META NET

White Paper Series Balto grāmatu sērija

THE LATVIAN LANGUAGE IN THE DIGITAL AGE

LATVIEŠU VALODA DIGITĀLAJĀ LAIKMETĀ

Inguna Skadiņa
Andrejs Veisbergs
Andrejs Vasiļjevs
Tatjana Gornostaja
Iveta Keiša
Alda Rudzīte

Springer

White Paper Series Balto grāmatu sērija

THE LATVIAN LANGUAGE IN THE DIGITAL AGE

LATVIEŠU VALODA DIGITĀLAJĀ LAIKMETĀ

Inguna Skadiņa Tilde
Andrejs Veisbergs Latvijas Universitāte
Andrejs Vasiļjevs Tilde
Tatjana Gornostaja Tilde
Iveta Keiša Tilde
Alda Rudzīte Tilde

Georg Rehm, Hans Uszkoreit
(editors, redaktori)

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 978-3-642-30875-8 ISBN 978-3-642-30876-5 (eBook)
DOI 10.1007/978-3-642-30876-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012947861

PRIEKŠVĀRDS PREFACE

Šī baltā grāmata ir daļa no dokumentu sērijas, kurā apkopota informācija par valodu tehnoloģijām un to iespējām. Tā ir paredzēta pedagogiem, žurnālistiem, politiķiem, valodniekiem un citiem sabiedrības locekļiem. Valodu tehnoloģiju pieejamība un lietojums dažādās Eiropas valodās atšķiras. Tādējādi katrai valodai nepieciešamas atšķirīgas darbības, lai tālāk izpētītu un attīstītu valodu tehnoloģijas. Tās ir atkarīgas no daudziem faktoriem, piemēram, konkrētās valodas sarežģītības un tās lietotāju skaita.

Šajās balto grāmatu publikācijās (91. lpp.) veikta pašreizējo valodas resursu un tehnoloģiju analīze. Tās vadītājs bija META-NET — Eiropas Komisijas finansētais izcilības tīkls. Šajā analīzē galvenā uzmanība tika pievērsta 23 Eiropas oficiālajām valodām, kā arī citām nozīmīgām Eiropas valstu un reģionālajām valodām. Analīzes rezultāti liecina, ka visu valodu pētniecībā ir daudz svarīgu izaicinājumu un problēmu. Lai turpmākajai pētniecībai būtu maksimāla atdeve un tiktu samazināti potenciālie riski, nepieciešama detalizēta un lietpratīga analīze, kā arī pašreizējās situācijas novērtējums. Tīklā META-NET ietilpst 54 pētniecības centri 33 valstīs [1] (87. lpp.). Tie sadarbojas ar pārstāvjiem no privātajiem uzņēmumiem, valsts aģentūrām, rūpniecības nozarēm, pētniecības iestādēm, programmatūras izstrādātājiem, tehnoloģiju nodrošinātājiem un Eiropas universitātēm. Visi šī tīkla dalībnieki strādā pie kopīga tehnoloģiju redzējuma. Tiek izstrādāta stratēģija, kā līdz 2020. gadam risināt visas ar pētniecību saistītās problēmas, izmantojot valodu tehnoloģiju lietojumprogrammas.

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

The availability and use of language technology in Europe varies among languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The 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. 91). 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 many significant research gaps for each language. A more detailed expert analysis and assessment of the current situation will help maximise the impact of additional research and minimise any risks.

META-NET consists of 54 research centres from 33 countries [1] (p. 87) that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers, and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

Šī dokumenta autori pateicas vācu valodas baltās grāmatas autoriem par atļauju atkārtoti izmantot daļu sava dokumenta materiālu, kas neskar konkrēto valodu [2].

Autori pateicas Daigai Deksnei, Kārlim Gobam un Raivim Skadiņam par vērtīgajiem ierosinājumiem, *Tildes* Lokalizācijas un dokumentācijas daļas tulkotājiem, īpaši Elitai Kalniņai, par sākotnējo dokumenta tulkojumu, Aivaram Bērziņam, Ievai Dātavai, Evitai Korņējevai, Indrai Sāmītei, Katrīnai Baltmanei un Lindai Staužai par neizsīkstošu palīdzību dokumenta galīgās versijas sagatavošanā.

Šīs baltās grāmatas sagatavošanu finansiāli atbalstīja Eiropas Komisijas Septītā pamatprogramma un IKT politikas atbalsta programma saskaņā ar līgumiem T4ME (dotācijas nolīgums 249 119), CESAR (dotācijas nolīgums 271 022), METANET4U (dotācijas nolīgums 270 893) un META-NORD (dotācijas nolīgums 270 899).

