Monica Busuioc, Dan Caragea

ROMTEXT, CORPUS ELECTRONIC DATAT ŞI ADNOTAT AL LIMBII ROMÂNE[1]

Având în vedere ediția a II-a a *Dicționarului limbii române* (DLR)[2], după actualizarea bibliografiei acestuia, am început, în 2017, elaborarea unui corpus intitulat Romtext[3], inspirat de Frantext, corpusul de referință al uneia dintre cele mai impresionante opere ale lexicografiei mondiale actuale: *Trésor de la langue française.*

Romtext urmează astfel unui proiect încheiat în urmă cu zece ani, CNR, a cărui realizare a fost condiționată de experiența şi de capacitățile informatice limitate de la acea dată[4]. Din acest proiect mai vechi, am păstrat neschimbat obiectivul. Astfel, Romtext, corpusul care urmează să conțină textele menționate în bibliografia DLR, va putea fi considerat, prin structură şi amploare, un model pentru *corpusul de referință al limbii române* pentru întreaga sa istorie (secolele XVI–XXI).

Numele a fost ales de directorul de proiect ca omagiu adus lexicografiei franceze, modelul avut în vedere de către Sextil Puşcariu şi, apoi, de către toți coordonatorii DLR. Dincolo de nume, am preluat modelul metadatelor, ideea adnotării corpusului şi, în consecință, reflexul acestora în criteriile de filtrate şi de căutare. Restul asemănărilor sunt fie inerente, fie întâmplătoare.

Limitările proiectului (500 de texte) ne obligă să ne mulțumim cu realizarea unui nucleu de corpus sau a unui corpus nuclear, cum îi spunem noi, şi care sperăm să se dezvolte în continuare, după încheierea proiectului.

Acest corpus îşi propune să răspundă următoarelor cerințe: să prezinte un flux de obținere şi tratare a textelor electronice plecând de la edițiile menționate în bibliografia DLR-ului; să prezinte un set de metadate care să permită delimitarea unor subcorpusuri de cercetare; să prezinte un model de adnotare semiautomată morfologică, sintactică şi semantică pentru facilitatea căutărilor şi filtrarea

[1] Această lucrare este finanțată printr-un grant al Autorității Naționale pentru Cercetare Ştiințifică şi Inovare, CNCS – UEFISCDI, proiect nr. PN-III-P4-ID-PCE-2016-0826.

[2] Dicționarul a fost început sub direcția lui Sextil Puşcariu, în 1906, cu sprijinul financiar al regelui Carol I. Primele tomuri şi fascicule au fost publicate între 1913 şi 1949 (literele *A, B, C, D–DE, F, I, J – LOJNIȚĂ*). După o întrerupere de câțiva ani, din 1965 şi până în 2010, sub mai mulți coordonatori, a apărut şi restul tomurilor şi al fasciculelor.

[3] *ROMTEXT, corpus electronic de texte, adnotat şi datat, al limbii române, secolele XVI–XXI*, finanțat de UEFISCDI, prin PN III-P4. Perioada de derulare: 2017–2019.

[4] *CNR – Corpus de referință al limbii române pentru constituirea de dicționare academice*, proiect finanțat de CNCSIS, București, 2007–2008.

LR, LXVIII, nr. 2, p. 179–184, București, 2019

rezultatelor; să prezinte o interfață de filtrare şi căutare pentru obținerea unor rezultate care să satisfacă atât necesitățile lexicografilor la noua ediție a DLR, cât şi cercetările altor specialişti, cercetări bazate pe corpusul lingvistic.

Am decis ca la elaborarea acestui corpus nuclear să avem permanent în vedere criteriul acoperirii şi cel al reprezentativității. Primul criteriu ne spune că acest corpus, chiar dacă este de o dimensiune redusă, trebuie să aibă acoperire istorică şi de limbaj. Au fost astfel selectate texte din toate secolele, cu preocuparea de a include cât mai multe varietăți stilistice şi o parte din repertoriile terminologice mai frecvente. Este evident că nu vom avea câte un text pentru tot ceea ce am dori să fie reprezentat, dar, în linii mari, proporțiile deduse din bibliografie au fost păstrate. Astfel, al doilea criteriu, cel al reprezentativității, ne-a impus să menținem analogia sugerată de bibliografie. Acest criteriu ne obligă să evităm includerea repetată a aceluiaşi tip de document (letopiseț, psaltire etc.) în defavoarea varietății de limbaj. Textele preferențiale pentru primele secole sunt mai degrabă cele cuprinse în antologii (precum Mareş 2016) decât cele de mari dimensiuni, să spunem.

