Lingvistica românească teoretică şi computaţională

Dan Tufiş
Membru corespondent al Academiei Române

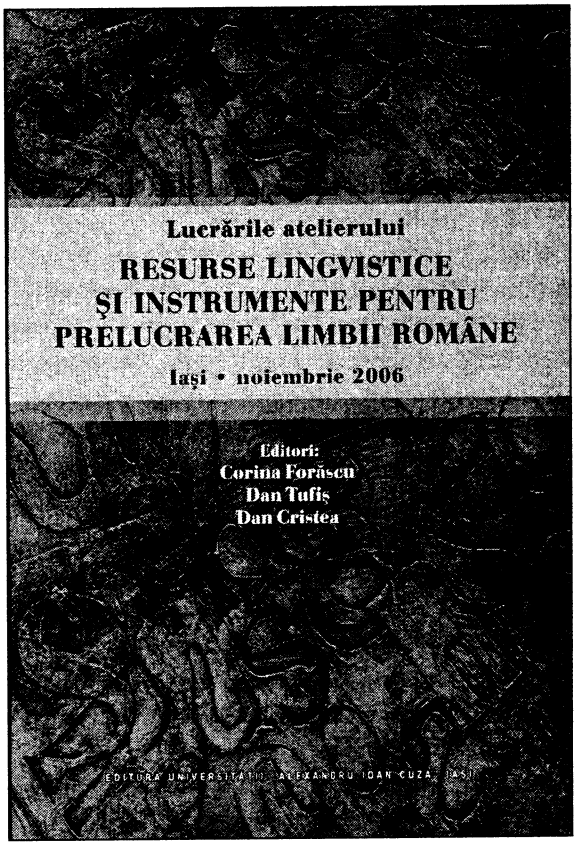
În contextul actual, al comunicării mediate de tehnologia informaţiei şi comunicaţiilor, prin contrast cu termenul deja încetăţenit de „limbă de circulaţie internaţională", putem vorbi, de „limbă de circulaţie electronică" în ciberspaţiu. Acest concept, pe lângă semnificaţia lui directă, are profunde implicaţii culturale, sociale, economice şi nu în ultimul rând etice statuând dreptul fiecărui cetăţean de a avea acces în propria limbă la cunoştinţele, informaţiile şi serviciile ciberspaţiului.

Voi încerca, în limita timpului disponibil, să schiţez câteva dintre elementele semnificative ale evoluţiei în ţara noastră în ultimii 5-6 ani a domeniului.

Această evoluţie a urmat în mod firesc priorităţile internaţionale născute din provocările şi necesităţile celei de a doua generaţii a comunităţilor bazate pe web şi servicii web, de trecerea de la un web al informaţiilor la un web al cunoştinţelor, al opiniilor şi sentimentelor, dar şi al afacerilor, al comerţului, al serviciilor de tot felul. Caracterul multilingual şi multicultural al noului web acordă tehnologiilor lingvistice o importanţă strategică.

Întâlnirea de azi este un moment important al evaluării domeniului tehnologiilor lingvistice pentru limba română, atât prin prisma onorantei participări la această masă rotundă, a domnului Comisar european Leonard Orban, prin deosebitul interes dovedit de prezenţa dumneavoastră numeroasă în Aula Academiei cât şi prin faptul că cercetarea românească de profil începe să culeagă roadele unei îndelungate perioade, pe care aş numi-o de construcţie infrastructurală.

Educaţia de specialitate a fost şi este o componentă esenţială a acestei construcţii. În anul 2001 au început primele programe de masterat în lingvistica teoretică şi computaţională (la Universitatea Bucureşti) şi prelucrarea limbajului natural (la Universitatea „A. I. Cuza" din Iaşi). Tot în anul 2001, Academia Română a aprobat înfiinţarea Comisiei de Informatizare pentru Limba Română, for consultativ al specialiştilor români. Câteva luni mai târziu Comisia a decis înfiinţarea unui organism mai larg, având responsabilităţi organizatorice şi executive: Consiliul de



