Daniela GHELTOFAN
(Universitatea de Vest
din Timișoara)

Despre corpusurile electronice românești. *Inimă*: câteva e-ocurențe[1] sintagmatice

Abstract: (*About Romanian Electronic Corpora*. Inimă: *Some Syntagmatic e-Occurrence*) The aim of this paper is to present some of the electronic corpora of Romanian language, especially since, over the past decades, corpora have increasingly been used in linguistics studies. Many scholars from different research branches acknowledge the value of lingual national e-corpus. Using it as a tool in research, this allows for easy queries and obtains interesting results in linguistics behaviour. However, the use of Romanian electronic corpora in the academic field is very recent; because the corpora have been created less than two decades ago. For example, Contemporary Romanian Language corpus (CoRoLa) was initiated in 2014 (*cf.* Tufiș 2018). In this paper, we provide some examples to illustrate the application of Romanian electronic corpus (CoRoLa) with the keyword "inimă" (heart).

Keywords: *electronic corpus, corpus-based linguistic studies, e-occurrence*

Rezumat: În această lucrare ne propunem să prezentăm câteva dintre corpusurile electronice ale limbii române, mai ales că, în ultima vreme, acestea sunt tot mai des utilizate în studiile lingvistice. Mulți cercetători au recunoscut valoarea importantă pe care o au e-corpusurile lingvistice în investigațiile științifice. Utilizat ca metodă de analiză, corpusul permite o chestionare rapidă, precum și obținerea de rezultate importante despre un anumit profil lingvistic. Totuși, folosirea corpusurilor electronice ale limbii române în cercetare este de dată recentă, mai ales că acestea sunt create de mai puțin de două decenii. De pildă, alcătuirea Corpusului de referință al limbii române contemporane (CoRoLa) a fost inițiată în 2014 (*cf.* Tufiș 2018). În această lucrare, vom ilustra utilizarea corpusului electronic (CoRoLa) cu un exemplu-aplicație pe baza cuvântului-cheie *inimă*.

Cuvinte-cheie: *corpus electronic*, *studii lingvistice pe baza corpusului*, *e-ocurențe*

1. Digitalizarea limbajului uman

1.0. În contextul societății actuale informaționale, este vital să poți utiliza instrumentele digitale, să fii un cunoscător al cyberspațiului, să ai competențe digitale. Și domeniul lingvisticii a beneficiat de tehnologizarea informatică, de inteligența artificială prin crearea, bunăoară, a domeniului *lingvistică computațională*. Aceasta a luat naștere în anii '50-'60, în Statele Unite ale Americii, când se căutau soluții avantajoase și rapide în procesarea limbajului uman, mai ales în traducerea automată a limbilor naturale. De atunci, cu o evoluție constantă, s-au făcut pași importanți în

[1] Termenul „e-ocurențe" este creat după modelul „e-dicționare", însemnând numărul de prezențe (electronice) dintr-un corpus digital.

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

această direcție prin dezvoltarea și a altor ramuri ale lingvisticii care fac uz de achizițiile informatice: lingvistica corpusului, studiile de traducere etc.; prin crearea de baze electronice, de corpusuri lingvistice electronice, de dicționare și gramatici electronice, de motoare de traducere automată etc. Practic, lingvistica computațională înseamnă cercetarea lingvistică cu ajutorul tehnicii inteligenței artificiale, a unor modele algoritmice, structurale și formale aplicate la diverse nivele ale limbii: fonologic, morfologic, sintactic, semantic, pragmatic. Însă, acest fapt nu înseamnă că teoriile și ideile anterioare, mai ales cele ale lingvisticii generativiste și structuraliste, nu au avut o înrâurire însemnată asupra problematicilor lingvisticii computaționale (vezi infra).

1.1. Nu mai puțin important este și faptul că digitalizarea limbajului uman a dus la dezvoltarea unei industrii de tip software lingvistic.

1.2. În consecință, și cercetarea românească s-a orientat spre această zonă a lingvisticii, devenind imperios necesară și pentru limba română crearea de resurse informaționale digitale sub formă de e-corpusuri[1] ale limbii române, scrise şi vorbite, de e-dicţionare ale limbii române, de e-dicţionare bi- şi multilingve, de e-dicţionare terminologice mono- şi multilingve, de e-gramatici ale limbii române etc., prin programe naționale în acord cu reglementările internaționale. Proiectele din aceasta zonă lingvistică se pot descrie ca fiind strategice sau de importanță strategică, întrucât această digitalizare a limbii române oferă nu numai posibilități de studiere a ei, ci, mai cu seamă, de conservare a limbii noastre. De altfel, acad. Dan Tufiș precizează, într-un interviu, că în anul 2012, în META-NET – o reţea internațională dedicată consolidării societății informaționale europene multilingve, s-a publicat un studiu, elaborat de peste 200 de experți, în care se arată că limba română „este în pericol de extincție în spațiul digital", alături de alte 20 de limbi din spațiul comunitar european. În consecință, s-a afirmat necesitatea imediată de tezaurizare digitală a limbii române (vezi **2**). Printre primele încercări de acest tip, la nivel național, au loc, la sfârșitul anilor '90 începutul anilor 2000.

2. E-corpusuri lingvistice ale românei

2.0. În spațiul occidental, digitalizarea unui important volum lexical a avut loc la începutul anilor '80, la Universitatea din Princeton, sub coordonarea renumitului psiholingvist George Miller, inițiindu-se proiectul WordNet (*cf.* Tufiș *et al*. 2004b, Tufiș 2008). O urmare firească a creării acestei rețele lexicale computaționale a fost lansarea, la nivelul Uniunii Europene, în 1996, a proiectului EuroWordNet, care a cuprins 6 limbi europene de circulație internațională (engleză, franceză, germană, italiană, olandeză, spaniolă) și care, în 1999, a fost extins prin cuprinderea limbilor bască, catalană, cehă şi estoniană (EuroWordNet II) (*cf*. Tufiș 2008).

2.1. În 2001, la inițiativa lui D. Tufiș (IIA) și a lui D. Cristea (UAIC), s-a lansat proiectul BalkaNet, fiind cuprinse cinci limbi din zona balcanică: bulgara, greaca, româna, sârba şi turca, plus limba cehă al cărei wordnet era deja dezvoltat în cadrul

[1] „E-corpusuri", după modelul „e-dicționare".

