UN INVENTAR AL LEXICOANELOR (DIGITALE) PENTRU LIMBA ROMÂNĂ. PERIOADA 1973–2022

CLAUDIUS-MARIAN TEODORESCU*

1. INTRODUCERE

În prezentul articol voi trece în revistă lexicoane sau resurse similare, digitale sau nu, concepute în România în perioada 1973–2022. Termenul *lexicon* este aici înțeles în sensul gramaticilor formale de tip generativ (Bidu-Vrănceanu *et al*. 2005: 289–290). Pentru această prezentare, am luat în considerație inițiativele modelare și de construire a unui lexicon, indiferent de gradul acestuia de cuprindere și indiferent de specificul unității de cercetare, de utilitate publică sau privată.

Criteriul de selecție al proiectelor de lexicoane prezentate (în număr de 19) este acela ca produsul final al proiectului analizat să fie un lexicon, eventual digital, sau o resursă, eventual digitală, care să poată fi asimilată cu un lexicon, chiar dacă resursa nu a fost elaborată efectiv, cum vom vedea că a fost cazul uneori. Prezentarea este cronologică, cu ordonarea făcută după anul de început al perioadei de elaborare a lexiconului. Pentru fiecare inițiativă am urmărit să compilez, din informațiile disponibile, modelul de date.

2. CRITERII PRIVIND ANALIZA, RESPECTIV SINTEZA DATELOR DE CERCETARE

Am ales opt seturi de informații despre fiecare proiect, fiecare set fiind inclus într-o secțiune separată, astfel încât să asigur o prezentare și o analiză caracterizate prin cuprindere și uniformitate. Secțiunile urmărite sistematic sunt următoarele: descriere, obiective, echipă, finanțare, date de intrare, metodologie de lucru, model pentru date și rezultate obținute. Mai jos se găsesc unele precizări despre câteva dintre aceste secțiuni.

2.1. Secțiunea *Finanțare*

Secțiunea despre finanțare a fost redactată urmărindu-se cu rigurozitate toate sursele de finanțare, eventual cu indicarea finanțatorilor și a sumelor denominate conform cu denominarea realizată în anul 2005. Acolo unde proiectele nu aveau

* Institutul de Filologie Română „Alexandru Philippide” al Academiei Române – Filiala Iași / Universitatea „Transilvania” din Brașov, România (claudius.teodorescu@gmail.com).

ALIL, t. LXIII, 2023, Bucureşti, p. 109–140
https://www.doi.org/10.59277/ALIL.2023.07

menționată nicio sursă de finanțare, am ales să consider, pornind de la contextul de dezvoltare a proiectului respectiv și chiar de la anumite mențiuni în planurile de cercetare anuale ale unității de cercetare în cadrul căreia s-a desfășurat proiectul, că finanțarea a fost asigurată prin includerea respectivului proiect în planul de cercetare al instituţiei.

2.2. Secțiunea *Date de intrare*

Am considerat necesară o secțiune referitoare la datele de intrare pentru proiect, pentru a se putea delimita și cuantifica într-un mod cât mai precis rezultatele proiectului, dar și pentru a avea o evaluare realistă a metodologiei de lucru. Această perspectivă a permis, de exemplu, să fie identificate cazurile fericite în care unele echipe de cercetare au primit finanțări pentru rafinări succesive ale metodologiei de lucru aplicate la aceleași date de intrare, cu rezultate din ce în ce mai bune.

2.3. Secțiunea *Metodologie de lucru*

Stabilirea metodologiei de lucru pentru fiecare proiect a implicat, în destule cazuri, un efort de a corela datele despre proiect din mai multe articole sau capitole de cărți. Rezultatele compilării cu rigurozitate a acestor informații ar putea să devină cu adevărat vizibile în contextul stabilirii metodologiei de lucru pentru realizarea unui lexicon digital pentru limba română.

2.4. Secțiunea *Model pentru date*

Pentru secțiunea privitoare la modelul pentru date, am făcut, dacă a fost necesar, conversii terminologice, deoarece unele proiecte, cele realizate în principal de cercetători informaticieni, folosesc mai curând o terminologie informatică decât una lingvistică ori filologică.

3. INVENTAR

3.1. Dicționar morfologic al limbii române (1973–1974)

Descriere

Acest dicționar a fost elaborat folosind un computer FELIX C-256 și limbajul de asamblare ASSIRIS.

Obiective

Obiectivul proiectului a fost realizarea unui algoritm pentru generarea de paradigme ale cuvintelor limbii române (Bocșa 1974b: 35).

Echipă

A fost formată din specialiști de la Facultatea de Matematică–Mecanică a Universității din Timișoara și de la Centrul Teritorial de Calcul Electronic Timișoara. Coordonarea a fost asumată de Minerva Bocșa.

Finanțare

A fost asigurată prin includerea în planul de cercetare al celor două instituții sus-menționate.

Date de intrare

Au fost inventariate „cca 2.000 cuvinte regulate, substantive, adjective și verbe, majoritatea cuprinse în listele de cuvinte din *Morfologia structurală a limbii române* de Valeria Guțu Romalo" (*ibidem*).

Metodologie de lucru

Au fost alese „doar *formele sintetice* (compuse dintr-un singur cuvânt)" dintre cuvintele din lucrarea *Morfologia structurală a limbii române* (*ibidem*). Pentru codificarea alternanțelor fonetice s-a folosit conceptul de *litere variabile*, elaborat de Gr. Moisil (*ibidem*: 36).

Model pentru date

Modelul pentru date este următorul: formă de bază; codificarea formei de bază, folosind sistemul de litere variabile; categoria gramaticală; numărul de alomorfe; numărul de caractere ale codificării formei de bază; numărul de caractere ale rădăcinii formei de bază; clasa flexionară (Bocșa 1974a: 43, Bocșa 1974b: 41).

Rezultate obținute

S-au obținut un dicționar morfologic cu 2.058 de cuvinte (Bocșa 1974a: 39), un automat pentru flexionare, respectiv un automat pentru lematizare sau lematizator (Bocșa 1974b: 50).

3.2. Banca de date fono-morfo-semantice a limbii române (1979–1996)

Descriere

În Dănăilă, Michăescu (1980) este prezentat un concept de bază de date pentru limba română, numită *Banca de date fono-morfo-semantice a limbii române* (BANDASEM) și destinată să fie „întocmită și exploatată exclusiv prin computerizare" (*ibidem*: 472), care să acopere „marile domenii de studiere a limbii române: fonetica și fonologia, ortografia și ortoepia, gramatica (morfologia și sintaxa), lexicul (semantica, formarea cuvintelor și etimologia) și stilistica" (*ibidem*: 475–476), din punct de vedere sincronic și diacronic. Această idee a fost prezentată consiliului științific al Institutului de Lingvistică din București în anul 1979 (Dănăilă 1996: 177).

Conceptul este foarte cuprinzător: „limita lui va coincide, în ultimă instanță, cu numărul de cuvinte distincte reperate în vocabularul limbii române" (*ibidem*: 473) sau „banca va include, în final, nu numai termenii comuni, ci și pe cei proprii (toponime și antroponime în sens larg)" (*ibidem*: 475).

Obiective

Proiectul și-a propus, pe de o parte, elaborarea de dicționare: dicționar general de frecvență a cuvintelor; dicționare de sinonime, analogii, asociații, antonime, omonime, paronime etc.; dicționar de frecvență grafo-fonematică; dicționar al tuturor formelor flexionare (*ibidem*: 471), *Dicționar confruntativ de*

omonime (DCO), *Dicționar confruntativ de sinonime, de analogii și de asociații* (DCSAAS) (ultimul în curs de elaborare în anul 1980) (*ibidem*: 472) şi, pe de altă parte, obținerea, pentru fiecare sens al unui cuvânt, de liste cu sinonime explicite (sinonimele propriu-zise) și cu sinonime implicite (analogii și asociații), ordonate după diverse criterii (Dănăilă 1993: 63).

Echipă

Acest concept i-a avut ca autori pe Ion Dănăilă și pe Radu Michăescu, iar inițiativa creării unei bănci de date semantice pentru limba română îi aparține acad. Ion Coteanu.

Finanțare

Finanțarea a fost asigurată prin includerea temei în planul de cercetare al Institutului de Lingvistică din București încă din 1979 (Dănăilă, Michăescu 1980: 471).

Date de intrare

S-a folosit *Dicționarul explicativ al limbii române*, deși se preconiza folosirea *Dicționarului limbii române* și a altor dicționare.

Metodologie de lucru

Se intenționa ca introducerea datelor să fie făcută în conformitate cu modelele tipurilor de intrări, după specific: date în clar sau date codificate (abrevieri, tipuri, subtipuri etc.). De asemenea, echipa își propunea să introducă mai întâi date sincronice, apoi date diacronice (*ibidem*: 476).

Model pentru date

Sunt propuse patru tipuri de modele (*ibidem*: 474–475): formă lexicală invariantă; formă lexicală variantă – sincronică (dublet literar, formă regională ori populară) sau diacronică (forme învechite); formă flexionară invariantă – formă gramaticală din paradigmele cuvintelor flexibile care este conformă cu normele limbii literare de astăzi; formă flexionară variantă – variațiile formale sincronice (dublete literare, forme gramaticale regionale sau populare) sau diacronice (forme gramaticale învechite).

Modelul unui articol de tipul formă lexicală invariantă este următorul: cuvânt-titlu; silabație; categorie lexico-gramaticală; indicații privind domeniile de folosire; rangul în vocabular; prima atestare; sens (număr de ordine; definiție lexicografică; sinonime; antonime; paronime; formele paradigmei; etimologie (limbă, etimon).

Modelul unui articol de tipul formă lexicală variantă este următorul: cuvânt-titlu; categorie lexico-gramaticală; indicații privind aria de răspândire; trimitere la forma lexicală invariantă de bază.

Modelul unui articol de tipul formă flexionară invariantă este următorul: cuvânt-titlu; categorie lexico-gramaticală; formele paradigmei.

Modelul unui articol de tipul formă flexionară variantă este următorul: cuvânt-titlu; categorie lexico-gramaticală; trimitere la forma flexionară invariantă de bază.

Rezultate obținute

Conform lui Dănăilă (1993: 61), au fost realizate modulele software pentru secțiunea de fonetică, respectiv pentru *Dicționarul confruntativ de sinonime,*

de analogii și de asociații al limbii române (DCSAAS). Partea de lexicografie a fost dezvoltată în cadrul Institutului de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București, iar secțiunile de fonetică-fonologie, morfologie și stilistică au revenit Institutului de Lingvistică și Istorie Literară „Sextil Pușcariu" din Cluj-Napoca (Șerban *et al*. 1996: 157).

În Vintilă-Rădulescu (2002: 29) se precizează că redactarea DCSAAS a ajuns până la litera S, apoi a fost întreruptă în favoarea lucrărilor prioritare ale Academiei; de asemenea, se adaugă că secțiunile de fonetică-fonologie, morfologie, respectiv stilistică ale BANDASEM au fost cedate Institutului de Lingvistică și Istorie Literară „Sextil Pușcariu" din Cluj-Napoca.

Din Dănăilă (1996: 178) aflăm că activitatea la DCSAAS s-a desfăşurat sporadic până în anul 1995 și că urma să continue cu mai multe resurse începând cu anul 1996.

