COMUNICAREA ACTUALĂ: ASPECTE SOCIOLINGVISTICE ȘI INTERCULTURALE

CZU: [81:001.891]:004.8

CÂT DE MULT ARE NEVOIE CERCETAREA LINGVISTICĂ DE *BIG DATA* ȘI INTELIGENȚA ARTIFICIALĂ?

Elena UNGUREANU
Institutul de Dezvoltare a Societății Informaționale,
Biblioteca Municipală „B.P. Hasdeu”
https://orcid.org/0000-0002-4326-6744

Rezumat. Studiul *Big data și inteligența artificială în cercetarea lingvistică* (Cluj-Napoca: Casa cărții de știință, 2024), semnat de Cristina Bleorțu și Lavinia Seiciuc (Universitatea „Ștefan cel Mare” din Suceava), reprezintă una din puţinele contribuţii destinate cercetătorului lingvist de generaţie nouă, care reuşeşte să utilizeze abil intrumentele oferite de tehnologiile computaţionale în cercetarea limbajului. În lucrare sunt prezentate 4 lingvistici de corpus de ţară (Spania, Franţa, Italia, Portugalia), la care adaugă o analiză pertinentă a lingvisticii de corpus din România. Toate aceste corpusuri se bazează pe *big data*, care sunt şi vor fi utilizate tot mai mult în analizele pe care le vor livra progresiv instrumentele de inteligenţă artificială. Componenta practică a lucrării se referă la ilustrarea modului de adnotare a scrisorilor din cadrul proiectul online de digitalizare a corespondenței lui Eugeniu Coșeriu cu titlul: *DiLeCos – Digitized Letters to Eugenio Coseriu*, precum și prezentarea unui model de procesare cu programul *R* a unui corpus literar (romanul *Maitreyi* de Mircea Eliade).
Cuvinte-cheie: *cercetare lingvistică tehnologizată, big data, inteligență artificială, lingvistica de corpus din Spania, Franţa, Italia, Portugalia, România, instrumente pentru adnotarea lingvistică.*

Abstract. The study *Big data and artificial intelligence in linguistic research* (Cluj-Napoca: Casa cărții de știință, 2024), signed by Cristina Bleorțu and Lavinia Seiciuc (University „Ștefan cel Mare” in Suceava) represents one of the few contributions intended for the new generation linguistic researcher, who manages to skillfully use the tools offered by computational technologies in language research. The paper presents 4 country corpus linguistics (Spain, France, Italy, Portugal), to which it adds a pertinent analysis of corpus linguistics from Romania. All these corpora are based on big data, which are and will be increasingly used in analyses that artificial intelligence tools will progressively deliver. The practical component of the paper refers to the illustration of the way of annotating letters within the online project of digitizing Eugeniu Coșeriu's correspondence with the title: *DiLeCos – Digitized Letters to Eugenio Coseriu*, as well as the presentation of a model of processing a literary corpus (the novel *Maitreyi* by Mircea Eliade) with the R program.
Key-words: *technological linguistic research, big data, artificial intelligence, corpus linguistics from Spain, France, Italy, Portugal, Romania, tools for linguistic annotation.*

Deschisă cu o dedicaţie adresată acelor cercetători care au „îmblânzit furtunile de date și au transformat textul brut în cunoaștere”, **Cristina Bleorțu** (cerc. șt. III la Universitatea „Ștefan

cel Mare” din Suceava) şi **Lavinia Seiciuc** (prof. univ. abilitat la aceeași universitate) lansează pe piaţa de carte lingvistică o lucrare mai mult decât actuală: *Big data și inteligența artificială în cercetarea lingvistică.* Cluj-Napoca: Casa cărții de știință, 2024, 108 p. [1].

