META-NET

White Paper Series Серия Бели книги

THE BULGARIAN LANGUAGE IN THE DIGITAL AGE

БЪЛГАРСКИЯТ ЕЗИК В ДИГИТАЛНАТА ЕПОХА

Diana Blagoeva
Svetla Koeva
Vladko Murdarov

Springer

White Paper Series Серия Бели книги

THE BULGARIAN LANGUAGE IN THE DIGITAL AGE

БЪЛГАРСКИЯТ ЕЗИК В ДИГИТАЛНАТА ЕПОХА

Diana Blagoeva Bulgarian Academy of Sciences
Svetla Koeva Bulgarian Academy of Sciences
Vladko Murdarov Bulgarian Academy of Sciences

Georg Rehm, Hans Uszkoreit
(редактори, 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 978-3-642-30167-4 ISBN 978-3-642-30168-1 (eBook)
DOI 10.1007/978-3-642-30168-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940341

ПРЕДГОВОР PREFACE

Бялата книга е част от серия документи, представящи развитието в областта на езиковите технологии и техния потенциал. Документите са предназначени за преподаватели, журналисти, политици, различни езикови общности и т. н. Достъпът до езикови технологии за езиците, които се говорят в Европа, е много различен. Затова и необходимите действия за подкрепа на изследванията и развитието на езиковите технологии също са различни. Те зависят от много фактори, например сложността на даден език и броя на неговите носители.

META-NET, мрежа за високи постижения, изградена с подкрепата на Европейската комисия, предлага анализ на съществуващите езикови ресурси и технологии в серията Бели книги (p. 81). Анализът е съсредоточен върху 23-те официални европейски езика, наред с други по-важни национални и регионални езици. Резултатите показват значителен недостиг за всеки език. Задълбоченият експертен анализ и оценка на актуалната ситуация ще помогнат за увеличаване на ефекта от изследванията и намаляване на риска от пропуски.

В META-NET участват 54 изследователски центъра от 33 страни (p. 77), работещи съвместно с търговски и правителствени организации, научни институции, софтуерни компании, фирми за информационни технологии и европейски университети. Те разработват заедно визия за технологично развитие и стратегическа програма за научни изследвания, които показват как езиковите технологии могат да отговорят на научните предизвикателства до 2020 г.

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. 81). 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. 77). 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.

Авторите на този документ благодарят сърдечно на авторите на Бялата книга за немски за предоставената възможност да използват избрани езиково независими части [1].

Разработката на настоящата Бяла книга е финансирана по Седма рамкова програма и Програма ICT Policy Support Programme на Европейската комисия, договори T4ME (договор за финансиране 249119), CESAR (договор за финансиране 271022), METANET4U (договор за финансиране 270893) и META-NORD (договор за финансиране 270899).

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

СЪДЪРЖАНИЕ CONTENTS

ЕЗИЦИТЕ В ЕВРОПЕЙСКОТО ЕЗИКОВО ИНФОРМАЦИОННО ОБЩЕСТВО

THE BULGARIAN LANGUAGE IN THE DIGITAL AGE

РЕЗЮМЕ

Информационните технологии променят живота ни. Ежедневно използваме компютри, за да пишем, редактираме, изчисляваме, търсим информация, и все по-често за да четем, слушаме музика, да разглеждаме снимки или да гледаме филми. Носим малки компютри в джобовете си и ги използваме, за да позвъним, напишем имейл, да получим необходима информация или просто за да се позабавляваме, където и да се намираме. Как широкообхватната дигитализация на информация, знание и ежедневна комуникация влияе върху езика ни? Дали езикът ни ще се промени и дали е възможно дори да изчезне?

