D2.3.ro
Language Report for Romanian
(Romanian version)

Version 1.2

2011-07-31

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	3-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu-Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Draft version
1.1	15-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu-Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Pre-final version
1.2	31-07-2011	Diana Trandabăț, Elena Irimia, Verginica Mititelu-Barbu, Dan Tufiș, Dan Cristea	UAIC, RACAI	Final version

METANET4U

D2.3.ro
Language Report for Romanian
(Romanian version)

Document METANET4U-2011-D2.3.ron
EC CIP project #270893

Deliverable
Number: D2.3.ro
Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: UAIC, RACAI

Authors: Diana Trandabăț, Elena Irimia, Verginica Mititelu-Barbu, Dan Tufiș, Dan Cristea

Reviewer: Amália Mendes, Adrian Iftene, Ionuț Pistol

Cuprins

Rezumat

Multe limbi europene riscă să devină victimele erei digitale, fiind sub-reprezentate şi dispunând de insuficiente resurse online. Oportunităţi uriaşe de piaţă regională rămân neutilizate astăzi datorită barierelor de limbă. Dacă nu acţionăm acum, vorbirea propriei limbi native va deveni un dezavantaj economic şi social pentru mulţi dintre cetăţenii europeni.

Inovativă, Tehnologia Limbajului (TL) este un intermediar care va permite tuturor cetăţenilor europeni să participe la o societate a cunoaşterii şi informaţiei egalitară, inclusivă şi de succes economic. Tehnologia Limbajului va fi poarta către comunicare şi interacţiune dincolo de graniţele limbii, de o manieră instantanee, ieftină şi fără efort.

Astăzi, serviciile lingvistice sunt oferite în general de către furnizori comerciali din Statele Unite. Google Translate, un serviciu gratuit de traducere automată, este doar un exemplu. Succesul recent al lui Watson, un sistem IBM care a câştigat un episod din spectacolul-joc TV Jeopardy împotriva unor candidaţi umani, ilustrează imensul potenţial al tehnologiei limbajului. Ca europeni, trebuie să ne punem câteva întrebări urgente:

- Infrastructura noastră de comunicare şi cunoaştere trebuie să fie dependentă de companii monopoliste?
- Ne putem cu adevărat baza pe servicii lingvistice care pot fi foarte uşor întrerupte de către alţii?
- Concurăm activ pe piaţa globală a cercetării şi dezvoltării în domeniul tehnologiei limbajului?
- Sunt părţile terţe de pe alte continente dispuse să adreseze problemele noastre de traducere precum şi alte subiecte legate de multilingvismul European?
- Este moştenirea culturală europeană capabilă de a modela societatea cunoaşterii oferindu-i tehnologie de înaltă calitate, mai bună, mai sigură, mai precisă, mai inovativă şi mai robustă?

Acest raport asupra limbii române demonstrează că cercetarea în universităţi şi în mediul academic a reuşit să proiecteze software specific de înaltă calitate precum şi modele şi teorii aplicabile pe scară largă. Totuşi, pentru ca dezvoltarea TL să continue, o implicare mai intensă a Guvernului prin finanţare adecvată ar trebui obţinută, iar colaborările atractive cu industriile care folosesc sau furnizează TL ar trebui promovate.

META-NET contribuie la construirea unui spaţiu digital informaţional European puternic şi multilingv. Prin realizarea acestui deziderat, o uniune multiculturală a naţiunilor poate prospera şi deveni un model pentru cooperarea internaţională paşnică şi egalitaristă. Dacă acest ţel nu poate fi atins, Europa va trebui să aleagă între a-şi sacrifica identităţile culturale sau a suferi o înfrângere economică.

Un risc pentru limbile noastre și o provocare pentru tehnologia limbajului

Suntem martorii unei revoluţii digitale care are un impact dramatic asupra comunicării şi societăţii. Dezvoltările recente din tehnologia comunicaţiilor digitizate în reţea sunt uneori comparate cu inventarea de către Gutenberg a tiparului. Ce ne poate spune această analogie despre viitorul societăţii informaţionale europene şi, în special, despre viitorul limbilor noastre?

Suntem martorii unei revoluţii digitale comparabilă cu inventarea tiparului de către Gutenberg.

Ulterior invenţiei lui Gutenberg au avut loc progrese reale în comunicare şi schimbul de informaţie, datorită unor eforturi precum traducerea de către Luther a Bibliei din Latină în limba enoriaşilor. În secolele următoare, au fost dezvoltate tehnici culturale pentru a îmbunătăţi prelucrarea limbajului şi schimbul de cunoştinţe:

- Standardizarea ortografică şi gramaticală a limbilor importante a permis diseminarea rapidă a noilor idei intelectuale şi ştiinţifice;
- Dezvoltarea limbilor oficiale a făcut posibilă comunicarea cetăţenilor în interiorul anumitor graniţe (adeseori politice);
- Predarea şi traducerea au facilitat schimbul între limbi;
- Crearea de principii jurnalistice şi bibliografice a asigurat calitatea şi disponibilitatea materialelor imprimate;
- Crearea diferitelor tipuri de media, precum ziarele, radioul, televiziunea, cărţile etc. a satisfăcut nevoile de comunicare.

În ultimii 20 de ani, tehnologia informaţiei a contribuit la automatizarea şi facilitarea multora dintre procese:

- Software-ul pentru tehnoredactare computerizată înlocuieşte dactilografierea şi culegerea;
- Microsoft PowerPoint înlocuieşte retroproiectorul;
- E-mail-ul trimite şi primeşte documente mai repede decât un fax;
- Skype permite convorbirile prin Internet şi găzduieşte întâlnirile virtuale;
- Formatele de codificare audio şi video uşurează schimbul de conţinut multimedia;
- Motoarele de căutare oferă acces la pagini web bazat pe cuvinte cheie;
- Serviciile online precum Google Translate produc traduceri rapide şi aproximative;
- Platformele de media socială facilitează colaborarea şi partajarea de informaţii.

Deşi astfel de instrumente şi aplicaţii sunt utile, acestea nu sunt suficiente pentru a implementa o societate informaţională europeană multilingvă şi sustenabilă, o societate inclusivă şi modernă unde informaţia şi bunurile pot circula liber.

Frontierele lingvistice frânează crearea unei societăți informaționale europene

Nu putem şti cu precizie cum va arăta viitoarea societate informaţională. Când vine vorba despre discutarea unei strategii energetice sau a unei politici externe comune, am dori să-i ascultăm pe miniş-

trii de externe europeni vorbind în limbile materne. Preferăm o platformă în care oamenii, vorbitori de limbi diferite şi cu diverse competenţe lingvistice, pot discuta un subiect specific în timp ce tehnologia le colectează automat opiniile şi generează scurte rezumate. Am dori de asemenea să putem vorbi cu un birou de asistenţă pentru asigurări de sănătate localizat într-o altă ţară.

Este limpede că este nevoie de o calitate a comunicării diferită celei de acum câţiva ani. Într-un spaţiu de informare şi economic global, suntem confruntaţi cu mai multe limbi, mai mulţi vorbitori şi mai mult conţinut şi suntem nevoiţi să interacţionăm rapid cu noi tipuri de media. Popularitatea actuală a mediilor sociale (Wikipedia, Facebook, Twitter şi YouTube) reprezintă doar vârful icebergului.

Un spaţiu economic şi de comunicare global ne confruntă cu mai multe limbi, mai mulţi vorbitori, mai mult conţinut.

Astăzi putem recepţiona gigaocteţi de text din orice colţ al planetei în câteva secunde, doar pentru a afla că textul este într-o limbă pe care nu o înţelegem. Potrivit unui raport recent solicitat de Comisia Europeană, 57% dintre utilizatorii de Internet din Europa achiziţionează bunuri şi servicii în limbi diferite de cea maternă. (Engleza este cea mai comună limbă străină, urmată de Franceză, Germană şi Spaniolă.) 55 % dintre utilizatori citesc conţinut 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[i]. Cu câţiva ani în urmă, engleza era privită ca *lingua franca* (limba de lucru) a Internetului – o vastă majoritate a conţinutului era scrisă în această limbă – dar situaţia s-a schimbat drastic acum. Cantitatea de conţinut online în alte limbi, în special asiatice şi arabe, a explodat.

În mod surprinzător, ubicua diviziune digitală provocată de frontierele lingvistice nu a câştigat încă prea multă atenţie în discursul public; totuşi, ea ridică o întrebare foarte presantă: „Care dintre limbile europene va reuşi să ajungă şi apoi să persiste în societatea virtuală a informaţiei şi cunoaşterii?"

Care dintre limbile europene va reuşi să ajungă şi apoi să persiste în societatea virtuală a informaţiei şi cunoaşterii?

Limbile noastre sunt în pericol

Tiparul a contribuit la un inestimabil schimb de informaţii în Europa, dar a condus de asemenea la extincţia multora dintre limbile europene. Limbile regionale şi minoritare au fost tipărite arareori. În consecinţă, multe limbi precum dalmata sau limba din Cornwall erau adeseori limitate la forme orale de transmitere, care le-au restricţionat adoptarea, răspândirea şi utilizarea.

Cele aproximativ 60 de limbi vorbite astăzi în Europa reprezintă unul dintre cele mai bogate şi importante bunuri culturale ale sale. Această multitudine de limbi este în acelaşi timp o componentă importantă a succesului său social[ii]. În timp ce limbi populare precum Engleza sau Spaniola vor rămâne cu siguranţă prezente în societatea şi pe piaţa digitală emergentă, multe limbi europene ar putea fi deconectate de la comunicarea digitală şi ar putea deveni irelevante pentru societatea Internetului.

Larga varietate de limbi a Europei este unul dintre cele mai importante bunuri culturale ale sale şi o componentă esenţială a succesului său social.

O astfel de evoluţie ar fi cu siguranţă de nedorit. În primul rând, ea ar însemna că o oportunitate strategică rămâne nefolosită, slăbind poziţia Europei pe piaţa globală. În al doilea rând, o astfel de evoluţie ar fi în contradicţie cu obiectivul esenţial al participării egale a fiecărui cetăţean european, indiferent de limba lui. Potrivit unui raport recent al UNESCO privind multilingvismul, limbile străine reprezintă un mediu esenţial pentru exercitarea drepturilor fundamentale precum exprimarea politică, educaţia şi participarea în societate[iii].

Tehnologia limbajului este cheia activării tehnologiei

În trecut, eforturile de investiţii financiare au fost concentrate asupra educaţiei lingvistice şi a traducerii. De exemplu, potrivit anumitor estimări, piaţa europeană de traducere, interpretare, software de localizare şi globalizare a paginilor de Internet a valorat 8.4 miliarde € în 2008 şi este de aşteptat să crească cu 10% pe an[iv]. Totuşi, această capacitate nu este suficientă pentru a satisface nevoile curente şi viitoare.

Tehnologia limbajului este o tehnologie cheie care poate proteja şi promova limbile europene. Ea ajută oamenii să colaboreze, să conducă afaceri, să împărtăşească cunoştinţe şi să participe în dezbateri politice şi sociale, independent de barierele lingvistice sau de competenţele de lucru cu calculatorul. Tehnologia limbajului asistă deja sarcinile de zi cu zi, precum scrierea e-mail-urilor, căutările online sau rezervarea unor bilete de avion. Beneficiem de pe urma tehnologiei limbajului atunci când:

Tehnologia Limbajului ajută oamenii să colaboreze, să conducă afaceri, să împărtăşească cunoştinţe şi să participe în dezbateri politice şi sociale în diverse limbi.

- Descoperim informaţii cu un motor de căutare pe Internet;
- Verificăm corectitudinea ortografică şi gramaticală într-un editor de text;
- Vizualizăm recomandări de produse într-un magazin online;
- Ascultăm instrucţiunile verbale ale unui sistem de navigare;
- Traducem pagini web cu un serviciu online.

Tehnologiile limbajului detaliate în continuare reprezintă o componentă esenţială a aplicaţiilor inovative viitoare. Tehnologia limbajului este de obicei o tehnologie auxiliară în interiorul unui cadru de aplicare mai larg, precum un sistem de navigare sau un motor de căutare. Cărţile albe realizate de METANET se concentrează pe disponibilitatea unor tehnologii lingvistice de bază pentru fiecare dintre limbi.

În viitorul apropiat, avem nevoie de tehnologie a limbajului care să fie disponibilă pentru toate limbile europene, accesibilă şi bine integrată în medii software mai largi. O experienţă de utilizator interactivă, multimedia şi multilingvă nu este posibilă fără tehnologia limbajului.

Oportunităţi ale tehnologiei limbajului

Tehnologia limbajului face posibile, pentru toate limbile europene, aplicaţii precum traducerea automată, generarea de conţinut, procesarea informaţiei şi managementul cunoştinţelor. De asemenea, ea poate continua dezvoltarea de interfeţe intuitive bazate pe limbă pentru dispozitive electrocasnice, utilaje, vehicule, computere şi roboţi. Deşi există deja multe prototipuri, aplicaţiile comerciale şi industriale sunt încă în stadii incipiente de dezvoltare. Realizări recente în cercetare şi dezvoltare au creat o adevărată fereastră de oportunitate pentru TL. De exemplu, traducerea automată (TA) oferă o acurateţe rezonabilă pentru domenii specifice iar aplicaţii experimentale asigură managementul informaţiei şi cunoştinţelor precum şi producerea de conţinut în multe din limbile europene.

Aplicaţii precum interfeţe bazate pe limbă şi pe voce sau sisteme de dialog sunt utilizate în mod tradiţional în domenii specializate şi prezintă adeseori performanţe limitate. Un câmp activ de cercetare este reprezentat de tehnologia dedicată operaţiilor de salvare în zonele sinistrate. În astfel de medii cu risc înalt, acurateţea traducerii poate fi o problemă de viaţă şi de moarte. La fel în ceea ce

privește utilizarea tehnologiei limbajului în sectorul de îngrijire a sănătății. Roboți inteligenți cu capabilități trans-linguale au potențialul de a salva vieți.

Există oportunități uriașe de piață în sectorul educației și al divertismentului pentru integrarea tehnologiei limbajului în jocuri, oferte "edutainment" (educație prin divertisment), medii de simulare sau programe de formare. Serviciile mobile de informații, softwareul pentru învățarea de limbi străine asistată de calculator, mediile e-learning, instrumentele de auto-evaluare și cele de detectare a plagiatului sunt doar câteva exemple de zone de aplicare 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 publicărilor, rezumarea discuțiilor, identificarea unor curente de opinie, detectarea răspunsurilor emoționale, descoperirea încălcărilor drepturilor de autor sau a situațiilor de abuz.

