Metode statistice aplicate crosdisciplinar în serii enumerative sau atributive nenumerice. Analize frecvențiale statistice în (mini)corpusuri filologice

Prof. univ. dr. habil. Gheorghe Săvoiu
Universitatea din Piteşti
Conf. univ. dr. habil. Constantin Manea
Universitatea din Piteşti

Rezumat

Una dintre cele mai simple serii de date statistice, apărută ca rezultat al prelucrării prin grupare a realităţii lingvistice sau filologice, aprofundată ştiinţific fie sub denumirea de serie atributivă nenumerică fie sub aceea de serie enumerativă este supusă în articol investigării specifice a gândirii statistice de început, respectiv analizei frecvenţelor de apariţie a unor cuvinte sau expresii specifice. Metoda analizei feecvenţiale în (mini) corpusuri filologice poate să devină astfel, prin aplicabilitate crosdisciplinară, o metodă utilă de validare în lingvistica modernă, cu şansa de a oferi soluţii de alegere criterială riguroasă sau argumente pertinente, extrase din uzul limbii în situaţii de ambiguitate şi chiar de incertitudine selectivă. Tipul de serie investigată în articol, tip dominant în universul filologic sau lingvistic, dobândeşte o utilitate specială ce poate fi asigurată prin prelucrarea rapidă cu ajutorul metodei sau analizei frecvenţiale. Metoda frecvenţială permite luarea de decizii prompte în context de incertitudine filologică sau lingvistică, oferind indicatori statistici diverşi capabili să confere veridicitate argumentaţiei probatorii axate pe uzul lingvistic, prin dezvoltarea statică, dinamică, spaţială şi structurală a cuantificării în (mini)corpusuri filologice a frecvenţelor de apariţie, prin confruntare de opţiuni, prin evaluarea concentrărilor sau diversificărilor şi în final chiar prin valorificarea profilului statistic în lingvistica modernă. Autorii oferă câteva exemplificări pertinente ale utilităţii metodei prin crosdisciplinaritate, care vizează în acest caz simplificarea deciziei, atât în filologia tradiţională, cât şi în lingvistica modernă, valorifică informaţii statistice trecute, pentru a desprinde tendinţele de utilizare de o manieră prospectivă cât mai corectă a limbajului ştiinţific clasic sau modern. Pentru aplicabilitatea (mini)corpusul filologic modern de tip Internet accesat cu motoare de căutare devine populaţia statistică observată şi prelucrată, iar prezentarea tabelară şi reprezentarea grafică adecvată delimitează imaginea *cantitativă a probatoriului. În final, autorii anticipează evoluţia către o lichiditate semnificativă sau chiar către o volatilitate excesivă a limbajului modern şi asmilând analiza fluxurilor de date ale unor serii enumerative, tratează şi identifică statistic soluţii subliniind impactul major al metodelor aplicate crosdisciplinar în cadrul oricărei cercetări ştiinţifice filologice sau lingvistice.*

Cuvinte cheie: crosdisciplinaritate, metodă statistică, seria enumerative, seria homogradă sau de attribute, analiză frecvenţială, (mini) corpus filologic sau lingvistic, indicatori de concentrare – diversificare, metoda profilului statistic, uzul lingvistic.

Coduri JEL: C46, C49.

1. Introducere

Metodele statistice teoretice îşi extind permanent aplicabilitatea lor practică şi devin în acest fel tot mai utile în variate domenii ale cercetării ştiinţifice, conturând o crosdisciplinaritate evidentă. Aşa cum este concepută şi descrisă aplicativ în acest articol, crosdisciplinaritatea "*constituie o abordare care selectează, îmbină, asociază, combină, aplică metode unice în realităţi ştiinţifice diverse sau pune în practica bine delimitată disciplinar metode ale unei ştiinţe în corpul metodologic al altora, devenind un concept generic în creativitatea demersului investigativ sau metodologic, şi constituie, în final, un prim pas către apariţia şi delimitarea de noi discipline sau ştiinţe*" (Săvoiu, et al., 2020, p. 8).

Crosdisciplinaritate modernă oferă soluţii complexe prin originalitatea abordărilor specifice sau prin creativitatea transpunerii lor în alte ştiinţe, pornind de la simpla constatare că simplifică un demers clasic devenit mult prea uzual şi uneori tot mai ineficient, în raport cu evoluţiile realităţii. În investigaţiile şi argumentaţiile cercetătorilor aflaţi în faţa unor fenomene tot mai variate şi complicate, de la cel sociologic, la cel filologic, de la cel biologic, la cel demografic etc., singura opţiune este transformarea viziunii izolatoare şi consacrate unidiscplinar într-o investigaţi ăn echipă de tip crosdisciplinar prin apelul la metode diverse, între care cele statistice, matematice şi fizice apar ca prioritare. Astfe, multe din problemele lingvistice destul de dificil de soluţionat pot fi rezolvate apelând la metode statistice dintre cele mai simple. Analiza textelor bazată pe recurenţe statistice devine uneori o chestiune de viaţă şi de moarte, aşa cum ar putea învăţa, astăzi, orice student aflat la primele sale lecţii de limbă engleză, de la un uimitor lingvist, cercetător şi scriitor, în acelaşi timp, ca David Crystal, care îi exemplifică expresia *a matter of life and death*, apelând la celebra sa lucrare *The English Language: A Guided Tour of the Language,* şi descriind modalitatea prin care același student atras de învăţarea corectă a limbii engleze ar putea scăpa de la o gravă acuzaţie de plagiat sau de pedepsire prin spânzurare, oferind drept probă salvatoare o analiză profundă – simultan lingvistică şi statistică – a unei scrisori care aparţine criminalului şi care i-a fost atribuită studentului în mod fals, urmărind doar frecvenţa utilizării unor expresii, a unui anumit stil sau coloratura distinctivă a limbajului ... (Crystal, 2002).

Acest articol este rezultatul unor căutări comune ale autorilor, apărute din dorinţa de a simplifica deciziile legate de uzul corect filologic sau lingvistic al limbajului și de a valida/invalida diverse ipoteze de analiză frecvenţială sau a contura ca indicatori utili datele referitoare la frecvența în uz pentru unii termeni sensibili, în ceea ce privește normarea/standardizarea lingvistică. Patima investigării crosdisciplinare a unor recurențe și convergențe lingvistice, spiritul de echipă şi dorinţa de a aplica metode statistice în domenii diferite, ca procese de investigaţie influenţate de spaţiu, timp şi structură (Manea, Săvoiu, 2014), au generat două perioade de invetigaţii prin accesări în 2013-2014 şi respectiv în 2019-2020.

2. Recenzia literaturii de specialitate

În cercetarea filologică românească, mai veche sau mai recentă, având la bază metoda statistică a analizelor frecvenţiale, primele studii au pornit fie de la dicţionare ale limbii române, fie de la anumite texte, alese în aşa fel încât să aibă un anume grad de reprezentativitate pentru limba română. Primul filolog preocupat de aplicarea cuantificărilor statistice cu rol de pionier în domeniul asigurării unei metode sau analize probatorii în studiul structurii etimologice a lexicului românesc a fost A. de Cihac. Astfel acest evaluator a realizat o falsă statistică prin consultarea incorectă metodologic a unor dicţionare şi glosare selectate special pentru a obţine anumite rezultate, fără justificarea ştiinnţifică referitoare la lista completă din care s-a realizat prelevarea nereprezentativă pentru perioada 1870-1879, precum şi la metoda de selecţie în sine, evident dirijată şi implicit subiectivă, mărturisită în prefaţa la volumul al doilea al dicţionarului său etimologic, intitulat *Dictionnaire d'étymologie daco-romane. Éléments slaves, magyars, turcs, grecs-modernes et albanais*, Francfort s/M, apărut în 1879, care constituie practic primul dicţionar ştiinţific al limbii române. A. de Cihac a realizat în final o apreciere aproximativă a unor frecvenţe relative, luând în calcul cuvintele „*nederivate*" din dicţionarul alcătuit de el, fără a se opri şi asupra unităţilor lexicale ale limbii române. Pornind de la cele „*aproximativ 500 de cuvinte latine, 1.000 de cuvinte slave, 300 de cuvinte turceşti, 280 de cuvinte greceşti moderne şi 20 până la 25 de cuvinte maghiare sau albaneze*" (Dimitriu, 1973, p. XIII), A. de Cihac apreciat eronat că elementul latin „*care constituie fără îndoială substanţa* *limbii române* […] *nu numai că a rămas aproape staţionar, după primirea sa, în ceea ce priveşte fondul vocabularului, dar acesta din urmă trebuie chiar să fi pierdut multe cuvinte ca urmare a atâtor tulburări la care aceste nefericite locuri au fost teatrul timp de secole*" (Dimitriu, 1973, p. VIII). Concluziile evaluărilor statistice aproximative şi dirijate ale lui A. de Cihac diminuează locul şi rolul elementului latin moştenit în vocabularul limbii române, în mod subiectiv. Din păcate, realitatea faptelor de limbă a fost grav viciată de datele furnizate de astfel de cuantificări de A. de Cihac, cât şi de aceea în care Sextil Puşcariu, o justifica pe prima, folosind procentaje apropiate şi o revalida în 1920. Frecvenţele relative calculate şi prezentate păreau să indice faptul că limba română ar avea o structură etimologică *preponderent slavă* a lexicului (cu o valoare cu puţin peste două cincimi), iar elementul *latinesc,* moştenit, nu era de fapt decât *una dintre celelalte* cincimi constitutive, celelalte *două cincimi* rămase reunind elemente lexicale turceşti şi elemente cu origine eterogenă: maghiară, neogrecească, albaneză etc).