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

EuroWordNet II (Tufiș & Cristea 2002, Tufiș *et al*. 2004b, Tufiș 2008). D. Tufiș (2008) precizează că „wordneturile (câte o reţea pentru fiecare limbă)" sunt colecţii de reţele semantice, „aliniate între ele prin intermediul unui index interlingual (ILI), conţinând reprezentări conceptuale ale înțelesurilor lexicalizabile în limbile ce formează ansamblul multilingv", aparținând hiperconceptului de „ontologii lexicale multilingve". Wordnetul românesc conținea, la finalul proiectului BalkaNet, 20.381 de sinseturi cu peste 36.000 de cuvinte-titlu (leme), precum şi o platformă software care conținea zeci de programe originale specializate pentru achiziționarea datelor relevante din corpusuri, pentru editarea, validarea şi corectarea structurilor semantice etc. (Tufiș *et al*. 2004b, Tufiș 2008).

2.2. Urmarea proiectului BalkaNet a fost Ro-Wordnet. O componentă unică, originală, a Ro-Wordnetului este extensia elaborată de colectivul proiectului prin care „toate cuvintele din definiţiile înţelesurilor reprezentate în WordNet au fost dezambiguizate morfo-lexical, lematizate (normalizate la forma de dicționar) şi au fost prelucrate sintactic (parsate) creându-se între ele legături de tip dependență sintactică", ceea ce a determinat digitalizarea/ prelucrarea semantică și pe axa sintagmatică, nu numai pe cea paradigmatică, și, în consecință, extensiile create de cercetătorii români au permis „detectarea contextuală a unor relaţii de tip sintagmatic între cuvinte (de pildă, între «a se căsători» şi «mire» sau «mireasă»)" (*cf*. Tufiș 2008).

2.3. La începutul anilor 2000, s-a încercat construirea de bănci de arbori sintactici (*treebank*) pentru limba română. Spre exemplu, colecția de 4042 de arbori (*i.e.*, de propoziții adnotate) din domeniul jurnalistic din cadrul proiectului RORIC-LING (*cf.* http://www.phobos.ro/roric/).

2.4 Proiectul SIASTRO (*Sistem informatic pentru analiza sintagmatică a textelor în limba română. Fundamentare teoretică şi implementare*, 2006-2008; director de proiect conf. dr. Emma Tămâianu-Morita) a avut ca obiectiv principal realizarea unui *analizor sintagmatic* care să identifice și să analizeze sintagme din textele scrise în limba română, în felul acesta continuându-se cercetările grupului RoLingva care au creat un *analizor morfologic*, în urma căruia a rezultat un dicționar morfologic electronic care acoperă lexicul din DEX. În vederea atingerii acestui obiectiv, cercetătorii au procedat la analiza modelelor existente pentru a expune informațiile lexicale şi terminologice[1]ale acestora, precum și la descrieri lexicale, morfologice şi sintactice necesare unei analizei de tip sintagmatic. Sistemul de descriere sintagmatică, ales în cadrul acestui proiect, s-a bazat pe modelul sintactic al profesorului universitar clujean D. D. Draşoveanu.

[1] De pildă, „datele lexicografice conţin date despre despărţirea în silabe – syllabification, despre structura şirului care formează intrarea – entryFormation (abreviere, acronim, cuvânt simplu, cuvânt compus, expresie, nespecificat), despre tipul expresiei – phraseType (expresie multi-cuvânt, expresie fixă, lexicalizată, colocaţie, idiom, nespecificată), autorul intrării – originator sau statutul intrării – adminStatus (intrare nouă, intrare verificată, implicită, exclusiv pentru MT, învechită, nespecificat)". (*cf*. Cherata și Mihăescu 2008, 164).

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

În teoria sa, D. D. Drașoveanu susține că limba, sub raport sincronic, dispune de trei nivele: *fonetic*, *lexical* și *gramatical* (= *sintagmic*[1]), întrucât, doar la aceste nivele, ne putem întâlni cu „fapte noi de limbă" (*cf.* Vîlcu 2008). *Sintagma* reprezintă, în cazul de față, unitatea minimală și maximală a gramaticii sale, rezultând *sintagmica flexională* și *sintagmica joncțională* (idem). Totodată, se subliniază că propoziția și fraza nu reprezintă fapte de limbă superioare sintagmei, iar „o propoziţie e doar o sintagmă în care relaţia e «purtată» de desinenţa unui verb finit (flectivul de acord verbal), iar o frază e sintagma în care relaţia e marcată de un conectiv cu regim personal-predicativ (conjuncţie, relativ etc.) sau de flectivul *să*" (idem, 121).

Viziunea lui D. D. Drașoveanu[2] simplificată, dar nu simplistă, asupra gramaticii oferă suportul teoretic ideal, după părerea cercetătorilor proiectului, în definirea și realizarea analizorului sintagmatic. Conform teoriei sale gramaticale, *morfologia* şi *sintaxa* reprezintă cele două fațete – *expresie* și *conţinut – ale uneia şi aceleiaşi realităţi gramaticale* (idem). De asemenea, gramatica este redusă la ideea centrală de *relație*. Relaţia sintagmatică dintre lexeme se sprijină pe *solidaritatea dintre un sens relaţional* şi *un relatem*, iar relatemul este *flectiv* sau *conectiv* (*cf. sintagmica flexională* și *sintagmica joncțională*) (idem).

Cercetătoarea Elena Tămâianu-Morita (2008) oferă, pe baza rezultatelor obținute de D. D. Drașoveanu, câteva schițe ale relatemelor subordonate și ale sintagmelor sub aspectul funcţionalităţii lor sintactice. Iată un exemplu edificator:

Construcţia 5. Part.-vb N2

Funcție sint.	Realizare	Exemplu
Atr.	+	*fetița premiată a venit*
NP	–	–
EPS1	+	*construcțiile au fost lăsate neterminate*
EPS2	+	*fetița vine premiată*
Compl.	–	–

Construcţia 10. D1'

Funcție sint.	Realizare	Exemplu
Atr.	+	*acordarea de premii elevilor*
NP	–	–
EPS1	–	–
EPS2	–	–
Compl.	+	C. Ind.: *s-au acordat premii elevilor* C. circ. de loc: *du-te naibii*

Trebuie să subliniem, de câte ori este nevoie, faptul că teoria profesorului clujean D. D. Drașoveanu, concepută în anii '70-'80, a reprezentat cadrul teoretic al unui proiect desfășurat în contextul tehnologizărilor actuale moderne și, prin aceasta, să

[1] Termenul *sintagmic* îi aparține lui D. D. Drașoveanu.

[2] Se regăsesc diverse afinități cu lingvistica integralistă coșeriană, care, însă, sunt rezultatul unor cercetări ale lui D. D. Drașoveanu fără ca acesta să aibă acces la ideile lui E. Coșeriu.

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

remarcăm limpezimea, rafinamentul și productivitatea teoretică ale unor demersuri analitice din secolul anterior.