Tehnologia limbajului reprezintă o oportunitate uriașă pentru Uniunea Europeană, atât din punct de vedere economic cât și din perspectivă culturală. Multilingvismul a devenit 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. Tehnologia limbajului poate ajuta la depășirea acestor bariere, sprijinind în același timp utilizarea liberă și deschisă a limbilor. Mai mult, o tehnologie a limbajului europeană inovativă și multilingvă ne poate ajuta să comunicăm cu partenerii noștri globali și comunitățile lor multilingve. Tehnologiile limbajului sprijină o varietate de oportunități economice internaționale.

Multilingvismul este regula, nu excepția.

Provocările tehnologiei limbajului

Deși tehnologia limbajului s-a dezvoltat considerabil în ultimii ani, ritmul actual al progresului tehnologic și al inovării este prea lent. Nu putem aștepta zece sau douăzeci de ani pentru a fi martorii unor îmbunătățiri semnificative în direcția unei mai bune comunicări și productivități în mediul nostru multilingv. 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.

Ritmul actual al progresului tehnologic și al inovării este prea lent pentru a obține produse software substanțiale în următorii douăzeci de ani.

Aplicațiile pentru comunicare multilingvă necesită un anumit nivel de sofisticare. Serviciile online de traducere automată precum Google Translate sau Bing Translator sunt excelente în a produce o bună aproximare a conținutului documentelor. Totuși, asemenea servicii, ca și aplicațiile profesionale de TA, întâmpină multe dificultăți atunci când este nevoie de traduceri precise și complete. Există multe exemple bine-cunoscute de traduceri greșite amuzante (ca de exemplu traducerile literale ale unor nume precum „Bush" sau „Kohl") care ilustrează provocările cărora tehnologia limbajului încă trebuie să le facă față.

Achiziția limbii

Ca să ilustrăm modul în care computerele interacționează cu limbajul și să explicăm de ce achiziția limbajului este o sarcină foarte dificilă, vom arunca o scurtă privire asupra modului în care oamenii achiziționează prima și a doua limbă, iar ulterior vom schița funcționarea sistemelor de traducere automată — există un motiv pentru care tehnologia limbajului este strâns legată de domeniul inteligenței artificiale.

Oamenii achiziţionează competenţele lingvistice în două moduri distincte. În primul rând, un copil învaţă o limbă ascultând interacţiuni între vorbitorii acelei limbi. Expunerea la exemple lingvistice concrete produse de utilizatorii limbii precum părinţii, fraţii sau alţi membri ai familiei ajută copiii să producă primele lor cuvinte sau fraze scurte la vârsta aproximativă de doi ani. Acest lucru este posibil pentru că oamenii au o predispoziţie genetică specială pentru învăţarea limbilor. Învăţarea unei a doua limbi presupune un efort mult mai mare atunci când copilul nu este introdus într-o comunitate lingvistică de vorbitori nativi. La vârsta şcolară, limbile străine sunt achiziţionate de obicei prin învăţarea structurii lor gramaticale, a vocabularului şi a ortografiei din cărţi şi materiale educaţionale care descriu cunoaşterea lingvistică prin reguli abstracte, tabele sau texte exemplu. Învăţarea unei limbi străine presupune mult timp şi efort şi devine din ce în ce mai dificilă cu înaintarea în vârstă.

Oamenii achiziţionează competenţe lingvistice în două modalităţi diferite: învăţând exemple şi învăţând regulile care stau la baza limbii.

Cele două tipuri principale de sisteme de TL achiziţionează capacităţi lingvistice într-o manieră similară oamenilor. Abordările statistice obţin cunoştinţe lingvistice dintr-o colecţie vastă de exemple concrete într-o anumită limbă sau din aşa numitele texte paralele disponibile în una sau mai multe limbi. Algoritmii de învăţare automată modelează acea facultate a limbajului care poate deriva şabloane de utilizare corectă a cuvintelor, frazelor scurte sau a propoziţiilor complete, într-o anumită limbă sau în traducerea dintr-o limbă în alta. Numărul de propoziţii necesare abordărilor statistice este uriaş iar calitatea performanţei creşte odată cu creşterea numărului de texte analizate. Nu este rar întâlnită antrenarea unor astfel de sisteme pe texte care conţin milioane de propoziţii. Acesta este unul dintre motivele pentru care furnizorii de motoare de căutare sunt dornici de a colecta cât mai mult material scris. Corectarea erorilor de ortografie în editoarele de text, disponibilitatea informaţiei online precum şi servicii ca Google Search şi Google Translate se bazează pe abordări statistice (orientate către date).

Cele două tipuri principale de sisteme de TL achiziţionează informaţia lingvistică într-o manieră similară oamenilor.

Sistemele bazate pe reguli reprezintă a doua direcţie majoră în domeniul TL. Experţi din Lingvistică, Lingvistică Computaţională sau Ştiinţa Calculatoarelor codifică gramatici şi compilează liste de tip vocabular (lexicoane). Realizarea unui sistem bazat pe reguli este o activitate care necesită mult timp şi efort intens, dar şi experţi cu specializare înaltă. O parte dintre cele mai performante sisteme de traducere automată bazate pe reguli se află în dezvoltare constantă de mai mult de douăzeci de ani. Avantajul acestor sisteme este că experţii pot avea un control mai detaliat asupra procesării limbajului. Asta face posibilă corectarea sistematică a greşelilor din software şi furnizarea de răspunsuri detaliate către utilizator, în special când sistemele bazate pe reguli sunt folosite pentru învăţarea unei limbi. Datorită constrângerilor financiare, tehnologia limbajului bazată pe reguli este fezabilă doar pentru limbile majore.

Limba română în societatea informațională europeană

Fapte generale

Vorbită de aproximativ **29.000.000 milioane de vorbitori,** limba română este limba maternă a 25.000.000 de vorbitori: în jur de 21.500.000 de vorbitori în România[v] plus aprox. 3.500.000 de vorbitori în Republica Moldova[vi] (unde limba este denumită în mod oficial moldovenească). Ţările vecine României (Albania, Bulgaria, Croaţia, Grecia, Ungaria, Fosta Republică Iugoslavă a Macedoniei, Serbia, Ucraina) şi comunităţi de imigranţi din Australia, Canada, Israel, America Latină, Turcia, S.U.A. şi alte ţări Europene şi Asiatice însumează în jur de 4.000.000 de vorbitori nativi de română[vii].

Româna este de asemenea o limbă oficială în Provincia Autonomă Voivodina din Serbia, în Muntele Athos autonom din Grecia, în Uniunea Europeană şi în Uniunea Latină; este recunoscută ca limbă minoritară în Ucraina.

Limba română are **4 dialecte**[viii]: Daco-Româna/ Româna, Aromâna (vorbită de aproximativ 600.000 de vorbitori în Albania, Bulgaria, Grecia şi Macedonia), Istro-Româna (15.000 de vorbitori în 2 zone mici din Peninsula Istria, Croaţia) şi Megleno-Româna (în jur de 5.000 de vorbitori în Grecia şi Macedonia). Datorită numărului mic de vorbitori, aceste dialecte sunt incluse în Cartea Roşie a Limbilor pe Cale de Dispariţie *UNESCO*.

În România există 18 minorităţi etnice recunoscute oficial; la ultimul recensământ (2002), cei mai numeroşi erau ungurii (1.431.807) şi rroma (535.140), urmaţi de germani, ucrainieni, ruşi lipoveni, turci, sârbi, croaţi, sloveni, tătari, slovaci, bulgari, evrei, cehi, polonezi, greci, armeni, etc. Pentru toate aceste minorităţi, politicile lingvistice oficiale în România garantează drepturile acestora de a fi protejate în calitate de comunităţi lingvistice şi de a utiliza limba maternă în medii private şi publice, culturale şi sociale, economice şi de comunicare. Totuşi, articolul 13 al Constituţiei prevede că "în România, limba oficială este româna". Mai mult, Legea nr. 500 din 12 noiembrie 2004 stipulează obligaţia ca orice text (fie el oral sau scris) de interes public să fie tradus sau adaptat în limba română[ix].

Particularităţile limbii române

Limba română este o limbă romanică orientală, care s-a format la distanţă de surorile sale occidentale. Elemente ale latinei populare, din care a evoluat, sunt mai bine păstrate în această limbă izolată geografic: s-au moştenit structura morfo-sintactică latinească, particularităţi pe care alte limbi romanice le-au pierdut (precum declinările), au fost întărite elemente morfologice (reflexivul) sau au fost preluate elemente non-romanice (vocativul în *-o*).

Cea mai mare parte a vocabularului limbii române are origine latină, fie moştenit din latina vulgară, fie împrumutat pe cale savantă, în epoca modernă. 60% din vocabularul fundamental (i.e. cuvintele cunoscute şi folosite curent de toţi vorbitorii) este moştenit din latină.

În timpul colonizării Daciei de către romani (106-271 d.Hr.), colonişţii au impus limba latină ca limbă oficială. Cu toate acestea, studii comparative între vocabularul românesc şi cel albanez dovedesc existenţa unui număr de aproximativ 100 de cuvinte păstrate din substratul traco-dac. Aceste cuvinte denumesc concepte fun-

damentale, precum părţi ale corpului, elemente naturale, hrană. Ele sunt folosite şi astăzi, sunt foarte frecvente, au dezvoltat o polisemie şi familii lexicale bogate.

În timpul migraţiei triburilor slave pe teritoriul României de astăzi, limba română a suferit un proces de transformare în toate compartimentele: fonetică, vocabular, morfologie şi sintaxă. Cu toate acestea, morfologia, care dă esenţa unei limbi, a rămas latinească în cele mai multe aspecte ale sale. Alfabetul chirilic a fost adoptat în această perioadă, mai ales datorită influenţei bisericeşti. Slavona a fost limba în care s-a oficiat serviciul religios în biserica ortodoxă până în secolul al XVIII-lea, când româna a început un proces de relatinizare, modernizare şi occidentalizare. Atunci, multe cuvinte de alte origini au fost înlocuite de cuvinte latineşti, împrumutate direct sau indirect, prin intermediul altor limbi romanice (franceză şi italiană). Franceza ca limbă de cultură în ultimele două secole şi Franţa ca ţara în care aristocraţia română îşi trimitea copiii la învăţătură justifică existenţa unui număr extrem de mare de cuvinte de această origine în limba română. În ultimul timp, limba engleză a luat locul francezei, iar româna are multe anglicisme, adaptate total, parţial sau deloc la sistemul său fonetic şi morfologic.

Aspecte politice, economice şi sociale din istoria poporului român explică existenţa cuvintelor de diverse origini: turcă, greacă, germană, maghiară, bulgară, rusă etc. În română au fost create cuvinte noi mai ales prin sufixare, deşi studiile recente reflectă creşterea importanţei pe care a căpătat-o în ultima vreme prefixarea.

Limba română are cinci litere cu diacritice: *ă*, *î*, *â*, *ş*, *ţ*. Pentru ultimele două au circulat două variante: una cu virgulă sub literă, alta cu sedilă. Însă numai prima variantă este recomandată astăzi de Asociaţia de Standardizare din România (ASRO). Multe texte electronice nu sunt scrise cu diacritice, însă au fost create programe pentru a introduce diacriticele în mod automat în astfel de texte.

Limba română are cinci litere cu diacritice: ă, î, â, ş, ţ. Pentru ultimele două au circulat două variante: una cu virgulă sub literă, alta cu sedilă. Însă numai prima variantă este recomandată.

Sistemul flexionar al limbii române este destul de bogat: pentru substantive, pronume şi adjective există cinci cazuri şi două numere, pentru verbe sunt două numere, fiecare cu câte trei persoane, cinci timpuri sintetice plus infinitivul, gerunziul si participiul. În medie, un substantiv poate avea cinci forme, un pronume personal şase, un adjectiv şase, iar un verb peste treizeci. În afară de sufixele morfologice şi de desinenţe, flexiunea cuvintelor mai prezintă şi alternanţe fonetice în interiorul rădăcinii.

Româna este o limbă care permite nelexicalizarea subiectului pronominal, ca cele mai multe limbi romanice, de altfel:

Ştiu.

Explicaţia rezidă în sistemul flexionar bogat al verbelor, care au desinenţe diferite pentru persoane şi numere diferite.

Cu toate acestea, şi dublarea subiectului este posibilă în română, atunci când un pronume personal dublează un grup nominal lexical:

Vine el tata imediat!

Structura este caracteristică limbajului familiar, marcând o anumită atitudine ilocuţionară a vorbitorului: ameninţare, promisiune, asigurare verbală.

Româna are în comun cu anumite dialecte spaniole şi cu câteva limbi balcanice o structură cunoscută sub numele de „dublare cliti-

că". Dublarea clitică pronominală în română se face cu pronume neaccentuate de dativ, de acuzativ sau ambele. De exemplu, în propoziţia

"I_i l_j-am dat mamei$_i$ pe Ion$_j$ la telefon."

substantivul *mamei* şi cliticul de dativ *i* se referă la aceeaşi persoană, iar cliticul de acuzativ *l-* şi substantivul în acuzativ *pe Ion* sunt tot coreferenţiale. Prezenţa cliticelor în asemenea construcţii este obligatorie, deşi ele nu complinesc valenţe verbale. Însă atunci când substantivele nu sunt prezente, pronumelor le revine sarcina de a satura valenţele verbale:

"I l-am dat la telefon."

Este obligatorie dublarea numelor proprii şi a substantivelor articulate hotărât, cu funcţie sintactică de complement direct sau indirect.

Limba română prezintă atât fenomenul concordanţei negative, cât şi dubla negaţie. Prezenţa marcatorului negativ *nu* în grupul verbal imprimă caracter negativ întregii propoziţii şi autorizează cuvintele negative în respectiva propoziţie (concordanţă negativă):

„Nu am văzut pe nimeni niciodată aici."

Limba română este o limbă cu un sistem bogat de flexionare, cu diferite particularităţi lingvistice: permite elipsa subiectului, dublarea cliticelor, permite concordanţă negativă şi negaţie dublă.

Totuşi, anumite configuraţii în care apar marcatorii şi cuvintele negative trebuie interpretate ca având dublă negaţie (adică, în ciuda formei negative a verbului predicativ, enunţul are un conţinut afirmativ). De exemplu, o propoziţie principală negativă urmată de o subordonată cu verbul la forma negativă a modului conjunctiv este o astfel de configuraţie cu sens afirmativ:

„Maria nu a vrut să nu spună nimic."=

„Maria a vrut să spună ceva."