În primul rând, cuantificările statistice amintite nu s-au sprijinit nici pe o metodologie coerentă şi nici pe un aparat statistico - matematic adecvat, deoarece, pe de o parte, ele nu corespund realităţii cifrelor date de „*indicele*" de la sfârşitul volumului al doilea, iar pe de altă parte numărarea cuvintelor nu a luat în consideraţie raportul dintre *cuvânt* şi *variantă*. În legătură cu primul aspect, care pune în cauză şi cuantificarea statistică făcută de Sextil Puşcariu după „*indicele*" de cuvinte al lui Cihac, Mircea Seche arată că numărătoarea nu corespunde realităţii din două motive clare (Seche, 1966, p. 107): i) în primul rând, cuvintele cuprinse în „indicele" lui A. de Cihac sunt în număr de peste 8900 (nu 5765, aşa cum fals apreciase Sextil Puşcariu); ii) suma totală a cuvintelor înregistrate de dicţionar (aşadar, nu numai de indicele de la finalul volumului al doilea) este de 17645, excluzându-se nu numai toponimele, aşa cum era şi firesc, dar şi variantele (fonetice sau lexicale) – Ambele argumente sunt de o importanţă capitală pentru a sublinia lipsa completă de acurateţe a cuantificării statistice şi în final lipsa unei aprecieri motivate şi dovedite ştiinţific. Mai mult, lipsa unui tratament metodologic unitar sau absenţa unei comparabilităţi statistice asigurate este la fel de evidentă, pentru cuvintele de origine latină, A. de Cihac a luat în consideraţie numai bazele, iar pentru celelalte, şi derivatele. Mircea Seche arată că totalitatea cuvintelor din dicţionarul lui A. de Cihac se repartizează pe straturi etimologice după cum urmează: „*elemente de origine latină (şi derivate ale acestora) – 6141; elemente de origine slavă – 4691; elemente de origine turcă (şi derivate ale acestora) – 1250; elemente de origine greacă (şi derivate ale acestora) – 1100; elemente de origne maghiară (şi derivate ale acestora) – 1026; elemente comune cu albaneza (şi derivate ale acestora) - 90*" (Seche, 1966, p.108). Procentajele corespunzătoare acestor noi date sunt: elementul latin (şi derivatele sale) reprezintă 45,6% din total, elementul slav (şi derivatele acestuia) reprezintă 34,8%, elementul turc (şi derivatele acestuia) reprezintă 7,1%, elementul neogrec (şi derivatele acestuia) reprezintă 6,2%, elementul maghiar (şi derivatele acestuia) reprezintă 5,8% din total, iar elementul comun cu albaneza (şi derivatele sale) reprezintă doar 0,5% din totalul cuvintelor cercetate. Se observă aşadar că diferenţa este sensibilă, atât faţă de frecvenţe relative indicate de A. de Cihac, cât şi faţă de cele indicate de Sextil Puşcariu.

În al doilea rând, statistica lui A. De Cihac pune pe acelaşi plan cuvinte inegale ca forţă *circulatorie* şi ca volum sau conţinut *semantic*. Numărătoarea lui Cihac (ca şi aceea a lui Puşcariu de mai târziu) nu ţine cont de faptul esenţial că unităţile lexicale ale unei limbi nu pot sta în nici un caz pe acelaşi plan în ceea ce priveşte importanţa lor *relativă*. *Ponderea* pe care o au cuvintele de diverse origini marchează identitatea unui vocabular, ca şi gradul lor de reprezentare în vocabularul de bază al limbii studiate. Această *pondere statistică,* măsurată ca frecvenţă relativă trebuie să fie studiată din punctul de vedere al dinamicii vocabularului limbii respective. Deşi Sextil Puşcariu arată în mai multe lucrări, de exemplu *Locul limbii române între limbile romanice*, *Limba română*, vol. I, că latinitatea limbii române, vizibilă din întrega sa structură, se poate deduce şi din „*materialul de construcţie*” al vocabularului, însă nu prin simpla numărare: „*Orice dicţionar etimologic e unilateral, căci el ţine seamă numai de originea, nu şi de circulaţia cuvintelor în limbă. Într-un dicţionar etimologic cuvinte cunoscute şi întrebuinţate zilnic de fiecare român, din orice parte a ţării, ocupă o unitate de loc, întocmai ca vorbele întrebuinţate numai în câte o regiune – şi acolo foarte rar – şi necunoscute tuturor celorlalte ţinuturi*” (Puşcariu, 1976, p. 181). În aceeaşi lucrare (Seche, 1966), este menţionat şi un posibil reper al aproximărilor sau o sursă probabilă a estimărilor false ale lui A de Cihac, respectiv o altă «statistică» din 1840 a lingvistului rus I. Hinculov, de unde se pare că s-ar fi inspirat controversatul A. de Cihac în ceea ce priveşte procentajul diverselor elemente etimologice ale lexicului.

Lingvistul care a sesizat prima oară falsitatea fundamentală a statisticii lui A. de Cihac a fost savantul român Bogdan Petriceicu Haşdeu, care a dezvoltat teoria circulaţiei ca o adaptare a teoriei circulaţiei în economie, aplicând-o la cuvintele din lexicul limbii române. Bogdan Petriceicu Haşdeu a arătat în mod lămurit că frecvenţa sau circulaţia cuvintelor are o importanţă decisivă în stabilirea fizionomiei lexicale a unei limbi. Criticând clasificarea etimologică făcută de A. de Cihac, pentru care: „*L'élément latin de la langue roumaine ne représente guère aujourd'hui qu'un cinquième de son vocabulaire, tandis que l'élément slave y entre pour le double ou pour*

