Corpusurile de limba română și importanța lor în realizarea de materiale didactice pentru limba română ca limbă străină

Carmen MÎRZEA-VASILE
Institutul de Lingvistică al Academiei Române „Iorgu Iordan-Al. Rosetti", București Facultatea de Litere, Universitatea din București

Abstract

The Romanian Corpora and their importance in creating teaching materials **for Romanian L2**

The article has two aims: 1. to describe the corpora of contemporary non-dialectal Romanian, including both electronic corpora — *The Romanian Balanced Annotated Corpus* (ROMBAC), RoCo_News (a Journalistic Corpus of Romanian), *The Reference Corpus of Contemporary Romanian Language* (CoRoLa), etc. — and raw oral corpora, available in print only — *Româna vorbită actuală* (ROVA), *Corpus de română vorbită* (CORV), *Interacțiunea verbală în limba română actuală* (IVLRA), *Corpus de limbă română vorbită actuală* (CLRVA), etc.; 2. to plead for using corpora for pedagogical purposes, especially in creating teaching materials for Romanian as a foreign / second language. The article gives a short general description of the corpora and their applications in Second Language Acquisition and Foreign Language Teaching. The Romanian corpora are hardly known even by the Romanian researchers; their presentation takes into account the stylistic structure, annotation, number of words and tokens, etc. (for electronical corpora); the number of texts, the period of time when the records were made, the type of texts, etc. (for oral corpora in print). The second part contains some examples of possible corpora applications for Romanian as a foreign/ second language: a list of the most frequent words; the refinement of the characteristics of various types of texts (medical, legal, journalistic, fiction, etc.); the most relevant contexts for the argumental structure of verbs, adjectives, etc.

In fact, the aim of the paper is to argue for developing annotated corpora for Romanian, easily accessible to researchers, professors and even students, and for using the existing corpora for pedagogical purposes.

Key-words: Romanian electronic Corpora, Romanian oral Corpora, Romanian as a foreign/ second language, pedagogical applications.

1. Preliminarii

Prima parte a acestui articol este o descriere sumară a corpusurilor românești (electronice și în format print), limitându-se la corpusurile lingvistice de română contemporană nedialectală. În a doua parte, sunt date câteva exemple de posibile aplicații ale corpusurilor în procesul de predare-învățare și elaborarea de materiale didactice pentru limba română ca limbă străină/ limbă a doua (L2). Aceste date sunt precedate de o scurtă introducere generală despre corpusuri și despre cele mai cunoscute moduri în care pot fi folosite în predarea-învățarea unei limbi străine. Articolul este, de fapt, o pledoarie pentru utilizarea corpusurilor românești în scopul menționat și, indirect, pentru dezvoltarea corpusurilor informatizate de limbă română și creșterea gradului lor de accesibilitate pentru publicul interesat.

2. Scurtă introducere generală

2.1. Definiție, caracterizare, tipuri de corpusuri

2.1.1. Despre corpusuri și aplicațiile lor, inclusiv în domeniul învățării unei limbi nenative și în predarea-învățarea limbii a doua, există o bibliografie străină enormă. Două dintre definițiile cele mai cunoscute sunt următoarele: „Corpus is a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language" (Crystal 2008) și „Corpus is an electronically stored collection of samples of naturally occurring language" (Hunston 2006: 234).

2.1.2. În definițiile curente, *corpus* înseamnă o colecție de texte, caracterizată prin (Stubbs 2013: 106):

(i) dimensiune mare — cele mai multe corpusuri moderne numără cel puțin 1 milion de cuvinte, Hunston 2006: 234, unele chiar sute de milioane);

(ii) posibilitatea de prelucrare electronică — cu ajutorul instrumentelor de exploatare electronică de tip concordanțier (care poate stabili concordanțe, care poate găsi, lista sau clasifica diverse contexte lingvistice, care poate număra ocurențe) sau de tip lematizator (prin intermediul căruia se poate adăuga o lemă, o intrare de dicționar fiecărei ocurențe);

(iii) scop: sunt destinate analizei lingvistice, fiind construite astfel încât să fie reprezentative (dimensiunea, structura, elaborarea urmează principiile unei teorii sociolingvistice de variație a limbii, astfel încât datele de

limbă să fie un eșantion reprezentativ pentru limba respectivă sau pentru un anumit registru, variantă a ei).

2.1.3. Criteriile generale de clasificare a corpusurilor nu variază esențial de la o lucrare la alta. În Cristea (2005) sunt reținute șase criterii:

(1) Tipul de text transcris/ înregistrat: a. colecții de texte, conținând texte inițial scrise; b. înregistrări de vorbire, conținând înregistrări ale limbajului (inițial) vorbit.

(2) Criteriul explicitării: a. corpusuri primare; b. corpusuri adnotate. Corpusurile primare (*raw corpora*) conțin texte în formatul inițial, dedicat uzului uman. În cazul corpusurilor adnotate, „textul primar este suplimentat cu adnotări ce reprezintă explicitarea în format inteligibil pentru mașină a informațiilor lingvistice și extralingvistice" (Cristea 2005: 4) considerate pertinente. Pentru adnotarea corpusurilor se folosesc limbaje specializate, cel mai utilizat fiind *Extended Mark-Up Language* (XML), ca și standarde de adnotare, precum *Corpus Encoding Standard* (CES) și *Text Encoding Initiative* (TEI).

(3) Criteriul cantității: a. corpusuri mai mari; b. corpusuri mai mici. Pentru a fi reprezentativ pentru o limbă, un corpus nu poate fi mai mic de 50 de milioane de cuvinte (Cristea 2005: 4). De exemplu, *British National Corpus* este un corpus reprezentativ (100 de milioane de cuvinte).

(4) Criteriul conținutului: a. corpusuri generale, de referință; b. corpusuri specializate. Corpusurile de referință (cum este *The Bank of English*) conțin date de limbă suficient de multe și de variate pentru a sta la baza unei gramatici de referință sau a unui dicționar-tezaur. În selectarea textelor care alcătuiesc un corpus de referință se ține cont de cât mai multe variabile sociolingvistice, iar proporția tipurilor de text urmează un anumit model. Ponderea limbajului vorbit în corpusurile moderne de referință este de aproximativ 10% (Bonelli, Sinclair 2006: 213). Corpusurile specializate (medicale, juridice, dialectale etc.) documentează un anumit tip de text, de registru, de variantă sau de stil al limbii.

(5) Criteriul temporalității: a. corpusuri pentru o anumită perioadă de timp; b. corpusuri-monitor; c. corpusuri atemporale. Primele sunt corpusurile care reprezintă limba caracteristică unei anumite perioade (spaniola medievală, engleza actuală etc.). Corpusurile-monitor (cum este *Global English Monitor Corpus*) au o dimensiune păstrată constantă în timp prin adăugarea de material nou care înlocuiește materialul mai vechi ce va fi arhivat; scopul este monitorizarea foarte precisă a evoluției limbii contemporane. În cazul corpusurilor atemporale este ignorat anul de apariție a textelor.