2.5. Una dintre entitățile cele mai importante care este angrenată în tehnologizarea limbajului românesc este Academia Română care, prin proiectele întreprinse, își asumă realizarea primului corpus de referință pentru limba română contemporană, proiectul CoRoLa (Contemporary Romanian Language corpus[1]), inițiat în anul 2014 și desfășurat la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București (ICIA) și Institutul de Informatică Teoretică din Iași (IIT). Cu trei interfețe specializate, accesibile publicului larg, corpusul conține texte-eșantion de limbă română, scrisă și vorbită de către nativi (cu excepția unor traduceri juridice ale legislației europene), la care se adaugă un set de instrumente de prelucrare și adnotare cu privire la datele de natură lingvistică, gramaticală (morfologice, lexicografice, sintactice etc.), precum și metadate-standard (autor, data publicării, editură, genul literar al textului etc.), datând din anul 1989 și până astăzi, însumând peste 1 miliard de cuvinte, 300 de ore de înregistrări, 4 domenii (social, știință, natură, artă-cultură) cu 70 de subdomenii științifice[2] și acoperind toate stilurile funcționale (*cf.* Barbu-Mititelu *et al.* 2018). Numărul cercetătorilor angrenați în proiect este de 17. În această „campanie" de alcătuire a tezaurului limbii române contemporane, cercetătorii proiectului susțin că au primit sprijinul multor specialiști din țară, de la institutele de lingvistică şi de la universităţile din Bucureşti, Iaşi și Craiova, de la experţi în problematicile lingvistice, în alcătuirea și clasificarea corpusurilor, dar adaugă și faptul că au existat și mulţi voluntari.

Pentru diseminarea rezultatelor obținute și pentru racordarea lor la noile achiziții teoretice și tehnologice informaționale din domeniu, acești cercetători participă în mod constant la conferințele și colocvii internaționale cu această tematică.

Conform obiectivelor aplicative ale proiectului CoRoLa, resursele informaționale dispuse pe platforma proiectului sunt deosebit de utile în:

> [...] studiile lingvistice (fonologie, morfologie, lexicologie, etimologie, sintaxă, semantică, pragmatică); modelarea limbajului pentru procesarea automată a limbii române; dezvoltarea de modele de traducere; învățarea limbii; indexare și recuperare inteligentă și multi-criterială de informație textuală și orală; clasificare semantică de volume mari de date (text și audio); extragere de cunoștințe din date (text și audio);

[1] Corpus de referință al limbii române contemporane.

[2] Spre exemplu, „architecture, art history, dance, design, fashion, film, folklore, literature, music, painting and drawing, poetry, sculpture and theatre are the 13 subdomains of the arts and culture domain; environment, natural disasters, natural resources and universe are the 4 subdomains of the nature domain; administration, army, economy, education, entertainment, family, gossip, health, law, politics, religion, social events, social movements, sports, and tourism are the 15 subdomains of the society domain; archaeology, astronomy, biology, chemistry, constructions, criminalistics, engineering, ethnology, geography, geology, history, informatics, juridical sciences, linguistics, logics, mathematics, medicine, metrology, military science, oenology, pedagogy, pharmacology, philology, philosophy, physics, political sciences, psychology, religious studies and theology, sociology, standards, technics/technology are the 31 subdomains of the science domain." (*cf.* Barbu-Mititelu *et al.* 2018).

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

rezumare automată de documente; sisteme de întrebare-răspuns; recunoaștere și sinteza automată a vorbirii etc. (www.corola.racai.ro).

2.5.1. Una dintre consecințele principale ale alcătuirii acestui corpus a fost și interesul manifestat de cel mai important creator de corpusuri din lume – Institutul Limbii Germane, care a devenit partener în acest proiect, oferindu-și serviciile sale software și platforma[1] sa care reprezintă un motor de interogare extrem de specializat. Totodată, această colaborare cu institutul german a condus spre un alt proiect comun – DRuKoLA, început în 2016, al cărui obiectiv este crearea unei analize contrastive între limbile germană și română[2] (*cf.* Cosma *et al.* 2016), care se va extinde prin cuprinderea limbilor poloneză și maghiară.

2.5.2. Alte extensii ale aceluiași proiect CoRoLa sunt *ReTeRom (*Resources and technologies for developing human-machine interfaces in Romanian[3]*), în care se implică Universitatea Tehnică din Cluj, Universitatea Politehnică din Bucureşti, Institutul de Informatică Teoretică din Iaşi şi ICIA, precum și proiectul în robotică cognitivă – ROBIN*[4]*, la care participă Universitatea Politehnică din Bucureşti, Universitatea Tehnică din Cluj, Institutul de Matematică al Academiei Române, Universitatea Dunărea de Jos din Galați şi ICIA, și care presupune programarea roboților și a autovehicule autonome şi vedere artificială pentru a putea interacţiona în limba română.*

Proiectul complex ReTeRom este alcătuit din 4 subproiecte: COBILIRO (Corpusul bimodal pentru limba română), TEPROLIN (Tehnologii pentru procesarea limbajului natural – text), TADARAV (Tehnologii pentru adnotarea automată a datelor audio și pentru realizarea interfețelor de recunoaștere automată a vorbirii) și SINTERO (Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate). Rezultatul proiectului COBILIRO va consta în alcătuirea unui „corpus bimodal (care este un caz particular de corpus multimodal, la rândul lui reprezentând un tip particular de corpus)", prin care se înțelege

[...] o colecție de înregistrări orale însoțite de transcrierile lor și de metadatele corespunzătoare. Un corpus bimodal este găzduit pe o platformă specializată,

[1] CoRoLa se află pe portalul KorAP al Institutului de Limbă Germană.

[2] Obiectivele specifice proiectului DruKoLA sunt: „the detected errors will be remedied (typographical errors, missing diacritics, missing or wrong metadata, wrong annotations, etc.) then, new processed texts (written and oral) both monolingual and multilingual will be added and several exploitation facilities (user defined sub-corpora, comparative analytics, etc.)" (*cf.* Păiș și Tufiș 2018).

[3] Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română.

[4] „(…) ne propunem să dezvoltăm o serie de scenarii pentru câteva micro-lumi și tehnologia de prelucrare a limbii române pentru dialoguri situaționale în aceste micro-lumi."; „Proiectul se referă la o gamă diversă de roboți: roboți asistivi pentru sprijinul persoanelor cu nevoi speciale, roboți de interacțiune cu clienții și roboți software care pot fi instalați pe vehicule în scopul realizării unei conduceri autonome sau semi-autonome. Proiectul combină tehnici și tehnologii avansate de inteligență artificială, interacțiune om-robot, interacțiune cu un mediu pervasiv și prelucrări în Cloud."
(*cf.* http://www.racai.ro/p/robin/rapoarte/ROBIN_Raport_Etapa-I-2018.pdf).