Cazul este sintetic în limba română: substantivul îşi schimbă forma pentru exprimarea cazului. Cu toate acestea, există şi trei prepoziţii care marchează cazul: *pe* pentru acuzativ (condiţionată de trăsăturile animat, hotărât şi specific ale grupului nominal), *la* pentru dativ şi *a* pentru genitiv (ambele condiţionate de prezenţa unui numeral în grupul nominal):

L-am văzut pe colegul meu.

Am dat cărţile la trei dintre copii.

Cărţile a trei copii erau noi.

Dezvoltări Recente

Asemănător cu procesul de relatinizare din secolul al nouăsprezecelea, de după eliberarea de sub dominaţia greacă şi otomană, limba română a parcurs în secolul al XX-lea un proces de trecere de la limbajul totalitar („limba de lemn", discursul unidirecţional etc.) la utilizarea deschisă, în care noi modele lingvistice trebuie să se adapteze la tranziţia socială şi culturală. Astfel, asemănător multor altor limbi, româna traversează un proces continuu de internaţionalizare, sub influenţa vocabularului anglo-saxon.

În domenii esenţiale, precum ştiinţele politice, administrative şi economice, în presă, publicitate, ştiinţa calculatoarelor etc., au fost împrumutate numeroase cuvinte sau cuvinte existente au căpătat sensuri noi, după model englezesc; terminologiile domeniilor noi

se bazează pe împrumuturi din engleză, vocabularul activ al oamenilor instruiţi conţine din ce în ce mai multe anglicisme, se pot observa noi modele intonaţionale (mai ales în presa vorbită).

În anumite domenii, anglicismele au început să înlocuiască vocabularul limbii române. Un exemplu este folosirea titlurilor englezeşti pentru anunţuri de locuri de muncă, în special pentru poziţii de conducere, de ex. 'Human Resource Manager' în loc de *Director de Resurse Umane*. O tendinţă puternică de a exagera folosirea anglicismelor poate fi observată în reclame. Bănci din România folosesc slogane promoţionale de genul: *Cu cine faci banking?* Sau *Prima modalitate de plată contactless*, deşi *banking* sau *contactless* snt anglicisme care nu au intrat în vocabularul comun şi cu care majoritatea românilor nu sunt obişnuiţi.

Exemplul de mai sus demonstrează importanţa ridicării unui semnal de alarmă asupra unei dezvoltări care riscă să excludă din societatea informaţională o mare parte a populaţiei, care nu este familiară cu limba engleză.

Cultivarea limbii în România

Academia Română, cel mai înalt forum cultural al ţării, are printre obiectivele sale principale cultivarea limbii naţionale. Scopul major al institutelor sale lingvistice a fost crearea şi publicarea *Dicţionarului Tezaur al Limbii Române*, proces care a durat aproape un secol. Seria mai veche, cunoscută sub numele de *Dicţionarul Academiei* (DA), include 5 volume cu 3146 de pagini şi 44890 de intrări lexicale şi a fost realizată între anii 1913 şi 1947. După o întrerupere, lucrul a fost reluat la mijlocul deceniului al şaptelea al secolului trecut cu o serie nouă, cunoscută sub numele de *Dicţionarul Limbii Române* (DLR). Ultimul volum a fost publicat la Editura Academiei la începutul lui 2009. În total, DA şi DLR au 33 de volume, peste 15000 de pagini şi în jur de 175000 de intrări. Dicţionarul a fost creat în stilul tradiţional, „cu creionul pe hârtie", cu citate adunate din peste 2500 de volume de literatură română scrisă.

Institutul de Lingvistică „Iordu Iordan – Al. Rosetti" are un program de cercetare ce urmăreşte cultivarea limbii, elaborează dicţionare normative (*Dicţionar ortografic, ortoepic şi morfologic al limbii române*, *Dicţionarul împrumuturilor neadaptate*, *Dicţionarul termenilor oficiali*) şi gramatici (*Gramatica limbii române*, *Dinamica limbii române actuale*).

Legea 500 din 12 noiembrie 2004 prevede ca toate textele scrise sau orale în limba română, care servesc interesul public, să respecte normele academice.

Institutul Limbii Române a fost creat cu scopul de a promova învăţarea limbii române peste hotare, de a-i sprijini pe cei care învaţă limba română şi de a le atesta cunoştinţele de română[x]. Există în străinătate peste 70 de centre în care se predă limba română ca limbă străină de către cadre didactice din învăţământul universitar românesc.

Există în străinătate peste 70 de centre în care se predă limba română ca limbă străină de către cadre didactice din învăţământul universitar românesc.

Şi în România se manifestă un interes crescând pentru studiul limbii române printre străini, nu doar la nivel diplomatic (de către reprezentanţii misiunilor diplomatice ale diverselor ţări), ci şi în mediul de afaceri. În afară de universităţi, care oferă cursuri de română ca limbă străină de obicei pentru studenţii străini din România, există şi numeroase firme particulare cu oferte mai ales pentru străini implicaţi în sectorul economic. Sunt organizate cur-

suri de vară de limba română pentru toate nivelurile, anual, în diverse locuri din ţară, de Fundaţia Culturală Română, precum şi de câteva instituţii de învăţământ superior (precum Universitatea „Al. I Cuza" din Iaşi).

Cultivarea limbii în contextul înnoirii accelerate este o prioritate şi pentru presă. Canalele naţionale de radio şi televiziune au emisiuni în care sunt discutate împreună cu specialişti şi explicate publicului aspecte mai complicate ale limbii.

Limba în educaţie

Conform Noului curriculum naţional (2000), româna se predă 4-5 ore obligatorii pe săptămână în şcoala gimnazială şi 3-4 ore în liceu. Aspectele prescriptive ale conservării limbii se combină cu comunicarea, comportament axat pe competenţe, accentuându-se relaţia limbă-cultură. Limba şi literatura română reprezintă o materie obligatorie la examenele naţionale (la absolvirea ciclului gimnazial şi liceal; bacalaureatul cuprinde două probe de limba română: una orală şi alta scrisă).

Limba şi literatura română se studiază ca specializări principale sau secundare în peste 30 de universităţi de stat şi particulare din România.

Aspecte Internaţionale

România este recunoscută pe plan internaţional pentru literatura sa, lucrările principale ale lui Eminescu (marele poet naţional al României) fiind traduse în peste 60 de limbi. Alte nume cunoscute din literatura română sunt: Mircea Eliade, primul istoric care a scris o istorie a religiilor; Eugen Ionesco, unul dintre promotorii Teatrul Absurdului, sau Emil Cioran, cunoscut pentru filosofia lui.

În prezent, o mare parte din publicaţiile ştiinţifice din domeniul TL sunt scrise în limba engleză, deşi Consorţiul pentru Informatizarea Limbii Române - ConsILR – organizează anual un atelier ştiinţific dedicat cercetărilor în TL, cu lucrările atelierului publicate în limba română. O situaţie similară se regăseşte şi în celelalte domenii, mai puţin proeminentă pentru discipline precum filozofie, lingvistică, teologie sau pentru domeniul juridic.

Aceeaşi situaţie se întâlneşte şi în lumea afacerilor. În multe companii mari internaţionale, engleza a devenit *lingua franca*, atât în comunicarea scrisă (e-mail şi documente) cât şi în cea orală, în special în companii multinaţionale cu directori străini.

Tehnologiile limbajului pot rezolva această provocare din altă perspectivă prin oferirea unor servicii precum traducerea automată sau regăsirea de informaţii multilingvă în texte în limbi străine, ajutând astfel la diminuarea dezavantajelor personale şi economice cu care se confruntă vorbitorii care nu au cunoştinţe avansate de limbă engleză.

Minorităţi române trăiesc în ţările vecine şi în diaspora peste tot în lume. România promovează politici pentru păstrarea identităţii lingvistice şi culturale de către comunităţile româneşti. Centrul Euxodius Hurmuzachi oferă sute de burse anual în România pentru minorităţile române din ţările vecine. Sunt multe schimburi şcolare şi academice, mai ales cu Republica Moldova. Primele extinderi în sistem franciză ale şcolilor şi universităţilor din România au apărut în Republica Moldova în 2000.

În diferite comunităţi din diaspora, există iniţiative diverse, prin care cei interesaţi pot studia limba şi cultura română. De exemplu, Şcoala de limba română din Kitchener, Canada, oferă ore de limbă şi cultură română pentru copii şi adolescenţi.

Institutele Culturale Române există în 19 oraşe din lume (inclusiv Bucureşti, New York, Paris, Londra, Roma, Istanbul etc.) şi toate au drept preocupare importantă promovarea limbii române prin cursuri şi evenimente culturale de diverse tipuri.

Limba română pe Internet

Piaţa Internetului în România este în continuă creştere. În 2010, 44,2% dintre români aveau acces la un calculator acasă, iar 35,5% (i.e. 7.786.700 de români) erau utilizatori de Internet[xi] (aproximativ 60% dintre ei fiind utilizatori zilnici), ceea ce plasează România pe locul 8 într-un top 10 al utilizatorilor de Internet din Europa[xii]. Peste 500.000 de website-uri sunt înregistrate cu domeniul *.ro*.

Comparând aceste date cu cele din 2000, când numai 3,6% din populaţie (adică 800.000 de români) foloseau Internetul, observăm o creştere de aproape zece ori.

Un studiu al Uniunii Latine din 2007 arată că, similar cu tendinţa celorlalte limbi neolatine, prezenţa limbii române pe Internet crescut din 1998 până în 2007. Împărţind procentul de pagini web pentru fiecare limbă la procentul de prezenţă relativă a vorbitorilor limbii din lumea reală, s-a calculat vigoarea fiecărei limbi (sau prezenţa limbilor studiate în spaţiul virtual). Deşi acest coeficient este considerat unul redus pentru limba română (0,6 în 2007, în comparaţie cu 4,44 pentru engleză, 2,24 pentru franceză şi 2,93 pentru italiană), româna este singura limbă care a cunoscut o creştere în vigoare în perioada 2005-2007 (înaintea integrării în Uniunea Europeană).

Bibliografie selectivă

Grigore Brâncuş, *Vocabularul autohton al limbii române*, Bucharest, Scientific and Encyclopaedic Ed., 1983.

Alf Lombard, *La langue roumaine: Une presentation*, Paris, Klincksieck, 1974.

Ioana Vintilă-Rădulescu, *Limba română din perspective integrării în Uniunea Europeană*, http://www.unibuc.ro/ro/limba_romn_din_perspectiva_integrri_europene

Uniunea Latină în colaborare cu Funredes, *Limbile şi culturile pe internet 2007* http://dtil.unilat.org/LI/2007/index_ro.htm

Suport tehnologic pentru limba română

Tehnologiile limbajului

Tehnologiile limbajului sunt tehnologii informatice specializate pentru lucrul cu limbile naturale. De aceea ele sunt adesea subsumate termenului Tehnologia limbajului uman. Limbajul uman se manifestă în scris şi oral. În timp ce vorbirea este modul cel mai vechi şi mai natural al comunicării umane, informaţiile complexe şi cea mai mare parte a cunoştinţelor omeneşti sunt păstrate şi transmise prin texte scrise. Tehnologia vorbirii şi a textelor scrise prelucrează şi produce limbaj în aceste două modalităţi de realizare. Dar limba are şi aspecte comune vorbirii şi scrierii, precum lexicul, cea mai mare parte a gramaticii şi semantica. De aceea, o mare parte a tehnologiilor limbajului nu poate fi subsumată nici tehnologiei vorbirii, nici tehnologiei textelor scrise. Printre acestea se află tehnologiile care leagă limba de cunoaştere. Imaginea din dreapta ilustrează peisajul tehnologiilor limbajului. În comunicare, oamenii combină limbajul cu alte moduri de comunicare şi cu alte mijloace de informare. Îmbinăm vorbirea cu gesturile şi expresiile faciale. Textele electronice se combină cu imagini şi sunete. Filmele pot conţine limbaj în formă scrisă şi vorbită. De aceea, tehnologia vorbirii şi a textelor scrise se suprapune şi interacţionează cu multe alte tehnologii care facilitează prelucrarea comunicării multimodale şi a documentelor multimodale.

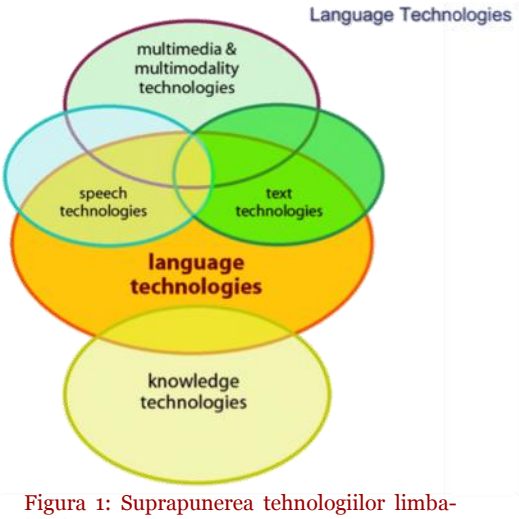

Figura 1: Suprapunerea tehnologiilor limbajului

Arhitecturile aplicațiilor din tehnologia limbajului

Aplicaţiile software tipice pentru prelucrarea limbii constau din câteva componente care reflectă diferite aspecte ale limbii şi ale sarcinii pe care o implementează. Figura din dreapta prezintă arhitectura foarte simplificată a unui sistem de prelucrare a textelor. Primele trei module abordează structura şi sensul textului introdus:

- Preprocesarea: curăţarea datelor, eliminarea formatărilor, recunoaşterea limbii introduse, înlocuirea diacriticelor greşite cu cele recomandate (de exemplu, înlocuirea lui ş cu sedilă cu ş cu virgulă).
- Analiza gramaticală: găsirea verbelor şi a argumentelor sale, a modificatorilor etc.; recunoaşterea structurii propoziţionale.
- Analiză semantică: dezambiguizare (cu ce sens sunt folosite cuvintele în context?), rezoluţia anaforei şi a expresiilor referenţiale precum „ea”, „maşina” etc.; reprezentarea sensului unei propoziţii într-un mod accesibil calculatorului.

Modulele specifice sarcinii efectuează apoi mai multe operaţii diferite precum rezumare automată a unui text introdus, căutări în baze de date şi multe altele. Mai jos vom ilustra principalele domenii de aplicaţii şi vom evidenţia anumite module ale diferitelor arhitecturi în fiecare secţiune. Arhitecturile sunt foarte simplificate şi idealizate, servind pentru ilustrarea complexităţii aplicaţiilor tehnologiei limbajului într-un mod inteligibil, la modul general.

