LAUDATIO

Domnului profesor universitar doctor **Dan TUFIŞ**,
membru corespondent al Academiei Române,
cu ocazia acordării titlului de *Profesor de Onoare* al
Universităţii ,,Alexandru Ioan Cuza" din Iaşi

16 decembrie 2010

www.uaic.ro

Laudatio

Domnului profesor universitar doctor **Dan TUFIŞ**,
membru corespondent al Academiei Române

Domnule academician Dan Tufiş,
Doamnelor şi domnilor membri ai Senatului,
Doamnelor, domnişoarelor şi domnilor,

Colorat, surprinzător, patetic, dar şi sobru, cuminte, ori... de lemn. Oricare dintre aceste epitete poate fi aplicat limbajului uman. Dar indiferent de exuberanţa ori de cuminţenia lui, limbajul are structură. Limba are reguli rigide de utilizare, care permit însă o remarcabilă libertate de exprimare. Această contradicţie a fascinat cercetătorii în ştiinţele limbii, care, de la Saussure şi până la Chomsky şi Coşeriu, au încercat să-i descrie osatura formală.

Mai nou, informatica, aşa cum a făcut-o şi în alte domenii, vine cu metode şi idei noi. Universitatea noastră omagiază în această după-amiază o personalitate românească de prim rang care lucrează la graniţele dintre două domenii, cel al ştiinţelor exacte şi al umanisticii.

Formația

Domnul profesor dr. Dan Tufiş a primit iniţial o formaţie de inginer în calculatoare din partea Facultăţii de Automatică şi

Calculatoare a Universităţii „Politehnica” Bucureşti, pe care, ulterior, a completat-o cu un masterat în lingvistică computațională, la Institutul Lingvistic, Universitatea Santa Cruz, California, titlul de doctor fiindu-i oferit de aceeaşi universitate bucureşteană care l-a găzduit ca student, pentru o teză în care cele două laturi, de inginer software şi inginer lingvist, se împleteau: „Mediu de dezvoltare a sistemelor de dialog în limbaj natural". Obţine titlul de cercetător ştiinţific gradul I în anul 1992 la Institutul Naţional de Cercetare în Informatică din Bucureşti, iar în 2003 Universitatea noastră îi conferă titlul de profesor universitar.

În anul 1997 este ales membru corespondent al Academiei Române, la Secția de Ştiința şi Tehnologia Informației.

Între 2005-2010, reprezintă Secția de Ştiința şi Tehnologia Informației în Prezidiul Academiei Române. Din 1997 conduce Institutul de Cercetări pentru Inteligență Artificială al Academiei Române (ICIA), din București[1].

Activitatea ştiinţifică a profesorului Dan Tufiş este dedicată domeniului tehnologiilor limbajului natural, o ramură a ştiinței şi tehnologiei informației. Domeniul este unul preponderent tehnologic şi aplicativ, cu un înalt grad de creativitate şi inovare. Ni se pare important să menționăm că în întreaga sa activitate ştiințifică, domnul Dan Tufiş a parcurs ciclul complet al cercetării informatice, de la analiza conceptuală şi modelarea problemelor, la definirea de concepte, metode şi algoritmi, la implementarea soluțiilor descoperite,

[1] Până la restructurarea lui din 2002, cunoscut sub numele de Centrul de Cercetări Avansate în Învățarea Automată, Prelucrarea Limbajului Natural şi Modelare Conceptuală.

continuând cu testarea şi evaluarea performanţelor computaţionale ale implementărilor şi terminând cu diseminarea rezultatelor.

Anii de început

A început să studieze problematica sistemelor inteligente de prelucrare a limbajului natural în 1981, în 1982 propunând prima temă de cercetare din programul naţional dedicată prelucrării limbajului natural prin metode ale inteligenţei artificiale. La început a fost preocupat cu precădere de aspectele logice ale comunicării prin intermediul limbajului natural, dezvoltând metode şi tehnici noi de reprezentare şi prelucrare a cunoştinţelor lingvistice. La scurt timp realizează primul sistem de dialog în limba română, SDLR, valorificat ulterior prin intermediul Bibliotecii Naţionale de Programe.

