankių ir išteklių, tačiau kol kas nepajėgta investuoti į semantinį teksto apdorojimą. Mums būtina sutelkti visas įmanomas pastangas, kad įgyvendintume itin ambicingą tikslą – sukurtume Europos kalboms aukštos kokybės automatinį vertimą.

KT būklė Lietuvoje teikia pagrindo nuosaikiam optimizmui. Lietuvos Respublikos Vyriausybė pabrėžia siekį užtikrinti KT plėtrą – tai rodo įvairių vyriausybinių institucijų ir Europos Sąjungos struktūrinių fondų finansuojamos programos, pagal kurias kuriamos ir tobulinamos KT. Keturi Lietuvos universitetai ir du mokslinių tyrimų institutai kuria KT mokslinę bazę. Verslo sektoriuje „Tilde IT“ yra pagrindinis dalyvis, kuriantis lietuvių kalbai pritaikytas technologijas.

Turima keletas bendrinei lietuvių kalbai skirtų technologijų, tai toli gražu neprilygsta pirmaujančiai šioje srityje anglų kalbai. Lietuvių kalba yra viena iš vadinamųjų „nekomercinių“ Europos kalbų, todėl plėtodama KT ji susiduria su sunkumais ir problemomis, būdingomis mažiau vartojamos kalbos raidai. Šių technologijų plėtra labai priklauso nuo kitų šalių patirties ir jų paramos bei tarptautinio bendradarbiavimo. Kita vertus, KT plėtojimas yra svarbiausia lietuvių kalbos funkcionalumo, žinomumo ir studijų bei lietuviškos kultūros sklaidos daugiakalbėje Europoje stiprinimo proceso sudedamoji dalis.

Tai rodo, kad būtina stengtis kaupti lietuvių kalbos išteklius, atlikti daugiau mokslinių tyrimų ir diegti naujoves. Be to, dėl būtinybės sukaupti didelį kiekį duomenų ir KT sistemų sudėtingumo reikia sukurti naujų informacijos mainų ir bendradarbiavimo infrastruktūrų.

Mūsų įžvalgos rodo, kad vienintelė alternatyva – sutelkti pastangas lietuvių kalbos išteklių kūrimui ir juos efektyviai panaudoti moksliniams tyrimams, inovacijoms ir plėtrai. Didesnių išteklių sankaupų poreikiui tenkinti ir ypač sudėtingoms KT sistemoms kurti būtinos naujos infrastruktūros ir sutelktesnė mokslinių tyrimų organizacija, užtikrinanti didesnę sklaidą ir bendradarbiavimą.

Be to, mokslinių tyrimų ir plėtros finansavimas dažnai trumpalaikis. Paprastai trumpalaikes suderintas programas keičia menko finansavimo ar netgi visiško nefinansavimo laikotarpis. Taip pat akivaizdžiai trūksta Europos Sąjungos šalių inicijuotų ir Europos Komisijos vykdomų programų koordinavimo.

Galime daryti išvadą, kad būtina didelės aprėpties koordinuota iniciatyva, skirta įveikti KT parengtumo skirtumus visose Europos kalbose.

META-NET tinklo ilgalaikis uždavinys – pristatyti kokybiškas KT, taikomas visose Europos kalbose, siekiant kultūrine įvairove pagrįstos politinės ir ekonominės vienybės. Šios technologijos padės sugriauti dabartinius barjerus ir nutiesti tiltus tarp Europos kalbų. Visos suinteresuotos šalys – politikai, tyrėjai, verslo ir visuomenės atstovai – turi suvienyti savo pastangas, kurdami bendrą ateitį.

Puikus palaikymas	Geras palaikymas	Vidutinis palaikymas	Fragmentiškas palaikymas	Menkas/nėra palaikymo
	anglų	vokiečių italų suomių prancūzų olandų portugalų ispanų čekų	baskų bulgarų danų estų galisų graikų airių katalonų norvegų lenkų švedų serbų slovakų slovėnų vengrų	islandų kroatų latvių **lietuvių** maltiečių rumunų

9: Šnekamosios kalbos apdorojimas: 30 Europos kalbų būklė

Puikus palaikymas	Geras palaikymas	Vidutinis palaikymas	Fragmentiškas palaikymas	Menkas/jokio palaikymo
	anglų	prancūzų ispanų	vokiečių italų katalonų olandų lenkų rumunų vengrų	baskų bulgarų danų estų suomių galisų graikų airių islandų kroatų latvių **lietuvių** maltiečių norvegų portugalų švedų serbų slovakų slovėnų čekų

10: Automatinis vertimas: 30 Europos kalbų būklė

Puikus palaikymas	Geras palaikymas	Vidutinis palaikymas	Fragmentiškas palaikymas	Menkas/nėra palaikymo
	anglų	vokiečių prancūzų italų olandų ispanų	baskų bulgarų danų suomių galisų graikų katalonų norvegų lenkų portugalų rumunų švedų slovakų slovėnų čekų vengrų	estų airių islandų kroatų latvių **lietuvių** maltiečių serbų

11: Teksto analizė: 30 Europos kalbų būklė

Puikus palaikymas	Geras palaikymas	Vidutinis palaikymas	Fragmentiškas palaikymas	Menkas/nėra palaikymo
	anglų	vokiečių prancūzų italų olandų lenkų švedų ispanų čekų vengrų	baskų bulgarų danų estų suomių galisų graikų katalonų kroatų norvegų portugalų rumunų serbų slovakų slovėnų	airių islandų latvių **lietuvių** maltiečių

12: Kalbos ir teksto ištekliai: 30 Europos kalbų būklė

5

APIE META-NET TINKLĄ

META-NET yra kompetencijos tinklas, finansuojamas Europos Komisijos [33]. Šiuo metu tinklą sudaro 54 nariai iš 33 Europos šalių. META-NET puoselėja Daugiakalbės Europos technologijos aljansą (angl. *Multilingual Europe Technology Alliance*, META) – gausėjančią KT profesionalų ir organizacijų bendruomenę.

META-NET bendradarbiauja su kitomis programomis, pavyzdžiui, CLARIN, *Bendrąja kalbos išteklių ir technologijų infrastruktūra* (angl. *Common Language Resources and Technology Infrastructure*), padedančia atlikti skaitmeninius humanitarinių mokslų tyrimus. META-NET puoselėja realiai daugiakalbės Europos informacinės visuomenės technologinius pagrindus, kurie:

- suteiks galimybę bendrauti ir bendradarbiauti skirtingomis kalbomis;
- užtikrins lygias galimybes naudotis informacija ir žiniomis, pateiktomis bet kuria kalba;
- pasiūlys Europos gyventojams pažangių, į tinklą sujungtų informacinių technologijų.

META-NET skatina ir propaguoja visoms Europos kalboms skirtas KT. Šios technologijos teikia galimybių atlikti įvairių sričių automatinį vertimą, kurti produktus, apdoroti informaciją ir valdyti žinias, naudojantis skirtingomis taikomosiomis programomis. Šio tinklo tikslas – tobulinti šiuo metu taikomus metodus, kad bendrauti ir bendradarbiauti skirtingomis kalbomis taptų lengviau. Europiečiai turi lygias teises į informaciją ir žinias, nesvarbu, kokia kalba jie kalba.

META-NET buvo pristatytas 2010 m. vasario 1 dieną. Tikslas – remti KT mokslinius tyrimus. Tinklas remia Europą, susijungusią į vieną skaitmeninę rinką ir informacinę erdvę. META-NET veikla apima kelias kryptis, padedančias siekti savo tikslų. Šios veiklos kryptys yra susijusios su KT plėtros vizijos kūrimu (META-VISION), sklaida (META-SHARE) ir moksliniais tyrimais (META-RESEARCH).

META-VISION skirta telkti dinamišką ir įtakingą suinteresuotųjų šalių bendruomenę, numatant KT plėtros viziją ir strateginių mokslinių tyrimų darbotvarkę (angl. *Strategic Research Agenda*, SRA). Pagrindinis šios veiklos tikslas – suburti susijusią ir darnią Europos KT bendruomenę, suvienijant itin suskaidytų ir skirtingų suinteresuotųjų šalių atstovus. Kartu su šia Baltąja knyga buvo parengti dar 29 tomai kitomis kalbomis. Bendrąją technologiją sukūrė trijų sektorių vizijų grupės. Rengiant šią viziją glaudžiai bendradarbiauta su visa KT bendruomene. Tuo tikslu buvo įsteigta ir META technologijų taryba.

META-SHARE kuria atvirą, visiems prieinamą įrankį keistis ir dalytis ištekliais. Saugyklų tinkle, kuriame ištekliais galės keistis visi naudotojai (angl. *peer-to-peer*), bus pateikiami kalbų duomenys, priemonės, teikiamos internetinės paslaugos, pagrįstos itin aukštos kokybės metaduomenimis ir suskirstytos į standartines kategorijas. Šiais ištekliais lengva naudotis, paieška galima pagal vienodus kriterijus. Tarp turimų išteklių – ir nemokama, atvirojo kodo medžiaga, ir riboto naudojimo mokami duomenys.

META-RESEARCH tiesia tiltus į su kalbos technologijomis susijusių technologijų sritis. Siekiama pasinaudoti kitų sričių pranašumais ir inovatyviais tyrimais, kurie galėtų praversti KT.

office@meta-net.eu – http://www.meta-net.eu

EXECUTIVE SUMMARY

During the last 60 years, Europe has become a distinct political and economic structure, yet culturally and linguistically it is still very diverse. This means that from Portuguese to Polish, Italian to Icelandic, everyday communication between Europe's citizens as well as communication in the spheres of business and politics is inevitably confronted by language barriers. The EU's institutions spend about a billion euros a year on maintaining their policy of multilingualism, i. e., translating texts and interpreting spoken communication. Yet does this have to be such a burden? Modern language technology and linguistic research can make a significant contribution to pulling down these linguistic borders. When combined with intelligent devices and applications, language technology will in the future be able to help Europeans talk easily to each other and do business with each other even if they do not speak a common language.

Language technology builds bridges.

The Lithuanian economy takes greater advantage than others of the European single market: in 2010, trade within the EU accounted for 61% of Lithuanian exports, and trade with other European countries totalled another 3%. But language barriers can bring business to a halt, especially for SMEs who do not have the financial means to reverse the situation.