Trebuie să spunem lucruri care s-au mai spus: în ultimele decenii tehnologiile computerizate s-au implântat/implantat deja în procesele de cercetare, deși lumea filologică încă mai păstrează rezerve faţă de ele, dintr-o obişnuinţă prea adâncă. Lucrarea proaspăt apărută e una din puţinele cu o astfel de tematică interdisciplinară (științe ale limbii – tehnologie) şi, de aici, importanța ei tot mai mare. Aflăm, din start, că „lucrarea are următoarele scopuri: 1. Să introducă studenții în domeniul vast al *big data*, evidențiind importanța acestuia în cercetarea lingvistică modernă; 2. Să îi încurajeze pe studenți să utilizeze aceste instrumente pentru crearea unui corpus, care să le permită analiza aprofundată a fenomenelor lingvistice de interes, pornind de la studii de caz concrete realizate cu ajutorul programului R; 3. Să îi pregătească pe studenți pentru abordări interdisciplinare, integrând cunoștințe și metode din diverse domenii pentru a sprijini o cercetare lingvistică complexă și inovatoare” [1, 12].

Fără a se opri la aspecte legate nemijlocit de provocările Inteligenţei Artificiale (cea mai în vogă în prezent, despre care se vor scrie cu siguranţă alte cărţi), autoarele intrigă cititorul prin informaţii de genul capacităţii unor astfel de tenologii de „a replica discursurile unor lideri mondiali, precum Obama sau Trump” [1, 9]. Se menționează că nivelul fără precedent al tehnologiilor de recunoaștere vocală, sintetizare și traducere revoluționează nu doar lingvistica, dar științele umane în general, făcând posibilă analiza și manipularea puternică a limbajului [1, 10].

Cercetările actuale tind să atragă pentru analize conţinut pus la dispoziţie în acces deschis din varii *corpusuri* alcătuite din milioane de cuvinte şi fraze, făcând posibilă interpretarea rapidă a mii şi mii de variabile, fapt de neimaginat acum câteva decenii (variația lingvistică mai cu seamă în corpusuri se vede!). E o regulă pe care savanţii vor să o respecte, spre deosebire de marile corporaţii, care se feresc în general de accesul deschis. *Lingvistica de corpus* (sau *lingvistica corpusurilor*) este aşadar disciplina (metoda) care îşi propune a analiza masive de date sub formă glosare sau texte (scrise sau orale), adunate pentru a reprezenta cât mai fidel limbajul utilizat de vorbitori în diverse situații. Astfel, un corpus poate include articole din ziare, conversații telefonice, postări de pe rețelele sociale sau chiar discursuri politice. Aceste texte sunt apoi stocate sub formă de fișiere pe calculator și analizate cu ajutorul unor programe speciale. Analiza acestor texte ne permite să descoperim modele recurente în utilizarea limbii, cum ar fi frecvența anumitor cuvinte sau expresii în anumite contexte [1, 17]. *British National Corpus*, de exemplu, conține peste o sută de milioane de cuvinte și a fost publicat la începutul anilor 90 [1, 18].

Pentru o poziționare mai clară a lingvisticii de corpus românești în spaţiul romanic, autoarele prezintă o diagramă cu corpusuri digitale în principalele limbi romanice, germană și engleză vizibile în *Virtual Language Observatory* (Observatorul Virtual al Limbilor) [2]. Cititorul este atenţionat că Infrastructura CLARIN [3] (formată în cea mai mare parte din țări europene, iar multe dintre corpusurile juridice reprezintă așa-numitul *acquis comunitar*, care cuprinde legislația, actele juridice și hotărârile judecătorești constituind legea Uniunii Europene) oferă acces la 33 de corpusuri juridice, bogat adnotate. Pentru comparaţie, corpusul românesc din CLARIN are doar 822 de documente în timp ce cel englez – 98 598”, deci de 100 de ori mai

puțin [1, 20]. Nici nu mai e nevoie să explicăm necesitatea urgentă a creșterii capacităților de aliniere la astfel de proiecte europene.