1981 reprezintă şi anul începerii colaborării cu Centrul de Calcul al Universităţii „Alexandru Ioan Cuza" din Iaşi. Din această colaborare se naşte, în 1983, un nou sistem de întrebare/răspuns în limbaj natural, independent de domeniul de aplicaţie şi de limba de interogare. Programul, numit IURES, a constituit nu numai o premieră naţională, dar în multe privinţe includea soluţii inedite pe plan mondial. Astfel, schema de reprezentare a cunoştinţelor realiza o îmbinare a metodelor de reprezentare declarative cu cele procedurale. Se extindea conceptul de gramatică semantică prin introducerea (şi implementarea) noţiunii de operator logico-lingvistic. IURES includea metode originale de navigare într-o reţea semantică cu moştenire multiplă. Bazele formale ale sistemului IURES au constituit obiectul a peste 30 de articole, comunicări ştiinţifice, rapoarte de cercetare şi tehnice, care s-au bucurat de o deosebită apreciere, fiind citat de

numeroşi cercetători din ţară şi străinătate. Sistemul IURES a fost omologat internaţional în 1988 şi a constituit primul produs românesc de inteligenţă artificială exportat.

În paralel cu activitatea în domeniul prelucrării limbajului natural, cercetătorul Dan Tufiş realizează, în anii ’80, un mediu de programare funcţională, numit TC-LISP, care s-a impus în ţară ca produs standard de programare LISP pe minicalculatoare. Pentru mulţi ani, toate realizările semnificative în domeniul inteligenţei artificiale în România, până la apariţia pe scară largă a calculatoarelor personale, s-au implementat în TC-LISP, limbaj care prezenta o serie de concepte de programare inedite în programarea LISP: spaţii virtuale multiple, aritmetică ,,chirurgicală”, utilizarea controlată de utilizator a memoriei virtuale, programare paralelă etc.

Morfologia paradigmatică

În aceeaşi perioadă (1987-1989) realizează un sistem original de gestiune a dicţionarelor de dimensiuni mari destinate sistemelor de prelucrare a limbajului natural. Cercetările în domeniul morfologiei şi lexicologiei computaţionale s-au concretizat în plan teoretic cu un model computaţional original, morfologia paradigmatică. Dintre lucrările în care domnul Dan Tufiş a descris modelul morfologiei paradigmatice, *It Would Be Much Easier if WENT Were GOED*[2], prezentată la Conferinţa Europeană de Lingvistică Computaţională în

[2] D.Tufiş. “It Would Be Much Easier If WENT Were GOED”, in *Proceedings of the 4th European Conference of the Association for Computational Linguistics*, Manchester, 1989.

1989, a fost cotată drept cea mai valoroasă contribuție, alături de comunicarea reputatului specialist american Ronald Kaplan de la Institutul de Cercetări Stanford. La aceeaşi conferință, independent de cercetările domnului Tufiş, dr. Jo Calder de la Universitatea din Edinburgh a propus un model similar numit tot ,,morfologie paradigmatică". În momentul de față morfologia paradigmatică, alături de morfologia derivativă pe 2 niveluri[3], este considerată una dintre cele două modele morfologice unanim practicate (sub diferite variante) în tehnologia limbajului[4]. Teoria morfologiei paradigmatice a stat la baza implementării unui sistem de învățare automată a morfologiei limbilor naturale, numit PARADIGM, cercetări care au fost răsplătite cu premiul „Traian Vuia" al Academiei Române pe anul 1989.

Între anii 1993 şi 1995, în colaborare cu Centrul de Studii Semantice şi Cognitive din Geneva, dr. Dan Tufiş a dezvoltat un sistem integrat de prelucrări lingvistice numit Mac-ELU, considerat ca fiind un sistem de generația a 3-a (cea mai evoluată la nivelul anului 1993). Pe baza acestui sistem, colectivul condus de dr. Tufiş a lucrat la realizarea primului dicționar computațional românesc (bazat pe unificare) de mare acoperire lexicală. Dicționarul conținea peste 40.000 de intrări în formă lemă, pe baza cărora şi a morfologiei paradigmatice a limbii române, puteau fi recunoscute şi generate peste 1.000.000 de forme flexionate.

---

[3] Kimmo Koskenniemi, ''Two-level Model for Morphological Analysis'' in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* August 1983, Karlsruhe, West Germany, ed. A. Bundy, 1983.

[4] Richard Sproat, "Morphology and Computation" MIT Press, 1992.

Standarde lexicale şi lexicografice, dezambiguizare morfo-lexicală

Între anii 1995-2000 a coordonat activitățile de cercetare în alte trei proiecte europene: MULTEXT-EAST, TELRI (*Trans European Language Resources Infrastructure*), ELSNET (*Excelency in Language and Speech NETwork*) şi ELSNET-Goes-East. Le menționăm, dintre multele la care a participat dr. Dan Tufiş în cariera sa, pentru că acestea au fost printre primele având ca obiectiv alinierea metodologică şi tehnologică la standardele şi recomandările internaționale în domeniul ingineriei limbajului şi sinergizarea europeană a activităților naționale în acest domeniu.