2/5 à peu près…" (Haşdeu, 1984, p. 73). Bogdan Petriceicu Haşdeu a relevat defectul unor false «statistici» care pun pe acelaşi plan cuvinte inegale ca volum semantic şi ca forţă circulatorie: „*Dicţionarul nu ne dă şi nu ne poate da circulaţunea limbei; şi tocmai acesta este punctul cel esenţial*". Dimitrie Macrea a realizat în 1942 o altă statistică bazată pe cuvintele cuprinse în *CADE* (*Dicţionarul enciclopedic ilustrat „Cartea Românească"*. Partea I: *Dicţionarul limbii române din trecut şi de astăzi* de I.-A. Candrea. Partea II: *Dicţionarul istoric şi geografic universal* de G. Adamescu, Bucureşti, 1926-1931), a cărui concluzie, considerată cel puţin uluitoare, după părerea regretatului academician Alexandru Graur) era aceea că elementele latine reprezintă 20,58% din total, cele slave 16,41%, cele franceze 29,69%, iar restul de 33,32% ar fi format din cuvinte avându-şi obârşia în limbi care nu contează procentual prea mult ori din cuvinte cu origine necunoscută. Este vorba aşadar despre semnificaţia crucială, recunoscută unanim în lingvistica generală, a lexicului fundamental al unei limbi sau partea ei cea mai rezistentă şi mai importantă, însuşi *nucleul* lexicului limbii; această noţiune se opune *masei vocabularului* sau *vocabularului secundar* şi pentru ea se folosesc mai multe denumiri, cum ar fi: *vocabular de bază, vocabular fundamental, vocabular esenţial, fond principal lexical, fond principal de cuvinte,* mai rar *fond lexical uzual*. (Hristea et al., 1984, p. 14). Rolul şi valoarea pe care le au diferitele elemente aflate în componenţa lexicală a limbii române pot fi mai bine precizate dacă se au în vedere datele pe care le oferă unele statistici ulterioare, efectuate pe baza unor lucrări de o mai mare profunzime, în aceeaşi măsură filologică, dar şi statistică.Într-o astfel de cuantificare statistică întocmită de Sever Pop, în anul 1948, asupra *Dicţionarului limbii române din trecut şi de azi* de I. A. Candrea (apărut în Editura *Cartea Românească*, Bucureşti, 1931), se constată ca numărul cuvintelor de origine latină se ridică la 8.800, la care trebuie să se adauge 14.000 de neologisme primite din limbile romanice, ceea ce dă un total de 22.800 de termeni (se impune o comparare cu situaţia din limba franceză, unde, după *Dictionnaire de l'Académie*, 1878, din 32.000 de cuvinte, 20.000 erau de origine savantă sau străină şi numai 12.000 constituiau cuvinte franţuzeşti de origine populară). Numărul cuvintelor de origine slavă (preluate din slava comună ori din limbi slave moderne ca bulgara, sârba, ruteana, rusa, polona) era de asemenea mare, de aproape 7.800, dar mulţi dintre aceşti termeni sunt căzuţi în desuetudine sau sunt termeni de uz regional, restrâns. O interesantă cercetare statistică a Mihaelei Bîrlădeanu, realizată în lucrarea *Structura etimologică a două vocabulare reprezentative: român şi francez*, apărută în *Studii şi cercetări lingvistice*, nr. 6/1983, aceasta compară vocabularele reprezentative ale limbilor română şi franceză, găsind următoarea structură etimologică pentru vocabularul reprezentativ al limbii noastre: latină: 1. moştenite: 30,45%, 2. savante 1,77%; formaţii interne 24,81%; substrat 0,96%; superstrat vechi slav 8,91%; neogreacă 1,11%; împrumuturi din limbile slave moderne 1,80%, împrumuturi romanice din: franceză 7,64%, italiană 0,54%; turcă 0,73%; maghiară 1,27%; germane 0,27%; engleză; onomatopeice 0,23%; etimologie multiplă 17,36%; origine incertă 2,08%. Dintre lucrările relativ recente de cuantificare şi de analiză statistică în domeniul filologic, câteva aparţin chiar autorilor articolului (Manea, 2004; 2009; Manea, Săvoiu, 2014). Metoda analizei statistice frecvenţiale a constituit metoda majoră aplicată într-unul dintre segmentele principale ale volumului *Structura etimologică a vocabularului neologic (cu specială referire la anglicismele din limba română)*, apărut în 2004, cât şi în Încercare statistică asupra ortografierii cu â *şi* î , în 2009.

3. Metodologie

Seria enumerativă este cea mai simplă formă de prezentare statistică a unei populaţii grupate după cele mai banale soluţii criteriale în universul noncantitativ, respectiv în lumea cuvintelor, specifică filologului sau lingvistului (Săvoiu, 2012). O serie de acest tip poate fi constituită din cea mai simplă listă a prenumelor şi numelor unor oameni, grupaţi după un anumit criteriu organizatoric, administrativ, structural, spaţial, temporal etc. În urma valorificării metodei grupării statistice a acestei populaţii în raport cu prenumele sau numele rezultă o serie de repartiţie sau de distribuţie de frecvenţe ale prenumelor sau numelor identificate ca distincte, serie cunoscută şi ca *serie homogradă sau atributivă nenumerică* primul şir fiind enumerativ sau calitativ iar cel de-al doilea cantittiv sau numeric (Săvoiu, 2003, p.118). *Într-o manieră simplificată de redare orice serie statististică enumerativă* prelucrată prin grupare devine o serie atributivă nenumerică - cu referire la primul şir care este esenţial fiind cel discriminant) – iar cel de-al doilea şir devine numeric ca urmare a faptului că s-au atribuit frecvenţe de apariţie unor cuvinte, expresii, calificative, ierahizări etc (Săvoiu, et al, 2006). În figura 1 sunt descrise *într-un model general* câteva serii atributive nenumerice sau enumerative uzuale:

Forme de existenţă a seriei statistice atributive nenumerice sau enumerative prelucrate prin grupare

Fig. nr. 1

Varianta (x_i)	cuvânt	expresie	calificativ	ierahie	corect/incorect	da/nu
Frecventa (n_i)	n_1	n_2	n_3	n_4	n_5	n_6

Sursa: Realizat de autori

Seria statistică de acest gen cuprinde două şiruri paralele, primul de variante ale variabilei atributive nenumerice sau enumerative discriminate după un criteriu lingvistic, filologic, ierarhic, de evaluare, de cunoaştere etc. iar cel de-al doilea de frecvenţe de apariţie a variantelor din primul şir. Varianta (x_i) poate reprezenta un cuvânt (substantiv, prenume, verb, adjectiv, adverb, etc), un substantiv (nume propriu uzual sau nu, un obiect comun) o formă de plural, o formă verbală, o formă de redactare sau pronunţare în uz, o soluţie de scriere de uz local sau general, o sursă de intrare dintr-o altă limbă a unui nelogism, o formă corectă sau incorectă de pronunţare sau redactare, un calificativ, o nuanţă sau un calificativ ierahic, un nivel de acoperire, un nivel de acceptare sau refuz, un raspuns mono sau plurisilabic (da, nu sau da, nu, nu ştiu/nu răspund), un mod de acceptare sau inacceptare etc. Frecventa (n_i) este o informaţie numerică rezultată din agregarea cantitativă variantelor identice, suma tuturor frecvenţelor reunind totalitatea cazurilor din serie.

Un exemplu la îndemână de serie enumerativă dintre cele mai simple şi relevante prin simpla ei constituire, exemplu redescoperit *în ultimele trei decenii*, considerat celebru pentru elitismul celor înregistraţi, dar şi pentru asocierea lor cu militantismul antireligios sau ateismul de tip comunist *în România*, este acela al seriei de personalităţi incinerate (arse) *şi nu* înhumate tradiţional ortodox (fără a respecta ritualuarile ortodoxe ale privegherii decedatului, *înmormântării şi pomenirilor devenite cu trecerea timpului pomeni* cu un conţinut dintre cele mai inexplicabile cu putinţă)... Nu numai legal, dar şi biblic, incinerarea constituie alături de înhumare o modalitate similară ca finalitate, prin care toţi aceia care mor se întorc în țărână, prima descriind un proces accelerat de descompunere a unui cadavru prin ardere şi transformare a unui corp uman în cenușă, iar cea de-a doua un proces mult mai lung în ani, unul natural de metamorfozare a omului în țărână ca o firească "întoarcere în pământul din care eşti luat; căci pământ eşti şi în pământ te vei întoarce". (*Biblia sau Sfânta scriptură*, 2001, versiune redactată și adnotată de Bartolomeu Valeriu Anania Facerea 3:19, p.22).

Această serie simplă enumerativă cuprinde ca variabile atributive nenumerice atât numele şi prenumele a 2153 de personalităţi incinerate, cât şi profesiunile acestora *şi prin gestul* implicit al incinerării. Seria denotă absenţa credinţei religioase ortodoxe în cazul tuturor celor înscrişi în cadrul acestei liste a indivizilor într-o societate românească declarată la recensăminte permanent dominant ortodoxă. Lista românilor consideraţi altfel decât majoritatea locuitorilor României ascunde aparent mai multe informaţii interesante, pe care le poate dezvălui prompt dacă este prelucrată statistic elementar sau supusă unei analize frecvenţiale şi transformată într-o serie atributivă nenumerică. În termenii statisticii clasice, o listă simplă de persoane incinerate se transformă

într-o *serie homogradă* a profesiunilor sau ocupaţiilor, amalgamate din păcate ***în corpul acesteia. Seria homogradă rezultată*** ofer***ă*** nu doar simple statistici, ci veritabile paradoxuri contrastând cu opinia comună despre procesul şi scopul incinerării aşa cum se desprinde din analiza celor care au apelat ceva mai frecvent la această soluţie ...