În continuare, lucrarea prezintă informaţii preţioase despre lingvistica de corpus din câteva țări europene:

Lingvistica de corpus în Spania este reprezentată de 19 corpusuri specializate (menţionăm doar câteva): *Biblia Medieval, BIDTEA/DCOST (Biblioteca Digital de Textos de Español Antiguo), CHARTA (Corpus Hispánico y Americano en la Red: Textos Antiguos), CORDE* (*El Corpus Diacrónico del Español*), *CORPES XXI* (*Corpus del Español del sigloXXI*), *EsTenTen* (*Spanish Corpus from the web*) *TeLeMe* (*Tesoro Lexicográfico Médico español*) etc. [1, 23-24];

Lingvistica de corpus în Franța are trecută în revistă în primul rând corpusul *Frantext* (http://www.frantext.fr/), care a constituit baza pentru *Trésor de la langue française*, dar și corpusurile specializate: *FrTenTen* (*Corpus of French Web*), *Clapi* (*Corpus de Langue Parlée en interaction*); *Corpus de Référence du Français parlé, Corpus of spoken French* etc. [1, 25-26];

Lingvistica de corpus în Italia – autoarele menționează primul corpus publicat online în Italia, intitulat *Corpus e Lessico di Frequenza dell'Italiano Scritto* (CoLFIS). Urmează o listă de corpusuri specializate, printre care: CORIS/CODIS (*CORpus di ItalianoScritto*), TLIO (*Tesoro della Lingua Italiana delle Origini,* Corpusul *La Repubblica*, cu articole din anii 1985-2000, cu peste 325 de milioane de cuvinte, dar şi impresionantul *WaCky – The Web-As-Corpus Kool Yinitiative*, cu peste un miliard și jumătate de cuvinte etc. [1, 28]. Nu e de prisos să adăugăm aici un corpus de texte biblice, intitulat INTRATEXT (https://www.intratext.com/), pe care am avut ocazia să-l studiem în câteva rânduri [4].

Lingvistica de corpus în Portugalia – cuprinde desrierea primului proiect elaborat în 1988 la *Centro de Linguística* al Universității din Lisabona: *Corpus de Referência do Português Contemporâneo*, cu peste 300 milioane de cuvinte), dar și câteva corpusuri specializate mai bogate, cum ar fi: *CIPM* (*Corpus Informatizado de Português Medieval*), *CTB* (*Corpus Histórico do português*), *Corpus do português*, cu 45 de miloane de cuvinte, *Tycho Brahe Parsed Corpus of Historical Portuguese* etc. [1, 30].

Lingvistica de corpus în România beneficiază de o descriere mai amplă. Aceasta a demarat mai târziu în comparație cu alte țări romanice. Astfel, *Corpusul Lexicografic Românesc Electronic* (*CLRE*) (un corpus de ediții digitale ale principalelor dicționare și lucrări lexicografice ale limbii române din toate timpurile) a fost inițiat de Academia Română abia în 2014, deci cu o întârziere de cca două decenii față de alte țări europene. Tot în 2014 a debutat și *Corpusul de referință pentru limba română contemporană* (CoRoLa) (https://corola.racai.ro), un proiect ambițios, cu un miliard de cuvinte și 300 de ore de înregistrări audio, acoperind 70 de subdomenii științifice. Existența CoRoLei se datorează eforturilor grupului de cercetători de la Academia Română din Iași, de la Institutul de Inteligență Artificială „Mihai Drăgănescu”. Ţinem să menţionăm faptul că la evenimentul de lansare a Corpusului CoRoLa la București și la Chișinău am avut fericita ocazie să participăm cu câte o comunicare şi să prezentăm publicului interesat cele două lansări [5; 6]. De asemenea, au fost obținute corpusuri importante prin tehnici moderne, cum ar fi *RoTenTen*, care au contribuit la extinderea resurselor lingvistice pentru studiul limbii române. [1, 36].

