

ACADEMIA ROMÂNĂ

INSTITUTUL DE FILOLOGIE ROMÂNĂ „ALEXANDRU PHILIPPIDE”, IAȘI
INSTITUTUL DE LINGVISTICĂ „IORGU IORDAN – ALEXANDRU ROSETTI”, BUCUREȘTI
INSTITUTUL DE LINGVISTICĂ ȘI ISTORIE LITERARĂ „SEXTIL PUȘCARIU”, CLUJ-NAPOCA

**LEXICOGRAFIA ACADEMICĂ ROMÂNEASCĂ.
PROVOCĂRILE INFORMATIZĂRII
AcadLexi 2025**

Volumul cuprinde lucrări prezentate la ediția a XIV-a a Colocviului cu participare internațională „Lexicografia academică românească. Provocările informatizării”. AcadLexi 2025 (<https://philippide.ro/acadlexi.html>).

Coordonatori de volum: Elena Isabelle Tamba, Mariana Nastasia

Editori:

Alexandru Anghelina
Iulia Barbu-Comaromi
Călina Ciocoiu
Alexandra Ilie
Mircea Minică
Andreea-Nora Pușcaș

Referenți științifici:

Conf. univ. dr. Ana-Maria Minuț
CS I dr. Gabriela Haja

Coperta: Lucian Săcrieru-Dragomir

ISSN: 3008-6159

ISSN-L: 3008-6159

DOI: https://doi.org/10.52257/30086159_2025

© 2025 Editorii volumului. Toate drepturile rezervate.

Universitatea Babeș-Bolyai

Presa Universitară Clujeană

Str. Hasdeu nr. 51

400371 Cluj-Napoca, România

Tel./fax: (+40)-264-597.401

E-mail: editura@ubbcluj.ro

<http://www.editura.ubbcluj.ro/>



ACADEMIA ROMÂNĂ

INSTITUTUL DE FILOLOGIE ROMÂNĂ „ALEXANDRU PHILIPPIDE”, IAȘI
INSTITUTUL DE LINGVICĂ „IORGU IORDAN – ALEXANDRU ROSETTI”, BUCUREȘTI
INSTITUTUL DE LINGVICĂ ȘI ISTORIE LITERARĂ „SEXTIL PUȘCARIU”, CLUJ-NAPOCA

ACTELE COLOCVIULUI INTERNAȚIONAL „LEXICOGRAFIA ACADEMICĂ ROMÂNEASCĂ. PROVOCĂRILE INFORMATIZĂRII” AcadLexi 2025

PRESA UNIVERSITARĂ CLUJEANĂ

2025

CUVINTE ȘI COD: LEXICOGRAFIA REINVENTATĂ DIGITAL

MARIUS-RADU CLIM*

1. Preliminarii

Dicționarele academice începute în secolul al XIX-lea reprezintă adevărate monumente culturale, cuprinzând tezaurul lexical al limbilor respective, dar și o descriere sistematică și destul de cuprinzătoare a întregului vocabular al unei limbi, pe baza observării empirice a utilizării autentice a limbii, în toate stilurile funcționale. Deși astfel de proiecte ambițioase sunt în mare parte realizate deja în majoritatea culturilor europene, rolul dicționarelor academice rămâne fundamental pentru istoria lexicală a unui popor, ele fiind parte din patrimoniul său cultural.

Motivul pentru care există dicționarele este implicit legat de comunicarea interumană. Ele sunt folosite pentru înțelegerea textelor, pentru a redacta un text corect gramatical, pentru a traduce un text, pentru cercetarea unor fenomene lingvistice, pentru înțelegerea variațiilor lexicale în evoluția istorică a limbii sau în creațiile literare, dar mai puțin în comunicarea spontană. Lucrările lexicografice au apărut inițial ca dicționare bilingve care facilitau înțelegerea de texte din alte limbi. Indiferent de scopul utilizatorului, dicționarele au fost de-a lungul secolelor un instrument normativ, necesar, care respectă anumite rigori, convenții, pentru a veni în ajutorul utilizatorilor. Însă, realizarea dicționarelor a constituit întotdeauna o activitate consumatoare de timp și de energie. Proiectele lexicografice majore, precum tezaurele sau dicționarele academice, au necesitat echipe mari de lucru, cât și zeci de ani pentru finalizare. De exemplu, prima ediție a *Oxford English Dictionary* a necesitat 70 de ani (între 1857–1928) sau *Das Deutsche Wörterbuch* realizat în 120 de ani, între 1838 și 1961. De asemenea, pentru publicarea dicționarului-tezaur românesc, *Dicționarul limbii române*, a fost necesar aproape un secol, între anii 1913–1948 și 1965–2010.