(6) Criteriul comparabilității: a. corpusuri monolingve; b. corpusuri multilingve. Corpusurile multilingve sunt folosite în principal pentru aplicații pentru traducerea automată. Corpusul paralel este un tip special de corpus multilingv; acesta este alcătuit din traduceri ale aceluiași text. Un exemplu de corpus multilingv aliniat este MultextEast, care conține

traducerile aliniate în 25 de limbi ale *Republicii lui Platon* și ale romanului *1984* de George Orwell, în 10 limbi (Cristea 2005: 4-5). Dacă folosim într-o accepție mai largă termenul *corpus*, incluzând și resursele textuale care sunt doar în format print, putem include printre corpusurile paralele și cunoscuta lucrare *Introduction à la morphologie comparée des langues romanes, basée sur des traductions anciennes des Actes des Apôtres*, ch. XX à XXIV (De Poerck, Mourin 1961-1964), precum și *Oratio Dominica Romanice* (Heinimann 1988) și *Die Bibel in der Romania: Matthäus 6, 5–13* (Heger 1967). De asemenea, merită menționate resursele de texte paralele disponibile on-line, precum Europarl (*A Parallel Corpus for Statistical Machine Translation*), JRC-ACQUIS (*Multilingual Parallel Corpus*), EMEA (*European Medicines Agency*). Pentru mai multe exemple de asemenea resurse, atât în format print, cât și informatizate, și pentru detalii, vezi Mîrzea Vasile (2015: 21-30). Corpusul de texte comparabile este, în general, tot multilingv, și conține texte originale comparabile ca gen, registru, temă, datare etc. Un exemplu de corpus comparabil este C-ORAL-ROM (*Integrated Reference Corpora for Spoken Romance Languages*), pentru a cărui descriere și trimiteri vezi tot Mîrzea Vasile (2015: 28-29).

2.1.4. Un tip special de corpus este corpusul vorbitorilor nenativi (*learner corpus*), construit cu principalul scop de a fi exploatat în activitățile ce privesc pedagogia unei limbi nenative. Una dintre definițiile acestuia este: „Learner corpora are electronic collections of authentic FL (Foreign Language)/SL (Second Language) textual data according to explicit desing criteria for particular SLA (Second Language Acquisition)/FLT (Foreign Language Teaching) purpose. They are encoded in a standardised and homogeneous way and are documented as to their origin and provenance" (Granger 2002: 7). În funcție de mai multe criterii (vezi și clasificarea generală anterioară), corpusurile celor care învață o limbă nefiind nativi (Díaz-Negrillo, Thompson 2013: 10) pot fi:

- corpusuri de texte scrise sau corpusuri orale, de limbă vorbită transcrisă — în funcție de tipul de producere;
- corpusuri adnotate sau neadnotate (*raw learner corpus*) — în funcție de gradul de adnotare;
- corpusuri multilingve sau monolingve — în funcție de numărul de limbi;
- corpusuri cu control, adică având informații despre cei care au produs textele, și corpusuri fără control — în funcție de condițiile în care au fost colectate datele;
- corpusuri longitudinale, diacronice, în care sunt colectate producții pe o perioadă mai lungă de timp, în general, de la câțiva subiecți, sau sincronice (*cross-sectional*), în care sunt colectate date de la mai mulți subiecți, în aceeași perioadă de timp — în funcție de perioada de timp avută în vedere;

• corpusuri generale sau specializate — în funcție de amploare, de tipul de conținut.

Cele mai multe corpusuri de limbă nenativă sunt scrise (în general, compuse din texte academice, de obicei, în format electronic) și sincronice. Dacă sunt adnotate, în general adnotarea privește greșelile. Aceste corpusuri sunt mai degrabă experimentale decât autentice.

În prezent există peste 150 de corpusuri de limbă nenativă. Acestea sunt inventariate pe site-ul Universității Catolice din Louvain (*https://www.uclouvain.be/en-cecl-lcworld.html*). Engleza nenativă beneficiază de cele mai multe corpusuri (CLC – Cambridge Learner Corpus, ICLE – International Corpus of Learner English, BELC – The Barcelona English Language Corpus, BICCEL – The Bilingual Corpus of Chinese English Learners, CALE – The Corpus of Academic Learner English etc.). Există corpusuri și pentru alte limbi vorbite de nenativi (franceză: FLLOC – French Learner Language Oral Corpora, FRIDA – French Interlanguage Database, CEFLE – Corpus Écrit de Français Langue Étrangère, IPFC – Interphonologie du Français Contemporain; germană: Falko – Fehlerannotiertes Lernerkorpus; WHiG – What's Hard in German?; spaniolă: CEDEL2 – Corpus Escrito del Español L2; finlandeză: ICLF — International Corpus of Learner Finnish; norvegiană: ASK – Language Learner Corpus of Norwegian etc.).

2.2. Utilizarea corpusurilor lingvistice în activități care implică predarea/învățarea unei limbi nenative

Analiza corpusurilor în activitățile care implică cercetarea și predarea/învățarea unei limbi nenative oferă informații legate de: (i) frecvența unui fenomen, a unui lexem etc.; (ii) colocații, expresii, structuri fixate, solidarități lexicale etc.; (iii) variația fenomenelor, a construcțiilor în funcție de registru; (iv) lexicul care trebuie utilizat în exersarea gramaticii; (v) autenticitatea faptelor de limbă (McEreny, Xiao 2010: 366).

Ca și pentru corpusuri în general, există o bibliografie extrem de bogată și pentru aplicațiile care se pot face pe baza corpusurilor în pedagogia limbii nenative. Dintre acestea, amintim doar volumele *How to use corpora in language teaching* (Sinclair 2004) și *From corpus to classroom: language use and language teaching* (O'Keeffe, McCarty, Carter 2007), precum și două secțiuni consistente din *The Routledge handbook of corpus linguistics* (O'Keeffe, McCarty (ed.) 2010) — *Using a corpus for language pedagogy and methodology* (pp. 317-384) și *Designing corpus-based materials for the language classroom* (pp. 385-487) și capitolul *What corpora can offer in language teaching and learning* (McEnery, Xiao 2010: 364-380) din *Handbook of research in second language teaching and learning* (Hinkel (ed.) 2010).

În domeniul care ne interesează, corpusurile pot fi folosite direct și indirect (Römer 2008: 113-114, McEnery, Xiao 2010):

(1) Aplicațiile directe se referă la utilizarea corpusurilor la ore, de către

studenți și profesor, sau în particular, de nenativii care învață o limbă străină (*data-driven learning*). Textele din World Wide Web pot constitui un tip special de corpus, ușor accesibil nenativilor care învață o limbă străină, dar utilizabil cu rezerve.

(2) Aplicațiile indirecte îi ajută pe cercetători, pe creatorii de materiale didactice și pe profesori să ia decizii în privința a ce trebuie învățat, a ordinii în care trebuie învățat și a modului cum trebuie predat. Altfel spus, pe baza corpusurilor, pot fi realizate dicționare, gramatici, culegeri de exerciții, teste, etc. destinate nenativilor. Ne vom limita aici la câteva exemple de lucrări realizare pe baza analizei de corpus (al vorbitorilor nativi sau nenativi) destinate nenativilor care învață limba engleză:

- *Collins COBUILD English course* (CCEC — Willis, Willis 1989) și alte materiale didactice COBUILD, construite pe baza unui corpus general enorm de engleză scrisă și vorbită (The Collins Corpus — COBUILD);
- *Longman student grammar of spoken and written English workbook* (Conrad, Biber, Leech 2002), o descriere a limbii engleze care se bazează, în întregime, pe datele oferite de analiza de corpus (Longman Spoken American Corpus și British National Corpus — BNC);
- *Learning from common mistakes* (Brook-Hart 2009), o culegere de exerciții realizată pornind de la Cambridge Learner Corpus (CLC);
- *Macmillan English dictionary advanced learner* (Rundell 2007), un dicționar bazat, în parte, pe International Corpus of Learner English (ICLE);
- *Longman Essential Activator* (1997), un dicționar cu casete care explică greșeli comune, pentru care a fost folosit Longman's Learner Corpus (LLC);
- *Focus on Vocabulary 2: Mastering the academic word list* (Schmitt, Schmitt 2011), lucrare destinată exersării vocabularului academic de către nenativi, realizată cu ajutorul unui corpus de engleză academică.