Capitolul al doilea al cărții este dedicat fenomenului *big data*, creării și folosirii corpusurilor. *Big data* înseamnă abundență de date, masive de date, volume extrem de mari de date și a devenit resursa esențială a secolului al XXI-lea. „*Big data* joacă un rol crucial, spun autoarele, în transformarea domeniului lingvisticii, oferind oportunități unice pentru înțelegerea și modelarea modului în care limbajul este utilizat și evoluează în contextul digital actual. Prin integrarea *big data* în cercetarea lingvistică, se deschid noi orizonturi pentru dezvoltarea unor teorii lingvistice și pentru aplicarea acestora în diverse domenii, de la sociolingvistică până la lingvistica computațională” [1, 40].

„Conceptul de big data este vast și complex, dar, în esență, se referă la acele seturi de date care pot fi colectat masiv și eficient prin intermediul tehnologiilor moderne, precum calculatoarele, telefoanele mobile, rețelele de socializare. Aceste date sunt caracterizate de cinci atribute principale, cunoscute sub denumirea de „cei cinci V”: **volum, viteză, varietate, valoare, veracitate** [1, 41-42]. Clasificarea și analiza *big data* scot în evidenţă faptul că există date structurate, semistructurate și nestructurate. În acest context, să nu uităm de Tim Berners-Lee, cel care a inventat WEB-ul, transformat în cel mai mare corpus de date al prezentului. Adevărat, date neprelucrate, nestructurate. În plus, sarcina cea mai dificilă – datele trebuie interconectate, legate între ele, spunea marele inventator. Nu le scapă, de asemenea, subiectul care privește problemele de etică în procesul creării corpusurilor în contextul big data. Un accent important este menționarea faptului că lingvistica corpusurilor nu mai seamănă cu lingvistica descriptivă tradițională, care era făcută de lingviștii „de fotoliu” (cum îi numește Fillmore, 1992) [1, 53]. Dar anume experiența lor prinde bine lingvisticii care utilizează instrumente digitale, o astfel de colaborare fiind una esențială: „Cele două tipuri de lingviști au nevoie unul de celălalt. Sau, și mai bine, că cele două tipuri de lingviști ar trebui, ori de câte ori este posibil, să coexiste în aceeași persoană” [1, 54].

Ca parte practică a lucrării, autoarele prezintă, pentru transcrierea unui corpus (2.2.1), câteva instrumente performante: EXMARaLDA; F4 (cu Windows) / F5 (cu macOs); ELAN; Whisper [1, 67-59]; instrumentul de transcriere automată a textului de la Microsoft Word (versiunea web) (p. [1, 59]. Surpriza lucrării a fost faptul că, pentru a extrage informațiile specifice dintr-un corpus și a le interpreta corespunzător, coerent și sistematic, un paragraf al studiului a fost dedicat adnotării corpusurilor (2.2.2), prezentând pentru ilustrare modul de adnotare a scrisorilor din cadrul proiectul online de digitalizare a corespondenței lui Eugeniu Coșeriu cu titlul: **DiLeCos – Digitized Letters to Eugenio** Coseriu (https://coseriu.uzh.ch/index.html) [1, 73-75].

Nu în ultimul rând, atenției cititorilor este propus, utilizând drept corpus romanul lui Mircea Eliade *Maitreyi,* un model de procesare a unui corpus cu *R* – un limbaj de programare utilizat pentru *data mining* (mineritul datelor), care se ocupă de separarea textului în cuvinte [1, 78], de frecvenţa cuvintelor; de separarea textului în fraze [1, 92]. Astfel, lucrarea mai constituie un ghid de inițiere în terminologie informatică tot mai necesară breslei lingviștilor, printre termenii care trebuie învățați numărându-se: *standardizrea textului, token* și *tokenizare, tag* și *tagging (etichetare), lemă* și *lematizare* etc.

În concluzie, lucrarea pe care o recomandăm aici pune în discuție aspecte ale unui rol covârșitor pe care îl vor avea în anii ce urmează instrumentele de analiză a datelor în cercetarea lingvistică.