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

împreună cu serviciile și aplicațiile web de acces, dezvoltare și întreținere ale lui, unde sunt specificați algoritmii de utilizare ai corpusului și, în unele cazuri, de unde pot fi descărcate exemple de aplicații care utilizează corpusul. Corpusurile pot conține texte scrise, înregistrări orale sau ambele modalități de redare a unei limbi naturale. În proiectul de față ne interesează ultimul caz. (*cf.* http://www.racai.ro/p/reterom/).

Iar cu privire la corpusurile orale, acelea se regăsesc sub două forme:

[...] voce-în-citire (*read speech*) (incluzând lecturi din cărți, știri, liste de cuvinte, secvențe de numere) și vorbire spontană (*spontaneous speech*) (incluzând: dialoguri între două sau mai multe persoane, narative, relatări despre trasee pe hartă, stabilirea de întâlniri, simulări „Vrăjitorul din Oz”. (idem).

Corpusurile aparținând acestor proiecte conțin peste 450 de ore de înregistrare, la care se adaugă 1871 de articole de ziare în format text.

Prin subproiectul TEPROLIN s-a creat o platformă ce conține un set de tehnologii avansate pentru procesarea limbajului natural (text) în limba română; astfel, lexiconul computațional, destinat aplicațiilor de prelucrare a vorbirii (cu peste 354.000 de intrări), colectat în COBILIRO, va conține pe lângă ocurențe, leme, descrierile morfo-lexicale, constituenții sintactici, analize sintactice cu relațiile de dependență, și informația specifică aplicațiilor de prelucrare fonematică și fonetică a vorbirii: silabificare, plasare accent, transcriere fonetică (*cf.* http://dev.racai.ro/ReTeRom/).

Următorul proiect TADARAV are ca obiectiv general dezvoltarea unui set de tehnologii avansate pentru adnotarea fonetică automată a semnalului vocal colectat în corpusul din COBILIRO, folosind modelele lingvistice generate în TEPROLIN, pentru realizarea interfețelor de recunoaștere automată a vorbirii în limba română (idem).

În final, proiectul SINTERO presupune dezvoltarea unei tehnologii avansate pentru sinteza text-vorbire de înaltă calitate si expresivitate în limba română pe baza lexiconului din COBILIRO, a adnotărilor obținute în TEPROLIN (text) și TADARAV (audio), având ca scop generarea de noi voci sintetizate, adaptarea unor aplicații dependente de stilul și expresivitatea vorbirii în diverse ipostaze: știri TV, discurs oratoric, voci cu emotivitate (idem).

La Universitatea „Al. I. Cuza” din Iași s-a derulat un proiect de creare a unui corpus digital al limbii române ce va conține texte vechi, scrise cu alfabet chirilic (*Electronic Corpus of The Ancient Romanian* (1521 – 1640) (CETRV)), echipa de cercetare fiind compusă din: Alexandru Gafton (directorul proiectului), Gheorghe Chivu, Adina Chirilă, George Bogdan Țâra, Roxana Vieru, Ionuț Vieru. Pe pagina electronică a proiectului se menționează că

CETRV (1521-1640) va cuprinde, în final, în variantă facsimilată şi în transcriere interpretativă cu alfabet latin, toate manuscrisele și tipăriturile românești scrise cu alfabet chirilic până la 1640, devenind astfel o sursă info-documentară de bază pentru lingviști, istorici şi teologi, pentru exegeţii literari şi pentru cercetătorii interesaţi de istoria

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

> tehnicii, a dreptului, a ştiinţei şi a culturii româneşti. (...) CETRV (1521-1640) va fi prima bază de date completă a textelor româneşti din perioada veche, ce va putea fi accesată, cu ajutorul internetului, de cercetătorii români şi străini de pretutindeni, ceea ce va deschide o nouă perspectivă asupra culturii româneşti scrise. (http://media.lit.uaic.ro/?page_id=3914; vezi și Chirilă, Țâra 2013).

2.7 ROMBAC (*The Romanian Balanced Annotated Corpus*) este un corpus balansat adnotat la nivel morfosintactic și sintactic, care conține 36 milioane de cuvinte din 5 domenii: jurnalistic, medical, juridic, de istorie literară și ficțiune, dezvoltat de către ICIA (Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București) și disponibil pe platforma META-SHARE (http://www.meta-share.eu/).

3. Corpusuri lingvistice tipărite

Pe lângă aceste corpusuri electronice, există 12 volume tipărite, care conțin transcrierea unor înregistrări de limbă română vorbită contemporană nedialectală (doar câteva parțial dialectale), dintre care amintim: Bochmann, K., V. Dumbrava (ed.), *Limba Română vorbită în Moldova istorică.* Vol. 2. *Texte*, Leipzig, Leipziger Universitätsverlag. Conține 48 de înregistrări transcrise, făcute între 1997-1998; Dascălu Jinga, L., *Corpus de română vorbită (CORV). Eşantioane*, București, Oscar Print. Conține 37 de înregistrări transcrise, făcute între 1993-2001 (cu 2 excepții); Ionescu-Ruxăndoiu, L. (coord.), *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, București, Ed. Univ. din București. Conține 81 de înregistrări transcrise, făcute între 1993-2002 (majoritatea în 2001); Hoarță Cărăușu, L. (coord.), *Corpus de limbă română vorbită actuală nedialectală*, Iași, Ed. Univ. „Alexandru Ioan Cuza". Conține 80 înregistrări transcrise, făcute între 2006-2013 etc. (*cf*. Mîrzea-Vasile 2017).

4. E-sintagmatica. Studiu de caz: *inimă*

4.0. Sintagmatica, în diversele ei segmente, de natură lexicală, semantică și gramaticală, reprezintă o problemă îndelung dezbătută. E-sintagmatica înseamnă circumscrierea problemei în universul digital/ computațional.

4.1. După prezentarea anterioară a corpusurilor computaționale, încercăm să vedem cum putem să întrebuințăm platforma CoRoLa (vezi **2.5**) în vederea extragerii unor metadate, dar și date lexico-semantice, sintagmatice și sintactice, pe care le discutăm sau le prezentăm sub formă de ilustrații-figuri. Pentru aceasta, alegem termenul-cheie *inimă*. Precizăm că, mai degrabă, suntem interesați să aflăm ce conține această platformă, ce fel de date putem să extragem, care este gradul de „fidelitate" al celor înregistrate electronic, decât să întreprindem o analiză pur lingvistică. De aceea, ne rămânem să facem interpretarea datelor într-un viitor studiu.

4.2. În imaginea de mai jos (fig. 1), se poate observa că a fost introdus lexemul *inimă* în bara corespunzătoare (http://89.38.230.10:5555/?q=inim%C4%83&ql=poliqarp). Ca urmare, au fost identificate 15788 de recurențe din cele peste 36 de milioane de

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

propoziții-eșantion, dispuse pe 632 de pagini. Se pot extrage referințele bibliografice și nu numai (*cf.* metadate-standard: autor, data publicării, editură, genul literar al textului etc.), referitoare la un anumit eșantion ce se dorește a fi selectat.