Cercetările legate de standardizarea descrierilor morfo-lexicale au debutat la începutul anilor '90, coordonate de EAGLES (*Expert Advisory Group on Language Engineering Standards*), una dintre cele mai influente organizații profesionale europene în prelucrarea automată a limbajului. Cooptat în acest grup în 1994, dr. Tufiş a elaborat specificațiile pentru codificarea dicționarelor morfo-lexicale pentru limba română (1996), singura limbă est-europeană inclusă (la vremea respectivă) în standardele EAGLES.

Exploatând descrierea morfologiei paradigmatice în termenii de atribut valoare, ca şi adecvarea ei la tehnicile de învățare automată, dr. Tufiş dezvoltă în anul 1997 un model de proiectare automată, independent de limbă, a adnotatoarelor morfo-lexicale (generatoare de programe de dezambiguizare automată, la nivel morfo-lexical, a cuvintelor din texte arbitrare). Modelul dezambiguizării morfo-lexicale ierarhizate (cunoscut sub numele de *tiered-tagging*) este prezentat în 1999 şi implementat prima dată în adnotatorul morfo-lexical Q-Tag.

Ele reprezintă contribuții larg citate în comunitatea internațională. Unul din motivele aprecierii de care se bucură aceste contribuții, pe lângă performanțele superioare altor abordări, este faptul că sunt independente de limbă şi sunt conforme unor standarde şi recomandări internaționale asupra codificării morfo-lexicale.

Standardul EAGLES a fost extins (inclusiv pentru limba română) în cadrul proiectului european MULTEXT-EAST (finalizat în 1998), în care dr. Dan Tufiş a coordonat echipa din România. În cadrul acestui proiect s-a realizat nu numai extensia specificațiilor de codificare, pe baza noilor recomandări ale ISO, şi TEI-P3 (*Text Encoding Initiative*), dar şi implementarea primului lexicon morfo-lexical (conținând peste 400.000 de intrări la vremea respectivă) şi a primului corpus de limbă română (cu aproape 500.000 de cuvinte), cu o codificare aliniată la practica internațională. Aceste resurse lingvistice, unice la vremea respectivă în limba română, atât prin cantitate dar mai ales prin calitate (prelucrările statistice au fost validate şi corectate manual) au constituit nucleul dezvoltărilor ulterioare ale corpusurilor şi lexicoanelor de limbă română existente actualmente în institutul pe care îl conduce. În prezent, lexiconul conține peste 1.400.000 de intrări iar corpusurile construite de-a lungul anilor includ peste 4.000.000.000 de articole lexicale (cuvinte şi semne de punctuație) adnotate.

Proiectarea optimă a categoriilor gramaticale şi a atributelor relevante pentru dezambiguizarea statistică este încă o problemă puțin studiată, prof. Dan Tufiş fiind unul dintre primii cercetători care au afirmat că acuratețea procesului automat de etichetare morfo-lexicală depinde mult mai puternic de modelarea lingvistică decât de algoritmul de optimizare a etichetării celei mai probabile. Mai mult,

domnia sa a fost primul autor care am descris complet şi a implementat un algoritm de proiectare a categoriilor morfo-lexicale (ctagset) optime pentru prelucrarea corpusurilor. Metodologia ca şi implementarea acesteia sunt independente de limbă şi au fost aplicate, în afara limbii române, la o mulțime de alte limbi, tipologic foarte diferite de română: maghiară, germană, cehă, bulgară, estoniană, slovenă etc.

Lexicografia computațională şi ontologii lexicale

În perioada 1997-1999, în cadrul proiectului european CONCEDE (***CON****sortium for* ***C****entral and* ***E****astern* ***D****ictionaries* ***E****ncoding*) dr. Dan Tufiş a fost unul din realizatorii schemei XML generice[5] de codificare standardizată a dicționarelor explicative. Schema de codificare, cunoscută sub numele CONCEDE, a fost folosită pentru implementarea unor dicționare explicative pentru mai multe limbi (bulgară, cehă, engleză, estoniană, maghiară, slovenă şi desigur română). Un rezultat foarte semnificativ al activității în domeniul lexicografiei computaționale a fost realizarea unui compilator pentru dicționare în format tipografic (de exemplu, Word) ce analizează textul respectiv şi generează codul XML conform cu descrierea CONCEDE. Compilatorul, numit ***DIC*** se bazează pe gramatica convențiilor tipografice specifice şcolii româneşti de lexicografie, fiind parametrizabil atât în raport cu convențiile tipografice, cât şi cu schema XML a codificării țintă. Cu ajutorul acestui compilator, în anul

[5] O schemă XML de codificare este o specificație formală a unui limbaj de adnotare textuală (un limbaj de programare) pentru care fiecare element de adnotare are un context de utilizare şi o semantică riguros descrise.