3.3. Lexiconul tezaur CEI (1981–1997)

Descriere

Acest lexicon tezaur cuprinde concepte din domeniul electrotehnicii și termeni asociați în mai multe limbi. A fost elaborat în cadrul Comisiei Electrotehnice Internaționale, care este cel mai vechi organism internațional pentru standardizare. Activitatea de standardizare a terminologiei din domeniul electrotehnicii a început încă din 1910 (Timotin, Tănăsescu 1997: 105).

Obiective

Obiectivul acestui proiect a fost acela de a se realiza un instrument complementar pentru dicționarele tehnice, deoarece „definițiile dintr-un dicționar tehnic, în special cele stabilite ca rezultat al unui acord internațional, rămân deschise la critică și adesea exprimă un punct de vedere eficient și pragmatic, dar particular, ceea ce le reduce durata de viață" (Timotin, Tănăsescu 1997: 105). Un astfel de instrument, „mai puțin precis, dar mai flexibil, bogat în corelații și mai durabil", ar fi „mai bine adaptat pentru o descriere semantică simplă a unei limbi" (Timotin, Tănăsescu 1997: 105).

Echipă

Echipa, compusă din specialiști din România, a fost coordonată de dr. Remus (Baziliu) Răduleț şi dr. Florin Teodor Tănăsescu (Timotin, Tănăsescu 1997: 105–106).

Finanțare

A fost asigurată, în mai multe tranșe, de către Comisia Electrotehnică Internațională (Timotin, Tănăsescu 1997: 105–106).

Date de intrare

Datele de intrare au fost constituite de texte de specialitate.

Metodologie de lucru

Arborele de concepte a fost elaborat menționându-se mai multe tipuri de relații, după cum urmează: relații logice (specie–gen), ontologice (vecinătate în spațiu, timp, determinare sau incluziune), ierarhice (ascendente şi descendente) și asociative (simetrice) (Tănăsescu 2009: 112; Timotin, Tănăsescu 1997: 108–112).

Model pentru date

Baza de date cuprinde o listă de concepte și liste de termeni asociați, în mai multe limbi.

Modelul pentru datele din listele de termeni este următorul (Timotin, Tănăsescu 1997: 113–116): forma de bază; sinonime acceptate; sinonime nerecomandate (învechite); abrevieri; simboluri literale; traduceri în alte limbi; domeniu de folosire; relații semantice (cu arborele de concepte); surse (tip; ediție; an etc.).

Rezultate obținute

Prima versiune, din octombrie 1983, conținea 120.000 de expresii, inclusiv sinonime, în două limbi (franceză și engleză) (Timotin, Tănăsescu 1997: 105). În 1986 a apărut o ediție în limba franceză, ca o „listă ordonată alfabetic de termeni, descriptori și sinonime, într-un volum de 900 de pagini, fiecare descriptor fiind însoțit de vecinătatea semantică cea mai apropiată și de referințe la publicațiile sursă" (Timotin, Tănăsescu 1997: 106). În 1989 a apărut o „ediție bilingvă, în franceză și engleză, cu 26.000 de concepte" (Timotin, Tănăsescu 1997: 106). În 1993 s-a început adăugarea la baza de date a definițiilor în cea de-a treia limbă, româna (Timotin, Tănăsescu 1997: 106). Ediția din 1997 a tezaurului cuprindea „circa 30.000 de concepte și circa 33.000 de termeni în fiecare dintre cele trei limbi" (Timotin, Tănăsescu 1997: 106). Fondul de termeni ai acestui lexicon tezaur a fost pus la dispoziția Academiei Române prin acad. Eugen Simion (Tănăsescu 2009: 119).

3.4. Inventarul lexical al limbii române (1984)

Descriere

Proiectul *Inventarul lexical al limbii române* (ILEX), elaborat de Ion Dănăilă, a fost inclus în planul de cercetare pentru anul 1984 al Institutului de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București, dar activitatea nu a început, deoarece au fost elaborate alte lucrări. ILEX reprezintă o modificare a proiectului BANDASEM (Dănăilă, Michăescu 1980), cu schimbări bazate pe diverse criterii, dintre care cel mai important este că „nu va fi o lucrare de semantică prin excelență" (Dănăilă 1993: 61), așa cum se dorea a fi BANDASEM, dar se intenționa să cuprindă „date grafo-fonematice, semantice, morfologice, stilistice, dialectologice, de istorie a limbii etc." (*ibidem*: 63).

Obiective

ILEX și-a propus obținerea de prime atestări, arii de folosire, liste de cuvinte cu diverse particularități fonetice, morfologice, stilistice (*ibidem*) şi, prin aceasta, „să înregistreze, cu ajutorul calculatorului, lexicul românesc de astăzi și din toate vremurile" (*ibidem*: 61).

Echipă

Echipa era constituită din cercetători de la Institutul de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București, sub coordonarea lui Ion Dănăilă.

Finanțare

A fost asigurată prin includerea proiectului în planul de cercetare al Institutului de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București.

Date de intrare

Ca date de intrare pentru ILEX, urmau să fie folosite „dicționarele explicative generale, enciclopediile, lexicoanele, dicționarele speciale [...], dicționarele de terminologii tehnico-științifice [...] glosarele regionale (din volume sau din periodice)" (*ibidem*: 62), atât din România, cât și „din Basarabia și din Bucovina" (*ibidem*: 67).

Metodologie de lucru

Se propunea ca ILEX „să cuprindă toate cuvintele (împreună cu variantele lor lexicale) comune și proprii din limba română de astăzi și din toate timpurile", estimările fiind după cum urmează: circa 240.000 de cuvinte comune, circa 60.000–160.000 de cuvinte din terminologiile tehnico-științifice și circa 200.000 de toponime și antroponime, rezultând astfel un total de circa 500.000–600.000 de cuvinte, împărțite în *Inventarul lexical de cuvinte comune al limbii române* (ILEXCOM) și, respectiv, *Inventarul lexical de onomastică al limbii române* (ILEXON) (*ibidem*: 65).

În ceea ce privește lexicul comun, se dorea înregistrarea, din perspectivă sincronică și diacronică, a tuturor cuvintelor comune din următoarele registre: limba română literară generală, limbajul literaturii artistice (culte sau populare), limbajele tehnico-științifice, limbajele speciale, limbajul familiar, vorbirea populară, graiuri, argouri, jargoane (*ibidem*: 66).

Model pentru date

Sunt propuse două tipuri de modele: articole propriu-zise; articole de trimitere a variantelor la forma-bază (*ibidem*: 67).

Modelul unui articol propriu-zis este următorul: cod de identificare; cuvânt-titlu; categorie lexico-gramaticală; indicații privind domeniile de folosire; indicații privind aria de răspândire; prima atestare (an; sursă); variante.

Modelul unui articol de trimitere a variantei la forma-bază este următorul: cod de identificare; cuvânt-titlu; trimitere la forma de bază; categorie lexico-gramaticală; indicații privind domeniile de folosire; indicații privind aria de răspândire; prima atestare (an; sursă).

Rezultate obținute

Proiectul nu a fost pus în practică.

3.5. Dicționar sintactic al verbelor românești (1991–1994)

Descriere

Acest dicționar este un proiect comun al Departamentului de gramatică al Institutului de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București și al Institutului de Lingvistică din Chișinău (Șerbănescu 1994: 133).

Obiective

Prin acest dicționar sintactic s-a dorit o „schimbare de perspectivă” față de dicționarele explicative tradiționale, în sensul concentrării pe „comportamentul sintagmatic al unităților lexicale, pe posibilitățile lor combinatorii” (Șerbănescu 1994: 134).

Echipă

A fost formată din specialiști de la cele două instituții menționate mai sus, coordonarea fiind realizată de dr. Mioara Avram.

Finanțare

A fost asigurată prin includerea în planul de cercetare al fiecărei unități de cercetare implicate.

Date de intrare

Au fost selectate verbele din *Dicționarul explicativ al limbii române* și din *Suplimentul* la *Dicționarul explicativ al limbii române*.

Metodologie de lucru

„S-a operat o selecție în sensul eliminării variantelor, al limitării regionalismelor și arhaismelor, a sensurilor regionale și arhaice” (*ibidem*: 135). Au fost enumerate formulele sintactice, dar nu au fost înregistrate „fenomenele de omonimie sintactică și sinonimie sintactică (în accepție generativă)” (*ibidem*). În unele cazuri, au fost comasate sensuri care aveau aceeași formulă sintactică sau au fost separate sensuri cu formule sintactice diferite (*ibidem*: 145).

Model pentru date

Modelul a fost următorul (*ibidem*: 137): forma de bază; sensuri (definiție; formulă sintactică; citate).

Rezultate obținute

Dicționarul a ajuns doar în faza fișării materialului.

3.6. Sistem lexico-morfologic computerizat (1992–1994)

Descriere

Proiectul *Sistem lexico-morfologie computerizat* (SILEX) cuprinde un set de produse informatice pentru studiul și prelucrarea automată a limbii române, considerat de autori a fi „o primă realizare de acest fel în cadrul limbii române” (Cherata, Mihăescu 1994–1995: 201).

Obiective

Conform aceleiaşi surse (*ibidem*: 203), SILEX permite următoarele aplicații: dicționar ortografic și morfologic în format digital; corector ortografic și morfologic pentru limba română; diverse studii statistice lexicale și gramaticale pentru limba română; concordanțe; verificare automată a textelor românești în privința greșelilor de ortografie (aplicație deosebit de utilă în activitatea editurilor); studierea asistată de calculator a limbii române.

Proiectul a vizat realizarea a trei resurse digitale (*ibidem*: 202): un dicționar care să cuprindă un lexic de 50.000 de intrări; un generator de forme flexionare (*ibidem*: 277–279); un lematizator (*ibidem*: 273–277).

Echipă

SILEX a fost rezultatul unei colaborări între S.C. Software ITC S.A. din Cluj-Napoca și Centrul de Analiză a Textului de la Facultatea de Litere a Universității „Babeș-Bolyai” din Cluj-Napoca (*ibidem*: 273).

Finanțare

A fost asigurată prin includerea în planurile de lucru ale celor două unități de cercetare.

Date de intrare

Datele de intrare pentru dicționar sunt cele cuprinse în *Dicționarul explicativ al limbii române*, *Dicționarul ortografic, ortoepic și morfologic al limbii române* și *Gramatica Academiei,* cu „corectarea erorilor de descriere lingvistică și integrarea unor soluții propuse și validate în lucrări de specialitate mai recente” (*ibidem*: 204).

Metodologie de lucru

Pentru a obține un dicționar mai restrâns ca mărime, nu au fost incluse: participiile, inclusiv participiile-adjective, substantivele provenite din infinitivul lung, substantivele și adjectivele derivate din radical verbal cu ajutorul sufixului *-tor*, substantivele omografe cu adjectivele, substantivele, adjectivele și verbele derivate din radical verbal cu prefixele *ne-* și *re-*, ceea ce a permis o economie de circa 17.000 de articole de dicționar.

Model pentru date

Autorii prezintă modelele pentru articolele din dicționar diferențiate după clasa lexicală a cuvântului-titlu. Din rațiuni care țin de unitatea prezentării tuturor acestor proiecte, voi include mai jos o sinteză a acestor modele.

