

THE SERBIAN LANGUAGE IN THE DIGITAL AGE
СРПСКИ ЈЕЗИК У ДИГИТАЛНОМ ДОБУ

Duško Vitas
Ljubomir Popović
Cvetana Krstev
Ivan Obradović
Gordana Pavlović-Lažetić
Mladen Stanojević



White Paper Series

Серија белих књига

THE SERBIAN
LANGUAGE IN
THE DIGITAL
AGE

СРПСКИ
ЈЕЗИК У
ДИГИТАЛНОМ
ДОБУ

Duško Vitas University of Belgrade

Ljubomir Popović University of Belgrade

Cvetana Krstev University of Belgrade

Ivan Obradović University of Belgrade

Gordana Pavlović-Lažetić University of Belgrade

Mladen Stanojević University of Belgrade

Georg Rehm, Hans Uszkoreit

(уредници, 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 978-3-642-30754-6 ISBN 978-3-642-30755-3 (eBook)
DOI 10.1007/978-3-642-30755-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012945734

© 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)



ПРЕДГОВОР

Ова бела књига је део серије која промовише знање о језичким технологијама и њиховим могућностима. Намењена је наставницима језика, новинарима, политичарима, језичким заједницама и другима. Покривеност језичким технологијама и начин њихове употребе се у Европи разликују од језика до језика. Због тога се разликују и активности које је потребно спровести да би се подржала истраживања и развој, а неопходни кораци зависе од многих фактора, као што су сложеност језика или величина заједнице која га користи. Пројекат МЕТА-НЕТ, мрежа изврности коју финансира Европска комисија, спровео је анализу текућих језичких ресурса и технологија. Анализа је била усмерена на 23 званична европска језика, као и на друге значајне националне и регионалне језике у Европи. Резултати анализе сугеришу постојање многих значајних празнина у истраживањима за сваки језик. Детаљнија експертска анализа и процена текуће ситуације за сваки језик помоћи ће да се повећа утицај нових истраживања и умање могући ризици. Према стању из новембра 2011, МЕТА-НЕТ повезује 54 истраживачка центра из 33 земље (стр. 81), који сарађују са заинтересованим странама из сфера предузетништва, државних институција, привреде, истраживачких организација, софтверских компанија, понуђача технологија и европских универзитета. Они заједно граде технолошку визију кроз развој стратешких истраживачких програма који показују како ће примене језичких технологија попунити постојеће празнине у истраживањима до 2020. године.

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 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. 84). The analysis focuses 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 future research.

As of November 2011, META-NET consists of 54 research centres in 33 European countries (p. 81). META-NET is working with stakeholders from economy (software companies, technology providers and 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.

Захваљујемо се ауторима беле књиге о немачком језику [1] што су дозволили да језички независне делове њиховог текста користимо у овом раду.

Израду ове беле књиге финансирали су Седми оквирни програм (FP7) и Програм подршке политици информационо-комуникационих технологија Европске комисије преко уговора T4ME (Уговор о финансирању 249 119), CESAR (Уговор о финансирању 271 022), METANET4U (Уговор о финансирању 270 893) и META-NORD (Уговор о финансирању 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 [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 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).



САДРЖАЈ CONTENTS

СРПСКИ ЈЕЗИК У ДИГИТАЛНОМ ДОБУ

1 Резиме	1
2 Опасност по наше језике и изазови пред језичким технологијама	4
2.1 Језичке границе представљају сметњу за европско информационо друштво	5
2.2 Наши језици су угрожени	5
2.3 Језичке технологије су кључне потпорне технологије	6
2.4 Могућности језичких технологија	6
2.5 Изазови пред језичким технологијама	7
2.6 Усвајање језика код људи и машина	8
3 Српски језик у европском информационом друштву	10
3.1 Општи подаци	10
3.2 Специфичности српског језика	11
3.3 Савремени развој	16
3.4 Неговање језика у Србији	16
3.5 Језик и образовање	17
3.6 Међународни аспекти	18
3.7 Српски језик на интернету	18
4 Језичке технологије за српски језик	20
4.1 Архитектуре апликација	20
4.2 Основна поља примене	21
4.3 Друге области примене	29
4.4 Образовни програми	31
4.5 Национални пројекти и иницијативе	32
4.6 Доступност алата и ресурса	34
4.7 Поређење језика	35
4.8 Закључци	36
5 МЕТА-НЕТ (МЕТА-NET)	40