2000 a fost finalizată implementarea conformă cu schema CONCEDE a întregului dicționar explicativ al limbii române (DEX, ediția 1996). Această implementare profesională a DEX-ului[6] permite regăsirea de informații lexicografice după o mulțime de criterii (categorie gramaticală, sufixe gramaticale sau lexicale, etimologie, variante, grupuri de litere conținute în cuvântul temă, definiții etc.). Aceste cercetări şi rezultatele obținute au facilitat lansarea în anul 2001 a proiectului de ontologie lexicală pentru limba română, proiect de un deosebit impact pentru comunitatea ştiințifică interesată de prelucrarea automată a limbii române.

Este vorba de proiectul european BalkaNet[7], în care grupul de cercetare al profesorului Tufiş şi cel de la Facultatea de Informatică a Universității ,,Alexandru Ioan Cuza" din Iaşi au fost responsabile de implementarea componentei de limbă română a Wordnet-ului – tezaur lexical, dezvoltat inițial pentru limba engleză, sub coordonarea profesorului George Miller de la Universitatea Princeton[8]. Global Wordnet Association[9] a indexat 63 de proiecte de dezvoltare de wordnet-uri în peste 50 de limbi şi multe din ele urmăresc principiile şi metodele proiectului BalkaNet. Prin amploarea mondială a mişcării

---

[6] Implementarea foarte populară a DEX-ului (în fapt o colecție de intrări din mai multe dicționare ale limbii române), finalizată în 2004 şi disponibilă pe web la adresa http://dexonline.ro/ constituie o simplă stocare într-o bază de date a textului amorf al intrărilor lexicale. Căutarea în dicționar se poate face numai după cuvântul titlu, şi mai recent folosind expresii regulate asupra textelor din definiții. În pofida simplității sale, dexonline este extrem de util, fiind consultat zilnic de foarte mulți utilizatori din întreaga lume.

[7] Proiectul a avut ca obiectiv crearea colecțiilor lexicale de tip Wordnet pentru 5 limbi din zona balcanică: bulgară, greacă, română, sârbă, turcă, aliniate la nivel de concept cu wordnetul englezesc.

[8] www.wordnet.com

[9] www.globalwordnet.org/

„wordnet", prin volumul de resurse umane şi financiare implicate, acest domeniu al lexicografiei computaționale îşi găseşte greu vreun rival în orice alt domeniu al informaticii.

De la finalizarea proiectului BalkaNet, dezvoltarea wordnet-ului românesc a continuat sub directa supervizare a prof. Tufiş, astfel că în prezent ontologia lexicală pentru limba română este printre cele mai mari din lume (conține peste 58.000 de clase de echivalență sinonimică, peste 84.500 de sensuri ale celor peste 51.600 de cuvinte distincte şi mai mult de 150.000 de relații semantice şi lexicale, neluând în considerare relația de sinonimie care leagă literalii în synset-uri), poate cea mai completă sub aspect lexicologic (de exemplu puține wordnet-uri ale altor limbi conțin definiții, sau dacă le conțin, de multe ori sunt în limba engleză, importate direct din wordnet-ul original, dezvoltat la Princeton).

Achiziţia automată de cunoştinţe lexicale; alinierea textelor paralele la nivel de propoziţie şi cuvânt

Cercetările prof. dr. Dan Tufiş în domeniul achiziției automate a cunoştințelor lexicale datează din anii 1997-1998, când domnia sa a dezvoltat un prim model inductiv, ce se baza pe echivalențele de traducere implicite existente între două texte reprezentând traduceri reciproce (bitext). În anii 2000-2002 au apărut şi primele rezultate competitive care au dovedit că modelele statistice de identificare a co-ocurențelor cros-linguale constituie o metodă robustă de aliniere lexicală a cuvintelor unui bitext, de extragere automată a dicționarelor bilingve şi, mai departe, de construire a modelelor de traducere (coloana vertebrală a unui sistem de traducere statistică). Problema

corectitudinii alinierii lexicale a fost identificată ca una din ştrangulările tehnologice ale progresului în traducerea automată, astfel încât în anul 2003, în cadrul Conferinței Asociației Nord-Americane de Lingvistică Computațională (NAACL-2003) de la Edmonton, a fost organizată o competiție pe această problemă, pentru mai multe perechi de limbi, printre care şi engleză-română.