Modelul sintetic este următorul: cuvânt-titlu; categorie lexico-gramaticală; posibilități de conversiune sau de derivare; formele paradigmei sau radicalul/ radicalii paradigmei plus liste de terminații; omografele (adjective cu substantive).

Rezultate obținute

Folosindu-se „formalizarea și codificarea unei descrieri lingvistice preexistente”, dar și „soluții descriptive originale” (*ibidem*: 212), s-au obținut liste de terminații pentru generarea automată de forme flexionare și pentru realizarea de derivări lexicale, iar dicționarul conține 31.000 de intrări. Prima utilizare reală a automatului pentru flexionare a fost „pentru verificarea corectitudinii datelor din dicționar și din listele de terminații” (*ibidem*: 279).

3.7. Lexicon EGLU (1993–1997)

Descriere

Acest lexicon a fost elaborat în două etape de dezvoltare: în prima (1993–1995) ca parte a sistemului Environnment Generique Linguistique d’Unification (EGLU), care mai conținea un analizor de text și un generator de text (Tufiș *et al*. 1997: 84), iar în a doua (1995–1997) – în cadrul proiectului MULTEXT-East[1].

[1] http://nl.ijs.si/ME/

Obiective

Prima etapă de dezvoltare a vizat realizarea unei „codificări a morfologiei limbii române și a unui lexicon asociat” (*ibidem*).

În cadrul proiectului MULTEXT-East, se urmărea ca lexiconul să fie dezvoltat cu mai multe secțiuni: fonologie, morfologie, sintaxă, clasificare terminologică, semantică lexicală (Tufiș *et al*. 1996: 94) și să fie folosit pentru prelucrarea de corpusuri, în cadrul platformei de prelucrare de corpusuri MULTEXT (*ibidem*: 98).

Echipă

În decursul ambelor etape de dezvoltare, acest lexicon a fost proiectat și dezvoltat în cadrul Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” din București, de un colectiv coordonat de către acad. Dan Tufiș.

Finanțare

Pentru prima etapă, finanțarea a fost asigurată prin includerea în planul de cercetare al Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” din București. În cea de a doua etapă, proiectul a fost finanțat de Comisia Europeană, prin grantul 0106-33181, în valoare de 320.000 de euro (https://cordis.europa.eu/project/id/0106-33181). Proiectul MULTEXT-East a avut mai multe obiective, în afara acestui lexicon.

Date de intrare

Datele de intrare sunt reprezentate de circa 80% dintre intrările din *Dicționarul explicativ al limbii române* (*ibidem*: 93).

Metodologie de lucru

Pentru prima etapă, s-a dezvoltat produsul informatic necesar; s-a realizat „o descriere cuprinzătoare a morfologiei limbii române folosind un limbaj neutru și reversibil (reprezentare atribut plat-valoare)”, reprezentare care a fost ulterior convertită în formalismul EGLU (*ibidem*: 84–85). În a doua etapă au fost inventariate toate intrările din *Dicționarul explicativ al limbii române*, apoi acestea au fost introduse în baza de date și au fost testate (*ibidem*: 98).

Model pentru date

A fost folosit formalismul Mac-ELU, care implică o clasificare după categorii majore și minore care sunt ordonate sub forma unui graf orientat aciclic (–: 94). Acesta este derivat din formalismul PATR-II (Shieber *et al*. 1983), destinat pentru scrierea de gramatici. Modelul propus este următorul: formă flexionară; formă de bază; categorie lexicală; set de categorii gramaticale (în cursul operațiilor automate de prelucrare de text, categoria gramaticală corectă este selectată în funcție de afixul lexical sau gramatical asociat); reguli de derivare; reguli de flexionare; formule sintactice (Tufiș *et al.* 1997: 87 și 94).

Rezultate obținute

S-au realizat un lexicon care cuprinde „aproximativ 50 000 cuvinte din fondul de cuvinte al limbii române” (*ibidem*: 98), un automat pentru lematizare, un automat pentru derivare, un automat pentru flexionare (*ibidem*: 87), precum și un corector ortografic (*ibidem*: 92). Lexiconul a permis generarea unui lexicon de forme

flexionare cu peste 1.000.000 de forme (*ibidem*), care este disponibil la adresa web https://clarin.si/repository/xmlui/bitstream/handle/11356/1041/wfl-ro.txt.gz.

3.8. Lexiconul din Baza de cunoștințe lingvistice pentru limba română (1996)

Descriere

Proiectul descrie un „model de reprezentare unitară a cunoștințelor lingvistice despre limba română"; aceste cunoștințe includ „un lexicon, regulile lexico-sintagmatice, precum și regulile gramaticale și reprezentările semantice" (Curteanu *et al*. 1996: 101).

Obiective

Această bază de cunoștințe este destinată pentru „analizarea și generarea textelor" (*ibidem*), pentru analiză morfologică, pentru consultarea lexiconului, a setului de reguli lexico-sintagmatice și a setului de reguli pentru analiză sintactică, pentru analiză semantică, precum și pentru analiza discursului (*ibidem*: 102). Generarea implică folosirea unei reprezentări logico-semantică a textului de generat, apoi se folosesc regulile semantice, cele sintactice și cele lexico-sintagmatice (*ibidem*: 102–103).

Echipă

Proiectul a fost elaborat în cadrul Institutului de Informatică Teoretică al Academiei Române, Filiala Iași, de o echipă coordonată de către dr. Neculai Curteanu.

Finanțare

A fost asigurată prin includerea în planul de cercetare al Institutului de Informatică Teoretică al Academiei Române, Filiala Iași.

Date de intrare

Nu există date de intrare, deoarece implementarea a fost parțială, iar lexiconul nu a fost elaborat.

Metodologie de lucru

Au fost implementate seturile de reguli și au fost dezvoltate automatele informatice.

Model pentru date

Modelul unei intrări din lexicon este următorul (codificarea informațiilor este făcută în conformitate cu teoria Head-Driven Phrase Structure Grammar): cuvânt-titlu; categorie gramaticală; categorie lexicală; caracteristici morfologice; valențe; sintagme și grupuri sintagmatice din care poate face parte cuvântul-titlu; informații semantice (*ibidem*: 103).

Rezultate obținute

Au fost dezvoltate un automat informatic pentru analiză morfologică și unul pentru analiză sintactică, un automat informatic pentru generarea de text plecând de la modelul de reprezentare a limbii, precum și unele seturi de reguli privind organizarea lexiconului și reprezentările sintactice și logice (*ibidem*: 102).

3.9. Lexiconul român (1996–2011)

Descriere

Este vorba despre un lexicon care conține „informații fonologice, morfologice, sintactice etc.” (Cojocaru 1996: 37).

Obiective

Pentru procesarea limbajului natural: restabilire diacritice, clasificare documente, dezambiguizare morfologică (Petic 2014: 248).

Echipă

Lexiconul a fost elaborat în cadrul Institutului de Matematică și Știința Computerelor al Academiei de Științe din Republica Moldova, de o echipă coordonată de dr. Constantin Ciubotaru.

Finanțare

A fost asigurată printr-o serie de granturi acordate de Consiliul Suprem pentru Ştiinţă şi Dezvoltare Tehnologică al Academiei de Ştiinţe a Republicii Moldova: *Aplicaţii ale resurselor reutilizabile pentru tehnologia limbajului natural*[2] (2003–2005) și RoLTech. *Platforme pentru tehnologia limbii române: resurse, instrumentar şi interfeţe*[3] (2006–2008).

Date de intrare

Dicționarul explicativ al limbii române și *Supliment* la *Dicționarul explicativ al limbii române*.

Metodologie de lucru

Pentru generarea paradigmei fiecărui cuvânt, a fost dezvoltat un automat informatic pentru flexionare, care folosește regulile din lucrarea *Dictionnaire morphologique de la langue roumaine* de Alf Lombard și Constantin Gâdei. Aceste reguli au fost formalizate, în vederea procesării automate, în cadrul unei gramatici automate, rezultând 866 de reguli şi 320 de seturi de terminaţii (Petic 2014: 246).

Dintre regulile de flexionare formalizate în cadrul acestui proiect, le menționez, pentru utilizări ulterioare, pe cele legate de numărul de forme flexionare pentru fiecare categorie gramaticală: douăsprezece pentru substantive, 20 pentru adjective și 35, 39 sau 40 pentru verbe (Boian *et al.* 2011: 684).

S-a realizat verificarea integrității datelor din baza de date cu ajutorul verificatorului ortografic al aplicației informatice MS Word și de către specialiști (*ibidem*: 683).

Model pentru date

Fiecare intrare conține următoarele informații: formă de bază; categorie gramaticală; indicație de folosire; forme flexionare (cu informații morfologice); posibile funcții sintactice (*ibidem*: 682–683, Petic 2014: 247).

[2] https://math.md/projects/064110303P/

[3] http://math.md/projects/INTAS05-104-7633/

Rezultate obținute

Un lexicon cu circa 100.000 de cuvinte-titlu, un corector ortografic, un automat de flexionare și un algoritm de căutare pentru pagini web (Boian *et al*. 2011: 685). Lexiconul este disponibil la adresa web https://math.md/elrr/res_main.php?main=Main+page.

3.10. Lexiconul TEZAROM (1992–2010)

Descriere

Lexiconul TEZAROM a fost dezvoltat succesiv în cadrul a șase proiecte de cercetare, prima etapă realizându-se după desprinderea din proiectul *Inventar lexical al limbii române* (vezi *supra*).

Cele șase etape sunt următoarele: *Banca de date pentru limba română*, în perioada 1992–1996; *Tezaur computerizat al limbii române pentru procesarea textelor scrise*, în perioada 1996–1998, în cadrul programului național ORIZONT 2000; *Model formalizat complet al morfologiei limbii române. Dicționar morfologic computerizat*, în perioada 2001–2002, în cadrul programului național RELANSIN; *Sinteza din text a vorbirii în limba română pe baza unui sistem expert lingvistic pentru interfețe multisenzoriale*, în perioada 2004–2006, în cadrul programului național INFOSOC; *Sistem informatic pentru analiza sintagmatică a textelor în limba română* (SIASTRO), nr. 86 CEEX II 03/ 31.07.2006, în perioada 2006–2008; *Sistem interactiv de analiză gramaticală pentru limba română scrisă. Model teoretic şi tehnologie de implementare* (SINTEGRO), în perioada 2007–2010, în cadrul programului național PNCDI.

Obiective

Obiectivul comun al celor șase proiecte de cercetare a fost elaborarea unei baze de date și a unor instrumente digitale pentru analiza textelor scrise în limba română.