Fig.1. *Inimă*: e-ocurențe

4.2.1. Totuși, pentru a avea numărul total al prezenței electronice a cuvântului „inimă”, este necesar să introducem în zona de chestionare și formele flexionare ale acestui lexem. Observăm o prezență de două ori mai mare a formei articulate „inima” (32951 de e-ocurențe) față de forma inițială, de dicționar („inimă”: 15788 de e-ocurențe). În final, obținem 63657 de recurențe din cele peste 36 de milioane de propoziții-eșantion, dispuse cf. fig. 2.

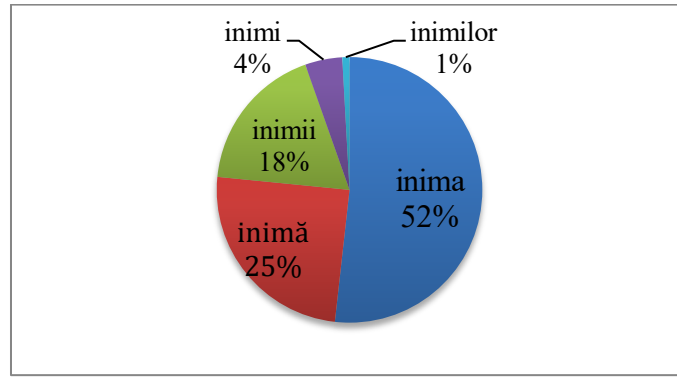


Fig. 2. E-diagrama lingvistică a „inimii”

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

4.2.2. Există posibilitatea de a introduce sintagme nominale. Bunăoară, selectând sintagma „inimă smerită” cu variantele „inima smerită”, „inimi smerite”, „inimii smerite” se identifică 23 de e-ocurențe (dispuse astfel:

6 e-ocurențe:
http://89.38.230.10:5555/?q=inim%C4%83+smerit%C4%83&ql=poliqarp;

13 e-ocurențe:
http://89.38.230.10:5555/?q=inima+smerit%C4%83&collection=&ql=po liqarp;

3 e-ocurențe: http://89.38.230.10:5555/?q=inimi+smerite&ql=poliqarp;

1 e-ocurență: http://89.38.230.10:5555/?q=inimii+smerite&ql=poliqarp);

pentru „inimă rănită” cu variantele „inima rănită”, „inimii rănite” se identifică 68 de e-ocurențe (dispuse astfel: 13 e-ocurențe: http://89.38.230.10:5555/?q=inim%C4%83 +r%C4%83nit%C4%83&ql=poliqarp;

50 de e-ocurențe
http://89.38.230.10:5555/?q=inima+r%C4%83nit%C4%83&ql=poliqarp;

3 e-ocurențe http://89.38.230.10:5555/?q=inimii+r%C4%83nite&ql=poliqarp;

2 e-ocurențe http://89.38.230.10:5555/?q=inimilor+r%C4%83nite&ql=poliqarp);

pentru „inimă de piatră”, cu variantele „inima de piatră”, „inimilor de piatră”, identificăm 103 (dispuse astfel: 56 de e-ocurențe:
http://89.38.230.10:5555/?q=inim%C4%83+de+piatr%C4%83&ql=poliqarp&p=3;

45 de e-ocurențe:
http://89.38.230.10:5555/?q=inima+de+piatr%C4%83&ql=poliqarp;

2 e-ocurențe:
http://89.38.230.10:5555/?q=inimilor+de+piatr%C4%83&ql=poliqarp;

pentru „inimă curată”, cu variantele „inima curată”, „inimii curate”, „inimi curate”, „inimilor curate”, identificăm 368 de (dispuse astfel: 139:
http://89.38.230.10:5555/?q=inim%C4%83+curat%C4%83&ql=poliqarp;
190: http://89.38.230.10:5555/?q=inima+curat%C4%83&ql=poliqarp;

12 e-ocurențe: http://89.38.230.10:5555/?q=inimii+curate&ql=poliqarp;

25 de e-ocurențe: http://89.38.230.10:5555/?q=inimi+curate&ql=poliqarp;

2 e-ocurențe: http://89.38.230.10:5555/?q=inimilor+curate&ql=poliqarp).

La fel am procedat pentru sintagma verbală: „asculta inima”

(7 e-ocurențe, în această variantă flexionară)
http://89.38.230.10:5555/?q=asculta+inima&collection=&ql=poliqarp).

4.2.3. Observăm faptul că putem întâmpina unele dificultăți, dacă folosim o anumită topică a cuvintelor sau dacă folosim varianta de dicționar a sintagmelor, deoarece răspunsul platformei va varia și nu va reflecta numărul real de ocurențe înregistrate în acest corpus (*cf*. și **4.2.2** „inimă de piatră” vs. „inima de piatră” vs. „inimilor de piatră”). Spre exemplu, introducând sintagma de dicționar „a asculta inima”, vom avea o singură ocurență (http://89.38.230.10:5555/?q=a+asculta+inima&coll ection=&ql=poliqarp) (vezi numărul de e-ocurențe la **4.2.9**.).

4.2.4. Pentru sintagma „asculta inima”, în momentul în care apăsăm butonul *Metadata*, sunt prezentate referințele extralingvistice în mod amănunțit (fig. 3).

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

4.2.5. La apăsarea butonului *Tokens*, pentru contextul selecționat („Două feluri de a asculta inima”), se pot afla diverse date de natură lingvistică, gramaticală (morfologice, lexicografice, sintactice etc.), vezi fig. 3, prezentate cu mare acuratețe.

