

O selecție de proiecte în care am participat¹

I. Proiecte românești

IURES – sistem de întrebare-răspuns la baze de cunoștințe semantice

Dezvoltat în perioada 1983-1985, în colaborare cu Dan Tufiș: prima realizare românească a unui sistem de întrebare/răspuns, independent de domeniul de aplicație și de limba de interogare. IURES (*I Understand and Replay Eliminating Syntax*) lucra pe baza unei rețele semantice (construită manual) de reprezentare a cunoștințelor, care îmbina metode de reprezentare declarative și procedurale. Sistemul era capabil să reconstituie elipse în exprimarea întrebărilor și să rezolve referințe anaforice. Aplicații implementate cu IURES: interfață de interogare a bazei de date a Bibliotecii Naționale de Programe a ICI București, aplicație de dialog liber asupra geografiei României, la Institutul de Hidrologie Iași, aplicație de dialog în domeniul medical.

QUERNAL I și QUERNAL II – sisteme de întrebare-răspuns la baze de date

Între 1985 și 1989 am realizat sistemele de dialog la baze de date QUERNAL I (în Lisp) și QUERNAL II (în C). QUERNAL (*QUERY by NATural Language*) prelua câteva elemente de la IURES (cum ar fi lexiconul de echivalenți semantici), dar extindea domeniul de acces către informații depozitate în baze de date. Sistemul prelucra întrebarea în mai multe etape (lexical, sintactic, semantic) până o transforma într-o expresie a limbajului QRL (*Query Representation Language*), propus de mine și descris la nivel sintactic și semantic. Odată forma QRL a întrebării generată, ea era transformată de un traducător într-o exprimare a unui limbaj de interogare a bazelor de date (DTR, pe vremea aceea, ulterior SQL). Lansarea aplicației de gestionare a bazei de date asupra întrebării primite astfel în intrare, genera apoi răspunsul.

Sistemele QUERNAL au avut implementări pe bază de contract la Institutul de Cercetări Metalurgice București, la Întreprinderea “Flamura Roșie” Sibiu și la Întreprinderea de Foraj-Extracție Țiței Moinești.

InterOb – avatar uman expresiv

Între anii 2006 și 2008 am fost responsabil din partea UAIC-FII pentru proiectul CEEC „InterOb – Interacțiunea gestuală cu sistemele informatice și robotice”, coordonat de Univ. „Ștefan cel Mare” Suceava. În cadrul acestui proiect am realizat modelul tridimensional al unui avatar reprezentând o față umană, capabil de a exprima emoții prin gestică feței.

eDTLR – marele Dicționar Tezaur al Limbii Române în format electronic

Între anii 2007 și 2010 am coordonat proiectul “eDTLR – Dicționarul Tezaur al Limbii române în format electronic”, finanțat prin programul Parteneriate în Domenii Prioritare. *Dicționarul limbii române* reprezintă opera lexicografică cea mai importantă pentru cultura românească, redactat și editat în două etape (seria cunoscută sub sigla DA, în perioada 1906–1944 și seria nouă – DLR, în perioada 1965–2010), în 37 de volume și cuprinde cca 175.000 de cuvinte și variante, cu peste 1.300.000 de citate.

După terminarea finanțării, proiectul eDTLR încă continuă. Varianta electronică a marelui dicționar va permite consultarea interactivă multicriterială. Pentru elaborarea eDTLR s-au

¹ Până în anul 2025

realizat multe tehnologii originale, nu numai în câmpul lexicografiei computaționale românești, dar și pe plan mondial (s-a experimentat fezabilitatea conceptului de lexicografie colaborativă, prin darea în exploatare a unor interfețe de colaționare interactivă deschise contribuabililor voluntari, s-au creat tehnologii originale de parsare pentru aducerea formatului Word al paginilor la TEI-P5², care să pună în evidență câmpurile intrării, s-au realizat interfețe de interogare după orice combinație de criterii).

Proiectul a fost nominalizat de CNMP în decembrie 2010 ca cel mai valoros din panopia de proiecte a ariei tematice IX și a fost prezentat în multe conferințe internaționale, printre care și în cea a selectei *Dictionary Society of North America* (Chicago, iunie 2007).

MappingBooks – Jump in the book!

Între 2014 și 2017 am coordonat proiectul “MappingBooks – Jump in the book!” finanțat de UEFISCDI a dezvoltat un nou tip de produs electronic cu un înalt impact în educație, care a inclus tehnici de prelucrare a limbajului natural în combinație cu resurse geografice și tehnici de realitate virtuală.

DeLORo – Deep Learning for Old Romanian

Între 2019-2022 am coordonat proiectul DeLORo, care și-a propus construirea de modele de inteligență artificială (învățare profundă) pentru transcrierea în alfabetul latin a documentelor vechi românești (tipărite și unciale) redactate în slove chirilice.