The authors of this document are grateful to the authors of the white paper on German for permission to re-use selected language-independent materials from their document [2].

The authors would like to thank Daiga Deksne, Kārlis Goba and Raivis Skadiņš for their valuable comments and contributions, the translators of Tilde's Localization and Documentation department, especially Elita Kalniņa for initial translation, and Aivars Bērziņš, Ieva Dātava, Evita Korņējeva, Indra Sāmīte, Katrīna Baltmane and Linda Stauža for their support in the editing and finalisation process.

The preparation 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).

SATURS CONTENTS

LATVIEŠU VALODA DIGITĀLAJĀ LAIKMETĀ

1 Kopsavilkums 1

2 Risks mūsu valodām un izaicinājums valodu tehnoloģijām 4
- 2.1 Valodu barjeras kā šķērslis Eiropas informācijas sabiedrībā . . . 5
- 2.2 Mūsu apdraudētās valodas . . . 5
- 2.3 Valodu tehnoloģijas – kritiski svarīgas tehnoloģijas . . . 6
- 2.4 Valodu tehnoloģiju iespējas . . . 6
- 2.5 Valodu tehnoloģiju iespējamie izaicinājumi . . . 7
- 2.6 Veids, kā valodas apgūst cilvēki un mašīnas . . . 8

3 Latviešu valoda Eiropas informācijas sabiedrībā 9
- 3.1 Vispārīgi fakti . . . 9
- 3.2 Latviešu valodas specifika . . . 10
- 3.3 Jaunākās attīstības tendences . . . 12
- 3.4 Valodas attīstība Latvijā . . . 13
- 3.5 Valoda izglītībā . . . 14
- 3.6 Starptautiskie aspekti . . . 16
- 3.7 Latviešu valoda internetā . . . 16

4 Valodu tehnoloģiju atbalsts latviešu valodai 19
- 4.1 Lietojumprogrammu arhitektūra . . . 19
- 4.2 Galvenās izmantošanas iespējas . . . 20
- 4.3 Citas izmantošanas iespējas . . . 29
- 4.4 Izglītības programmas . . . 31
- 4.5 Projekti un sasniegumi . . . 31
- 4.6 Rīku un resursu pieejamība . . . 33
- 4.7 Starpvalodu salīdzinājums . . . 34
- 4.8 Secinājumi . . . 35

5 Par META-NET 39

THE LATVIAN LANGUAGE IN THE DIGITAL AGE

1 Executive Summary 41

2 Risks for Our Languages and a Challenge for Language Technology 44
- 2.1 Language Borders Hinder the European Information Society . . . 45
- 2.2 Our Languages at Risk . . . 45
- 2.3 Language Technology is a Key Enabling Technology . . . 45
- 2.4 Opportunities for Language Technology . . . 46
- 2.5 Challenges Facing Language Technology . . . 47
- 2.6 Language Acquisition in Humans and Machines . . . 47

3 Latvian in the European Information Society 49
- 3.1 General Facts . . . 49
- 3.2 Particularities of the Latvian Language . . . 50
- 3.3 Recent Developments . . . 52
- 3.4 Language Cultivation in Latvia . . . 53
- 3.5 Language in Education . . . 55
- 3.6 International Aspects . . . 56
- 3.7 Latvian on the Internet . . . 57

4 Language Technology Support for Latvian 60
- 4.1 Application Architectures . . . 60
- 4.2 Core Application Areas . . . 61
- 4.3 Other Application Areas . . . 70
- 4.4 Educational Programmes . . . 71
- 4.5 National Projects and Efforts . . . 72
- 4.6 Availability of Tools and Resources . . . 74
- 4.7 Cross-language Comparison . . . 74
- 4.8 Conclusions . . . 76

5 About META-NET 80

A Atsauces – References 81

B Dalīborganizācijas – META-NET Members 87

C META-NET balto grāmatu sērija – The META-NET White Paper Series 91

1

KOPSAVILKUMS

Informācijas tehnoloģija maina mūsu ikdienu. Mēs lietojam datorus, lai rakstītu, sazinātos, veiktu aprēķinus, meklētu informāciju un — arvien vairāk — lai lasītu, klausītos mūziku, skatītos fotoattēlus un filmas. Kabatā sev līdzi mēs nēsājam mazus datorus — viedtālruņus —, no kuriem zvanām, kuros rakstām un saņemam e-pasta vēstules, iegūstam informāciju un izklaidējamies neatkarīgi no atrašanās vietas. Informācija, zināšanas un ikdienas saziņa masveidā tiek digitalizēta. Kā tas ietekmē valodu? Vai mūsu valoda mainīsies vai pat izzudīs?

