META NET

White Paper Series Zbirka Bela knjiga

THE SLOVENE LANGUAGE IN THE DIGITAL AGE

SLOVENSKI JEZIK V DIGITALNI DOBI

Simon Krek

Springer

White Paper Series Zbirka Bela knjiga

THE SLOVENE LANGUAGE IN THE DIGITAL AGE

SLOVENSKI JEZIK V DIGITALNI DOBI

Simon Krek "Jožef Stefan" Institute, Amebis, d. o. o.

Georg Rehm, Hans Uszkoreit
(urednika, 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 978-3-642-30635-8 ISBN 978-3-642-30636-5 (eBook)
DOI 10.1007/978-3-642-30636-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940564

PREDGOVOR

Bela knjiga je del zbirke, s katero širimo zavedanje o jezikovnih tehnologijah in o možnostih, ki jih ponujajo. Namenjena je izobraževalcem, novinarjem, politikom, jezikovnim skupnostim in vsem ostalim, ki jih zanima jezik. Dostopnost in raba jezikovnih tehnologij v Evropi se razlikuje od jezika do jezika. V skladu s tem se dejanja, potrebna za podporo raziskovanju in razvoju, med seboj razlikujejo in so odvisna od različnih dejavnikov, na primer od zahtevnosti jezikov ali velikosti njihovih skupnosti.

V projektu META-NET, mreži odličnosti, ki jo financira Evropska komisija, smo analizirali obstoječe stanje na področju jezikovnih virov in tehnologij (glej str. 79). Analiza zajema 23 uradnih evropskih jezikov in ter nekatere druge pomembne evropske nacionalne in regionalne jezike. Rezultati analize kažejo, da pri vsakem jeziku obstaja precej vrzeli, detajlna strokovna analiza in ocena trenutnega stanja pa bo pripomogla k najboljšemu izkoristku novih raziskav in zmanjšanju s tem povezanih tveganj.

Mrežo META-NET sestavlja 54 raziskovalnih centrov iz 33 držav (stanje novembra 2011, glej str. 75). V projektu sodelujemo z déležniki iz gospodarstva (računalniška podjetja, ponudniki tehnologij, uporabniki), državnih institucij, raziskovalnih organizacij, nevladnih organizacij, jezikovnih skupnosti in evropskih univerz. Skupaj z navedenimi skupnostmi v projektu META-NET ustvarjamo skupno tehnološko vizijo in strateški raziskovalni načrt za večjezično Evropo 2020.

PREFACE

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 differs. 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. 79). 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. 75). 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.

Avtor se zahvaljuje dr. Marku Stabeju (Filozofska fakulteta, Univerza v Ljubljani) in dr. Tomažu Erjavcu (Institut "Jožef Stefan") za njun prispevek pri nastanku te publikacije. Poleg tega se zahvaljuje avtorjem bele knjige o nemškem jeziku za dovoljenje glede uporabe jezikovno neodvisnih delov publikacije [1].

Izdelava bele knjige je bila financirana s sredstvi Sedmega okvirnega programa in Programa za podporo razvoju politik informacijsko-komunikacijskih tehnologij Evropske komisije v okviru pogodb T4ME (sporazum o dodelitvi sredstev 249119), CESAR (sporazum o dodelitvi sredstev 271022), METANET4U (sporazum o dodelitvi sredstev 270893) in META-NORD (sporazum o dodelitvi sredstev 270899).

The author of this document would like to thank Marko Stabej (Faculty of Arts, University of Ljubljana) and Tomaž Erjavec ("Jožef Stefan" Institute) for their contributions to this white paper. Furthermore, the author is grateful to the authors of the white paper on German for permission to re-use selected language-independent materials from their document [1].

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 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

KAZALO CONTENTS

SLOVENSKI JEZIK V DIGITALNI DOBI

1 Povzetek 1

2 Tveganje za naše jezike in izziv za jezikovne tehnologije 3
- 2.1 Jezikovne meje ovirajo evropsko informacijsko družbo . . . 4
- 2.2 Naši jeziki so ogroženi . . . 4
- 2.3 Jezikovne tehnologije so ključne podporne tehnologije . . . 5
- 2.4 Priložnosti za jezikovne tehnologije . . . 5
- 2.5 Izzivi za jezikovne tehnologije . . . 6
- 2.6 Usvajanje jezika pri ljudeh in strojih . . . 7

3 Slovenščina v evropski informacijski družbi 9
- 3.1 Splošni podatki . . . 9
- 3.2 Značilnosti slovenskega jezika . . . 10
- 3.3 Razvoj v zadnjem času . . . 11
- 3.4 Skrb za jezik v Sloveniji . . . 12
- 3.5 Jezik v izobraževanju . . . 13
- 3.6 Mednarodni vidiki . . . 15
- 3.7 Slovenščina na internetu . . . 16