Prin corpus generativ înțelegem un corpus care poate fi extins prin adăugiri de texte sau de microcorpusuri, întrucât acesta oferă soluțiile de tratament pentru orice tip de text, chiar dacă el nu este încă prezent în corpus. Astfel, un text de drept sugerează dezvoltarea unui microcorpus de texte de drept pentru observarea evoluției terminologice şi conceptuale a domeniului. Includerea unei opere precum *Cartea românească de învățătură* (1646) sugerează posibilitatea de includere, în viitor, a oricărui cod mai recent de drept, de exemplu, iar, pe lângă acesta, a textelor analoge, formând, laolaltă, microcorpusul de drept. Tratamentul lingvistico-informatic este asemănător, indiferent de epocă sau de domeniu. De aceea, un corpus generativ permite definirea etapelor de dezvoltare ulterioară, a rutelor de creştere, altfel spus, orientează dezvoltarea acestuia până la nivelul de saturație. Romtext va fi considerat un corpus saturat nu doar când toate lucrările din bibliografie vor fi incluse, ci atunci când toate intrările din DLR vor fi, în mod corespunzător, sprijinite pe corpus (excepție fac variantele înregistrate după alte lucrări, precum *Atlasul lingvistic* etc.). Din datele afişate online, Frantext conținea 5118 de texte de referință, în august 2017, ceea ce ne poate da o idee asupra nivelului de saturație. Totodată, un corpus generativ permite identificarea termenilor noi, a „plus-valorii" pe care un text o aduce corpusului şi, în consecință, ne permite calibrarea acestuia prin succesive analize ale „valorii" lingvistice a textelor. În sfârşit, un corpus care a permis dezvoltarea tuturor soluțiilor de tratament permite absorbția mult mai rapidă a restului textelor.

La terminarea proiectului, Romtext va conține, aşadar, cel puțin 500 de texte (80% literare, 20% neliterare), repartizate pe şase secole de limbă română scrisă. Graficul distribuțional în funcție de perioadele limbii şi cel al tipologiei textuale vor putea fi disponibilizate la încheierea proiectului, ambele cunoscând până atunci succesive rectificări.

În ceea ce priveşte fluxul de lucru, menționăm: scanarea ediţiilor, aplicarea de procedee de corectare a imaginii, de curățare a acesteia, urmată de recunoaşterea caracterelor prin lectură optică antrenată şi, apoi, de corectare a textului rezultat, cu ajutorul programului ArqCorr[5], tinzând spre fidelitate în raport cu originalul. Nu se poate vorbi de identitate între textul rezultat şi imaginea acestuia întrucât pentru corpus au fost eliminate liniuțele de despărțire în silabe la capăt de rând (cu excepția ultimului rând de pe pagină) şi a fost actualizată ortografia, conform normelor expuse în DOOM[2] (2005). Am menținut însă în corpus legătura dintre text şi imagine, permițând ca fraza să poată fi localizată de către lexicografi, dacă este necesar, pe imaginea paginii.

Fiecare text este, aşadar, un text corectat şi pe care aplicăm o serie de marcaje: izolăm textele „străine" cum ar fi un moto, un citat sau prefața semnată de o altă persoană etc. Marcăm, de asemenea, structurarea textului: titlu şi subtitlu, capitol, subcapitol etc. Astfel de separări sau izolări sunt deosebit de utile în analiza sintactică a textului şi în filtrarea căutărilor (evităm ca atunci când cineva caută lema „capitol" să obțină o listă enormă a separărilor intitulate „capitolul x" sau, dimpotrivă, putem cerceta o lemă doar într-un capitol al unei anumite opere). În Romtext vom putea localiza nu doar textele semnate de Tudor Arghezi, ci şi citatele altora din Arghezi. Astfel, un termen („nehotar") care apare într-o carte (Emil Hurezeanu, *Pe trecerea timpului,* 2015), dar care provine dintr-un citat (Tudor Arghezi, *Icoane de lemn,* 1929), nu va fi interpretat eronat de lexicograf şi nici de statistici (a se vedea paragraful următor).