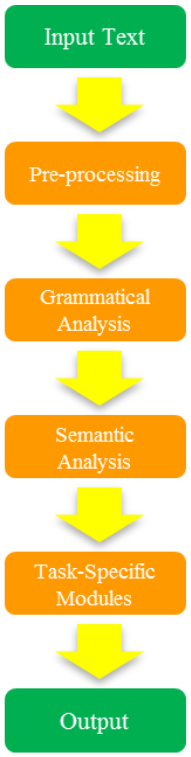

Figure 2: A Typical Text Processing Application Architecture

După introducerea principalelor domenii de aplicaţii, vom face o scurtă prezentare a situaţiei din cercetarea şi predarea din domeniul tehnologiei limbajului, încheind cu o enumerare a programelor de finanţare (din trecut). La finalul acestei secţiuni vom prezenta evaluarea de către experţi a instrumentelor şi resurselor din tehno-

logia limbajului, pe baza unor criterii precum disponibilitate, maturitate sau calitate. Acest tabel reprezintă un tablou corect al situaţiei tehnologiei limbajului pentru limba română.

Cele mai importante instrumente şi resurse din domeniu sunt subliniate în text şi pot fi găsite de asemenea în tabelul de la sfârşitul capitolului. Secţiunile care descriu principalele aplicaţii conţin de asemenea o trecere în revistă a companiilor active în domeniul respectiv în România.

Principalele domenii de aplicații

Corector de limbă

Oricine foloseşte un instrument de prelucrare a textului precum Microsoft Word a întâlnit o componentă care verifică ortografia, identifică greşelile de scriere şi face sugestii de corectură. După 40 de ani de la apariţia primului corector ortografic al lui Ralph Gorin, astfel de programe nu doar compară lista cuvintelor extrase cu cele dintr-un dicţionar cu cuvinte scrise corect, ci au devenit extrem de sofisticate în zilele noastre: utilizează algoritmi dependenţi de limbă pentru tratarea morfologiei (de exemplu, formarea pluralului), sunt acum capabile să recunoască greşeli de sintaxă, precum lipsa unui verb sau dezacordul în număr şi persoană dintre verb şi subiect, de exemplu „Ei scrie o scrisoare”. Cu toate acestea, cele mai multe corectoare ortografice disponibile (inclusiv cel din Microsoft Word) nu vor găsi nicio greşeală în următoarea strofă a unei poezii de Jerrold H. Zar (1992), poem bazat pe omofonie (pronunţia identică a unor cuvinte cu grafie diferită) şi lipsit de sens:

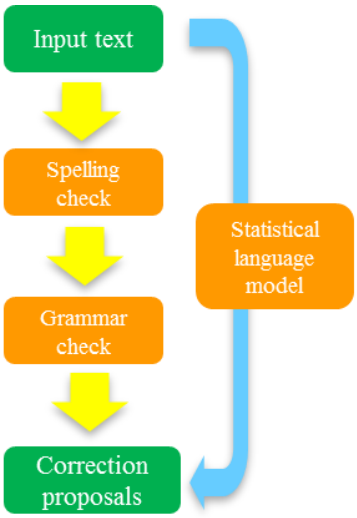

Figure 3: Language Checking (left: rule-based; right: statistical)

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

Pentru a da seama de astfel de greşeli este necesară analiza contextului în multe cazuri, de exemplu, pentru a decide dacă un cuvânt trebuie scris cu sau fără cratimă în română, precum în

Plouă întruna de ieri.

Într-una din zile am să merg la Paris.

Aceasta presupune fie formularea unor reguli gramaticale specifice limbii, adică un nivel înalt de expertiză şi muncă manuală, fie utilizarea aşa-numitelor modele lingvistice statistice. Acestea pot calcula probabilitatea ca un cuvânt să apară într-un anumit context (i.e. cuvintele dinainte şi de după). De exemplu, *într-una din zile* este o secvenţă de cuvinte mult mai probabilă decât *întruna din zile*, iar *plouă întruna* este mai frecventă decât *plouă într-una*, deci în al doilea caz se recomandă scrierea fără cratimă. Un model de limbă statistic poate fi creat automat pe baza unei cantităţi mari de date (corecte) de limbă (adică un corpus). Şi totuşi, sunt cazuri când nici măcar acesta nu este util::

Plouă întruna din primele zile ale lui martie.

Plouă într-una din primele zile ale lui martie.

Singurul element discriminatoriu aici este verbul. În prima propoziţie acesta este la prezent, având un sens durativ. În a doua este la trecut. Cele două forme verbale sunt omografe (se scriu la fel), dar

nu şi omofone în română. Numai adnotarea morfo-sintactică are valoare discriminatorie în asemenea cazuri.

Până acum, aceste abordări au fost dezvoltate şi aplicate mai ales pe date de limbă engleză. Ele nu pot fi transferate direct în limba română, care are o morfologie mai bogată şi construcţii specifice.

Utilitatea verificării limbii nu se limitează la instrumentele de prelucrare a textelor, ci se regăseşte şi în sistemele suport pentru autori. Urmare a creşterii numărului de produse tehnice, cantitatea de documentaţie tehnică a crescut vertiginos în ultimele decenii. Pentru a evita reclamaţiile clienţilor în legătură cu utilizarea incorectă şi pretenţiile pentru pagube rezultate din instrucţiunile greşite sau din înţelegerea greşită a acestora, companiile au început să se concentreze tot mai mult pe calitatea documentaţiei tehnice, ţintind în acelaşi timp pieţele internaţionale. Evoluţia prelucrării limbajului natural a dus la dezvoltarea unui software pentru sprijinul autorilor, care îl asistă pe cel care scrie documentaţii tehnice să folosească vocabularul şi structurile sintactice conforme cu anumite reguli şi restricţii terminologice (ale corporaţiilor).

Astăzi nu există companii româneşti sau furnizori de servicii lingvistice care să ofere astfel de produse, deşi cercetătorii din diverse grupuri de prelucrare a limbajului natural au dezvoltat modele de limbă ajustate la particularităţile Românei. La Institutul de Cercetări pentru Inteligenţă Artificială al Academiei Române (RACAI) au fost create modele de limbă pentru română pe baza unor corpusuri de mari dimensiuni. Întrucât majoritatea textelor de pe Web sunt scrise fără diacritice, RACAI a mai dezvoltat şi o aplicaţie de inserare a diacriticelor[xiii], care are scopul de a indica diacriticele corecte ale unui cuvânt scris iniţial fără diacritice; această aplicaţie foloseşte un lexicon românesc de mari dimensiuni dezvoltat în cadrul Institutului şi un model de ferestre de 5 caractere pentru a găsi cea mai probabilă interpretare în termeni de diacritice a unui cuvânt necunoscut. Metoda de lucru ia în considerare contextul unui cuvânt în faza de preprocesare prin adnotare morfo-sintactică, esenţială pentru alegerea cuvântului corect din lexicon. De exemplu, cuvântul „peste" este transformat în „peşte" în

Am cumparat peste.

dar este păstrat ca „peste" în

Era un pod peste rau.

Această decizie se bazează pe o etapă anterioară de adnotare morfo-sintactică, în care „peste" din prima propoziţie este adnotat cu o etichetă substantivală, iar acelaşi cuvânt din a doua propoziţie este adnotat cu o etichetă prepoziţională.

În română, cel puţin 30% dintre cuvintele dintr-o propoziţie folosesc semne diacritice, cu o medie de 1.16 semne diacritice per cuvânt. Doar aproximativ 12% dintre aceste cuvinte pot fi transformate imediat în versiunea lor cu diacritice (întrucât forma fără diacritice nu este un cuvânt valid în dicţionarul limbii române). Pentru restul cuvintelor, este util programul de descoperire a diacriticelor.

În afară de corectoarele de limbă şi sistemele suport pentru autori, verificarea limbii este importantă şi în domeniul învăţării limbilor cu ajutorul calculatorului şi se foloseşte la corectarea automată a întrebărilor introduse în motoarele de căutare pe Web: vezi sugestiile *„Aţi vrut să scrieţi…"* din Google.

Căutarea pe Web

Căutarea pe Web, în Intranet sau în biblioteci digitale este, probabil, cea mai folosită şi totuşi cea mai subdezvoltată tehnologie a limbajului astăzi. Motorul de căutare Google, care a apărut în 1998, este folosit pentru 80% dintre căutările la nivel mondial[xiv]. Nici interfaţa de căutare, nici prezentarea rezultatelor căutării nu s-au schimbat semnificativ de la prima versiune. În actuala versiune, Google oferă corectarea grafică a cuvintelor scrise greşit şi, din 2009, a încorporat abilităţi de căutare semantică elementară în pachetul de algoritmi[xv], ceea ce poate îmbunătăţi acurateţea căutării prin analiza sensului termenilor din fraza de interogare în context. Povestea de succes a Google dovedeşte că, dispunând de o cantitate uriaşă de date şi de tehnici eficiente de indexare a acestora, o abordare în principal statistică poate conduce la rezultate satisfăcătoare.

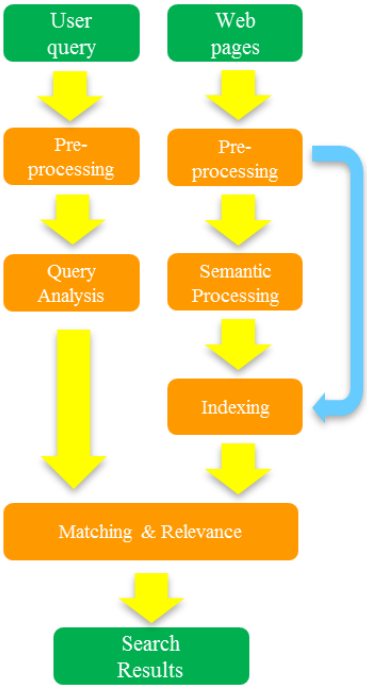

Figure 4: Web Search Architecture

Cu toate acestea, pentru o căutare mai sofisticată de informaţii, este esenţială integrarea cunoştinţelor lingvistice mai profunde. În laboratoarele de cercetare, experimentele care folosesc tezaure în format electronic şi resurse lingvistice de tip ontologie precum WordNet (sau echivalentul său românesc Romanian WordNet[xvi]) au demonstrat îmbunătăţiri ale rezultatelor prin apelul la sinonime ale termenilor de căutare, de exemplu *energie atomică* ori *energie nucleară* sau chiar termeni mai îndepărtaţi semantic.

Generaţia următoare a motoarelor de căutare va trebui să includă tehnologii ale limbajului mult mai sofisticate. Dacă o frază de interogare constă dintr-o întrebare sau alt tip de propoziţie, nu dintr-o listă de cuvinte-cheie, găsirea răspunsurilor relevante la această frază necesită analiza propoziţiei la nivel sintactic şi semantic, precum şi existenţa unui index care să permită găsirea rapidă a documentelor relevante. De exemplu, imaginaţi-vă că un utilizator introduce fraza de interogare:

> *Dă-mi o listă cu toate companiile care au fost preluate de alte companii în ultimii cinci ani.*

Pentru a găsi un răspuns satisfăcător, trebuie efectuată analiza sintactică pentru a identifica structura propoziţiei şi a stabili faptul că utilizatorul caută companii care au preluat alte companii. De asemenea, expresia *în ultimii cinci ani* trebuie prelucrată pentru a stabili la ce ani se referă.

În sfârşit, trebuie încercată potrivirea dintre fraza prelucrată şi o cantitate uriaşă de date nestructurate pentru a găsi informaţia căutată de utilizator. Acest proces este cunoscut sub numele de regăsirea informaţiei şi presupune căutarea şi ordonarea documentelor relevante. În plus, generând o listă de companii trebuie să extragem şi informaţia că un anumit şir de cuvinte dintr-un document se referă la numele unei companii. Acest tip de informaţie ne este furnizat de programele de recunoaştere a entităţilor numite.

Şi mai solicitantă este încercarea de potrivire a frazei de interogare cu documente scrise în altă limbă. Pentru regăsirea informaţiei la nivel interlingual trebuie să traducem automat fraza de interogare în toate limbile sursă posibile şi să transferăm informaţia regăsită în limba ţintă. Procentul în creştere de date disponibile în format non-text necesită servicii care să permită regăsirea informaţiei multimedia, i.e. căutarea informaţiei în date de tip imagine, audio şi video. Pentru fişierele audio şi video, aceasta presupune un modul de recunoaştere a vorbirii care converteşte conţinutul de vorbi-

re în text sau într-o reprezentare fonetică, la care se încearcă potrivirea frazei de interogare.

În România, tehnologiile de căutare bazate pe limba naturală nu sunt încă vizate de aplicaţiile industriale. În schimb, tehnologiile de tip *open source* precum Lucene sunt adesea folosite de companiile care fac căutări pentru a furniza infrastructura elementară de căutare. Totuşi, grupurile de cercetare de la Universitatea Al. I. Cuza (UAIC) şi de la RACAI au dezvoltat diverse module care constituie partea centrală a unui instrument de căutare semantică, precum analizoare morfo-sintactice, analizoare sintactice, analizoare semantice, programe de recunoaştere a entităţilor numite, instrumente de indexare, programe de regăsire a informaţiei multimedia etc. Acoperirea şi eficienţa lor sunt, totuşi, destul de limitate.

Astfel, la RACAI, un analizor morfo-sintactic capabil să identifice forma de dicţionar şi partea de vorbire a cuvintelor din text este disponibil ca serviciu web[xvii]. De exemplu, dacă fraza de interogare a unui utilizator pentru o căutare pe web conţine cuvântul „evenimente", poate fi utilizată rădăcina („eveniment") pentru a efectua căutarea[xviii].

Alt modul dezvoltat de cercetătorii de la UAIC şi de la RACAI este un program de recunoaştere a entităţilor numite, care, dându-se un text conţinând nume de persoane, de companii, de organizaţii, de evenimente etc. (toate acestea cunoscute ca entităţi numite), poate să le recunoască în text. Pentru propoziţia

Maria şi-a luat bilet la concertul trupei din vară de la Paris.

acest sistem recunoaşte „Maria" ca nume de persoană, „din vară" ca o referinţă temporală, iar „Paris" ca nume de loc.

Un analizor semantic dezvoltat la UAIC[xix], disponibil pentru limba română, poate identifica într-o propoziţie rolurile semantice diferite ale entităţilor. De exemplu, pentru propoziţia de mai sus, sistemul identifică „Maria" ca persoana care face acţiunea, „bilet la concertul trupei" ca obiectul care a fost cumpărat. Asemănător, în exemplul

Maria şi-a luat fără ezitare bilet pentru a-şi vedea trupa preferată

sistemul recunoaşte „fără ezitare" ca *modalitatea* în care Maria şi-a cumpărat biletul, iar „pentru a-şi vedea trupa preferată" reprezintă *scopul* pentru care biletul a fost achiziţionat.