The alternative to this kind of multilingual Europe would be to allow a single language to take a dominant position and end up replacing all other languages. However, this would create difficulties for the multilingual citizens of Europe. One classic way of overcoming the language barrier is to learn foreign languages. Yet without technological support, mastering the 23 official languages of the member states of the European Union and some 60 other European languages is an insurmountable obstacle for the citizens of Europe and its economy, political debate, and scientific progress.

The solution is to build key enabling technologies. These will offer European actors tremendous advantages, not only within the common European market but also in trade relations with third countries, especially emerging economies. In the long run, language technology solutions will eventually serve as a unique bridge between Europe's languages.

Language technology is a key for the future.

Information technology changes our everyday lives. We typically use computers for writing, editing, calculating, and information searching, and increasingly for reading, listening to music, viewing photos and watching movies. We carry small computers in our pockets and use them to make phone calls, write emails, get information and entertain ourselves, wherever we are. How does this massive digitization of information, knowledge and everyday communication affect our language? Will our language change or even disappear?

Many of the world's 6,000 languages will not survive in a globalized digital information society. It is estimated that at least 2,000 languages are doomed to extinction in the decades ahead. Others will continue to play a role in

families and neighbourhoods, but not in the wider business and academic world. What are the Lithuanian language's chances of survival? The status of the language depends not only on the number of speakers or books, films and TV stations that use it, but also on the presence of the language in the digital information space and software applications.

This is important to the Lithuanian language, which is one of the European languages that have a somewhat lower market appeal and a small user base, with just 4 million people speaking it, most of them residents of the Republic of Lithuania. The Lithuanian language enjoys the status of the state language, which is embedded in the Constitution. The enforcement of this status is regulated by the Law on the State Language and other legislation items. On top of that, the language has been included in the legislation on protecting cultural and ethnic heritage as part of the cultural identity. The *Programme on the Expansion of the Lithuanian Information Society* in 2011–2019 lists a strategic goal of improving the quality of living for the Lithuanian people and the condition of the corporate environment when it comes to using IRT possibilities and to achieve that at least 85 per cent of Lithuanian population have Internet access by 2019. This goal prioritises on expanding electronic content and services and promoting their usage. To that end, the government of Lithuania has set two objectives: (1) digitalising Lithuanian cultural heritage objects and use them as a basis for the development of digital products that would be available to the public, thus ensuring the conservation and dissemination of digital content in the electronic space; (2) integrating digital products of the Lithuanian language into IRT to ensure the full-scale functioning of the Lithuanian language in both its written and spoken form across the spheres of life of the nation. Will these political efforts be enough for the Lithuanian language to gain a toehold in the European multilingual information space?

After Lithuania had joined the European Union, the Lithuanian language entered a new phase of development, its new status of an official EU language standing as a guarantee that the Lithuanian language will be used and dispersed at European Union institutions. The development and application of language resources and technologies needed to ensure the full-scale functioning of the language in a multilingual environment has picked up pace as well. Still, the Lithuanian language is one of the so-called non-commercial European languages and is therefore facing the challenges and difficulties that are typical of the development of a language that has limited use. The development of such technology relies heavily on the experience of and assistance from other countries and international cooperation. On the other hand, developing language technologies is the most important element in the process of strengthening the functionality, recognition and learning of the Lithuanian language as well as the dissemination of the Lithuanian culture across the multilingual Europe. The full-on functioning of the Lithuanian language has become a particularly relevant consideration for the surviving and the development of the language.

Within the information society, the viability and appeal of a language is determined by the possibility to exchange multilingual information, receive services, and so on, in a prompt and convenient manner. Information technology is opening up new ways of communication, corpus development, information dissemination and retrieval for the Lithuanian language. The speed and geographical reach of modern communications makes it easier to use the Lithuanian language for human interaction, the quantity of Lithuanian content and services available on the Internet is growing and there are tools being developed that will help people use the language correctly, will serve to satisfy the special needs of the users, etc. On the other hand, the rate of change in this area is so high that the efforts aimed at the planning

and expansion of the Lithuanian language are no longer able to address every challenge on time. The fact that the users are able to access products and information in the English language faster and in a more user-friendly way result in the relatively low popularity of Lithuanianized software, the slow deployment and dispersion of language technology and tools, the inadequate development of digital language resources and tools.

Just like in many other European states, the development of the space of language technology in Lithuania is rather bumpy. Research has allowed designing, with some success, acceptable software for basic textual analysis, like parsers. However, advanced technology requiring more thorough knowledge of linguistic processing and semantics is still in its embryonic stage. The development efforts have produced quite a few digital language resources (e-dictionaries, corpuses, terminological dictionaries) and essential linguistic analysis tools (tools for determining and generating morphological attributes, spellcheckers), a Lithuanian synthesiser, Lithuanian – English machine translation systems, Lithuanianized software, an original Lithuanian computer font called *Palemonas*, which is geared towards scientific applications. Yet a lot of the available resources, products and systems require ongoing upgrades and development to keep up with the shifting user demands. The low grade of development of semantics research has resulted in stunted advancement on the field of language generation, textual interpretation and textual analysis. Even though there are quite voluminous and thorough lexicological databases available, there is still no WordNet, thesaurus, etc. Furthermore, there is no adequate Lithuanian grammar geared towards language technology or treebanks available.

Development of smart and sophisticated tools like machine translation requires resources and technology to cover more linguistic aspects and allow for a more detailed semantic analysis of text input. Improvement of the quality and scope of some basic resources should enable us to find new ways to blast our way into the areas of application of advanced technology.

So far, in terms of the quality, the space of Lithuanian language technology is rather fragmented and possessed of a very low degree of interaction. The available language resources that could be used as a basis to build language technology on have been developed by separate institutions, groups of researchers or businesses that did not necessarily follow the generally accepted standards, and therefore their compatibility with language technology is somewhat limited or economically not viable, bearing in mind that the resources should have to be rearranged to conform to the new standards.

Currently, there are several projects in progress in Lithuania aimed at applying the international standards to the older resources (like the Corpus of Modern Lithuanian) or designing new products. A higher degree of interaction would allow building integrated products that the common European linguistic space needs, such as machine translation tools, dictionaries, tools to search for semantic information, and would reduce the isolation of the Lithuanian-speaking community as well as boost the international prestige and accessibility of the Lithuanian language.

Language technology helps to unify Europe.

The expansion of the information society started picking up pace and a lively interest in language technologies occurred and the development of resources in Lithuania began just a few decades ago, and therefore, building language technology that is really effective and can be readily used in everyday life will require a substantial amount of research. The small market of users of language technology and tools, the underdeveloped and fragmented infrastructure of research and studies, the lack of clear priorities and coordination does little to promote ini-

tiative in private business. Currently, there are several companies involved in the field of language technology, and the number of orders for research coming from the business is really very small indeed.

The condition of language technology in Lithuania prompts cautious optimism. The government of the Republic of Lithuania emphasises the goal to ensure the expansion of language technology, as demonstrated by the programmes funded by various governmental institutions and the European Union structural funds, which are aimed at designing and improving language technology. Breakthroughs in language technology would promote its industrial application, help developing and improving public services and so on, and would allow using the Lithuanian language in every sphere of life and communication media.

The number and the applicability of technologies designed for and customised to European languages vary by quite a margin. Obviously, there is also a dramatic difference between Europe's member states in the efforts needed to promote research and development of the technology for specific languages as well as in terms of both the maturity of the research and in the state of readiness with respect to language solutions. Before truly effective language technology solutions are ready for everyday use, Lithuania still needs to conduct further research and come up with additional resources, tools, which will need to be integrated to ensure the highest degree of interaction possible.

META-NET's long-term goal is to introduce high-quality language technology for all European languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down the existing barriers and build bridges between Europe's languages. This requires all stakeholders – in politics, research, business, and society – to join their efforts for the future.

This white paper series complements other strategic actions taken by META-NET (see the appendix for an overview). Up-to-date information such as the current version of the META-NET vision paper or the Strategic Research Agenda (SRA) can be found on the META-NET website: http://www.meta-net.eu.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnessing a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLDING BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is an increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

The global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the web [2]. A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages?

Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [3].

The variety of languages in Europe is one of its richest and most important cultural assets.

While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [4].

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focussed primarily on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [5]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use the appropriate technology, just as we use technology to solve our transport and energy needs, among others.

Europe needs robust and affordable language technology for all European languages.

Language technology targeting all forms of written text and verbal discourse can help people collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It already often operates invisibly inside complex software systems to help us to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

To maintain our position at the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image or a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison came along to be able to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simu-

lation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggests a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

Technological progress needs to be accelerated.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce

their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems 'acquire' language capabilities in a similar manner. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete sample texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct errors in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we will describe the role of Lithuanian in the European information society and assess the current state of language technology for the Lithuanian language.

3

THE LITHUANIAN LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

The Lithuanian language is one of the least commonly used European languages. Only about four million people speak it, and most of them live in the Republic of Lithuania. Lithuanian, as the state language, is the common written and spoken language for all citizens of the Republic of Lithuania. Based on the 2011 census, Lithuania's population totals about 3,2 million, including roughly 2,7 million people who are Lithuanian by nationality.

The population of Lithuania consists of: Lithuanians (84%), Poles (6.1%), Russians (4.9%), Belarusians (1.1%), Ukrainians (0.6%), Jews (0.1%), Germans (0.1%), Latvians (0.1%), Tatars (0.1%), Karaits among others. There is also a Roma community of about 3,000 people, mainly settled in the Vilnius region (2001 census). Unfortunately, the Lithuanian population has been decreasing by the year since 2007 and the 2011 data indicate a figure of around 3.2 million people. Such changes are caused by the shrinking birth rate and emigration, which leads to a decrease in the number of users of the language who live in Lithuania.

Providing an accurate number of people who speak Lithuanian all over the world is quite difficult. It is estimated that some 500,000 Lithuanian speakers could be living abroad, while other sources point to at least 15 per cent of speakers. The Lithuanian language is spoken by Lithuanian ethnic minorities in Belarus, Poland, Latvia as well as the vast emigrant communities in the United States, Canada, the United Kingdom, Ireland, Spain, South America, etc. By the number of its speakers, the Lithuanian language places 144th in the world.

The Lithuanian language is one of the least commonly used European languages. Only about four million people speak it, and most of them live in the Republic of Lithuania.

