

Promovarea limbii române în SI-SC

Dan Tufis

Motto: În era electronică, *este esențial pentru supraviețuirea unei limbi ca ea să fie folosită în sistemele de informare electronică.* A.Danzin:
Towards a European Language Infrastructure, March 1992, Raport Special
al Comisiei Europene

1. Introducere

Viteza cu care societatea informațională evoluează a creat o creștere fără precedent a gamei și numărului de servicii electronice și de resurse informaționale sub formă textuală, audio, grafică și/sau video. Largul acces la astfel de servicii și resurse a născut speranța că societatea informațională va rezolva mai ușor problemele cu care se confruntă actualmente societatea pre-informațională și că soluțiile își vor găsi rezolvări creative, inovative [1].

Limbajul este o premisă a dezvoltării comunicării, educației și abilităților individuale de toate felurile (mai ales tehnologice) adică exact acele obiective considerate a fi factorii vitali ai viitoarei competitivității a Europei, zonă geo-politică, economică și socială ce este și trebuie să rămână multilingvă. Realizarea unei piețe europene unice va crea presiuni pentru îmbunătățirea comunicației între statele membre ale acestui spațiu. Libera circulație a persoanelor, a bunurilor, a serviciilor și capitalului, precum și dorința de creștere a coeziunii sociale în cadrul Comunității largite implică necesitatea ca oamenii săi să se înțeleagă la toate nivelurile, să schimbe informația scrisă sau orală cu un minim de bariere lingvistice în comunicare.

Limbajul constituie fundamentul comunicării între oameni și pentru foarte mulți dintre ei, acesta poartă conotații emoționale și culturale profunde, valori conținute într-o vastă moștenire literară, istorică, filozofică și educațională. Tocmai din acest motiv, limba maternă nu trebuie să constituie un obstacol în calea accesului la cunoașterea multiculturală umană disponibilă în cyberspațiu [7, 24]. Dezvoltarea armonioasă a societății informaționale bazată pe cunoștințe este deci posibilă doar prin promovarea informației și accesului cu caracter multilingv și multicultural [20,21,22,24].

În contextul societății informaționale, al comunicării mediate de tehnologia informației și de telecomunicații, limba devine obiect al investigației tehnice. Tehnologia limbajului impune metodologii specifice de cercetare/dezvoltare, dezvoltarea sau adaptarea resurselor lingvistice [7, 8] fundamentale cum ar fi dicționarele, tezaurele, corpusurile și gramaticile computerizate, în conformitate cu standardele sau recomandările existente. În funcție de resursele lingvistice disponibile, de volumul și calitatea lor, de compatibilitatea

codificării lor în raport cu recomandările și standardele internaționale etc., se poate vorbi de *nivelul de tehnologizare* al unei limbi naturale.

Prin prisma nivelului de tehnologizare există decalaje foarte mari între limbile vorbite actualmente în Europa sau în alte părți ale lumii. În conformitate cu un raport al Directoratului General XIII al Comisiei Comunităților Europene [20], în afara limbii engleze și într-o oarecare măsură și a celei franceze și germane, nivelul de tehnologizare al celorlalte limbi europene era la începutul anilor '90 foarte slab sau practic nul. Lucrurile au evoluat în cei aproape 10 ani care au trecut de la acel raport, dar evoluția a fost semnificativă doar în privința unui număr restrâns de limbi europene: franceza, germana, italiana, spaniola, și într-o bună măsură ceha și poloneza.

Promovarea limbii române în SI-SC presupune informatizarea limbii române ca factor infrastructural fundamental și precum și stimularea utilizării curente a limbii române în utilizarea tehnologiilor și a serviciilor informatice.

În cele ce urmează ne propunem să prezentăm o serie de noțiuni terminologice legate de informatizarea limbii și să schițăm câteva dintre măsurile ce se impun pentru accelerarea procesului de informatizare a limbii române.

2. Informatizarea limbii

Prin procesul de informatizare a unei limbi naturale se înțelege ansamblul programelor de cercetare specifice și a măsurilor tehnice, organizatorice și legislative privitoare la dezvoltarea și utilizarea de programe software pentru prelucrarea automată a limbii respective. Procesul de informatizare a unei limbi naturale nu înseamnă nicidecum stâlcirea limbii, sau așezarea ei într-un pat al lui Procust de tip orwellian, ci potențarea și diseminarea ei prin mijloacele tehnologice ale societății informaționale. Sigur, limba este un fenomen extraordinar de complex, iar comunicarea om-calculator prin limbaj natural complet nerestricționat este o utopie (cel puțin la nivelul cunoașterii științifice actuale). Dar pentru anumite registre lingvistice, și universuri de discurs precizate, prelucrarea automată a limbajului natural este o realitate, o necesitate în afara oricărei discuții.