În plus, forma de organizare a dicționarelor a fost și ea criticată tocmai din cauza rigidității ei și a faptului că nu reflectă limbajul în situațiile de comunicare. Această structură a dicționarelor a avut ca rol tocmai facilitarea accesului oricărui utilizator la informațiile lexicografice (cum ar fi informații gramaticale, etimologice, de pronunție, citate, variante

* Institutul de Filologie Română „Alexandru Philippide”, Academia Română, Filiala Iași, str. Th. Codrescu, nr. 2.

etc.). Unul dintre criticii acestei structuri a fost și lingvistul Dwight Le Merton Bolinger, profesor de limbi romanice la Universitatea Harvard, care, într-un articol publicat în 1985, susținea următoarele:

Lexicography is an unnatural occupation. It consists in tearing words from their mother context and setting them in rows – carrots and onions and beetroot and salsify next to one another – with roots shorn like those of celery to make them fit side by side, in an order determined not by nature but by some obscure Phoenician sailors who traded with Greeks in the long ago. Half of the lexicographer's labor is spent repairing this damage to an infinitude of natural connections that every word in any language contracts with every other word, in a complex neural web knit densely at the center but ever more diffusely as it spreads outward. A bit of context, a synonym, a grammatical category, an etymology for remembrance' sake, and a cross-reference or two – these are the additives that accomplish the repair (Bolinger 1985: 69).

Trecerea în mediul digital a eliminat, în primul rând, constrângerile de spațiu ale dicționarilor tipărite, nemaexistând limitări pentru stocarea informațiilor legate de cuvinte. În al doilea rând, mediul digital a permis ca dicționarele să nu mai fie dependente nici măcar de ordinea alfabetică în organizarea articolelor lexicografice. Opțiunile de căutare permit interogarea atât prin scrierea cuvântului dorit, cât și prin dictarea sau realizarea unei poze, ordinea alfabetică devenind nerelevantă. De asemenea, căutarea nu se limitează doar la un cuvânt, ci pot fi găsite inclusiv expresii sau chiar unități frazeologice. Și, nu în ultimul rând, accesul la un corpus mare de texte permite ca intrările să poată fi exemplificate cu cât mai multe citate din stiluri diferite ale limbii. În ciuda acestor schimbări, nu toate limbile europene au o dezvoltare digitală în domeniul lexicografiei. Lexicografia are un rol încă de neînlocuit în proiectele care depind de resurse nedigitalizate, și anume lexicografia istorică și cea diacronică, incluzând aici și realizarea de lucrări fundamentale pentru culturile europene, cum ar fi tezaurele academice. De aceea, prin acest articol ne propunem să evidențiem pe scurt noi modele de automatizare a redactării dicționarilor, limitele actuale ale acestora, precum și modalități de îmbunătățire pentru a eficientiza activitatea lexicografică.

2. Modelul COBUILD

Încercările de a automatiza lucrul la realizarea de dicționare, având ca efect reducerea timpului și economia de resurse, sunt mai mult decât binevenite. Merită evidențiat aici un proiect de pionierat în domeniul lexicografiei, intitulat COBUILD, inițiat la Universitatea din Birmingham, în 1980. COBUILD este acronim pentru *Collins Birmingham University International Language Database* și reprezintă un proiect finanțat de editura Collins împreună cu Universitatea Birmingham, coordonat, la început, de profesorul John Sinclair. Cea mai importantă realizare a acestui proiect o reprezintă crearea unui corpus adnotat de texte contemporane în limba engleză și redactarea dicționarului monolingv pentru elevi *Collins COBUILD English Language Dictionary*, bazat pe studiul corpusului COBUILD și

publicat pentru prima dată în 1987. Inovațiile aduse de acest proiect s-au dovedit definitorii în evoluția lexicografiei spre domeniul digital. În primul rând, realizarea dicționarului s-a bazat pe un corpus de texte digitale care a constituit modelul de semnificație și utilizare a cuvintelor, fiind primul dicționar care a avut o astfel de bibliografie (Lew 2023: 2). În al doilea rând, a adus o nouă viziune asupra limbii, tratând sensul și structura în mod unitar, nefăcând o delimitare, după modelul tradițional de până atunci, între lexic și gramatică. Acest lucru a însemnat o nouă viziune și asupra sensurilor articolelor de dicționar, bazată pe asocierea sensului cu frecvența și cu modele specifice de utilizare. În al treilea rând, ca o consecință a acestei viziuni, multe intrări au fost definite în dicționarul COBUILD după un format nou, numit *Full-Sentence Definition*, care presupune două părți ale definiției: o definiție explicativă și una prin parafrază. Această nouă abordare a fost determinată de publicul țintă al dicționarului, și anume elevii, și s-a bazat într-o mare măsură pe un format care să ilustreze termenii definiți în realitățile de comunicare, iar definițiile să fie date într-un mod mai puțin rigid, ca și cum profesorii ar explica elevilor semnificația cuvintelor. De asemenea, citatele oferite nu erau inventate de redactori pentru a ilustra sensul sau utilizarea cuvintelor, ci erau excerptate din corpus, fiind exemple autentice de limbă folosite de vorbitori nativi. În cazul în care au existat referințe neclare la context, nume proprii sau cuvinte rare, citatele au fost modificate pentru a nu crea confuzie, mai ales pentru cursanții de limbi străine.