The Lithuanian language is part of the Baltic branch of the Indo-European language family. Its next-of-kin is the Latvian language, which is spoken in neighbouring Latvia.

When it comes to classing languages into dominant and dominated in terms of the history of social development of European languages, the Lithuanian language should be considered as one of the latter. As a rule, dominant languages picked up one dialect as a basis for the formation of the standard language not later than during the Renaissance (as was the case with the English, French, Italian, Portuguese languages), while dominated languages, such as Bulgarian, Croatian, Lithuanian, Slovakian took shape during the Spring of Nations in the 19th century. The Lithuanian standard language evolved in the late 19th century – early 20th century.

The vast variety of regional strains is a defining feature of the Lithuanian language. There are two major dialects, the Highland (aukštaičiai) and the Lowland (žemaičiai, Samogitians). The strains of the Lithuanian language

are different both in terms of their phonetic characteristics and grammar as well as vocabulary. These dialects break down into fourteen sizeable regional sub-dialects, each consisting of branches typical of smaller regional units. Sub-dialects differ from each other by various sounds, word forms and other characteristics.

The common Lithuanian language evolved on the basis of one of the Higher Lithuanian sub-dialects in early 20th century; however, the regional identity and dialectal differences are still very obvious.

As of 1995, sign language was officially recognised as the vernacular of the deaf in the Republic of Lithuania. Ever since then, the Lithuanian sign language has been evolving as an independent language.

3.2 PARTICULARITIES OF THE LITHUANIAN LANGUAGE

In the 19th century, Indo-European linguists glorified the remarkable resemblance between the Lithuanian language and Sanskrit, and the language was honoured as a spoken Indo-European language with the least mutated structure. This is why those who can speak classical European languages (Latin, ancient Greek) tend to understand the Lithuanian grammar easier. Lithuania takes pride in the statement by French linguist Antoine Meillet that everyone who wants to hear the way the forefathers of the Indo-Europeans used to talk should go and listen to a Lithuanian peasant.

Certain linguistic characteristics of Lithuanian are challenges for computational processing.

Lithuanian is the most conservative of the living Indo-European languages: it has best preserved many of its archaic features. From the typological viewpoint, Lithuanian is important because of its many unique features, including its rich inflection, a distinctive synthesis of tonic and dynamic accent and an extremely variable word order that reflects the complicated relations between the communicative and the syntactic levels of discourse.

Letters make up millions of words of written Lithuanian. These millions require as many as 32 letters. This is the exact number of letters in the alphabet of Standard Lithuanian. It was fixed by Jonas Jablonskis in his Lithuanian Grammar (1901). Therefore, the present alphabet is over a hundred years old, but the history of its evolution is much longer. The alphabet of the Lithuanian language is based on Latin; nevertheless, it has its own peculiar characters, some of them original (like the letter *ė*). Others are borrowed from other languages (like the Czech *š, ž* or *ą, ę* borrowed from Polish). However, it is as yet difficult to resolve the complicated issue of stressed vowels in the Lithuanian language, which has particular relevance when it comes to transferring various dictionaries or stressed corpuses to the digital space – emphasis in the Lithuanian language is free and distinctive, which means that it can determine the lexical or grammatical meaning of the word, e. g.:

nãmo gen. sg. 'house' – *namõ* adv. 'home'.

Besides, all long syllables have one of two accents that also distinguish the meaning of words, e. g.:

áukštas 'high, tall' – *aũkštas* 'storey', etc.

These are the things to consider for developers of speech technologies as well.

In Lithuanian, which is an inflectional language, the majority of word forms are constructed with affixes, viz. endings and inflectional suffixes. The endings are the principal means of marking the syntagmatic relations between words in a sentence and/or the relations between word forms in a paradigm.

Endings mostly are fused, i. e., an ending encodes two or more grammatical meanings and thus a word form enters into the same number of morphological categories.

Suffixes are also widely used in Lithuanian to make up word forms. They mainly indicate paradigmatic relations between word forms rather than syntagmatic relations. Inflectional suffixes are used to mark the degrees of comparison in adjectives and many adverbs, some tense and mood forms in verbs, and also the non-finite verb forms: the infinitive, participles (including gerunds) and verbal adverbs (*būdinys*).

In word forms affixation is often (especially in the verbal paradigm) conjoined with changes in the root.

The grammatical means of marking syntactic relations in Lithuanian are endings and, less commonly, inflectional suffixes, often supplemented by structural words, viz. prepositions, conjunctions, and particles.

Alongside simple (synthetic) word forms, made with affixes, a paradigm may contain periphrastic (analytical) word forms comprised of the main word and an auxiliary. According to the shared morphological, syntactic and semantic properties, words are classified into grammatical classes, traditionally termed parts of speech. In Lithuanian 11 parts of speech are distinguished: the noun, the adjective, the numeral, the pronoun, the verb, the adverb, the particle, the preposition, the conjunction, the interjection and onomatopoeic words.

The term syntactic relation is used here to refer to the immediate relations between word forms, word group, and clauses in a sentence. The grammatical means of marking syntactic relations in Lithuanian are endings and, less commonly, inflectional suffixes, often supplemented by structural words, viz. prepositions, conjunctions, and particles.

Word order is of secondary importance as a means of expressing grammatical relationships in Lithuanian. For instance, it signals the syntactic function of the adjective in phrases like *Gražios gėlės* (attribute; cf. *Gražios gėlės auga sode* 'Beautiful flowers grow in the garden') and *Gėlės gražios* (predicative), meaning *Gėlės yra gražios* 'The flowers are beautiful.' Within a sentence, intonation binds word forms into groups and serves to reinforce their syntactic relations (immediately related word forms usually form an intonational unit); it also signals communicative sentence types. Three principal types of syntactic relations are distinguished: interdependence, subordination and coordination.

Furthermore, the Lithuanian language can be defined by a free order of words, which means that the same idea can be expressed in different ways (even though there are structures that can only be used for the purposes of stylistics).

Then there are elliptical sentences, where missing words can only be guessed from the context. Besides, sentences can be very long and have a complex structure, which also leads to some difficulties in automated processing.

Lithuanian language can be defined by a free order of words.

There are plenty of polysemantic words and therefore a learner of the Lithuanian language might find it difficult to recognise the meaning and form of one word or another.

Many grammatical forms, like Lithuanian nominal words, have the grammatical categories of case, gender and number. Furthermore, adjectives can be pronominal or not and have degrees of inflection as well as being neuter.

The verb of the Lithuanian language is even more complicated as it has inflective and non-inflective forms that possess the qualities of both nominal words and verbs, i. e., they can be inflected by type, time, number, gender and case.

3.3 RECENT DEVELOPMENTS

Even though its written tradition stems from the 16th century, the Lithuanian language became a standard language as late as in the beginning of the 20th century, when its first normative grammar was written and a Lithuanian thesaurus was launched, with its final volume (20) published in 2002.

Just as it started to gain a foothold, the common Lithuanian language was faced with a lot of challenges. Since the very beginning of writing, it had been heavily affected by Slavic languages. The Soviet era saw the learning and usage of the Russian language being promoted, while the role of the Lithuanian language in certain areas, like science and governmental administration, was quite limited. The current senior and middle-aged generation of the public grew up naturally surrounded by the Russian language and culture. As words of a foreign origin reflect both the language and the social life, a lot of borrowed words, and terms, and administrative language structures and so on found their way into the Lithuanian language at that time. The influence of the Russian language is still very strong when one considers some of the peripheral species of the language: slang, substandard vocabulary, etc.

The political, economic, social and cultural processes of the most recent fifteen years in Lithuania have led to particularly drastic changes in the Lithuanian vocabulary. Over 700 new foreign roots were registered in press in 1993-1997 [6]. Mostly these were words borrowed from or through the English language. This was caused by the rapid expansion of information technologies that started after the restoration of independence, as well as by the emergence of new cultural, social and economic options. The current Lithuanian corpus indicates that better than 10,000 new words were added to the vocabulary of the Lithuanian language over the period between 1991 and 1996 alone. It is therefore likely that these processes are currently progressing at an even faster pace. Ever since 1990, the information space has been flushed with pop culture in the form of television series, shows, music, etc., mostly of American export. Even though foreign films and television series or shows come with translation into Lithuanian, such a cultural shift has had significant influence on the Lithuanian language and culture. The English language is seen as the most important mediator on the global tangible and intellectual market and hence its role in the economic, social and cultural life of Lithuania is growing as well, and along with it is the motivation to learn this language and use it in specialty studies, work and intellectual work. Presently the English language is more acceptable and prestigious to the younger generation as a *lingua franca*, which connects both with cultural integration and studies, career opportunities, and so on. So far, there have not been any sufficient studies, but it is probable that most of the words and entire structures that have been borrowed from the English language are used by young people, and by those who belong to certain subcultures in particular.

In Lithuania the English language affects the same areas of usage of the language as everywhere else. English probably has the broadest use when it comes to addressing a young audience. For instance, 75% of movie trailers are presented in English or in mixed languages. The situation in other areas is better. For instance, Lithuanian television channels broadcast more than 60% of commercials in Lithuanian. Mixed commercials usually only feature English product names [6].

The development of scientific language is a major concern. To attain international recognition and dissemination, some scientific fields publish very little of their research results in the Lithuanian language in Lithuania. The internationalism of science is greatly encouraged in spite of there being fears that this impoverishes the scientific terminology of the Lithuanian language and causes the Lithuanian language to be pushed out of

specific areas of usage, leading to reduced motivation to brush up the specialty language skills at the university level.

When discussing the possible future development and outlook of the Lithuanian language, it would be a proper argument that the language anchoring itself within the information society would be a modern and functional tool of communication.

3.4 OFFICIAL LANGUAGE PROTECTION IN LITHUANIA

The Lithuanian language enjoys the status of state language, which is fixed in the Constitution of the Republic of Lithuania. The enforcement of this status, i. e., the usage of the state language in public life, as well as its protection and control and the liability for its violations, is regulated by the Law on the State language (1995). Its enforcement rests with the State Commission of the Lithuanian Language, which files motions with regard to legal regulation and deliberates matters of normalisation and the usage of the language.

State and municipal institutions, establishments, enterprises and organizations must conduct correspondence with each other in the state language. Heads of communications, transportation, health and social security, police and law-enforcement services, trade and other establishments providing services to the population must ensure that the population were provided with services in the state language.