II. Proiecte finanțate de organizații europene

Am 16 participări în proiecte internaționale, majoritatea dintre ele finanțate de organisme europene, în unul dintre ele în calitate de coordonator, iar în 15 – în calitate de responsabil din partea UAIC sau ARFI. Menționez pe cele mai semnificative.

MICH & MULTI-MORPH – morfologie prin clasificare

În 1993, la Universitatea Edinburgh, în laboratorul prof. Chris Mellish, și apoi la LIMSI – Universitatea Paris-Sud Orsay, în laboratorul dr. Michael Zock, am propus și implementat în Lisp un limbaj de clasificare pe care l-am denumit MICH. Setul de proprietăți ce caracterizau o anumită flexiune a unui cuvânt rezulta în urma clasificării caracteristicilor morfologice și ale rădăcinii într-o ierarhie de clase anterior definită. Cu ajutorul acestui limbaj am dezvoltat apoi sistemul Multi-Morph, capabil să facă generare morfologică în mai multe limbi. La LIMSI am realizat module pentru limbile spaniolă, română, rusă și franceză.

PROSODICS – pachet de prelucrare a semnalului vocal

Între anii 1994 și 1995 în Laboratorul de Lingvistică Computațională al Universității din Veneția, condus de prof. Rodolfo Delmonte, am propus și realizat (în colaborare cu 3 studenți de-ai mei), sistemul de prelucrare a vorbirii PROSODICS, care realiza analiza formantică a

² *Text Encoding Initiative* (<http://www.tei-c.org/index.xml>) este un consorțiu care menține standarde de reprezentare a textelor în format digital.

semnalului vocal, cu identificarea frecvenței fundamentale F0 și desenarea, pe baza ei, a traseelor melodice ale enunțurilor vocale.

BALKANET – rețeaua balcanică de wordneturi

Între 2001 și 2004 am fost responsabilul din partea UAIC al proiectului FP5 *BALKANET*, coordonat de Univ. Patras, Grecia, care a construit o rețea de wordnet-uri aliniate ale limbilor din Balcani: bulgara, greaca, româna, sârba, turca (în proiect mai fiind implicate și limbile cehă și engleză). S-a realizat prima versiune a wordnet-ului românesc, 25.000 de synseturi³ aliniate cu cele ale limbii engleze.

LT4eL – tehnologii lingvistice pentru eLearning

Între 2005 și 2008 am fost responsabilul din partea UAIC al proiectului FP6 *LT4eL – Language Technology for e-Learning*, coordonat de Univ. Utrecht, Olanda. Proiectul a produs tehnologii capabile să dezvolte resurse de eLearning din documente textuale, în mai multe limbi (engleză, germană, polonă, română, cehă, bulgară).

RoLTech – platformă pentru tehnologia limbii române: resurse, instrumente și interfețe

Între 2006 și 2008 am coordonat proiectul INTAS *RoLTech - Platform For Romanian Language Technology: Resources, Tools And Interfaces*, care a urmărit achiziționarea de resurse electronice pentru limba română și dezvoltarea de instrumente corespunzătoare pentru întreținerea și utilizarea lor în aplicații. S-a construit un portal web unde au fost depozitate aceste resurse lingvistice: dicționare, tezaure, corpusuri, modele statistice și simbolice, instrumente de prelucrare a limbii române, informații de interes pentru prelucrări asupra limbii române, precum și aplicații dedicate persoanelor care doresc să învețe limba română, vorbitorilor nativi de română și utilizatorilor experți de limba română (lexicografi în activitatea lor de creare de dicționare și lexicoane).

CLARIN –infrastructura europeană de cercetare lingvistică

Între 2008 și 2010 am fost responsabil din partea UAIC și membru al *Executive Board* în proiectul FP7 *CLARIN – Common Language Research Infrastructure*, coordonator Univ. Utrecht – Olanda. CLARIN a fost și continuă să fie unul dintre cele mai vizibile inițiative europene din domeniul PLN în secolul nostru. Scopul lui este crearea unei infrastructuri de cercetare care să permită integrarea și interoperabilitatea resurselor lingvistice și a tehnologiilor domeniului PLN. Resursele și serviciile oferite prin intermediul unei rețele de centre specializate sunt bazate pe tehnologii ale web-ului semantic, în așa fel încât să se facă față deosebirilor de format, structură ori terminologie. Proiectul își propune să servească, cu prioritate, cercetătorii din zona științelor sociale și umaniste într-o Europă multiculturală și multilingvă, platformele interactive inteligente create de CLARIN facilitând acestor categorii de cercetători accesul la o varietate mare de resurse lingvistice și instrumente de prelucrare a limbajului.

Am fost coeditor al publicației electronice CLARIN Newsletter, care a produs 13 numere (seria de articole Tadic&Cristea din anii 2008, 2009, 2010, 2011 în <https://profs.info.uaic.ro/dan.cristea/publications.html>).