Компютрите ни са свързани един с друг в непрекъснато развиваща се компактна и мощна глобална мрежа. Момиче от Ипанема, ученик от Жеравна и инженер от Катманду могат да общуват с приятелите си във Фейсбук, но е малко вероятно да се срещнат един с друг в онлайн общности или форуми. Ако ги интересува как да излекуват главоболието си, те могат да потърсят повече информация в Уикипедия, но дори и тогава няма да прочетат една и съща статия. Когато европейските потребители на интернет обсъждат във форуми и чатове последствията от катастрофата във Фукушима за европейската енергийна политика, те го правят в ясно разграничени езикови общности. Всичко, което интернет съдържа и предлага, е все още разделено от езика на потребителите. Винаги ли ще бъде така? В научно-фантастичните филми всички говорят един език – английски, китайски или български – в зависимост от това, къде се излъчва филмът. Възможно ли е езикът на космонавтите да бъде български, въпреки че те рядко биха употребявали български думи толкова естествено, колкото английски? Много от съществуващите в момента 6 000 езика едва ли ще оцелеят в глобализираното дигитално информационно общество. Предполага се, че поне 2 000 езика са обречени на изчезване в идните десетилетия. Други ще продължат да играят роля на семейно и регионално ниво, но не и в по-широки делови или академични кръгове. Какви са шансовете на българския език да оцелее? Българският език се говори от близо 9 милиона души предимно в България, но също и в Гърция, Македония, Румъния, Турция (европейската част), Украйна, Австралия, Канада, САЩ, Германия и Испания. За малка страна като България съществува относително голямо количество телевизионни канали на български език – седем национални телевизии, 16 кабелни и сателитни телевизии с многорегионално покритие и 46 – с регионално покритие. Повечето чужди филми са дублирани на български език. Книгите се връщат на мода, въпреки констатациите, че през последните години българинът е спрял да се интересува от литература. Българският е първият славянски език, който разполага със своя собствена писмена система, датираща от 9-ти век. На 1 януари 2007 г., когато България е приета за пълноправен член на Европейския съюз, кирилицата става третата официална азбука на Европейския съюз след латинската и гръцката. Някои среди изразяват недоволство от нарастващата употреба на чужди думи, особено английски, и дори съществуват

страхове, че българският език ще се „прояде" от множество английски думи и изрази. През вековете българският език е устоял на влиянието на думи и термини от гръцки и латински – езиците на познанието, както и на навлизането на френски думи през 18-ти и 19-ти век. Добро противодействие срещу изчезването на обичаните от нас български думи е наистина да ги използваме – често и съзнателно. Главното ни притеснение не трябва да е нарастващото английско влияние върху езика, а пълното му изчезване от някои основни области на личния ни живот. Нито науката, нито авиацията или глобалният финансов пазар се нуждаят от език, разпространен по целия свят – lingua franca. В много области на живота е по-важно общуването с гражданите на страната, отколкото с международните партньори – вътрешната политика, например административните процедури, правото, културата и търговията. Статутът на езика зависи не само от броя на неговите носители, създадените книги и филми, телевизионните канали, които го използват, но и от присъствието на езика в дигиталното информационно пространство и софтуерните приложения. В това отношение българският език е относително добре представен: всички важни международни софтуерни продукти са локализирани за български, българската Уикипедия е на 34-а позиция сред 270 в света. Потребителите на интернет в България през 2009 г. са се увеличили с 31% спрямо 2007 г. и вече са 46% от цялото население. В областта на езиковите технологии за български също съществуват редица продукти, технологии и ресурси. Има приложения за възпроизвеждане на реч, проверка на правописа и граматиката. Съществуват и програми за автоматичен превод, макар че не винаги се предлагат лингвистично коректни преводи, особено когато преводът е от друг език на български. Това се дължи основно на специфичните езикови характеристики на българския език. Информационните и комуникационните технологии се подготвят за следваща революционна стъпка. След персоналните компютри, мрежите, миниатюризацията на техниката, мултимедията, мобилните устройства и паралелната обработка на информация, идва епохата на технологии, които ще разбират не просто букви или звукове, но и словосъчетания и изречения. Така те ще подпомагат в много по-голяма степен потребителите, тъй като ще говорят, знаят и разбират техния език. Пионери в тази сфера са например Гугъл преводачът, който предлага безплатен онлайн автоматичен превод между 57 езика, супер компютърът на IBM Watson, който победи шампиона на САЩ в играта „Jeopardy" или мобилният асистент Siri на iPhone, който реагира на гласови команди и отговаря на въпроси на английски, немски, френски и японски.

Следващото поколение информационни технологии ще се усъвършенства в употребата на естествения език до такава степен, че потребителите ще общуват, използвайки технологиите на собствения си език. Устройствата ще могат автоматично да намерят най-важните новини и информация в световното дигитално изобилие от познание само с помощта на гласови команди. Езиковите технологии ще предлагат автоматичен превод или ще подпомагат превода, ще осигуряват резюмиране на диалог или на различни документи, а компютърно подпомогнатото обучение ще съдейства за по-лесното интегриране на малцинствени групи и чужденци. Следващото поколение информационни и комуникационни технологии ще създаде индустриални и обслужващи роботи (в момента все още в научните лаборатории), които точно ще разбират какво искат техните потребители и ще рапортуват за изпълнението на задачите си. Такова равнище на работа надхвърля простите множества от символи и речници, програми за проверка на правописа и правила за