Visas mūsu skaitļošanas ierīces ir savstarpēji saistītas globālā tīklā, kas kļūst arvien blīvāks un jaudīgāks. Tomēr to, kā Fukušimas atomreaktorā notikusī avārija ietekmēs Eiropas enerģētikas politiku, eiropieši tiešsaistes forumos apspriež katrs savā valodā atsevišķās kopienās. Izmantojot internetu, cilvēki var sazināties, taču viņus joprojām šķir valodas barjera. Vai tā būs vienmēr?

Daudzas no pasaules 6900 valodām globalizētajā digitālās informācijas sabiedrībā neizdzīvos. Tiek lēsts, ka nākamajos gadu desmitos izzudīs vismaz 2000 valodu. Vēl daudzas citas tiks lietotas tikai ģimenes lokā un ikdienas saziņā, bet ne uzņēmējdarbības vidē vai zinātnē. Kādas izredzes izdzīvot ir latviešu valodai?

Latviešu valoda, ko visā pasaulē lieto aptuveni 1,5 miljoni cilvēku, valodas lietojuma ziņā ir apmēram 150. vietā pasaulē. Latviešu valoda ir vienīgā valsts valoda Latvijas Republikā un viena no Eiropas Savienības oficiālajām valodām.

2010. gadā Latvijā tika izdotas 2035 grāmatas un bukleti, kas ir visai daudz, tomēr kopējais izdoto eksemplāru skaits bija tikai 3,33 miljoni — ievērojami mazāk nekā 1991. gadā, kad tika izdoti 28,355 miljoni eksemplāru [3]. Programmas latviešu valodā piedāvā daudzas radiostacijas, divi sabiedriskās televīzijas kanāli un vairākas privātās telekompānijas. Daudzas ārzemju filmas tiek dublētas latviešu valodā.

Latvijā vēl joprojām jārisina problēmas, ko rada valstij "mantojumā" atstātā 20. gs. 50.–80. gados padomju varas piekoptā masveida imigrācijas un izglītības sistēmas segregācijas politika. Gandrīz trešās daļas Latvijas iedzīvotāju dzimtā valoda ir krievu valoda. Daudzās Latvijas skolās izglītību savulaik varēja iegūt tikai krievu valodā. Galu galā 1989. gadā tikai piektā daļa krieviski runājošo iedzīvotāju prata latviešu valodu [4]. Latviešu valodas nozīmes mazināšanās radīja bažas, ka tā pakāpeniski izzudīs.

Latviešu valodu aizsargā valsts valodas politika. Tās pamatprincips: latviešu valoda ir vienīgā Latvijas valsts valoda un dažādu Latvijā dzīvojošo etnisko grupu integrācijas valoda. Vienlaikus valsts valodas politika nodrošina iespēju saglabāt, attīstīt un lietot minoritāšu valodas dažādās jomās. Valdība cenšas risināt lingvistiskās segregācijas problēmu, veicinot bilingvālo izglītību un nosakot prasību vidusskolām vismaz 60% mācību priekšmetu pasniegt latviešu valodā.

Šo pasākumu rezultātā pašlaik vairāk nekā 75% iedzīvotāju, kuru dzimtā valoda ir krievu valoda, ir labas vai viduvējas latviešu valodas zināšanas, tostarp gandrīz visiem (94%) jauniešiem vecumā no 17 līdz 25 gadiem ir ļoti labas latviešu valodas zināšanas [5].

Latvijā nereti dzirdamas sūdzības par nemitīgi pieaugošo anglicismu lietošanu latviešu valodā, un dažkārt

pat paustas bažas, ka latviešu valodā ieviesīsies pārmērīgi daudz angļu valodas vārdu un frāžu. Tomēr latviešu valodas iekšējā sistēma ir izdzīvojusi pat pēc apjomīgas un daudzveidīgas saskares ar citām valodām (krievu, angļu, vācu, poļu, zviedru), un valoda ir saglabājusi stabilitāti. Neraugoties uz to, ir jāatzīst, ka pēc gadsimtiem ilgas svešzemju kundzības mūsdienu latviešu valodas leksikā un morfoloģijā var novērot plašu minēto svešvalodu ietekmi — aizguvumus, kalkus un aizgūtas, pilnībā asimilētas idiomas.

