META-NET

White Paper Series Seria de studii

THE ROMANIAN LANGUAGE IN THE DIGITAL AGE

LIMBA ROMÂNĂ ÎN ERA DIGITALĂ

Diana Trandabăț
Elena Irimia
Verginica Barbu Mititelu
Dan Cristea
Dan Tufiș

Springer

White Paper Series Seria de studii

THE ROMANIAN LANGUAGE IN THE DIGITAL AGE

LIMBA ROMÂNĂ ÎN ERA DIGITALĂ

Diana Trandabăț [1, 2]
Elena Irimia [3]
Verginica Barbu Mititelu [3]
Dan Cristea [1, 2]
Dan Tufiș [3]

[1] University "Alexandru Ioan Cuza" of Iași
[2] Romanian Academy, Institute of Computer Science
[3] Romanian Academy, Research Institute for AI

Georg Rehm, Hans Uszkoreit
(editori, 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 978-3-642-30702-7 ISBN 978-3-642-30703-4 (eBook)
DOI 10.1007/978-3-642-30703-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012942726

PREFAȚĂ

Acest studiu face parte dintr-o serie de studii care promovează cunoașterea tehnologiilor limbajului și a potențialului lor. El se adresează jurnaliștilor, politicienilor, comunităților lingvistice și tuturor celor interesați de limba română. În Europa, disponibilitatea și utilizarea tehnologiilor limbajului variază de la o limbă la alta. În consecință, sunt necesare și acțiuni diferite pentru a sprijini în continuare cercetarea și dezvoltarea acestor tehnologii. Acțiunile necesare depind de mai mulți factori, cum ar fi complexitatea unei anumite limbi sau dimensiunea comunității care o folosește. META-NET, o rețea de excelență finanțată de Comisia Europeană, a efectuat o analiză a resurselor și tehnologiilor lingvistice actuale prin intermediul studiilor de față (vezi lista lor la pag. 87). Această analiză s-a concentrat pe cele 23 de limbi oficiale ale Uniunii Europene, precum și asupra altor limbi naționale și regionale importante din Europa. Rezultatele acestei analize indică faptul că există un deficit enorm în sprijinirea tehnologiei și lacune de cercetare semnificative pentru fiecare limbă. Analiza detaliată prezentată și evaluările experților vor contribui la maximizarea impactului cercetărilor ulterioare. META-NET este formată din 54 de centre de cercetare din 33 de țări (în luna noiembrie 2011, vezi pag. 83), care colaborează cu persoane cheie din domeniul afacerilor (companii de software, furnizori de tehnologie, utilizatori), din agenții guvernamentale, organizații de cercetare, organizații nonguvernamentale, comunități lingvistice și universități europene. Împreună cu aceste comunități, META-NET dezvoltă o viziune comună asupra tehnologiei și o agendă strategică de cercetare pentru o Europă multilingvă la nivelul anului 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. 87). 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. 83). 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.

Autorii acestui document sunt recunoscători autorilor studiului pentru limba germană, care le-au permis să (re)folosească în prezentul document anumite materiale independente de limbă [1].

Acest studiu a fost finanțat prin Programul Cadru nr. 7 și prin Programul de sprijinire a politicii în domeniul Tehnologiilor Informației și Comunicațiilor (ICT Policy Support Programme) al Comisiei Europene prin proiectele T4ME (contract nr. 249 119), CESAR (contract nr. 271 022), METANET4U (contract nr. 270 893) și META-NORD (contract nr. 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).

CUPRINS CONTENTS

LIMBA ROMÂNĂ ÎN ERA DIGITALĂ

THE ROMANIAN LANGUAGE IN THE DIGITAL AGE

REZUMAT

În ultimii 60 de ani, Europa a devenit o structură politică și economică distinctă, păstrându-și însă diversitatea culturală și lingvistică. Acest lucru înseamnă că, de la portugheză la poloneză și de la italiană la islandeză, comunicarea de zi cu zi între cetățenii europeni, precum și comunicarea din domeniile economic și politic, se confruntă inevitabil cu barierele lingvistice. Instituțiile Uniunii Europene cheltuiesc aproximativ un miliard de euro pe an pentru menținerea politicii lor asupra multilingvismului, de exemplu, traducerea textelor și interpretarea discursurilor. Trebuie însă să fie multilingvismul o astfel de povară? Tehnologiile moderne ale limbajului și cercetarea lingvistică pot avea o contribuție semnificativă la reducerea acestor frontiere lingvistice. Combinate cu dispozitive și aplicații inteligente, tehnologiile limbajului vor fi în măsură în viitor să-i ajute pe europeni să comunice cu ușurință unul cu altul și să facă afaceri împreună, chiar dacă nu vorbesc aceeași limbă.

Tehnologiile limbajului construiesc punți de legătură pentru viitorul Europei.