The Lithuanian language enjoys the status of state language, which is fixed in the Constitution of the Republic of Lithuania.

There are 34 national and local television channels and 52 radio stations that broadcast in Lithuanian [7]. Audio-visual programmes and motion pictures publicly shown in Lithuania must be translated into the state language or shown with subtitles in Lithuanian. So, translation is a very relevant and important area considering that translated books account for nearly a third of all books that are published in Lithuanian (according to 2010 data, of the 2,962 books published in the Lithuanian language, 982 were translated books [8]). By the way, the mass media of Lithuania (the press, television, radio, etc.), all publishers of books and other publications must observe the norms of the correct Lithuanian language. Compliance with the requirements for usage and correctness of the state language is monitored by the State Language Inspectorate.

The principal guidelines of the Lithuanian language policy are the following:

- The Lithuanian language is an instrument of communication of the state and its people in every area of social life and one of the critical characteristics of the sovereignty and integrity of the state.
- The policy of the Lithuanian language must satisfy the need for social, national and cultural unity of the public, including Lithuanian nationals residing abroad.
- The policy of the Lithuanian language must stand in harmony with the European Union policy on languages that encourages the preservation of a linguistic variety in multicultural Europe, which is seen as one of the greatest values in Europe.
- The policy of the Lithuanian language must nurture a conscious and creative attitude towards the usage of the Lithuanian language as well as an understanding of the value and distinction of the Lithuanian language within the society.
- The dialects of the Lithuanian language are the wealth of the Lithuanian language and culture and therefore they should be protected and maintained as such.

- The policy of the state language must extend to systems of communication designed for people with special needs.
- The expansion of the Lithuanian language prioritises digital linguistic systems and resources available on the Internet. The Lithuanian language should be developed as a constituent part of multilingual terminologies and usage resources of the EU. Automated translation to and from the Lithuanian language is an important element of using the language in the EU space.

The policy of the Lithuanian language must stand in harmony with the European Union policy on languages that encourages the preservation of a linguistic variety in multicultural Europe, which is seen as one of the greatest values in Europe.

The amount of political effort to protect and support the Lithuanian language is also enough: the usage of the state language and the protection of its status is regulated by the Law on the State language (1995), the control of its usage and correctness is governed by the Law on the State language Inspectorate (2001), and the expansion of terminological resources is the subject of the Law on Term Bank (2003). Besides, as part of cultural identity, the language is also covered by legislation on protection of cultural and ethnic heritage.

After Lithuania joined the European Union, the Lithuanian language entered a new phase of development: its newly acquired status of an official EU language ensured that the Lithuanian language will be used and disseminated across the multilingual space of the EU, development of language resources required for the language to function properly in the multilingual environment picked up pace, and so on.

The usage and promotion of the Lithuanian language has the support of governmental bodies and public organisations, like the State Lithuanian Language Commission, the State Language Inspectorate, the Lithuanian Language Fellowship, etc. There are a lot of government supported programmes aimed at promoting linguistic research and dissemination underway as well.

The Institute of the Lithuanian Language with its Language Museum, which is open to the public, is one of the important centres for research and the dissemination of the Lithuanian language. Various aspects of the Lithuanian language are the subjects of different university research programmes, ranging from traditional empiric research to language technologies.

To top it off, efforts are being made to involve the public, school students, and young people in particular, into initiatives on the usage and dissemination of the Lithuanian language. These initiatives are staged by governmental bodies, educational establishments and businesses. Such initiatives include the National Dictation and various competitions, like calligraphy contests, a contest titled "I Create the Language" to replace foreign words with Lithuanian counterparts, and a contest for usage of the correct Lithuanian language in information technologies that goes under the title of "Clean Language – Clear Head" and is aimed at encouraging school students to use Lithuanian letters and the correct Lithuanian language in the digital space. Furthermore, there are annual contests to elect the most beautiful corporate name, the most beautiful Lithuanian word, etc.

3.5 LANGUAGE IN EDUCATION

According to the Law on the State Language, the state guarantees education of all degrees in Lithuanian, the mother-tongue. The state final exam of the Lithuanian language is the only obligatory exam in all schools of secondary education that teach in Lithuanian. The very same kind of examination will be taken by pupils in ethnic minority schools as well. The country has schools for the children of ethnic minorities that teach in languages

other than the state language, like Russian, Polish, Belarusian, or in mixed languages [9]. Furthermore, there are schools that teach in English, German, French and Hebraic.

Despite the apparently significant degree of attention to the teaching of the Lithuanian language, achievements in learning the language deteriorate across all educational concentres every year. According to a 2007 national survey, only 39% of eight-graders managed to attain the general level of skill in the Lithuanian language [9].

Achievements in mastering the Lithuanian language are very scattered and, compared to the 2006 results, are much worse as well [10]. An OECD PISA research study has shown the average results of reading skills among fifteen-year-old Lithuanian pupils to be rather below the international average as well [11]. By the way, there has been a significant decline in reading skill achievements among boys. The results of primary education are somewhat better: as many as 99% of Lithuanian fourth-graders attained the threshold level in the 2006 international reading skills survey PIRLS (Progress in International Reading Literacy Study); however, only 5% made it to the top level compared to the international average of 9% [12].

Quite a few pupils believe they have no aptitude for the Lithuanian language: the results of the 2008 national pupils' achievement study are that 52% of tenth-graders said they did not consider themselves apt at learning the Lithuanian language and only one-half of all pupils said they enjoyed their Lithuanian language classes [11].

However, the shifting attitude towards the teaching of the Lithuanian language and literature is expected to bring some break-through. The Strategy for Teaching the Lithuanian Language in Comprehensive Schools in 2010–2014 [13] was approved in 2010; the strategy provides that the school must develop the kind of Lithuanian identity that gives the 21st century human being everything they need to grow free, confident, and to think critically, creatively and responsibly. The strategy perceives the Lithuanian language as a modern tool of communication in the society that is and should be used in all areas and settings of life. The mutation of the language and its customisation to the needs of the modern information society, as well as its openness and ability to replenish itself, is considered to be the precondition for its survival.

> Lithuanian language must be perceived as a modern tool of communication in society that is and should be used in all areas and environments of life.

A lot of attention is being paid to the teaching and dissemination of the Lithuanian language in Lithuania and abroad. With the extent of migration growing, attempts are being made to provide conditions for children who are returning from abroad to learn the Lithuanian language and to continue their studies in Lithuania, while those who are leaving are free to choose between self- or long-distance education [13]. However, there is a consensus that without there being any modern long-distance education systems a part of emigrant children lose their ties with the Lithuanian language and culture. There are efforts to expand the teaching of the Lithuanian language in ethnic minority schools, thus offering speakers of other languages the possibility of integration into the Lithuanian labour market, of obtaining information and communicating in social space.

3.6 INTERNATIONAL ASPECTS

Although when asked about what puts Lithuanian on the world map, a substantial part of the Lithuanians will say, without much consideration, that it is basketball (in fact, the Lithuanian names of famous Lithuanian basketball players and coaches is a proper exercise in

Lithuanian pronunciation and a chance to get to know the Lithuanian language for basketball fans all over the world), it is the Lithuanian language itself that is traditionally considered one of the core national values and objects of presenting Lithuania to the world.
To this day, to some Lithuanians, their understanding of their nationality is based on their linguistic identity.

> The Lithuanian language, its dialects, and its folklore, has been described and studied by scholars from various countries with the geography of such studies ranging from the neighbouring countries to Japan, Australia, and the United States.

Lithuania has been luring the famous artists Adam Mickiewicz, Prosper Mérimée, Johann Wolfgang von Goethe and others both with the archaic character of its language and the traditions of pagan religions. The pagan beliefs and customs that probably persevered in Lithuania the longest in Europe have always greatly interested mythologists and it is no accident that Marija Gimbutas, one of Europe's most prominent researchers of prehistory originated from Lithuania; mythology fell into the field of studies of Algirdas Julien Greimas, the founder of the Parisian school of semiotics as well.

> The tradition of studying the Lithuanian language at major centres of Indo-European studies lives to this day.

The Lithuanian language, its dialects, and its folklore, has been described and studied by scholars from various countries, the geography of such studies ranging from the neighbouring countries to Japan, Australia, and the United States. The tradition of studying the Lithuanian language at major centres of Indo-European studies lives to this day. Foreign universities have some 10 lituanistic centres that currently offer independent curricula of lituanistics and baltistics. On the global scale, there are a total of over 30 lituanistics centres of varying sizes, most of them clustered in Europe that study or teach the Lithuanian and the Baltic languages and culture.
The Ministry of Education and Science supports and promotes centres of lituanistics. The scholarship of Kazimieras Būga is granted to foreigners who study the Lithuanian language in foreign colleges and universities every year.

3.7 LITHUANIAN ON THE INTERNET

Figures for 2011 show that almost 64.1% of Lithuanian people use the Internet in their home or workplace, and the percentage of users amongst those 16-24 years of age was even higher – 96.6% [14]. However, no accurate figures are available as to which language these people use to access the Internet. The IRT infrastructure is developed particularly well: Lithuania is leading by penetration of light-diode broadband network (23%) in Europe, places first in the world in terms of mobile communications subscribers per population 100 and second by the speed of the Internet connection; on top of that, the country has the most dense network of wireless Internet access points (875) in Europe [15].

> Lithuania is leading by penetration of light-diode broadband network in Europe, places first in the world in terms of mobile communications subscribers per population and second by the speed of the Internet connection.

A total of 126,000 domains ending with *.lt* were registered in 2010. Of those some 1,400 are registered using specific Lithuanian characters (*ė*, *š*, etc.). Most of them display Lithuanian content. One can find websites offering Lithuanian content in domains ending with *.eu*, *.org*, *.com* as well.

A total of 126,000 domains ending with *.lt* were registered in 2010. Of those some 1,400 are registered using specific Lithuanian characters.

The abundance of public services available on the Internet in the Lithuanian language is growing. The level of transfer of the main public services to electronic media in Lithuania stood at 64% in 2005. The rate at which business-driven services are being transferred on to the Internet is even higher, their level standing at 76% in 2005 compared to 56% of public-driven services [16]. Within the framework of implementing various national programmes, more public services will be transferred to the web, expanding the volume of Lithuanian content on the Internet: digitalising and dispersing Lithuanian cultural heritage, creating conditions for the people in Lithuania to use IRT with Lithuanian links. Moreover, efforts are made to reduce digital isolation and to make technologies user-friendly and easily accessible for people with disabilities.