Informatizarea limbii, ce include în sfera sa de interes atât limbajul scris cât și cel vorbit, este adeseori referită sub numele de inginerie a limbajului. “De ce inginerie?” se pot întreba pe bună dreptate unii oameni ai literelor, așa cum cu ceva vreme în urmă reprezentanți ai științelor mentalului sau ai științelor viului se întrebau “de ce inginerie a cunoștințelor”, sau de ce “inginerie genetică”. Răspunsul trebuie căutat în însăși dezvoltarea științei în general, și a științelor aplicate în particular. Ingineria, în contextul unor astfel de modificări terminologice, vine să sublinieze aspectele legate de validarea experimentală a ipotezelor științifice, de necesitatea ca stările de lucruri anticipate de teorie să poată fi realizate și reproduse experimental ori de câte ori este nevoie. Calculatorul electronic este fără îndoială un instrument aproape perfect al științei aplicate (și nu numai). Pentru a simula un proces fizic sau mental, modelul acestuia trebuie să fie riguros specificat și corect transpus în reprezentarea internă a calculatorului. Teorii sau modele ale căror transpuneri pe calculator sunt imposibil de realizat (neexistând o

descriere algoritmică) sau de experimentat (necesitând resurse de calcul imposibil de asigurat sau inacceptabil de mari) se plasează de regulă în afara disciplinelor științifice căror nume pot coloca cu termenul “inginerie”. Pe de altă parte, chiar dacă modele formale cu proprietăți computaționale adecvate au fost definite și implementarea lor efectuată, până la realizarea unui sistem care să realizeze automat o prelucrare lingvistică semnificativă rămâne o distanță uriașă, de multe ori ignorată, reprezentată de *instanțierea* modelului sau a teoriei respective în raport cu o limbă anume.

Din acest punct de vedere este semnificativ a arăta că însuși numele domeniului de cercetare a prelucrării automate a limbajului natural a suferit modificări reflectând progresele științifice și tehnologice: inițial, desprinzându-se din lingvistica formală, *lingvistica matematică* a încercat dezvoltarea unor modele matematice de reprezentare a limbajelor naturale sau formale (în general al aspectului lor sintactic, gramatical), căutând soluții abstracte de modelare generativă de tip universal a ceea ce se presupunea (la nivelul cunoașterii științifice a anilor '50 - '60) a fi facultatea limbajului. Curând metodele lingvisticii matematice și-au atins limitele drept care în anul 1966 la propunerea lui David Hays, domeniul de cercetare al limbajelor naturale, din perspectiva utilizării acestora în interacțiunea cu calculatoarele electronice, este individualizat sub numele de *lingvistică computațională*. Scopul declarat al noii discipline era cel al modelării și construirii efective a bazelor de cunoștințe lingvistice și extralingvistice necesare prelucrării mecanice a limbilor naturale (cu alte cuvinte al instanțierii modelelor lingvistice). Pe lângă analiza complexității algoritmilor de analiză și generare, devenite în lingvistica computațională filtrul oricărui nou model sau al oricărei teorii lingvistice, dimensiunea fundamentală a lingvisticii computaționale devine fezabilitatea instanțierii unei descrieri lingvistice cât mai complete, mentenabilitatea acestei instanțieri și desigur conformanța cu realitatea uzului limbii. Teorii sau formalisme lingvistice în vogă în anii '70 (diverse variante ale teoriei transformaționale, rețele de tranziție extinse) au sucombat din cauza puterii lor generative excesive (echivalente celei ale unei mașini Turing), al dificultăților de mentenanță a instanțierilor lingvistice sau a altor deficiențe computaționale (de pildă, dependența modelului de direcția (analiză sau generare) de prelucrare).

În paralel cu termenul *lingvistică computațională* se impune sintagma *prelucrarea limbajului natural* ca un sumum al teoriilor și modelelor de *inteligență artificială* ce abordează problematica comunicării între om și calculator. Din perspectiva *inteligenței artificiale* prelucrarea limbajului natural presupune îmbinarea cunoștințelor lingvistice (morfologie, lexic, sintaxă și discurs) cu cele extra-lingvistice (cunoștințe despre domeniul discursului, cunoștințe generale despre lume, etc.) înțelegerea și producerea limbajului natural fiind considerate ca manifestări fundamentale ale inteligenței. Semantica și pragmatica computațională precum și reprezentarea cunoștințelor sunt domeniile cele mai fecunde ale inteligenței artificiale în domeniul prelucrării și modelării limbajelor naturale.

