META NET

White Paper Series Niz Bijele Knjige

THE CROATIAN LANGUAGE IN THE DIGITAL AGE

HRVATSKI JEZIK U DIGITALNOM DOBU

Marko Tadić
Dunja Brozović-Rončević
Amir Kapetanović

Springer

White Paper Series Niz Bijele Knjige

THE CROATIAN LANGUAGE IN THE DIGITAL AGE

HRVATSKI JEZIK U DIGITALNOM DOBU

Marko Tadić [1]
Dunja Brozović-Rončević [2]
Amir Kapetanović [2]

[1] Filozofski Fakultet, Zagreb
[2] Institut za hrvatski jezik i jezikoslovlje

Georg Rehm, Hans Uszkoreit
(urednici, 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 978-3-642-30881-9 ISBN 978-3-642-30882-6 (eBook)
DOI 10.1007/978-3-642-30882-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012946921

© 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)

PREDGOVOR PREFACE

Ova bijela knjiga dio je niza koji promiče jezične tehnologije i njihove mogućnosti. Namijenjena je novinarima, političarima, jezičnim zajednicama, učiteljima, predavačima i ostalima. Dostupnost i uporaba jezičnih tehnologija u Europi različita je od jezika do jezika. Sukladno, različite su i aktivnosti potrebne za daljnju potporu istraživanjima i razvoju jezičnih tehnologija od jezika do jezika. Potrebne akcije ovise o mnogo čimbenika kao što su složenost pojedinoga jezika i veličina dotične jezične zajednice.

Mreža izvrsnosti META-NET, koju podupire Europska komisija, provela je analizu trenutačno raspoloživih jezičnih resursa i tehnologija u ovome nizu bijelih knjiga (s. 93). Ta je analiza usredotočena ponajprije na 23 službena jezike Europske unije, ali i na ostale važne nacionalne i regionalne jezike u Europi. Rezultati ove analise ukazuju na nesrazmjerne nedostatke u tehnološkoj potpori i značajne istraživačke nedostatke za svaki od promatranih jezika. Predstavljena podrobna stručna analiza i procjena trenutačne situacije pomoći će u učinkovitosti dodatnih istraživanja u tome smjeru.

Od mjeseca studenoga 2011. META-NET se sastoji od 54 istraživačka središta iz 33 europske zemlje (s. 89). META-NET surađuje s ključnim dionicima iz gospodarstva (tvrtke koje izgrađuju programsku podršku, tehnološki isporučitelji, korisnici), vladinim agencijama, istraživačkim organizacijama, nevladinim organizacijama, jezičnih zajednicama i europskim sveučilištima. Zajedno s njima META-NET stvara zajedničku tehnološku viziju i strateški plan za višejezičnu Europu 2020.

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. 93). 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. 89). 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.

Autori ovoga dokumenta zahvalni su autorima Bijele knjige o njemačkome jeziku za dopuštenje uporabe odabrane jezično-neovisne građe iz njihovoga teksta [1].

Izradba ove bijele knjige poduprta je od strane Sedmoga okvirnoga programa i ICT programa za podršku politici Europske komisije u skladu s ugovorima T4ME (opći ugovor 249 119), CESAR (opći ugovor 271 022), METANET4U (opći ugovor 270 893) i META-NORD (opći ugovor 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).

SADRŽAJ CONTENTS

HRVATSKI JEZIK U DIGITALNOM DOBU

THE CROATIAN LANGUAGE IN THE DIGITAL AGE

1

SAŽETAK

Informacijske tehnologije mijenjaju naš svakodnevni život. Svakodnevno se služimo računalima za pisanje, uređivanje, računanje, pretragu obavijesti i sve više za čitanje, slušanje glazbe, pregledavanje fotografija i gledanje filmova. U svojim džepovima nosimo mala računala koja koristimo za obavljanje telefonskih poziva, pisanje e-pošte, prikupljanje obavijesti i za zabavu gdje god se nalazili. Kako ta masovna digitalizacija obavijesti, znanja i svakodnevnih komunikacija utječe na naš jezik? Hoće li se naš jezik promijeniti ili čak nestati? Kakve su mogućnosti hrvatskoga jezika za preživljavanje?