The popularity of news portals, the websites of Lithuania's main periodicals, some of the scientific magazines, among other things, is increasing. Of all the major projects on the dissemination of Lithuanian content, the most noteworthy are the cultural heritage portal www.epaveldas.lt, which offers the resources of libraries, museums and other heritage institutions in a virtual space, the teaching resource portal www.emokykla.lt, www.emokymas.lt, etc. There are plans to develop a portal that would host easily accessible linguistic resources and technologies designed under the programme of the Lithuanian Language in the Information Society, which has been launched recently.

Within the framework of implementing various national programmes, more public services will be transferred to the web, expanding the volume of Lithuanian content on the Internet.

The next chapter gives an introduction to language technology and its core application areas, together with an evaluation of current language technology support for Lithuanian.

4

LANGUAGE TECHNOLOGY SUPPORT FOR LITHUANIAN

Language technology is used to develop software systems designed to handle human language and is therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and, in terms of human evolution, the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include:

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

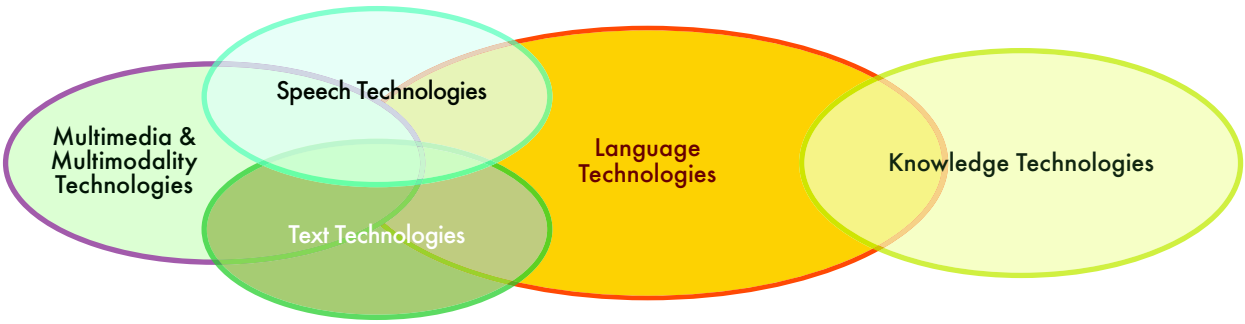
Language technology is an established area of research with an extensive set of introductory literature. Those interested in finding out more about it should see the following references: [17, 18, 19, 20, 21].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.



1: Language technology in context

2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns in the sentence); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

In the remainder of this section, we will firstly introduce the core application areas for language technology, and will follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Lithuanian in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Lithuanian language is summarised in figure 7 (p. 70) at the end of this chapter. This table lists all tools and resources that are boldfaced in the text. LT support for Lithuanian is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

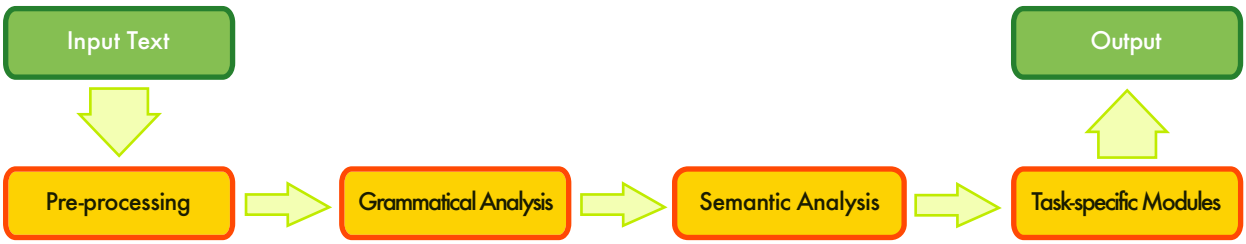
In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities in Lithuania.

4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling errors and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [22]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

This type of analysis either needs to draw on language-specific **grammars** laboriously coded into the software by experts, or on a statistical language model. In this



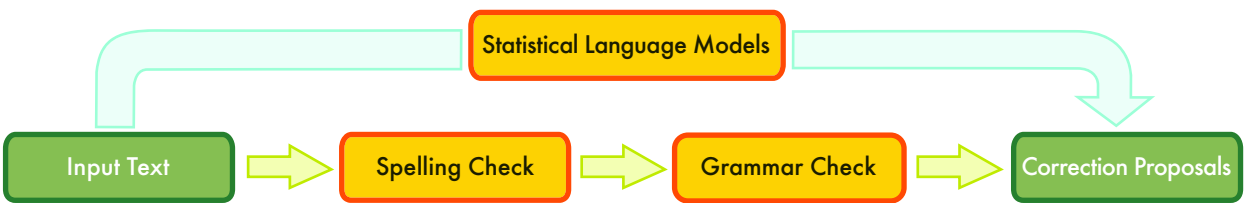
2: A typical text processing architecture

case, a model calculates the probability of a particular word as it occurs in a specific position (e. g., between the words that precede and follow it). For example: *englische Buch* is a much more probable word sequence than *Englisch Buch*. A statistical language model can be automatically created by using a large amount of (correct) language data, a **text corpus**. Most of these two approaches have been developed around data from English. Neither approach can transfer easily to Lithuanian because the language has a flexible word order and a richer inflection system (ever since 2002, scientists at Vytautas Magnus University have worked consistently with the statistical model of the Lithuanian language, which has its tools available for download [23]).

Language checking is not limited to word processors; it is also used in "authoring support systems", i. e., software environments in which manuals and other types of technical documentation for complex IT, healthcare, engineering and other products, are written. To offset customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation to use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

Language checking is not limited to word processors but also applies to authoring support systems.

Only a few Lithuanian companies offer products in this area. In 1992-1994, Fotonija, UAB developed a spellchecking program *Juodos Avys*, which has been subjected to ongoing upgrades. Automated spellchecking is based on an algorithm that helps avoiding sponta-



3: Language checking (statistical; rule-based)

neous correction of a name or title for something else. Erroneous words are recognised and their correct versions are prompted. The program also prompts to correct missing specific Lithuanian characters like *š*, *ž*, *ū*, *ė*. The spellchecker has an integrated hyphenation tool for Lithuanian syllables.

In 2001, a spelling checker for Lithuanian was developed by Tilde IT, UAB. Tilde IT is continuously improving its spelling checker and is developing a new grammar checker that will analyse the sentence structure, identify missing and unnecessary commas or other punctuation marks, and check the syntax and lexical errors. The grammar checker will operate not only in Microsoft Office, but also in Open Office and on the Internet. It will be easily installable in other programs that use language items (e. g., enterprise resource management system, e. business solutions, etc.). Users will be able to try out the new grammar checker in 2012.

Beside spell checkers and authoring support, language checking is also important in the field of computer assisted language learning. Language checking applications also automatically correct search engine queries, as found in Google's *Did you mean...* prompts.

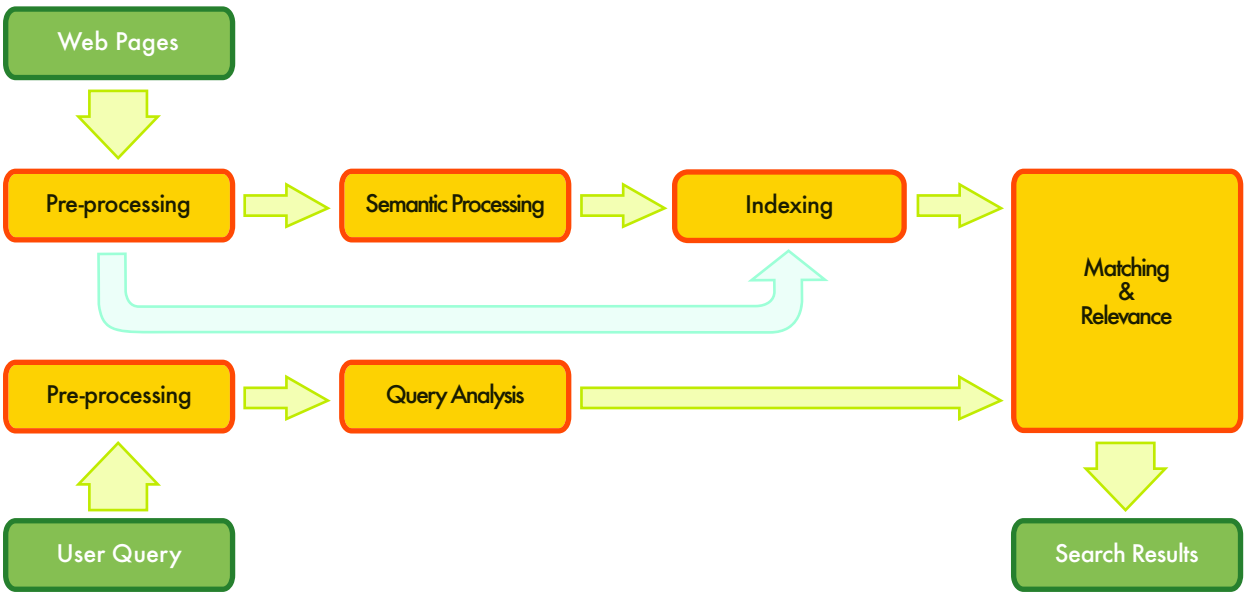
4.2.2 Web Search

Searching the web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which was launched in 1998, now handles about 80% of all search queries [24]. Since 2004, the verb *googeln* has even had an entry in the Duden dictionary. The Google search interface and results page display has not significantly changed since the first version. However, in the current version, Google offers spelling correction for misspelled words and incorporates basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [25]. The Google success story shows that a large volume of data and efficient indexing techniques can deliver satisfactory results using a statistical approach to language processing.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate text interpretation. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources (e. g., WordNet for English or GermaNet for German) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *Atomkraft* [atomic energy], *Kernenergie* [atomic power] and *Nuklearenergie* [nuclear energy], or even more loosely related terms.

The next generation of search engines will have to integrate much more sophisticated language technology.