Dintre obiectivele specifice fiecăruia dintre cele șase proiecte, menționez următoarele (RoLingva 2009):

1. „procesare electronică a unei analize, realizate în prealabil de lingvist (pe foi de programare), asupra întregului stoc de cuvinte” (Șerban *et al*. 1996: 157); „studierea structurii fonetice a limbii române, asistată de calculator”; „elaborarea unui program de analiză automată a versificației în limba română”; „realizarea unui dicționar ortografic-ortoepic al limbii române” (*ibidem*: 159); realizarea unor dicționare de autor, pentru „Dosoftei, Eminescu, Sadoveanu, Blaga sau Agârbiceanu” (*ibidem*), în vederea cunoașterii „inventarului lexical și a sensurilor contextuale”, adică a stilului de autor (*ibidem*), și a obținerii de atestări mai vechi, de cuvinte noi și de sensuri noi;

2. începerea formalizării morfologiei limbii române;

3. finalizarea formalizării morfologiei limbii române, formalizarea silabației, precum și a accentuării cuvintelor;

4. realizarea *Dicționarului morfologic român*;

5. formalizarea fonologiei și adăugarea de reguli și atribute fonologice la resursele lingvistice existente;

6. extinderea *Dicţionarului morfologic român*, precum şi „analiza sintactico-semantică a textelor româneşti, cu aplicaţii dintre cele mai diverse: corectoare gramaticale, sisteme de asistare a învăţării limbii române (atât de către vorbitorii nativi, cât şi de către cei străini), sisteme de adnotare a corpusurilor, sisteme de traducere automată etc.” (Tămâianu-Morita *et al*. 2006–2007: 83).

Analizorul sintagmatic dezvoltat în cadrul celui de al șaselea proiect se bazează pe opera gramaticală a lui D.D. Drașoveanu, permițând „punerea în evidenţă a tuturor flectivelor şi conectorilor dintr-un text”, determinându-se astfel toate sintagmele din textul respectiv (Vîlcu 2008: 117).

Echipă

Pentru cele șase proiecte, echipele au avut varii componențe, fiind formate din specialiști de la Universitatea „Babeș-Bolyai” din Cluj-Napoca, de la S.C. Software ITC S.A. din Cluj-Napoca, de la Universitatea Tehnică din Cluj-Napoca, respectiv de la Institutul de Lingvistică și Istorie Literară „Sextil Pușcariu” al Academiei Române, Filiala Cluj-Napoca. Primul proiect a fost coordonat de dr. Felicia Șerban, al patrulea proiect de dr. Sanda Cherata, al cincilea proiect de dr. Emma Tămâianu-Morita, iar al șaselea de dr. Manuela Mihăescu.

Finanțare

A fost asigurată prin granturile menționate mai sus. Pentru prima etapă, finanțarea a fost asigurată prin includerea proiectului în planul de cercetare al Institutul de Lingvistică și Istorie Literară „Sextil Pușcariu”. Pentru etapa a cincea, finanțarea a fost de 1.438.155 lei, iar pentru etapa a șasea, de 2.000.000 lei.

Date de intrare

Pentru primul proiect, au fost 50.000 de intrări ale *Dicționarului explicativ al limbii române*, ediția din 1975, iar pentru celelalte au fost datele din *Dicţionarul morfologic român*, precum și diversele formalisme elaborate.

Metodologie de lucru

Au fost elaborate formalizările morfologică (Cherata, Mihăescu 2008: 154), respectiv fonologică, iar pe baza lor au fost elaborate dicționarul morfologic și celelalte produse informatice. Intrărilor din *Dicţionarul morfologic român* li s-au adăugat informații morfosintactice. Au fost studiate formatele de reprezentare a datelor lexico-morfologice şi sintactice care să permită şi extinderea spre bazele de date terminologice (*ibidem*: 165).

Model pentru date

Modelul global pentru intrările din lexicon este următorul: forma de bază; limba articolului; categoria gramaticală; indicații privind domeniul de folosire; definiție; informații morfologice; informații sintactice; informații „despre elemente fonetice, cum ar fi silabaţia fonetică, accent, împreună cu regulile de migrare a acestuia în cadrul flexiunii” (*ibidem*: 161).

Rezultate obținute

Voi menționa în mod global rezultatele proiectelor, deoarece unele dintre rezultate au fost obținute în cursul a două sau mai multe etape ale proiectului global:

• pachetul lingvistic *Ortograf*, care cuprinde un corector de texte pentru limba română, „un despărțitor în silabe la sfârșit de rând și un dicționar de sinonime” și care „a fost cumpărat de firma Microsoft pentru a fi inclus în produsele de tip Office” (RoLingva 2009); în RoLingva 2009 se precizează un fapt remarcabil: „respectarea teoriei lingvistice clasice a permis formalizarea regulilor de silabație și obținerea unui algoritm de silabisire cu rezultate 100% corecte”; unele detalii despre modelul de formalizare a generării automate de forme flexionare sunt date în Peev *et al.* 1997;

• *Dicționar morfologic român*, cu peste 80.000 de cuvinte-titlu și peste 2.000.000 de forme flexionare;

• un automat pentru analize sintagmatice care permite realizarea unui extractor de termeni, descrierea formală a sintagmelor limbii române, un automat pentru analize morfologice și un automat pentru analize sintactice (Cherata, Mihăescu 2008: 165 și Tămâianu-Morita *et al.* 2006–2007: 80–82);

• în privința formalizării fonologiei limbii române, în RoLingva 2009 este menționat un alt fapt remarcabil, și anume: „corectitudinea confruntării modelului fonetic cu realitatea a fost de 100%”;

• un extractor de termeni-candidați pentru terminologii (Peev, Șerban 2008).

3.11. RoWordNet (2001–2013)

Descriere

Acest lexicon este o „implementare în limba română a șirurilor sinonimice din wordnet-ul englezesc” (Tufiș *et al.* 2006: 17). Lexiconul a fost dezvoltat în două etape de lucru: prima etapă a fost în cadrul proiectului BalkaNET[4], cod IST-2000 29388, desfășurat în perioada 2001–2004, finanțat de Uniunea Europeană și cofinanțat de Ministerul Educației și Cercetării prin programul CORINT; a doua etapă s-a desfăşurat în cadrul unor proiecte de cercetare (ROTEL[5], STAR[6], SIR-RESDEC[7], ACCURAT[8], METANET4U[9]) și în cadrul planului de cercetare al Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” din București, în perioada 2004–2008 (Tufiș 2008: 3, Mititelu *et al.* 2014: 268).

Obiective

Pentru prima etapă, obiectivul proiectului BalkaNET a fost elaborarea a câte o bază de date lexicală de tipul WordNet pentru fiecare dintre limbile din Balcani. Aceste baze de date urmau să fie aliniate prin intermediul *Indexului Interlingual*[10]. Încea de a doua etapă, proiectul a vizat îmbunătățirea bazelor de date.

[4] http://dblab.upatras.gr/balkanet/

[5] http://ai.ici.ro/rotel_eng/index.htm

[6] https://racai.ro/en/research-activities/national-projects/

[7] https://racai.ro/en/research-activities/national-projects/

[8] http://accurat-project.eu/

[9] http://metanet4u.eu/

[10] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/beschreibung/interlingual-index/

Echipă

Consorțiul proiectul BalkaNET a fost format din următorii parteneri: Universitatea din Patras (Grecia), Institutul de Cercetare Academică în Tehnologia Computerelor din Patras (Grecia), Universitatea din Atena (Grecia), Institutul Limbii Bulgare al Academiei Bulgare de Științe din Sofia (Bulgaria), Universitatea din Plovdiv (Bulgaria), Universitatea Sabanci din Istanbul (Turcia), Universitatea „Alexandru Ioan Cuza" din Iași, Centrul pentru Cercetări Avansate în Învățarea Automată al Academiei Române din București, Universitatea Masaryk – Facultatea de Informatică din Brno (Cehia), firma Memodata din Dumont d'Urville, Franța. Activitatea a fost coordonată de dr. Dimitris Christodoulakis. Pentru etapa a doua, echipa a fost formată din specialiști de la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București, iar coordonarea a fost asigurată de dr. Dan Tufiș.

Finanțare

Pentru prima etapă, finanțarea din partea Uniunii Europene a fost de 1.542.253 de euro. Pentru cea de a doua etapă, finanțarea a fost asigurată prin includerea proiectului în planul de lucru al Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București.

Date de intrare

Pentru prima etapă s-au avut în vedere *Dicționarul explicativ al limbii române* (ediția din 1996), *Dicționarul de sinonime al limbii române*, precum și un dicționar român–englez elaborat în cadrul Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București (Tufiș *et al.* 2004: 108). La acestea s-au adăugat un corpus de limbă română contemporană, segmentat și adnotat morfosintactic, și un corpus paralel român–englez (*ibidem*: 110). Pentru cea de a doua etapă, datele de intrare au fost datele de ieșire ale primei etape, datele din corpusul EuroWordNet[11], lexiconul VerbNet[12], corpusul 1984[13], corpusul Acquis Communautaire și tezaurul Eurovoc[14], precum și cuvinte din inventarul lexical al versiunii în limba română a Wikipedia (Mititelu *et al.* 2014: 268).

Metodologie de lucru

Pentru prima etapă: toate dicționarele, odată aduse în formă digitală, au fost codificate cu ajutorul standardului XML[15]. *Dicționarul de sinonime al limbii române* a fost digitizat și stocat într-o bază de date de tip ACCESS (Tufiș, Cristea 2002: 3); din el au fost excluse variantele arhaice și variantele regionale (Tufiș *et al.* 2004: 110); dicționarul român–englez a fost extras în mod automat dintr-un corpus paralel român–englez (*ibidem*). La seriile de sinonime din *Dicționarul de sinonime al limbii române* au fost adăugate definițiile din *Dicționarul explicativ al limbii române* și au fost etichetate sensurile (Tufiș *et al.* 2004: 111). Pentru a doua etapă s-au realizat „clasificarea tuturor șirurilor de sinonime în conformitate cu

[11] https://archive.illc.uva.nl/EuroWordNet/
[12] https://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[13] http://nl.ijs.si/ME/Vault/CD/docs/1984.html
[14] https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en
[15] https://w3.org/TR/xml/

taxonomia DOMAINS 3.1[16], alinierea tuturor șirurilor de sinonime cu conceptele ontologiei SUMO&MILO[17], etichetarea fiecărui șir de sinonime cu un marcaj de subiectivitate de tip SentiWordNet"[18] și s-au stabilit şi relații de tip sintagmatic între cuvinte, nu doar de tip paradigmatic (Tufiș 2008: 3–4).

Model pentru date

Pentru prima etapă, modelul pentru date a fost: identificator WordNet (pentru punere în corespondență cu WordNet-ul); șir de sinonime (pentru fiecare sinonim: formă de bază, număr de sens); identificator și tip de relație cu Indexul interlingual; definiție; citate de redactor (Tufiș *et al.* 2004: 111). În cea de-a doua etapă, modelul de date a fost următorul: identificator WordNet (pentru punere în corespondență cu WordNet-ul); categorie gramaticală; șir de sinonime (pentru fiecare sinonim: formă de bază; număr de sens); definiție; identificator și tip de relație cu Indexul interlingual; marcator pentru seturi de concepte de bază; domeniu de folosire; identificator pentru concept SUMO și tip relație de corespondență (RoWordNet 2013).

Rezultate obținute

„La sfârşitul celor trei ani ai proiectului BalkaNet, WordNet-ul românesc conținea 20.381 de șiruri de sinonime cu peste 36.000 de cuvinte-titlu (leme) şi o platformă software dedicată" (Tufiș 2008: 3), iar lexiconul aferent avea circa 70.000 de intrări (Tufiș *et al.* 2004: 108). La sfârșitul celei de-a doua etape, RoWordNet-ul conținea 48.541 de șiruri de sinonime și 43.037 de leme (Tufiș 2008: 4). Baza de date, în formă binară, se găsește la adresa web https://github.com/dumitrescustefan/RoWordNet/blob/master/rowordnet/ rowordnet.pickle.

