

UN RISC PENTRU LIMBILE NOASTRE: O PROVOCARE PENTRU TEHNOLOGIA LIMBAJULUI

Suntem martorii unei revoluții digitale care are un impact dramatic asupra comunicării și societății. Dezvoltările recente din tehnologia informațiilor și comunicațiilor digitale sunt uneori comparate cu inventarea tiparului de către Gutenberg. Ce ne poate spune această analogie despre viitorul societății informaționale europene în general, și despre viitorul limbilor noastre în particular?

Revoluția digitală este comparabilă cu inventarea tiparului de către Gutenberg.

Ulterior invenției lui Gutenberg au avut loc progrese reale în comunicare și în schimbul de informație, datorită unor eforturi precum traducerea textelor religioase în limba enoriașilor. În secolele următoare, au fost dezvoltate tehnici culturale pentru a îmbunătăți prelucrarea limbajului și schimbul de cunoștințe:

- Standardizarea ortografică și gramaticală a limbilor importante a permis diseminarea rapidă a noilor idei culturale și științifice.
- Dezvoltarea limbilor oficiale a făcut posibilă comunicarea dintre cetățeni în interiorul anumitor granițe (adeseori politice).
- Predarea și traducerea limbilor străine au facilitat schimburile dintre limbi.

- Crearea de principii jurnalistice și bibliografice a asigurat calitatea și disponibilitatea materialelor impri- mate.
- Crearea diferitelor tipuri de media, precum ziarele, radioul, televiziunea, cărțile etc. a satisfăcut nevoia de comunicare.

În ultimii 20 de ani, tehnologia informației a contribuit la automatizarea și facilitarea mai multor procese:

- software-ul pentru tehnoredactare computerizată înlocuiește acum dactilografierea și culegerea textelor.
- Microsoft PowerPoint înlocuiește retroproiectorul.
- Serviciile de e-mail permit trimiterea și primirea de documente mai rapid decât folosind un fax.
- Skype permite convorbiri prin Internet și găzduiește întâlniri virtuale.
- Formatele de codificare audio și video ușurează schimbul de conținut multimedia.
- Motoarele de căutare oferă acces bazat pe cuvinte cheie la un număr din ce în ce mai mare de pagini web.
- Serviciile online precum Google Translate produc traduceri rapide, chiar dacă aproximative.
- Platformele sociale de media, precum sunt Facebook, Twitter și Google+, facilitează colaborarea și partajarea de informații.

Deși astfel de instrumente și aplicații sunt utile, acestea nu sunt suficiente pentru a implementa o societate informațională europeană multilingvă și sustenabilă, în care informația și bunurile să poată circula liber.

2.1 FRONTIERELE LINGVISTICE FRÂNEAZĂ CREAREA UNEI SOCIETĂȚI INFORMAȚIONALE EUROPENE

Nu putem ști cu precizie cum va arăta viitoarea societate informațională. Dar există o mare probabilitate ca revoluția în tehnologia comunicațiilor să faciliteze apropierea între oameni, vorbitori de limbi diferite, în noi moduri. Necesitatea de a comunica forțează oamenii să învețe limbi străine noi, iar pe dezvoltatori îi obligă să creeze noi aplicații tehnologice pentru a asigura înțelegerea reciprocă și accesul la cunoștințele comune. Este limpede că progresul societății impune acum o calitate a comunicării diferită de cea de acum câțiva ani.

Economia și spațiul informațional global ne confruntă cu mai multe limbi, mai mulți vorbitori, mai mult conținut.

Într-un spațiu economic și informațional global, suntem confrunțați cu mai multe limbi, mai mulți vorbitori și mai mult conținut și suntem nevoiți să interacționăm rapid cu noi tipuri de media. Popularitatea actuală a mediilor sociale (Wikipedia, Facebook, Twitter și YouTube) reprezintă doar vârful aisbergului.

Astăzi putem recepționa gigaocteți de text din orice colț al planetei în câteva secunde, doar pentru a afla că textul este într-o limbă pe care nu o înțelegem. Potrivit unui raport recent solicitat de Comisia Europeană, 57% dintre utilizatorii de Internet din Europa achiziționează bunuri și servicii în limbi diferite de cea maternă (engleza

este cea mai cunoscută limbă străină, urmată de franceză, germană și spaniolă). 55% dintre utilizatori citesc într-o limbă străină, în timp ce doar 35% utilizează o altă limbă pentru a scrie e-mail-uri sau a publica comentarii pe web [3]. Cu câțiva ani în urmă, engleza era privită ca lingua franca (limba de lucru) a Internetului – o vastă majoritate a conținutului era scrisă în această limbă – dar situația s-a schimbat drastic acum. Cantitatea de conținut online în alte limbi ne-europene (precum cele asiatice sau cele din Orientul Mijlociu) a explodat.