3. Corpusurile lingvistice de română actuală

Atât la Academia Română (Institutul de Cercetări pentru Inteligență Artificială și Institutul de Informatică Teoretică – Iași), cât și în anumite centre universitare (Universitatea din București, Universitatea Transilvania din Brașov, Universitatea „Alexandru Ioan Cuza" din Iași, Universitatea „Babeș-Bolyai" din Cluj), prin proiecte naționale sau în colaborare internațională, s-au realizat sau sunt în lucru diverse corpusuri electronice sau volume în format print conținând texte considerate a alcătui un anumit tip de corpus. Volumele în format print conțin, majoritatea, transcrieri de texte orale.

În continuare vor fi prezentate foarte succint cele mai importante corpusuri electronice și volumele de corpus oral care ilustrează româna contemporană de uz general.

Prin cantitatea de material și prin posibilitățile de investigare, corpusurile electronice, în comparație cu cele în format print (de dimensiuni reduse, care nu pot fi interogate electronic), sunt, în mod clar, mai ușor de folosit în cercetarea lingvistică. Totuși, corpusurile orale în format print prezintă un interes deosebit în stadiul actual al dezvoltării corpusurilor de română, reprezentând, momentan, un tip de text complementar tipurilor conținute în corpusurile electronice finalizate și, într-o oarecare măsură, accesibile. După cum se va vedea (în secțiunea 3.1), corpusurile electronice disponibile nu includ și texte orale, componentă obligatorie a oricărui corpus reprezentativ pentru o limbă standard (vezi secțiunea 2.1, criteriul 4).

3.1. Corpusurile electronice

ROMBAC (*The Romanian Balanced Annotated Corpus*, descris detaliat în Ion, Irimia, Ștefănescu, Tufiș 2012) este realizat la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" al Academiei Române, în contextul proiectului internațional METANET4U. Este cel mai mare corpus adnotat pentru română. Conform licenței, această resursă este gratuită, dacă este folosită pentru cercetarea lingvistică, și contra cost, dacă este folosită în scopuri comerciale.

Acest corpus conține cinci părți egale de texte reprezentând genurile: a. jurnalistic (știri și editoriale), b. juridic, c. ficțiune (romane și poezie, scrise sau traduse), d. medicină și farmacie (texte scurte) și e. biografii ale diferitelor personalități române și lucrări de critică literară. Întregul corpus conține în jur de 36 de milioane de cuvinte (prin repetare, echivalente cu peste 44 de milioane de ocurențe), distribuite în proporții aproximativ egale în cele cinci tipuri de texte. Acest corpus are nevoie de un spațiu de stocare de 4 GB. Sunt folosite diacriticele.

ROMBAC este adnotat la nivel de paragraf, de frază, de cuvânt și de grup sintactic, oferind informații morfosintactice. Conține metadate de tip metanet. Sunt identificate grupurile sintactice, dar nu sunt marcate relațiile dintre ele. Se pot face analize statistice (lexicale, morfologice și sintactice sau analize combinând nivelurile) relevante pentru domeniile reprezentate în corpus. ROMBAC este destinat, în primul rând, uzului de către mașină, și mai puțin celui uman, pentru că nu e o interfață pentru căutări. Se pot face căutări manuale, cu ajutorul funcției Search, oferită de aplicația cu care se deschide fișierul.

La același institut de cercetare este realizat și Corpusul RoCo_News (descris în Tufiș, Irimia 2006). RoCo_News reprezintă, de fapt, partea de texte jurnalistice din ROMBAC și este, așadar, un corpus de texte jurnalistice din perioada 2003-2006. Are dimensiune medie (aproximativ 7 milioane de ocurențe). Abundă în nume proprii, numerale, nume de instituții.

CoRoLa (*The Reference Corpus of Contemporary Romanian Language*, descris în Mititelu, Irimia, Tufiș 2014) este un proiect al Academiei Române,

în derulare la același institut menționat din București și la Institutul de Informatică Teoretică din Iași. Este prevăzut ca acest corpus să fie reprezentativ, conținând mai mult de 500 de milioane de cuvinte și reprezentând toate stilurile funcționale, prin intermediul textelor inițial scrise și orale. Durata materialului oral transcris este planificată să fie de 300 de ore. Va fi un corpus pre-procesat și adnotat (cel puțin la nivel lexical) și va conține metadate (autor, perioadă, tipul de text, număr de cuvinte etc.). CoRoLa se dorește a fi destinat uzului uman (va avea interfață de căutare în corpus, subcorpus, după anumite criterii etc.).

Corpusul ZiareRom (descris foarte succint pe site-ul Institutului de Lingvistică din București „Iorgu Iordan − Al. Rosetti" al Academiei Române, www.lingv.ro) este o resursă electronică creată din inițiativă privată. Acesta este alcătuit din texte culese din variantele on-line ale unor ziare românești (*Adevărulonline, BBC-Romanian, Bursa, Capital-RO, Cotidianul, Crainou, Euractiv-ro, Evenimentul Zilei, Jurnalul, Libertatea, Ziarul de Iași, Ziua, 7Plus*) din perioada 2004-2007 și conține peste 86 de milioane de cuvinte.

3.2. Corpusurile în format print

Relativ recent (2000-2015), au apărut 12 volume de corpus lingvistic de română vorbită contemporană nedialectală (câteva sunt dialectale numai parțial și numai la nivel fonetic și lexical). Termenul *corpus* este folosit, în acest caz, într-o accepție largă (orice colecție de texte întocmită cu scop lingvistic). În lista cronologică de mai jos, pentru fiecare lucrare au fost date, în afară de indicațiile bibliografice, următoarele informații de bază: număr de texte transcrise și echivalentul total în pagini, durata înregistrărilor transcrise, perioada când s-au efectuat înregistrările.

Bochmann, Dumbrava (2000) = Bochmann, K., V. Dumbrava (ed.), *Limba Română vorbită în Moldova istorică.* Vol. 2. *Texte*, Leipzig, Leipziger Universitätsverlag. Conține 48 de transcrieri (= 295 p.), echivalentul a 13h,39'25" de înregistrări făcute, în general, între 1997-1998.

CORV (2002) = Dascălu Jinga, L., *Corpus de română vorbită (CORV). Eșantioane*, București, Oscar Print. Conține 37 de transcrieri (= 245 p.), echivalentul a 3h,19'43" de înregistrări făcute în perioada 1993-2001 (cu 2 excepții).

IVLRA (2002) = Ionescu-Ruxăndoiu, L. (coord.), *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, București, Ed. Univ. din București. Conține 81 de transcrieri (= 241 p.). Înregistrările au fost făcute în perioada 1993-2002 (majoritatea în 2001).

Bochmann (2004) = Bochmann, K. (ed.), *Gesprochenes Rumänisch in der Ukraine. Soziolinguistische Verhältnisse und linguistische Strukturen*, Leipzig, Leipziger Universitätsverlag. Conține 37 de transcrieri (= 139 p.). Înregistrările au fost făcute în 1998 și în perioada 2000-2004.

CLRVA I (2005) = Hoarță Cărăușu, L. (coord.), *Corpus de limbă română vor-*

bită actuală, Iași, Ed. Tehnică, Științifică și Didactică Cermi. Conține 36 de transcrieri (= 216 p.), echivalentul a 3h,19′58″ de înregistrări făcute în perioada 2004-2005 (2 înregistrări nu sunt datate).

IV II (2007) = Ionescu-Ruxăndoiu, L. (coord.), *Interacțiunea verbală (IV II). Aspecte teoretice și aplicative. Corpus*, București, Ed. Univ. din București. Conține 24 de transcrieri (= 133 p.). Înregistrările au fost făcute în perioada 2000-2007 (o înregistrare nu e datată).