Incinerarea a devenit istoric o practică recunoscută oficial prin legile României, având statut juridic legal ca ***şi*** înhumarea, un statut tot mai actual *în condiţii de pandemie Covid-19, trăite în* prezent cu mare intensitate şi de populaţia *ţării noastre.* Incinerarea în raport cu înhumarea oferă unele avantaje nu numai sanitare, ci mai extinse, asigurând o ieşire "*mai demnă din scena vieţii*" într-un context viciat de impactul unor boli infecţioase tot mai agresive, de boli transmisibile sau contagioase apte să genereze epidemii şi chiar pandemii, de accidente ce desfigurează cadavrele, de rămăşitede corp uman grav afectate de războaie, explozii, inundaţii etc.: i) presupune costuri mult mai mici decât cele ale înhumării pe durată mai mare de timp eliminând costurile tradiţiilor clasicei înhumări**;** ii) este o practică mult mai estetică şi ecologică; iii) induce un sentiment al egalităţii între oameni în final; iv) extinde cultul morţilor, dilatând accentul pus pe suflet comparativ cu cel pus pe trup; v) urna cu cenușa umană poate fi păstrată acasă sau îngropată în mormântul oricărui cimitir; vi) incinerarea presupune respectarea voinţei exprimate liber de persoana decedată de a nu constitui ulterior mai mult decât o amintire şi nu o obligaţie pentru familie etc. (http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati).

Fără a intenţiona *în niciun fel* o promovare a inciner*ării* în detrimentul înhumării clasice sau a face reclamă unor societăţi comerciale al căror obiect de activitate este incinerarea şi nu înhumarea, o analiză statistică a impactului celi dintâi *în Europa relevă* că peste 2/3 ajungând chiar şi la 70% din oameni aleg soluţia inciner*ării,* în Marea Britanie, Suedia, Danemarca, Cehia, Ungaria, atingând chiar proporţii optim paretiene în Elveția (85%). Se identifică multe ierarhii frecvenţiale surprinzătoare sau aparent neaşteptate ***şi în*** seria enumerativă a incineraţilor în România, conform datelor disponibile online la http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati *şi prelucrate de autori. Seria homogradă rezultată* este prezentată parţial prin segmentarea în raport cu profesiunea sau ocupaţia, declarată de familiile celor incineraţi în două subgrupe limitative ce oferă informaţii statistice simple axate pe frecvenţe maxime şi minime (Tabel 1).

Frecvenţele minime sau apariţiile cele mai rare (stânga) şi cele maxime sau cele mai dese (dreapta) calculate din seria enumerativă a incinerărilor în România

Tabel 1

Profesiune/ocupaţie	Frecvenţă minimă (n_i)		Profesiune/ocupaţie	Frecvenţă maximă (n_i)
Oameni de afaceri	2		Profesori	333
Politolog	2		Generali	164
Piloţi de curse	2		Scriitori	158
Tipografi	2		Actori	125
Preot ortodox, deputat	1		Ingineri	111
Bancher	1		Cercetători	109
Statistician	1 (Mircea Biji)		Militanţi comunisti	108

Sursa: Realizat de autori prin prelucrarea frecvenţelor de apariţie la: http://www.incinerareamurg.ro/romani-celebri-care-au-fost-incinerati

Sistemul de indicatori statistici pentru analiza frecvenţială într-o serie de repartiţie sau distribuţie de tip homograd este alcătuit dintr-o diversitate de componente: "*frecvenţe absolute (n_i), frecvenţe relative (n^*_i), frecvenţe cumulate crescător ($n_i\uparrow$ sau $n^*_i\uparrow$) sau descrescător ($n_i\downarrow$ sau $n^*_i\downarrow$), densităţi de repartiţie a frecvenţelor (n_i/h_i sau n_i^*/h_i). Frecvenţele cumulate crescător sau descrescător permit identificarea cuantilelor (Cv) din a căror numeroasă familie fac parte: mediana (Me), quartilele (Q_1, Q_2, Q_3), decilele (D_1, ..., D_9) şi centilele (C_1, ..., C_{99}), care împart populaţia statistică în două, în patru, în zece şi într-o sută de părţi egale*" (Săvoiu, et al., 2020, p. 130).

O varietate de concepte şi instrumente *statistice* sunt aplicabile și, aşa cum s-a dovedit adesea, foarte utile practic în toate domeniile de cercetare și în toate tipurile de abordări științifice. Acest acest adevăr este dificil sau chiar imposibil de contestat în filologie sau lingvistică, pornind de la valorificarea sistemului de indicatori frecvenţiali, trecând prin coeficienţii de concentrare şi diversificare Herfindahl – Hirschman şi Gini – Struck, iar în final valorificând inovativ profilul statistic în filologie sau lingvistică, în cazul concret al unui (mini)corpus lingvistic dezpovărându-l de ambiguităţi şi inceritudini legate de noramre, uz, corectitudine etc. conceptele care sunt supuse în continuare observării, cuantificării și analizelor statistice frecvenţiale sunt de dimensiuni mici cu rol mai degrabă metodologic şi de amplificare a rolului crosdisciplinarităţii (dimensiunea lor fiind cuprinsă între 25 şi 50 de itemuri, aproape toate fiind termeni aparținând vocabularului științifico-tehnic al limbii engleze contemporane (aşa cum reiese şi din tabelele detaliate ale rezultatelor). Motoarele de căutare utilizate au fost *Google* şi *Ask*, iar corpusurile de cuvinte (texte) accesate au fost constituite din material (texte, articole etc.) de tip academic, găsite pe Internet.

O problemă metodologică majoră cum este aceea a subiectivismului sau a falsificării rezultatelor unei investigaţii *directe* într-un (mini)corpus filologic, într-un anumit sens dorit de cel ce realizează construcţia (mini) corpusului filologic şi prelevarea de eşantioane, observate şi prelucrate ulterior, nu poate fi eliminată în lipsa onestităţii cercetătorului, indiferent de gradul de corectitudine a analizei frecvenţiale sau de prelevare a eşantionului cu respectarea extracţiilor aleatore ale fiecărui cuvânt sau expresii pornind de la probabilităţi practice cunoscute matematic. La toate aceste atitudini legate de aplicarea incorectă a teoriei sondajului, acţiuni premeditate sau nu de coordonatorii unor cercetări directe ce devin astfel neştiinţifice sau simple opinii filologice sau lingvistice, se adaugă şi existenţa unor premise permanente de apariţie a efectului Hawthorne. Într-o cercetare direct sau pe teren, efectul Hawthorne este un efect clasic cunoscut ca formă de reactivitate psihologică a respondenţilor prin care subiecţii unei cercetări parţiale sau experimentale îşi modifică anumite aspecte ale comportamentului lor, în uzul limbajului cu impact filologic sau lingvistic, ca urmare a faptului conştientizat că sunt studiați fără ca această atitudine să constituie un răspuns la manipuări cu scopuri subiective (Sesardić, 2018, p. 21).

Categoriile cele mai importante de erori metodologice grave, legate de apariţia în cercetări *indirecte* sau *documentare*, axate pe analize frecvenţiale pot fi considerate: i) erori cauzate de tehnica de generare de (mini)corpusuri filologice, respectiv ca urmare a inadecvării formei sau tehnicii de eşantionare aleatoare, dirijată sau mixtă; ii) erori metodologice sau sistematice, în strictă conexiune cu dificultăţile de prelevare şi cuantificare din Internet, tehnici cu rezultate eronate atâta timp cât sunt lipsite de instrumente şi softuri de control sau verificare a înregistrărilor specifice, de la forme de plural, la forme verbale, de la expresii, la exprimări integrale sau abrevieri specifice etc; iii) erori cauzate de neîndeplinirea constrângerilor de programare temporară spaţială şi structurală a căutărilor; iv) erori generate de limitările diferentiate ale motoarelor de căutare; v) erori provocate de algoritmii de optimizare a căutărilor; vi) erori de pre-procesare inadecvată utilizată pentru a reduce dimensiunea spațiului soluției şi în final rezultatele cercetărilor la cele validate, credibile, verificate etc.

4. Analize frecvenţiale statistice în (mini)corpusuri filologice

Analiza statistică frecvenţială statică (absolută) este practic cea mai simplă modalitate de aplicare a unei metode crosdisciplinar şi investigativ cu scopul de a soluţiona o problemă filologică (lingvistică), cum poate fi aceea a identificării pluralului gramatical corect conform uzului dominant într-un (mini)corpus filologic sau într-o bază de date distinctivă (Internet). În tabelul 2 este prezentată o astfel de analiză cu relevanţa şi irelevanţa ei specifică, pornind de la ideea de confruntare suplimentară pe principii statistice a cel puţin două motoare de căutare în cadrul aceloraşi (mini)corpusuri.