Recent, un grup de cercetători de la UAIC au început o cercetare pentru detectarea şi adnotarea automată a imaginilor, în vederea dezvoltării unui instrument de căutare a imaginilor[xx]. Sistemul este încă într-o fază incipientă.

Interacţiunea vocală

Tehnologia interacţiunii vocale reprezintă baza pentru crearea de interfeţe care să permită utilizatorului să interacţioneze cu maşinile utilizând limba vorbită mai degrabă decât, de exemplu, o interfaţă grafică, o tastatură şi un mouse. Astăzi, interfeţele vocale cu utilizatorul (VUI – Vocal User Interface) sunt de obicei utilizate pentru servicii complet sau parţial automatizate furnizate de companii, prin telefon, clienţilor, angajaţilor, sau partenerilor. Domenii de afaceri care se bazează foarte mult pe VUI sunt băncile, logistica, transportul public şi telecomunicaţiile. Alte utilizări ale tehnologiei interacţiunii vocale sunt interfeţele pentru anumite dispozitive, ca

de exemplu sistemele de navigare ale autovehiculelor, şi utilizarea limbii vorbite ca alternativă la modalităţile de input/output ale interfeţelor grafice, ca de exemplu pe smartphone-uri.

Interacţiunea vocală cuprinde următoarele patru tehnologii diferite:

- Recunoaşterea automată a vorbirii (RAV) este responsabilă pentru identificarea cuvintelor care au fost rostite într-o secvenţă de sunete rostită de utilizator.
- Analiza sintactică şi interpretarea semantică presupune analizarea structurii sintactice a enunţului utilizatorului şi interpretarea acestuia conform scopurilor sistemului în care este integrată tehnologia.
- Managementul dialogului este necesar pentru determinarea acţiunii care va fi efectuată dat fiind inputul utilizatorului şi funcţionalitatea sistemului.
- Sinteza vorbirii (Text-to-Speech – TTS) este utilizată pentru transformarea cuvintelor unui enunţ în sunetele care vor constitui output-ul pentru utilizator.

Una dintre provocările majore este realizarea unui sistem RAV care să recunoască cuvintele enunţate de utilizator cât mai precis cu putinţă. Acest lucru necesită fie o restrângere a domeniului enunţurilor posibile la un set limitat de cuvinte cheie, fie crearea manuală a unor modele de limbă care să acopere un interval larg de enunţuri în limbaj natural. În timp ce prima soluţie rezultă într-o utilizare a unei VUI mai degrabă rigide şi inflexibile şi produce o acceptare slabă din partea utilizatorilor, opţiunea creării, reglării şi menţinerii unor modele de limbă poate creşte semnificativ costurile. Totuşi, VUI care utilizează modele de limbă şi permit iniţial utilizatorului să-şi exprime intenţia în mod flexibil prezintă atât o rată mai mare de automatizare cât şi o acceptare mai mare din partea utilizatorului şi pot fi, de aceea, considerate ca avantajoase în raport cu o abordare mai puţin flexibilă, precum cea a dialogului dirijat.

Pentru componenta de output a unei VUI, companiile tind să utilizeze enunţuri pre-înregistrate ale unor vorbitori profesionişti. Pentru enunţuri statice, în care rostirea nu depinde de contextul particular de utilizare sau de datele personale ale unui anumit utilizator, această soluţie va conduce la o experienţă utilă pentru utilizator. Totuşi, cu cât conţinutul unui enunţ este mai dinamic, cu atât experienţa utilizatorului are de suferit datorită unei prozodii sărace, rezultate din concatenarea diferitelor fişiere audio conţinând diverse silabe-cuvinte. Prin contrast, sistemele TTS de azi se dovedesc a fi superioare (chiar dacă optimizabile) în ceea ce priveşte naturaleţea prozodică a enunţurilor dinamice.

În ceea ce priveşte piaţa tehnologiei interacţiunii vocale, ultimul deceniu a adus o puternică standardizare a interfeţelor dintre diferitele componente tehnologice. A avut loc o puternică consolidare a pieţei în ultimii zece ani, cu precădere în domeniile RAV şi TTS. Aici, pieţele naţionale din ţările blocului G20 – altfel spus ţări puternice din punct de vedere economic şi cu o populaţie considerabilă – sunt dominate de mai puţin de 5 actori mondiali, Nuance şi Loquendo fiind cei mai proeminenţi din Europa.

Domeniul recunoaşterii şi analizei vorbirii este unul dintre cele mai puţin reprezentate în România. Pe piaţa românească de TTS, există soluţii comercializate de companii internaţionale (precum MBROLA sau IVONA), dar rezultatele prezintă o acurateţe şi o

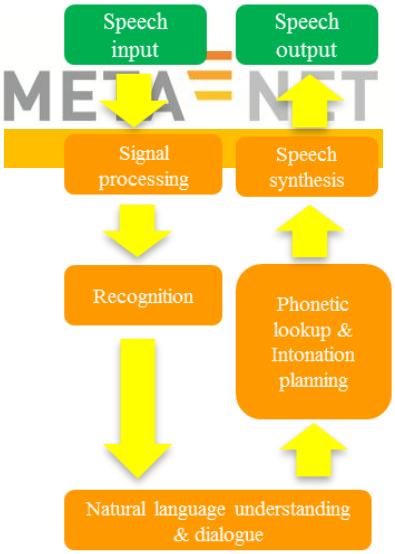

Figure 5: Simple Speech-based Dialogue Architecture

fluenţă redusă. Companiile de echipamente auto sau de telecomunicaţii, precum Continental şi Orange, au început recent să aloce resurse pentru departamente specializate în procesarea vorbirii, adaptând soluţii deja existente nevoilor lor specifice. Pe de altă parte, cercetări în această direcţie au loc la Universitatea Bucureşti şi la Institutul de Informatică Teoretică al Academiei Române, Filiala Iaşi. Majoritatea cercetătorilor se concentrează pe sinteza vorbirii, în timp ce interpretarea vorbirii nu este încă dezvoltată.

Privind dincolo de starea actuală a tehnologiei, preconizăm schimbări semnificative datorită răspândirii smartphone-urilor ca o nouă platformă pentru administrarea relaţiilor cu clienţii, alături de canalele mai vechi precum telefon, internet şi email. Această tendinţă va afecta şi utilizarea tehnologiei pentru Interacţiune prin Voce. Pe de o parte, cererea pentru VUI bazate pe telefonie va scădea pe termen lung. Pe de altă parte, utilizarea limbii vorbite ca o modalitate facilă de input pentru smartphone-uri va căpăta o importanţă semnificativă. Această tendinţă este sprijinită de progresul evident al acurateţei recunoaşterii vorbirii independente de vorbitor în cadrul serviciilor de dictare a vorbirii, care sunt deja oferite ca servicii centralizate utilizatorilor de smartphone-uri. Având în vedere "externalizarea" sarcinii de recunoaştere a vorbirii către infrastructura aplicaţiilor, utilizarea tehnologiilor lingvistice într-o manieră specifică unei anumite aplicaţii va câştiga în importanţă în raport cu situaţia actuală.

Traducerea automată

Ideea de a folosi calculatoarele pentru traducere a apărut în 1946 la A.D. Booth şi a fost urmată de finanţare substanţială pentru cercetări în acest domeniu între anii 1950 - 1980. Cu toate acestea, traducerea automată (TA) încă nu s-a ridicat la nivelul aşteptărilor ridicate stabilite în primii ani de la apariţia domeniului.

În cea mai simplă formă, TA înlocuieşte pur şi simplu cuvintele dintr-o limbă cu cuvintele din altă limbă. Acest lucru poate fi util în domenii cu limbaj foarte restrâns, formalizat, cum sunt de exemplu rapoartele meteo. Însă, pentru o traducere bună a unor texte mai puţin standardizate, trebuie potrivite elemente mai lungi din text (expresii, propoziţii, sau chiar pasaje întregi) cu fragmentele lor echivalente din limba ţintă. Dificultatea majoră aici constă în faptul că limbajul uman este ambiguu, ceea ce ridică provocări pe mai multe niveluri, de exemplu dezambiguizarea sensurilor cuvintelor la nivel lexical (Jaguar poate însemna fie o maşină fie un animal) sau ataşarea corectă a grupurilor prepoziţionale la nivel sintactic, ca în:

Poliţistul a văzut omul cu telescopul.

Poliţistul a văzut omul cu arma.

Una din modalităţile de abordare a traducerii automate se bazează pe reguli lingvistice. Pentru traduceri între limbi strâns legate, o traducere directă poate fi fezabilă în cazuri precum cele din exemplul de mai sus. Dar, de cele mai multe ori, sistemele bazate pe reguli (sau bazate pe cunoştinţe) analizează textul de intrare şi creează o reprezentare intermediară, simbolică, pe baza căreia este generat textul pentru limba ţintă. Succesul acestor metode depinde în mare măsură de disponibilitatea unor lexicoane extinse cu informaţii morfologice, sintactice şi semantice, precum şi de existenţa unor seturi mari de reguli gramaticale atent proiectate de lingvişti calificaţi.

Începând cu sfârşitul anilor 1980, pe măsură ce puterea de calcul a crescut şi a devenit mai puţin costisitoare, au început să atragă interes modelele statistice pentru TA. Parametrii acestor modele statistice sunt derivaţi din analiza corpusurilor de texte bilingve, cum este corpusul paralel Europarl, care conţine lucrările Parlamentului European în 11limbi europene. Având date suficiente, TA statistică funcţionează suficient de bine pentru a obţine un înţeles aproximativ al unui text într-o limbă străină. Cu toate acestea, spre deosebire de sistemele bazate pe cunoştinţe, sistemele de TA statistică (sau bazate pe date) generează de multe ori texte incorecte gramatical. Pe de altă parte, pe lângă avantajul că este necesar mai puţin efort uman pentru scris reguli gramaticale, sistemele de TA bazate pe date pot de asemenea acoperi particularităţi ale limbii care lipsesc în sistemele cunoştinţe bazate pe cunoştinţe, de exemplu expresii idiomatice.

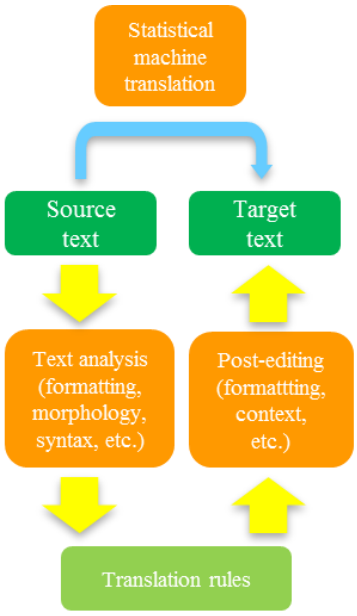

Figure 6: Machine translation (top: statistical; bottom: rule-based)

Deoarece avantajele şi dezavantajele sistemele de TA bazate pe date şi a celor bazate pe cunoştinţe sunt complementare, cercetătorii folosesc în prezent aproape în unanimitate abordări hibride, care combină cele două metodologii. Acest lucru poate fi realizat în mai multe moduri. Unul presupune folosirea atât a sistemelor de TA bazate pe cunoaştere cât şi a celor bazate pe date, iar apoi un modul de selecţie decide care dintre cele două traduceri este mai bună pentru fiecare propoziţie. Cu toate acestea, pentru propoziţii mai lungi, nici una dintre traduceri nu va fi perfectă. O soluţie mai bună este de a combina, pentru fiecare propoziţie, secvenţele traduse corect de sisteme diferite. Sarcină destul de complexă, deoarece nu este totdeauna evident care sunt părţile traduse corect de fiecare sistem, şi este necesară o aliniere.

Folosirea traducerii automate poate îmbunătăţi semnificativ productivitatea cu condiţia unei bune adaptări la terminologia specifică utilizatorului. Sisteme speciale pentru ajutor la traducerea interactivă au fost dezvoltate, în vreme ce portaluri lingvistice oferă acces la dicţionare şi terminologii specific companiilor, memorii de traducere şi suport pentru TA.

Tabelul 1, prezentat în cadrul proiectului European Euromatrix+, arată performanţele obţinute de sisteme la traducerile automate încrucişate în cele 22 de limbi oficiale ale Uniunii Europene (galeza irlandeză lipseşte), raportate la scorul BLEU[xxi].

Cele mai bune rezultate (în verde şi albastru) le au limbile care beneficiază de eforturi de cercetare considerabile în domeniul TA în cadrul unor programe coordinate, şi de existenţa unor corpusuri substanţiale paralele (ex. engleză, franceză, olandeză, spaniolă, germană). Rezultatele cele mai slabe (în roşu) sunt obţinute de limbi care nu beneficiază de eforturi similar, sau care sunt foarte diferite din punct de vedere al comportamentului lingvistic faţă de alte limbi (ex. Ungară, malteză, finlandeză).

Calitatea sistemelor de TA este considerată ca având potenţial de îmbunătăţire. Modificările includ adaptabilitatea resurselor lingvistice la diferite domenii sau utilizatori, precum şi integrarea în platforme existente cu memorii de traducere sau forme de bază a termenilor.

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Tabel 1: Performanţe obţinute la traducerile automate în cele 22 de limbi oficiale ale Uniunii Europene (sursa: Euromatrix+)

Domeniul traducerii automate este, în ochii investitorilor, cel mai atractiv domeniu dintre tehnologiile limbajului. Astfel, companii precum Language Weaver lucrează în domeniul traducerilor din/spre română folosind diferite tehnologii lingvistice. Sistemul major online de traducere automată cuprinde limba română atât ca limbă sursă, cât şi ca limbă ţintă. De asemenea, există în mediul online o multitudine de dicţionare pentru limba română.

Eforturi importante de cercetare au fost şi continuă să fie dedicate de cercetători din diferite centre domeniului traducerii automate cu româna ca limbă sursă sau limbă ţintă. Rezultate bune sunt raportate pentru un experiment de traducere bazată de date pentru perechea de limbi română-engleză, raportate la rezultatele sistemului Google Translate pentru aceeaşi pereche de limbi.[xxii].

La RACAI, de mai bine de 5 ani se experimentează cu diferite abordări de genul traducere automată bazată pe exemple, traducere automată statistică, extragerea de traduceri din corpusuri paralele etc. Două teze de doctorat, însoţite de mai multe articole ştiinţifice şi susţinute de diferite proiecte naţionale sau internaţionale, precum STAR şi ACCURAT, sunt dedicate acestui domeniu[xxiii],[xxiv].