3.12. Lexiconul RoMorphoDict (2002–2005)

Descriere

Acest lexicon a fost constituit sub forma unui dicționar de forme flexionare, la care au fost adăugate formele despărțite în silabe (Barbu 2008: 1937). Activitatea de elaborare a lexiconului s-a desfășurat în două etape, 2002–2004, respectiv 2004–2005.

Obiective

Obiectivul comun al celor două proiecte a fost realizarea unui lexicon de forme flexionare.

Echipă

Pentru prima etapă, echipa a fost formată din specialiști de la Universitatea din București și de la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București. Coordonarea proiectului a fost asigurată de dr. Emil Ionescu. Pentru a doua etapă, activitatea de cercetare a fost asumată exclusiv de dr. Ana-Maria Barbu.

[16] https://wndomains.fbk.eu/
[17] https://ontologyportal.org/
[18] https://github.com/aesuli/SentiWordNet

Finanțare

Pentru prima etapă – proiectul *Dicţionar morfologic în format electronic al limbii române contemporane* –, finanțarea a fost asigurată prin grantul nr. 33549/18A/2002, acordat de Consiliul Național al Cercetării Științifice din Învățământul Superior (CNCSIS), iar pentru a doua etapă – proiectul *Silabisitor* –, prin grantul nr. 8/2005, acordat autoarei, dr. Ana-Maria Barbu, de Institutul Limbii Române (*ibidem*: 1941).

Date de intrare

În prima etapă, au fost cele circa 65.000 de intrări din versiunea digitală a *Dicţionarului ortografic, ortoepic şi morfologic al limbii române* (ediția din 1989) (*ibidem*: 1937). Pentru cea de a doua etapă, datele de intrare au fost datele de ieșire ale primei etape (*ibidem*: 1940).

Metodologie de lucru

Pentru prima etapă, au fost elaborate două gramatici pentru „analizarea intrărilor verbale, respectiv a intrărilor nominale (substantive și adjective)”, iar pronumele au fost tratate manual; după aceea, paradigma a fost completată (*ibidem*: 1938). În cea de a doua etapă, s-au folosind indicațiile privind silabația, care sunt prezente în *Dicţionarul ortografic, ortoepic şi morfologic al limbii române*, un automat informatic pentru silabație și un inventar cu diftongii și triftongii limbii române (*ibidem*: 1940).

Model pentru date

Pentru prima etapă, modelul a fost următorul: categorie gramaticală; paradigmă (cod; forme flexionare cu informații morfologice pentru fiecare formă flexionară); formă de bază; glosă (*ibidem*: 1939). Pentru cea de a doua etapă, s-a adăugat forma despărțită în silabe pentru fiecare formă flexionară; forma despărțită în silabe prezintă și accent (*ibidem*: 1940–1941).

Rezultate obținute

După prima etapă, lexiconul de forme flexionare conținea 775.969 de forme flexionare, care corespund la circa 65.000 de forme de bază (*ibidem*: 1938). Pentru cea de a doua etapă, s-a adăugat forma despărțită în silabe pentru un număr de 525.530 de forme flexionare (*ibidem*: 1940).

3.13. DCVLR (2005–2018)

Descriere

Dicționarul de contexte verbale pentru limba română (DCVLR) conține „reprezentări de tip HPSG[19] ale valențelor verbale pentru limba română” (Barbu *et al*. 2006: 2467). Dezvoltarea acestui lexicon s-a făcut în două etape: perioada 2005–2007, respectiv perioada 2008–2018.

[19] Head-Driven Phrase Structure Grammar (Gramatica centrilor de sintagmă) (Barbu, Ionescu 2008: 296).

Obiective

Pentru ambele etape de dezvoltare s-a urmărit „umplerea unui gol în domeniul resurselor și instrumentelor pentru limba română”, deoarece, în momentul respectiv, limba română nu avea „un dicționar de valențe, atât în versiune pe hârtie, cât și în format electronic” (*ibidem*).

Echipă

Pentru prima etapă, echipa a fost formată din „lingviști, informaticieni și studenți de la Universitatea din București, de la Academia Română, precum și din domeniul industrial” (*ibidem*). Pentru cea de a doua etapă, echipa a fost formată doar din dr. Ana-Maria Barbu (Barbu 2018: 412).

Finanțare

Pentru prima etapă de dezvoltare, finanțarea a fost asigurată de Consiliul Național al Cercetării Științifice din Învățământul Superior (CNCSIS), prin grantul nr. 1156/2005, denumit *Bază de date sintactico-semantică în format XML: valenţele combinatorii ale verbelor româneşti în reprezentare HPSG* (*ibidem*: 411). În cea de a doua etapă nu a existat finanțare (*ibidem*: 412).

Date de intrare

Pentru ambele etape de dezvoltare, datele de intrare au fost preluate din publicații online (*ibidem*: 411).

Metodologie de lucru

Prima etapă a implicat faptul că „descrierea valențelor a fost limitată în principal la complementele obligatorii (valențele minimale) și la sensurile verbelor” (*ibidem*). În cadrul celei de-a doua etape, s-a „acordat o atenție specială complementelor facultative sau opționale și elementelor similare cu adjuncții (sau modificatorii)” (*ibidem*: 412).

Model pentru date

Pentru ambele etape de cercetare, modelul pentru date este următorul: forma de bază; formule sintactice; sensuri (pentru fiecare sens sunt date: definiție; citate); îmbinări de cuvinte (*ibidem*: 414).

Rezultate obținute

Lexiconul cuprinde „628 de verbe românești pentru care au fost create în mod manual 2.372 de grile de subcategorizare și au fost definite 2.476 de sensuri sau subsensuri” (Barbu *et al.* 2022: 3). Lexiconul este disponibil la adresa web http://dcv.lingv.ro/.

3.14. Lexiconul GRAALAN (2007–2013, 2014–2017)

Descriere

Acest lexicon a fost elaborat în cadrul a trei proiecte de cercetare succesive, și anume: *Pachet de aplicații lingvistice pentru limba română* (PALIROM), care s-a desfășurat în perioada 2007–2010, *Dezambiguizarea sensurilor cuvintelor folosind reţele lexicon* (SenDiS), realizat în perioada 2010–2013, și *Analiza fonetică a limbii române: studiu şi aplicaţii informatice* (AFLR), desfășurat în perioada 2014–2017.

Obiective

Proiectul PALIROM a avut următoarele obiective: crearea unui limbaj de programare necesar pentru codificarea informațiilor lingvistice (alfabet, silabificare, morfologie, flexiune, lexicon, sintaxă), numit GRAALAN (Grammar Abstract Language); elaborarea unui generator de forme flexionare; elaborarea unui automat pentru despărțirea în silabe; crearea unui lexicon. De asemenea, în cadrul proiectului PALIROM se propunea ca resursele digitale elaborate să fie folosite pentru obținerea de analizatoare morfologice, verificatoare gramaticale, generatoare de forme flexionare, aplicații pentru indexare / căutare, lematizatoare, corectoare ortografice, automate pentru despărțirea în silabe, lexicoane, traducerea asistată de calculator (Diaconescu, Dumitrașcu 2007: 228, 237).

Scopul proiectului SenDIS a fost „concepţia, proiectarea şi implementarea unui sistem de dezambiguizare cu caracter general (i.e. utilizabil pentru orice limbă naturală)", care să fie integrat „într-un set de aplicaţii de prelucrare a limbajului natural (Natural Language Processing – NLP), dintre care un rol important îl va avea sistemul de traducere automată" (SenDIS 2017).

Obiectivele proiectului AFLR au fost: actualizarea lexiconului GRAALAN, prin adăugarea, cu ajutorul unui produs informatic dezvoltat tot în cadrul proiectului, a transcrierilor fonetice, atât pentru formele de bază, cât şi pentru formele flexionare; generarea „unui dicţionar fonetic al tuturor silabelor limbii române şi al cuvintelor ce acoperă aceste silabe"; „dezvoltarea unei aplicaţii de recunoaştere a vorbirii pentru limba română bazată pe analiza fonetică a silabelor cuvintelor" (AFLR 2017).

Echipă

Pentru proiectul PALIROM, echipa a fost constituită din angajaţi ai următoarelor companii şi instituţii: SC SOFTWIN SRL din București, Universitatea „Politehnica" din București – Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei, Universitatea „Politehnica" din București – Facultatea de Automatică şi Calculatoare, Institutul de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București, Universitatea din Bucureşti – Facultatea de Litere și Institutul de Matematică „Simion Stoilow" al Academiei Române din București (PALIROM 2010). Echipa a fost coordonată de către ing. Ştefan Stelian Diaconescu.

Pentru proiectul SenDIS, membrii echipei au fost de la Universitatea din Bucureşti – Facultatea de Matematică și Informatică, SC SOFTWIN SRL din București, Institutul de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București, Universitatea din Bucureşti, Institutul de Matematică „Simion Stoilow" al Academiei Române din București și Universitatea „Politehnica" din București – Facultatea de Automatică şi Calculatoare (SenDIS 2017). Echipa a fost coordonată de către prof. dr. Alin Ştefănescu.

Pentru proiectul AFLR, membrii echipei au fost de la SC SOFTWIN SRL din București, Universitatea „Politehnica" din București – Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei și Institutul de Lingvistică „Iorgu Iordan – Alexandru Rosetti" din București (AFLR 2017). Echipa a fost coordonată de ing. Ştefan Stelian Diaconescu.

Finanțare

Pentru proiectul PALIROM, finanțarea a fost asigurată prin grantul nr. 54/26.09.2007, acordat prin Planul Național de Cercetare, Dezvoltare și Inovare (PNCDI). Proiectul SenDIS a fost finanțat prin grantul nr. 207/20.07.2010, în valoare de 2.175.000 de lei, acordat de Fondul European de Dezvoltare Regională (SenDIS 2017). Proiectul AFLR a fost finanţat prin grantul cu codul PN-II-PT-PCCA-2013-4-1451, în valoare de 1.559.405 lei, acordat de Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării (UEFISCDI).

Date de intrare

Pentru proiectul PALIROM, au fost intrările din *Dicționarul explicativ al limbii române*, ediția din anul 1998 și, parțial, intrările din *Micul dicționar academic*, respectiv din *Dicţionarul tezaur al limbii române* (Holban *et al*. 2014: 128). Pentru proiectul SenDIS, datele de intrare au fost datele de ieșire din proiectul PALIROM (SenDIS 2017). În sfârşit, pentru proiectul AFLR, datele de intrare au fost datele de ieșire din proiectul SenDIS (AFLR 2017).

Metodologie de lucru

În cadrul proiectului PALIROM, au fost parcurse următoarele etape de lucru: elaborarea limbajului GRAALAN și a unui interpretor pentru acest limbaj, proiectarea bazelor de date și a interfețelor de acces la acestea, introducerea datelor, realizarea prototipurilor produselor informatice asociate: verificator ortografic și verificator gramatical (PALIROM 2010).