Prof. Tufiş îşi mobilizează echipa şi, în mai puțin de 2 săptămâni, reuşesc să adapteze sistemul lor de aliniere, TREQ-AL, la cerințele organizatorilor, în două variante diferite. Sistemele ICIA[10] ocupă primele două locuri, devansând competitori celebri, cum ar fi: XEROX Research Europe (XRCE-locurile 3, 5, 6, 7), Universitatea din Alberta, Canada (Proalign-locul 4), Universitatea din Montreal, Canada (Ralign-locul 8), Universitatea Carnegie Melon, Institutul de Tehnologii ale Limbii, USA (BiBr – locurile 9, 10, 11), Universitatea din Minnesota, Duluth, USA (UMD - locurile 12, 13), MIT Research Corporation, USA (Fourday – locul 14).

Doi ani mai târziu, prof. Tufiş şi echipa sa propun un al doilea sistem, numit COWAL, care, combinând mai multe soluții de aliniere obținute independent, se dovedeşte a fi din nou cel mai performant la următoarea competiție, organizată de data aceasta de către Asociația de Lingvistică Computațională (ACL), în iunie 2005 la Ann Arbor, Michigan. În competiția din SUA au fost înscrise în concurs 37 de sisteme de la universități şi companii celebre (ISI-Universitatea din

[10] Institutul de Cercetări în Inteligență Artificială al Academiei Române, cunoscut în comunitatea internațională sub sigla RACAI (*Romanian Academy Centre for Artificial Intelligence*).

California, Universitatea din Maryland, Microsoft Research, Carnegie Mellon etc).

Dezambiguizarea semantică automată

Dezambiguizarea semantică (WSD – ***W**ord **S**ense **D**isambiguation*) este o altă problemă cheie în traducerea automată. În ipoteza identificării din context a sensului cuvântului polisemantic din limba sursă, relațiile de echivalență semantică interlinguale codificate de wordnet-urile multilingve de tip BalkaNet permit identificarea exactă a cuvântului potrivit de traducere în limba țintă. Cu cât distincțiile între sensuri sunt mai fine, cu atât este mai dificilă însă rezolvarea problemei WSD.

Metodologia dezvoltată de prof. Tufiş pentru rezolvarea problemei WSD în texte paralele este printre cele mai avansate din lume. Sistemul dezvoltat[11] a fost principalul mijloc de validare a corectitudinii semantice a wordnet-urilor dezvoltate în cadrul proiectului BalkaNet, fiind utilizat pentru toate cele 6 limbi ale proiectului. Mai multe lucrări publicate şi conferințe invitate au tratat pe larg diferite aspecte conceptuale, algoritmice sau noi dezvoltări în dezambiguizarea automată a sensurilor cuvintelor în texte paralele.

WEB Semantic şi servicii web de prelucrare a limbajului natural

Una dintre direcțiile cele mai pregnante ale cercetării actuale în tehnologiile limbajului se încadrează domeniului web-ului semantic.

[11] WSDTool.

Cercetările în această direcție ale prof. Dan Tufiş s-au focalizat pe problematica serviciilor web şi a aplicațiilor complexe bazate pe prelucrări distribuite geografic.

Începând din anul 2006, el coordonează activitatea de implementare a unei platforme de calcul pentru web-ul semantic, ce asigură servicii web lingvistice pentru limba română şi limba engleză. De curând, în colaborare cu Universitatea Marc Bloch din Strasbourg, serviciile ICIA asigură şi prelucrarea limbii franceze. Această platformă permite accesul de la distanță la majoritatea instrumentelor şi resurselor dezvoltate de ICIA în ultimii 15 ani.

Începând cu data de 1 iulie 2009, platforma de servicii web a fost deschisă comunității Internet, accesul fiind nerestricționat şi gratuit. Conform unei statistici cerute site-ului Google Analytics, în data de 12 decembrie 2010, la circa 1 an şi jumătate de la inaugurarea lui, situl oficial al Institutului de Cercetări pentru Inteligență Artificială al Academiei Române (ICIA), a fost vizitat de 11.301 de ori de 7.231 de utilizatori distincți ce au investigat 45.356 de pagini, cu o medie de peste 4 pagini la fiecare vizită şi un total de 1.443 ore UC de prelucrări pe serverul institutului. Vizitatorii au provenit din 815 de oraşe, din 85 de țări efectuând operații.