3. Reinventarea lexicografiei

3.1. După modelul proiectului COBUILD, Robert Lew a încercat generarea de exemple pentru articolele de dicționar folosind ChatGPT¹ Plus (*ibid.*: 5). În urma acestui demers cercetătorul englez a observat că nu există diferențe notabile între exemplele generate cu inteligența artificială și cele generate de experții lexicografi. E adevărat că pentru aceste rezultate au fost necesare îmbunătățiri permanente ale setului de instrucțiuni transmise sistemului. Prin aceste ajustări, prin expertiza experților, sistemul, în accepția lui Lew, pare capabil să genereze exemple de propoziții care sunt clare și sună autentic, imitând comunicarea umană.

Preocuparea de a cere unui sistem informatic să redacteze intrări de dicționar au avut-o și cercetătorii Miloš Jakubiček și Michael Rundell care au încercat, cu ajutorul ChatGPT, folosind o bază mare de date lexicale și toate instrumentele lexicografice digitale realizate până acum, să genereze un mic dicționar englezesc având 99 de intrări generate complet automat (Jakubiček, Rundell 2023: 518). Cei doi au urmărit performanța sistemului în ceea ce privește următoarele elemente: dezambiguizarea semantică a cuvintelor, definițiile, informațiile gramaticale, elementele lexicale marginale (cum ar fi cuvinte formale, arhaice

¹ GPT este acronim pentru *Generative Pre-trained Transformer*, un tip de rețea neuronală care este antrenată pe un corpus mare neanalizat (adică text simplu), producând un model lingvistic, adică o distribuție probabilistică a cuvintelor având în vedere cuvintele anterioare. Un astfel de model face posibilă realizarea a ceea ce formal se numește decodare sau inferență și, în termeni practici, generează cea mai probabilă secvență de cuvinte având în vedere un indiciu.

sau ofensatoare) și propozițiile folosite ca exemple de utilizare. În fiecare caz, au fost comparate rezultatele ChatGPT cu intrările echivalente din două dicționare de referință: *Oxford Dictionary of English* (ODE), care este în prezent sursa implicită pentru o căutare Google de tipul „definește X”; și *Macmillan English Dictionary* (MED). Provocările legate de sensurile cuvintelor și de organizarea semantică a intrărilor într-un dicționar sunt bine cunoscute. Stabilirea unui set de sensuri pentru un cuvânt dat este considerată, în general, cea mai dificilă sarcină din lexicografie, nu în ultimul rând pentru că sensul este atât de determinat contextual. Sensurile numerotate ale unui articol de dicționar sunt în realitate o construcție lexicografică și, în multe cazuri, nu există două dicționare care să prezinte același inventar de sensuri pentru un cuvânt polisemantic. Cu toate acestea, în cadrul acestei paradigme convenționale, putem verifica dacă analiza sensurilor unui cuvânt într-un anumit dicționar este o reflectare corectă – și practic utilă – a modului în care cuvântul este utilizat în situații de comunicare reale. Cei doi cercetători au observat că ChatGPT nu are performanțe bune în acest domeniu, deși gradul de dificultate nu a fost unul ridicat. În plus, eșantionul propus de cei doi nu a inclus niciun cuvânt de mare complexitate (cum ar fi cuvintele cu șase sau mai multe sensuri într-un dicționar tradițional), astfel încât putem presupune că sistemul, în varianta actuală, nu poate oferi soluții lexicografice pentru cuvintele polisemantice.

În ce privește definirea cuvintelor, ChatGPT reușește să ofere răspunsuri potrivite, într-o formă accesibilă, comparabilă cu cea din dicționarele tradiționale. Dar, în ce privește informațiile gramaticale au existat erori ale sistemului în a identifica corect termenii interogați, în special în a deosebi valoarea tranzitivă sau intransitivă a unui verb. În unele cazuri, adverbele au fost identificate drept prepoziții, ceea ce a generat atât erori de răspuns din partea sistemului, cât și soluții eronate oferite utilizatorului. Referindu-se la exemplele de utilizare a termenilor analizați, cei doi cercetători au observat faptul că sistemul ChatGPT are o problemă evidentă și persistentă în modul în care generează contexte de comunicare. Exemplele din ChatGPT sunt, în cea mai mare parte, neconvingătoare, iar atunci când există un set de exemple, acestea prezintă prea puțină diversitate în ceea ce privește structurile și chiar subiectul, observându-se o paradigmă uneori supărătoare, de tipul pronume de persoana a III-a, urmat de verb la trecut. Dicționarele tradiționale urmăresc toate contextele de comunicare și expun utilizatorul la o varietate de stiluri și modele de utilizare a cuvintelor.