The next generation of search engines will have to integrate much more sophisticated language technology, especially to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies. For the expression *last five years*, the system needs to determine the relevant range of years, taking into account the present year. The query then needs to be matched against a huge amount of unstructured data to find the pieces of information that are relevant to the user's request. This process is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular



4: Web search

string of words in a document represents that a company name, using a process called named entity recognition.

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the user's target language.

Now that data can be found in non-textual formats, increasingly often there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

In terms of the Lithuanian language, such technology is only in its early stage of development. Research and projects related to this area are conducted by Vytautas Magnus University (project *Semantic Engine for Information Management* is supported by EU Structural Funds within the programme of economic growth activities), the Institute of Mathematics and Informatics of the University of Vilnius, Kaunas University of Technology. Several IT companies, such as Sintagma, UAB that has designed the document handling system *Avilys* are taking their first steps in the field of designing ontologies, knowledge and document handling. Tilde IT, UAB has been developing semantic system projects since 2008 and is currently involved in the linguistic semantic network project SemTi as well as the international project SOLIM (*Spatial Ontology Language for Multimedia Information Modeling*). Implementation of ontologies for some areas, like in library science is only now gaining speed.

So far, any efforts in this area have been fragmented and any greater breakthrough is expected from the 2009-2013 programme of the Lithuanian Language in Information Society as initiated by the government of the Republic of Lithuania, which programme envisages the development of tools designed to accommodate the providing of the syntactic/semantic analysis service,

analysing Lithuanian website content, using it as a filter in searches, etc.

4.2.3 Speech Interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of using a graphical display, keyboard and mouse. Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces with car navigation systems and the use of spoken language as an alternative to the graphical or touchscreen interfaces in smartphones.

Speech interaction technology comprises four technologies:

1. Automatic **speech recognition** (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. **Speech synthesis** (text-to-speech or TTS) transforms the system's reply into sounds for the user.

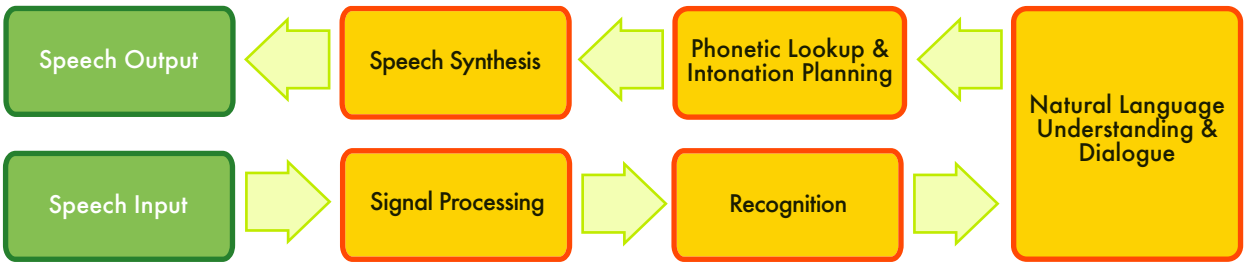
One of the major challenges for ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly — prompted by a *How may I help you?* greeting — tend to be automated and are better received by users.

Speech interaction is the basis for interfaces that allow a user to interact with spoken language.

Companies tend to use utterances pre-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Through optimisation, today's TTS systems are getting better at producing natural-sounding dynamic utterances.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.

Research into speech technology in Lithuania has been carried out at Kaunas University of Technology since



5: Speech-based dialogue system

1980, with the Institute of Mathematics and Informatics of the University of Vilnius working in this field for many years as well; research in this area has been launched at Vytautas Magnus University too.

In the Speech Research Laboratory (Kaunas University of Technology), ASR research was first started in 1980. The Laboratory has developed a corpus of commands and digital sequences. Lithuanian computer dialogs are being developed and a Lithuanian speech corpus, LT-DIGITS, has been compiled and is now undergoing improvements. It contains continuous digit sequences and Lithuanian computer control words. Characteristics of Lithuanian speech are being researched at the Institute of Mathematics and Informatics of the University of Vilnius, which has compiled a corpus of broadcast news called LRNO. A universal corpus of spoken Lithuanian has been compiled by Vytautas Magnus University, which also compiles corpora of a smaller extent, like those designed to study the language, such as SACODEYL, a corpus of spoken teenager language, etc., and research automated segmentation of spoken Lithuanian; development of automated transcription of spoken Lithuanian language is underway as well.

Even though speech recognition research continues with the aim of improving the quality of this service, ASR software already has found successful application in law enforcement, telephony, education, transport, the Internet, etc.

Investigations in the text-to-speech synthesis and its application for the blind and partially sighted were carried out at Vilnius University. The most important components of the Lithuanian text-to-speech synthesiser *Aistis* are: a) the automatic division of Lithuanian words into syllables; b) the automatic stressing of words in a Lithuanian text; c) the automatic transcription of Lithuanian text; d) a phonetic units base; e) Lithuanian text-to-speech synthesis quality evaluation. It is easily accessible and developed with a focus on applications for user groups with specific demands, such as physically handicapped people and the elderly. The synthesiser MBROLA is easily accessible on the Internet and runs on a base of phonetic units developed at the Vilnius University by Aleksas Girdenis and Pijus Kasparaitis.

The components of *Aistis* were used in the development of the Lithuanian language synthesiser WinTalker Voice with two voices, *Gintaras* and *Aistis2*, which was produced by the Czech company Rosasoft on order from the Lithuanian Association of the Blind and Partially Sighted (there are a total of some 7,000 people with special needs in Lithuania and as of March 1, 2010 there were 258 blind persons using computers across Lithuania). Another free TTS synthesiser has been developed by Etalinkas, UAB. The synthesiser runs under Windows and Linux OS.

For the purposes of speech recognition, Vytautas Magnus University has developed corpus-based statistical

models of the Lithuanian language, a prototype of a continuous speech recognition tool that encompasses more than 1 million word forms, an automated accentuation programme with unambiguous homographs, which is available on the Internet, and models of durations of the Lithuanian language sounds.

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

4.2.4 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet **machine translation** (MT) still cannot deliver on its initial promise of providing across-the-board automated translation.

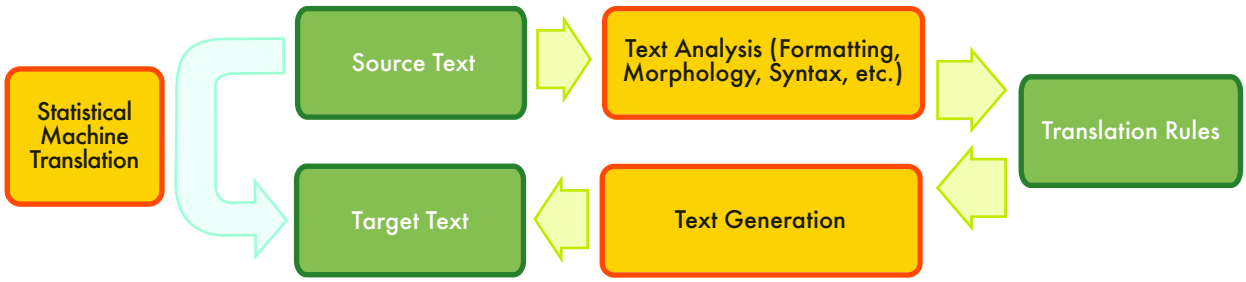
> At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:

- *I was happy to read a book.*
- *Aš buvau laimingas:*
  1. *perskaitęs knygą.*
  2. *skaitydamas knygą.*
  3. *galėdamas perskaityti knygą.*

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process.

In the late 1980s, when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, **parallel corpora**, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical out-



6: Machine translation (statistical; rule-based)

put. Data-driven MT is advantageous because less human effort is required, and it can also cover special particularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.

The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be far from perfect. A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Machine Translation is particularly challenging for the Lithuanian language.

Machine translation is particularly challenging for the Lithuanian language. Free word order and verb constructions pose problems for analysis, and inflection is a challenge for generating words with proper gender and case markings.

With languages that have a smaller user base, such as the Baltic languages MT research tools, as well as language technologies in general, are less developed. There have been several activities related to MT in Lithuania. There are three translation tools currently available on the web: WIFTA project [26], Google translator and *Vertimo Vedlys* [27]. The first system was developed in 2008 in cooperation with the Russian company ProMT and is based on rule-based technology. It performs MT that takes into account the morphological, syntactic and semantic properties of texts. This project was finished successfully. The English-Lithuanian MT system has been in operation at http://vertimas.vdu.lt since 2008. It generates 127 million hits and attracts approximately 1 million unique users yearly. Registered users have access to computer and business thesauruses.

The Corpus of the Modern Lithuanian, which has been compiled at Vytautas Magnus University and contains ap. 140 million words, is open to the public over the Internet [28]. Furthermore, a parallel corpus of the Lithuanian language and other languages (English, German, Czech) is being compiled, other special corpora have been developed as well (for instance, the University of Vilnius has compiled a Corpus of Academic Lithuanian *CorALit* [29]). However, the current corpus of the Lithuanian language cannot accommodate the needs of the development of modern technologies (like search for information, automated translation and other systems) of the Lithuanian language. The existing and future corpora require software customised to the Lithua-

nian language that would allow making better use of the available language resources and the digital descriptions that they produce. One of the prerequisites for classical SMT systems is the availability of a large parallel corpus which computer then uses in the training process. The lack of a large parallel corpus is the main reason why experiments with SMT in the Baltic countries have only started recently. The Google MT software uses the SMT approach and provides MT for about 30 languages, including the Lithuanian language.

Vertimo Vedlys is an experimental machine translation tool developed by the Institute of the Lithuanian Language in association with Tilde IT, UAB. The trial version provides translation from Lithuanian to English. The automated translation tool analyses the structure of sentences and automatically prompts a translation of a sentence, its part or individual words. It is based on SMT and uses Giza ++ and Moses engines. Tilde IT is developing a Lithuanian-English machine translation system by integrating statistical and rule-based MT methods as well as by applying an innovative processing of multiword expressions. The accuracy of translation is 30%. The system is constantly improved. Tilde IT's MT tool has been used not only for full text translation but also in cross-lingual search applications.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that already have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support a few languages that can be translated to Lithuanian and vice – versa. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns help to compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 7 (p. 28), which was prepared during the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [30]. A human translator would normally achieve a score of around 80 points.