Am precizat încă din titlu că Romtext este un corpus datat. Acest lucru permite ca vocabularul să poată fi, în mod implicit, datat integral. Printr-un astfel de demers, putem obține, în viitor, rezultate fiabile privind primele atestări în limbă, fapt ce poate limpezi unele interpretări etimologice (calea de pătrundere). Mai mult, putem obține o cronogramă a unui cuvânt în raport cu textele dintr-o anumită perioadă de timp, o istorie a diverselor grafii folosite (vezi mai jos conceptul de arhilemă), putem cerceta convenabil semantismul în diacronie şi separa cuvintele autorului de cuvintele incluse în citări.

Trecem acum la chestiunea adnotării. Aşa cum am precizat în titlu, toate textele care formează Romtext vor fi adnotate semiautomat, în regim asistat de către calculator. Acest lucru înseamnă antrenarea programului de adnotare pe cel puțin 6 000 de fraze. Programul va oferi sugestii, iar lexicograful va decide în cazurile ambigue sau necunoscute. Programul de adnotare va conține informații pentru un număr de peste 50 000 de cuvinte, inclusiv pentru unele variante.

Formele flexionate sunt gestionate de către un lematizator care permite controlul şi prezentarea acestora la o căutare cu lema. Lematizatorul foloseşte, de fapt, arhileme (reuniune sub aceeaşi etichetă, care reprezintă forma canonică, a

[5] Software de corectură a textelor electronice obținute prin lectură optică, dezvoltat de Arheus.ro (http://www.archeus.ro/lingvistica/main).

variantelor istorice, regionale sau chiar stilistice). Acelaşi procedeu de structurare va fi extins şi la clasele neflexibile datorită dimensiunii istorice şi geografice. De reținut că toate formele pot fi datate. Altfel spus, nu doar forma canonică, ci şi o anume formă de plural, un anumit timp verbal etc. vor fi datate.

Gramatica folosită în adnotare este gramatica tradițională a limbii române, aplicată local. Etichetele morfologice au fost preluate din DEX (2009), prin urmare clasele sunt cele indicate de acest dicționar, coroborate cu DOOM^2 (2005). Când invocăm caracterul local avem în vedere cuvântul în context, altfel spus, în propoziția gramaticală. Formele aglutinate (de exemplu, substantivele articulate cu articol hotărât), sunt înțelese ca forme flexionate ale unei singure forme canonice (substantivul). La fel, formele compuse scrise despărțit vor primi o etichetă suplimentară locală („Baia Mare” față de „baie” + „mare”).

În privința sintaxei, cuvintele vor fi adnotate prin tehnica adnotării asistate, pe baza relațiilor de dependență. Această adnotare va permite vizualizarea analizei unei propoziții din corpus sub formă de dependențe şi reprezentată grafic.

Cuvintele considerate semantic „pline” (substantive, adjective, verbe, adverbe) vor fi adnotate, ceea ce va permite regăsirea unui câmp semantic printr-o singură căutare (printr-unul dintre membrii săi). Astfel, dacă vom scrie cuvântul „pălărie”, vom putea cerceta toate cuvintele din câmpul semantic al accesoriilor pentru acoperirea capului (unele cunoscute, altele pe care poate le ignorăm). Putem stabili inclusiv o relație între o clasă morfologică (un verb oarecare, de exemplu), şi un câmp semantic.

Cele mai spinoase probleme le ridică, firește, textele româneşti vechi (secolele XVI–XVIII). Adnotarea acestor texte se face cu ajutorul glosarelor de termeni ale edițiilor din care am extras textul electronic. Va fi probabil mult mai greu de controlat etichetarea sintactică, dar este momentul să testăm limitele interacțiunii om–maşină.