În România, cercetările în domeniul lingvisticii computaționale și al prelucrării limbajului natural, precum și primele rezultate practice au apărut la începutul anilor '80 [3, 4, 5, 6].

Tendința majoră a ultimului deceniu în informatizarea limbilor poate fi considerată reorientarea cercetărilor de la abordarea de tip introspectiv spre cea bazată pe evidența datelor lingvistice furnizate de volume mari de texte sau înregistrări vocale organizate sub forma corpusurilor lingvistice. Disponibilitatea în domeniul public al Internet-ului a unor volume mari de date adnotate în conformitate cu o sintaxă și o semantică ce converg rapid spre standardizare, ca și dezvoltarea tehnologiilor WEB asociate, au creat premisele apariției unor noi paradigme de investigație științifică și tehnologică cum ar fi lingvistica e-corpusului și respectiv lingvistica WEB-ului. Bazate în principal pe metode statistice și inductive, noile tehnologii ale limbajului vin în sprijinul eliminării (sau cel puțin al diminuării) fenomenului cunoscut sub numele “ștrangularea procesului de achiziție a cunoștințelor” (knowledge acquisition bottleneck). Ipoteza fundamentală ce instrumentează noile orientări în cercetarea legată de prelucrarea automată a limbajului natural este că limbajul scris sau vorbit, transpus într-o reprezentare electronică, conține suficientă informație implicită sau explicită (atunci când reprezentarea textuală conține adnotări) pentru a permite extragerea automată a cunoștințelor lingvistice necesare prelucrării limbajului natural. Acest proces de extragere a cunoștințelor lingvistice din corpus/corpusuri [10, 12, 13, 16, 17, 18] este în esență un proces de învățare automată al cărui rezultat depinde, ca în orice proces de învățare, de calitatea și cantitatea surselor de informație, pe de o parte, și de inteligența modelului de învățare pe de altă parte. Informatizarea unei anumite limbi naturale se referă într-o măsură covârșitoare la crearea surselor (sau mai precis, după cum se va vedea în continuare, a resurselor) lingvistice, pentru limba în cauză, adecvate procesului de învățare automată.

2.1. Programe de cercetare și măsuri tehnice

Programele de cercetare au ca obiect modelarea computațională și construirea de baze de cunoștințe lingvistice. Aceste cunoștințe lingvistice, numite generic resurse lingvistice, trebuie descrise într-un format exact, prelucrabil mecanic. Mai mult, într-un context multilingv, această descriere trebuie să fie compatibilă din punctul de vedere al formalizării cu descrierile altor limbi prezente în contextul comunicațional respectiv. Resursele lingvistice specifice fiecărei limbi naturale trebuie dezvoltate în conformitate cu standardele, recomandările și practicile internaționale. Alinierea la aceste standarde, recomandări și practici internaționale este esențială pentru motive legate de independența față de diferitele platforme hardware și software pe care vor fi utilizate programele de prelucrare automată a limbii precum și pentru asigurarea inter-operabilității unor sub-sisteme de prelucrare a unor limbi diferite (de exemplu pentru traducere automată).

Printre resursele lingvistice fundamentale pentru informatizarea unei limbi pot fi menționate:

- baze de date fonetice și fonologice;
- baze de date lexicale: dicționare electronice, tezaure terminologice, dicționare bi-și multilingve, ontologii;
- corpusuri electronice mono- și multilingve;
- modele formale ale limbii la toate nivelurile ei (fonetică, morfologie, sintaxă, discurs): gramatici formale, modele probabiliste, modele euristice.

Tot în categoria programelor de cercetare specifice și a măsurilor tehnice sunt cuprinse dezvoltarea de programe software generice de prelucrare a unei limbi sau mai multor limbi naturale. Distincția netă care se face între resursele lingvistice (specifice unui anumite limbi) și programele generice de prelucrare ale acestor resurse subliniază încă o dată necesitatea standardizării în dezvoltarea resurselor lingvistice.