The best results highlighted (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese and Finnish).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics.

Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Language technology applications often provide significant service functionalities behind the scenes of larger software systems.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify specific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Text summarisation and **text generation** are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create a summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text.

This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation.

For Lithuanian, the situation in all these research areas is much less developed than it is with the English language: some experiments have been performed on Lithuanian text summarization, automatic identification of educational and scientific terminology (in Vytautas Magnus University), etc.

Lithuanian has been included in the international projects that the Latvia-based CIA Tilde carries out. Prototypes of Lithuanian information retrieval engines were developed as part of FP5 project CLARITY: A proposal for cross-language information retrieval and organisation of text and audio documents. The CLARITY cross-language information retrieval system was developed for the following language pairs: English-Latvian, Latvian-English, German-Latvian, Latvian-German, Russian-Latvian, Latvian-Russian, Lithuanian English, English-Lithuanian, German-Lithuanian, Lithuanian-German, Lithuanian-Russian and Russian German. With respect to the Baltic languages, the results for document retrieval using direct query translation indicate that the average precision can reach a level of more than 70% compared to monolingual retrieval.

4.4 EDUCATIONAL PROGRAMMES

Language technology is a very interdisciplinary field that involves the combined expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists among others. As a result, it has not acquired a clear, independent existence in the Lithuanian faculty system. Some universities have established separate centres, e. g., Centre for Computational Linguistics (CL) in Vytautas Magnus University, or laboratories, e. g., the Speech Research Laboratory in Kaunas University of Technology. Currently, there is only one curriculum for bachelor studies of Computational Linguistics in the Faculty of Humanities in Kaunas University of Technologies. The program was launched in 2003, and had had 73 graduates by 2010. The steadily rising demand of qualified personnel specialised in the field of language technology cannot be met by the comparably low number of graduates.

In Vilnius University and Vytautas Magnus University some CL- and LT-related courses are taught as part of other studies. As of 2011, the Kaunas Humanities Faculty of the University of Vilnius offers master courses of audiovisual translation. Vytautas Magnus University has had a master programme of digital linguistics since 2006 (accredited until 2015). As yet, no university offers consistent studies of every level, and therefore the field of linguistic technologies employs scientists who have completed linguistic or informatics studies (some of them both).

A scientific research base is being developed and resources compiled at the Institute of Mathematics and Informatics and the Institute of the Lithuanian Language; the latter establishment founded a Digital Language Resources Laboratory in 2010.

4.5 NATIONAL PROJECTS AND INITIATIVES

A higher degree of expansion of the information society, a lively interest in language technologies, and the development of resources in Lithuania only started a few decades ago. Since the Lithuanian language has a rather limited number of users, the commercial market for language technologies is not very big and besides, Lithuania has no such modern technological giants as BMW or NOKIA, and there are merely a few commercial businesses operating on the LT field.

Most of the initiative and commitment with regard to the functioning of the Lithuanian language within the information society and LT development originates on the national level. The year 2000 saw the launch of the first national programme of the Lithuanian Language in the Information Society for the period of 2000–2006. The programme was coordinated by the State Commission of the Lithuanian Language and dealt with localisation, resource and tool creation, documentation and some other activities:

- ASR expansion, involving research of the traits of the Lithuanian speech, development of a prototype recognition tool for separate spoken words, improvement of the spoken corpus of Lithuanian broadcast news LRNO, studies of Lithuanian computerised voice dialogs, improvement of the quality of TTS synthesis, development of pilot samples of applications of Lithuanian speech technologies, automated segmentation of the Lithuanian spoken language and development of automated transcription of the Lithuanian speech.
- Standardisation of the Lithuanian language subjects in IT (e. g., development of the computer font *Palemonas*, localisation, etc).
- Translation and development of the necessary resources and tools, involving the development of a

computerised system for literal translation of specialised texts, development of a parallel corpus of the Lithuanian and the Czech languages, as later updated, development of a tool of morphological analysis and generation.

- Work has been started in the field of syntactic and semantic analysis of texts in the Lithuanian language.

The Information Society Development Committee under the Ministry of Transport and Communications is responsible for the second phase of the program of the Lithuanian Language in the Information Society 2009-2013. The programme provides for the creation of an Internet portal with free access to all the available language resources and technologies, augmentation of the existing and newly created linguistic resources, improvement of the ASR and TTS technologies, new MT tools, improvement and development of semantic and syntactic analysis and search tools.

Research any resource generation in this area is promoted as well. The Research Council of Lithuania has launched the first national program "State and Nation: Heritage and Identity" that encompasses digitalization of intangible heritage (this program saw the implementation of the project called "Development of a Lituanistic Digital Resource Metadata System and its Compatibility with CLARIN"). Recently, the Research Council of Lithuania also finances the programme for the Development of National Lituanistics 2009–2015, aimed at developing and promoting lituanistic research, helping meet the priority of lituanistic research, strengthening the input of lituanistic research data into the overall expansion of nation-wide humanistic, providing a scientific base for nurturing national self-consciousness and protecting lituanistic heritage. Companies doing business on the field of language technology are few and include Tilde IT, Fotonija, Microsoft Lietuva, CID Baltic, Synergium, Sintagma, TokenMill, HLTech. Tilde IT, UAB is the clear leader in the area of language technology, with 12 years of experience of operating on the Lithuanian market. The company is giving a lot of attention to software localisation, translations of technical documentation, development of software to support the Lithuanian language. Tilde IT is one of the largest providers of localisation services in Lithuania. The company works together with international localisation and translation companies on a continuous basis.

At this time, Tilde IT is engaged in improving the quality of machine translation and the development and upgrade of spellchecking systems. The company initiates research and technological development projects aimed at developing prototype software in cooperation with the Institute of the Lithuanian Language and the Institute of Mathematics of Informatics, as well as the Philological Faculty of the University of Vilnius.

Tilde IT has been conducting semantic system projects since 2008. Since Tilde IT is involved in providing machine translation technologies for the European market, the new technology will be used as an addition to the set of techniques that improve machine translation results. Tilde IT aims to create a database of links between Lithuanian words, also known as a linguistic semantic database. A Lithuanian semantic web would greatly help marketing professionals because it would help to predict public reaction to the proposed product promotions, packaging or name. Therefore, such a thinking map can be used to create new products and the generation of new or original ideas. One word can have more than 15 synonyms, although ordinarily people only know just 5 to 6 of them. A detailed semantic web will help save the Lithuanian language's synonymic diversity.

Tilde IT has joined SOLIM (Spatial Ontology Language for Multimedia Information Modelling) project of the *Eurostars* program. The project is aimed to improve context aware information analysis to venture beyond a static world, by adding the concepts of space and

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech recognition	2	0	2	1	1	0	2
Speech synthesis	3	2	2,5	2,5	1,5	1	2
Grammatical analysis	2	1,5	2,5	2	1,5	1	2
Semantic analysis	1,3	1	1,3	1	0	0	0,3
Text generation	0	0	0	0	0	0	0
Machine translation	2	3	2,5	2,5	2	2	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	1,5	1,5	2,5	2,5	2	2,5	2,5
Speech corpora	2	1	2	2	1	1	2
Parallel corpora	2	2	1,5	1,5	2	2	4
Lexical resources	2,5	2	2,5	2	2	0,5	2,5
Grammars	0	0	0	0	0	0	0

7: State of language technology support for Lithuanian

change. The goal of the project is to extend the Web Ontology Language OWL to support effective storage and reasoning on spatial information, and to demonstrate the power of such extension for automatic processing of textual and graphical information in real proof of concept applications.

As of 1991, Fotonija, UAB has been integrating Lithuanian into computers by developing and upgrading drivers (*WinLika*, *Lika*), designing a Lithuanian font dubbed *Aistika*, the text management application *Mainukai*, the document converter, the text creation, editing and proofing application, and the spellchecker *Juodos Avys*. An important area of Fotonija's business is the development of monolingual and multilingual dictionaries, which include the international dictionary *Interleksis*, TŽŽ, the English-Lithuanian dictionary *Anglonas*, and its French counterpart, *Frankonas*.

There are other companies engaging in localisation, ontology development and other LT projects as well, including Microsoft Lietuva, CID Baltic, Synergium, Sintagma, TokenMill, HLTech,and so on.

As we have seen, previous programmes have led to the development of a number of LT tools and resources for the Lithuanian language. In the following section, the current state of LT support for Lithuanian is summarised.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 7 provides a rating for language technology support for the Lithuanian language. This rating of existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from

0 (very low) to 6 (very high) using seven criteria. The key results for Lithuanian language technology can be summed up as follows:

- Research has successfully led to the design of medium-quality software for basic text analysis, such as tools for morphological analysis and syntactic parsing. But advanced technologies that require deep linguistic processing and semantic knowledge are still in their infancy.
- The more linguistic and semantic knowledge a tool takes into account, the more gaps exist (see, e. g., information retrieval, text semantics, etc.), and more efforts for supporting deep linguistic processing are needed.
- While some specific corpora of comparably good quality exist, they are not fully developed, some of them are available only via specialised, individual access tools, or even inaccessible. A very large syntactically annotated corpus is not available.
- Many of these tools, resources and data formats do not meet industry standards and cannot be sustained effectively. A concerted programme is required to standardise data formats and APIs.
- There is a lack of parallel corpora for machine translation. Translation from Lithuanian to English works best because this language pair has the most data available.
- There is a huge gap in multimedia data.

In a number of specific areas of Lithuanian language research, we have software with limited functionality available today. Advanced tools, like treebanks, lexical semantic knowledge base or taxonomies of concepts, such as WordNet are yet to be designed for the Lithuanian language. Even though automated translation tools have recently been developed, the most advanced resources, or general applications, are only entering the phase of development [32]. Obviously, further research will probably fill in the gap of detailed semantic analysis of texts and see to it that the missing resources, such as parallel texts for machine translation, WordNet, etc. are compiled.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics),

coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 8 to 11 show that Lithuanian is falling behind the LT leaders, such as the English language, which is in the lead in almost all LT areas. When it comes to clusters, it is faced with other European languages that have fewer users and are hence not so commercially attractive, like Latvian, Slovakian, Slovenian. On the other hand, the Lithuanian language resources and technologies are developed quite unevenly, for instance, in the resource domain there are a few rather sizeable terminology databases, yet there is no WordNet or thesaurus. Moreover, no Lithuanian grammar suitable for language technologies exists. This inhibits successful development of language models that could be applied to specific language technologies.