Tehnologia informației ne schimbă viața de zi cu zi. Scriem deja folosind calculatorul, editām, facem calcule, căutăm informații, dar și, din ce în ce mai des, citim, ascultăm muzică, vedem fotografii și urmărim filme pe calculator. Purtăm calculatoare mici în buzunare și le utilizăm pentru a efectua apeluri telefonice, a scrie e-mailuri, pentru a obține informații de pe Internet și pentru a ne ține de urât, oriunde ne-am afla. În ce mod este afectată limba română de această digitalizare masivă a informațiilor, cunoștințelor și comunicării de zi cu zi? Se va schimba ea sau chiar va dispărea?

Toate calculatoarele noastre sunt legate într-o rețea globală din ce în ce mai densă și puternică. Fata din Buenos Aires, ofițerul vamal din Constanța și inginerul din Katmandu pot discuta cu prietenii lor de pe Facebook, dar este puțin probabil să se întâlnească în comunitățile online și pe forumuri. Dacă vor să afle cum pot trata un țiuit în urechi, probabil vor căuta un răspuns pe Wikipedia, dar chiar și atunci ei nu vor citi același articol. Când internauții Europei discută în forumuri și pe chat efectele accidentului nuclear Fukushima asupra politicii energetice europene, ei fac acest lucru în comunități lingvistice distincte. Deși Internetul conectează, există încă o separare evidentă în funcție de limba folosită de fiecare utilizator. Va fi mereu așa?

Tehnologiile limbajului – cheia spre viitor.

În filmele SF, toată lumea vorbește aceeași limbă. Ar putea fi româna, chiar dacă am avut doar un singur astronaut român? Multe dintre cele 6.000 de limbi nu vor supraviețui într-o societate a informațiilor digitale globale. Se estimează că cel puțin 2.000 de limbi sunt condamnate la dispariție în deceniile următoare. Altele vor continua să joace un rol important în familii și în zone restrânse, dar nu și în lumea academică sau în lumea afacerilor. Care sunt șansele de supraviețuire a limbii române?

Vorbită de aproximativ 29.000.000 de vorbitori în întreaga lume, limba română este prezentă nu doar în cărți,

filme sau canale TV, ci și în spațiul informațional digital. Piața Internetului în România este în continuă creștere. Din ce în ce mai mulți români au acces la un calculator acasă, fiind și utilizatori de Internet. Domeniul *.ro* înregistrează 0.4% din paginile web existente în acest moment, comparabil cu domeniul *.eu*.

Limba română prezintă un număr de caracteristici specifice care contribuie la bogăția limbii, dar care pot fi, de asemenea, o provocare pentru prelucrarea computațională a limbajului natural.

Instrumentele de traducere automată și de prelucrare a vorbirii disponibile în prezent pe piață sunt încă departe de standardele la care se așteaptă să ajungă. Actorii dominanți în domeniu sunt, în principal, întreprinderi private cu sediul în America de Nord, axate pe profit. De la sfârșitul anilor 1970, Uniunea Europeană a înțeles importanța tehnologiilor lingvistice ca motor al unității europene și a început finanțarea primelor proiecte de cercetare, cum a fost EUROTRA. În același timp, au fost inițiate proiecte naționale, care au generat rezultate valoroase, dar nu au condus niciodată la acțiuni concertate la nivel european. În contrast cu acest efort de finanțare extrem de selectiv, alte societăți multilingve, cum ar fi India (cu 22 de limbi oficiale) și Africa de Sud (cu 11 limbi oficiale) au înființat de curând programe naționale pe termen lung de cercetare a limbii și dezvoltare tehnologică.

Există unele îngrijorări privind utilizarea din ce în ce mai largă a anglicismelor, și unii lingviști chiar se tem că limba română va fi sufocată de cuvinte și expresii în limba engleză. Studiul nostru indică totuși că această îngrijorare nu este fondată.

Similar procesului de relatinizare din secolul al XIX-lea, de după eliberarea de sub dominația greacă și otomană, limba română a parcurs, în ultimii douăzeci de ani, un proces de trecere de la limbajul totalitar („limba de lemn", discursul unidirecțional etc.) la utilizarea deschisă, în care noi modele lingvistice trebuie să se adapteze la tranziția socială și culturală. Astfel, asemănător multor altor limbi, româna traversează un proces continuu de internaționalizare, sub influența vocabularului anglo-saxon.

Principala noastră grijă nu ar trebui să fie anglicizarea treptată a limbii române, ci dispariția sa completă din domeniile majore ale vieții noastre personale. Nu îngrijorează domenii precum științele, aviația și piețele financiare mondiale, care chiar au nevoie de o *lingua franca* la nivel mondial, ci multe domenii ale vieții de zi cu zi, în care este mult mai important să fii aproape de cetățenii unei țări decât de partenerii internaționali, cum sunt, de exemplu, politicile interne, procedurile administrative, dreptul sau cultura.