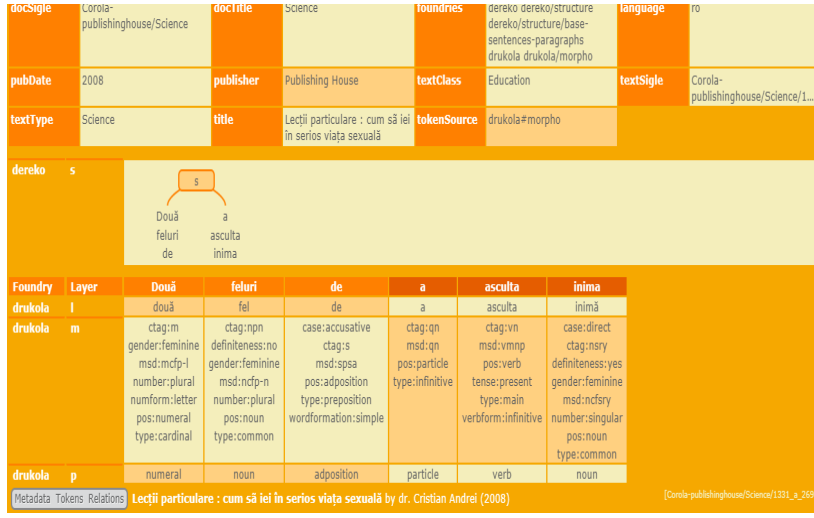


Fig. 3. *Asculta inima*: metadate și diagrama lingvistică

4.2.6. Dacă ne oprim puțin asupra propoziției-eșantion ca răspuns la interogarea-impuls „a asculta inima”, observăm că putem întâmpina unele dificultăți dacă dorim să stabilim sensul pe care îl are sintagma dată în enunțul „Două feluri de **a asculta inima**....” (vezi infra). Pe de o parte, sintagma poate avea sensul propriu, încadrabil limbajului medical, de „a asculta (bătăile) inima (inimii)”, pe de altă parte, poate fi vorba de sensul figurat pe care îl are lexemul *inimă*: „sediu al sentimentelor umane”, deci, sintagma e posibil să semnifice „a răspunde unei chemări sentimentale”. În plus, trebuie să subliniem și faptul că nu putem să spunem cu precizie care este sensul sintagmei, chit că avem acces și la contextul lărgit:

> […] copilului este marcată și de sincretism, adică de tendința de a judeca pornind de la concluziile altora. Aceasta face din copil o persoană naivă în aparență, iar din partenerul sexual foarte tânăr un partener în aparență naiv. Două feluri de **a asculta inima**... Deunăzi vorbeam cu o domnișoară de șaptesprezece ani despre o pățanie de-a ei cu un doctor. Doctorul, văzând că are o sumedenie de probleme neclare și nu tocmai severe, s-a decis la un moment dat să-i asculte…

4.2.7. Sesizând această situație neclară, curiozitatea ne împinge să ne reîntoarcem la ocurențele anterioare (7 e-ocurențe) ale sintagmei „asculta inima”. Astfel, ne oprim asupra primei recurențe:

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

(1) [...] doar o moarte mai blândă despre iubire numai de bine în amurgul făcut pumn plânge o metaforă nenuntită nu-ți voi spune cuvintele care fac luna să descrească un înger mai trece prin ochii mei abia șoptit așa cum s-ar **asculta inima** mea de către inima mea când una din ele nu bate apa degustă grafica albă a norilor vântul prin valuri un strabism al durerii pierzi amintirea mea ca pe o floare căzută de la butonieră pe canale venețiene când mor pescărușii de.

În secțiunea metadate, ni se precizează că fragmentul (1) aparține Violetei Deminescu, extras dintr-un text intitulat *Călătorie de nuntă* și că se poate accesa la un anumit link: „NUNTĂ CĂLĂTORIE a VIOLETA DEMINESCU în ediția nr. 860 din 09 mai 2013 de către http://confluente.ro/Calatorie_de_nunta_violeta_deminescu_13681 21297.html".

În continuare, am procedat la identificarea textului la linkul specificat, însă aflarea acestuia nu a fost posibilă, deși, ajunși la pagina respectivă, am introdus în zona de căutare și aceste sintagme: *Violeta Deminescu*, *Călătorie de nuntă*. Totuși, am reținut faptul că pagina aceasta http://confluente.ro/ reprezintă un blog pe care sunt postate texte din domenii foarte diverse, având o componentă socială accentuată, conținând știri despre împrejurările economico-sociale cu impact imediat, dar și articole de tip lifestyle-magazin.

Mai departe, am luat o parte a fragmentului (1) și l-am căutat cu motorul GOOGLE. Aici, primele două intrări ne trimiteau spre corpusul CoRoLa (platforma KorAP), iar cea de-a treia spre blogul autoarei textului. Am selectat cea de-a treia variantă (http://violetademinescu.blogspot.com/2013/05/) și, astfel, am ajuns la fragmentul căutat (1), vizualizând textul în întregime. Prima constatare importantă este că fragmentul (1) face parte dintr-un text poetic ce are versuri albe, fapt ce nu putea fi stabilit doar vizualizând eșantionul platformei. Această nouă realitate textuală contribuie la descifrarea sensului sintagmei „ascultă inima", adică a înțelegerii sale efective ca exprimare figurată. Apoi, prin corespondența datei la care apare textul (9 mai 2013) din blog și din secțiunea metadate a platformei avem confirmarea că este vorba despre același text.

Următoarele două ocurențe ale sintagmei „asculta inima" sunt expuse în fragmentele (2), (3), extrase de pe același blog (http://confluente.ro/). Și în cazul acestora nu se mai pot identifica textele la httpurile menționate pe platformă, însă sensul sintagmei poate fi determinat cu ușurință din fragmentele-eșantion; la (2), avem un sens direct, iar la (3), un sens figurat:

(2) [...] Pe vremuri, în România, puteam cel mult ***asculta inima*** celui nenăscut cu o pâlnie metalică după patru-cinci luni. (…). (CÂTEVA NOȚIUNI DE BAZĂ DESPRE VENIREA PE LUME de OCTAVIAN CURPAȘ în ediția nr. 262 din 19 septembrie 2011 de către http://confluente.ro/Cteva_notiuni_de_baza_despre_future _pe_lume_.html);

(3) [...] nu avem destul timp? De ce nu mai avem puterea și voința de-a ne opri pentru o clipă ca să ne analizăm mai serios și mai temeinic cursul desfășurării vieților noastre?

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

Oare este atât de devastator să începem prin a ne ***asculta inima*** mai atent*?* (…). (DEZAMĂGIRI ȘI ÎMPLINIRI ÎN PLANUL SPIRITUAL! de la MARIANA DUMITRESCU în ediția nr. 1256 din 09 iunie 2014 de către http://confluente.ro/Mariana_dumitrescu_1402311922.html).

A patra ocurență (4) nu este însoțită la secțiunea Metadate decât de titlul textului din care este extras fragmentul, anul și autorul acestuia, cu toate că introducerea titlului și autorului pe motorul GOOGLE duce la identificarea precisă a linkului (https://www.dcnews.ro/ce-spune-forma-sanilor-despre-o-femeie_483831.html), unde putem să stabilim și tipul textului: lifestyle-magazin. O precizare deosebit de importantă este că textul apare, la linkul inițial, fără diacritice, iar, pe platformă, apare, ca în majoritatea cazurilor, cu diacritice, excepție făcând chiar sintagma analizată, fapt ce are ca efect schimbarea numărului de ocurențe, după cum vom vedea mai jos. De asemenea, și acest fragment permite deslușirea sensului (sens figurat):

(4) [...] mereu în centrul atenției și știi cum să faci față acestui gen de situație. Îți place să seduci și să fii sedusă, să dai și să primești atenție. Nu-ți plac nici rutină, nici banalitatea, nici oamenii care nu-și ***asculta inima***. Mottoul vieții tale: „Love is all we need. – Dacă dragoste nu e, nimic nu e”.