În mod surprinzător, diviziunea digitală accentuată datorată frontierelor lingvistice nu a câștigat încă prea multă atenție în discursul public; totuși, ea ridică o întrebare foarte presantă: „Care dintre limbile europene vor prospera în societatea virtuală a informației și cunoașterii și care sunt sortite dispariției?”

2.2 LIMBILE NOASTRE SUNT ÎN PERICOL

Tiparul, deși a contribuit la un inestimabil schimb de informații în Europa, a condus de asemenea la extincția multora dintre limbile europene. Limbile regionale și minoritare au fost tipărite arareori, iar limbi precum dalmata sau limba din Cornwall au cunoscut doar forme orale de transmitere, care le-au restricționat adoptarea, răspândirea și utilizarea. Va avea Internetul același efect asupra limbilor noastre?

Cele aproximativ 80 de limbi vorbite astăzi în Europa reprezintă unul dintre cele mai bogate și importante bunuri culturale ale sale, dar și o componentă importantă a modelului său social unic [4]. În timp ce limbi populare precum engleza sau spaniola vor rămâne cu siguranță prezente pe piața digitală emergentă, multe limbi europene ar putea fi deconectate de la comunicarea digitală și ar putea deveni irelevante pentru societatea Internetului. O astfel de evoluție ar slăbi poziția Europei pe piața globală și ar fi în contradicție cu obiectivul stra-

tegic de asigurare a participării egale a fiecărui cetățean european, indiferent de limba lui.

Marea varietate de limbi ale Europei este unul dintre cele mai importante bunuri culturale ale sale și o componentă esențială a succesului său social.

Potrivit unui raport recent al UNESCO privind multilingvismul, limbile reprezintă un mediu esențial pentru exercitarea drepturilor fundamentale precum exprimarea politică, educația și participarea în societate [5].

2.3 TEHNOLOGIA LIMBAJULUI ESTE CHEIA ACTIVĂRII TEHNOLOGIEI

În trecut, eforturile de investiții financiare în preservarea limbilor s-au concentrat asupra educației lingvistice și a traducerii. De exemplu, potrivit anumitor estimări, piața europeană de traducere, interpretare, localizare de software și globalizare a paginilor de Internet a fost estimată la 8,4 miliarde € în 2008 și este de așteptat să crească cu 10% pe an [6]. Totuși, această cifră acoperă doar o mică parte din nevoile curente și viitoare în comunicarea dintre cetățeni. Soluția cea mai convingătoare pentru a asigura amploarea și extinderea utilizării limbilor în Europa de mâine este de a utiliza tehnologiile adecvate, așa cum folosim tehnologiile pentru transport, energie sau alte nevoi.

Tehnologiile limbajului (care acoperă toate formele de texte scrise și discursuri rostite) pot ajuta oamenii să colaboreze, să conducă afaceri, să împărtășească cunoștințe și să participe în dezbateri politice și sociale, independent de barierele lingvistice sau de competențele de lucru cu calculatorul. Tehnologia limbajului operează de obicei în culise, în cadrul unor sisteme complexe, care ne ajută, de exemplu:

- să găsim informații cu un motor de căutare pe Internet;
- să verificăm ortografia și corectitudinea gramaticală cu un editor de texte;
- să vizualizăm recomandări de produse oferite într-un magazin virtual;
- să ascultăm instrucțiunile unui sistem de navigație;
- să traducem pagini web cu un serviciu online.

Tehnologia limbajului constă într-o serie de aplicații de bază care activează procese auxiliare în cadrul unei aplicații mai complexe. Scopul realizării seriei de studii în cadrul proiectului METANET este să descopere cât de avansate sunt aceste tehnologii pentru fiecare dintre limbile europene.

Europa are nevoie de tehnologii ale limbajului robuste și accesibile, adaptate tuturor limbilor europene.

Pentru a-și menține poziția în prima linie a inovării globale, Europa are nevoie de tehnologii ale limbajului adaptate tuturor limbilor europene, care să fie robuste, accesibile financiar și bine integrate în medii software complexe. Experiența utilizatorului cu mediul virtual în regim interactiv, multimedia și multilingv nu este posibilă fără tehnologia limbajului.

2.4 OPORTUNITĂȚI ALE TEHNOLOGIEI LIMBAJULUI

În lumea tiparului, realizarea tehnologică proeminentă a constat în copierea rapidă a imaginii unei pagini de text folosind un dispozitiv de tipărire. Oamenilor le-a rămas munca grea de a căuta, citi, traduce și rezuma cunoștințe. A trebuit să așteptăm până la Edison pentru a înregistra limba vorbită și, din nou, tehnologia lui a făcut pur și simplu copii analogice.