Gheorghe, Măda, Săftoiu (2009) = Gheorghe, M., S. Măda, R. Săftoiu, *Comunicarea la locul de muncă. Corpus de interacțiune verbală în mediul profesional* (+ CD), Brașov, Ed. Univ. Transilvania din Brașov. Conține 29 de transcrieri (= 177 p.). Înregistrările au fost făcute în perioada 2006-2009.

ROVA (2011) = Dascălu Jinga, L. (coord.), *Româna vorbită actuală (ROVA). Corpus și studii*, București, Ed. Acad. Române. Conține 32 de transcrieri (= 178 p.), echivalentul a 5h,44′7″ de înregistrări făcute în perioada 2008-2010.

CLRVA II (2013) = Hoarță Cărăușu, L. (coord.), *Corpus de limbă română vorbită actuală nedialectală*, Iași, Ed. Univ. „Alexandru Ioan Cuza". Conține 80 de transcrieri (= 569 p.), echivalentul a 22h,6′38″ de înregistrări făcute în perioada 2006-2013.

Corpus POP I (2004) = Pop, L. (ed.), *Verba Volant. Recherche sur l'oral*, Cluj-Napoca, Echinox. Conține 20 de transcrieri (= 49 p.). Unele înregistrări au fost făcute în 2003. Cele mai multe înregistrări nu sunt datate.

Corpus POP III (2010) = Pop, L. (ed.), *Où va la communication?(Comunicarea, încotro?)*, Cluj-Napoca, Echinox. Conține 11 transcrieri (= 28 p.). Înregistrările au fost făcute în 2009 (4 înregistrări nu sunt datate).

Corpus POP IV (2011) = Pop, L., M. Duma, C. Pașcalău (ed.), *Façons de parler.ro*, Cluj-Napoca, Echinox. Conține 22 de transcrieri (= 97 p.). Înregistrările au fost făcute în perioada 2007-2010.

La aceste volume, pot fi adăugate:

Corpus POP II (2008) = Pop, L. (ed.), *La langue virtuelle. Recherches sur les forums des jeunes*, Cluj-Napoca, Echinox. Conține 15 decupaje tematice din româna din spațiul virtual (= 135 p.), datate în perioada 2006-2007.

AF I (1988) = Avram, A. (coord.), *Antologie fonetică a limbii române*, București, Institutul de Cercetări Etnologice și Dialectale. Acesta este un corpus fonologic. Conține 38 de texte (= 25 p.) de română vorbită în secolul al XX-lea (limita superioară fiind anul 1976).

Aceste corpusuri sunt disponibile doar în formatul print și nu sunt însoțite de înregistrările materialului audio(-video) transcris, nici de o variantă electronică a textelor. Volumul Gheorghe, Măda, Săftoiu (2009) constituie o excepție.

Înregistrările sunt transcrise cu grafia standard, deci lectura este larg accesibilă. Sunt înregistrate diverse informații prozodice și paralingvistice și, de asemenea, sunt codificate informații legate de oralitatea interacțiunii verbale (repetiții, ezitări, reveniri, suprapuneri etc.).

În general, varianta ilustrată este româna standard sau de uz mediu, nedialectală, nepopulară, vorbită de adulți cu un nivel mediu de educație (CORV 2002, IVLRA 2002, IV II 2007, CLRVA II 2013 etc.). Corpusurile POP I (2004), POP II (2008) și POP III (2010) sunt trilingve; acestea pun în contrast româna, engleza și franceza. Corpusul mixt (scris și oral) POP IV (2011) ilustrează câteva modalități de comunicare specifice românilor: bârfa, gâlceava, scenarita, zeflemeaua etc. Româna standard, urbană din Republica Moldova și Ucraina este ilustrată în Bochmann, Dumbrava (2000), respectiv Bochmann (2004).

În majoritatea volumelor, spațiul destinat interacțiunii publice depășește spațiul consacrat interacțiunii private. De asemenea, limba transmisă (prin radio, televiziune sau prin canalul YouTube) este privilegiată în raport cu comunicarea directă, față în față. Tipologia și proporția tipurilor de înregistrări transcrise variază de la un volum la altul. De exemplu: în CLRVA II (2013), din totalul de 22h,6′ de înregistrări transcrise, discursul religios ocupă 10h; în CORV (2002), monologul reprezintă a treia parte din totalul duratei înregistrărilor (1h,6′ din 3h,9′); în Bochmann, Dumbrava (2000), din 13h,39′ de înregistrări transcrise, interacțiunea în context instituțional (școală, universitate, policlinică) ocupă 7h,27′, româna transmisă prin radio și televiziune, 2h,36′, iar interacțiunea din context neinstituțional (pe stradă, la piață, în familie, în campusul universitar), 3h,35′.

4. Posibile aplicații ale corpusurilor de limba română în elaborarea de materiale didactice pentru limba română ca limbă nenativă

După cum s-a putut vedea și în secțiunea 2.2 de mai sus, corpusurile pot fi utilizate în foarte multe moduri în activitățile care implică o limbă străină/limba a doua. În continuare, sunt date câteva exemple practice pentru română.

4.1. Cele mai frecvente cuvinte

La actualizarea inventarului vocabularului minimal al românei contemporane, în afară de alte metode și criterii, utilizarea indicațiilor de frecvență dintr-un corpus reprezentativ este indispensabilă. Cel mai potrivit corpus electronic care poate fi folosit este ROMBAC.

Structura stilistică a corpusului electronic ROMBAC este următoarea (Ion, Irimia, Ștefănescu, Tufiș 2012: 343):

Texte	Fraze	Ocurențe	Cuvinte (lexicale + funcționale)	Cuvinte lexicale (vb., subst., adj., adv.)
Jurnalistice	651.872	10.294.016	8.558.619	4.662.528
medicale + farmaceutice	603.161	10.950.271	9.163.029	5.226.837
Juridice	659.646	9.067.516	7.482.484	4.247.737
biografice + critică literară	314.368	5.802.961	4.298.493	2.567.427
literare	517.803	8.002.596	6.773.648	3.531.156
Total	2.746.850	44.117.360	36.276.273	20.235.685

S-au extras cinci liste cu cele mai frecvente 3.000 de cuvinte (ca intrări de dicționar), însoțite de numărul de ocurențe; aceste liste corespund celor cinci tipuri de texte din tabelul anterior. Lipsesc textele orale transcrise, care ar fi dat un profil mai corect al românei contemporane standard.

Astfel, primele 50 de cuvinte și sintagme fixe din fiecare tip de texte sunt:

(i) Texte jurnalistice (8.558.619 cuvinte / 10.294.016 ocurențe): fi, 61.841 de ocurențe; an, 30.257; putea, 25.444; avea, 20.527; nr., 18.647; oră, 16.811; persoană, 14.008; face, 12.082; zi, 12.029; dată, 11.523; lună, 11.362; zonă, 11.119; program, 10.115; mare, 10.062; țară, 9.861; perioadă, 9.760; vinde, 9.707; național, 9.668; nou, 9.560; public, 8.985; cameră, 8.969; privi, 8.946; proiect, 8.912; oferi, 8.819; serviciu, 8.405; casă, 8.166; afla, 8.159; două, 8.117; județ, 8.017; loc, 7.955; român, 7.882; activitate, 7.830; organiza, 7.825; apartament, 7.789; caz, 7.649; centru, 7.575; curs, 7.554; începe, 7.466; muncă, 7.421; lucrare, 7.383; preț, 7.347; trebui, 7.259; lege, 6.830; avea loc, 6.693; urma, 6.687; executa, 6.426; str., 6.384; local, 6.320; desfășura, 6.313; astfel, 6.296.