Analiză statistică frecvenţială statică a unui plural gramatical corect conform uzului majoritar într-un (mini)corpus filologic (Internet)

Tabel 2

Cuvinte cu forme multiple de plural în limba engleză	Rezultate obţinute cu motoare de căutare		Observaţii
	Google (search)	Ask (search)	
apsides	96,300	8,330	
apses	427,000	76,300	Relevant
apsises	12,200	-	
octopuses	651,000	225,000	Relevant
octopodes	124,000	12,800	
octopi	523,000	133,000	
addenda	5,700,000	430,000	Relevant
addendums	468,000	105,000	
addendas	115,000	-	
criteria	445,000,000	42,800,000	Relevant
criterions	511,000	131,000	
criterias	679,000	244,000	
antennae	7,570,000	832,000	***Irelevant****
antennas	2,840,000	4,770,000	
apexes	416,000	67,600	
apices	607,000	193,000	Relevant
apparatus	169,000,000	12,700,000	Relevant
apparatuses	7,670,000	515,000	
appendixes	2,580,000	264,000	
appendices	17,400,000	2,150,000	Relevant
aquariums	29,300,000	3,210,000	Relevant
aquaria	9,790,000	1,260,000	
automatons	528,000	1,260,000	
automata	29,200,000	1,450,000	Relevant
bureaux	73,700,000	1,150,000	***Irelevant*****
bureaus	30,600,000	3,210,000	
cerebellums	44,800	6,510	
cerebella	393,000	50,200	Relevant
curricula	17,700,000	2,440,000	Relevant
curriculums	6,860,000	694,000	
formulas	57,000,000	8,200,000	Relevant

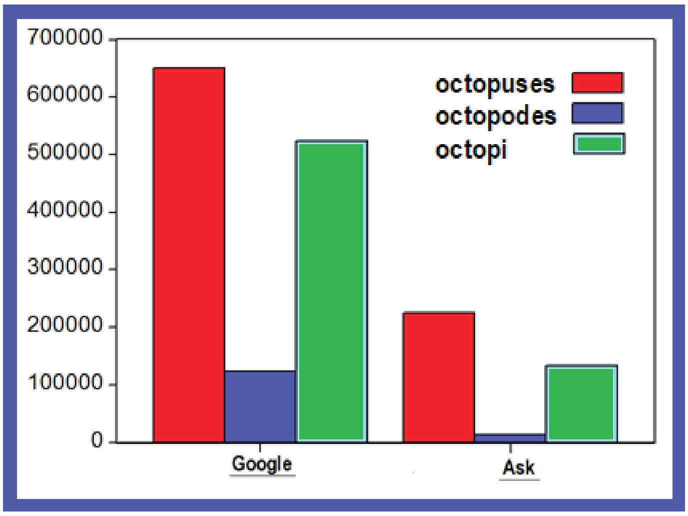
formulae	13,100,000	1,940,000	
genera	96,300,000	5,450,000	Relevant
genuses	117,000	15,900	
hiatuses	279,000	42,500	
hiatus	34,900,000	4,140,000	Relevant
maximums	2,860,000	407,000	
maxima	131,000,000	6,640,000	Relevant
minimums	8,030,000	1,080,000	
minima	63,500,000	1,340,000	Relevant
nuclei	22,200,000	4,050,000	Relevant
nucleuses	73,200	8,270	
phenomena	68,900,000	10,800,000	Relevant
phenomenons	462,000	99,900	
syllabuses	532,000	144,000	
syllabi	4,930,000	799,000	Relevant
strata	65,600,000	4,060,000	Relevant
stratums	336,000	24,200	
vortexes	486,000	113,000	
vortices	2,610,000	448,000	Relevant

Sursa: Preluat de autori dintr-o investigaţie proprie mai veche (Manea, Săvoiu, 2014, pp.13-17).

Evaluând static sau absolut distribuțiile de frecvență ale pluralului gramatical adecvat în conformitate cu uzul sau utilizarea specifică a majorității subiecților într-un (mini)corpus filologic (Internet) pentru cuvinte ca *apse*, *addendum*, *antenna*, *apex*, *apparatus*, *appendix*, *automaton*, *bureau*, *criterion*, *cerebellum*, *curriculum*, *formula*, *genus*, *hiatus*, *maximum*, *nucleus*, *octopus*, *phenomenon*, *syllabus*, *stratum*, *vortex*), s-au identificat date relevante pentru majoritatea absolută (cu doar două excepții: *antenna* și *bureau*), s-au putut trage concluzii practice valoroase asupra utilității metodei analizei frecvenţiale, abordată prin cel puţin două motoare de căutare. Reprezentările grafice ale unor valori frecvenţiale comparate pot contribui printr-o mai bună vizibilitate la aprecierea pluralului relevant şi probabil implicit corect, acolo unde informaţia cantitativă este relevantă, detaliind frecvenţele absolute după mai multe motoare de cautare (de exemplu: Google si Ask).

Analiză statistică grafică comparată a unui plural gramatical cu două motoare de căutare într-un (mini)corpus lingvistic (Internet)

Fig. nr. 2



Sursa: Preluat de autori după o lucrare proprie mai veche (Manea, Săvoiu, 2014, pp.13-17)

Figura 2 subliniază că un grafic care compară distribuțiile formelor de plural studiate şi cu ajutorul a două motoare de căutare poate fi mai util, prin vizibilitate superioară, în ceea ce priveşte evaluarea ipotezelor referitoare la adecvarea gramatical-fonetică a uneia dintre cele trei forme de plural, analizate statistic frecvenţial. Informațiile cantitative relevante asigură că şi frecvenţa ocurenţelor va deveni relevantă, iar dacă apar practic stări de ambiguitate se poate apela firesc la mai multe motoare de căutare (Google și Ask).

În analize complexe se poate trece la o subîmpărțire a listei extinse de termeni specializați sau tehnici (dintre care unii au fost științifici doar inițial), după cum: a) ilustrează o chestiune legată de morfofonematică; b) ilustrează o simplă problemă de ortografie; c) ilustrează o categorie de termeni care nu pot fi numiţi chiar termeni tipic tehnici sau științifici, deși sunt fără îndoială termeni livrești.

“Iată câteva exemple de termeni din subcategoriile deduse din lista extinsă: a) *antenna*, *apex*, *apparatus*, *appendix*, *automaton*, *cactus*, *calyx*,*cerebellum*, *cerebrum*, *cicada / cicala*, *colloquium*, *cranium*, *criterion*, *curriculum*, *dilettante*, *discus*, *fauna*, *flora*, *formula*, *fungus*, *genus*, *hiatus*, *iambus*, *larynx*, *libretto*, *memorandum*, *novella*, *nucleus*, *palazzo*, *phenomenon*, *radius*, *radix*, *retina*, *rhombus*, *stratum*, *syllabus*, *tableau*, *tempo*, *trapezium*, *vacuum*, *vertebra*, *vertex*, *vortex*; b) *bureau*, *flamingo*, *fresco*, *grotto*, *halo*, *manifesto*, *memento*, *motto*; c) *aquarium*, *candelabrum*, *cicerone*, *colossus*, *focus*, *grotto*, *gymnasium*, *hippopotamus*, *maximum*, *millennium*, *minimum*, *narcissus*, *persona grata*, *referendum*, sanatorium, *symposium*, *terminus*, *ultimatum*" (Manea, Săvoiu, 2014).

Observațiile de detaliu pe care le-au putut face autorii în sprijinul ideii de concretizare faptică a unor ipoteze privind recurența cuvintelor în (mini) corpusuri de date încă din 2014, au relevat, că există de pildă foarte puțini termeni care au trei forme de plural dovedibile (viz. *octopus* „caracatiță"- pl. *octopuses*, *octopodes*, *octopi*, deși ultima formă este improprie / nerecomandată). La fel de instructive pot deveni şi alte investigaţii axate pe o căutare asemănătoare privind distribuția substantivelor *apsis* „*absidă*" (pl. *apsides, apses, apsises*), *addendum* „*adendă*" (pl. *addenda, addendums, addendas*) și *criterion* „*criteriu*" (pl. *criteria, criterions, criterias*), posibil și *frustrum*.