În ciuda acestor inadvertențe, cei doi cercetători scot în evidență și câteva argumente pentru folosirea ChatGPT în activitatea lexicografică, subliniind faptul că sistemul este capabil să se îmbunătățească prin sugestiile și indicațiile oferite, este mai rapid și are deja integrate funcții de traducere automată și, nu în ultimul rând, este ușor de utilizat.

Totuși limitările sistemului nu pot fi trecute cu vederea, de aceea Miloš Jakubiček și Michael Rundell, preluând o remarcă făcută de Adam Kilgarriff, din 2007 („Googleology is a bad science. And GPTology too.”), insistă pe faptul că folosirea exclusivă a ChatGPT nu este în niciun caz o soluție pentru realizarea de dicționare, din cel puțin 3 motive (Jakubiček, Rundell 2023: 530), și anume:

a) în primul rând, sistemul ChatGPT utilizează surse de date necunoscute, cu răspunsuri nedeterministe, cu stabilitate și reproductibilitate foarte limitate. Utilizarea sa ca sistem de căutare cu scop general într-un context științific suferă în mod inevitabil de toate problemele pe care le are o abordare bazată pe căutarea Google;

b) ChatGPT nu are încă acces la toate dicționarele unei limbi și nici nu poate distinge între lucrări lexicografice de referință și alte lexicoane, de aceea, de multe ori modelul ales de sistem nu este cel mai potrivit pentru utilizator. În plus, este necesar ca datele de limbă să fie prelucrate lexicografic, adică să ofere informații despre sensuri, forme, frecvență etc., tocmai pentru a fi identificate corect de către sistemul analitic;

c) Generarea de dicționare utilizând exclusiv instrumente precum GPT implică abandonarea rolului actual al lexicografului de a examina și verifica dovezile sugerate de un sistem analitic. Majoritatea instrumentelor de prelucrare a limbajului pentru lexicografie interoghează un corpus, efectuează o analiză (adesea foarte complexă), dar se raportează la dovezile din corpus sub formă de linii de concordanță, astfel încât lexicograful să poată determina dacă rezultatele automate corespund cu ceea ce se află în corpus (și să verifice conținutul corpusului, metadatele, adnotarea etc.). În aceste sisteme, precum GPT, factorul uman este eliminat din procesul de generare a datelor, iar lexicografii nu pot verifica cum a folosit sistemul datele. De asemenea, generarea de exemple ar trebui să reflecte ceea ce datele ne arată a fi contextele și modelele în care un cuvânt apare cel mai des: exemplele nu ar trebui inventate, ci ar trebui să fie găsite în corpus (și scurtate sau ușor modificate, dacă este necesar). Exemplele generate de ChatGPT nu pot fi găsite nicăieri și nici nu avem o siguranță că ele au fost produse vreodată de un scriitor sau de vreun vorbitor uman sau că reprezintă modelul potrivit pentru definiția aleasă.

3.2. Pentru a înțelege oportunitățile și limitările unui astfel de sistem, am încercat antrenarea lui pentru definirea unui cuvânt rar din limba română – *cab* ca unitate de măsură. De aceea, am atașat niște capturi de ecran care să evidențieze modul de lucru și de interpretare din partea ChatGPT (versiunea GPT-4) a informațiilor la care are acces.

Read aloud

Ce înseamnă cab, unitate de măsură?

„Cab” nu este o unitate de măsură standard cunoscută în sistemele internaționale de măsurare. Totuși, ar putea fi o abreviere sau un termen specific unui domeniu anume. Ai un context anume în care ai întâlnit acest termen? Astfel, te pot ajuta să-l interpretezi corect. 😊

cab este o unitate de măsură evreiască, veche

Ai dreptate! „Cab” este o unitate de măsură tradițională evreiască, folosită în antichitate, în special în perioada Bibliei. Aceasta era folosită pentru măsurarea volumului și era echivalentă cu aproximativ un sfert dintr-o „seah” (altă unitate de măsură evreiască veche).

Un cab era folosit pentru măsurarea cerealelor, lichidelor și altor produse, iar în unele surse, se estimează că un cab ar fi echivalent cu aproximativ 0,6 litri (sau 600 ml).

Această unitate nu este utilizată în prezent, dar apare în texte istorice și biblice.