The extremely low level of semantics research has resulted in stunted advancement in the areas of language generation, textual interpretation and analysis. For speech processing current technologies perform well enough to be successfully integrated into a number of industrial applications. With speech synthesis research and application progressing at a quicker pace, speech recognition still represents a rather more complicated field.

However, for building more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

4.8 CONCLUSIONS

In this series of white papers, we have made an important effort by assessing the language technology support for 30 European languages, and by providing a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large-scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support among the various European languages. While there is good-quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of, for example, semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

The situation of Lithuania concerning language technology support gives rise to cautious optimism. The government of the Republic of Lithuania has placed an exclusive emphasis on developing language technologies as evidenced by programmes funded by various governmental institutions and the European structural funds' financial resources for the development of LT. A scientific base for language technologies is being developed by four universities and two research institutes of Lithuania. Within the business sector, Tilde IT is the key player on the field of developing the Lithuanian LT. For standard Lithuanian, a number of technologies and

resources exist, albeit much fewer than for English. The Lithuanian language is one of the so-called non-commercial European languages and is therefore facing the IT challenges and difficulties that are typical of the development of a less widely used language. The development of the Lithuanian LT relies heavily on the experience of and assistance from other countries and international cooperation. On the other hand, developing language technologies is the most important element in the process of strengthening the functionality, recognition and learning of the Lithuanian language as well as the dissemination of the Lithuanian culture across the multilingual Europe.

From this, it is clear that more efforts need to be directed into the creation of resources for Lithuanian and into research, innovation, and development.

Our findings show that the only alternative is to make a substantial effort to create LT resources for Lithuanian, and use them to drive research, innovation and development forward. The need for large amounts of data and the extreme complexity of language technology systems makes it vital to develop a new infrastructure and a more coherent research organization to spur greater sharing and cooperation.

Finally there is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level.

The long term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders - in politics, research, business, and society - to unite their efforts. The resulting technology will help tear down the existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian **Lithuanian** Maltese Romanian

8: Speech processing: State of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian **Lithuanian** Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

9: Machine translation: State of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian **Lithuanian** Maltese Serbian

10: Text analysis: State of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian **Lithuanian** Maltese

11: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence partially funded by the European Commission [33]. The network currently consists of 54 research centres in 33 European countries. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that centers around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present white paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

LITERATŪRA REFERENCES

[1] Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] User Language Preferences Online (Vartotojo kalbos pasirinkimas internete), 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[3] Multilingualism: an Asset for Europe and a Shared Commitment (Daugiakalbystė – Europos turtas ir bendras rūpestis), 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[4] Intersectoral Mid-term Strategy on Languages and Multilingualism (Tarpsektorinė kalbų ir daugiakalbystės strategija), 2007. http://ec.europa.eu/dgs/translation/publications/studies.

[5] Size of the language industry in the EU (Kalbos industrijos mastas Europos Sąjungoje), 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[6] Loreta Vaicekauskienė. *Naujieji lietuvių kalbos svetimžodžiai (New Borowings in Lithuanian)*. Lietuvių kalbos institutas (Institute of the Lithuanian Language), 2007.

[7] Source: Lietuvos radijo ir televizijos Licencijavimo ir kontrolės skyrius (Licensing and Control Department, Radio and Television Commission of Lithuania).

[8] Source: Nacionalinės Martyno Mažvydo bibliotekos Bibliografijos ir knygotyros centras (Bibliography and Book Science Centre, Martynas Mažvydas National Library of Lithuania).

[9] Lietuvos švietimas. Tik faktai (Education in Lithuania), 2010. http://www.smm.lt/svietimo_bukle/docs/apzvalgos/Lietuvos%20svietimas%202010.pdf.

[10] Bendrosios 2008 m. nacionalinio 6 ir 10 klasių mokinių pasiekimų tyrimo išvados (The general findings of the 2008 national 6th and 10th grade student achievement test), 2008. http://www.smm.lt/svietimo_bukle/docs/tyrimai/nmp/2008%20metu%20pagrindines%20tyrimo%20isvados.pdf.

[11] Tarptautinis penkiolikmečių tyrimas. Programme for International Student Assesment OECD PISA 2009, 2010. http://www.nec.lt/failai/1810_PISA_Rezultatai.pdf.

[12] Tarptautinio skaitymo gebėjimų tyrimo ataskaita (Progress in International Reading Literacy Study), 2007. http://www.smm.lt/svietimo_bukle/docs/tyrimai/sb/PIRLS_ataskaita.pdf.

[13] Lietuvių kalbos ugdymo bendrojo lavinimo programas vykdančiose mokyklose 2010–2014 metų strategija (The Strategy for Teaching the Lithuanian Language in Comprehensive Schools in 2010–2014), 2010. http://www.smm.lt/ugdymas/docs/Lietuviu%20kalbos%20strategija%20(1).pdf.

[14] 16–74 m. amžiaus asmenys, kurie naudojosi kompiuteriu, internetu (Computer and internet users aged 16 to 74), 2011. http://db1.stat.gov.lt/statbank/selectvarval/saveselections.asp.

[15] Investuok Lietuvoje (Invest Lithuania). http://www.investlithuania.com/lt/investuok/isvystyta-infrastruktura.

[16] 2007–2013 m. Ekonomikos augimo veiksmų programa (The economy growth action programme for 2007–2013). http://www.esparama.lt/es_parama_pletra/failai/fm/teises_aktai/Stebesenos_komiteto_nutarimai/VP2-2009-05-14.pdf.

[17] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer, and Ralf Klabunde, editors. *Computerlinguistik und Sprachtechnologie: Eine Einführung (Computational Linguistics and Language Technology: An Introduction)*. Spektrum Akademischer Verlag, 2009.

[18] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice Hall, 2009.

[19] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[20] Language Technology World (LT World). http://www.lt-world.org/.

[21] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*. Cambridge University Press, 1998.

[22] Jerrold H. Zar. Candidate for a pullet surprise. *Journal of Irreproducible Results*, page 13, 1994.

[23] Statistiniai kalbos modeliavimo įrankiai (Statistical language modeling tools). http://donelaitis.vdu.lt/~airenas.

[24] Google zieht weiter davon. *Spiegel Online*, 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[25] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[26] Mašininio (automatinio) vertimo sistema (English–Lithuanian machine translation). http://vertimas.vdu.lt.

[27] Mašininio vertimo laboratorija (Machine Translation Laboratory). http://mvlab.lki.lt.

[28] Dabartinės lietuvių kalbos tekstynas (Corpus of the Contemporary Lithuanian Language). http://tekstynas.vdu.lt/tekstynas/.

[29] CorALit: Lietuvių mokslo kalbos tekstynas (CorALit: the Corpus of Academic Lithuanian). http://coralit.lt.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[31] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for europe. In *Proceedings of MT Summit XII*, 2009.

[32] Rūta Marcinkevičienė. Two decades of lithuanian hlt. In *Proceedings of 17th Nordic Conference of Computational Linguistics*, 2009.

[33] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

B

META-NET NARIAI META-NET MEMBERS

Airija	Ireland	School of Computing, Dublin City University: Josef van Genabith
Austrija	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgija	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
Bulgarija	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Čekija	Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Danija	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
JK	UK	School of Computer Science, University of Manchester: Sophia Ananiadou
		Institute for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
Estija	Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider
Graikija	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Islandija	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Ispanija	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo
Italija	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Kipras	Cyprus	Language Centre, School of Humanities: Jack Burston

Kroatija	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Latvija	Latvia	Tilde: Andrejs Vasiļjevs
		Institute of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa
Lenkija	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Department of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Lietuva	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Liuksemburgas	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Nyderlandai	Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk
		Computational Linguistics, University of Groningen: Gertjan van Noord
Norvegija	Norway	Department of Linguistic, Literary and Aesthetic Studies, University of Bergen: Koenraad De Smedt
		Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Portugalija	Portugal	University of Lisbon: António Branco, Amália Mendes
		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Prancūzija	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur and Institute for Multilingual and Multimedia Information: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Rumunija	Romania	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Serbija	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Slovakija	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Slovėnija	Slovenia	Jožef Stefan Institute: Marko Grobelnik
Suomija	Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela

		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Švedija	Sweden	Department of Swedish, University of Gothenburg: Lars Borin
Šveicarija	Switzerland	Idiap Research Institute: Hervé Bourlard
Vengrija	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh, Gábor Olaszy
Vokietija	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Department of Computational Linguistics, Saarland University: Manfred Pinkal

Apie šimtą kalbos technologijų ekspertų – META-NET tinkle dalyvaujančių šalių ir kalbų atstovų – diskutavo ir apibendrino Baltųjų knygų serijos rezultatus META-NET susitikime Berlyne (2011 m. spalio 21–22 d.) – About 100 language technology experts representing the countries and languages covered by META-NET discussed and finalised the key results and messages of the white paper series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

META-NET BALTŲJŲ KNYGŲ SERIJA

THE META-NET WHITE PAPER SERIES

Airių	Irish	Gaeilge
Anglų	English	English
Baskų	Basque	euskara
Bulgarų	Bulgarian	български
Čekų	Czech	čeština
Danų	Danish	dansk
Estų	Estonian	eesti
Galisų	Galician	galego
Graikų	Greek	ελληνικά
Islandų	Icelandic	íslenska
Ispanų	Spanish	español
Italų	Italian	italiano
Katalonų	Catalan	català
Kroatų	Croatian	hrvatski
Latvių	Latvian	latviešu valoda
Lenkų	Polish	polski
Lietuvių	Lithuanian	lietuvių kalba
Maltiečių	Maltese	Malti
Norvegų Bokmål	Norwegian Bokmål	bokmål
Norvegų Nynorsk	Norwegian Nynorsk	nynorsk
Olandų	Dutch	Nederlands
Portugalų	Portuguese	português
Prancūzų	French	français
Rumunų	Romanian	română
Serbų	Serbian	српски
Slovakų	Slovak	slovenčina
Slovėnų	Slovene	slovenščina
Suomių	Finnish	suomi
Švedų	Swedish	svenska
Vengrų	Hungarian	magyar
Vokiečių	German	Deutsch