Includerea limbii române în competiții internaționale

Am menționat mai sus participarea echipei prof. Tufiş în competiții internaționale. Într-adevăr, de câtva timp, progresul în domeniul tehnologiilor lingvistice se apreciază în condiții riguroase de concurs. Limbile care pot oferi corpusuri suficient de mari de antrenament şi testare devin limbi de concurs. Aceste corpusuri sunt

dificil de achiziționat, întrucât necesită adnotări asupra fenomenelor lingvistice supuse probelor, care trebuie realizate de experți. Profesorul Tufiş a contribuit la includerea limbii române în competițiile internaționale, prin punerea la dispoziția celor interesați a unor corpusuri de limbă română, de mari dimensiuni, prelucrate adecvat pentru antrenarea sistemelor din competiții, precum şi la dispoziția participanților a o serie de instrumente de prelucrare a limbii române.

În afară de includerea limbii române în competițiile de alinieri lexicale interlinguale (**ACL-WA** de la Edmonton, 2003 şi de la Ann Arbor, Michigan, 2005) şi de dezambiguizare automată (**ACL-SENSEVAL III**, Barcelona, 2004), din anul 2006, limba română este prezentă şi în competițiile europene **CLEF**, care testează o gamă largă de problematici de inginerie lingvistică**.** Prof. Tufiş participă împreună cu doctoranzii săi la competițiile CLEF2006, ACL-SEMEVAL2007, CLEF 2007, CLEF2008, CLEF 2009 şi CLEF 2010 (pe care de altfel, cu excepția competițiilor din 2008 şi 2010, le câştigă, la concurență cu unele dintre cele mai reputate centre de cercetare sau companii de software din lume). Dar introducerea limbii române între limbile de concurs poate fi considerat mai important decât faptul că sistemele elaborate sub conducerea prof. Tufiş la ICIA s-au dovedit cele mai performante.

Sisteme de interogare în limbaj natural

În anul 2006 colectivul prof. Dan Tufiş elaborează un sistem de întrebare-răspuns în univers de discurs deschis (web) cross-lingual (întrebarea este pusă în limba română, iar răspunsul este căutat în documente de limbă engleză). Cele două grupuri de cercetare „surori”

(ICIA şi UAIC) au fost protagoniştii tuturor ediţiilor CLEF în care limba română a fost limbă de concurs.

În anul 2009, pentru prima oară, rezultatele competiţiei pentru sisteme de întrebare-răspuns în limbaj natural (**CLEF-ResPubliQA**) au putut fi comparate interlingual, întrucât întrebările de test (500) au fost aceleaşi în 7 limbi (bulgară, engleză, franceză, germană, italiană, română şi spaniolă) răspunsurile trebuind a fi căutate în corpusul paralel al legislaţiei europene „Acquis Communautaire", disponibil în 22 din limbile oficiale ale Uniunii Europene. Sistemul realizat în colectivul coordonat de prof. Tufiş a câştigat din nou detaşat, cu cel mai ridicat scor pe toate limbile, devansând toate celelalte 43 de sisteme competitoare.

Recuperarea automată a diacriticelor în textele de limbă română

Recuperarea diacriticelor în limba română este o problemă lingvistică netrivială. Pentru rezolvarea ei automată trebuie să se facă apel la o gamă largă de metode (analiză morfologică contextuală, dezambiguizare automată, modelul morfologiei paradigmatice, modelul combinat *Hidden Markov Model* şi *Maximum Entropy* ce implementează algoritmul *tiered-tagging*, corectare ortografică contextuală etc.). Începută în 1998, abia recent problema a putut fi rezolvată satisfăcător, în colectivul prof. Dan Tufiş[12].

Traducerea automată în şi din limba română

[12] Sistemul DIAC-PLUS, integrat în editorul MS Word, poate fi descărcat gratuit, de pe situl Institutului (www.racai.ro/diac).

Problema traducerii automate, veche de peste jumătate de secol, a înfierbântat mințile informaticienilor, fiind reluată în decursul anilor de nenumărate personalități. Actualitatea ei este reflectată şi în faptul că ultimul apel al Comisiei Europene pentru proiecte de cercetare în domeniul Tehnologiilor Limbajului a fost dedicat aproape exclusiv acestui domeniu. Abordările s-au orientat inițial spre crearea de modele simbolice, în care primordiale erau reguli de natură combinată sintaxă-semantică, care descriau particularitățile limbii sursă, ale celei țintă, precum şi modele de transfer a structurilor sintactice între cele două limbi.