La o primă analiză, sistemul a identificat termenul ca fiind o abreviere, deși i s-a sugerat deja domeniul de utilizare. La a doua încercare, prin rafinarea informației oferite, ChatGPT a putut găsi date despre termenul *cab* drept unitate de măsură. Din datele extrase de sistem, el a concluzionat că valoarea cabului este puțin peste jumătate de litru, ceea ce era diferit față de informațiile pe care le avea lexicograful. De aceea, am cerut ChatGPT să îmi ofere sursele bibliografice care susțin valoarea de aproximativ doi litri pentru un cab, așa cum fusese excerptată informația din dicționarele de referință aflate în bibliografia DLR. În urma propriei analize, realizată de sistem a reieșit că majoritatea dicționarelor academice susțin valoarea sugerată de lexicograf. Acest lucru sugerează că sistemul ChatGPT are capacitatea de a oferi răspunsuri din ce în ce mai adaptate cerințelor utilizatorului, prin rafinări permanente:

 Surse care indică o valoare mai apropiată de 2 litri:

1. Herbert Danby – în traducerea și comentariile sale la *Mishna*, unele interpretări indirecte pot sugera un **cab mai mare** (nu dă direct 2L, dar din raporturi reiese posibil).
2. Victor Tcherikover și alți istorici ai perioadei greco-romane din Israel menționează că:
„În funcție de zona geografică și perioada istorică, un cab ar fi putut avea între 1,2 și 2 litri.”
3. *Archaeological Study Bible* (Zondervan, 2005):
„Some scholars suggest a cab could be as much as 2 liters, depending on how one interprets the size of a log or seah.”
4. *Manuale de arheologie biblică* – cum ar fi cele ale Israel Finkelstein sau Amihai Mazar – uneori sugerează o variație mai largă pentru unități ca log, cab, hin etc.

 Calcul indirect:

Dacă iei 1 seah = 13.2 litri (așa cum sugerează unii cercetători), atunci:

- 1 seah = 6 cabs → 1 cab = 13.2 / 6 = 2.2 litri

 Concluzie:

Da, există surse și argumente care susțin că 1 cab ar putea fi ≈ 2 litri, deși cele mai comune estimări moderne sunt între 1,2 și 1,5 litri. Valoarea mai mare (2L) este acceptată în unele cercuri academice, în special când se bazează pe proporții mai mari pentru unități ca seah sau hin.

Am cerut apoi sistemului să redacteze o definiție, asemenea unui lexicograf, pentru termenul analizat. ChatGPT a menționat faptul că a realizat o definiție după modelul DEX pe care l-a avut disponibil online. Și aceasta, deși termenul *cab* ca unitate de măsură nu este definit în DEX.

cab s. n.

(învechit, rar) Unitate de măsură pentru volum folosită în vechiul sistem evreiesc, egală cu aproximativ 1,2–2 litri; era utilizată în special pentru produse alimentare, cum ar fi cerealele sau lichelele.

– Din ebr. *kab*.

Prin acest demers, sistemul a putut adapta o definiție preluată din texte din mai multe limbi, la un dicționar explicativ românesc, învățând modul de organizare a informațiilor lexicografice. După această definiție, sistemul a oferit el alternative pentru lexicograf,

formulând definiții după alte tipuri de dicționare specializate, cum ar fi un dicționar biblic, istoric sau teologic, preluând modele din lucrările lexicografice internaționale.

Pentru a testa capacitatea sistemului ChatGPT de a genera fraze, i-am cerut formularea unui enunț în care să fie folosit *cab* drept unitate de măsură:

Îmi poți genera o frază în care să folosești cuvântul „cab” cu sensul de unitate de măsură?

Desigur! Iată o frază corectă contextual și istoric:

„Pentru ofranda de făină adusă la templu, fiecare familie trebuia să aducă o cantitate de un cab de grâu curat.”

Vrei și o variantă într-un context modernizat sau explicativ?

o variantă cu formulare biblică?

Desigur! Iată o variantă cu un stil de redactare inspirat din limbajul biblic:

„Și a măsurat un cab de floare de făină, curată și fără prihană, precum era rânduit după legea Domnului, și a adus-o înaintea preoților spre jertfă.”

Dorești și o versiune care să includă un citat stilizat sau o referință fictivă la un verset?

Dă-mi și o referință fictivă

Desigur! Iată fraza completă cu o referință fictivă în stil biblic:

„Și a luat un cab de untdelemn curat, stors din măslinile cele dintâi, și l-a turnat în vasul sfânt, după rânduiala dată în cartea Leviticului (Leviticul 9:14f).”

♦ Notă: Versetul „Leviticul 9:14f” este fictiv, dar formulat în stil autentic pentru a păstra atmosfera unui text sacru sau liturgic.

Vrei și alte exemple din alte „cărți fictive” sau o formulare stil Talmud?

