Menţinerea identităţii prin limbă: instrumente şi resurse lingvistice în format electronic

Gabriela Haja

Cercetările în planul prelucrării limbajului natural şi a limbii vorbite, al tehnologiei limbajului în genere, sunt de maximă importanță în contextul exploziei informaționale contemporane, în cadrul căreia Internetul funcționează ca un mijloc cu posibilități încă insuficient exploatate de noi. De aceea, implicarea specialiştilor români din domenii devenite complementare – informatica şi lingvistica – este firească.

În anul 2001 s-au creat şi circumstanțele acestei colaborări, atunci când a fost înființată, la Academia Română, Comisia de Informatizare pentru Limba Română (CILR), cu scopul declarat al apărării identității, a individualității şi a particularităților limbii române, prin susținerea studiilor lingvistice dintr-o perspectivă informațională. Încă înainte de această dată, a fost semnalată, la întâlniri şi discuții care au avut loc în Bucureşti, Iaşi şi Chişinău, necesitatea creării unui Consorțiu care să realizeze un cadru propice comunicării dintre specialiştii lingvişti şi informaticieni. Grație unei finanțări asigurate printr-un proiect național, intitulat „Societatea Informațională – Societatea Cunoaşterii", s-a realizat un sit electronic dedicat Consorțiului de Informatizare pentru Limba Română (http//consilr.info.uaic.ro). Principiile şi obiectivele acestui Consorțiu au fost prezentate de Dan Cristea în comunicarea *Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române*[1], la una dintre secțiunile Simpozionului Internațional *Identitatea limbii şi literaturii române în perspectiva globalizării*, organizat de Institutul de Filologie Română „A. Philippide" şi Asociația pentru Literatura Română şi Cultura Poporului Român „Astra" – Despărțământul „Mihail Kogălniceanu", la Iaşi, în 2002.

Potrivit acestor principii şi obiective, cercetători din domeniul informaticii de la Facultatea de Informatică a Universității „Alexandru Ioan Cuza" şi de la Institutul de Cercetări pentru Inteligență Artificială al Academiei Române din Bucureşti, împreună cu lingvişti de la Institutul de Filologie Română „A. Philippide" din Iaşi, s-au implicat în proiecte naționale şi internaționale, care contribuie la realizarea de resurse, instrumente şi tehnologii de prelucrare a limbii române, în toată complexitatea acesteia.

---

[1] Dan Cristea, Dan Tufiş, *Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române*, în Ofelia Ichim, Florin-Teodor Olariu (eds.), *Identitatea limbii şi literaturii române în perspectiva globalizării*, Iaşi, Editura „Trinitas", 2002, p. 193–210.

Unul dintre aceste proiecte, care poate asigura o importantă cale de integrare a limbii române în circuitul global al comunicării informatizate, este BalkaNet[2]. Un alt proiect, în care sunt implicate Academia Română, prin trei dintre institutele sale (Institutul de Informatică Teoretică şi Institutul de Filologie Română „A. Philippide" din Iaşi şi Institutul de Cercetări pentru Inteligență Artificială din Bucureşti) şi Universitatea ieşeană, prin Facultatea de Informatică, este cel intitulat *Tehnologii lingvistice pentru Web semantic*, care îşi propune să realizeze, pe de o parte, corpusuri adnotate şi, pe de altă parte, tehnologii pentru realizarea unui Web semantic pentru limba română. În esență, acesta va funcționa ca instrument indispensabil traducerilor automate.

S-au făcut până acum eforturi în acest sens, dar, datorită insuficienței resurselor, rezultatele sunt modeste[3].

Este evident că participarea lingviştilor la realizarea acestor resurse (baze de date) este esențială şi salutară. Proiectul amintit mai sus, *Tehnologii lingvistice pentru Web semantic*, lansat în acest an, sub programul-cadru INFOSOC, are drept principal scop ameliorarea situației actuale. Obiectivele detaliate ale proiectului nostru sunt realizarea de *corpusuri de texte adnotate*, de *standarde textuale* şi de *instrumente de prelucrare* a limbii.