A cincea ocurență (5) are doar următoarele mențiuni că este un text literar imaginativ și că este publicat, fără a fi precizate alte metadate precum autorul, titlul etc.

(5) [...] Bătrânul era mort când l-au adus înăuntru și, în timp ce doctorul îi ***asculta inima*** cu o chestie pe care și-o băgase-n urechi, am auzit o împușcătură afară, asta însemnând că l-au omorât pe Gilford. M-am întins lângă bătrân când au cărat targa la spital și m-am ținut bine de.

Apelăm, din nou, la GOOGLE și reușim să identificăm un fragment-sursă, aproape identic cu cel din platforma analizată, fiind vorba despre un eșantion din textul literar *Bătrânul* de E. Hemingway, tradus în limba română. Odată cu acest fapt se ridică o problemă intens dezbătută în studiile recente de specialitate din spațiul occidental în care se utilizează metoda corpusului, referitoare la faptul că unii dintre cercetători consideră că textele traduse nu ar putea să formeze un corpus reprezentativ pentru limba traducerii, ci doar un corpus reprezentativ al evidențelor traductive, al particularităților traductive etc. Ca urmare, în accepțiunea acestora, digitalizarea unei limbii trebuie să însemne crearea unui corpus de referință al limbii date în mod separat de corpusul de texte traduse în acea limbă (*cf*. Baker 1995, Olohan 2002). De pildă, în limba engleză există *BNC* (British National Corpus) și *TEC* (Translational English Corpus) (idem).

Revenind la ocurențele sintagmatice cu lexemul-cheie „inimă”, (6) face parte dintr-un text medical în limba română *Radio-oncologia cancerului genital feminin*, scris de Eduard Bild[1]. Sintagma „asculta inima” are un sens propriu:

[1] Renumit oncolog român.

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

(6) [...] În continuare, bolnavul este culcat pe spate, trecându-se la examinarea sistemului cardio- vascular: – se palpează vârful cordului; – se **asculta inima** la vârf, în focarul pulmonar și aortic (…).

Ultima ocurență este cea prezentată la subpunctul **4.2.6.**, care reprezenta răspunsul platformei la o formă de dicționar a verbului din sintagma supusă analizei noastre și care punea probleme în identificarea sensului propriu sau figurat al întrebuințării.

4.2.8. La o nouă căutare pe platforma proiectului CoRoLa, dar, de data aceasta, folosind forma flexionară „ascultă inima", regăsim tot 7 ocurențe (http://89.38.230.10 :5555/?q=ascult%C4%83+inima&collection=&ql=poliqarp). Primele 2 ocurențe se află, de fapt, în același fragment. Redau eșantionul în care se află sintagma analizată: „Apoi îl așează la orizontală, îi **ascultă inima**, îi luă pulsul (FRAGMENT DIN NUVELA RASCRUCEA DESTINULUI de la VASILICA ILIE în ediția nr. 222 din 10 august 2011 de către http://confluente.ro/Fragment_din_nuvela_rascrucea_destin ului_.html)". A treia ocurență nu este însoțită de multe date standard, doar se specifică faptul că este un fragment literar: „Îmi **ascultă inima** și plămânii, iar capul chel i se înclină pe gâtul lung precum capul unui brontozaur gustând frunze". A patra ocurență apare într-un fragment literar ce aparține lui G. Călinescu, fiind prezentă în romanul Enigma Otiliei: „Făcu o auscultație sumară, palpă și percută ficatul, **ascultă inima**". Următoarea ocurență este prezentă într-un fragment literar tradus, din romanul *Ultima iubire a președintelui* de Andrei Kurkov (remarcăm că se precizează și traducătorul: Antoaneta Olteanu): „Controlul medical trecu destul de repede. (…) După ce **ascultă inima**, dădu din cap satisfăcut.". A șasea ocurență apare în fragmentul-eșantion din romanul *Celsius: 41,1* de Victor Cojocaru: „Doctorul vârstnic e din nou lângă mine și îmi **ascultă inima**". Ultima ocurență aparține unui fragment de text folcloric în versuri „Dragă mi-i frunza, iarba, / Că-mi **ascultă inima**". Remarcăm faptul că de data aceasta se indică prezența unor versuri prin folosirea barei oblice. Cu excepția ultimei ocurențe, în toate celelalte exemple-eșantion, sintagma „ascultă inima" are sensul propriu.

4.2.9. Mai trebuie să menționăm că la alte chestionări ale platformei cu aceeași sintagmă dar cu diverse flexiuni vom întâlni alte e-ocurențe: „ascult inima": 12 e-ocurențe (http://89.38.230.10:5555/?q=ascult+inima&ql=poliqarp); „asculți inima": 10 e-ocurențe (http://89.38.230.10:5555/?q=ascul%C8%9Bi+inima&ql=poliqarp); „ascultau inima": 1 e-ocurență (http://89.38.230.10:5555/?q=ascultau+inima&ql=poliqarp); „ascultăm inima": 1 e-ocurență (http://89.38.230.10:5555/?q=ascult%C4%83m+inima& ql=poliqarp); „ascultați inima": 2 e-ocurențe (http://89.38.230.10:5555/?q=asculta%C8%9B i+inima&ql=poliqarp); „ascultând inima": 1 e-ocurență (http://89.38.230.10:5555/?q=ascu lt%C3%A2nd+inima&ql=poliqarp). În final, însumând toate prezențele electronice, sub diverse forme flexionare, obținem 41 de e-ocurențe. Trebuie să spunem că această chestionare, cantitativă îndeosebi, ne-a făcut să remarcăm absența unor forme flexionare ale verbului „a asculta" din sintagma supusă atenției precum formele de perfectul simplu, de mai mult ca perfect sau de viitor, fapt ce poate fi interpretat cu ajutorul unor factori intra- și extralingvistici. De asemenea, nici îndemnul „ascultă-ți

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

inima!” nu este prezent în acest corpus (http://89.38.230.10:5555/?q=ascult%C4%83-%C8%9Bi+inima&ql=poliqarp). Într-un alt studiu, se impune interpretarea acestor observații.