Pentru a distinge dar și pentru a sublinia interdependența noțională, vom folosi în continuare termenul de **sursă lingvistică** pentru orice izvor informațional produs în lingvistica tradițională sau discipline conexe (cărți, dicționare, atlase, studii, indecși, glosare de termeni, etc.) și cea de **resursă lingvistică** pentru orice reprezentare electronică a conținutului (*exact/modificat, total/parțial*) unei surse lingvistice, reprezentare care să permită utilizarea algoritmică, neambiguă și deterministă a informației lingvistice din sursă. De pildă, un dicționar explicativ cum este DEX-ul, reprezintă o sursă extrem de valoroasă de cunoștințe lexicale asupra limbii române. Dar DEX-ul în forma sa tipărită pe hârtie nu reprezintă încă o resursă lingvistică (în sensul prezentat aici) pentru limba română. Nici măcar rezultatul scanării, urmată de o prelucrare de tip OCR, sau al dactilografierii textului conținut în volumul tipărit nu reprezintă o resursă lingvistică (acest format se numește **sursă în format electronic**) Transformarea volumului publicat într-o resursă lingvistică este un proces extrem de laborios, care necesită printre altele explicitarea tuturor informațiilor implicite sublimite în convențiile tipografice și eliminarea tuturor inconsecvențelor inerente muncii manuale. Pentru a evidenția și mai puternic distincția dintre **sursa DEX** (destinată uzului uman) și **resursa DEX** (destinată prelucrării automate) este suficient a spune că dacă s-ar tipări textul resursei lingvistice, ar rezulta circa 8-10 volume de dimensiunea DEX-ului [11, 14, 15].

Dacă programele de prelucrare, datorită standardizării, pot fi în mare măsură preluate și adaptate de la o limbă la alta, sau altfel spus pot fi făcute de americani, francezi sau nemți și utilizate de unguri, cehi sau români, în schimb sursele sau resursele lingvistice nu pot fi dezvoltate decât de vorbitori nativi ai limbii respective. Și de regulă, realizarea acestora intră sub incidența și responsabilitatea autorităților naționale.

Resursele lingvistice trebuie să fie dezvoltate pentru toate limbile societății globale. Ele sunt indispensabile în funcționalitatea societății informaționale, începând cu editarea și prelucrarea documentelor, traducerea acestora și sfârșind cu publicarea și distribuția lor. Sistemele educaționale, inclusiv cele destinate persoanelor cu handicapuri psiho-motorii, nu pot funcționa în absența unor resurse lingvistice dezvoltate adecvat. În plus, resursele lingvistice ar trebui să reprezinte referința fundamentală pentru autoritățile naționale responsabile de urmărirea evoluției limbii și totodată principala sursă de material lingvistic pentru toate ramurile lingvisticii.

Este deja cunoscut faptul că activitatea de creare și întreținere a resurselor lingvistice implică costuri ridicate. Pe măsură ce presiunea exercitată de implementarea conceptelor societății informaționale globale va crește, costurile lansării programelor de creare a resurselor lingvistice naționale (acolo unde astfel de proiecte nu au fost lansate deja) vor fi din ce în ce mai mari.

Având în vedere rezultatele obținute de cercetarea românească în domeniul informatizării limbii române, se impune corelarea acestora și agrearea lor într-un program național având ca obiective pe termen scurt (2004-2005):

- realizarea unor e-corpusuri de referință ale limbii române scrise și vorbite, adnotate standardizat la nivel morfologic, lexical, sintactic și discursiv;
- realizarea dicționarilor de referință ale limbii române în formă electronică standardizată (noul DEX, Dicționarul Tezaur al Limbii Române, dicționarul de sinonime, etc);
- realizarea de dicționare bi- și multilingve (româna una dintre limbi) în formate compatibile, respectând standardele și recomandările internaționale în domeniul lexicografiei computaționale;
- realizarea de dicționare terminologice mono- și multilingve în cât mai multe domenii, folosind standarde și tehnologii comune;
- realizarea unei ontologii lexicale pentru limba română integrabile în EURO-WordNet și Global-WordNet (cele mai mari proiecte multilinguale în domeniul ontologiilor lexicale);
- realizarea de gramatici (incrementale) ale limbii române.

În spiritul integrator al societății globale resursele lingvistice specifice diferitelor limbi vor fi disponibilizate pentru uzul general [1], desigur cu respectarea drepturilor de proprietate intelectuală. Utilizatorii potențiali ai tehnologiei limbajului trebuie conștientizați de beneficiile tehnice și economice ale utilizării în comun a resurselor.