(ii) Texte literare (6.773.648 de cuvinte / 8.002.596 de ocurențe): fi, 125.710; avea, 31.241; putea, 30.664; face, 29.568; da, 18.439; vedea, 16.894; spune, 16.207; ști, 15.213; om, 15.098; mare, 14.864; trebui, 12.721; zice, 12.561; numai, 11.841; lua, 9.826; veni, 9.461; bine, 8.840; crede, 8.372; vrea, 8.311; acum, 8.192; mult, 8.175; ochi, 8.116; zi, 8.058; așa, 8.058; viață, 8.001; părea, 7.963; chiar, 7.937; an, 7.868; lume, 7.746; pune, 7.719; țară, 7.227; apoi, 7.178; rămâne, 6.995; atunci, 6.966; mână, 6.950; lucru, 6.889; atât, 6.840; trece, 6.791; sta, 6.647; lăsa, 6.618; începe, 6.447; care, 6.424; singur, 6.400; vorbi, 6.315; duce, 6.102; uita, 6.084; privi, 6.083; timp, 6.015; simți, 5.601; bun, 5.585; ajunge, 5.488.

(iii) Texte biografice și de critică literară (4.298.493 de cuvinte / 5.802.961 de ocurențe): fi, 39.074; literar, 14.963; român, 10.685; literatură, 10.062;

revistă, 9.900; poezie, 8.219; românesc, 7.573; an, 6.909; viață, 6.860; avea, 6.614; apărea, 6.388; scriitor, 6.299; volum, 6.233; carte, 6.050; face, 6.030; roman, 5.903; vers, 5.871; ș.a., 5.471; istorie, 5.268; autor, 5.216; studiu, 5.025; mare, 4.989; poet, 4.979; proză, 4.979; limbă, 4.797; publica, 4.448; articol, 4.327; scrie, 4.191; putea, 4.155; text, 4.088; lume, 4.053; nou, 3.917; redactor, 3.816; deveni, 3.596; apoi, 3.595; urma, 3.571; operă, 3.559; critic, 3.539; cultural, 3.517; cultură, 3.505; număr, 3.471; artă, 3.468; om, 3.441; teatru, 3.437; timp, 3.377; universitate, 3.320; școală, 3.320; cronică, 3.303; liceu, 3.277; facultate, 3.269.

(iv) Texte juridice (7.482.484 de cuvinte / 9.067.516 ocurențe): fi, 48.108; articol, 38.505; regulament, 33.329; caz, 29.493; trebui, 29.291; privi, 27.136; prezent, 26.828; putea, 25.681; alineat, 25.508; stat membru, 21.006; prevedea, 20.267; produs, 19.696; comisie, 18.531; directivă, 16.070; dată, 15.856; avea, 14.601; măsură, 14.499; punct, 14.191; anexă, 13.881; aplica, 13.445; parte, 13.440; decizie, 13.165; acord, 12.895; stabili, 12.844; aplicare, 12.333; consiliu, 12.303; valoare, 11.925; autoritate, 11.757; condiție, 11.340; informație, 11.320; următor, 11.290; respectiv, 11.274; necesar, 11.171; perioadă, 11.075; dispoziție, 10.586; bază, 10.584; menționa, 10.031; adopta, 9.523; comunitar, 9.365; vedere, 8.929; cauză, 8.876; comunitate, 8.794; prezenta, 8.638; țară, 8.464; face, 8.435; sistem, 8.144; în special, 8.021; procedură, 7.871; an, 7.801; utiliza, 7.780.

(v) Texte medicale și farmaceutice (9.163.029 de cuvinte / 10.950.271 de ocurențe): fi, 105.546; pacient, 82.320; trebui, 63.053; doză, 59.796; putea, 55.097; tratament, 52.417; medicament, 44.114; studiu, 36.210; mg, 34.681; administrare, 32.592; reacție, 27.104; administra, 26.502; utiliza, 26.420; caz, 24.068; medic, 22.256; avea, 22.049; insulină, 21.411; clinic, 20.264; advers, 20.003; vedea, 19.766; utilizare, 17.879; valoare, 16.721; săptămână, 16.572; creștere, 16.426; efect, 16.217; conține, 16.088; observa, 15.529; necesar, 15.262; trata, 15.185; privi, 15.176; apărea, 15.071; timp, 14.879; mare, 14.855; lua, 14.362; concentrație, 14.271; zi, 14.052; risc, 13.925; prezenta, 13.174; dată, 13.064; informație, 12.729; crește, 12.424; copie, 11.863; renal, 11.850; frecvent, 11.547; oră, 11.526; recomanda, 11.402; special, 11.374; plasmatic, 11.150; comprimat, 11.098; soluție, 11.066.

Cele mai frecvente 15 cuvinte în cele cinci tipuri de texte sunt:

texte jurnalistice	texte literare	biografii + critică literară	texte medicale + farmaceutice	texte juridice
fi, 61.841; an, 30.257; putea, 25.444; avea, 20.527; nr., 18.647; oră, 16.811; persoană, 14.008; face, 12.082; zi, 12.029; dată, 11.523; lună, 11.362; zonă, 11.119; program, 10.115; mare, 10.062; țară, 9.861.	fi, 125.710; avea, 31.241; putea, 30.664; face, 29.568; da, 18.439; vedea, 16.894; spune, 16.207; ști, 15.213; om, 15.098; mare, 14.864; trebui, 12.721; zice, 12.561; numai, 11.841; lua, 9.826; veni, 9.461.	fi, 39.074; literar,14.963; român, 10.685; literatură, 10.062; revistă, 9.900; poezie, 8.219; românesc, 7.573; an, 6.909; viață, 6.860; avea, 6.614; apărea, 6.388; scriitor, 6.299; volum, 6.233; carte, 6.050; face, 6.030.	fi, 105.546; pacient, 82.320; trebui, 63.053; doză, 59.796; putea, 55.097; tratament, 52.417; medicament, 44.114; studiu, 36.210; mg, 34.681; administrare, 32.592; reacție, 27.104; administra, 26.502; utiliza, 26.420; caz, 24.068; medic, 22.256.	fi, 48.108; articol, 38.505; regulament, 33.329; caz, 29.493; trebui, 29.291; privi, 27.136; prezent, 26.828; putea, 25.681; alineat, 25.508; stat membru, 21.006; prevedea, 20.267; produs, 19.696; comisie, 18.531; directivă, 16.070; dată, 15.856.

Fără a fi suficiente singure pentru a stabili cuvintele cele mai relevante care trebuie introduse în etape în predare-învățare, aceste liste de frecvență sunt indispensabile în elaborarea de materiale profesioniste pentru limba română ca limbă străină. Pentru moment, pentru limba română, inventarul din *Vocabularul minimal al limbii române curente* (ediția a treia din 1994, refăcută în întregime) poate fi folosit foarte eficient, deși autoarele (M. Iliescu și A. Costăchescu) nu au beneficiat de ajutorul cercetării informatizate. Totuși, există necesitatea ca aceste liste de frecvență să fie mereu aduse la zi, lucru destul de ușor de realizat cu ajutorul unui corpus-monitor (vezi supra 2.1.3, criteriul 5). De exemplu, în ediția din 1994, în dicționarul menționat au fost introduse cuvinte care făceau trimitere la realități noi sau devenite mult mai importante, precum *cec, cont, delegat, demagog, implementare, parlament, poliție*, fiind eliminate cuvinte ca *milițian* și *miliție* (vezi *Cuvânt-înainte*, p. 7). O listă actualizată ar trebui, probabil, să conțină elemente noi, precum *euro* (al 471-lea cuvânt, ca frecvență, în corpusul jurnalistic), *internet* (al 495-lea, în același corpus), *mobil* (al 865-lea), *romano-catolic* (al 871lea), *asistent(ă)* (al 1.194-lea), șomaj (al 1.201-lea), *rom* (al 1.558lea, cu omonimia nume etnic - băutură),

patron (al 2.613lea), poate și *facebook, carte de identitate, ajutor social, telenovelă, anticoncepționale*; în schimb, ar trebui probabil eliminate din lista celor mai frecvente cuvinte potrivite pentru a fi învățate de străinii începători elemente precum *telegramă, telegrafia, tutun, spanioloaică*.