Tehnologia limbajului digital permite dezvoltarea de aplicații precum traducerea automată, generarea de conținut, procesarea informației și managementul cunoștințelor pentru toate limbile europene. De asemenea, ea poate îmbogăți cu interfețe intuitive, bazate pe limbaj, dispozitive electrocasnice, utilaje, vehicule, computere și roboți. Deși există deja multe prototipuri, aplicațiile comerciale și industriale sunt încă în stadii incipiente de dezvoltare. Realizările recente din cercetare și dezvoltare au creat o adevărată avalanșă de oportunități de aplicare a tehnologiei limbajului (TL). De exemplu, traducerea automată (TA) oferă o acuratețe rezonabilă pentru domenii specifice, iar o serie de aplicații experimentale pot asigura managementul informației și cunoștințelor, precum și producerea de conținut în multe din limbile europene.

Tehnologiile limbajului ajută la depășirea „handicapului” indus de diversitatea lingvistică europeană.

Ca în majoritatea cazurilor, primele aplicații lingvistice, precum interfețe vocale și sisteme de dialog, au fost dezvoltate pentru domenii foarte specializate și prezintă adeseori performanțe limitate. Există oportunități uriașe de piață în sectorul educației și al divertismentului pentru integrarea tehnologiei limbajului în jocuri, site-uri de patrimoniu cultural, oferte „edutainment” (educație prin divertisment), medii de simulare sau programe de formare. Serviciile mobile de informații, software-ul pentru învățarea limbilor străine asistată de calculator, mediile e-learning, instrumentele de auto-evaluare și cele de detectare a plagiatului sunt doar câteva exemple de zone ale aplicațiilor în care tehnologia limbajului poate juca un rol important. Popularitatea aplicațiilor de media socială precum Twitter și Facebook sugerează încă o ocazie în care tehnologii sofisticate ale limbajului sunt necesare pentru monitorizarea publicațiilor, rezumarea discuțiilor, identificarea unor curente

de opinie, detectarea răspunsurilor emoționale, descoperirea încălcărilor drepturilor de autor sau a situațiilor de abuz.

Tehnologiile limbajului reprezintă o oportunitate uriașă pentru Uniunea Europeană, atât din punct de vedere economic, cât și din perspectivă culturală. Multilingvismul a devenit o regulă în Europa. Companiile, organizațiile și școlile europene sunt, de asemenea, multinaționale și diverse. Cetățenii doresc să comunice dincolo de frontierele de limbă care persistă pe Piața Comună Europeană, iar tehnologiile limbajului pot ajuta la depășirea acestor bariere, sprijinind în același timp utilizarea liberă și deschisă a limbilor. Privind chiar mai departe, o tehnologie europeană a limbajului, inovativă și multilingvă, va putea fi un punct de referință pentru partenerii noștri globali și comunitățile lor multilingve. Tehnologiile limbajului pot fi văzute ca o formă de tehnologie „de asistență” care ajută la depășirea „handicapului” indus de diversitatea lingvistică și face comunitățile lingvistice mai accesibile. Un câmp activ de cercetare este reprezentat de tehnologia dedicată operațiilor de salvare în zonele sinistrate. În astfel de medii cu risc înalt, acuratețea comunicării poate fi o problemă de viață și de moarte. Roboți inteligenți cu capacități multilingve au potențialul de a salva vieți.

2.5 PROVOCĂRILE TEHNOLOGIEI LIMBAJULUI

Deși tehnologia limbajului s-a dezvoltat considerabil în ultimii ani, ritmul actual al progresului tehnologic și al inovării este prea lent. Tehnologiile elementare care sunt utilizate pe scară largă, precum opțiunile de corectare gramaticală și ortografică din editoarele de text, sunt de obicei monolingve și sunt disponibile doar pentru câteva limbi.

Serviciile online de traducere automată, deși sunt utile pentru generarea rapidă a unei aproximări rezonabile a

conținutului unui document, întâmpină multe dificultăți atunci când este nevoie de traduceri precise și complete. Datorită complexității limbajului uman, modelarea limbilor noastre în programe software și testarea lor în lumea reală este o întreprindere costisitoare, care necesită angajamente de finanțare susținută. Europa trebuie astfel să își mențină rolul de pionierat în confruntarea cu provocările tehnologice ridicate de o comunitate multilingvă, prin inventarea de noi metode pentru a accelera dezvoltarea. Acestea ar putea include atât noi direcții în tehnici și calcule computaționale, cât și crowdsourcing (exploatarea cunoștințelor maselor).

Ritmul actual al progresului tehnologic
este prea lent.