În destule cazuri, marcarea uzului lingvistic, așa cum este înțeles și ilustrat de dicționarele anglo-americane, a fost într-atât de importantă încât lexicografii respectivi au găsit de cuviință să gloseze (*MacMillan*) forme de plural neregulate drept *leme* legitime, de sine stătătoare (de exemplu, *bacteria*, *criteria*, *alga*, *data*). Când însă se ia în considerare și problema sensului, procedura în sine pare să își dovedească neputința totală: cum am putea verifica semnificația și utilizarea fiecărei ocurențe care apare în texte sau (mini)corpusuri pe care s-au făcut căutări? [ex: *genius* (pl. *geniuses* vs. pl. *genii*), *domino* – pl. *dominoes* („*joc*") / *dominos* („*mantie*"), *index* – pl. *indexes*/*indices*, *stamen* – pl. *stamens*/ *stamina*, *milieu* – pl. *milieus*/ ranc. *milieux* [‘mi:ljø], *calculus* – pl. *calculuses*; med. *calculi* [‘kælkjulai], *polypus* – pl. *polypi*; *data* (pl., deși de obicei considerat nenumărabil) → sg.*datum*, *agenda* (pl. lui *agendum*; azi considerat sing.].

Analiza statistică frecvenţială dinamică şi comparată instrumental descrie o modalitate mai evoluată de investigaţie crosdisciplinară cu acelaşi rol de identificare a pluralului gramatical corect conform uzului dominant abordat în două momente diferite ale evoluţiei unui (mini)corpus filologic sau o bază de date distinctivă (Internet). În tabelul 3 este prezentată această analiză pe acelaşi fond tematic al pluralului corect (2014) dar de o manieră repetitivă (2020), cu un calcul suplimentar al *indicelui evolutiv al relevanţei aprecierii:*

Analiză statistică frecvențială dinamică a pluralului gramatical corect conform uzului majoritar într-un (mini)corpus filologic (Internet)

Tabel 3

Cuvinte cu forme multiple de plural în limba engleză	Rezultate obţinute cu motoare de căutare în ani diferiţi				Observaţii
	2014		2020		
	Google (search)	Ask (search)	Google (search)	Indice evolutiv (%)	
apsides	96300	8330	127000	131,88	
apses	427000	76300	**637000**	**149,18**	Relevant
apsises	12200	-	*5680*	*46,56*	
octopuses	651000	225000	**4450000**	**683,56**	Relevant
octopodes	124000	12800	162000	130,65	
octopi	523000	133000	2130000	407,27	
addenda	5700000	430000	**8560000**	**150,18**	Relevant
addendums	468000	105000	1170000	250,0	
addendas	115000	-	200000	173,91	
criteria	445000000	42800000	**698000000**	**154,83**	Relevant
criterions	511000	131000	1060000	207,43	
criterias	679000	244000	2060000	303,39	
antennae	7570000	832000	8890000	113,21	
antennas	2840000	4770000	**74100000**	**26091,55**	**Relevant
apexes	416000	67600	597000	143,51	
apices	607000	193000	**2100000**	**349,42**	Relevant
apparatus	169000000	12700000	**190000000**	**112,43**	Relevant
apparatuses	7670000	515000	5320000	69,36	
appendixes	2580000	264000	3110000	120,54	
appendices	17400000	2150000	**22800000**	**131,03**	Relevant
aquariums	29300000	3210000	**59100000**	**201,71**	Relevant
aquaria	9790000	1260000	10600000	108,27	
automatons	528000	1260000	1370000	259,47	
automata	29200000	1450000	**52100000**	**178,42**	Relevant
bureaux	73700000	1150000	72800000	98,78	
bureaus	30600000	3210000	**118000000**	**385,62**	*Relevant
cerebellums	44800	6510	65500	146,21	
cerebella	393000	50200	**460000**	**117,05**	Relevant
curricula	17700000	2440000	**28000000**	**158,19**	Relevant
curriculums	6860000	694000	8100000	118,08	
formulas	57000000	8200000	**155000000**	**271,93**	Relevant
formulae	13100000	1940000	52000000	396,94	
genera	96300000	5450000	**156000000**	**161,99**	Relevant
genuses	117000	15900	238000	203,42	
hiatuses	279000	42500	396000	141,94	
hiatus	34900000	4140000	**59400000**	**170,20**	Relevant
maximums	2860000	407000	3930000	137,41	
maxima	131000000	6640000	**548000000**	**418,32**	Relevant
minimums	8030000	1080000	9210000	114,70	
minima	63500000	1340000	**276000000**	**434,65**	Relevant
nuclei	22200000	4050000	**34800000**	**156,76**	Relevant
nucleuses	73200	8270	96600	131,97	
phenomena	68900000	10800000	**112000000**	**162,55**	Relevant
phenomenons	462000	99900	1210000	261,90	
syllabuses	532000	144000	1500000	281,95	
syllabi	4930000	799000	**6080000**	**123.33**	Relevant

strata	65600000	4060000	**66500000**	**101,37**	Relevant
stratums	336000	24200	928000	276,19	
vortexes	486000	113000	1020000	209,88	
vortices	2610000	448000	**4260000**	**163,22**	Relevant

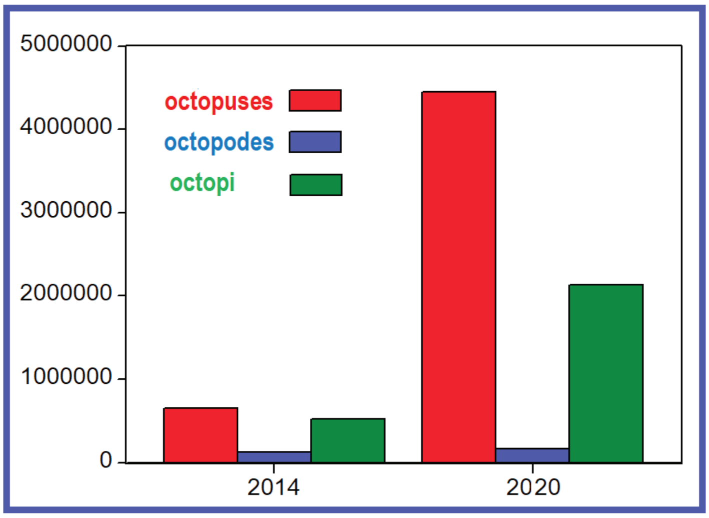
Sursa: Realizat de autori pentru coloanele referitoare la 2020 şi după (Manea, Săvoiu, 2014).
Notă*: Agregarea în timp a datelor frecvenţiale transformă situaţiile irelevante în relevante într-o manieră dacă nu promptă cel puţin într-una neaşteptată.
Notă **Căutarea **antennae** a returnat 8890000 rezultate fiind devansată de căutarea **antennas** care a returnat 74100000 rezultate, conferind o extensie în uz multiplicată maximal în 6 ani de 260 de ori, respectiv mai mult cu **26091,55 – 100,00 = 25991,55 %, în** raport cu trecutul.

Timpul ucide irelevanţa iniţială şi, aşa cum se poate constata, se obţin soluţii pentru toate exemplele, şi chiar mai mult se pot identifica dinamici diferite conform analizei statistice frecvenţiale. Astfel, unele cuvinte sau forme de plural exemplificate şi în tabelul 2 şi în 3 deţin o relevanţă mult mai mare ca urmare a unei explozii în uzul acestora în (mini)corpusul investigat (Internet). Reluând analiza frecvenţială în cazul triplului plural potenţial: ***octopuses, octopodes şi octopi*** se identifică, din datele finale, trei dinamici complet diferite, care reconfirmă relevanţa formei ***octopuses,*** dar se mai constată şi o accelerare a polarizării între formele ***octopuses şi octopi,*** ceea ce oferă dovezi lingvistice ale continuităţii uzului lor lingvistic accentuat.

Reprezentările statistice grafice ale valorilor frecvenţiale comparate în timp în figura 2 contribuie la aprecierea pluralului relevant (corect), printr-o mai bună vizibilitate, acolo unde informaţia cantitativă este relevantă, detaliind frecvenţele absolute, după mai multe investigaţii realizate în momente diferite (ex. 2014 si 2020).

Analiză statistică grafică de tip comparat în două momente diferite în timp a variantelor unui plural gramatical într-un (mini)corpus lingvistic (Internet)

Fig. nr. 3



Sursa: Realizat de autori după datele din tabelul 3.

