Contribuţii la depăşirea frontierelor lingvistice şi la integrarea limbii române în spaţiul informaţional digital European

Irimia Elena

Într-o epocă în care tehnologia informației digitale este omniprezentă și din ce în ce mai interconectată cu toate aspectele existenței umane, limbajul natural, în calitatea sa de transmițător de informație, este menit digitalizării. Pentru a supraviețui în societatea informațională a viitorului, o limbă are nevoie de tehnologii dedicate înțelegerii, procesării și generării sale.

Recensământul din 2011[1] anunța că populația stabilă a României se ridica la 20.121.641 de cetățeni, dintre care 385.729 erau plecați, temporar, în străinătate. La aceste cifre se adaugă numărul de 2,3 milioane de români emigrați după 1989, estimat de Institutul Național de Statistică la 1 ianuarie 2013[2]. Majoritatea acestor români concurează pe piața muncii și/sau pe piața economică europeană și intră în relații complexe cu instituțiile europene. Pentru asigurarea egalității de șanse a tuturor cetățenilor europeni în toate aspectele vieții, de la educație la muncă, de la acces la servicii de sănătate și sociale la participare publică și politică, multilingvismul pieței și instituțiilor europene este o condiție esențială. De aceea, politica de multilingvism a Uniunii Europene se sprijină pe trei scopuri prioritare: 1. încurajarea învățării limbilor străine și promovarea diversității lingvistice în societate; 2. promovarea unei economii multilingve; 3. facilitarea accesului pentru toți cetățenii la legislația și procedurile Uniunii Europene și informarea cetățenilor în propriile limbi materne. Implicit, una dintre principalele arii de activitate a Comisiei Europene este sprijinul activităților de cercetare și dezvoltare tehnologică în domeniul tehnologiilor lingvistice.

Comisia Europeană are ca prioritate dezvoltarea unei Piețe Digitale Unice (Digital Single Market), dar, în același timp, rămâne fidelă strategiei sale de promovare a multilingvismului în societatea Europeană. În acest sens, în aprilie 2015 a avut loc la Riga un summit european dedicat Pieței Digitale Unice Multilingve[3], la care România a participat și unde s-a angajat la producerea și promovarea de tehnologii digitale pentru înlăturarea barierelor lingvistice[4].

România are un dramatic deficit tehnologic de recuperat în acest domeniu, așa cum o semnala în 2012 studiul dedicat limbii române[1] în seria de studii META-NET asupra disponibilității și utilizării tehnologiei limbajului pentru 31 de limbi europene. Totuși, anterior acestui studiu și de atunci încoace, multe eforturi individuale, instituționale sau prin colaborarea mai multor instituții au avut loc în direcția micșorării acestor diferențe tehnologice. Mă voi concentra asupra proiectelor inițiate și susținute la Institutul de Cercetări pentru Inteligență Artificială în cadrul căruia lucrez și în care am avut ocazia de a-mi aduce modesta contribuție la integrarea limbii române în spațiul digital European, dar nu pretind să dau socoteală de toate eforturile depuse în

[1] http://www.recensamantromania.ro/rezultate-2/
[2] http://www.insse.ro/cms/ro/content/populatia-rezidenta-si-numarul-estimat-de-emigranti
[3] http://www.rigasummit2015.eu/multilingual-dsm
[4] http://www.meta-net.eu/projects/cracker/multimedia/mdsm-sria-draft.pdf

această instituție, și cu atât mai puțin de cele depuse de alte organisme de cercetare din România. O enumerare a acestor eforturi se regăsește în studiul META-NET deja menționat.

Una dintre cele mai importante și susținute inițiative la ICIA, care a debutat la începutul anilor 2000 în cadrul proiectului internațional BalkanNet și continuă și astăzi, este dezvoltarea wordnetului românesc, RoWordnet, o ontologie lexicală monolingvă aliniată printr-un index interlingual la Princeton Wordnet (Wordnetul original, a cărui dezvoltare a început în 1985), și, prin acesta, la o rețea globală de wordneturi, cunoscută sub numele de Global Wordnet[5]. RoWordnet este o resursă esențială în dezvoltarea a numeroase aplicații monolingve și multilingve, precum dezambiguizarea semantică, sistemele de traducere automată, sistemele întrebare-răspuns, etc. Echipa ICIA îl dezvoltă continuu, în direcția celorlalte interese de cercetare ale sale: de exemplu, pentru o vreme ne-am concentrat exclusiv pe implementarea unor sinseturi[6] pentru verbe, datorită preocupării pentru crearea de cadre de subcategorizare pentru acestea și utilizarea cadrelor pentru dezvoltarea unui analizor sintactic (en., parser) pentru limba română.

Traducerea automată este o altă preocupare importantă a cercetării noastre, susținută și de participarea la proiectul internațional ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation) în perioada 2010-2012. Scopul acestui proiect a fost dezvoltarea de metodologii și tehnologii prin care corpusuri comparabile de mari dimensiuni să fie exploatate pentru creșterea performanțelor aplicațiilor de traducere automată prin metode statistice[4]. Alte direcții de cercetare abordate au fost dezvoltarea unui sistem de traducere automată bazat pe exemple [5], dezvoltarea unui sistem de traducere automată pentru limbaj vorbit [6], dezvoltarea de corpusuri paralele care servesc drept resurse de antrenare pentru traducătoare statistice, oferirea online[7], spre utilizare în scopuri de cercetare, a unui sistem de traducere statistic fiabil si performant, pentru perechi de limbi precum engleză-română, germană-română, spaniolă-română.