THE SERBIAN LANGUAGE IN THE DIGITAL AGE

1	Executive Summary	41
2	Languages at Risk: a Challenge for Language Technology	43
2.1	Language Borders Hold back the European Information Society	44
2.2	Our Languages at Risk	44
2.3	Language Technology is a Key Enabling Technology	44
2.4	Opportunities for Language Technology	45
2.5	Challenges Facing Language Technology	46
2.6	Language Acquisition in Humans and Machines	46
3	The Serbian language in the European Information Society	48
3.1	General Facts	48
3.2	Particularities of the Serbian Language	49
3.3	Recent Developments	54
3.4	Official Language Protection in Serbia	54
3.5	Language in Education	55
3.6	International Aspects	56
3.7	Serbian on the Internet	56
4	Language Technology Support for Serbian	58
4.1	Application Architectures	58
4.2	Core Application Areas	59
4.3	Other Application Areas	66
4.4	Educational Programmes	67
4.5	National Projects and Initiatives	68
4.6	Availability of Tools and Resources	70
4.7	Cross-language comparison	71
4.8	Conclusions	72
5	About META-NET	76
A	Литература – References	77
B	Чланице META-NET-а – META-NET Members	81
C	META-NET серија белих књига – The META-NET White Paper Series	84

РЕЗИМЕ

У последњих 60 година Европа је постала јединствена политичка и економска структура, мада је културно и језички веома разноврсна. То значи да је, од португалског до пољског, од италијанског до исландског, свакодневна комуникација становника Европе, као и комуникација у сфери пословања и политике, нужно суочена са језичким препрекама. Институције Европске уније троше око милијарду евра годишње на одржавање своје политике вишејезичности, тј. на превођење текстова и говорне комуникације. Питање које се поставља јесте да ли је толико оптерећење неопходно. Модерне језичке технологије и лингвистичка истраживања могу значајно да допринесу брисању језичких граница. Језичке технологије у комбинацији са интелигентним уређајима и апликацијама могу у будућности да помогну Европљанима да се међусобно споразумевају једноставно и лако и да обављају послове и кад не говоре истим језиком.

Језичке технологије граде мостове за европску будућност.

Главни трговински партнери Србије су земље Европске уније, са уделом од преко 50% у укупној трговинској размени, при чему је извоз из Србије на тржиште ЕУ ослобођен царине у складу са Споразумом о стабилизацији и придруживању. Али језичке препреке могу да зауставе пословање, посебно за СМП (средња и мала предузећа) која немају финансијских средстава да их превазиђу. Једина (незамислива) ал-

тернатива за вишејезичну Европу била би да један језик почне да доминира и на крају замени све остале језике.

Један традиционални начин превазилажења језичких баријера јесте учење страних језика. Међутим, без технолошке подршке, савладавање 23 званична језика земаља чланица Европске уније и око 60 других европских језика, за становнике Европе представља непремостиву препреку, баш као и за њену економију, политичке дебате и научни напредак.

Решење лежи у изградњи кључних потпорних технологија. Оне ће европским актерима понудити огромне предности, не само у оквиру заједничког европског тржишта већ и у трговинским односима са трећим земљама, посебно са привредама које се брзо развијају. Да би се постигао тај циљ и очувала европска културна и језичка разноврсност, неопходно је да се прво спроведе систематска анализа језичких специфичности свих европских језика, као и текућег стања њихове опремљености језичким технологијама. На тај начин ће језичке технологије послужити као јединствени мост међу европским језицима.

Језичке технологије као решење за будућност.