Calculul coeficienţilor de concentrare–diversificare, de tip HH (Herfindahl–Hirschman) sau C_{G-S} (Gini–Struck), descrie prin valorile ascendente (în cazul HH) sau care tind tot mai clar spre 1 (în cazul C_{G-S}) o concentrare în corpusul lingvistic (Internet) analizat sau într-un (mini)corpus lingvistic specific (o bază de date distinctivă). Limitele unei concentrări în exces sunt conferite de valorile de 0,667 (HH) şi 0,409 (C_{G-S}) pentru situaţia în care oferta lingvistică se rezumă la 3 soluţii (n = 3) aşa cum rezultă din calculele realizate în lucrarea publicată de Săvoiu, Crăciuneanu, Ţaicu, în 2010.

În tabelul 4 se constată că ***octopuses*** este apt prin uzul lingvistic cuantificat cu ajutorul analizei statistice frecvenţiale să reprezinte o soluţie unică de plural, generând valori peste limitele unei concentrări excesive în 2020 (atât după valoarea efectivă a coeficientului HH cât şi după aceea a C_{G-S}).

Analiză statistică frecvenţială dinamică a pluralului gramatical corect cu ajutorul coeficienţilor statistici de concentrare - diversificare

Tabel 4

Varianta	2014	(gi)	$(gi)^2$		2020	(gi)	$(gi)^2$	
octopuses	651,000	0,501	0,251	**HH = 0,650**	4,450,000	0,660	0,436	**HH = 0,732**
octopodes	124,000	0,096	0,009		162,000	0,024	0,001	
octopi	523,000	0,403	0,163		2,130,000	0,316	0,099	
Total	1,298,000	1,000	0,423	**C_{G-S} = 0,367**	6,742,000	1,000	0,536	**C_{G-S} = 0,551**

Sursa: Realizat de autori

Metoda statistică anterioară motivează frecvenţial riguros alegerea formei corecte *octopuses,* conform uzului lingvistic dominant (cu valori structurale g_i situate peste 0,5). Totodată, această metodă validând un proces de concentrare în exces, garantat de cuantificările coeficientului Gini-Struck (C_{G-S}) care epăşesc valori de 0,41 (Săvoiu, Crăciuneanu, Ţaicu, 2010, pp. 15-27) conduce în final la idea că analiza frecvenţială completată cu analiza concentrării – diversificării asigură o credibilitate mult mai mare decât aplicarea exclusivă a analizei frecvenţiale simple în filologie în investigarea variaţiei uzului lingvistic în timp şi spaţiu.

O altă valorificare a analizelor statistice frecvenţiale se poate realiza pentru identificarea sursei de unde a fost preluat un neologism, cu referire la literatura ştiinţifică ce l-a consacrat prin utilizare extinsă sau maximală. Pentru a exemplifica acest aspect, în tabelul 5 sunt sintetizate apariţiile unui neologism în limba română în paralel cu sinonimele (aproape omonime) din limbile franceză şi din engleză:

Analiză statistică frecvenţială a sursei neologismelor, preluate în română din franceză sau engleză, în cadrul unui (mini)corpus lingvistic (Internet)

Tabel 5

Limba română	Limba franceză	Limba engleză	Sursa (abrev.)
fenomen 30100000	*phénomène* 62300000	*phenomenon* 647000000	eng.
teoremă 12800000	*théorème* 3370000	*theorem* 60100000	eng.
dilemă 29600000	*dilemme* 3480000	*dilemma* 101000000	eng.

Sursa: Realizat de autori

Cu siguranţă că etimologia sau originea greacă a celor trei cuvinte exemplificate nu poate fi pusă în discuţie, dar analiza frecvenţială subliniază că aceşti termeni ştiinţifici deşi au intrat la noi prin limba franceză, conform uzului lingvistic aparţin cu siguranţă de literatura ştiinţifică engleză. În mod evident, nu poate fi abandonată etimologia sau originea cuvintelor, deoarece se poate astfel ajunge astfel la a deveni nu informat ci dimpotrivă prost informat. În cazul descris în tabelul 5, analizei statistice frecvenţiale nu i se pot aduce nici acuzaţii de rea informare şi cu atât mai puţin de dezinformare, aceasta probând cantitativ valorificarea limbajului ştiinţific într-o limbă sau în alta.

Cât de mare este utilizarea unor expresii incorecte după normele aparent în vigoare, fără să uităm că limba şi limbajul deţin o vitalitate extraordinară şi când începe ca dinamica lor să depăşească pe aceea a formelor considerate corecte? La o astfel de întrebare, tot o analiză cantitativă statistică de tip frecvenţial găseşte răspuns pornind de la echilibrul paretian 20/80 şi credem că, de îndată ce forma incorectă (F_{INC}) depăşeşte 20 % din utilizările totale agregate şi deţine un *indice evolutiv al relevanţei aprecierii* mai mare decât cea declarată corectă (F_{COR}), o competiţie evidentă se instalează în uzul celor două. În tabelul 6 sunt prezentate două exemple ale valorificării metodei analizei statistice frecvenţiale în identificarea unor competiţii lingvistice referitoare la uzul alternativ (corect-incorect) dar şi unul de concurenţă neinstalată între o formă consacrată şi alte două forme considerate toate împreună ca fiind corecte:

Analiză statistică frecvenţială a raportului paretian sau non paretian între forme corecte şi incorecte ale aceluiaşi cuvânt într-un **(mini)corpus filologic (Internet)**

Tabel 6

Frecvenţa F_{COR}	Frecvenţa F_{INC}	Uz agregat	Raport F_{INC}/F_{COR}
Preşedinţie 819000	*Preşedenţie* 78300	897300	8,7% vs 91,3%
Să aibă 25300000	*Să aivă + Să aibe* 2780000	28080000	9,9% vs 90,1%
Găluşti 45700	*Găluşte* 387000	*432700	10,6% vs 89,4%

Sursa: Realizat de autori.

*Notă : Ambele forme sunt considerate corecte, dar chiar şi aşa nu s-a depăşit limita de optim paretian care ar fi putut deschide o competiţie lingvistică reală legată de uzul lor alternativ.

Argumentaţia care justifică folosirea raportului optim de tip paretian porneşte de la conceptul de variaţie şi implicit de la cel derivat de omogenitate în cazul variabilelor alternative sau binare, unde limita coeficientului de omogenitate se atinge în evaluările de tip 20% cu 80% (DA vs NU). La valori mai mici de 80% comparative cu limita paretiană acestea probează posibilităţi certe de eterogenitate în cadrul oricărei populaţii, inclusive a unui (mini) corpus filologic.

Metoda profilului statistic, mai ales în varianta clasică de confruntare a unor profile statistice deţine o aplicabilitate ligvistică sau filologică potenţială destul de mare. Beneficiind de o mare putere de sinteză, metoda confruntării prin profile statistice poate sintetiza dialogul şi argumentaţia filologică în mod esenţial, aşa cum se poate constata din tabelul nr. 7.

Confruntarea filologică a profilurilor statistice realizată în acelaşi (mini) corpus filologic

Tabel 7

Profilul statistic fals al structurii limbii române realizat de A. de Cihac şi Sextil Puşcariu plagiat după lingvistul rus I. Hinculov	Profilul statistic al structurii limbii române realizat obiectiv şi echilibrat al lui Mircea Seche
Cuvinte de origine latină: 20,2%	Cuvinte de origine latină: 45,6%
Cuvinte de origine slavă: 41,0%	Cuvinte de origine slavă: 34,8%
Cuvinte de origine turcă: 16,7%	Cuvinte de origine turcă: 7,1%
Cuvinte de origine neogreacă: 11,0%	Cuvinte de origine neogreacă: 6,2%
Cuvinte de origine maghiară:10,2%	Cuvinte de origine maghiară: 5,8%
Cuvinte de alte origini: 0,9%	Cuvinte de alte origini: 0,5%

Sursa: Realizat de autori cu rol sintetic şi de confruntare finală.

Există, evident, mult mai multe metode statistice aplicabile în filologie sau lingvistică, ceea ce conferă un spaţiu cu mult mai larg de mişcare crosdisciplinarităţii şi chiar şi multidisciplinarităţii, iar studiile şi cercetările prezente şi viitoare vor dovedi cu siguranţă acest adevăr.