Pentru proiectul SenDIS, pornind de la datele rezultate din proiectul PALIROM, „textul definițiilor a fost actualizat în conformitate cu regulile morfologice, ortoepice și ortografice din *Dicționarul ortografic, ortoepic și morfologic al limbii române*, ediția a II-a” (Holban *et al*. 2014: 128) și a fost realizat produsul informatic pentru dezambiguizare (SenDIS 2017). Apoi, fiecare cuvânt din fiecare definiție a fost adnotat (Holban *et al*. 2014: 130).

În cadrul proiectului PALIROM, au fost actualizate cuvintele-titlu din lexiconul GRAALAN, au fost realizate transcrierile fonetice pentru formele de bază și pentru formele flexionare din lexicon și au fost extrase dicționarele menționate ca obiective ale proiectului (AFLR 2017).

Model pentru date

Conform informațiilor din Diaconescu 2006: 2, Diaconescu, Dumitrașcu 2007: 234 și Diaconescu *et al*. 2009: 102, lexiconul cuprinde următoarele categorii de intrări: morfeme (rădăcini, prefixe, sufixe, prefixoide, sufixoide etc.); leme și forme flexionare; unități frazeologice, cărora li se indică și structura sub forma unui arbore de dependențe.

Fiecare intrare a fost proiectată să conțină următoarele informații: cuvânt-titlu; transcrierea fonetică a cuvântului-titlu; informații privind despărțirea în silabe a cuvântului-titlu (despărțirea în silabe a cuvântului-titlu scris în alfabet obișnuit, despărțirea în silabe a cuvântului-titlu scris în alfabet fonetic și despărțirea în silabe a cuvântului-titlu scris în alfabet fonetic, dar cu unele restricții legate de structura morfematică a cuvântului-titlu); arborele de dependențe pentru unitățile

frazeologice; regula de flexionare, în formă codificată; informații semantice (sinonime, antonime, paronime, hiperonime, hiponime, meronime, omonime, conotații), etimologice (limba originară, forma originară, transliterarea formei originare în alfabetul limbii curente), morfologice. Pentru același cuvânt-titlu, intrările din lexicon sunt separate conform cu categoria gramaticală, cu „genul, tranzitivitatea, reflexivitatea etc." (Holban *et al.* 2014: 129).

În cadrul proiectului SenDIS, au fost adăugate semnături semantice pentru fiecare sens al fiecărei intrări din lexicon (Mincă, Diaconescu 2013: 444).

În cadrul proiectului AFLR, au fost adăugate transcrierile fonetice ale formelor de bază și ale formelor flexionare din lexicon (AFLR 2017).

Rezultate obținute

Pentru proiectul PALIROM, lexiconul GRAALAN conține circa 76.000 de leme, 115.000 de sensuri, 103.000 de forme flexionare și 12.700 de unități frazeologice. Automatul pentru despărțirea în silabe conține 723 de reguli pentru despărțirea în silabe a cuvântului-titlu scris cu alfabet normal și același număr de reguli pentru despărțirea în silabe a cuvântului-titlu scris cu alfabet fonetic. A fost produs și un lexicon de forme flexionare, care conține 9.946.686 de forme flexionare (Diaconescu *et al.* 2009: 102–104).

Pentru proiectul SenDIS, s-a obținut un lexicon adnotat semantic, cu 130.000 de sensuri interconectate de peste 600.000 de relații semantice, precum și un automat pentru adnotare semantică. S-a elaborat un set de reguli specifice de adnotare a sensurilor gloselor românești (Holban *et al.* 2014: 131). Pentru realizarea scopului proiectului, dezambiguizarea sensurilor, au fost calculate semnături semantice pentru fiecare sens al fiecărei intrări din lexicon (Mincă, Diaconescu 2013: 444).

În cadrul proiectului AFLR, au fost publicate lucrările *Dicționarul morfologic și fonetic al limbii române* (ISBN 978-1514315125), respectiv *Fonetica limbii române* (ISBN 978-1514315422), a fost îmbunătățit lexiconul GRAALAN și a fost realizat un automat pentru recunoaşterea vorbirii (AFLR 2017).

După încheierea celor trei proiecte, lexiconul GRAALAN număra circa 100.000 de leme și peste 13.700.000 de forme flexionare (AFLR 2017).

3.15. Lexicon UaicPosTagger (2009–2016)

Descriere

Acest lexicon face parte dintr-un set de instrumente pentru procesarea limbajelor naturale, numit *UaicNlpToolkit*.

Obiective

Obiectivul proiectului a fost dezvoltarea unui set de instrumente pentru administrare și procesare de corpusuri, prin segmentare, lematizare și adnotare morfosintactică.

Echipă

Activitatea în cadrul proiectului a fost desfășurată de dr. Radu V. Simionescu, în cursul studiilor sale masterale, respectiv doctorale.

Finanțare

Finanțarea a fost asigurată, după cum am arătat mai sus, de la bugetul de stat, în cadrul studiilor universitare efectuate de Radu V. Simionescu.

Date de intrare

Sunt reprezentate de intrările din mai multe dicționare în format digital, la care s-au adăugat 100.000 de nume proprii de persoane, orașe, companii și țări (Simionescu 2011: 26).

Metodologie de lucru

Datele de intrare au fost corectate manual, apoi au fost adăugate 100.000 de nume proprii (*ibidem*).

Model pentru date

Conform informațiilor din Simionescu 2011: 26, modelul este următorul: formă flexionară; informații morfologice; formă de bază.

Rezultate obținute

Un lexicon de forme flexionare, care cuprinde 1.925.022 de forme flexionare, disponibil la adresa web https://github.com/radsimu/UaicNlpToolkit/blob/master/Resource Data/unzip_me_here.zip.

3.16. Lexicon chirilic român (2013–2021)

Descriere

Lexicon chirilic român (LexCYR) este un proiect de generare a unui lexicon românesc scris cu alfabet chirilic, prin transliterare automată, pornind de la un lexicon românesc scris cu alfabet latin.

Obiective

Este destinat a fi folosit pentru digitizarea și transliterarea textelor cu alfabet chirilic publicate în Republica Moldova în perioada 1967–1989 (Ciubotaru *et al.* 2019: 310).

Echipă

Proiectul de lexicon a fost elaborat în cadrul Institutului de matematică și știința computerelor al Academiei de Ştiințe din Republica Moldova, de către o echipă coordonată de dr. Constantin Ciubotaru.

Finanțare

Finanțarea studiilor pentru elaborarea algoritmilor de transliterare și a testelor de transliterare a fost asigurată prin includerea proiectului în planul de cercetare al Institutului de Matematică și Știința Computerelor al Academiei de Științe din Republica Moldova.

Date de intrare

Pentru teste au fost folosite, cu unele adaptări, datele din Lexiconul român (LexROM) elaborat de Universitatea „Alexandru Ioan Cuza” din Iași (Ciubotaru 2021: 137). LexROM conține 1.096.674 cuvinte-titlu (Ciubotaru *et al.* 2019: 314).

Metodologie de lucru

În cadrul testelor, datele de intrare (articolele de la litera C din lexiconul de intrare) au fost filtrate, eliminându-se numele proprii și cuvintele de origine străină, iar ortografia tuturor cuvintelor-titlu și a tuturor formelor flexionare a fost actualizată prin folosirea lui „â”, conform cu normele Academiei Române (Ciubotaru 2021: 137). S-a aplicat, la datele de intrare, un algoritm de transliterare din alfabet chirilic (*ibidem*: 138). Rezultatele obținute trebuie verificate de specialiști, iar algoritmul de transliterare trebuie modificat, după caz.

Model pentru date

Modelul datelor din LexCYR este următorul (*ibidem*: 141): formă flexionară (transliterată din alfabet chirilic); indicații gramaticale; formă de bază (transliterată din alfabet chirilic).

Rezultate obținute

În urma activității de cercetare s-au obținut un algoritm pentru transliterarea cuvintelor românești din alfabet latin în alfabet chirilic (Ciubotaru *et al.* 2019: 311) și un algoritm pentru transliterarea cuvintelor românești din alfabet latin în alfabet chirilic (*ibidem*: 312).

3.17. Lexicon MaRePhoR (2014–2017)

Descriere

Lexiconul Machine-Readable Phonetic Dictionary for Romanian (MaRePhoR) este un lexicon fonetic pentru limba română, cu acces liber (Toma *et al.* 2017: 1).

Obiective

Acest lexicon este destinat pentru a servi ca instrument în cadrul activităților de cercetare în domeniile sinteză vocală sau recunoaștere vocală (*ibidem*).

Echipă

A fost formată din cercetători de la Academia Tehnică Militară din București și Universitatea Tehnică din Cluj-Napoca.

Finanțare

A fost asigurată parțial prin grantul PN-II-PT-PCCA-2013-4/2014, acordat de Ministerul Educației și Cercetării.

Date de intrare

Sunt cuvintele-titlu din lista oficială de cuvinte a Federației Române de Scrabble și cele 15.517 cuvinte-titlu ale unui lexicon elaborat în cadrul Academiei Tehnice Militare în anul 2013.

Metodologie de lucru

S-a realizat transcrierea fonetică în mod automat, apoi aceasta a fost verificată și corectată manual de patru specialiști (*ibidem*: 1–2).

Model pentru date

Pentru notarea cuvintelor-titlu în baza de date, s-a ales convenția de scriere cu minuscule a literelor cu diacritice și cu majuscule a literelor fără diacritice (*ibidem*).

Modelul este următorul: cuvânt-titlu; forma despărțită în silabe a cuvântului-titlu; forma accentuată a cuvântului-titlu; transcrierea fonetică a cuvântului-titlu, folosind codificarea SAMPA[20].

Rezultate obținute

S-a elaborat un lexicon care conține 72.375 cuvinte-titlu, cu transcrierea fonetică a acestora.

3.18. Lexicon MARCELL (2018–2020)

Descriere

Acest lexicon a fost dezvoltat în cadrul proiectului *Multilingual Resources for CEF.AT in Legal Domain* (MARCELL).

Obiective

Obiectivele acestui proiect au fost: procesarea tezaurului multilingual bazat pe ontologie EUROVOC[21] (EU Vocabularies, bază de date multilingvală și multidisciplinară a UE) și procesarea corpusului de legi în limbile respective, în vederea asigurării de resurse digitale pentru traducere automată pentru CEF.AT[22].

Echipă

Echipa a fost constituită din specialiști de la Institutul de Cercetări Lingvistice al Academiei Maghiare de Științe, Institutul pentru Limba Bulgară „Prof. Lyubomir Andreychin” (Bulgaria), Facultatea de Științe Umaniste și Sociale a Universității din Zagreb (Croația), Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu”, București, Institutul de Lingvistică al Academiei Slovace de Științe și Institutul „Jožef Stefan” din Liubliana (Slovenia). Coordonarea a fost asigurată de Institutul de Cercetări Lingvistice al Academiei Maghiare de Științe.

Finanțare

Finanțarea a fost asigurată printr-un grant în valoare de 1.883.715 euro, acordat de EU – CEF (Connecting Europe Facility).

Date de intrare

Peste 144.000 de texte legislative, emise începând cu anul 1881 (Tufiș *et al.* 2020: 2773).