Tehnologia informației și comunicației se pregătește acum pentru următoarea revoluție. După calculatoare personale, rețele, miniaturizare, multimedia și dispozitive mobile, următoarea generație de tehnologie va include programe care înțeleg nu doar litere și sunete vorbite sau scrise, ci cuvinte și fraze întregi, și care vin în sprijinul utilizatorului pentru că vorbesc și înțeleg limba lui. Precursorii acestei evoluții sunt serviciul online gratuit Google Translate, care traduce din și spre 57 de limbi, Watson, supercomputerul IBM care a fost capabil să-l învingă pe campionul SUA în jocul „Jeopardy", dar și Siri, asistentul mobil de la Apple pentru iPhone, care poate reacționa la comenzi vocale și poate răspunde la întrebări în limbile engleză, germană, franceză și japoneză.

Următoarea generație de tehnologii informaționale vor stăpâni limbajul uman într-o asemenea măsură, încât utilizatorii umani vor fi capabili să comunice folosind tehnologia în propria lor limbă. Dispozitivele vor fi capabile să găsească în mod automat, la simpla solicitare a utilizatorului printr-o comandă vocală, cele mai importante știri și informații de la magazinul digital de cunoștințe. Tehnologiile bazate pe limbaj vor fi capabile să traducă automat sau să asiste interpreții, să rezume con-

versații și documente, dar și să asiste activ utilizatorii în procesul de învățare.

Noile tehnologii informaționale și de comunicații vor permite roboților industriali și de servicii (în curs de dezvoltare în prezent în laboratoarele de cercetare) să înțeleagă cu exactitate ceea ce utilizatorii își doresc de la ei și apoi să raporteze cu mândrie realizările lor.

Acest nivel de performanță presupune să trecem cu mult dincolo de simple seturi de caractere și lexicoane, programe de corectare a limbii și reguli de pronunție. Tehnologia trebuie să depășească abordările simpliste și să înceapă să modeleze limbajul într-un mod atotcuprinzător, luând în considerare deopotrivă sintaxa și semantica pentru a înțelege întrebări și a genera răspunsuri complete și relevante.

În cazul limbii române, cercetările din universități și institute de cercetare din România și Republica Moldova au dus la dezvoltarea de sisteme de înaltă calitate, precum și modele și teorii aplicabile pe scară largă. Cu toate acestea, domeniul de aplicare al resurselor, precum și gama de instrumente sunt încă foarte limitate în raport cu resursele și instrumentele existente pentru limba engleză și nu sunt suficiente din punct de vedere calitativ și cantitativ pentru a dezvolta tehnologiile necesare sprijinirii unei societăți a cunoașterii cu adevărat multilingve. Subdezvoltarea care se resimte în zona resurselor lingvistice (cantitativă și calitativă) îngreunează enorm eforturile de dezvoltare a tehnologiilor limbajului și a aplicațiilor.

Tehnologiile limbajului ajută la unificarea Europei.

O situație neclară din punct de vedere juridic restricționează utilizarea textelor digitale, cum ar fi cele publicate on-line de ziare, pentru cercetări empirice lingvistice și pentru tehnologiile limbajului, de exemplu pentru construirea modelelor statistice de limbă. Împreună cu politicienii și factorii de decizie politică, cercetătorii ar trebui să poată contribui la stabilirea unor legi sau reglementări care să le permită să utilizeze textele puse la dispoziția publicului pentru activități de cercetare și dezvoltare legate de limbaj.

Se observă, de asemenea, o lipsă a continuității în finanțarea cercetării și dezvoltării. Programe coordonate pe termen scurt tind să alterneze cu perioade de finanțare insuficientă sau deloc. În plus, există în general o slabă coordonare cu programe din alte țări ale UE și la nivelul Comisiei Europene (cum se întâmplă, de exemplu, cu programele PSP-ICT, care au ca protagoniști și universități din România, dar care nu sunt sprijinite de guvern pentru asigurarea coerentă a cofinanțării). Nevoia de mari cantități de date și complexitatea extremă a sistemelor ce folosesc tehnologia limbajului fac să fie vitală dezvoltarea unei noi infrastructuri și a unei organizări mai coerente a finanțării cercetării în domeniul tehnologiilor limbajului natural, dacă dorim să putem spera la folosirea noii generații de tehnologii ale comunicării și informației în domeniile vieții private sau publice în care vorbim în limba română.

În concluzie, putem considera că deocamdată limba română nu este în pericol. Cu toate acestea, întreaga situație s-ar putea schimba dramatic atunci când o nouă generație de tehnologii începe să stăpânească într-adevăr eficient limbajul uman. Prin îmbunătățiri în traducerea automată, tehnologia limbajului va ajuta la depășirea barierelor lingvistice, dar va fi capabilă să opereze doar între acele limbi care au reușit să supraviețuiască în lumea digitală. Dacă este disponibilă o tehnologie adecvată a limbajului, atunci aceasta va fi în măsură să asigure supraviețuirea limbii, altfel, chiar și limbile „mai mari" vor intra sub o presiune severă.

Dacă ne bazăm pe experiența dobândită până acum, tehnologiile „hibride" de astăzi ale limbajului, care combină prelucrări de adâncime cu metode statistice, par să fie capabile să elimine decalajul dintre limbile europene.