Imposibilitatea de a formaliza exhaustiv extrem de vasta diversitate de exprimări, dar şi progresele realizate în abordărilor statistice şi a metodelor de procesare ghidate de date, au dus la mutarea centrului de greutate în cercetările de traducere automată către metode statistice. Majoritatea rezultatelor obținute după anul 2000 în domeniul lingvisticii corpusului au permis lansarea unor cercetări sistematice privind traducerea automată din şi în limba română, începând cu anul 2003.

Între anii 2005 şi 2010 au fost elaborate trei teze de doctorat în acest domeniu sub îndrumarea prof. Dan Tufiş, finalizate cu prototipuri funcționale de sisteme de traducere din limba engleză în limba română şi invers, clădite pe modele, metode, algoritmi şi resurse lingvistice create sub coordonarea dumnealui în perioada anilor 1995-2008.

Sistemele, antrenate pe resursele multilinguale, demonstrează un mare grad de generalitate şi reprezintă premise solide pentru realizarea unui sistem profesional de traducere automată, uşor adaptabil la orice pereche de limbi. Metoda alinierii lexicale prin

reificare a textelor comparabile, modelele de combinare a diferitelor ipoteze de traducere a unor fragmente de text, metodele de optimizare a calității traducerii şi alte probleme specifice traducerii automate prin metode statistice, constituie obiective de cercetare avansată în cadrul unor proiecte naționale şi europene, aflate în curs de desfăşurare.

Implicarea în cercetarea națională şi europeană

În cursul anilor, profesorul Dan Tufiş a coordonat ori a participat în 35 de proiecte de cercetare internaționale, iar după anul 1997, cel al primirii în Academie, a fost responsabilul a 13 teme anuale de cercetare, înscrise în planul de cercetare al Academiei Române.

A fost implicat direct în configurarea multor programe naționale. De exemplu, a fost directorul Programului Național INFOSOC - ,,Strategii şi soluții pentru Societatea Informațională - Societatea Cunoaşterii în România", între 2001-2002, o continuare a programului fundamental al Academiei Române dedicat problemelor strategice ale prelucrării automate a limbii române.

A coordonat colectivul de elaborare al subprogramului ,,Tehnologia Limbajului" din strategia şi planul național de cercetare al MCT ,,Societatea Informațională" (2005).

Din aprilie 2009 face parte din Comisia de monitorizare a proiectelor de cercetare şi diseminare a rezultatelor, organism al Consiliului Național al Cercetării Ştiințifice Universitare.

Din 1994 este expert UNESCO în domeniile inteligenței artificiale, lingvisticii computaționale şi al programării funcționale (LISP).

În 2001 a fost ales în comisia guvernamentală de experți UNESCO responsabili de elaborarea proiectului de recomandări privind ,,Promovarea multilingvismului şi a accesului universal în spațiul informațional".

În perioada ianuarie 1997- ianuarie 1999 a fost membru al *Advisory Board* al Asociației Europene de Lingvistică Computațională, cea mai importantă asociație profesională în domeniul Prelucrării Limbajului Natural.

În anul 2001 a înființat Comisia de Informatizare pentru Limba Română în subordinea Secției de Știința şi Tehnologia Informației, al cărei preşedinte este de la înființare. Această comisie constituie un organism consultativ şi un forum pentru discutarea priorităților şi problematicilor specifice prelucrării automate a limbii române. Ca organism executiv, mult mai larg, a fost înființat, în acelaşi an, Consorțiul pentru Informatizarea Limbii Române[13], cu scopul de a disemina contribuțiile publice (resurse şi instrumente de prelucrare a limbii române) dar şi ca forum de discuții cu toți partenerii interesați. Conferința Consorțiului a ajuns în anul 2010 la cea de a șaptea ediție.

A participat adesea, ca expert al CE, la pregătirea planurilor de finanțare a cercetării europene în tehnologia limbajului şi, desigur, în repetate rânduri, la evaluarea proiectelor propuse spre finanțare. În luna noiembrie a acestui an a fost invitat la Luxemburg, alături de alți 16 specialişti europeni, de către directorul general al „*INFSO.E1 Language Technologies & Machine Translation*", ca membru în comisia de stabilire a direcțiilor prioritare de cercetare în planul de lucru pe anii 2011-2012 pentru obiectivele tematice: tehnologii multilinguale şi

[13] http://consilr.info.uaic.ro/

managementul conținutului documentelor, traducerea automată (scris şi vorbit), interfețe inteligente etc.