Metodologie de lucru

S-au parcurs următoarele etape de lucru: elaborarea unui corpus monolingual național de texte legislative, segmentat și adnotat morfosintactic, pentru fiecare dintre cele șapte limbi incluse în proiect; s-au identificat termenii IATE[23] (Interactive Terminology for Europe) și EUROVOC în texte și au fost clasificate textele în conformitate cu cele 21 de domenii principale din EUROVOC (*ibidem*: 2773).

[20] https://phon.ucl.ac.uk/home/sampa/romanian.htm
[21] https://eur-lex.europa.eu/browse/eurovoc.html?locale=ro
[22] https://cef-at-service-catalogue.eu/
[23] https://iate.europa.eu

Model pentru date

Modelul este următorul (*ibidem*: 2776): identificator; formă flexionară; formă de bază; categorie gramaticală; informații morfologice; cap de grup sintagmatic; relație cu capul de grup sintagmatic; identificator IATE; identificator EUROVOC.

Rezultate obținute

S-a obținut un corpus cu 163 274 de texte adnotate, disponibil la adresa web https://live.european-language-grid.eu/catalogue/corpus/19464/download/. Forma de adnotare permite o extragere facilă a unui lexicon de forme flexionare.

3.19. Lexicon RoLEX (2018–2022)

Descriere

Acest lexicon a fost dezvoltat în cadrul proiectului complex ReTeRom, format din patru proiecte.

Obiective

Acest proiect a avut următoarele obiective: „crearea unui corpus bimodal pentru limba română adnotat pe multiple niveluri (COBILIRO)"; elaborarea unui set de „tehnologii pentru procesarea limbajului natural – text (TEPROLIN)"; elaborarea unui set de „tehnologii pentru adnotarea automată a datelor audio şi pentru realizarea interfeţelor de recunoaştere automată a vorbirii (TADARAV)"; elaborarea unui set de „tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate (SINTERO)" (ReTeRom 2022).

Echipă

A fost formată din cercetători de la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București, Universitatea Tehnică din Cluj-Napoca, Universitatea „Politehnica" din Bucureşti și Universitatea „Alexandru Ioan Cuza" din Iași. Echipa a fost coordonată de dr. Dan Tufiș.

Finanțare

A fost asigurată prin grantul PN-III-P1-1.2-PCCDI-2017-0818 – 73/2018, acordat de Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării (UEFISCDI).

Date de intrare

S-au folosit transcrieri ale unor înregistrări audio, precum și lexiconul TBL dezvoltat în cadrul Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București. Alte informații lexicale (silabizare, accent, transcriere fonetică) au fost parțial preluate din RoSyllabiDict (Barbu 2008) și MaRePhor (Toma *et al.* 2017) și au fost calculate parțial în mod automat.

Metodologie de lucru

S-au corectat, actualizat și îmbunătățit datele de intrare, atât în mod automat, cât și manual. Datele de intrare au fost apoi agregate (ReTeRom 2022).

Model pentru date

Modelul este următorul: formă flexionară; formă de bază; informații gramaticale; forma despărțită în silabe; forma transcrisă fonetic, folosind codificarea SAMPA (*ibidem*).

Rezultate obținute

A fost elaborat un lexicon de forme flexionare cu 330.866 de intrări (*ibidem*), disponibil la adresa web https://racai.ro/p/reterom/rapoarte/lexicon_reterom_final.zip.

4. SINTEZĂ A REZULTATELOR OBȚINUTE

Dintre cele 19 proiecte analizate, un număr de 15 au produs lexicoane. Pe lângă acestea, au mai fost dezvoltate următoarele tipuri de produse informatice:

– descriere formală a sintagmelor limbii române (vezi 3.10);
– descriere formală a fonologiei limbii române (vezi 3.10);
– descriere formală a regulilor de silabație (vezi 3.10);
– descriere formală a regulilor de flexionare (vezi 3.10);
– algoritm pentru flexionare (vezi 3.6);
– algoritm pentru transcriere în și din alfabet chirilic în alfabet latin (vezi 3.16);
– algoritm pentru adnotare semantică (vezi 3.14);
– algoritm pentru despărțire în silabe (vezi 3.10);
– automat pentru flexionare (vezi 3.1, 3.6, 3.7, 3.10);
– automat pentru lematizare (vezi 3.1, 3.7);
– automat pentru derivări lexicale (vezi 3.6, 3.7);
– corector ortografic (vezi 3.7, 3.10);
– automat pentru analize morfologice (vezi 3.8, 3.10);
– automat pentru analize sintactice (vezi 3.8, 3.10);
– automat pentru analize sintagmatice (vezi 3.10);
– automat pentru analize semantice (vezi 3.14);
– automat pentru analize fonetice (vezi 3.10);
– automat pentru generarea de texte (vezi 3.8);
– automat pentru despărțire în silabe (vezi 3.10, 3.14);
– automat pentru recunoaşterea vorbirii (vezi 3.14);
– extractor de termeni pentru terminologii (vezi 3.10).

5. CONCLUZII

Se observă o producție destul de bogată de lexicoane, care arată interesul crescut pentru asigurarea de instrumente (digitale) pentru sistematizarea cunoașterii privind limba română. De asemenea, există o producție crescută de produse informatice adiționale, de o complexitate destul de crescută, realizate, în cea mai mare parte, conform unor standarde cu largă răspândire pe plan internațional.

Din punctul de vedere al metodologiei de lucru, se observă că proiectele se pot împărți în două categorii: proiecte cu metodologie informatică și proiecte cu metodologie lingvistică. Astfel, în cazul automatelor pentru flexionare, doar trei dintre proiecte (vezi 3.6, 3.12, 3.14) conțin algoritmi și liste de terminații, arătând astfel o abordare lingvistică asistată de calculator, pe când alte cinci (vezi 3.7, 3.9, 3.15, 3.18, 3.19) conțin liste de forme flexionare, în unele cazuri în număr de milioane, aceasta denotând o abordare informatică.

Această subtilă divergență se manifestă, de asemenea, în privința terminologiei și a finalității temelor de cercetare.

BIBLIOGRAFIE

AFLR 2017 = AFLR. *Analiza fonetică a limbii române: studiu şi aplicaţii informatice* (http://softwinresearch.ro/index.php/ro/proiecte/aflr, accesat la data de 29.07.2022).

Barbu *et al.* 2006 = Ana-Maria Barbu, Emil Ionescu, Verginica Barbu Mititelu, *Romanian Valence Dictionary in XML Format*, în *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, mai 2006, Genova, Italia, p. 2467–2470 http://www.lrec-conf.org/proceedings/lrec2006/pdf/295_pdf.pdf, accesat la data de 07.08.2022).

Barbu, Ionescu 2008 = Ana-Maria Barbu, Emil Ionescu, *Proiectarea unui dicționar de valențe verbale destinat prelucrării limbajului natural*, în „Studii și cercetări lingvistice", LIX, nr. 2, p. 295–306.

Barbu 2008 = Ana-Maria Barbu, *Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries*, în *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, mai 2008, Marrakech, Maroc, p. 1937–1941 (http://www.lrec-conf.org/proceedings/lrec2008/pdf/495_paper.pdf, accesat la data de 07.05.2023).

Barbu 2018 = Ana-Maria Barbu, *Dictionary of Verbal Contexts for the Romanian Language*, în *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Ljubljana University Press, Facultatea de Arte, p. 411–422 (https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2912-1-10-20180820.pdf, accesat la data de 07.08.2022).

Barbu *et al*. 2022 = Ana-Maria Barbu, Verginica Barbu Mititelu, Cătălin Mititelu, *Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs*, în *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, 20–25 iunie 2022, Marsilia, Franța (https://aclanthology.org/2022.lrec-1.714, accesat la data de 07.05.2023).

Bidu-Vrănceanu *et al.* 2005 = Angela Bidu-Vrănceanu, Cristina Călărașu, Liliana Ionescu-Ruxăndoiu, Mihaela Mancaș, Gabriela Pană Dindelegan, *Dicționar de științe ale limbii*, ediția a doua, București, Nemira & Co.

Bocșa 1974a = Minerva Bocșa, *Technique: Letters with Variable Values and the Mechanical Inflection of Rumanian Words*, în „American Journal of Computational Linguistics", decembrie, p. 38–52 (https://aclanthology.org/J74-3003.pdf, accesat la data de 13.08.2022).

Bocșa 1974b = Minerva Bocșa, *Un algoritm pentru generarea cuvintelor limbii române*, în „Revista de analiză numerică și teoria aproximației", vol. 3, fascicula 1, p. 35–51 (https://ictp.acad.ro/ranta-ro-72-74/journal/article/view/40/40, accesat la data de 13.08.2022).

Boian *et al.* 2011 = Elena Boian, Constantin Ciubotaru, Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Malahov, Mircea Petic, *Creation and Development of the Romanian Lexical Resources*, în *Proceedings of Recent Advances in Natural Language Processing*, RANLP 12–14 septembrie 2011, Hissar, Bulgaria, p. 678–685.

Cherata *et al.* 1994–1995 = Sanda Cherata, Teodor Vușcan, Emma Tămâianu, *SILEX – Un sistem lexico-morfologic computerizat pentru analiza textelor românești*, în „Dacoromania" (serie nouă), I, p. 201–212.

Cherata *et al.* 1996–1997 = Sanda Cherata, Teodor Vușcan, Emma Tămâianu, *Silex – Funcțiile de lematizare și de generare a paradigmelor*, în „Dacoromania" (serie nouă), II, p. 273–285.

Cherata, Mihăescu 2008 = Sanda Cherata, Manuela Mihăescu, *Modele formale de reprezentare a informațiilor lexicale și terminologice în proiectul* SIASTRO, în „Dacoromania" (serie nouă), XIII, nr. 2, p. 151–169.

Ciubotaru *et al.* 2019 = Constantin Ciubotaru, Valentina Demidova, Tudor Bumbu, *Generation of the Romanian Cyrillic lexicon for the period 1967 – 1989,* în P*roceedings of the Fifth Conference of Mathematical Society of Moldova IMCS-55,* 28 septembrie – 1 octombrie 2019, Chișinău, Republica Moldova, p. 309–316 (https://ibn.idsi.md/sites/default/files/imag_file/309-316_1.pdf, accesat la data de 13.07.2022).

Ciubotaru 2021 = Constantin Ciubotaru, *Backtracking algorithm for lexicon generation*, în „Computer Science Journal Of Moldova", vol. 29, nr. 1 (85), p. 135–152 (http://math.md/files/csjm/v29-n1/v29-n1-(pp135–152).pdf, accesat la data de 13.07.2022).

Cojocaru 1996 = Svetlana Cojocaru, *Lexicon român: instrumentar, implementare, utilizare,* în Dan Tufiș (editor), *Limbaj și tehnologie*, București, Editura Academiei Române, p. 37–39.

Curteanu *et al.* 1996 = Neculai Curteanu, G. Holban, S. Negulescu, C. Păvăloi, I. Păvăloi, A. Todirașcu, *Bază de cunoștințe ligvistice pentru limba română*, în Dan Tufiș (editor), *Limbaj și tehnologie*, București, Editura Academiei Române, p. 101–108.

Dănăilă, Michăescu 1980 = Ion Dănăilă, Radu Michăescu, *Banca de date fono-morfo-semantice a limbii române (BANDASEM)*, în „Limba română", vol. XXIX, nr. 5, p. 471–476.