📄 📄 📄 📄 📄 📄



După cum se poate observa, citate nu sunt excerptate din texte online sau site-uri, ci sunt inventate pur și simplu. Din această cauză, sistemul a inventat fraze care nu conțin adevăruri, ci ficțiuni. De exemplu, la Templul evreiesc se aduceau animale ca jertfă, însoțite de un dar de mâncare, și anume turte. Dar nu se aduceau ofrande de făină, ci făina era frământată cu untdelemn și coaptă. De asemenea, chiar dacă pare o metaforă interesantă, nu putem spune despre făină că era „fără prihană”, acesta fiind un atribut prin care pot fi caracterizați doar Dumnezeu și, în anumite cazuri, ființa umană. De aceea, nici cea de-a doua frază nu poate fi considerată potrivită. La ultima frază sistemul a întrebat dacă utilizatorul dorește un citat stilizat sau o referință fictivă, ceea ce a fost o propunere puțin șocantă pentru un lexicograf. Încercând să înțeleg până unde merge, am rămas uimit cu câtă larghețe mi s-a oferit o frază în care, pe lângă informațiile false extrase din cartea *Levitic*, am primit și o referință fictivă, argumentată de sistem ca fiind un mijloc de a păstra „atmosfera unui text sacru sau liturgic”. Am insistat asupra unei fraze dintr-un text autentic în care să apară termenul discutat, dar sistemul a preferat să realizeze el una nouă, susținând că a fost formulată în stil biblic și că reflectă utilizarea tradițională a acestei unități de măsură în contextul ritualurilor descrise în textele sacre evreiești. Totuși, la o simplă căutare această

unitate de măsură nu este deloc folosită în textele care descriu activitatea la Templu sau lucrările preoților. Insistând ca sistemul să îmi ofere un citat cu pagină, sistemul a găsit ceea ce trebuia să-mi ofere de la început:

Îmi poți da și sursa, cu pagina?

Desigur! Iată un citat din Biblie în care apare cuvântul „cab” cu sensul de unitate de măsură:

„În timpul asediului, un sfert de cab de găinaț de porumbel se vindea cu cinci sicli de argint.”

dictionarbiblic.blogspot.com

Însă a preferat să îmi ofere adresa unui site, decât să citeze direct din textul biblic.

Dacă ar fi să comparăm ceea ce am reușit să realizăm cu ChatGPT și cum arată intrarea pentru termenul *cab* în dicționarul-tezaur, diferențele sunt foarte mari ceea ce dovedește faptul că mai este mult de lucru până când un sistem ca acesta va putea înlocui total munca lexicografilor, cel puțin pentru limba română:

CAB¹ s. m.

Ebraism Unitate de măsură pentru capacitate (egală cu aproximativ doi litri), la vechii evrei, folosită inițial pentru determinarea volumului materiilor solide, iar în perioada talmudică și pentru lichide; vas care are capacitatea de un cab. V. c a u ș, p i n t â.

<1665–1685> *Biblia* <1665–1685>, VIII, 149: S-au făcut capul de măgar 50 de sicli de argint și a patra parte de **cavă** [sic!] a găinațului de porumbel, 50 [sic!] sicile de argint.

1688 *Biblia* (1688), 270: Să făcu foamete mare în Samaria. Și iată, șădea prejur ea, până unde au ajuns capul de măgar de 50 de sicile de argint și a patra parte de **cav** de găinaț de porumbel, 5 sicile de argint.

<1760–1761> *Biblia* <1760–1761>, II, 498: Până într-atâta s-au conăcit, până când să vindea un cap de asin cu optzeci de arginți și a patra parte de **cavi** de gunoiu de porumbel cu cinci arginți.

1862 ANTONESCU, D.

1913 I. BRĂESCU, *Măsurătoarea*, 3: Sistemul lui Carol cel Mare ... nu era decât o amestecătură a sistemelor preexistente; piciorul său e piciorul arab; prăjina (de suprafață) – actus al romanilor; pinta – **cabul** evreu; livra de 12 uncii – livra arabă.

1938 *Biblia* (1938), 391: Bântuia foamete mare în Samaria. Și ei au ținut-o împresurată, încât un homer de vin costa optzeci de sicli de argint și un sfert de **cab** de făină albă costa cinci sicli de argint.

1988 *TDRG*², II, 211.

2001 *Biblia* (2001), 598: Atât de mult au ținut-o împresurată, încât o căpățână de măgar ajunsese la prețul de cincizeci de sicli de argint, iar un sfert de **cab** de găinaț de porumbel [se vindea] cu cinci sicli de argint.

2004 *Septuaginta*, 2, 558: A fost foamete mare în Samaria și, iată, au împresurat-o până ce o căpățână de asin a ajuns să prețuiască cincizeci de sicli de argint, iar un sfert de **cab** de găinaț de porumbel – cinci sicli de argint.

– Pl.: **cabi**.

– Var.: *inv.*, *rar* **cav** s. m.

– Din. ebr. **qab**, gr. **κάβοϛ**. Cf. lat. c a b u s.

Referințe bibliografice:

<1691–1697> T. CORBEA, *Dictiones*, 80; 1905 PUȘCARIU, *Îns. Etym. Wb.*, 31; 1923 GIUGLEA, *Cuv. Rom.*, 163; 2005 în *Biblia* <1760–1761>, II, 526.