4 Jezikovne tehnologije za slovenščino 17
- 4.1 Procesna arhitektura . . . 17
- 4.2 Ključne aplikacije . . . 18
- 4.3 Druge aplikacije . . . 26
- 4.4 Izobraževalni programi . . . 27
- 4.5 Nacionalni projekti in pobude . . . 28
- 4.6 Dostopnost virov in orodij . . . 30
- 4.7 Primerjava med jeziki . . . 30
- 4.8 Zaključek . . . 31

5 O projektu META-NET 35

THE SLOVENE LANGUAGE IN THE DIGITAL AGE

1 Executive Summary 37

2 Languages at Risk: a Challenge for Language Technology 39
- 2.1 Language Borders Hold back the European Information Society . . . 40
- 2.2 Our Languages at Risk . . . 40
- 2.3 Language Technology is a Key Enabling Technology . . . 41
- 2.4 Opportunities for Language Technology . . . 41
- 2.5 Challenges Facing Language Technology . . . 42
- 2.6 Language Acquisition in Humans and Machines . . . 42

3 The Slovene Language in the European Information Society 44
- 3.1 General Facts . . . 44
- 3.2 Particularities of the Slovene Language . . . 45
- 3.3 Recent Developments . . . 47
- 3.4 Official Language Protection in Slovenia . . . 48
- 3.5 Language in Education . . . 49
- 3.6 International Aspects . . . 50
- 3.7 Slovene on the Internet . . . 52

4 Language Technology Support for Slovene 53
- 4.1 Application Architectures . . . 53
- 4.2 Core Application Areas . . . 54
- 4.3 Other Application Areas . . . 61
- 4.4 Educational Programmes . . . 63
- 4.5 National Projects and Initiatives . . . 63
- 4.6 Availability of Tools and Resources . . . 65
- 4.7 Cross-language comparison . . . 65
- 4.8 Conclusions . . . 67

5 About META-NET 70

A Bibliografija – References 71

B Članstvo v META-NET – META-NET Members 75

C Zbirka Bela knjiga META-NET – The META-NET White Paper Series 79

POVZETEK

V zadnjih 60 letih je Evropa postala prepoznavna politična in ekonomska danost, vendar je kulturno in jezikovno še vedno zelo raznolika. To pomeni, da se od portugalščine do poljščine, od italijanščine do islandščine neizogibno soočamo z jezikovnimi mejami pri vsakodnevni komunikaciji med prebivalci Evrope, kot tudi znotraj poslovne in politične sfere. Evropske institucije potrošijo približno milijardo evrov na leto za vzdrževanje politike večjezičnosti, torej za prevajanje besedil in za tolmačenje pri govorni komunikaciji. Pa je nujno, da takšno breme ostaja še naprej? Sodobne jezikovne tehnologije in jezikoslovne raziskave lahko pomembno prispevajo k rušenju jezikovnih meja. Kombinirane s pametnimi napravami in računalniškimi programi bodo jezikovne tehnologije v prihodnosti pripomogle, da bodo prebivalci Evrope lahko govorili drug z drugim ali skupaj poslovali, tudi če ne bodo govorili skupnega jezika.

Jezikovne tehnologije gradijo mostove.

Slovensko gospodarstvo je v juliju 2011 v države EU izvozilo 71,9 % od celotnega izvoza blaga. Nemško gospodarstvo kot največje evropsko gospodarstvo je v letu 2010 v države EU izvozilo 60,3 % blaga, z dodatnimi 10,8 % izvoza v ostale evropske države. Jezikovne meje lahko poslovanje povsem zaustavijo, kar velja predvsem za mala in srednja podjetja, ki nimajo finančnih sredstev za prilagoditev stanju. Edina (nezamisljiva) alternativa večjezični Evropi bi bila, če bi dovolili, da en jezik prevzame dominantni položaj in na koncu nadomesti vse ostale jezike. Tradicionalna pot za premagovanje jezikovnih ovir je učenje tujih jezikov. Toda brez tehnološke podpore je obvladovanje 23 uradnih in približno 60 drugih evropskih jezikov nepremostljiva ovira za evropske državljane, evropsko gospodarstvo, politične razprave in znanstveni razvoj. Rešitev je v razvoju ključnih podpornih tehnologij. Te bodo evropskim akterjem zagotovile prednost ne le v okviru skupnega evropskega trga, temveč tudi pri trgovanju s tretjimi državami, predvsem s hitro rastočimi gospodarstvi. Da bi ta cilj dosegli in ohranili evropsko kulturno in jezikovno raznolikost, je najprej treba sistematično analizirati jezikovne značilnosti vseh evropskih jezikov in trenutno stanje jezikovnotehnološke podpore za vsakega od njih. Jezikovnotehnološke rešitve bodo na koncu služile kot most med evropskimi jeziki. Orodja za strojno prevajanje in procesiranje govora, ki so na voljo na tržišču, še ne izpolnjujejo tega zahtevnega cilja. Prevladujoči igralci na tem področju so predvsem zasebna tržno usmerjena severnoameriška podjetja. Že v poznih 70-ih letih je EU prepoznala pomen jezikovnih tehnologij kot gonila evropske enotnosti in začela financirati prve raziskovalne projekte, kakršen je bil npr. EUROTRA. Hkrati se je začelo financiranje nacionalnih projektov, katerih rezultatih so bili dragoceni, toda skupna usklajena evropska akcija ni bila nikoli izpeljana. V nasprotju z omenjenimi nepovezanimi napori pri financiranju so druge večjezične družbe, kot sta Indija (22 uradnih jezikov) ali Južna Afrika (11 uradnih jezikov), v zadnjem času izdelale dolgoročne nacionalne programe raziskovanja jezikov in tehnološkega razvoja.