Алати за аутоматско превођење и обраду говора који се могу наћи на тржишту још увек не омогућавају остварење овог амбициозног циља. Главни актери на овом пољу су пре свега приватна профитна предузећа из Северне Америке. Још крајем 1970-их Европска

унија је препознала суштински значај језичких технологија као покретача европског јединства, и почела је са финансирањем првих истраживачких пројеката као што је био пројекат EUROTRA. У исто време, започели су национални пројекти, који су дали вредне резултате, али нису покренули и заједничку усклађену европску акцију. Насупрот овим појединачним и неповезаним напорима у финансирању, друга вишејезична друштва, као што су Индија (22 званична језика) и Јужна Африка (11 званичних језика) [2], недавно су почела дугорочне националне програме језичких истраживања и технолошког развоја.

Данас се главни актери на подручју језичких технологија ослањају на непрецизне статистичке приступе, који не користе дубље лингвистичке методе и знања. На пример, реченице се аутоматски преводје тако што се нове реченице пореде са хиљадама претходно „ручно” преведених реченица. Квалитет резултата у великој мери зависи од квантитета и квалитета расположивог корпуса узорака. Мада машинско превођење једноставних реченица може да пружи употребљиве резултате у језицима са довољном количином расположивог текстуелног материјала, ове плитке статистичке методе нужно доживљавају неуспех у случају језика са мањим обимом узорака или у случају реченица комплексне структуре.

Европска унија је због тога одлучила да финансира пројекте као што су EuroMatrix и EuroMatrixPlus (од 2006) и iTranslate4 (од 2010) који спроводе основна и примењена истраживања и стварају ресурсе за успостављање језикотехнолошких решења високог квалитета за све европске језике. Анализа дубљих структурних својстава језика је једини начин да се изграде апликације које дају добре резултате на целом распону европских језика.

Европска истраживања у овој области већ су постигла бројне успехе. На пример, преводилачки

сервиси Европске уније користе софтвер отвореног кода за машинско превођење MOSES, који је претежно развијен кроз европске истраживачке пројекте. Суштински пробој у области синтезе и препознавања говора на српском језику начинила је група са Факултета техничких наука Универзитета у Новом Саду. Развијен је низ апликација у области TTS и ASR на бази говорних и лексичких база података акценатованих облика речи. Препознавање и генерисање говора за српски комерцијализовала је фирма AlfaNum која је потекла са Универзитета у Новом Саду. AlfaNum има значајан број корисника међу српским фирмама. С друге стране, први корпус савременог српског језика, електронски морфолошки речник, паралелни француско-српски и енглеско-српски корпуси литерарних текстова, као и различити софтверски алати развијени су у оквиру заједничких пројеката Математичког факултета и Одсека за српски језик Филолошког факултета у Београду.

Језичке технологије помажу уједињењу Европе.

Према увиду у досадашње стање, сви су изгледи да ће „хибридне” језичке технологије које комбинују дубинску обраду са статистичким методама бити у могућности да премосте јаз између свих европских језика, и шире. Како показује ова серија белих књига, постоји драматична разлика у степену припремљености када су у питању језичка решења и стање истраживања међу европским језицима. Српски језик је један од „мањих” европских језика и потребна су даља истраживања која ће омогућити да ефикасна решења која нуде језичке технологије уђу у свакодневну употребу.

Дугорочни циљ МЕТА-НЕТ-а јесте да уведе језичке технологије високог квалитета за све језике, како би се постигло политичко и економско јединство кроз

културну разноврсност. Те технологије ће помоћи да се уклоне постојеће баријере и да се изграде мостови међу европским језицима. Ово захтева од свих заинтересованих учесника – у политици, истраживању, привреди и друштву – да уједине своје напоре за будућност. Ова серија белих књига допуњује друге

стратешке активности које предузима МЕТА-НЕТ (видети преглед у додатку). Ажурне информације као што су текућа верзија текста МЕТА-НЕТ визије [2] или стратешки истраживачки план рада (Strategic Research Agenda, SRA) могу се наћи на МЕТА-НЕТ веб локацији: <http://www.meta-net.eu>.