3.3. Așa cum menționau și Miloš Jakubiček și Michael Rundell, chiar dacă rezultate experimentelor cu ChatGPT au demonstrat că sistemele analitice nu pot încă înlocui

expertiza lexicografilor în realizarea de dicționare, astfel de sisteme nu pot fi ignorate. Așa cum până acum lexicografia a ținut cont de tendințele tehnologice emergente, cu scopul de a produce dicționare mai bune la un cost mai mic de timp și resurse, este necesar acum să se analizeze contribuția ChatGPT la aceste activități. Modelele GPT pot fi antrenate pe corpusuri corect adnotate pentru ca rețelele neuronale să poată învăța automat din seturi mari de date prelucrate pentru a oferi soluții personalizate lexicografilor.

Astăzi, marile dicționare englezești au înglobat sistemele de procesare a limbajului natural și de inteligență artificială pentru a oferi utilizatorilor răspunsuri cât mai potrivite, dar și cât mai cuprinzătoare despre definiții, etimologii, pronunție, exemple și contexte de utilizare. Putem menționa aici *Oxford English Dictionary Online* sau *Cambridge Dictionary Online* sau *Miriam-Webster Dictionary Online* care permit diverse facilități pentru căutări personalizate pentru a ajuta fiecare utilizator, într-un mod cât mai potrivit. Pe lângă acestea și alte programe de traducere automată, precum *Reverso* sau *WordReference*, au integrat pe platformele lor sisteme de inteligență artificială pentru a oferi traduceri cât mai adaptate contextului de comunicare.

4. În loc de concluzii

Lexicografia a contribuit la dezvoltarea corpusurilor și la crearea de sisteme de scriere de dicționare, care au eficientizat activitatea lexicografică și au eliminat multe din activitățile care necesitau mult timp din partea lexicografilor. În plus, publicarea online a eliminat limitările de dimensiune ale lucrărilor tipărite, a permis adăugarea de conținut audio-video și a introdus o multitudine de criterii de căutare în text, ca alternative la căutarea pe baza cuvintelor-titlu. Este adevărat faptul că dicționarele academice au fost inițial create pentru experți, având ca rol pentru publicul larg, mai degrabă, furnizarea de informații cât mai detaliate și constituind un punct de referință pentru alte dicționare uzuale. Dar digitizarea acestor lucrări a permis integrarea informației în diverse aplicații precum asistenți de scriere digitală, programe de învățare a limbilor străine sau instrumente de traducere asistată. Pe lângă acestea, numeroase proiecte de cercetare au contribuit la dezvoltarea și reinventarea lexicografiei. O inițiativă care a urmărit o nouă orientare în lexicografie a fost în cadrul proiectului COST ENeL *European Network on e-Lexicography* (<https://www.elexicography.eu/>), care și-a propus nu doar enunțarea de idei, ci punerea lor în practică prin diverse proiecte. Amintim aici realizarea consorțiului ELEXIS (<https://elex.is/>), care a urmărit interconectarea dicționarelor științifice, ca Linked Open Data (https://en.wikipedia.org/wiki/Linked_data), astfel încât să devină disponibile ca resurse de date deschise pentru cercetători și lexicografi. Proiectul Prêt-à-LLOD (<https://pret-a-llod.github.io/>) este și el un mijloc prin care s-a urmărit corelarea datelor și tehnologiilor lingvistice, adică Linguistic Linked Open Data (LLOD) pentru a crea date multilingve corecte și disponibile online. Pe lângă acestea a fost realizată și o rețea numită Nexus Linguarum (<https://nexuslinguarum.eu/>), tot în cadrul unui proiect COST intitulat *The European network for Web-centred linguistic data science*, încheiat în 2024, prin care s-a

urmărit construirea și consolidarea unei rețele europene pentru diverși creatori și utilizatori de date lingvistice, precum: lingviști, lexicografi, informaticieni, terminologi etc.

În ciuda acestor inițiative diverse, lexicografia academică românească nu este încă valorificată la potențialul ei maxim, în primul rând, din cauza stadiului scăzut de digitalizare, ceea ce o face să nu fie interconectată la bazele imense de date și la tehnologiile care procesează limbajul natural. Dar rolul ei rămâne fundamental în ce privește știința cuvintelor și modul cum ar trebui ele folosite. Schimbarea de paradigmă presupune și a vedea dicționarele nu ca pe lucrări lexicografice de sine stătătoare, ca până acum, ci ca lucrări interconectate într-o vastă rețea digitală, în care dicționarele servesc drept resursă de bază pentru multiple aplicații dezvoltate în diverse domenii de activitate umană. De asemenea, este necesar ca datele lexicografice să fie adaptate pentru diverse scopuri, în diferite forme, pentru a fi astfel relevante în funcție de modelul de aplicație folosit.