Informatizare pentru Limba Română (ConsILR). ConsILR este o organizaţie deschisă, incluzând între membrii săi pe lângă specialişti români din ţară şi din străinătate, numeroşi studenţi, masteranzi sau doctoranzi de la principalele instituţii de cercetare, învăţământ sau industrie. Majoritatea manifestărilor ştiinţifice importante din România în domeniul tehnologiilor lingvistice au fost organizate de sau cu participarea nemijlocită a Comisiei şi Consiliului pentru Informatizarea Limbii Române: Conferinţele anuale ale ConsILR, şcolile internaţionale de Vară EUROLAN sau Conferinţele internaţionale SPED. Numeroşi absolvenţi români ai programelor de masterat amintite, sau a şcolilor de Vară EUROLAN sunt actualmente membri ai unor echipe de cercetare în universităţi sau institute prestigioase, majoritatea lor lucrând la proiecte multilinguale ce includ limba română sau fiind înscrişi în programe doctorale, pregătindu-şi teze asupra prelucrării limbii române

la importante universităţi din SUA, Canada, Italia, Spania, Marea Britanie, Franţa, Germania, Elveţia ş.a.

O serie de tineri specialişti români şi-au făcut deja un nume important în viaţa ştiinţifică internaţională a domeniului şi mulţi dintre cei care acum sunt în străinătate sunt membri activi ai Consiliului de Informatizare pentru Limba Română, făcând o foarte bună propagandă cercetării şi şcolii româneşti. Aceşti tineri deosebiţi păstrează o strânsă legătură cu formatorii lor din ţară încercând şi reuşind de multe ori să atragă echipe din România în proiectele în care instituţiile lor actuale sunt implicate. În mediile academice internaţionale se vorbeşte cu deosebit respect despre fenomenul românesc în domeniul PLN. Nu există conferinţă majoră a acestui domeniu, în care printre protagonişti să nu fie prezenţi şi cercetătorii români, din ţară sau din străinătate.

Ca o consecinţă firească a vizibilităţii cercetării româneşti, limba română a intrat de curând în circuitul select al limbilor pentru care se organizează competiţii de evaluare a tehnologiilor lingvistice, de regulă în context multilingual: alinierea lexicală – organizată de asociaţia nord-americană de lingvistică computaţională Edmonton – Canada 2003, dezambiguizarea semantică automată a cuvintelor, organizată de asociaţia mondială de lingvistică computaţională, Barcelona, Spania 2004, alinierea lexicală organizată de asociaţia mondială de lingvistică computaţională Ann-Arbor, SUA 2005, întrebare-răspuns cross-lingual română-engleză, organizată de Forumul European de Evaluare Cross-lingvală CLEF 2006 (Alicante, Spania), întrebare-răspuns în limba română organizată de Forumul European de Evaluare Cross-lingvală CLEF 2007 (Budapesta, Ungaria). Sistemele dezvoltate de specialiştii români nu numai că au câştigat toate competiţiile pentru limba română, dar s-au dovedit printre cele mai performante şi în competiţiile pentru limba engleză organizate în jurul unor probleme dificile ale PLN: rezolvarea automată a anaforelor, organizată de DAARC (Discourse Anaphora and Anaphor Resolution Colloquium) în 2007, detectarea implicaţiei lexicale şi respectiv dezambiguizarea semantică automată a cuvintelor organizate de asociaţia mondială de lingvistică computaţională, Praga, Republica Cehă 2007.

Prezenţa cercetătorilor români în proiectele de cercetare europeană din domeniul tehnologiilor lingvistice s-a îmbunătăţit (de pildă numai ICIA şi FII, instituţii a căror activitate o cunosc foarte bine, participând în 18 astfel de proiecte, 4 aflându-se în derulare, iar alte 4 urmând să înceapă în lunile următoare). O menţiune specială aş dori să fac asupra programului CLARIN, unul din cele 28 de programe incluse în planul strategic de dezvoltare a infrastructurii europene de cercetare. CLARIN (Common Language Research Infrastructure) are un orizont de realizare de 10 ani şi un buget de aproape 1/4 miliard de Euro, din care contribuţia UE va fi de peste 50%. Programul CLARIN reprezintă cea mai amplă acţiune europeană de până acum pentru armonizarea cercetărilor intra- şi intercomunitare în domeniul resurselor şi a tehnologiilor lingvistice. Acesta include actualmente 92 instituţii din 33 de ţări europene (3/4 din România). Ca o recunoaştere a contribuţiilor româneşti în domeniu, la nivelul decizional al acestui program România este reprezentată printr-un vicepreşedinte al consiliului ştiinţific şi printr-un membru al comitetului executiv.

Afirmam la începutul discursului meu că până nu demult cercetarea românească în domeniul tehnologiilor putea fi caracterizată ca traversând perioada construcţiei infrastructurale. În această perioadă de acumulări teoretice şi experimentale sprijinul Comisiei Europene a fost fundamental.