2.6 ACHIZIȚIA LIMBII DE CĂTRE OM ȘI MAȘINĂ

Pentru a ilustra modul în care computerele prelucrează limbajul și pentru a explica de ce achiziția limbii este o sarcină foarte dificilă, vom arunca o scurtă privire asupra modului în care oamenii achiziționează prima și a doua limbă și apoi asupra modului de funcționare a sistemelor bazate pe tehnologiile limbajului.

Oamenii achiziționează competențe lingvistice în două moduri diferite: învățând din exemple și învățând regulile care stau la baza limbii.

Oamenii achiziționează competențele lingvistice în două moduri distincte. Copiii învață o limbă ascultând interacțiuni dintre părinți, frați sau alți membri ai familiei. La vârsta de aproximativ doi ani, copiii ajung să producă primele lor cuvinte sau fraze scurte. Acest lucru este posibil pentru că oamenii au o predispoziție genetică pentru a imita și a înțelege ceea ce aud.

Învățarea unei a doua limbi presupune un efort cognitiv mult mai mare atunci când copilul nu este introdus într-o comunitate lingvistică de vorbitori nativi. La vârsta școlară, limbile străine sunt însușite de obicei prin învățarea structurii lor gramaticale, a vocabularului și a ortografiei din cărți și materiale educaționale care descriu cunoașterea lingvistică prin reguli abstracte, tabele sau texte exemplu. Învățarea unei limbi străine presupune mult timp și efort și devine din ce în ce mai dificilă cu înaintarea în vârstă.

Cele două tipuri principale de sisteme de TL achiziționează capacități lingvistice într-o manieră similară oamenilor. Abordările statistice (sau bazate pe date) obțin cunoștințe lingvistice dintr-o colecție vastă de exemple concrete. Dacă pentru anumite sisteme, precum corectoarele de limbă, sunt suficiente texte într-o singură limbă, alte aplicații necesită texte în două sau mai multe limbi, cum este cazul sistemelor de traducere automată. Algoritmii statistici de învățare automată „învață” șabloane de traducere corectă a cuvintelor, a frazelor scurte sau chiar a propozițiilor întregi.

Abordarea statistică poate avea nevoie de milioane de exemple, iar calitatea performanței crește odată cu numărului de texte analizate. Acesta este unul dintre motivele pentru care furnizorii de motoare de căutare sunt dornici să colecteze cât mai mult material scris. Corectarea erorilor de scriere în editoarele de text și servicii ca Google Search și Google Translate se bazează pe abordări statistice. Marele avantaj al statisticii este faptul că mașina învață repede, în cicluri repetate de antrenare, deși calitatea învățării poate varia arbitrar.

A doua abordare a tehnologiilor limbajului este dezvoltarea de sisteme bazate pe reguli. Experți din lingvistică, lingvistică computațională sau informatică codifică analize gramaticale (reguli de traducere) și compilează liste de tip vocabular (lexicoane). Realizarea unui sistem bazat pe reguli este o activitate care necesită mult timp și efort intens, dar și experți cu specializare înaltă.

O parte dintre cele mai performante sisteme de traducere automată bazată pe reguli se află în dezvoltare constantă de mai mult de douăzeci de ani. Avantajul acestor sisteme este că experții pot avea un control mai detaliat asupra procesării limbajului. Aceasta face posibilă corectarea sistematică a greșelilor din software și furnizarea de răspunsuri detaliate către utilizator, în special când sistemele bazate pe reguli sunt folosite pentru învățarea unei limbi. Datorită constrângerilor financiare, sisteme de tehnologia limbajului bazate pe reguli au fost până acum dezvoltate doar pentru câteva limbi majore. Deoarece punctele forte și punctele slabe ale sistemelor statistice și ale sistemelor bazate pe reguli tind să fie complementare, cercetările actuale se concentrează pe abordări hibride, care combină cele două metodologii. To-

tuși, aceste abordări nu au avut până în prezent același succes în aplicațiile industriale ca cel din laboratoarele de cercetare.

După cum am văzut în acest capitol, multe aplicații utilizate pe scară largă în societatea informațională de astăzi se bazează pe tehnologii ale limbajului. Datorită comunității sale multilingve, acest lucru este valabil în special în spațiul economic și informațional din Europa. Deși tehnologia limbajului a făcut progrese considerabile în ultimii ani, există încă un potențial uriaș în îmbunătățirea calității sistemelor bazate pe tehnologii lingvistice. În cele ce urmează, vom descrie rolul limbii române în societatea informațională europeană și vom evalua stadiul actual al cercetărilor în domeniul tehnologiei limbajului pentru limba română.

The Romanian Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)

2012, VI, 87 p. 24 illus. in color., Softcover

ISBN: 978-3-642-30702-7