Tehnologiile limbajului

Construirea de aplicaţii bazate pe tehnologiile limbajului implică o varietate de sub-probleme care nu apar întotdeauna la nivelul interacţiunii cu utilizatorul, dar oferă funcţionalităţi semnificative "în cilisele" sistemului. Din acest motiv, ele constituie domenii importante de cercetare care au devenit discipline de sine stătătoare ale lingvisticii computaţionale.

Sistemele de Întrebare-Răspuns (ÎR) reprezintă o zonă importantă a cercetării, pentru care au fost construite corpusuri adnotate şi au fost iniţiate competiţii ştiinţifice. Idea este trecerea de la căutarea bazată pe cuvinte cheie (la care sistemul răspunde print-o colecţie de documente cu posibilă relevanţă) la scenariul în care utilizatorul pune o întrebare concretă şi sistemul oferă un singur răspuns : 'La ce vârstă a păşit Neil Armstrong pe lună?' - '38'. În vreme ce acest domeniu este în mod evident legat de domeniul Căutării pe Internet, sistemele ÎR sunt acum un termen general pentru cercetări de

genul ce *tipuri* de întrebări există şi cum trebuie tratate, cum poate o colecţie de documente cu un posibil răspuns să fie analizată şi comparată (în cazul răspunsurilor conflictuale) şi cum poate fi extras dintr-un document răspunsul (informaţie specifică) fără a ignora contextul. ÎR pot fi folosite cu succes pentru a identifica răspunsuri de tipul locaţie, persoană, organizaţie, dată, măsură, număr.

Acest domeniu este strâns legat de cel al extragerii informaţiei (EI), o zonă care a fost extrem de populară şi influenţială în perioada statistică a lingvisticii computaţionale, la începutul anului 1990. EI tinde să identifice bucăţi de informaţie în diferite clase de documente; de exemplu, detectarea persoanelor cheie în preluările de companii, după cum sunt raportate în ziare. Alt scenariu care a fost luat în considerare e reprezentat de rapoartele asupra incidentelor teroriste, unde problema se reduce la potrivirea pe text a unui şablon care specifică atentatorul, ţinta, locaţia şi momentul incidentului, precum şi rezultatul acestuia. Caracteristica principală a EI este completarea unor şabloane specifice fiecărui domeniu, din acest motiv fiind un exemplu de tehnologie din culise care constituie o arie de cercetare bine delimitată, dar pentru a avea aplicabilitate practică trebuie inclusă într-un mediu specific.

Două zone de limită, care uneori joacă rolul de aplicaţii independente, alte ori de componentă din culise, sunt rezumarea şi generarea de text. Rezumarea se referă evident la scurtarea unui text lung, şi este oferită ca funcţionalitate de exemplu în MS Word. Una din abordările rezumării automate are baze statistice, identificând cuvinte "importante" din text (de exemplu cuvinte care au frecvenţă mare în text şi care sunt mai puţin frecvente în utilizarea comună a limbajului) şi apoi determinând acele propoziţii care conţin aceste cuvinte importante. Propoziţiile sunt apoi marcate în document, sau extrase din el, pentru a constitui rezumatul. În acest scenariu, rezumatul este o extragere de propoziţii iar textul este redus la un subset din propoziţiile sale. Majoritatea sistemelor de re zumare automată folosesc această metodă.

Un dezavantaj al acestei abordări este faptul că ignoră expresiile referenţiale care pot apărea în textul iniţial şi să fie păstrate în rezumat. Dacă datorită eliminării de propoziţii, antecedentul acestor referinţe nu mai este prezent, rezumatul rezultat poate deveni de neînţeles. De exemplu, pentru textul:

> *Hercule, dintre toţi copii nelegitimi ai lui Zeus, părea să fie centru mâniei Herei. Pe când el era doar un copil, ea a trimis un şarpe cu două capete să-l atace.*

rezumatul acestui fragment ar putea fi, folosind metoda de eliminare a propoziţiilor:

> *Ea a trimis un şarpe cu două capete să-l atace.*

ceea ce este destul de greu de înţeles dacă nu există nici o explicaţie despre cine este "ea" sau "el" (din cliticul –l se înţelege că există o persoană atacată care este de genul masculin).

O modalitate de a spori coerenţa acestor rezumate este de a deriva iniţial structura de discurs a textului şi de a ghida selecţia propoziţiilor care urmează a fi introduse în rezumat folosind un scor care să ţină cont şi de relevanţa propoziţiei în discurs, dar şi de coerenţa textului, rezultată din rezoluţia anaforelor[xxv]. Pentru rezumatul dat ca exemplu mai sus, rezoluţia anaforelor presupune identificarea

relaţiei dintre “ea” şi “Hera” şi dintre “-l” şi “Hercule”. Astfel, rezumatul devine inteligibil:

Hera a trimis un şarpe cu două capete să-l atace pe Hercule.

Sistemul de rezumare automată dezvoltat de UAIC a adoptat această metodă, producând rezumate foarte bune pentru texte de dimensiuni reduse[xxvi].

O metodă alternativă căreia îi sunt dedicate multe cercetări este sintetizarea de *noi* propoziţii, adică construirea unui rezumat din propoziţii care nu sunt neapărat în forma din rezumat şi în textul iniţial. Această metodă necesită o înţelegere mai profundă a textului (ceea ce este mai costisitor din punct de vedere al resurselor computaţionale şi mai greu de realizat), dar poate fi aplicată cu succes pentru texte mai lungi. De exemplu, pentru romane nu este relevant calculul celor mai frecvente cuvinte (pentru că acestea vor fi cuvintele funcţionale gen *şi*, *iar*, *dar*, *al,* etc.), şi nici structura de discurs (aceasta fiind mult prea stufoasă). În aceste cazuri, alte metode, exemplu expandarea unui set de şabloane flexibile predefinite (bazate de exemplu pe genul romanului, sau pe anumite informaţii despre personajele principale, timpul sau locaţia intrigii) pot fi aplicate.

Un generator de text nu este în majoritatea cazurilor o aplicaţie de sine stătătoare, ci inclusă într-o platformă software mai largă, aşa cum într-un sistem de management medical sunt colectate, stocate şi procesate informaţii despre pacient, iar generarea rapoartelor este doar o funcţionalitate.

Limba română, ca limbă ţintă pentru cercetările din toate aceste domenii, este mai puţin atractivă decât limba engleză, unde sistemele de întrebare-răspuns, de extragere de informaţii sau de rezumare automată au fost încă din anii 1990 subiectul a numeroase competiţii, precum cele organizate de DARPA/NIST în Statele Unite sau campaniile CLEF în Europa. Totuşi, echipe de cercetători români de la UAIC şi RACAI au participat începând cu anul 2006 la competiţii de întrebare-răspuns cu rezultate destul de bune[xxvii]. Principalul dezavantaj este dimensiune redusă a corpusurilor adnotate sau alte resurse necesare dezvoltării acestor domenii. Sistemele de rezumare automată, dacă folosesc doar metode statistice, sunt în mare măsură independente de limbă, astfel că există prototipuri care pot fi aplicate şi pentru limba română. La UAIC, un instrument de rezumare bazat pe structura de discurs şi rezoluţia anaforei este disponibil pentru texte în limba română.

Domenii adiacente în care cercetători români au fost implicaţi cuprind lexicologie computaţională, e-learning şi analiza sentimentelor şi a opiniilor.

Un consorţiu de cinci institute de cercetare din România şi o universitate (UAIC) a fost implicat recent în transformarea în format electronic a Dicţionarului Tezaur al Limbii Române, care însumează 35 de volume, redactate din 1913 până în prezent. Obiectivul principal a fost transformarea celor aprox. 15.000 de pagini ale dicţionarului într-un format electronic structurat, care să permită căutări complexe, dar şi o editare şi activitate de actualizare mai uşoară[xxviii].

Alt acces la materialul lexicografic al limbii este facilitat de reţelele semantice sub formă de wordnets (reţele de cuvinte). WordNetul românesc a fost în lucru timp de 8 ani, şi totalizează mai mult de 52.000 de serii sinonimice (synset-uri) în care apar aprox. 60.000

de cuvinte, distribuite în patru categorii de părţi de vorbire: substantive, verbe, adjective şi adverbe. Fiecare synset conţine un set de cuvinte (cu un număr de sensuri asociate) care sunt sinonime. Synset-urile sunt noduri ale reţelei, în timp ce arcele sunt relaţiile semantice dintre synset-uri: hiponimie (relaţia *is-a,* care specifică că X este un fel de Y), meroninime, înlănţuire, cauză, şi altele. WordNetul românesc este aliniat cu Princeton WordNet[xxix] (varianta pentru limba engleză), primul şi cel mai mare wordnet dintre cele existente pentru diferite limbi. Synset-urile au etichete DOMENIU: fiecare synset este etichetat cu numele domeniului în care este folosit. Mai mult, WordNetul românesc este aliniat cu cea mai mare ontologie disponibilă gratuit, SUMO&MILO[xxx], şi este folosit în diverse aplicaţii dezvoltate pentru limba română: sisteme de întrebare-răspuns, dezambiguizare a sensurilor cuvintelor, traducere automată.

Un alt domeniu în care cercetătorii din UAIC au fost implicaţi este e-learning, prin încorporarea instrumentelor multilingve de tehnologie a limbajului şi tehnici de semantică web pentru îmbunătăţirea regăsirii de materiale de învăţare. Tehnologia dezvoltată facilitează accesul personalizat la cunoaştere în cadrul sistemelor de gestionare a învăţării şi ajută la operarea colectivă a datelor în gestionarea conţinutului.

Cel mai nou domeniu de interes pentru tehnologiile limbajului este analiza sentimentelor şi a opiniilor. Astfel, fiind dat un text, un program identifică dacă textul are o încărcătură emoţională pozitivă sau negativă. Cercetări în acest domeniu au început la RACAI cu dezvoltarea SentiWordNet, o adnotare la sentimente a WordNetului românesc[xxxi]. La UAIC, cercetări în această direcţie au implicat colaborarea cu o companie privată, Sentimatrix, pentru dezvoltarea unui sistem capabil să monitorizeze web-ul şi să extragă opinia utilizatorilor (din forumuri, bloguri, reţele sociale, etc.) referitoare la diferite produse[xxxii].

Tehnologiile limbajului în educaţie

Tehnologiile limbajului sunt un domeniu interdisciplinar, care implică expertiza lingviştilor, informaticienilor, statisticienilor, psiholingviştilor. Până acum nu şi-a stabilit un loc fix în sistemul de învăţământ din România. Multe universităţi din România şi din Republica Moldova au introdus recent cursuri de prelucrare a limbajului natural şi lingvistică computaţională la nivelul studiilor universitare, de masterat şi doctorat. Din 2001, un masterat în lingvistică computaţională a fost introdus în curricula Facultăţii de Informatică a Universităţii Alexandru Ioan Cuza din Iaşi. Totuşi trebuie conceput un sistem consolidat de educaţie superioară în procesarea limbajului natural şi lingvistică computaţională.

Industria TL şi programe

Industriile care folosesc şi furnizează TL în România sunt cu siguranţă importante (BitDefender, Continental, Nokia, etc.), dar este necesară o mai bună colaborare între ele. O problemă importantă care trebuie rezolvată este "caracterul secret" al TL, care ar putea fi rezolvată printr-o strategie bună de marketing. Industria limbajului nu este un angajator important în România, puţine companii din domeniul Tehnologiilor Informaţiei şi Comunicării (TIC) având deja departamente de TL.

Programe naţionale anterioare au avut un impuls iniţial, dar lipsa ajutorului financiar consecvent sau destul de atractiv a dus la pierderea interesului marilor companii de TIC şi a tinerilor cercetători,

formaţi de universităţi şi de institutele de cercetare. Unul dintre programele de colaborare dintre industrie şi educaţie care a avut un impact pozitiv şi rezultate bune în România în domeniul TL este Alianţa Academică MSDN, care oferă acces gratuit studenţilor la diferite tehnologii Microsoft.

Principalele laboratoare de cercetare cu activitate în domeniul TL în România sunt: RACAI la Academia Română din Bucureşti; Departamentul de cercetare al Facultăţii de Informatică al Universităţii Alexandru Ioan Cuza University din Iaşi; Institutul de Informatică Teoretică al Academiei Române din Iaşi, care găzduieşte arhiva Sunetele Limbii Române – un repozitoriu online de sunete ale limbii române înregistrate. În ceea ce priveşte programele de cercetare, UAIC şi RACAI au fost implicate în mai multe proiecte de cercetare naţionale şi internaţionale care-şi propun să dezvolte tehnologii ale limbajului existente sau noi. Printre acestea pot fi menţionate proiectele europene: ACCURAT-RO (Analiza şi evaluarea corpusurilor comparabile pentru domenii cu puţine resurse pentru traducere automată), STAR (Sistem pentru traducere automată pentru limba română), proiectul PC7 CLARIN (Infrastructură interoperabilă de resurse lingvistice pentru limba română), BALKANET (Construirea unei reţele de wordnet-uri pentru limbile balcanice), proiectul PC6 LT4eL (Tehnologii ale limbajului pentru e-learning), proiectul INTAS RoLTech (platformă pentru tehnologiile limbajului pentru limba română: resurse, instrumente şi interfeţe) etc. Au existat de asemenea proiecte cu finanţare naţională precum: SIR-RESDEC (Sistem de întrebare răspuns pentru domeniu deschis pentru limbile română şi engleză), ROTEL (Sisteme inteligente pentru web semantic, bazate pe logica ontologiilor şi pe TL), eDTLR (Dicţionarul Tezaur al Limbii Române în format electronic), printre altele.

Piaţa pentru tehnologiile limbajului poate fi doar estimată, şi mai mult ca sigur va primi un impuls prin platformele mobile, de tipul Apple iPad şi alte produse similare, jocuri (educaţionale) etc.

Cercetarea în domeniul TL și educația

Cele mai reprezentative centre în lingvistica computaţională a limbii române sunt în România la Bucureşti, Iaşi, Cluj-Napoca, Timişoara şi Craiova, iar în Republica Moldova la Chişinău.

Punctele comune de întâlnire ale celor mai mulţi cercetători din domeniul TL sunt, pe lângă conferinţele internaţionale din străinătate, o serie de evenimente internaţionale şi naţionale care adună tinerii şi cercetători cu experienţă, lingvişti şi informaticieni ţinute periodic în România: Evenimentele consorţiului pentru Informatizarea Limbii Române – ConsILR[xxxiii], seria de şcoli de vară internaţionale EUROLAN, conferinţele SPED – Tehnologiile vorbirii şi dialog om – calculator, conferinţele KEPT – Ingineria cunoaşterii: principii şi tehnici, conferinţele ECIT – conferinţa europeană pe domeniul sistemelor şi tehnologiilor inteligente etc.