Pe baza informațiilor despre frecvența cuvintelor într-o limbă, se poate aproxima mai bine repartizarea lexicului pe niveluri și, în consecință, acesta poate fi introdus progresiv mai eficient. De asemenea, pot fi identificate solidaritățile lexicale (expresii, locuțiuni, compuse etc.) cele mai frecvente; de exemplu, în lista de 3.000 de unități extrase din subcorpusul jurnalistic ROMBAC se numără și: *avea loc, cel puțin, în prezent, de asemenea, în vârstă, mai ales, în special, în/din urmă, punct de vedere, în continuare, cel mult, din față, pe termen lung, din nou, de curând, în comun, la cerere, în general, în plus, de exemplu, pentru prima dată, de fapt, pe loc, de altfel, pe zi* etc. Foarte frecvente, și deci de luat în seamă, sunt următoarele abrevieri și sigle: *nr., str., S.R.L., dvs., O.U.G., H.G.R., ap.* (din textele jurnalistice); *etc., pref., ș.a.m.d., ș.a.* (din textele academice).

Lexemele cel mai des utilizate pot fi folosite ca exemple în componenta de gramatică. O observație punctuală se referă la frecvența foarte mare a substantivelor care denumesc unități de timp (*oră, minut, secundă, zi, an* etc.), a căror marcare pentru gen și număr pune probleme străinilor și care ar trebui exersate mai mult chiar la nivelul A1.

O listă de frecvență a cuvintelor în româna standard este, deci, indispensabilă în construirea de dicționare pentru cei care învață limba română. Există destul de multe dicționare destinate străinilor care învață româna, de exemplu: Lombard, Gâdei (1981), Costăchescu, Iliescu (1994), Barbu (2009) —dicționare morfologice; Ionescu, Steriu (1999), Drăghicescu (coord., 2002) — dicționare sintactice; Biriș, Burlacu, Șoșa (2013) — dicționar semantic. Eficiența folosirii acestor dicționare (de altfel, din păcate, în general greu accesibile sau total necunoscute) va fi sporită prin aducerea la zi a intrărilor selectate și a informațiilor (lexico-semantice și sintactice) oferite.

4.2. Descrierea mai precisă a diverselor stiluri funcționale

Prin cercetări statistice făcute pe corpusuri electronice (vezi, de exemplu, Ion, Irimia, Ștefănescu, Tufiș 2012), se pot stabili cu precizie anumite caracteristici ale stilurilor funcționale. De exemplu:

(i) Tipul de lexic folosit (terminologie, solidarități lexicale, conectori speciali etc.).

(ii) Lungimea medie a frazei și tipul de punctuație caracteristic. De exemplu, textele biografice și de critică literară se disting prin lungimea frazei și prin utilizarea frecventă a virgulei și a semnului punct și virgulă (Ion, Irimia, Ștefănescu, Tufiș 2012: 343).

(iii) Ponderea cuvintelor lexicale în raport cu cele funcționale. În texte-

le de ficțiune, raportul este de aproximativ 1:1, în timp ce în cele biografice și de critică literară, cuvintele funcționale sunt mai rare, 1:5; în textele medicale și în cele juridice, ponderea este asemănătoare, 1:3 (Ion, Irimia, Ștefănescu, Tufiș 2012: 343).

(iv) Ponderea categoriilor lexicale specifice. Se observă că textele de ficțiune se diferențiază prin frecvența mai mare a verbelor și adverbelor, în timp ce textele biografice și de critică literară, prin frecvența substantivelor și a adjectivelor. De asemenea, se remarcă numărul mare de adjective din textele medicale și faptul că ponderea claselor lexicale este similară pentru textele jurnalistice și cele juridice (Ion, Irimia, Ștefănescu, Tufiș 2012: 343).

4.3. Descrierea unor probleme gramaticale specifice

Cercetarea pe baza corpusului a unor probleme gramaticale punctuale ale unei limbi poate da rezultate importante pentru descrierea mai precisă a problemei respective și, în continuare, pentru prezentarea acestei probleme nenativilor care învață această limbă. Mai jos sunt selectate câteva informații obținute din analiza de corpus de română contemporană pentru concurența formelor de viitor (Mîrzea Vasile 2012) și pentru valorile prezumtivului (Siminiciuc 2016). Cele două autoare au folosit corpusuri întocmite personal.

4.3.1. Concurența formelor de viitor românesc

În legătură cu utilizarea formelor de viitor standard (tipul *voi pleca*) și colocvial (tipurile *o să plec* și *am să plec*), dintre informațiile oferite de analiza unui corpus de texte beletristice și eseistice, jurnalistice, juridice, științifice și de română vorbită merită selectate câteva (Mîrzea Vasile 2012):

(i) Viitorul de tipul *voi pleca* apare și în varianta orală a românei, inclusiv în româna neliterară vorbită.

(ii) Viitorul colocvial nu este folosit în textele juridice și în cele științifice.

(iii) Dintre cele două tipuri de viitor colocvial, este mai frecvent tipul *o să plec*.

(iv) În cazul viitorului colocvial de tipul *am să plec*, cele mai frecvente sunt formele de persoana I sg., a IIa sg., formele de persoana I pl. și a IIa plural nefiind utilizate, iar cele de persoana a IIIa fiind rare. În multe situații, valoarea modală a acestei perifraze se suprapune peste cea temporală.

(v) Viitorul anterior cu valoare strict temporală este rar utilizat, valoarea formelor de tipul *voi fi plecat* fiind, în general, modală (de prezumtiv).

(vi) În româna orală, în contextul adverbului *mâine*, formelor marcate de viitor le este preferat net prezentul indicativ.

(vii) Perifrazele negramaticalizate cu valoare de viitor de tipul *aveam să plec* și *urma să plec* (formate pe baza imperfectului verbului *a avea* și *a*

urma) sunt suficient de frecvente pentru a nu fi ignorate în descrierea viitorului românesc. Primul tip este specific stilului beletristic. Al doilea tip este frecvent în limbajul jurnalistic.

4.3.2. Aspecte privind prezumtivul

În prezentarea valorilor prezumtivului străinilor care învață română, dintre cele șase valori identificate de Siminiciuc (2016), în acord cu valorile viitorului epistemic din franceză și italiană (valoarea epistemică de incertitudine, valoarea concesivă, valoarea de întărire, valoarea de „împrumut", valoarea enunțiativă și valoarea temporală), ar trebui insistat asupra a trei valori:

(1) *Nu s-o fi trezit, de întârzie atât! Sau n-o fi mai venind deloc*. (incertitudine)
(2) *O fi (fiind) ea pisică, dar nu-i place peștele*. (concesie)
(3) *Am muncit 40 de ani în uzină! Oi fi având și eu dreptul la pensie acum, nu? Mi-o fi ajuns și mie!* (valoare de întărire)

De asemenea, relevante sunt și informațiile privind frecvența verbelor care sunt folosite la prezumtiv (în ordinea frecvenței: *a fi, a avea, a vrea, a putea*, apoi alte verbe) și contextele fixate (*Ce-o fi o fi; Fie cum o fi; Ce-o da Dumnezeu; Când o fi mai rău, așa să ne fie; C-o fi, c-o păți* etc.), în care este vizibilă valoarea temporală specifică viitorului din care derivă prezumtivul.

4.4. Verificarea și rafinarea descrierii grupurilor sintactice

Un dicționar sintactico-semantic de verbe (dar și un dicționar explicativ general, cu atât mai mult unul pentru nenativi) trebuie să se bazeze pe datele pe care le oferă un corpus, pentru aspecte precum: grila sintactico-semantică a verbului; frecvența pozițiilor obligatorii dependente; sinonimia pozițiilor sintactice obligatorii și preferința pentru una sau alta; exemplificarea prin cele mai frecvente unități lexicale, exemple naturale.