Dănăilă 1993 = Ion Dănăilă, *Pentru un inventar general al lexicului limbii române,* în „Limba română", vol. XLII, nr. 2, p. 61–68.

Dănăilă 1996 = Ion Dănăilă, *Primul proiect românesc de dicționar lingvistic computerizat. Dicționarul confruntativ de sinonime, de analogii și de asociații al limbii române (DCSAAs)*, în Dan Tufiș (editor), *Limbaj și tehnologie*, București, Editura Academiei Române, p. 177–179.

Diaconescu 2006 = Ștefan Diaconescu, *Crearea resurselor lingvistice cu ajutorul unui limbaj specializat,* în *Workshop on Linguistic resources and tools for Romanian language processing*, Iași, Editura Universității „Alexandru Ioan Cuza".

Diaconescu, Dumitrașcu 2007 = Ștefan Diaconescu, Ionuț Dumitrașcu. *Complex Natural Language Processing System Architecture*, în Corneliu Burileanu, Horia-Nicolai Teodorescu (ed.), *Advances in Spoken Language Technology*, București, Editura Academiei Române, p. 228-240.

Diaconescu *et al.* 2009 = Ștefan Diaconescu, Cristi Ingineru, Felicia Codîrlașu, Monica Rizea, Oana Bulibașa, *General System for Normal and Phonetic Inflection*, în Corneliu Burileanu, Horia-Nicolai Teodorescu (ed.), *From Speech Processing to Spoken Language Technology. Proceedings of the 5th Conference on Speech Technology and Human-Computer Dialogue "SpeD 2009"*, București, Editura Academiei Române, p. 98–107.

Holban *et al.* 2014 = Corina Holban, Felicia Codîrlașu, Andrei Mincă, Ștefan Diaconescu, *Sense-tagging of Romanian glosses*, în Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiș (ed.), *Proceedings of the 10th International*

Conference on Linguistic Resources and Technologies for Romanian Language (ConsILR 2014), Iași, Editura Universității „Alexandru Ioan Cuza”, p. 125–132.

Mincă, Diaconescu 2013 = Andrei Mincă, Ștefan Diaconescu, *An Approach to Reduce Part of Speech Ambiguity Using Semantically Annotated Lexicon Definitions*, în *Proceedings of the 2nd International Conference on Management Science and Industrial Engineering (MSIE 2013*, noiembrie 2013 (https://www.atlantis-press.com/proceedings/msie-13/9764, accesat la data de 08.06.2023; DOI: 10.2991/msie-13.2013.94).

Mititelu *et al.* 2014 = Verginica Barbu Mititelu, Ștefan Daniel Dumitrescu, Dan Tufiș, *News about the Romanian Wordnet*, în *Proceedings of the Seventh Global Wordnet Conference*, ianuarie 2014, Tartu, Estonia, p. 268–275 (https://aclanthology.org/W14-0137.pdf, accesat la data de 06.08.2022).

PALIROM 2010 = *PALIROM. Pachet de aplicații lingvistice pentru limba română* (http://softwinresearch.ro/index.php/ro/proiecte/palirom, accesat la data de 28.07.2022).

Peev *et al.* 1997 = Luciana Peev, Lidia Bibolar, Jodal Endre, *A Formalization Model of the Romanian Morphology*, în Dan Tufiș, Poul Andresen (ed.), *Recent Advances in Romanian Language Technology*, București, Editura Academiei Române, p. 72–77.

Peev, Șerban 2008 = Luciana Peev, Felicia Șerban, *Metode de analiză lingvistică a textelor în limba română pentru extragerea terminologică. Instrumente şi resurse*, în *Actele Seminarului Internaţional „Instrumente pentru asistarea traducerii”*, 28–29 februarie 2008, București (https://unilat.org/Library/Handlers/File.ashx?id=c5da92d6-ef18-4e7d-b853-0769a17a4f0c, accesat la data de 13.08.2022).

Petic 2014 = Mircea Petic, *Tendinţe actuale în procesarea limbajului natural pentru limba română*, în Mircea Petic, *Tradiţie şi inovare în cercetarea ştiinţifică*, ediția a 3-a, Materialele Colloquia Professorum, 12 octombrie 2012, Bălți, Editura Universității de Stat „Alecu Russo”, p. 245–249.

ReTeRom 2022 = *ReTeRom. Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română* (https://racai.ro/p/reterom, accesat la data de 30.07.2022).

RoLingva 2009 = *RoLingva. Colectiv RoLingva – SC Software ITC Cluj SA, Cluj-Napoca* (https://web.archive.org/web/20090227061255/http://rolingva.ro/rolingva.php, accesat la data de 12.08.2022).

RoWordNet 2013 = *RoWordNet. Romanian WordNet* (http://ws.racai.ro:9191/narratives/batch1/Ro WordNet.pdf, accesat la data de 06.08.2022).

SenDIS 2017 = *SenDIS. Dezambiguizarea sensurilor cuvintelor folosind reţele lexicon* (http://softwinresearch.ro/index.php/ro/proiecte/sendis, accesat la data de 28.07.2022).

Shieber *et al.* 1983 = Stuart Shieber, Hans Uszkoreit, Fernando Pereira, Jane Robinson și Mabry Tyson, *The Formalism and Implementation of PATR-II*, în Barbara J. Grosz and Mark E. Stickel (ed.), *Research on Interactive Acquisition and Use of Knowledge (Final report of SRI Project 1894)*, Menlo Park, California, SRI International Publishing House (https://dash.harvard.edu/bitstream/handle/1/23492376/formalism%20and%20implementation%20of%20patr-ii%20-%20shieber%20-%201983.pdf, accesat la data de 14.07.2022).

Simionescu 2011 = Radu Simionescu, *Hybrid POS Tagger*, în *Proceedings of the Workshop “Language Resources and Tools with Industrial Applications”*, 30–31 august 2011, Cluj-Napoca, Iași, Editura Universității „Alexandru Ioan Cuza” din Iași, p. 21–28.

Șerban *et al.* 1996 = Felicia Șerban, Luciana Peev, Lidia Bibolar, Dana Bucerzan, *Baza de date a limbii române. Fonetică și fonologie*, în Dan Tufiș (editor), *Limbaj și tehnologie*, București, Editura Academiei Române, p. 157–160.

Șerbănescu 1994 = Andra Șerbănescu, *Pentru un dicționar sintactic al verbelor românești*, în „Studii și cercetări lingvistice", XLV, nr. 3–4, București, p. 133–150.

Tămâianu-Morita *et al.* 2006–2007 = Emma Tămâianu-Morita, Sanda Cherata, Cornel Vîlcu, *Analiza sintagmatică a textelor româneşti prin mijloace informatice: proiectul* SIASTRO, în „Dacoromania" (serie nouă), XI–XII, p. 77–87.

Tănăsescu 2009 = Florin-Teodor Tănăsescu, *Remus Răduleț și terminologia tehnică*, în *Studii și comunicări*, vol. II, Cluj-Napoca, Editura Mega Print, p. 97–120 (https://studii.crifst.ro/doc/2009/2009_06.pdf, accesat la data de 11.08.2022).

Timotin, Tănăsescu 1997 = Alexandru Timotin, Florin Teodor Tănăsescu, *Structures for a Thesaurus of Technical Terminology*, în Dan Tufiș, Poul Andresen (ed.), *Recent Advances in Romanian Language Technology*, București, Editura Academiei Române, p. 105–120.

Toma *et al*. 2017 = Ștefan-Adrian Toma, Adriana Stan, Mihai-Lica Pura, Traian Bârsan, *MaRePhoR – An Open Access Machine-Readable Phonetic Dictionary for Romanian*, în *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, București (http://adrianastan.com/papers/2017_SPED_Marephor.pdf, accesat la data de 29.07.2022).

Tufiș *et al.* 1996 = Dan Tufiș, Lidia Diaconu, Călin Diaconu, Ana Maria Barbu, *Dicționar al limbii române destinat traducerii automate*, în Dan Tufiș (editor), *Limbaj și tehnologie*, București, Editura Academiei Române, p. 93–100.

Tufiș *et al.* 1997 = Dan Tufiș, Ana-Maria Barbu, *A Reversible and Reusable Morpho-Lexical Description of Romanian*, în Dan Tufiș, Poul Andresen (ed.), *Recent Advances in Romanian Language Technology*, București, Editura Academiei Române, p. 83–93.

Tufiș, Cristea 2002 = Dan Tufiș, Dan Cristea, *Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet*, în *Proceedings of The Third International Conference on Language Resources and Evaluation*, Las Palmas, Insulele Canare, Spania, 29 mai – 31 mai 2002 (https://racai.ro/media/Tufis-Cristea-LREC2002.pdf, accesat la data de 03.08.2022).

Tufiș *et al.* 2004 = Dan Tufiș, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, Luigi Bozianu, *The Romanian Wordnet*, în *Romanian Journal of Information Science and Technology*, vol. 7, nr. 1–2, 2004, p. 107–124(http://dblab.upatras.gr/balkanet/journal/12_Romanian1RJIST-RACAI1corect.pdf, accesat la data de 03.08.2022).

Tufiș *et al.* 2006 = Dan Tufiș, Verginica Barbu Mititelu, Alexandru Ceauşu, Luigi Bozianu, Cătălin Mihăilă, Margareta Manu Magda, *Noi dezvoltări ale WordNET-ului românesc*, în Corina Forăscu, Dan Tufiș, Dan Cristea (ed.), *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Iaşi, Editura Universității „Alexandru Ioan Cuza", p. 17–21.

Tufiș 2008 = Dan Tufiș. *Ro-WordNet: ontologie lexicală pentru limba română*, în „Academica", XVIII, nr. 208–209, februarie–martie,, p. 30–34.

Tufiș *et al.* 2020 = Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, Andrei Coman, *Collection and Annotation of the Romanian Legal Corpus*, în *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 11–16 May 2020, Marseille, p. 2773–2777 (http://lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.337.pdf, accesat la data de 31.07.2022).

Vintilă-Rădulescu 2002 = Ioana Vintilă-Rădulescu, *Resurse lingvistice pentru limba română elaborate la Institutul de Lingvistică „Iorgu Iordan”*, în Dan Tufiș, Florin Gh. Filip (coordonatori), *Limba română în societatea informațională – societatea cunoașterii*, București, Editura Expert, p. 21–33.

Vîlcu 2008 = Cornel Vîlcu, *Preliminarii teoretice la analiza gramaticală în proiectul SIASTRO: nivelul sintagmatic*, în „Dacoromania” (serie nouă), XIII, nr. 2, p. 117–126.

INVENTORY OF (DIGITAL) LEXICONS FOR THE ROMANIAN LANGUAGE. PERIOD 1973–2022

ABSTRACT

This article aims to present the lexicons (or resources that are similar to lexicons), which were elaborated in Romania during the period 1973–2022, in all media forms, from the written to the digital. The term *lexicon* is hereby understood as it is used within the field of generative grammar.

For each project I considered the following pieces of information: description, objectives, team, funding, input data, work methodology, data model and, respectively, the results. As to the data model, I have made terminological conversions, due to the fact that some projects, especially those realised by researchers with an Information Technology background, use a terminology that is more specific to Information Technology field than to linguistics or philology.

Keywords: *digital lexicon, Romanian language, XML, natural language processing*, linguistics.