Lingvistica computaţională este un domeniu exotic şi este localizat fie la facultăţi de informatică fie la facultăţi de ştiinţe umaniste. Acest lucru este un dezavantaj pentru domeniul LT, deoarece studiul lingvisticii computaţionale este astfel orientată fie pe aspecte lingvistice, fie pe cele de inginerie, iar suprapunerile sunt doar parţiale. Alt dezavantaj al acestui peisaj este implicarea minoră a companiilor din domeniul TIC în cercetarea în TL (deşi recent au început să fie prezente în viaţa educaţională).

Situația instrumentelor și resurselor pentru limba română

Tabelul următor oferă o privire de ansamblu asupra situației prezente a tehnologiilor limbajului pentru limba română. Evaluarea tehnologiilor şi resurselor existente este bazată pe estimarea mai multor experţi din domeniu, folosind următoarele criterii (fiecare de la 0 la 6).

1. **Cantitate**: Există pentru limba română instrumentul sau resursa respectivă? Cu cât există mai multe instrumente /resurse cu atât este mai mare scorul.
   - 0: nici un instrument/resursă;
   - 6: multe instrumente/resurse, varietate mare.
2. **Disponibilitate**: Sunt accesibile instrumente/resurse, sunt Open Source, gratuite, pot fi folosite pe orice platformă sau sunt disponibile doar pentru un preţ ridicat sau în condiţii foarte restrictive?
   - 0: toate instrumentele/resursele sunt disponibile pentru un preţ foarte mare;
   - 6: o mare parte dintre instrumente/resurse sunt disponibile gratuit cu licenţă Open Source sau Creative Commons care permit refolosirea lor.
3. **Calitate**: cât de bine sunt respectate criteriile de performanţă ale instrumentelor şi indicatorii de calitate ale resurselor de către cele mai bune instrumente, aplicaţii sau resurse? Sunt actualizate şi menţinute aceste resurse/instrumente?
   - 0: resursă/instrument de tip proof-of-concept;
   - 6: instrument de calitate înaltă, resursă cu adnotare manuală verificată.
4. **Acoperire**: În ce măsură cele mai bune instrumente îndeplinesc criteriile respective pentru acoperire (stil, gen, tipuri de texte, fenomene lingvistice, tipuri de intrări/ieşiri, număr de limbi pentru sistemele de TA etc.)? În ce măsură sunt reprezentative pentru limba ţintă resursele?
   - 0: resursă sau instrument pentru scop specific, caz particular, acoperire foarte mică, utilizabil doar pentru cazuri foarte specifice;
   - 6: acoperire foarte largă, instrument foarte robust, aplicare pe scară largă, multe limbi suportate.
5. **Maturitate**: poate fi considerat matur instrumentul / resursa, stabil, pregătit pentru piaţă? cele mai bune instrumente / resurse pot fi folosite aşa cum sunt sau trebuie adaptate? Performanţele unei astfel de tehnologii sunt adecvate şi pregătite pentru folosirea în producţie sau este doar un prototip? Un indicator poate fi dacă instrumentul/ resursa sunt acceptate de comunitate şi folosite cu succes în sisteme de TL.
   - 0: prototip preliminar, exemplu de resursă;
   - 6: componentă integrabilă/aplicabilă imediat.
6. **Sustenabilitate**: Cât de bine poate fi integrat în sistemele IT existente? Resursa/instrumentul îndeplineşte un anumit nivel de sustenanţă în ceea ce priveşte documentaţia/manuale, explicarea diferitelor cazuri de utilizare, interfeţelor etc.? Foloseşte medii de programare standard (pre-

cum Java EE)? Există standarde de cercetare /industriale şi dacă da instrumentul / resursa le respectă (format al datelor etc)?

- 0: format al datelor, API specific, ad hoc;
- 6: respectă standardele, documentat.

7 **Adaptabilitate**: Cât de uşor poate fi adaptat / extins la domeniul nou / problemă nouă / tip diferit de text instrumentul respectiv?

- 0: practic imposibil de adaptat instrumentul/resursa la altă problemă, chiar dacă sunt implicate multe resurse umane sau financiare;
- 6: nivel înalt de adaptabilitate, adaptare foarte uşoară şi eficientă.

Tabel al instrumentelor și resurselor pentru limba română

	Cantitate	Disponibilitate	Calitate	Acoperire	Maturitate	Sustenabilitate	Adaptabilitate
Tehnologia Limbajului (instrumente, tehnologii, aplicaţii)							
Tokenizare, Morfologie (tokenizare, parsarea părţilor de vorbire, analiza/generarea morfologică)	5	4	5	5	5	4	4
Parsare (sintactică de adâncime)	3	3	4	4	4	3	4
Semantica Propoziţii (WSD, structura argumentelor, roluri semantice)	4	3	4	4	3	4	4
Semantica Textului (coreferenţă, rezoluţia anaforei, context, pragmatică, inferenţă)	3	3	5	4	5	5	5
Procesare avandată de discurs (structura textuală, coereţă, stctura retorică /RST, argumentare, şabloane de text, typuri de text etc.)	3	3	4	3	3	3	3
Regăsirea de informaţii (indexarea de text, sisteme ÎR multimedia, sisteme ÎR multilingve)	3	4	5	5	5	5	5
Extragerea de informaţii (recunoaşterea numelor de entităţi, extragerea relaţiei dintre evenimente, recunoaşterea sentimentelor)	3	4	4	5	4	4	5
Generrea de Limbaju (generarea de propoziţii, generarea de rapoarte, generarea de text)	0	0	0	0	0	0	0
Rezumare, Sisteme de Întrebare-Răspuns, Tehnologii avasate de acces informaţional.	5	4	5	4	5	4	4
Traducere automată	3	4	4	3	4	4	4
Recunoaşterea vorbirii	2	1	3	2	2	2	2

	Cantitate	Disponibilitate	Calitate	Acoperire	Maturitate	Sustenabilitate	Adaptabilitate
Sinteza vorbirii	1	1	2	2	2	2	1
Gestionarea dialogului	0	0	0	0	0	0	0
Resurse Lingvistice (Resurse, date, baze de date, baze de cunoştinţe)							
Corpus de referinţă	1	1	1	2	2	1	2
Corpus sintactic (treebanks, arbori de dependenţă)	3	3	5	4	4	4	4
Corpusuri semantice	2	3	3	2	3	3	3
Corpusuri de discurs	2	2	3	2	2	3	2
Corpusuri paralele, memorii de traducere	5	6	5	4	6	6	5
Corpusuri pentru vorbire	3	2	4	2	3	3	3
Date multimedia şi multimodale (date text combinate cu date audio)	0	0	0	0	0	0	0
Modele de limbă	4	1	4	4	4	3	3
Lexicoane, terminologii	4	3	5	4	5	4	4
Gramatici	2	2	3	2	2	3	3
Thezaure, WordNets	4	3	4	4	5	5	4
Resurse ontologice pentru cunoaşterea asupra lumii	1	2	2	2	2	2	2

Principalul scop al tabelului nu este de a oferi o privire exhaustivă a domeniului, ci de a sprijini mesaje de nivel înalt, care sunt explicate în această secţiune:

- Chiar dacă în general toate domeniile TL sunt acoperite, există domenii care nu sunt încă avute în vedere pentru limba română de cercetători: generarea de limbaj, sisteme de gestionare a dialogului şi construirea de corpusuri multimodale.
- Deşi sunt disponibile diferite tehnologii de parsare pentru limba română, un corpus de referinţă care să fie refolosit pentru evaluarea automată a parsărilor nu există încă
- Procesarea vorbirii este momentan mult mai puţin dezvoltată decât alte domenii ale TL, în ceea ce priveşte corpusurile şi instrumentele.
- Dacă poate fi observată o atenţie semnificativă pentru domenii precum tokenizarea, semantica propoziţiilor sau sisteme de întrebare-răspuns, nu acelaşi lucru este valabil şi pentru domenii care se ocupă de domenii mai complexe precum analiza sintactică de adâncime sau procesarea avansată a discursului.

- Resursele pentru limba română sunt mai puţin reprezentative decât instrumentele, deşi sunt esenţiale pentru testarea instrumentelor create.
- Cu câteva excepţii cum ar fi serviciile web pentru procesări de bază ale limbajului, analiză morfologică, instrumente de întrebare-răspuns şi sisteme de traducere automată, sistemele existente pentru limba română nu sunt disponibile.
- Instrumentele pentru limba română au o acoperire largă pentru domenii privind semantica propoziţiei şi regăsirea de informaţii, dar sunt restrânse pentru celelalte probleme.
- Printre instrumentele existente de TL pentru limba română, cele mature sunt disponibile gratuit.
- Dacă instrumentele nu sunt în mod necesar menţinute activ, resursele pentru limba română au o calitate bună şi sunt în general sustenabile.
- Deoarece majoritatea instrumentelor sunt bazate pe modele de limbă sau folosesc tehnici de învăţare automată, adaptarea lor este în general posibilă, ceea ce nu se întâmplă în cazul resurselor.
- Scorurile pe care diferiţi experţi le-au dat aceluiaşi domeniu din TL au fost în general asemănătoare, în special în ceea ce priveşte disponibilitatea, ceea ce indică faptul că instrumentele şi resursele existente pentru limba română sunt diseminate pe scară largă. Uneori totuşi, privind sustenabilitatea şi acoperirea, experţii au dat scoruri care diferă cu mai mult de jumătate din scorul total. Principalele zone de dezacord au fost: corpusul de referinţă, corpusuri semantice, gramatici şi resurse ontologice.
- Rândul care conţine informaţii despre modele de limbă poate fi interpretat diferit, deoarece unii experţi au dat scoruri ţinând cont de modele pentru limbajul scris, în timp ce alţii au dat scoruri mai mici gândindu-se la modele pentru limbajul vorbit.

Concluzii

Acest document descrie stadiul actual al tehnologiei limbajului în general cu aplicaţii pentru limba română în particular, şi evidenţiază sprijinul pe care TL îl pot aduce limbii.

Cercetările din universităţi şi institute de cercetare au dus la dezvoltarea de sisteme de înaltă calitate, precum şi modele şi teorii aplicabile pe scară largă. Totuşi este aproape imposibil să introducă soluţii sustenabile şi standardizate în contextul dezvoltării reduse a resurselor lingvistice. Există o necesitate majoră de resurse, de la texte în limba română până la corpusuri adnotate, în care fenomene lingvistice particulare sunt evidenţiate de experţi. Cum cea mai bună sursă de texte sunt copiile electronice ale publicaţiilor, o campanie de conştientizare adresată editurilor, cu scopul de a le convinge să doneze o parte din textele lor pentru cercetare este mai mult decât necesară[xxxiv].

Multe resurse nu sunt standardizate; programe şi iniţiative de standardizare a datelor şi a formatelor interschimbabile sunt necesare.

Generarea de limbaj şi sistemele de gestionare a dialogului sunt domenii ale TL la început de drum pentru limba română, pentru care se pot dezvolta încă multe tehnologii, aplicaţii şi resurse. Tehnologiile şi corpusurile pentru vorbire necesită o atenţie deosebită în vederea alinierii limbii române la standardele celorlalte limbi europene.

Pentru dezvoltarea domeniilor TL în România, trebuie obţinută o implicare mai accentuată a guvernului prin mecanisme politice şi financiare adecvate. Acest document nu conţine o evaluare a fondurilor alocate de Guvernul României domeniilor TL, dar sprijinul redus al statului pentru acest domeniu a fost observat în mod repetat de cercetătorii activi din domeniu, şi se reflectă în numărul redus de proiecte finanţate din fonduri naţionale, în implicarea timidă în crearea de infrastructuri europene de genul CLARIN-ERIC, şi în lipsa interesului pentru cofinanţarea proiectelor ICT-PSP din domeniu (precum proiectul METANET4U, în cadrul căruia a fost luată iniţiativă creării acestui raport).

Despre META-NET

META-NET este o reţea de excelenţă finanţată de către Comisia Europeană. Reţeaua conţine în prezent 47 de membri din 31 de ţări europene. META-NET promovează Alianţa Tehnologică pentru o Europă Multilingvă (Multilingual Europe Technology Alliance - META), o comunitate de profesionişti şi organizaţii din domeniul tehnologiei limbajului din Europa în continuă creştere.

The Multilingual Europe Technology Alliance (META)

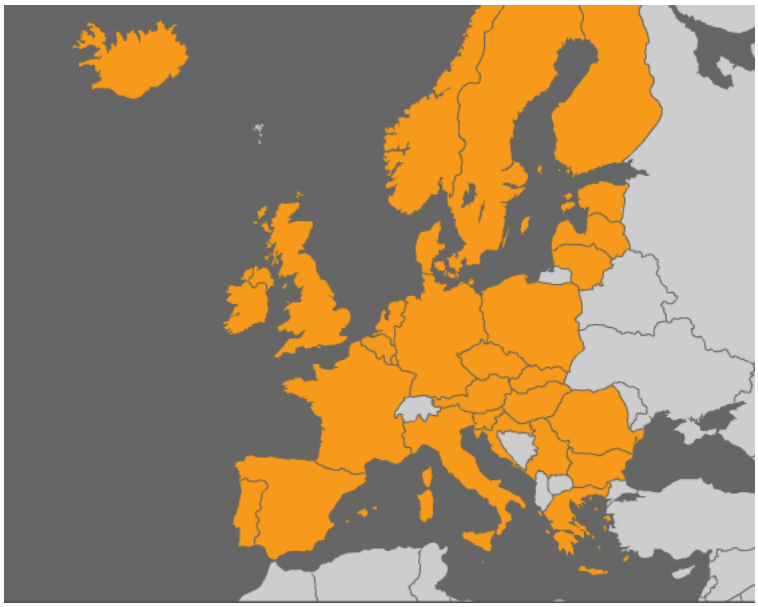

Figura 7: Ţările reprezentate în META-NET

META-NET cooperează cu alte iniţiative precum Infrastructura Comună pentru Resurse şi Tehnologii ale Limbajului (Common Language Resources and Technology Infrastructure - CLARIN), care ajută la stabilirea cercetării în domeniul ştiinţelor umaniste digitalizate în Europa. META-NET promovează fundamentele tehnologice pentru stabilirea şi menţinerea unei societăţi informaţionale europene cu adevărat multilingve, care:

- vor facilita comunicarea şi cooperarea între limbi diferite;
- vor asigura acces egal la informaţii şi cunoaştere în orice limbă;
- vor oferi funcţionalităţi ale tehnologiei informaţiei cetăţenilor europeni.