La selectarea butonului pentru a identifica lexemele similare, analogice pentru cuvântul „inimă” (http://89.38.230.23/word_embeddings/index.php?similarity_w=inima&do=similarity), apar:

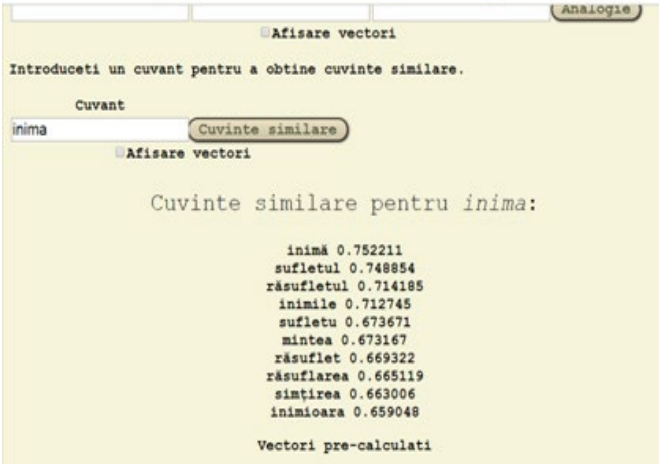


Fig. 3. *Inimă*: e-sinonime și e-cuvinte similare

4.4. În e-corpusul oral al proiectului CoRoLa, lexemul „inimă”, cu variantele „inima”, „inimii”, „inimi”, „inimilor” are 106 e-ocurențe, dispuse astfel:

36 de e-ocurențe: http://89.38.230.23/corola_sound_search/index.php?search_type=0&search=inim%C4%83&search2_type=0&search2=&start=0&count=20&show_word=on&context=5;

60 de e-ocurențe: http://89.38.230.23/corola_sound_search/index.php?start=20&count=20&search=inima&search_type=0&search2_type=0&search2=&show_word=on&context=5

9 e-ocurențe: http://89.38.230.23/corola_sound_search/index.php?search_type=0&search=inimii&search2_type=0&search2=&start=0&count=20&show_word=on&context=5;

1 e-ocurență: http://89.38.230.23/corola_sound_search/index.php?search_type=0&search=inimi&search2_type=0&search2=&start=0&count=20&show_word=on&context=5.

5. Concluzii

Corpusul electronic poate oferi varii date și informații relevante pentru o limbă, care să ofere un suport real cercetărilor din diverse domenii. Iar, dacă există segmente lingvistice importante care nu pot fi evidențiate cu ajutorul lingvisticii corpusului, atunci se poate apela la metode conjugate de analiză.

În era digitalizării, munca întreprinsă de cercetătorii din zona prelucrării limbajului natural și a lingvisticii computaționale este deosebit de însemnată, mai ales

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)

că s-a reușit, în timp relativ scurt, să se facă pași importanți în informatizarea limbii române scrise și vorbite, îndeosebi în format nedialectal, oferindu-se suport pentru viitoarele înregistrări.

Referințe bibliografice

Baker, Mona. 1995. *Corpora in Translation Studies: An Overview and Some Suggestions for Future Research*, in „Target", 7(2), p. 223-243.

Barbu Mititelu, Verginica, Tufiș, Dan, Irimia, Elena. 2018. *The Reference Corpus of the Contemporary Romanian Language (CoRoLa)*, in „Proceedings of LREC", p. 1178-1185.

Cherata, Sanda, Mihăescu, Manuela. 2008. *Modele formale de reprezentare a informaţiilor lexicale şi terminologice în Proiectul SIASTRO*, in „Dacoromania", serie nouă, XIII (2), p. 151-169.

Chirilă, Adina, Țâra, George Bogdan. 2013. *Opțiuni și constrîngeri lexico-semantice în traducerile textului biblic*, in Emanuel Gafton, Sorin Guia, Ioan Milică (ed.), „Perspective asupra textului și discursului religios", Iași: EUAIC, p. 111-123.

Cosma, R., Cristea, D., Kupietz, M., Tufiș, D., Witt, A. 2016. *DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora*, in „Proceedings of LREC", p. 28-32.

Ion, R., Irimia, E., Ștefănescu, D., Tufiș, D. 2012. *ROMBAC: The Romanian Balanced Annotated Corpus*, in „Proceedings of LREC", p. 339-344.

Irimia, Elena, Barbu Mititelu, Verginica. 2015. *RACAI-RoTb: nucleu de corpus de limbă română adnotat sintactic cu relaţii de dependenţă*, in „Revista Română de Interacţiune Om-Calculator", 8 (2), p. 101-120.

Mîrzea-Vasile, Carmen. 2017. *Corpusurile de limbă română şi importanţa lor în realizarea de materiale didactice pentru limba română ca limbă străină*, in „Romanian Studies Today", I, p. 74-95.

Olohan, M. 2002. *Corpus Linguistics and Translation Studies: Interaction and Reaction*, in „Linguistica Antverpiensia", 1, p. 419-429.

Păiș, V., Tufiș, D. 2018. *Computing Distributed Representations of Words using the CoRoLa Corpus*, in „Proceedings of the Romanian Academy", Series A, 19(2), p. 403-409.

Tămâianu-Morita, Emma, Cherata, Sanda, Vîlcu, Cornel. 2006-2007. *Analiza sintagmatică a textelor româneşti prin mijloace informatice: Proiectul SIASTRO*, in „Dacoromania", serie nouă, XI-XII, p. 77-87.

Tămâianu-Morita, Emma. 2008. *Tipologia sintagmelor în modelul D. D. Draşoveanu. Posibile aplicaţii în proiectul SIASTRO*, in „Dacoromania", serie nouă, XIII, p. 137-150.

Tufiș, D. 2008. *Ro-WordNet: ontologie lexicală pentru limba română*, in „Academica", XVIII (208-209), p. 30-34.

Tufiş, D., Barbu, E., Mititelu, V., Ion, R., Bozianu, L. 2004a. *The Romanian Wordnet*, in Tufiş, D. (ed.) „Romanian Journal on Information Science and Technology. Special Issue on BalkaNet", vol. 7 (2-3), p. 107-124.

Tufiş, D., Cristea, D. 2002. *Methodological Issues in Building the Romanian Wordnet and Consistency Checks in Balkanet*, in „Proceedings of LREC 2002 Workshop on Wordnet Structures and Standardization", p. 35-41.

Tufiş, D., Cristea, D., Stamou, S. 2004b. *BalkaNet: Aims, Methods, Results and Perspectives: A General Overview*, in Tufiş, D. (ed.) „Romanian Journal on Information Science and Technology. Special Issue on BalkaNet", Romanian Academy, vol. 7 (2-3), p. 9-43.

Tufiș, Dan. 2018. *CoRoLa Primul corpus computațional de referință pentru limba română contemporană*, in „Market Watch", nr. 205, p. 28-29.

Vîlcu, C. 2008. *Preliminarii teoretice la analiza gramaticală în proiectul SIASTRO: Nivelul sintagmatic*, in „Dacoromania", serie nouă, XIII, nr. 2, p. 117-126.

BDD-A31434 © 2020 Jatepress
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 16:23:33 UTC)