De exemplu, descrierea verbului *a (se) ocupa* într-un dicționar de tipul celui menționat, realizată mai mult sau mai puțin intuitiv, având ca bază de pornire DEX, trebuie verificată cu ajutorul unui corpus reprezentativ sau monitor.

1. GN [N, + persoană / + grup /+ instituție]
   GN [Ac, + teritoriu / + populație]

Sens: „a cuceri": *Turcii au ocupat Țara Românească. Armata rusă i-a ocupat pe români.*

2. GN [N, + persoană / + grup / + instituție]
   GN [Ac, + spațiu (locativ)]

Sens: „a folosi (temporar)": *Firma ocupă etajul doi al clădirii. Ion și-a ocupat un loc în sală.*

3. GN [N]
   GN [Ac, + suprafață]

Sens: „a se întinde": *Viile ocupă cea mai mare parte din teritoriul cultivat.*

4. GN [N, + persoană]
   GN [Ac, + post / + funcție]

Sens: „a lua în primire, a deține": *Ion ocupă funcția de director.*

5. V [se]
   GN [N, + persoană / + instituție]
   GP [*cu*, + activitate]

Sens: „a se îndeletnici, a lucra într-un anumit domeniu": *Ion / firma se ocupă cu recrutarea de personal. Țăranii se ocupă cu agricultura.*

6. V [se]
   GN [N, + persoană / + instituție]
   GP [*de*]

Sens: „a avea grijă, a se îngriji, a avea ca responsabilitate": *Mă ocup de un bolnav / de grădiniță / de strângerea de fonduri.*

Exemplul este extras din materialul lucrat în cadrul unui grant care a avut ca scop realizarea unui dicționar de aproximativ 3.000 de verbe, pentru fiecare verb fiind indicate valențele combinatorii și sensurile principale (*Bază de date sintactico-semantică în format XML: valențele combinatorii ale verbelor românești în reprezentare HPSG*, grant CNCSIS nr. 1156/2005, coord.: Ana-Maria Barbu, 2005-2007).

5. În loc de concluzii

Prin această lucrare s-a dorit, pe de o parte, să se aducă în atenție o metodă nelipsită din studiul limbilor de cultură importante, inclusiv în descrierea lor pentru nenativi și în elaborarea de materiale didactice și, pe de altă parte, să se facă cunoscute corpusurile românești, acestea fiind puțin utilizate în cercetările lingvistice. Chiar dacă pentru română nu există încă un corpus reprezentativ întocmit după criteriile consacrate în bibliografie (conținând și texte orale transcrise, printre altele), utilizarea corpusurilor electronice existente, ca și a celor în format print, poate ajuta la profesionalizarea activității de elaborare a materialelor didactice și de predare/învățare în timpul orelor de curs.

Cele câteva exemple de aplicații făcute pe baza corpusurilor românești

— lista celor mai frecvente cuvinte, caracteristicile distinctive ale românei în funcție de tipul de stil funcțional, utilizarea formelor de viitor standard și colocvial, valorile prezumtivului — arată utilitatea cercetărilor de acest tip pentru domeniul pedagogiei românei ca limbă nenativă.

Bibliografie:

AVRAM, Andrei (coord.), 1988, *Antologie fonetică a limbii române (AF I)*, Bucureşti, Institutul de Cercetări Etnologice și Dialectale;

BARBU, Ana-Maria, 2009, *Conjugarea verbelor românești. Dicționar: 7500 de verbe românești grupate pe clase de conjugare*, ed. a 5-a, București, Editura Coresi ;

BARBU, Ana-Maria (coord.), 2005-2007, *Bază de date sintactico-semantică în format XML: valențele combinatorii ale verbelor românești în reprezentare HPSG*, grant CNCSIS nr. 1156/ 2005;

BARBU MITITELU, Verginica, Elena IRIMIA, Dan TUFIȘ, 2014, „CoRoLa – The Reference Corpus of Contemporary Romanian Language", in Nicoletta CALZOLARI (conference chair), Khalid CHOUKRI, Thierry DECLERCK *et alii* (ed.), *Proceedings* LREC '14, ELRA, pp. 1235-1239 (http://www.lrec-conf.org/proceedings/lrec2014/index.html);

BIBER, Douglas, Susan CONRAD, Geoffrey LEECH, 2002, *Longman student grammar of spoken and written English workbook*. Harlow, Pearson Education;

BIRIȘ, Gabriela, Diana-Viorela BURLACU, Elisabeta ȘOȘA, 2013, *Antonime, sinonime, analogii. Vocabular minimal al limbii române (cu traducere în limba engleză)*, Cluj-Napoca, Casa Cărții de Știință;

BOCHMANN, Klaus (ed.), 2004, *Gesprochenes Rumänisch in der Ukraine. Soziolinguistische Verhältnisse und linguistische Strukturen*, Leipzig, Leipziger Universitätsverlag;

BOCHMANN, Klaus, Vasile DUMBRAVA (ed.), 2000, *Limba română vorbită în Moldova istorică*, Vol. 2. *Texte*, Leipzig, Leipziger Universitätsverlag;

BONELLI- TOGNELLI, Elena, John SINCLAIR, 2006, „Corpora", in Keith BROWN (ed.), *Encyclopedia of language and linguistics*, Amsterdam, Elsevier, pp. 206-219;

BROOK-HART, Guy, 2009, *Learning from common mistakes*, Cambridge, Cambridge University Press;

Corpusul ZiareRom – http://www.lingv.ro;

CRISTEA, Dan, 2005, „Resurse lingvistice și tehnologii ale limbajului natural. Cazul limbii române", *Prelegerile Academiei Române. Filiala Iași* (http://profs.info.uaic.ro/~dcristea/papers/cristea-prelegeri.pdf), 23 p.;

CRYSTAL, David, 2008, *A dictionary of linguistics and phonetics*, 6th ed., Blackwell Publishing;

DASCĂLU JINGA, Laurenția, 2002, *Corpus de română vorbită (CORV).*

Eșantioane, București, Oscar Print;

DASCĂLU JINGA, Laurenția (coord.), 2011, *Româna vorbită actuală (ROVA). Corpus și studii*, Academia Română, Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti", București, Editura Academiei Române;

DE POERCK, Guy, Louis MOURIN, 1961-1964, *Introduction à la morphologie comparée des langues romanes, basée sur des traductions anciennes des Actes des Apôtres, ch. XX à XXIV*, tome I : *Ancien portugais* [par L. Mourin] *et ancien castillan* [par G. De Poerck], 1961, tome IV : *Sursilvain et engadinois anciens, et ladin dolomitique*, par L. Mourin, 1964, tome VI : *Ancien roumain*, par L. Mourin, 1962, Rijksuniversiteit Gent, Handboeken uitgegeven door de Faculteit van de Letteren en Wijsbegeerte, 1e/4e/6e Aflevering, Bruges, De Tempel, tome II (partim) : *Ancien catalan*, par L. Mourin, 1961, tome V (partim) : *Sarde*, par L. Mourin, 1963, Bruxelles, Presses universitaires;

DRĂGHICESCU, Janeta (coord.) 2002, *Dicționar de construcții verbale român-francez-italian-englez*, Craiova, Editura Universitaria;

GHEORGHE, Mihaela, Stanca MĂDA, Răzvan SĂFTOIU, 2009, *Comunicarea la locul de muncă. Corpus de interacțiune verbală în mediul profesional* (+ CD), Brașov, Editura Universității Transilvania din Brașov;