META-NET stimulează şi promovează tehnologiile multilingve pentru toate limbile europene. Tehnologiile limbajului activează traducerea automată, producerea de conţinut, procesarea informaţiilor şi gestionarea cunoştinţelor pentru o gamă largă de aplicaţii şi domenii. Reţeaua îşi propune să îmbunătăţească abordările curente, pentru a facilita o comunicare şi cooperare mai bună între limbi. Cetăţenii Europei au drept egal la informaţie şi cunoaştere, indiferent de limba vorbită.

Linii de acțiune

META-NET a fost lansată pe data de 1 Februarie 2010 cu scopul de a avansa cercetarea din domeniul tehnologiilor limbajului. Reţeaua sprijină o Europă unită într-o singură piaţă digitală şi spaţiu informaţional. META-NET a desfăşurat mai multe activităţi care au

avansat scopurile sale. META-VISION, META-SHARE şi META-RESEARCH sunt cele trei linii de acţiune ale reţelei.

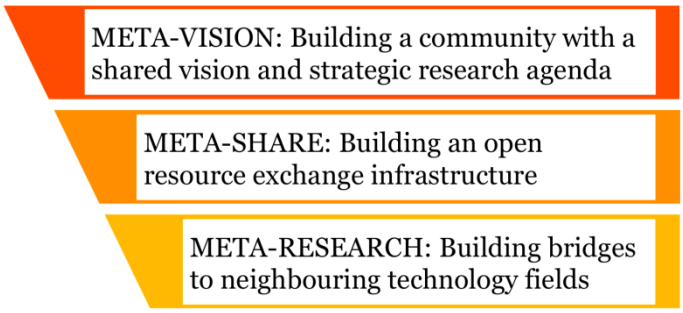

Figura 8: Cele trei linii de acţiune ale META-NET

META-VISION promovează o comunitate dinamică şi influenţială, unită în jurul unei viziuni comune şi a unei agende strategice comune de cercetare. Principalul scop al acestei activităţi este constituirea unei comunităţi de TL coerente şi coezive în Europa prin persoane cheie din diferite grupuri reprezentative. În primul an al META-NET, prezentările de la FLaReNet Forum (Spania), Language Technology Days (Luxemburg), JIAMCATT 2010 (Luxemburg), LREC 2010 (Malta), EAMT 2010 (Franţa) şi ICT 2010 (Belgia) s-au axat pe ridicarea numărului de adepţi din public. În conformitate cu estimările iniţiale, META-NET a contactat deja mai multe de 2.500 profesionişti din domeniul TL pentru a-şi dezvolta viziunea împreună cu ei. La META-FORUM 2010 în Bruxelles, META-NET a comunicat rezultatele iniţiale ale procesului de construire a viziunii unui public de peste 250 de participanţi. În cadrul unor serii de sesiuni interactive, participanţii au avut ocazia să discute viziunea prezentată de reţea.

META-SHARE creează o infrastructură publică distribuită pentru schimbul şi partajarea de resurse. Reţeaua de depozite va conţine date lingvistice, instrumente şi servicii web documentate cu metadate de nivel înalt, organizate în categorii standardizate. Resursele pot fi accesate şi permit căutări uniformizate. Resursele disponibile includ materiale gratuite, cu acces Open Source sau restricţionat, precum şi resurse disponibile contra unei sume. META-SHARE ţinteşte datele lingvistice, instrumentele şi sistemele existente, dar şi produsele noi necesare pentru construcţia sau evaluare de noi tehnologii sau servicii. Refolosirea, combinarea sau remodelarea datelor şi instrumentelor lingvistice joacă un rol crucial. META-SHARE va deveni o componentă critică a pieţei de tehnologii ale limbajului pentru dezvoltatori, experţi în localizare, cercetători, traducători şi profesionişti ai limbajului, de la firmele mici şi mijlocii până la cele mari. META-SHARE vizează întregul proces de dezvoltare a TL – de la cercetare la produse şi servicii inovatoare. Un aspect important al acestei activităţi este stabilirea META-SHARE ca componentă de valoare a unei infrastructuri globale europene pentru comunitate TL.

META-RESEARCH construieşte punţi de legătură între domenii tehnologice învecinate. Această activitate încearcă să aplice descoperirile recente din alte domenii pentru TL. În particular, această activitate este dedicate incorporării de mai multă semantică în Traducerea Automată (TA), optimizării diviziunii muncii în TA hibridă, exploatării contextului în vederea traducerii şi pregătii unei fundaţii empirice pentru TA. META-RESEARCH lucrează cu alte discipline, precum învăţarea automată şi comunitatea web-ului semantic. META-RESEARCH se concentrează pe colectarea date-

lor, pregătirea seturilor de date şi organizarea resurselor lingvistice pentru evaluare; crearea unor inventare de instrumente şi metode; şi organizarea de ateliere şi evenimente de formare pentru membrii comunităţii. Această activitate a identificat deja aspecte ale TA unde informaţia semantică poate avea un efect substanţial în sistemele existente. În plus, activitatea a dezvoltat recomandări privind integrarea semanticii în TA. META-RESEARCH finalizează o nouă resursă lingvistică pentru TA, corpusul hibrid adnotat de mostre de TA, care oferă date pentru perechile de limbi engleză-germană, engleză-spaniolă şi engleză-cehă. META-RESEARCH a dezvoltat de asemenea un instrument de colectare a corpusurilor multilingve ascunse pe web.

Organizații membre

Tabelul următor prezintă lista organizaţiilor şi a reprezentanţilor lor care participă în META-NET.

Țara	Organizaţia	Participant/Participanţi
Austria	Universitatea din Viena	Gerhard Budin
Belgia	Universitatea din Antwerp	Walter Daelemans
	Universitatea din Leuven	Dirk van Compernolle
Bulgaria	Academia de Ştiinţe din Bulgaria	Svetla Koeva
Croaţia	Universitatea din Zagreb	Marko Tadić
Cipru	Universitatea din Cyprus	Jack Burston
Republica Cehă	Universitatea Charles din Praga	Jan Hajic
Danemarca	Universitatea din Copenhaga	Bolette Sandford Pedersen and Bente Maegaard
Estonia	Universitatea din Tartu	Tiit Roosmaa
Finlanda	University Aalto	Timo Honkela
	Universitatea din Helsinki	Kimmo Koskenniemi and Krister Linden
Franţa	CNRS/LIMSI	Joseph Mariani
	Agenţia de Evaluare şi Distribuire a Resurselor Lingvistice (ELRA)	Khalid Choukri
Germania	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	University din Saarland	Manfred Pinkal
Grecia	Institutul pentru Prelucrarea Limbajului şi a Vorbirii, "Athena" R.C.	Stelios Piperidis
Ungaria	Academia de Ştiinţe din Ungaria	Tamás Váradi

Țara	Organizația	Participant/Participanți
	Budapest Universitatea de tehnologii și Științe din Budapesta	Géza Németh and Gábor Olaszy
Islanda	Universitatea din Islanda	Eirikur Rögnvaldsson
Irlanda	University din Dublin City	Josef van Genabith
Italia	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Letonia	Tilde	Andrejs Vasiljevs
	Institutul de Matematică și Informatică, Universitatea din Letonia	Inguna Skadina
Lituania	Institutul pentru Limba Lituaniană	Jolanta Zabarskaitė
Luxemborg	Arax Ltd.	Vartkes Goetcherian
Malta	Universitatea din Malta	Mike Rosner
Olanda	University din Utrecht	Jan Odijk
	Universitatea din Groningen	Gertjan van Noord
Norvegia	Universitatea din Bergen	Koenraad De Smedt
Polonia	Academia de Științe din Polonia	Adam Przepiórkowski and Maciej Ogrodniczuk
	Universitatea din Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pęzik
Portugalia	Universitatea din Lisabona	Antonio Branco
	Institutul pentru Ingineria Sistemelor și Calculatoare	Isabel Trancoso
România	Academia Română	Dan Tufis
	Universitatea Alexandru Ioan Cuza din Iași	Dan Cristea
Serbia	Universitatea din Belgrad	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institutul Mihailo Pupin	Sanja Vranes
Slovacia	Academia de Științe din Slovacia	Radovan Garabik
Slovenia	Institutul Jozef Stefan	Marko Grobelnik
Spania	Barcelona Media	Toni Badia
	Universitatea Tehnică Catalonia	Asunción Moreno
	Universitatea Pompeu Fabra	Núria Bel

Țara	Organizația	Participant/Participanți
Suedia	Universitatea din Gothenburg	Lars Borin
Regatul Unit	Universitatea din Manchester	Sophia Ananiadou
	Universitatea din Edinburgh	Steve Renals

References

[i] Diresctoratul general pentru Societatea Infromaţională şi Media, Comisia Europeană, *User language preferences online*, Flash Eurobarometer #313, 2011 http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[ii] Conisia Europeană, *Multilingualism: an asset for Europe and a shared commitment*, Bruxelles, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[iii] Director-General UNESCO, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[iv] Directoratul General pentru traduceri al Comisiei Europene, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[v] Anuar statistic 2009 al Institutului Naţional de Statistică http://www.insse.ro/cms/files/Anuar statistic/02/02 Populatie_en.pdf

[vi] Bază de date statistic a Biroului Naţional de Statistici al Republicii Moldova http://statbank.statistica.md

[vii] http://en.wikipedia.org/wiki/Romanian_diaspora

[viii] Marius Sala (ed.), *Encyclopaedia of the Romanian Language* (in Romanian), 2nd Edition, Bucharest, Univers Enciclopedic Publishing House, 2006.

[ix] http://www.efnil.org/documents/language-legislation-version-2007/ romania

[x] http://www.ilr.ro/plr.php?lmb=1

[xi] http://www.internetworldstats.com/eu/ro.htm

[xii] http://www.internetworldstats.com/stats9.htm

[xiii] Tufiş Dan şi Ceauşu Alexandru (2008). *DIAC+: A Professional Diacritics Recovering System*, In Proceedings of Language Resources and Evaluation Conference, LREC 2008, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0

[xiv] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[xv] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_ semantic_search_capabilities.html

[xvi] Tufiş, D., Radu I., Bozianu, L., Ceauşu, A., Ştefănescu, D. (2008). *Romanian Wordnet: Current State, New Applications and Prospects*. In Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen (eds.), Proceedings of 4th Global WordNet Conference, GWC-2008, pp. 441-452, ISBN 978-963-482-854-9.

[xvii] www.racai.ro/WebServices.

[xviii] Tufiş, D., Radu I., Ceauşu, A., Ştefănescu, D. (2008). *RACAI's Linguistic Web Services*, In Proceedings of Language Resources and Evaluation Conference, LREC 2008, Marakkech, Morocco. ELRA - European Language Resources Association. ISBN: 2-9517408-4-0

[xix] Trandabăţ D. (2011) *Towards automatic cross-lingual transfer of semantic annotation*, in 6e Rencontres Jeunes Chercheurs en Recherche d'Information RJCRI-CORIA 2011, 16-18 March, Avignon, France.

[xx] Iftene, A., Vamanu, L., Croitoru, C. (2010). *UAIC at ImageCLEF 2009 Photo Annotation Task*. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6242, Part II (Multilingual Information Access Evaluation Vol. II Multimedia Experiments). Pp. 283-286. ISBN: 978-3-642-15750-9. Springer, Heidelberg.

[xxi] Cu cât scorul este mai mare, cu atât traducerea este mai bunăş un traucător uman ar obţine în jur de 80. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA.

[xxii] Munteanu D.S. şi D. Marcu. (2005). *Statistical Machine Translation: An English-Romanian Experiment*. Tutorial invitat la EUROLAN 2005 Summer School on NLP & HLT: The Multilingual Web: Resources, Technologies, and Prospects, Cluj Napoca, Romania.

[xxiii] Tufiş, Dan, Ceauşu Alexandru (2009). *Factored Phrase-Based Statistical Machine Translation*, In Corneliu Burileanu, Horia Nicolai Teodorescu (eds.) Proceedings of the 5th Conference "Speech Technology and Human-Computer Dialogue" SpeD 2009, IEEE Catalogue number:CFP095H-CDR

[xxiv] Irimia. Elena (2009). *EBMT experiments for the English-Romanian Language Pair*. International Joint Conference Intelligent Information Systems (IIS 2009). Kraków, Poland, June 15-18, 2009.

[xxv] O analiză detaliată a coerenţei diferitelor texte este prezentată în Cristea, D., Iftene, A. (2011) *If you want your talk be fluent, think lazy! Grounding coherence properties of discourse*. Lucrare invitată la the University of Sussex, March.

[xxvi] Cristea, D., Postolache, O., Pistol, I. (2005): *Summarisation through Discourse Structure*. In Alexander Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, Proceedings of CICLing 2005, LNSC, vol. 3406, ISBN 3-540-24523-5, pp. 632-644.

[xxvii] Vezi de exemplu sistemul prezentat în Iftene, A., Trandabăţ, D., Moruz, A., Pistol, I., Husarciuc, M., Cristea, D. (2010). *Question Answering on English and Romanian Languages*. In C. Peters et al. (Eds.): CLEF 2009, LNCS 6241, Part I. Pp. 229-236. ISBN 978-3-642-15753-0. Springer, Heidelberg.

[xxviii] Vezi mai multe detalii despre informatizarea Dicţionarului tezaur al Limbii Române în Cristea, D. (2009): *Steps towards an electronic version of the Tesaurus Dictionary of the Romanian language* (in Romanian), ASTRA 2009, Iaşi.

[xxix] http://wordnet.princeton.edu/

[xxx] Conţinutul acestei resurse lexical poate fi găsit la http://www.racai.ro/wnbrowser/

[xxxi] Baccianella Stefano, Andrea Esuli, Fabrizio Sebastian (2008). *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, In Proceedings of LREC 2008, pp. 2200-2204.

[xxxii] Gînscă, A. L., Boroş, E., Iftene, A., Trandabăţ, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D. (2011). *Sentimatrix - Multilingual Sentiment Analysis Service*. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-WASSA2011) Portland, Oregon, USA, June 19-24, 2011.

[xxxiii] Vezi colecţia de volume cu lucrările publicate la Atelierul de lucru Instrumente şi resurse lingvistice pentru limba română din 2006 până în 2010, Editura Universităţii "A.I. Cuza" Iasi, ISBN 978-973-703-208-9.

[xxxiv] Vezi detalii ale soluţiei propuse în Cristea, Dan (2010). *Resurse lingvistice în flux continuu*. In Iftene, A. et al. (ed). Linguistic resources and instruments for Romanian language processing, Bucharest, University "Al. I. Cuza" Iaşi Ed. ISSN 1843-911X.