De asemenea, suntem angajați, împreună cu Institutul de Informatică Teoretică – Iași, într-un program prioritar ale Academiei Române: realizarea unui corpus computațional de referință pentru limba română contemporană, denumit CoRoLa[7]. Acesta va fi o colecție de texte în format digital (scrise și orale) de dimensiune mare (cinci sute de milioane de cuvinte). Adnotate cu metainformații – precum autor, data publicării, etc. – și cu date lingvistice – precum părți de vorbire, forma din dicționar a cuvântului adnotat, dependențe sintactice etc –documentele vor fi disponibile liber online, spre consultare și valorificare în scopuri de cercetare.

În final, menționez proiectul care este obiectul stagiului meu postdoctoral: realizarea unei bănci de arbori sintactici pentru limba română (en. treebank), adnotați în formalismul gramaticilor de dependențe. În dezvoltarea acestuia, am avut un sprijin esențial în consultanța și resursele oferite

[5] http://globalwordnet.org/
[6] Mulțime de cuvinte cu același sens, mulțime de sinonime.
[7] http://www.racai.ro/tools/translation/racai-translation-system/

de o echipă a Institutului de Cercetare pentru Lingvistică Aplicată (IULA) din cadrul Universității Pompeu Fabra, Barcelona, vizitată prin stagiul postdoctoral de mobilitate internațională. Anterior, la IULA se accelerase construirea unui treebank pentru catalană[8] (o altă limbă europeană cu sprijin tehnologic redus) folosind un model statistic de analiză sintactică antrenat pe un corpus de limbă spaniolă. Bazându-ne pe similaritatea tipologică între cele trei limbi (catalană, spaniolă și română), aparținând familiei limbilor romanice, am utilizat aceeași strategie pentru a adnota automat corpusul românesc și am continuat printr-o etapă de corectare manuală a erorilor. Aceste erori provin în principal din diferențele structurale între limba spaniolă și limba română dar și, firesc, din natura procesului de adnotare statistică.

Proiectul reprezintă doar un pas dintr-o strategie menită să asigure resurse şi instrumente dedicate analizei sintactice a limbii române, domeniu în care suportul tehnologic este insuficient în acest moment. La ICIA, ne propunem să distribuim online acest treebank de dimensiuni modeste (5000 de propoziții) în cadrul corpusului CoRoLa și să îl folosim pentru a genera un model statistic de analiză sintactică, cu care să adnotăm ulterior întreaga componentă scrisă a CoRoLa. De asemenea, ne dorim să aliniem resursa noastră unei inițiative internaționale de standardizare a treebank-urilor de dependențe, Universal Dependency[8], care își propune să dezvolte treebank-uri într-o manieră consistentă cross-lingual, și la care participă în acest moment 25 de limbi.

Această lucrare a fost realizată în cadrul proiectului "Cultura română şi modele culturale europene: cercetare, sincronizare, durabilitate", cofinanţat de Uniunea Europeană şi Guvernul României din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013, contractul de finanţare nr.POSDRU/159/1.5/S/136077

1. TRANDABĂŢ, D., IRIMIA, E., BARBU MITITELU, V., CRISTEA, D., TUFIŞ, D. *The Romanian Language in the Digital Age. Limba română în era digitală*. In White Papers Series (Rehm, Georg and Uszkoreit, Hans). Springer-Verlag, Berlin, Heidelberg, 2012.
2. TUFIŞ, D., Cristea, D. *Methodological issues in building the Romanian Wordnet and consistency checks in BalkaNet.* In Proceedings of LREC 2002 Workshop on Wordnet Structures and Standardisation (Christodoulakis, Dimitris, N. and Kunze, Claudia and Lemnitzer, Lothar). Las Palmas, Spain, pp. 35--41, may 2002
3. BARBU MITITELU, V., DUMITRESCU, Ş.D., TUFIŞ, D. *News about the Romanian Wordnet*. In Proceedings of the 7th International Global WordNet Conference. Tartu, Estonia,2014
4. TUFIŞ, D., ION R., DUMITRESCU, Ş.D. *Wikipedia as an SMT Training Corpus*. In Proceedings of the International Conference on Recent Advances on Language Technology (RANLP 2013). Hissar, Bulgaria, September 2013

[8] http://universaldependencies.github.io/docs/introduction.html

5. IRIMIA, E., *EBMT experiments for the English-Romanian Language Pair*. In Recent Advances in Intelligent Information Systems (Klopotek et al.). Springer, Warsaw, pp. 91-102, 2009
6. TUFIȘ, D., BOROȘ, T., DUMITRESCU, Ș.D. *The RACAI Speech Translation System.* In Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SPED 2013). Cluj-Napoca, October 2013
7. BARBU MITITELU, V., IRIMIA, E.. The Provisional Structure of the reference Corpus of the Contemporary Romanian Language (CoRoLa). In Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language" (Colhon, Mihaela and Iftene, Adrian and Barbu Mititelu, Verginica and Cristea, Dan and Tufiș, Dan). Editura Universității „Alexandru Ioan Cuza", Iași, pp. 57–66, September 2014
8. ARIAS, B., BEL, N., FOMICHEVA, M., LARREA, I., LORENTE, M., MARIMON, M., MILA, A., VIVALDI, J., PADRO, M., *Boosting the creation of a treebank*, Proceedings of LREC 2014, Reykjavik, Iceland