Neizzust mūsu skaistajiem latviešu valodas vārdiem un frāzēm var palīdzēt to bieža un apzināta lietošana; valodnieku polemika par svešvalodu ietekmi un oficiāli noteikumi parasti nav iedarbīgi. Visvairāk mums jāuztraucas nevis par valodas pakāpenisko pārangliskošanos, bet par tās pilnīgu izspiešanu no galvenajām sadzīves jomām.

Valodas situācija ir atkarīga ne tikai no tā, cik cilvēku tajā runā, cik grāmatu tajā izdots un filmu uzņemts vai cik televīzijas kanālu tajā pārraida, bet arī no valodas lietojuma digitālās informācijas telpā un datorprogrammās. Šajā jomā latviešu valodas pozīcijas nav tik labas. Mazāk nekā 0,1% pasaules tīmekļa vietņu ir latviešu valodā, un tas ir mazāk nekā lietuviešu vai slovēņu valodā pieejamo vietņu [6]. Kaut gan ir pieejamas vairāku globālu programmproduktu versijas latviešu valodā, daudz lietotāju labprātāk izvēlas angļu vai krievu valodas versiju.

Valodu tehnoloģijas jomā latviešu valodai nav īpaši laba tehnoloģiju un resursu nodrošinājuma. Kaut gan ir izstrādātas lietojumprogrammas un rīki, kas paredzēti pareizrakstības un gramatikas pārbaudei, teksta marķēšanai un vārdšķiru noteikšanai, tomēr ir arī pietiekami būtiski un steidzami novēršami trūkumi. It īpaši pietrūkst runas tehnoloģiju risinājumu un lielu un kvalitatīvu valodas resursu. Ir pieejamas elektroniskās vārdnīcas un lietojumprogrammas, kas paredzētas mašīntulkošanai latviešu valodā un no latviešu valodas svešvalodā. Kaut gan tās lieti noder, lai gūtu vispārīgu priekšstatu par svešvalodā sarakstīta teksta jēgu, tās vēl nevar izmantot, lai iegūtu lingvistiski un idiomātiski pareizus tulkojumus.

Informācijas un sakaru tehnoloģijas joma gatavojas nākamajai revolūcijai. Nākamās paaudzes tehnoloģija, kas mūsu dzīvē ienāks pēc personālajiem datoriem, tīkliem, miniaturizācijas, multivides, mobilajām ierīcēm un mākoņdatošanas, būs programmatūra, kas uztvers tekstuālus vai balsī izteiktus teikumus un lietotājiem būs daudz noderīgāka, jo "sapratīs" lietotājus un varēs sazināties ar viņiem lietotāju dzimtajā valodā. Šādu gaidāmo risinājumu priekšteči ir bezmaksas tiešsaistes pakalpojums *Google tulkotājs*, kas tulko tekstu daudzās valodās, IBM superdators *Watson*, kas uzvarēja spēles *Jeopardy* ASV čempionu, un produktam *iPhone* paredzētais *Apple* mobilais palīgs *Siri*, kas reaģē uz balss komandām un atbild uz jautājumiem angļu, vācu, franču un japāņu valodā.

Nākamā informācijas tehnoloģijas paaudze būs apguvusi cilvēku valodas tādā pakāpē, ka dažādu tautību cilvēki spēs sazināties, izmantojot šo tehnoloģiju savā dzimtajā valodā. Ierīces pratīs automātiski atrast svarīgākās ziņas un informāciju pasaules digitālajā zināšanu krātuvē, reaģējot uz viegli lietojamām balss komandām. Tehnoloģija, kas prot lietot valodu, varēs tulkot automātiski vai palīdzēt tulkiem darbā, sagatavot sarunu un dokumentu kopsavilkumus un būs noderīga mācībās. Piemēram, tās izmantošana vietējiem uzņēmumiem atvieglos klientu atrašanu ārzemēs, bet imigrantiem — latviešu valodas apguvi un pilnvērtīgāku integrēšanos sabiedrībā.

Nākamā informācijas un sakaru tehnoloģijas paaudze ļaus rūpniecības un pakalpojumu sfēras robotiem (kas pašlaik tiek izstrādāti pētniecības laboratorijās) "saprast" lietotāju vēlmes un sarunāties ar tiem.