Cooperarea intra- și interdisciplinară în domeniul realizării resurselor lingvistice ar trebui să se manifeste cu atât mai mult cu cât resursele lingvistice necesare cercetării fundamentale sunt la fel de necesare și în procesul de realizare a programelor comerciale de către firme specializate.

În paralel cu disponibilizarea resurselor fundamentale ale limbii române va putea începe procesul de dezvoltare a aplicațiilor cu utilizarea limbii române ca limbă de interacțiune cu calculatorul. În perioada imediată (2001-2003) vor putea apare sisteme autoriale inteligente care să asiste utilizatorul în redactarea documentelor scrise (îmbunătățirea verificatoarelor ortografice, dezvoltarea de corectoare sintactice și stilistice, verificatoare de consistență terminologică etc). În perspectiva anilor 2005-2010 se preconizează apariția primelor sisteme comerciale de clasificare automată a documentelor electronice în limba română, sisteme de interogare în limba română a informației de pe WEB, sisteme de rezumare automată a documentelor, traducere din și în limba română a limbajului scris sau vorbit, servicii publice de tip chioșc-electronic cu interacțiune în limba română (scris sau vorbit).

2.2. Măsuri organizatorice

Cele mai noi abordări în domeniul informatizării limbilor naturale sunt cele lexicalizate, adică cele ce pun în centrul modelelor lingvistice bazele de cunoștințe lexicale. Unul din motivele preponderenței acestor modele lingvistice este faptul că din punct de vedere conceptual, este mai ușor a se controla complexitatea fenomenelor lingvistice structurând

cunoștințele asupra limbii în jurul elementelor lexicale. În schimb dezvoltarea unor resurse lingvistice lexicalizate, semnificative din punctul de vedere al acoperirii lingvistice, devine o muncă herculeană.

Din acest motiv, procesul informatizării unei limbi naturale, în speță al limbii române, trebuie realizat ca un proiect complex, organizat pe diferite niveluri astfel încât pe de o parte componente ale proiectului să se poată desfășura separat și în paralel, iar pe de altă parte interacțiunea/integrarea acestor sub-proiecte să se realizeze simplu și ușor de controlat/validat. Un astfel de demers nu se poate realiza spontan, ci este nevoie de un cadru organizatoric care să ofere elemente de atracție dar și de control. Un astfel de cadru organizatoric îl reprezintă, la nivel european, Comisia Europeană, care prin programe cadru și prin mijloace financiare adecvate, a reușit și continuă să controleze direcțiile de cercetare/dezvoltare considerate prioritare.

Impunând ideea consorțiilor multi-naționale pentru proiectele din domeniul informatizării limbilor europene, s-au creat acele structuri de cercetare/dezvoltare care au permis în ultimii 10 ani progrese mai mari decât în toată istoria de circa 50 de ani a domeniului prelucrării automate a limbajului natural. Dacă perioada anilor '60-'80 a acestui domeniu poate fi considerată ca perioada "one-man-show", ultimii 10 ani au demonstrat că fără colaborarea unor grupuri mari de specialiști, trecerea de la stadiul de prototip la cel de produs industrial este imposibilă. Informatizarea unei limbi revine în ultimă instanță la crearea unor produse industriale, continuu perfectibile este adevărat, dar utilizabile la orice moment în aplicații reale.

Cercetarea în domeniul limbii române a fost și este una din preocupările fundamentale ale Academiei Române. Informatizarea limbii române a devenit de curând o direcție de interes în Academia Română în cadrul căreia a fost înființată "Comisia pentru informatizarea limbii române". Acesta este un îmbucurător pas înainte. Formarea de noi specialiști, specializarea sau respecializarea, sunt condiții esențiale ale informatizării rapide a limbii române. Începând din anul 2000 a fost lansat la Universitatea București, Catedra de Limba Română, primul program de Masterat în lingvistică computațională. Din toamna acestui an va fi creat și la Facultatea de Informatică a Universității A.I. Cuza din Iași un program de Masterat în lingvistica computațională. Aceste două inițiative ale unor specialiști ai Academiei Române, sprijinite de Ministerul Educației Naționale constituie un alt element important al procesului informatizării limbii române.

Colaborarea interdisciplinară deschisă a tuturor specialiștilor, accesul neîngrădit la surse și resurse lingvistice, utilizarea tehnologiilor lingvistice moderne, lansarea de proiecte prioritare sunt alți câțiva vectori esențiali ai procesului pe care Academia Română poate, trebuie și este cea mai în măsură să-l organizeze.