Lexicografia academică este datoare să continue monitorizarea tendințelor și schimbărilor importante din limbă, dar nu doar a celor extrase din bibliografii sau corpusuri bine delimitate și selectate cu atenție ca să reprezinte utilizarea generală a limbii. Ci este nevoie ca lexicografia să se raporteze la o bază empirică în continuă schimbare și creștere exponențială. Totuși cantitatea de date lingvistice disponibile devine aproape imposibil de prelucrat în cadrul fluxurilor de lucru tradiționale. De exemplu, pentru termenul „ca” în corpusul actual al *Dicționarului limbii române* au fost găsite, în luna mai 2025, mai mult de 4 355 900 de ocurențe. De aceea, mijloacele de automatizare a prelucrării datelor lingvistice, de sistematizare și de analiză a lor care există în prezent sau care vor fi dezvoltate în viitor trebuie adoptate și adaptate pentru a oferi lexicografilor un material util pentru redactare. Dacă până acum un rol fundamental al oricărei academii a fost realizarea unui dicționar-tezaur național, astăzi cercetarea academică cuprinde și alte activități complexe precum compilarea de resurse lingvistice, crearea de corpusuri de texte, baze de date adnotate sau instrumente de prelucrare a limbii, cu scopul nu doar de a înregistra istoric o realitate lexicală, ci de a oferi resurse practice, corecte științific și cât mai cuprinzătoare, celor interesați. Așadar, este necesară crearea unei infrastructuri de date lexicale prelucrate, ce ar putea fi folosite pentru diferite scopuri și în diferite moduri.

BIBLIOGRAFIE

- Bolinger 1985 = Dwight Le Merton Bolinger, *Defining the Indefinable*, in Robert Ilson (ed.), *Dictionaries, Lexicography and Language Learning*, ELT Documents 120, Oxford, Pergamon Press, 1985, 69–73. Reprinted in Thierry Fontenelle (ed.), *Practical Lexicography: A Reader*, Oxford, Oxford University Press, 2008.
- Jakubíček, Rundell 2023 = Miloš Jakubíček, Michael Rundell, *The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?* in Medve, M. & Měchura, M. & Tiberius, C. & Kosem, I. & Kallas, J. & Jakubíček, M. & Krek, S. (eds.) (2023). *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, Brno, 27–29 June 2023, Brno, Lexical Computing, p. 518–533.

Lew 2023 = Robert Lew, *ChatGPT as a COBUILD lexicographer*, in „Humanities and Social Sciences Communication”, nr. 10, <https://doi.org/10.1057/s41599-023-02119-6>.

**WORDS AND CODE:
LEXICOGRAPHY REINVENTED WITH DIGITAL TECHNOLOGY**

ABSTRACT

Nineteenth-century academic dictionaries are veritable cultural monuments, offering not only a collection of words, but also a scientific description of language based on actual usage. Although such projects are rare today, they remain fundamental landmarks for understanding and preserving linguistic heritage. Dictionaries support communication by defining words, facilitating translation and correct writing, functioning as normative tools, built according to rigorous conventions. The creation of these works is complex and time-consuming. The *Oxford English Dictionary* took over 70 years to complete, and the *Romanian Language Thesaurus* took almost a century to develop. Sometimes criticised for their rigidity, these dictionaries are nevertheless organised in such a way as to provide clear information on meanings, grammar, etymology and stylistic usage. The transition to the digital environment has removed physical limitations and enabled quick searches, access to expressions and examples from various corpora, and the integration of multimedia content.

However, the digitisation of lexicography is uneven across Europe. Academic thesauruses remain indispensable for in-depth language description and cannot be completely replaced by automated solutions. One relevant project is COBUILD (1980), which used authentic corpora to define meanings and provide examples, becoming a model in pedagogical lexicography. Recent research has tested the ability of AI, particularly ChatGPT, to generate lexicographical content. The results were promising for definitions but poor for generating realistic examples, grammatical classification and organising meanings. Although AI offers speed and flexibility, there are risks: the lack of transparent sources, the instability of responses, and the elimination of human control affect the quality and reliability of information. However, the integration of artificial intelligence into lexicographical projects is inevitable. Major publishers and online applications are already using NLP and AI technologies to offer advanced features: customised definitions, contextual examples, pronunciations and automatic translations.

European initiatives such as ELEXIS and ENeL promote collaboration between linguists, computer scientists and developers to create open, interconnected lexical resources tailored to modern needs. Even if academic lexicography is not yet being exploited to its full potential, its future depends on a balance between human expertise and technological innovation. Thus, reinventing dictionaries becomes not only possible, but necessary.

Keywords: *lexicography, academic dictionaries, ChatGPT, digital technology, annotated databases, multilingual corpus, NLP.*