Mnogi od šest tisuća jezika na svijetu ne će preživjeti u globaliziranom digitalnom informacijskom društvu. Procjenjuje se kako je barem dvije tisuće jezika osuđeno na izumiranje u sljedećem desetljeću. Preostali će nastaviti igrati ulogu u privatnome krugu obitelji ili susjedstva, ali ne nužno i na razini općega poslovanja ili na akademskoj razini. Status jezika ne ovisi samo o broju njegovih govornika ili broju knjiga, filmova i TV-postaja koje se njime služe, nego i o prisutnosti toga jezika u digitalnome informacijskome prostoru i u adekvatnoj programskoj podršci.

U današnjem informacijski usmjerenom društvu, mogućnost dostupa obavijestima na vlastitome jeziku smatra se dosegnutom civilizacijskom razinom nezaobilaznom za prevladavaju digitalnoga jaza. Naime, jezične zajednice, koje za svoj jezik ne budu imale razvijene jezične tehnologije, ostat će s druge strane digitalne razdjelnice. Kad je riječ o hrvatskome jeziku i jezičnim tehnologijama, onda ponajprije valja imati na umu ne samo osiguranje njegova ravnopravnoga sudjelovanja s drugim jezicima u globaliziranome informacijskome društvu, nego i promjenu njegovih sociolingvističkih okolnosti koja se može očekivati u 2013. kad će postati 24. službeni jezik Europske unije. Od toga trenutka za hrvatski se jezik očekuje dostupnost čitavoga niza jezičnotehnoloških resursa, alata i usluga kakve već postoje, ali se isto tako i dalje nesmetano razvijaju za ostale službene jezike EU-a. Tražilice koje mogu pretraživati puni tekst prema svim oblicima u kojima se hrvatske riječi mogu pojavljivati, sustavi za diktiranje tj. automatsko pretvaranje govora na hrvatskome u tekst, ili, možda najvažniji, sustavi za strogo prevođenje na i sa hrvatskoga, samo su neki od primjera uporabivosti jezičnih tehnologija koje se očekuju ne samo kao istraživački prototipovi, nego i kao korisni komercijalni proizvodi. Ne možemo očekivati kako će ih za hrvatski jezik izraditi istraživači koji se bave engleskim, francuskim, njemačkim, češkim, slovenskim ili srpskim, već te jezične resurse, alate i usluge moramo razviti sami. Međutim, utoliko će nam biti lakše ako te napore uskladimo i koordiniramo sa sličnim takvim naporima za druge EU jezike, a upravo tome služi inicijativa opisana u ovoj tiskovini.

Ova bijela knjiga o hrvatskome jeziku pokazuje kako u Hrvatskoj postoji temeljno okružje za istraživanje jezičnih tehnologija, međutim to do sada nije rezultiralo i razvojem jezične industrije. Unatoč tome što su za hrvatski izrađeni neki jezični resursi i tehnologije, znatno ih je manje nego za druge slavenske jezike, npr. češki, a još ih je manje razvijeno u usporedbi s većim europskim jezicima kao što su engleski, njemački ili francuski.

Premda u Hrvatskoj postoji već polustoljetna tradicija istraživanja na području računalnoga jezikoslovlja, računalne obradbe teksta i korpusne lingvistike (uz nastanak tako značajnih resursa kao što su Hrvatski čestotni rječnik, Hrvatski nacionalni korpus, Hrvatsko-engleski usporedni korpus, Hrvatski morfološki leksikon, Hrvatska ovisnosna banka stabala, itd.), ne može se reći da je sadašnje stanje jezičnih tehnologija zadovoljavajuće. Uz nacionalno podupirane projekte, koji su na žalost još uvijek malobrojni, od 2008. započinje se ozbiljnija potpora kroz pet projekata Europske komisije: CLARIN, ACCURAT, LetsMT!, ATLAS, XLike; ali i oni su mahom usmjereni na rješavanje pojedinačnih problema ili pružanja tehnoloških rješenja, a rijetko na ukupnost jezičnih tehnologija za hrvatski jezik. Tu ulogu za hrvatski jezik preuzima šesti projekt – CESAR – kao i šira META-NET inicijativa, stvaranjem ove bijele knjige.