Adnotarea textelor (scrise şi vorbite) se va face la mai multe niveluri: *fonetic*, *morfo-sintactic*, la nivelul *grupurilor nominale*, la nivel *sintactic* (specificându-se atât structurile sintactice, cât şi relațiile dintre ele, aşa numitele *tree-banks*), la nivel *semantic* (bănci propoziționale, *prop-banks*, care presupun specificarea structurilor de tip predicat-argument, cu identificarea statutului – opțional sau obligatoriu – şi a rolului fiecărui argument: agent, beneficiar, loc, mod etc.), la nivelul „formal", *al segmentării* (propoziții în cadrul frazelor, unități de discurs), la nivelul *relațiilor anaforice*: coreferința (echivalențe referențiale), referințe de tip parte-întreg, obiect-proprietate, posesor-obiect posedat etc., explicite sau implicite, la nivelul *sensului cuvintelor* (nivel care

---

[2] Vezi în acest sens: Dan Tufiş, *BalkaNet – tezaur lingvistic multilingv pentru limbile din Balcani,* în *op. cit*., p. 177–192 şi Dan Tufiş, Dan Cristea, *RO-BALKANET – ontologie lexicalizată, în context multilingv, pentru limba română,* în volumul *Societatea Informațională – Societatea Cunoaşterii*, Bucureşti, Academia Română, Secția de Ştiință şi Tehnologia Informației, Institutul de Cercetări pentru Inteligența Artificială, 2003, p. 137–164.

[3] Există în Internet maşini de tradus, dintre care amintim câteva: Softchim Translator, Millenium Computers, Edison Translator şi cele realizate de InterTran, care face următoarea precizare: „InterTran pentru perechile română – engleză funcționează ceva mai bine decât pentru română – [celelalte limbi]. Traducerile sunt de slabă calitate (chiar şi pentru română – engleză), translatorul fiind în curs de dezvoltare pentru limba română şi alte limbi din țările în curs de dezvoltare. Actualmente *nu există translatoare on-line profesionale, gen SysTran, care să includă limba română.* InterTran-ul are interfața on-line, dar *translatorul pentru română are baza de date extrem de mică*" (s.n.) (http: webenterprise.hast.gk/php/rotrans.php).

implică chestiuni de *dezambiguizare semantică*[4]) şi la nivelul *structurilor de discurs* (de exemplu, structura relațiilor retorice).

Aceste corpusuri pot fi adnotate semi-automat: pe baza unui model de analiză/adnotare realizat de specialistul lingvist, se pot genera adnotări automate, a căror corectitudine trebuie, fireşte, verificată. Adnotarea automată este facilitată de existența în format electronic a unor dicționare precum *Dicționarul explicativ al limbii române* (DEX), realizat de Institutul de Lingvistică „Iorgu Iordan" din Bucureşti, şi *Dicționarul de sinonime al limbii române* (DSR), redactat de Luiza şi Mircea Seche. La acest nivel intră în funcțiune standardele textuale, care facilitează adnotarea elementară automată.

Pentru prelucrarea inteligentă/informatizată a limbajului sunt necesare instrumente. Acestea sunt componente sofware, a căror caracteristică este independentă de limbă, iar aplicarea lor la limba română este dependentă de folosirea resurselor lingvistice specifice, româneşti. Domeniul instrumentelor aparține informaticienilor, dar buna lor funcționare nu poate fi realizată decât cu ajutorul asistenței specialiştilor lingvişti.

Realizarea acestor elemente implică, desigur, eforturi considerabile. Dar odată puse în practică rezultatele cercetărilor de acest tip, problema amenințării identității lingvistice şi implicit culturale a unei națiuni nu se mai pune, de vreme ce comunicarea vorbitorului de limbă română, prin Internet, cu orice vorbitor, indiferent de limba lui maternă, devine astfel posibilă.

The preservation of identity through language: instruments and linguistic resources in electronic format

In the context of new informational communities and of globalization, it appears very clear that it is necessary to create different types of instruments for the Romanian (spoken and written) language. The collaboration between specialists in these two fields – linguistics and computer science – is very important in order to accomplish this task. BalkaNet and Linguistic Technologies are two of the results of this collaboration. The former is an international project and its aim is to create a lexical ontology for the Balcanic languages which could be related to EuroWordNet. The Romanian Academy, by three of its institutes (the Institute of Theoretical Computer Science, the "A. Philippide" Institute of Romanian Philology, Iasi and the Research Institute for Artificial Intelligence, Bucharest), and the "Al.I. Cuza" University of Iasi achieve the latter. Its main goal is to create a large database for computerized translations. If these goals are fulfilled, the problem of the threat of the globalization against linguistic and cultural identity becomes irrelevant.

[4] Rezultate şi soluții foarte interesante în domeniul dezambiguizării semantice prezintă Dan Tufiş, în articolul *Dezambiguizarea automată a cuvintelor din corpusuri paralele, folosind echivalenții de traducere*, din volumul *Societatea Informațională – Societatea Cunoaşterii*, ed. cit., p. 235–268.