GRANGER, Sylviane, 2002, „A Bird's-eye view of learner corpus research", in Sylviane GRANGER, Joseph HUNG, Stephanie PETCH-TYSON (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam, Philadelphia, John Benjamins Publishing Company, pp. 3-33;

HEGER, Klaus (ed.), 1967, *Die Bibel in der Romania: Matthäus 6, 5-13*, Tübingen, Niemeyer;

HEINIMANN, Siegfried (ed.), 1988, *Oratio Dominica Romanice. Das Vaterunser in den romanischen Sprachen von den Anfängen bis ins 16. Jahrhundert mit den griechischen und lateinischen Vorlagen herausgegeben und eingeleitet von… ,* Tübingen, Niemeyer;

HOARȚĂ CĂRĂUȘU, Luminița (coord.), 2005, *Corpus de limbă română vorbită actuală*, Iași, Editura Tehnică, Științifică și Didactică CERMI;

HOARȚĂ CĂRĂUȘU, Luminița (coord.), 2013, *Corpus de limbă română vorbită actuală nedialectală*, Iași, Editura Universității „Alexandru Ioan Cuza";

HUNSTON, Susan, 2006, „Corpus linguistics", in Keith Brown (ed.), *Encyclopedia of language and linguistics*, Amsterdam, Elsevier, pp. 234-248;

ILIESCU, Maria, Adriana COSTĂCHESCU, 1994, *Vocabularul minimal al limbii române curente. Cu indicații gramaticale complete. Tradus în germană, franceză, italiană, spaniolă*, ediția a treia, București, Editura Demiurg;

ION, Radu, Elena IRIMIA, Dan ȘTEFĂNESCU, Dan TUFIȘ, 2012, „ROMBAC: The Romanian Balanced Annotated Corpus", in Nicoletta CALZOLARI (conference chair), Khalid CHOUKRI, Thierry DECLERCK *et*

alii (ed.), *Proceedings* LREC '14, ELRA, pp. 1235-1239 (http://www.lrec-conf.org/proceedings/lrec2014/index.html);

IONESCU, Adriana, Maria STERIU, 1999, *Verbul românesc. Dicționar sintactic*, ed. a II-a, București, Editura Universității din București;

IONESCU-RUXĂNDOIU, Liliana (coord.), 2002, *Interacțiunea verbală în limba română actuală. Corpus (selectiv). Schiță de tipologie*, București, Editura Universității din București;

IONESCU-RUXĂNDOIU, Liliana (coord.), 2007, *Interacțiunea verbală (IV II). Aspecte teoretice și aplicative. Corpus*, București, Editura Universității din București;

LOMBARD, Alf, Constantin GÂDEI, 1981, *Dictionnaire morphologique de la langue Roumaine : permettant de connaître la flexion entière des mots qui en possèdent une : substantifs, adjectifs, pronoms, verbes*, București, Editura Academiei, Lund, CWK Gleerup;

McENERY, Tony, Richard XIAO, 2010, „What corpora can offer in language teaching and learning", in Eli HINKEL (ed.), *Handbook of research in second language teaching and learning*, vol. 2. London, New York, Routledge, pp. 364-380;

McENERY, Tony, Richard XIAO, YuKio TONO, 2006, *Corpus-based language studies: An advanced resource book* (secțiunea A. 10.8. Language learning and teaching), London, New York, Routledge, pp. 97-103;

MÎRZEA VASILE, Carmen, 2012, „Utilizarea viitorului în limba română actuală. Note pe baza studiului de corpus", in Elena PLATON, Antonela ARIEȘAN (ed.), *Noi perspective în abordarea românei ca limbă străină – ca limbă nematernă*. Cluj-Napoca, Casa Cărții de Știință, pp. 103-114;

MÎRZEA VASILE, Carmen, 2015, „Anthologies et corpus pan-romans", in Maria ILIESCU, Eugeen ROEGIEST (ed.), *Manuel des anthologies, corpus et textes romans*, Berlin, De Gruyter Mouton (Manuals of Romance Linguistics 7), pp. 9-33;

O'KEEFFE, Anne, Michael McCARTY (ed.), 2010, *The Routledge handbook of corpus linguistics* (secțiunile V. Using a corpus for language pedagogy and methodology, VI. Designing corpus-based materials for the language classroom), London, Routledge pp. 317-384, 385-487;

O'KEEFFE, Anne, Michael McCARTY, Ronald CARTER, 2007, *From corpus to classroom: language use and language teaching*, Cambridge, Cambridge University Press;

POP, Liana (ed.), 2004, *Verba Volant. Recherche sur l'oral* (Seria „Atelier"), Cluj-Napoca, Echinox;

POP, Liana (ed.), 2008, *La langue virtuelle. Recherches sur les forums des jeunes* (Seria „Atelier"), Cluj-Napoca, Echinox;

POP, Liana (ed.), 2010, *Où va la communication* (Seria „Atelier"), Cluj-Napoca, Echinox;

POP, Liana, Melania DUMA, Cristian PAȘCALĂU (ed.), 2011, *Façons de parler.ro* (Seria „Atelier"), Cluj-Napoca, Echinox;

RÖMER, Ute, 2008, „Corpora and language teaching", in Anke LÜDELING, Kytö MERJA (ed.), *Corpus linguistics. An international handbook (volume 1),* Berlin, Mouton de Gruyter, pp. 112-130;

RUNDELL, Michael, 2007, *Macmillan English dictionary advanced learner*. MacMillan;

SCHMITT, Diane, Norbert SCHMITT, 2011, *Focus on Vocabulary 2: Mastering the academic word list*. White Plains, NY, Pearson Education;

SIMINICIUC, Elena, 2016, „Le présomptif roumain à la lumière d'une étude quantitative de corpus journalistique", in Actes du colloque international *Normes et grammaticalisation*, Sofia, 21–22 novembre 2015 (ms);

SINCLAIR, John McH. (ed.), 2004, *How to use corpora in language teaching*, Amsterdam, Philadelphia, John Benjamins Publishing Company;

STUBBS, Michael, 2013, „Language corpora", in Alan DAVIES, Catherine ELDER (ed.), *The handbook of applied linguistics*. Blackwel Publishing Ltd, pp. 106-132;

TUFIȘ, Dan, Elena IRIMIA, 2006, „RoCo_News – A hand validated journalistic corpus of Romanian", in *Proceedings of* LREC 2006, pp. 869-872;

WILLIS, Jane, Dave WILLIS, 1989, *Collins COBUILD English course*, London, Harper Collins.

***1998, *Dicționarul explicativ al limbii române (DEX)*, ediția a II-a, București, Univers Enciclopedic;

*** 1997, *Longman Essential Activator*, London, Longman;

Resurse online:

EMEA — *http://www.ema.europa.eu/ema*;

Learner corpora around the world— https://www.uclouvain.be/en-cecl-lcworld.html;

CS III Carmen MÎRZEA VASILE
Institutul de Lingvistică al Academiei Române „Iorgu Iordan-Al. Rosetti", București Facultatea de Litere, Universitatea din București
carmen_marzea@yahoo.fr.

Cercetător la Institutul de Lingvistică al Academiei Române „Iorgu Iordan – Al. Rosetti" din Bucureşti (din 2005, asistent de cercetare; din 2013, cercetător III) şi asistent la Centrul de Studii Româneşti al Universităţii din Bucureşti (din 2011), doctor al Universităţii din Bucureşti (2010). A publicat două cărți despre adverbul românesc, precum și articole şi recenzii în reviste româneşti şi capitole în volume colective apărute în România și în străinătate (la Oxford University Press, Cambridge Scholars Publishing, Mouton de Gruyter). Domenii de interes științific: morfosintaxa (în special a adverbului), lingvistica romanică, schimbarea şi variaţia lingvistică, tipologia, învăţarea/predarea limbii române ca limbă nenativă.