Prema procjenama podrobnije iznesenim u ovome izvješću, potrebno je poduzeti niz ciljanih mjera kako bi se hrvatski jezični resursi i alati doveli na istu razinu razvijenosti glede njihove kakvoće i količine, kakva je razina već dosegnuta za druge europske jezike.

Vizija META-NET-a su visokokvalitetne jezične tehnologije za sve jezike koje podupiru političko i gospodarsko jedinstvo kroz kulturnu raznolikost. Ove će tehnologije pomoći u uklanjanju prepreka i u izgradnji mostova između jezika u Europi. To, međutim, traži od svih dionika ovoga procesa – politike, istraživanja, gospodarstva i društva u cjelini – objedinjavanje svojih napora u budućnosti.

Ovaj niz bijelih knjiga nadopunjuje ostale strateške aktivnosti koje poduzima META-NET. Najnovije obavijesti, kao što su trenutačna inačica vizije META-NET-a [2] ili Strateški istraživački plan (SIP) može se pronaći na META-NET-ovim mrežnim stranicama: http://www.meta-net.eu.

2

JEZICI U OPASNOSTI: IZAZOV ZA JEZIČNE TEHNOLOGIJE

U ovome trenutku svjedočimo digitalnoj revoluciji koja korjenito utječe na našu komunikaciju i naše društvo. Najnoviji razvoj digitalnih i mrežnih komunikacijskih tehnologija ponekad se uspoređuju s Gutenbergovim izumom tiska pomičnim slovima. Što nam ta analogija može reći o budućnosti europskoga informacijskoga društva i o našim vlastitim jezicima?

Digitalna revolucija usporediva je s Gutenbergovim izumom tiska pomičnim slovima.

Nakon Gutenbergova izuma pravi su proboji u komunikaciji i razmjeni znanja postignuti pothvatima kao što je Lutherov prijevod Biblije na narodni jezik (ili u hrvatskome slučaju, glagoljički prvotisak Misala iz 1483. kao prve tiskanje knjige na hrvatskome jeziku). U nadolazećim stoljećima razvijeni su razni kulturni postupci koji su omogućili obradbu jezika i razmjenu znanja:

- pravopisno i gramatičko normiranje većih jezika omogućilo je brzu razmjenu novih znanstvenih ideja;
- uspostavljanje službenih jezika omogućilo je građanima komu-ni¬kaciju unutar određenih (često političkih) granica;
- poučavanje jezika i prevođenje omogućilo je razmjenu preko jezičnih granica;
- stvaranje uredničkih i bibliografskih normi osiguralo je kakvoću tiskovina;
- stvaranjem različitih medija kao što su knjige, novine, radio, televizija i drugi, zadovoljavaju se komunikacijske potrebe pučanstva.

U zadnjih je dvadeset godina informacijska tehnologija omogućila olakšavanje i automatizaciju mnogih procesa:

- računalna priprema za tisak zamijenila je tipkanje i grafički slog;
- Microsoft PowerPoint zamijenio je projiciranje s prozirnica;
- e-pošta omogućuje odašiljanje i primanje dokumenata brže od telefaks uređaja;
- Skype nudi jeftine internetske telefonske pozive i održavanje virtualnih sastanaka;
- zajednički formati zapisa zvučnih i vizualnih podataka omogućuju jednostavnu razmjenu multimedijskih sadržaja;
- tražilice omogućuju pristup www-stranicama na temelju pretrage uporabom ključnih riječi;
- mrežne usluge poput Google prevoditelja nude brze, ali zato približne prijevode;
- društvene mreže kao što su Facebook, Twitter i Google+ pospješuju komunikaciju, omogućuju suradnju i dijeljenje obavijesti.

Premda su takve aplikacije i usluge višestruko korisne, ipak još ne mogu podupirati u cijelosti održivo, više-