Jezikovne tehnologije kot ključ za prihodnost.

Sedanji prevladujoči igralci na področju jezkovnih tehnologij se zanašajo na nenatančne statistične pristope, pri katerih ne uporabljajo zahtevnejših jezikoslovnih metod in znanja. Stavki so denimo prevedeni avtomatsko zgolj s primerjavo novonastalega stavka s tisoči stavkov, ki so jih prevedli ljudje. Kvaliteta rezultata je v veliki meri odvisna od količine in kakovosti dostopnega korpusa vzorcev. Če z avtomatskim prevajanjem preprostih stavkov pri jezikih, za katere je na voljo zadostna količina besedilnega gradiva, lahko pridemo do uporabnih rezultatov, so statistične metode obsojene na neuspeh pri jezikih, za katere je na voljo precej manjša količina vzorčnega gradiva ali pri stavkih z zapleteno strukutro.

Jezikovne tehnologije pomagajo združevati Evropo.

Evropska unija je zato sklenila, da bo financirala projekte, kot sta EuroMatrix in EuroMatrixPlus (od l. 2006) in iTranslate4 (od l. 2010), v okviru katerih se izvajajo temeljne in aplikativne raziskave in ki ustvarjajo vire, potrebne za vzpostavljanje kvalitetnih jezikovnotehnoloških rešitev za vse evropske jezike. Analiza globjih strukturnih značilnosti jezikov je edina pot naprej, če želimo zgraditi aplikacije, ki dobro delujejo pri celotnem razponu evropskih jezikov. Dosedanje evropske raziskave so bile na tem področju že zelo uspešne. Prevajalske službe Evropske unije tako uporabljajo prosto dostopni strojni prevajalnik MOSES, ki je bil razvit pretežno v okviru evropskih raziskovalnih projektov.

Po dosedanjih dognanjih se zdi, da bodo današnje "hibridne" jezikovne tehnologije, pri katerih se zahtevnejša analitična obdelava meša s statističnimi metodami, lahko premostile vrzeli med vsemi evropskimi jeziki ter med drugimi jeziki. Kot kaže ta zbirka belih knjig, med članicami Evropske unije v zvezi z jezikovnimi rešitvami in stanjem raziskav obstajajo dramatične razlike glede pripravljenosti. Po natančnem pregledu in primerjavi z drugimi jeziki lahko ugotovimo, da je stanje pri jezikovnih tehnologijah in virih za slovenščino dokaj zaskrbljujoče, in sicer iz dveh razlogov. Prvi razlog je razumljiv in izhaja iz števila govorcev slovenščine. Teh je približno 2 milijona, kar ne zagotavlja, da bi se viri in tehnologije lahko razvijali zgolj znotraj komercialnega okolja. Na drugi stani država Slovenija oz. institucije, ki znotraj slovenske jezikovne skupnosti skrbijo za razvoj jezika, v zadnjem desetletju niso uspele zagotoviti ustreznega institucionalnega okvira, kjer bi potekal načrten in sistematičen dolgoročni razvoj tehnologij, virov in orodij, ki so jezikovno specifični. Brez tega ni mogoče pričakovati, da bo slovenščina obdržala enakovreden status v prihodnjem digitalnem okolju. Posledica pomanjkanja trajnega institucionalnega okvira je tudi ta, da je v slovenskem akademskem okolju študij računalniškega procesiranja naravnih jezikov bistveno premalo prisoten. Najpomembnejši korak pri zagotavljanju kvalitetnih jezikovnih tehnologij in virov za slovenščino bi bila torej čimprejšnja izdelava programa njihovega razvoja in zagotovitev ustreznega institucionalnega okvira, ki bi ta program izvajal. Dolgoročni cilj mreže META-NET je uvedba kakovostnih jezikovnih tehnologij za vse jezike, da bi vzpostavili politično in ekonomsko enotnost skozi kulturno različnost. Tehnologije bodo pomagale podreti zidove in zgraditi mostove med evropskimi jeziki. Za to je potrebno, da vsi deležniki – v politiki, raziskovanju, gospodarstvu in v družbi – združijo svoje napore za prihodnost.

Zbirka Bela knjiga dopolnjuje strateške akcije, ki jih izvaja mreža META-NET (za pregled glej prilogo). Sveže informacije, kot npr. zadnjo verzijo Strateške vizije [2] ali Strateški raziskovalni načrt, je mogoče najti na spletni strani mreže META-NET: http://www.meta-net.eu.