2.3. Măsuri legislative

Statele și organizațiile internaționale interguvernamentale trebuie să reafirme și să promoveze respectul pentru folosirea tuturor limbilor în cyperspațiu, să contribuie la păstrarea bogăției și diversității moștenirii umane universale și la coexistența pașnică,

obiective care sunt stipulate în multe declarații și convenții internaționale și în multe constituții naționale.

Este datoria fiecărui stat să formuleze politici naționale în legătură cu problema crucială a supraviețuirii limbii în cyberspațiu. Asistența internațională în formularea și implementarea politicilor lingvistice pe rețelele de informații globale menite să promoveze limbile native și învățarea acestora, trebuie pusă în aplicare respectând diversitatea culturală și întărirea solidarității naționale și internaționale.

Aspectele legislative ale informatizării limbii române, sunt extrem de importante, ele putând impulsiona dezvoltarea procesului sau dimpotrivă contribuind substanțial la încetinirea și rămânerea lui în urma proceselor similare din alte țări.

Un exemplu aparent minor, dar în realitate semnificativ, este faptul că tastaturile cu claviatură românească sunt rarități în spațiul comercial al României. Cea mai mare parte a sistemelor de operare în uz nu sunt localizate. Acest lucru se datorează în primul rând unei carențe legislative.

Conținutul Internet, cu precădere la nivelul portalurilor naționale și al siturilor autorităților publice trebuie realizat în primul rând în limba română și apoi și în alte limbi de circulație (și nu invers). Unul din criteriile evaluării nivelului de informatizare al unei limbi naturale este printre multe altele și volumul informației disponibile pe Internet în limba respectivă. O serie de estimări, cu rezultate congruente [1, 2, 19], au putut permite o ierarhizare a limbilor “vizibile” pe Internet. De pildă, în studiul realizat de XEROX Europe-Research Center [2], limba română ocupă un modest loc 21 din 32 de limbi cu grafie latină (devansând doar islandeza, irlandeza, estoniana, latina, basca, esperanto, letoniana, lituaniana, bretona, albaneza și galeza!).

Sunt necesare măsuri legislative care să reglementeze, să asigure și să încurajeze finanțarea, atât prin bugetul de stat cât și prin contribuția sectorului privat, pentru crearea, dezvoltarea, conservarea și menținerea website-urilor în limba română și alte limbi de circulație internațională.

Statul român și organizațiile internaționale trebuie să sprijine finanțarea instituțiilor publice pentru a asigura conservarea și digitizarea informațiilor din domeniul public în conformitate cu standardele și sistemele adecvate pentru schimbul de informații, portabilitate, operabilitate și acces on-line. Aceste instituții trebuie încurajate să pună la dispoziția tuturor celor interesați, prin rețelele globale de informații, rezultatele obținute.

Statul român și organizațiile internaționale non-guvernamentale și interguvernamentale trebuie să adopte strategii pentru dezvoltarea și pentru distribuirea on-line a materialelor liber accesibile de educație lingvistică.

Așa cum arătam mai devreme, resursele lingvistice au la bază cercetări și rezultate ale lingvisticii teoretice, dicționare realizate în zeci de ani de muncă de un mare număr de lexicologi sau lexicografi. Editurile care au tipărit aceste surse esențiale de cunoștințe asupra limbii, au făcut investiții importante. Editurile moderne din străinătate apelează tot

mai des la tehnologii ale limbajului pentru a-și transforma sursele textuale de informație lingvistică în resurse lingvistice.

Pe de altă parte limitarea accesului pentru grupurile de cercetare la sursele și resursele lingvistice este contra-productivă atât pentru procesul informatizării limbii în general, pentru progresele în cercetarea lingvistică cu mijloace computerizate cât și pentru dezvoltarea industriei de software lingvistic. Este deci necesară o serie de reglementări în acest sens, care să specifice în ce condiții pot fi folosite legal sursele și resursele lingvistice și să asigure protecția intelectuală.

Practica arată că puține sisteme de prelucrare a limbajului natural au avut vreo șansă comercială atunci când nu s-au bazat pe surse și resurse lingvistice de bună calitate. În aceste condiții nu este de mirare (în alte părți) de ce din beneficiile financiare ale unei firme de software lingvistic, o parte semnificativă se cheltuiește pentru drepturile de copyright ale lexicografilor, ale editurilor, ale specialiștilor în lingvistică computațională.