Contribuții la formarea unei școli românești de lingvistică computațională

Profesorul Dan Tufiş are o contribuţie însemnată la crearea unei şcoli româneşti de lingvistică computaţională, cu largă recunoaştere internaţională. Un impresionant număr de tineri colaboratori din colectivele pe care le-a condus şi-au obținut doctorate şi lucrează actualmente în importante universități sau institute de cercetare din lume. Numeroşi studenți ai cursurilor de master ale Universității Bucureşti şi „Alexandru Ioan Cuza" din Iaşi sau la Şcolile de Vară Eurolan au obținut titlul de doctor sau sunt doctoranzi la universități de prestigiu.

Directorul şi academicianul Dan Tufiş a reuşit să creeze în institutul pe care l-a condus timp de 13 ani un climat de cercetare incitant, să formeze şi să păstreze în jurul său un colectiv, mereu reînnoit, pe care l-a format într-un spirit de competitivitate, care a stimulat înalta performanță. Credem că prof. dr. Dan Tufiş este un exemplu demn de urmat asupra modului în care un conducător de colectiv reuşeşte să implice tinerii cercetători în toate etapele activității de cercetare, de la aprofundarea şi dezvoltarea propriilor idei, la elaborarea de lucrări ştiinţifice, de la analiza şi evaluarea unor articole ştiinţifice, până la preluarea responsabilității unor componente importante în proiecte naţionale şi internaţionale de cercetare, tinerii merituoşi recunoscându-i aceste calităţi şi faptul că întotdeauna au fost promovați deschis, pe criterii exclusiv profesionale.

Colaborarea cu Universitatea ,,Alexandru Ioan Cuza" din Iaşi

Profesorul Tufiş a contribuit la stabilirea unui parteneriat exemplar de cercetare-învățământ între ICIA şi Facultatea de Informatică a Universității „Alexandru Ioan Cuza" din Iaşi. Acest parteneriat îşi are începuturile în urmă cu 30 de ani, concretizându-se în numeroase proiecte de cercetare comune, atât naționale cât şi internaționale, schimburi de doctoranzi şi cercetători, organizarea în comun de cursuri şi seminarii, la nivel de masterat, doctorat şi postuniversitare, precum şi a unor manifestări internaționale sau naționale, devenite deja tradiționale: Şcolile bienale de Vară ,,euroLAN", atelierul de lucru anual, devenit apoi conferinţă internaţională, „Resurse lingvistice şi instrumente pentru prelucrarea limbii române". Seria Şcolilor de Vară euroLAN a început în 1993, din 1995 prof. Tufiş fiind constant unul din co-directori (alături de inițiatorul acestora, prof. Dan Cristea, iar mai târziu de d-na Nancy Ide). La cele 9 ediții organizate până acum, peste 140 de mari personalități au susținut prelegeri la cel mai înalt nivel ştiinţific.

Concluzii

Prima dintre universităţile României are azi privilegiul de a reprimi în rândurile sale, în calitate de *Profesor de Onoare*, un om de ştiinţă cu merite excepţionale în cercetare şi formarea tinerilor cercetători.

Vă urăm, domnule Profesor, să aveți o lungă viață activă şi plină de satisfacții.

COMISIA DE ÎNTOCMIRE A LAUDATIO

Preşedinte
Profesor univ. dr. **Vasile IŞAN**,
Rectorul Universităţii „Alexandru Ioan Cuza” din Iaşi

Membri:
Profesor univ. dr. **Henri LUCHIAN**,
Prorector al Universităţii „Alexandru Ioan Cuza” din Iaşi

Profesor univ. dr. **Gheorghe GRIGORAŞ**,
Decan al Facultăţii de Informatică,
Universitatea „Alexandru Ioan Cuza” din Iaşi

Profesor univ. dr. **Dan CRISTEA**,
Prodecan al Facultăţii de Informatică,
Universitatea „Alexandru Ioan Cuza” din Iaşi
Directorul Departamentului de Cercetare - Facultatea de Informatică

Profesor univ. dr. **Dumitru OPREA**
Facultatea de Economie şi Administrarea Afacerilor
Universitatea „Alexandru Ioan Cuza” din Iaşi

Profesor univ. dr. **Florin Gheorghe FILIP**,
Preşedintele Secţiei de Ştiinţa şi Tehnologia Informaţiei a Academiei Române
Directorul General al Bibliotecii Academiei Române

Profesor univ. dr. **Eugen MUNTEANU**,
Facultatea de Litere, Universitatea „Alexandru Ioan Cuza” din Iaşi
Director al Institutului de Filologie Română „Alexandru Philippide” al Academiei Române
Director al Centrului de Studii Biblico-Filologice „Monumenta Linguae Dacoromanorum”, Universitatea „Alexandru Ioan Cuza” din Iaşi

Iaşi, 16 decembrie 2010