³ Un synset este un set de lexicali (adnotați la sensuri) pentru care există contexte în care aceștia pot fi considerați sinonimi.

ALEAR – evoluția limbajului studiată pe roboți autonomi

Între anii 2008 și 2010 am fost responsabil din partea UAIC în proiectul FP7 *ALEAR – Artificial Language Evolution on Autonomous Robots*, coordonat de Univ. Humboldt Berlin, Germania, care și-a propus să folosească agenți inteligenți pentru a descifra mecanisme fundamentale ale cogniției și evoluției limbajului, aplicând un model al profesorului Luc Steels⁴. Roboți autonomi umanoizi își auto-organizează sisteme de comunicare, cu caracteristici similare celor întâlnite în limbile umane. Formarea conceptelor și invențiile lexicale și gramaticale sunt încorporate în agenți în cursul interacțiunilor acestora cu agenți similari.

Colectivul condus de mine s-a ocupat de propunerea de modele care să explice competențele agenților inteligenți în elaborarea discursului, adică a secvențelor de dialog mai lungi de un singur enunț, precum și apariția „spontană” în limbaj a referențelor pronominali și menținerea coerenței discursului. Modelele construite de noi au probat că pronumele poate apărea spontan în urma unei serii de interacțiuni între agenți, însă nu același lucru se întâmplă în cazul coerenței discursului. Concluziile noastre arată că abilitatea de a produce discursuri coerente nu reprezintă o adaptare prin învățare în urma unei șir de experimente de comunicare cu agenți similari, ci ar putea fi manifestarea directă a unui aparat cognitiv de genul memoriei imediate (asemănător stivei) și a unui principiu lingvistic minimalist, care favorizează economia în exprimare.

ATLAS – tehnologii lingvistice aplicate în biblioteci și publicistica online

Am fost responsabilul UAIC al proiectului PSP ICT #250467 *ATLAS – Applied Technology for Language-Aided Content Management Systems*, coordonat de *Tetacom Interactive Solutions*, Sofia, derulat între martie 2010 și februarie 2013, cu scopul de a dezvolta tehnologii lingvistice care să ajute persoanele care administrează și publică pe web conținut în limbile bulgară, engleză, germană, greacă, polonă și română. Platforma ATLAS, prin cele trei sisteme create, *i-Librarian*, *i-Publisher* și *EUDocLib*, reduce semnificativ timpul și eforturile utilizatorilor în crearea, gestionarea și publicarea de conținut multilingv, oferind totodată acces intuitiv la o vastă colecție de documente ale legislației UE, clasifică automat, rezumă, adnotează și traduce documente, indiferent de format și de limbă.

METANET4U – resurse și instrumente de PLN

Am fost responsabilul UAIC al proiectului PSP ICT #270893 *METANET4U*, coordonat de Univ. Lisabona, care s-a derulat între februarie 2010 și ianuarie 2012. *METANET4U* a reprezentat una dintre cele 4 componente ale rețelei de excelență META –NET. O realizare însemnată a consorțiului o reprezintă rețeaua de portaluri META-SHARE, pe care s-au stocat resursele și a instrumentele elaborate de parteneri, o parte din ele fiind românești.

Participarea în alte proiecte și rețele internaționale

- 2014 – 2018: responsabil din partea ARFI-IIT în proiectul DrUKoLA, finanțat de Fundația Humboldt, în partenerial cu AR-ICIA și Universitatea București.
- 2014 – 2018: E-READ (*Evolution of Reading in the Age of Digitization*) – un proiect COST care a vizat îmbunătățirea cunoașterii asupra implicațiilor pe care digitizarea poate să le aibă în asimilarea informațiilor citite.

⁴ Director al Laboratorului de Inteligență Artificială al Universității Vrije Bruxelles și al Laboratorului *Sony Computer Science* din Paris.

- 2013 – 2017: ENeL (*European Network of e-Lexicography*) – un proiect COST care a dus la crearea unei rețele europene de lexicografi și informaticieni interesați de găsirea de metode de acces mai facile la dicționare în format electronic și de dezvoltarea unui punct de vedere comun asupra e-lexicografiei și al exploatării mai eficiente a dicționarelor limbilor vorbite în Europa.
- 2007-2010: responsabil național și membru în Comitetul de Coordonare al proiectului COST A31: *Stability and Adaptation of Classification Systems in a Cross-Cultural Perspective*, coordonat de CNRS Paris.
- din 2008: am adus UAIC ca membru instituțional în FlaReNet (*Fostering Language Resources Network*) o rețea de excelență cu misiunea de a identifica priorități, obiective strategice pe termen scurt, mediu și lung în domeniul resurselor lingvistice și de a elabora recomandări consensuale sub forma unui plan de acțiune adresat CE, organizațiilor naționale și industriei.