Academia Română, ca principala realizatoare a surselor și resurselor lingvistice ale limbii române, ca instituție responsabilă de păstrarea, îngrijirea și dezvoltarea limbii române este poate cea mai îndreptățită instituție a țării care să inițieze ansamblul de măsuri legislative necesare derulării normale a procesului de informatizare a limbii.

În sensul considerațiilor de mai sus trebuie luate urgent măsuri atât la nivel național în concordanță cu politicile internaționale pentru incurajarea punerii în practică a principiilor fundamentale care vor determina în viitor societatea informatică globală.

Avansul rapid al inovațiilor în tehnologiile informației și comunicării a dat naștere unei competiții pentru revendicarea cunoștințelor, ceea ce conduce la riscuri dacă se încearcă anexarea și privatizarea informației în domeniul public. În principal este responsabilitatea instituțiilor publice (biblioteci, arhive, agenții guvernamentale etc.) să faciliteze accesul la acest tip de informație, prin implicarea contribuabililor din sectorul privat și chiar prin participarea cetățenilor.

Statul român în colaborare cu organizațiile internaționale, interguvernamentale și non-guvernamentale trebuie să reafirme și să sublinieze principiile care promovează excepțiile de la protejarea proprietății intelectuale, în particular cele care au ca scop educația și cercetarea științifică. Menținerea echilibrului dintre protecția copyrightului și accesul la informație este o mare provocare pentru societatea informațională. Aceasta implică atât reglementări naționale cât și internaționale. Anumite principii ale copyrightului (de exemplu limitarea duratei și scopului protecției) reprezintă căutarea aceluși echilibru. Multe tratate internaționale confirmă extinderea recentă a prerogativelor autorilor și a deținătorilor drepturilor lor dar au impus însă și limitări sau dispense (pentru domeniile menționate mai sus) ca parte a procesului de îmbunătățire a drepturilor.

3. Concluzii

Societatea informațională globală constituie un concept generos, al egalității drepturilor fundamentale ale oamenilor viitorului.

Accesul neîngrădit de bariere tehnologice, lingvistice sau culturale la informația publică, la educație, asistență medicală, servicii sociale sau economice este o realitate aflată deocamdată pe platformele tehnologice de cercetare/dezvoltare. Experimentele realizate în unele dintre cele mai dezvoltate țări ale lumii au demonstrat deja cu prisosință realismul tehnologic al societății informaționale globale. Este indiscutabil că informatizarea generalizată va avea un impact extraordinar asupra umanității. Dacă acest impact va fi benefic sau dimpotrivă va accentua decalajele actuale (“digital divide”), depinde de conștientizarea factorilor de decizie superioară asupra demersurilor ce trebuie întreprinse, de evaluarea corectă a priorităților de cercetare/dezvoltare. Între acestea, tehnologiei limbajului îi revine statutul de premiză a societății informaționale globale.

Referințe bibliografice

- [1] E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli (eds) “Multilingual Information Management: Current Levels and Future Abilities”, NSF Report, 1999.
- [2] G. Grefenstette, J. Nioche “Estimation of English and non-English Language Use on the WWW”, Proceedings of RIAO2000, June 2000.
- [3] D.Tușiș, “Câteva aspecte ale interacțiunii om-calculator prin intermediul limbajului natural”, în Buletinul Român de Informatică, 1980.
- [4] D.Tușiș, “Demonstrarea automată, un mod de abordare a sistemelor de întrebare/răspuns” în volumul AI III-lea Simpozion Național de Informatică INFO' IAȘI, Iași, 1981.
- [5] D.Tușiș, “SDLR: A Dialogue System For Romanian Language”, in J.Miklosko (ed.) Computers and Artificial Intelligence, VEDA Publishing House, Bratislava, 1983.
- [6] D.Tușiș, D.Cristea, “IURES: A Human Engineering Approach to Natural Language Question Answering”, in W. Bibel, B.Petkoff (eds), Artificial Intelligence: Systems, Applications, Methodology, North Holland, 1985.
- [7] D. Tușiș, P. Andersen, (eds.) “Recent Advances in Romanian Language Technology”, Editura Academiei Române, București 1997, ISBN 973-27-0626-0.
- [8] D. Tușiș, “A Generic Platform for Developing Language Resources and Applications”, in W. Teubert, R. Markincevicene (eds), Proceedings of the 3rd European TELRI Conference in Language Resources, Kaunas, 1997.
- [9] D. Tușiș, “Yet Another Head Driven Generator of Natural Language Generator”, in Journal on Information and Control, vol.3, 1999.
- [10] D. Tușiș, A. Chițu, “Automatic Diacritics Insertion in Romanian Texts”, In F. Kiefer, G. Kiss, J. Pajzs (eds) Papers in Computational Lexicography COMPLEX'99, Hungarian Academy Publishing House, 1999.
- [11] D. Tușiș, G. Rotariu, A. M. Barbu, “TEI-Encoding of a Core Explanatory Dictionary of Romanian” In F. Kiefer, G. Kiss, J. Pajzs (eds) Papers in Computational Lexicography COMPLEX'99, Hungarian Academy Publishing House, 1999.
- [12] D. Tușiș, P. Dienes, C. Oravecz, T. Varadi “Principled Tagset Design for Tiered Tagging of Hungarian” in Proceedings of the LREC'2000 International Conference, Athens, June 2000.