Darbība šajā līmenī nozīmē krietni vairāk par rakstzīmju apstrādi vai vienkāršu leksikonu izpratni, pareizrakstības vai pareizrunas pārbaudi. Tehnoloģijas izstrādē vairs nepietiek ar vienkāršotu pieeju, ir jāķeras pie visaptverošas valodas modelēšanas, ņemot vērā sintaksi un semantiku, lai izprastu cilvēka uzdoto jautājumu

būtību un spētu sniegt pilnvērtīgas un precīzas atbildes. Ne visas Eiropas valodas ir līdzvērtīgi sagatavojušās šim nākotnes uzdevumam. Angļu valodas tehnoloģiskais nodrošinājums ir ievērojami plašāks nekā latviešu valodai, un šī nevienlīdzība arvien palielinās. Tas ir vērojams ne tikai salīdzinājumā ar lielākajām valodām, bet arī ar mazākām valodām, kam ir bijis pieejams sistemātisks valsts atbalsts valodu tehnoloģiju izstrādē.

Valodu tehnoloģija Latvijā nekad nav bijusi prioritāra pētniecības joma. Mūsu valstī nav īpašas valodas tehnoloģijas programmas, pētniecības un izstrādes darbs ir fragmentārs un galvenokārt tiek organizēts īstermiņa projektos, kas sarežģī lielāka apjoma resursu izstrādi un iestāžu sadarbību ilgtermiņā. Ir maz mācību kursu, kas būtu saistīti ar valodas tehnoloģiju. Tomēr 2005.–2009. gadā valsts pētniecības programmās informācijas un komunikāciju tehnoloģiju (IKT) jomā un latviešu valodas pētniecības programmās ir īstenoti vairāki sekmīgi projekti. Pēc tam valodas tehnoloģijai sniegtais atbalsts ievērojami samazinājās, tāpēc tika īstenoti tikai daži pasākumi semantikas, kontrolētās valodas un mašīntulkošanas jomā. Tomēr pētniecības institūtu un universitāšu izpētes potenciāls joprojām ir liels.

Līdztekus pētniecības centriem un universitātēm vērā ņemami sasniegumi ir bijuši novatoriskiem valodas tehnoloģijas izstrādes uzņēmumiem. Pievēršot uzmanību praktiski izmantojamām lietojumprogrammām un strādājot nozīmīgos Eiropas mēroga sadarbības projektos, īpaši ievērojams progress sasniegts tulkošanas tehnoloģiju jomā.

Katrā starptautiskā tehnoloģiju salīdzinājumā ir vērojama tendence, ka angļu valodas automātiskās analīzes rezultāti ir ievērojami labāki nekā citu valodu, arī latviešu valodas, analīzes rezultāti. Daudzi pētnieki uzskata — atpalicības cēlonis ir tas, ka pēdējos piecdesmit gadus datorlingvistikas metožu un algoritmu izstrādē un valodas tehnoloģijas lietojamības pētījumos uzmanība pirmām kārtām tiek pievērsta angļu valodai. Savukārt citi pētnieki uzskata, ka angļu valoda savu īpašību dēļ ir labāk piemērota datorapstrādei. Izmantojot pašlaik pieejamās metodes, arī tekstu franču un spāņu valodā ir daudz vieglāk apstrādāt nekā tekstu latviešu valodā. Tas nozīmē — ja vēlamies tajās privātās un darba dzīves jomās, kurās lietojam latviešu valodu, izmantot nākamās paaudzes informācijas un komunikācijas tehnoloģiju, ir nepieciešami mērķtiecīgi, sistemātiski un ilgtspējīgi pētījumi.

Latviešu valodai nedraud tūlītējas briesmas; tādas nerada pat angļu valodai izstrādāto valodas tehnoloģiju lielais pārākums. Tomēr situācija var radikāli mainīties, ja jaunās paaudzes tehnoloģijas patiešām efektīvi apgūs cilvēku valodu. Tādi valodas tehnoloģijas sasniegumi kā kvalitatīva mašīntulkošana palīdzēs pārvarēt valodas barjeras, taču šie sasniegumi būs lietojami tikai valodās, kas izdzīvos digitālajā pasaulē. Ja būs pieejams pietiekami daudz pieņemamas kvalitātes valodas tehnoloģijas risinājumu, valoda spēs izdzīvot arī tad, ja tās lietotāju skaits būs neliels. Ja šis priekšnosacījums netiks izpildīts, pat lielākas valodas var kļūt apdraudētas. Lai latviešu valoda arī turpmāk būtu dzīvotspējīga valoda attīstītajā pasaulē, tai jābūt pieejamiem atbilstošiem IT risinājumiem. Tāpēc valsts valodas politikai jānodrošina sistemātisks darbs valodas tehnoloģijas jomā un tam nepieciešamie ieguldījumi.