Un câştig remarcabil al perioadei la care mă refer îl reprezintă introducerea cercetării în domeniul tehnologiilor limbajului, ca una dintre priorităţi în mai toate programele naţionale ale principalilor finanţatori ai cercetării româneşti.

În contextul programului european CLARIN, multe din decalajele infrastructurale ale tehnologiilor limbajului pentru limba română vor fi recuperate mai repede şi mai coerent decât până acum. Cu atât mai mult cu cât începând cu programul CEEX şi continuând cu PNII, finanţarea internă a devenit şi sperăm că va rămâne semnificativă.

Ceea ce s-a realizat până în prezent în domeniul resurselor lingvistice standardizate (corpusuri adnotate, lexicoane multilinguale, modele statistice ale limbii române etc.) precum şi numeroase programe specializate în prelucrarea unor segmente critice ale lanţului complet de prelucrare lingvistică presupus de realizarea unor aplicaţii practice, constituie „cărămizi" cu care se poate trece la „asamblarea" sau proiectarea şi construirea unor sisteme complexe pentru prelucrarea textelor sau vorbirii în limba română.

Traducerea automată este numită pe drept cuvânt regina tehnologiilor lingvistice multilinguale, subsumând practic toate tipurile de resurse lingvistice şi cele mai robuste metode şi tehnici de prelucrare lingvistică.

Avansurile ştiinţifice din ultimii 10-15 ani, sprijinite de marile agenţii de finanţare a cercetării (Comisia Europeană, NSF, Darpa, NICT in Japonia etc.) sau de a marile companii (IBM, Microsoft, Google, Altavista, ş.a.) au permis ca pentru acele limbi pentru care existau resurse lingvistice

computaţionale adecvate să se realizeze în intervale de timp relativ scurte, sisteme automate de traducere foarte promiţătoare, mult mai robuste, cu o mai largă acoperire lingvistică, mai flexibile, mult mai rapide şi evident mai economice decât toate încercările celor 50 de ani de istorie anterioară a domeniului. Deşi sistemele dezvoltate sau în curs de realizare în Uniunea Europeană, în SUA sau în Asia, nu reuşesc deocamdată să atingă acurateţea şi fineţea traducătorilor umani, ele produc traduceri inteligibile, uneori chiar de bună calitate. Iar aceste traduceri se obţin în timpi de sute sau chiar mii de ori mai scurţi şi cu costuri neglijabile în raport cu traducerile umane.

Pentru limba română încercările, foarte puţine care au fost până acum, au eşuat sau sunt foarte departe de performanţele sistemelor de traducere amintite. Credem însă că stadiul actual al cunoaşterii ştiinţifice şi tehnologice, ca şi realizările obţinute în România constituie pentru prima dată premise realiste de lansare a unui proiect de traducere automată pentru limba română. Dintre aceste realizări, în ultima parte a întâlnirii de azi, vor fi prezentate câteva instrumente pentru prelucrarea automată a textelor, componente esenţiale în implementarea unui sistem competitiv de traducere automată pentru limba română:

- platforma web de servicii lingvistice bilingve (română şi engleză) incluzând identificarea automată a limbii unui text (actualmente sunt recunoscute 22 de limbi), segmentarea lexicală, dezambiguizarea morfo-sintactică şi lematizarea;
- ontologia lexicală Ro-wordnet, aliniată la nivel conceptual cu ontologiile lexicale pentru limbile engleză, bulgară, cehă, greacă, sârbă şi turcă. Prin utilizarea tranzitivităţii relaţiei de echivalenţă conceptuală cu limba engleză, se pot deriva alinieri conceptuale ale intrărilor din ontologia lexicală Ro-wordnet cu intrările ontologiilor lexicale alte altor zeci de limbi;
- sistemul DIAC^{+} un program complex de inserare automată a diacriticelor în textele româneşti scrise fără sau parţial cu diacritice;
- MT-KIT – un ansamblu de instrumente pentru construcţia modelelor de traducere, pentru alinierea lexicală automată, vizualizarea şi editarea structurilor de control într-un sistem de traducere automată bazat pe abordări statistice.

Demonstraţiile vor fi făcute prin intermediul unor scurte filme, dar persoanele interesate ne pot contacta ulterior pentru demonstraţii „live" şi detalii suplimentare.