- [13] D. Tufiş, “Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging”, in Proceedings of the LREC’2000 International Conference, Athens, June 2000.
- [14] D. Tufiş, C. Ştefan, “DIC: Gramatica pentru Dicţionarul Explicativ al Limbii Române”, Raport de cercetare RACAI, iunie 2000.
- [15] D. Tufiş, C. Ştefan, “DIC: Compiler pentru Dicţionarul Explicativ al Limbii Române”, Raport de cercetare RACAI, 2 volume, iunie 2000.
- [16] D. Tufiş, C. Popescu, R. Roşu, “Automatic Classification of Documents by Random Sampling”, in Proceedings of the Romanian Academy Series, A, Vol. 1 no. 2, 2000.
- [17] D. Tufiş, A.M. Barbu, “Extracting multilingual lexicons from parallel corpora”, in Proceedings of the ACH-ALLC International Conference, New York, June 2001.
- [18] D. Tufiş, A.M. Barbu, “Computational bilingual lexicography: automatic extraction of translation dictionaries”, in Romanian Journal on Information Science and Technology, vol. 4 no. 3, 2001.
- [19] J.L. Xu “Multilingual Search on the World Wide Web. In Proceedings of the Hawaii International Conference on System Sciences, HICSS-33, Maui, Hawaii, January 2000.
- [20] *** Language and Technology, Report of DGXIII to Commission of the European Communities, September 1992.
- [21] *** The Multilingual Information Society, Report of Commission of the European Communities, COM(95) 486/final, Brussels, November 1995.
- [22] *** Multilingualism in an Information Society, International Symposium organized by EC/DGXIII, UNESCO and Ministry of Foreign Affairs of the French Government, Paris 4-6 December 1997.
- [23] *** Les Frontieres du droit d’auteur ses limites et exceptions, ALAI Workshop, 14-17 Septembrie 1998, Ed. Australian Copyright Council, 1999.
- [24] *** Promotion and Use of Multilingualism and Universal Access to Cyberspace, UNESCO 31st session, November 2001.

Documente internaționale relevante pentru problema informatizării limbilor naturale și a multilingvismului în societatea informațională

- Universal Declaration of Human Rights, 1948 <http://www.unhchr.ch/udhr/index.htm>
- ACC Statement on Universal Access to basic Communication and Information Services <http://www.itu.int/acc/rtc/acc-rep.htm>
- Universal Copyright Convention, 1952 text and text revised in 1971 http://www.unesco.org/culture/laws/copyright/html_eng/page2.htm#ARTICLE
- Universal Declaration of Linguistic Rights/ Unesco/ Plurilinguisme <http://www.linguistic-declaration.org/index-gb.htm>
- Convention Establishing the World Intellectual Property Organization , signed at Stockholm on July 14, 1967 and as amended on September 28, 1979 <http://www.wipo.org/eng/main.htm>
- Berne Convention for the Protection of Literary and Artistic Works, Paris Act of July 24, 1971, as amended on September 28, 1979 http://www.wipo.org/eng/ipler/wo_ber0_.htm
- Treaty on Intellectual Property in Respect of Integrated Circuits, adopted at Washington, D.C., on May 26, 1989 http://www.wipo.org/eng/ipler/wo_top0_.htm
- WIPO Copyright Treaty, adopted by the Diplomatic Conference on December 20, 1996 <http://www.wipo.org/eng/diplconf/distrib/94dc.htm>
- Okinawa Charter on Global Information Society, July 23, 2000 <http://www.g8kyushu-okinawa.go.jp/e/documents/it1.html>