În concluzie, scopul acestor adnotări pe trei niveluri: morfologic, sintactic şi semantic − o premieră în tratamentul corpusurilor de texte româneşti −, este acela de a facilita căutarea şi, mai ales, de a evita restituirea unui număr prea mare de forme cu caracteristici identice (restituire redundantă). Numărul de forme găsite în corpus trebuie să-i răspundă lexicografului din perspectiva acoperirii şi a reprezentativității lingvistice, în limitele acceptabile ale muncii la dicționar. Astfel, vom încerca să dezambiguizăm forma „vie”, forma de feminin a adjectivului „viu” ca să nu apară în răspunsuri laolaltă cu forma „(să) vie” de la verbul „a veni”.

La sfârşitul acestei intervenții, dorim să atragem atenția atât asupra nevoii de informatizare a noii ediții a DLR-ului, cât şi a tranziției textelor aflate încă în format tradițional în Romtext, apt nu numai pentru a restitui rapid şi fiabil date lingvistice, ci şi pentru a permite cercetări felurile, din diverse domenii, atât de necesare într-o societate ca a noastră, numită şi informațională.

SIGLE

DLR: *Dicţionarul limbii române*. Serie nouă. Redactori responsabili: acad. Iorgu Iordan, acad. Alexandru Graur şi acad. Ion Coteanu, 1965–2010 (din anul 2000, redactori responsabili: acad. Marius Sala şi acad. Gheorghe Mihăilă), Bucureşti, Editura Academiei Române.

DEX (2009): *Dicţionarul explicativ al limbii române*. Ediţia a II-a revăzută şi adăugită. Bucureşti, Editura Univers Enciclopedic Gold, 2009.

DOOM^2 (2005): *Dicţionar ortografic, ortoepic şi morfologic al limbii române*. Ediţia a II-a revăzută şi adăugită, Bucureşti, Editura Univers Enciclopedic, 2005.

Frantext: 5.118 texte de referinţă, 297.586.781 de termeni, secolele X–XXI (în august 2017), http://www.frantext.fr/.

BIBLIOGRAFIE

⁎ ⁎ ⁎, *Trésor de la langue française informatisée,* ATILF–CNRS & Université de Lorraine, http://atilf.atilf.fr/

Mareş Alexandru (coord.), 2016, *Crestomaţia limbii române vechi*, vol. I (1521–1631), Bucureşti, Editura Academiei Române.

ROMTEXT, AN ANNOTATED AND DATED DIGITAL CORPUS OF TEXTS OF THE ROMANIAN LANGUAGE

(Abstract)

Romtext is a dated and annotated corpus consisting of texts selected from the bibliography of the *Romanian Language Dictionary*, 16th-21st centuries, designed to support the creation of the new, computer-based edition of the thesaurus dictionary. This corpus is prepared as part of a project developed by the "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics of the Romanian Academy, project funded by the Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI), between 2017 and 2019.

Romtext will include over 500 literary and non-literary texts, obtained by optical reading, using the best editions, and corrected in computer-assisted mode. Subsequently, these texts will be annotated morphologically, syntactically and semantically, in assisted mode, in a semi-automatic process.

Romtext will have two search interfaces: one for lexicographers working on the new edition of the *Dictionary*, and another for the general public, the results being displayed in concordance mode. There will be ways of narrowing down and sorting search results based on textual metadata in case of too many results.

Romtext is, in this phase, a *nuclear corpus*. A series of mini-corpuses containing specialized texts will gravitate around it, and their treatment will be assisted, on sample texts, for all language varieties.

Romtext is also a *generative corpus*, meaning it can indicate growth paths by comparing the resulting vocabulary with the macrostructure of the *Romanian Language Dictionary.*

Considered, therefore, a nuclear and generative corpus, Romtext tends to become *the reference corpus* of the Romanian language.

Cuvinte-cheie: corpus nuclear, corpus generativ, corpus datat, corpus adnotat, corpus de referință, *Dicționarul limbii române.*

Keywords: nuclear corpus, generative corpus, dated corpus, annotated corpus, reference corpus, *Romanian Language Dictionary***.**

Institutul de Lingvistică al Academiei Române
„Iorgu Iordan – Alexandru Rosetti”, Bucureşti
monica.busuioc@yahoo.com
dccaragea@yahoo.com.br