Acest studiu ne-a dat următoarea idee: În Republica Moldova, din câte cunoaştem, nu există la ora actuală preocupări privind elaborarea vreunui corpus reprezentativ. Propunem unul: Corpus al limbii și literaturii române din Basarabia. O echipă în care cercetătorii s-ar (re)profila pe acest domeniu (studierea teoriei şi metodologiei corpusurilor) – crearea unui corpus dedicat sau completarea corpusurilor care sunt elaborate în România, cu un segment important care să fie lizibil și identificabil în corpusul MARE al limbii române, pentru a putea descrie mai pertinent particularitățile limbii române din Basarabia, pentru a se putea opera cu un tezaur de exemple mult mai bogat și mai variat, pentru a se putea regăsi pe o platformă unică texte scrise și semnate de autori din Basarabia laolaltă cu autori din România sau din diaspora. Există numeroase contribuţii care stau pe dinafara corpusurilor – trebuie să stea la un loc.

În concluzie, putem vorbi despre puterea şi povara (ştiinţei) corpusurilor. **PUTEREA** corpusurilor vine, în primul rând, din informația mai **exactă** ca niciodată, mai **rapidă**, mai ușor **verificabilă**, **monitorizabilă**, **cuantificabilă** ș.a.m.d. În al doilea rând, puterea vine, după Berners-Lee, din legarea datelor. Interacțiunea lor permite multiplicarea rapidă a informației, iar cunoașterea anume de acest lucru are nevoie. Pui două cunoștințe alături ca să fie generat un tip nou de cunoștințe. **POVARA** corpusurilor se referă la pregătirea specialiștilor de noua generație, la familiarizarea cu tehnicile de alcătuire și de studiere a corpusurilor. Muncă migaloasă, pe alocuri muncă de bijutier, cere o pregătire pe măsură. Crearea corpusurilor solicită multă dedicație, personal și finanțe.

În aşteptarea următoarelor studii de acest gen, suntem nevoiţi să specificăm următoarele. Deși în adresa AI se fac auzite tot mai multe voci, avizate sau deloc, răspunsul la întrebarea din titlul comunicării noastre este: foarte mult! Și loc pentru discernământul pe care e obligat să-l livreze creierul cercetătorului bine documentat.

BIBLIOGRAFIE

1. BLEORȚU, Cristina, SEICIUC, Lavinia. Big data și inteligența artificială în cercetarea lingvistică. Cluj-Napoca: Casa cărții de știință, 2024, 108 p.
2. VIRTUAL LANGUAGE OBSERVATORY. [Citat 25 noiembrie 2024]. Disponibil: https://www.clarin.eu/content/virtual-language-observatory-vlo
3. CLARIN (European Research Infrastructures for Language Resources and Technology). [Citat 25 noiembrie 2024]. Disponibil: https://www.clarin.eu/.
4. UNGUREANU, Elena. Corpus de texte româneşti din Biblioteca digitală INTRATEXT: un studiu comparativ. In: Limba română scrisă şi vorbită în contextul noilor tehnologii informatice. Realizări şi perspective: Simpozion Aniversar „150 de ani de la înfiinţarea Academiei Române & 25 de ani de la înfiinţarea Secţiei de Ştiinţa şi Tehnologia Informaţiei”. Iaşi, 24 martie 2016. [Citat 1 decembrie 2024]. Disponibil: http://iit.academiaromana-is.ro/RoNLP/index_ro.html.
5. Lansarea Corpusului computaţional de referinţă pentru limba română CoRoLa [Citat 1 decembrie 2024]. Disponibil: https://idsi.md/lansarea-corpusului-computational-de-referinta-pentru-limba-romana-contemporana-CoRoLa.
6. UNGUREANU, Elena. CoRoLa lansată la București și la Chișinău. In: Revista de ştiinţă, inovare, cultură şi artă „Akademos”, 2018, nr. 2(49), p. 18-19. ISSN 1857-0461.