5. Concluzii

Articolul este rezultatul aplicării crosdisciplinare a metodelor statistice, autorii fiind călăuziţi de dorinţa de a identifica soluţii utile, fie ele uneori şi exclusiv parțiale, la câteva probleme controversate, cu rezultate obiective și fiabile, care să confirme sistematica logică și analogică a funcționării limbii naturale. Orice analiză frecvenţială de tip statistic aplicată adecvat în corpusuri şi (mini)corpusuri filologice conduce la final la sprijinirea utilizatorului comun al limbii, ale cărui capacităţi multiple de asociere anticipată, de comparare simplă cu scop de ierarhizare şi de confruntare cu scop de eliminare, pot fi permanent optimizate oferind soluţiile lingvistice probate şi dovedind prin argumente statistice uzul practic al acestora. În opinia autorilor, această direcție de cercetare crosdisciplinară poate aduce cu sine clarificări suplimentare privind adecvarea metodelor statistice la investigaţia filologică, prin valorificarea directă a instrumentelor frecvenţiale, a indicilor evolutivi ai relevanţei aprecierii uzului lingvistic, a coeficienţilor de concentrare - diversificare pentru a identifica pluralul dominant al unor anglicisme, sursele de unde au fost preluate unele neologisme care deţin deja un caracter internaţionalizat sau intensitatea competiţiei dintre formele corecte şi incorecte ale aceluiaşi cuvânt în uzul lingvistic (Săvoiu, et al., 2020, p. 146)

Autorii îşi propun să continue investigaţiile realizate cu altele similare prin crosdisciplinaritate, şi chiar să iniţieze într-o echipă mai largă ce tinde spre multidisciplinaritate, cercetări mai extinse prin multiplicarea metodelor, simultan cu extensia gradului de acoperire prin corpusuri de dimensiuni tot mai mari. Pe termen scurt, merită analizate, prin prisma uzului și a frecvenței de utilizare, variantele semantice hibride („*barbarismele*”) care apar frecvent în limba română, cel mai adesea împrumutate direct din limba engleză ori cel puțin calchiate după modele anglo-americane, cum ar fi: *oneros*, *intrepid*, *vocal*, *versatil*. Din păcate, nu se pot identifica nici chiar pe termen mediu soluţii optime de cercetare a modalităţilor distincte de pronunţare a unor astfel de termeni căutați, întrucât pronunția/fonetica nu poate fi înregistrată în textele din corpusurile mari cu acces universal de tip Internet (Săvoiu, et al., 2020, p. 147)

Singura viabilă în viitorul apropiat rămâne aşadar studierea *metodică* a (mini)corpusurilor filologice.

Ce lucruri utile ar mai putea intra în vizorul analizelor statistice frecvenţiale? Probabil că se pot identifica unele proiecte de lucru în echipă multidisciplinară în domeniul standardizării/ normării și al didacticii de tip comparativ-contrastiv devenite tot mai utile sau relevante. Analog simpla constituire a unor corpusuri de date, în principal, prin căutări efectuate pe multitudinea de texte găzduite de Internet, ar putea dovedi prompt şi concret, care este situația reală a unor cuvinte care prezintă dificultăți de uz, cum se face acordul verbului/predicatului cu subiectul sau folosirea unor prepoziţii sau conjuncții, precum şi prevalența sau nu, în uzul real al limbii, a anumitor forme de plural declate incorecte dar evolutive sau dinamice etc. Pornind de la modelul dicționarului compus de J. C. Wells, se pot realiza comparaţii lingvistice sau filologice internaţionale a unor cuvinte românești cu a celor similar ca sens, folosire de exemplu în limba engleză, tot prin evaluare statistică frecvenţială, dar şi structurări sau ponderări ale unor cuvinte ce încep cu o anumită literă de dicționar, precum şi cât la sută dintre substantive sau verbe au forme sau utilizări neregulate sau considerate „*aberante*”, pornind de la o riguroasă cuantificare statistică cu softuri tot mai performante.

O carte apărută destul de recent în cadrul unor cercetări crosdisciplinare şi pe alocuri chiar multidisciplinare (Ross, Greenhill, Atkinson, 2013), unde s-au folosit simultan şi genetica dar şi istoria poveştilor populare a expus un fapt interesant cu consecinţe legate de dinamica teritorială a uzului limbajului, constatând că poveştile sunt grupate geografic asemănător sau chiar similar genelor. Acest aspect ar putea permite determinarea ca areal în uz a limbii şi a graniţelor ei etnico-lingvistice, pornind pe urmele variantelor poveştilor, în paralel cu evoluţia ADN-ului, impactul limbii şi al poveştilor fiind mai mare decât acela al genelor sau cum exprima sintetic jurnalista australiană Christine Kenneally, în 2019: "*un cuplu de oameni î*şi pot mai uşor amesteca genele fără a împărtăşi aceeaşi limbă, cu mult mai uşor decât o poveste poate să treacă peste o barieră lingvistică*"* (Kenneally, 2019, p. 385).

"*Cultura şi limba sunt şi rămân solide în esenţa lor, în timp ce genele sunt lichide (Khan, 2013), iar o metodă statistică poate studia aparent mai uşor un fenomen stabil sau momentul (...), dar nu trebuie să uite să o facă* şi *de o manieră* [metodologică] *dinamică (la intervale periodice), deci cu intenţii generatoare sau dătătoare de vitalitate şi curgere filosofică (...)"* (Săvoiu, et al., 2020, p. 150)

Bibliografie

1. Graur, A., 1989. *Iarăşi â şi î*, in *Puţină gramatică*, vol. II, Bucureşti: Editura Academiei R.S.R.
2. Haşdeu, B. P. 1984. *Cuvente den bătrâni*, Bucureşti: Editura Didactică şi Pedagogică,
3. Hristea, T. (coord.) 1984. *Sinteze de limba română*, Bucureşti: Editura Albatros
4. Iordan, I., 1978. *Istoria lingvisticii româneşti* (Coordonator: acad. Iorgu Iordan), Bucureşti: Editura Ştiinţifică şi Enciclopedică.
5. Iordan, I., Robu, V. 1978. *Limba română contemporană*, Bucureşti: Editura Didactică şi Pedagogică, Bucureşti.
6. Kenneally, C., 2019. *Povestea secretă a speciei umane, Cum ne sunt modelate identitatea şi viitorul de ADN şi de istorie*, Bucureşti: Editura Humanitas
7. Khan, R. 2013. Why Culture is Chunky and genes are Creamy, *Gene Expression*.
8. Lombard, A., 1992. Despre folosirea literelor â şi î, *Limba română*, nr. 10. pp. 531-540.
9. *MacMillan Dictionary*, MacMillan, 2012
10. Macrea, D., 1943. Fizionomia lexicală a limbii române, *Dacoromania*, vol 10/2. pp. 362-373.
11. Manea, C. 1993. *Consideraţii statistice asupra ortografiei lui â din a şi î din i*, comunicare prezentată la Sesiunea de comunicări ştiinţifice „Rolul presei şi învăţământului în relansarea economico-socială a României", de la Universitatea din Sibiu, 21-22 mai 1993.
12. Manea, C. 1997. Structura etimologică a limbii române literare contemporane în lumina frecvenţei cuvintelor, *Studii şi cercetări lingvistice*, anul XVIII, nr. 2.

13. Manea, C. 2009. ***Încercare statistică asupra ortografierii cu â şi î***, în vol. *Distorsionări în comunicarea lingvistică, literară şi etnofolclorică românească şi contextul european*, Academia Română, Institutul de Filologie Română „*A. Philippide*", Iaşi: Editura Alfa, pp. 209-218
14. Manea, C., 2004. S*tructura etimologică a vocabularului neologic (cu specială referire la anglicismele din limba română)*, Piteşti: Editura Universităţii din Piteşti.
15. Manea, C., 2009. ***Încercare statistică asupra ortografierii cu â şi î***, în vol. *Distorsionări în comunicarea lingvistică, literară şi etnofolclorică românească şi contextul european*, Academia Română, Institutul de Filologie Română „A. Philippide", Ed. *Alfa*, Iaşi, pp. 209-218.
16. Manea, C., Săvoiu, G. 2014. *Frequency Analysis of the Use of a Number of Morphological Variants in Scientific Language*, in ESMSJ, vol. IV, no. 2 (special issue), pp. 13-17.
17. Marcu, F. 1997. *Noul dicţionar de neologisme*, Bucureşti: Editura Academiei Române.
18. Marcu, F. 2000. *Dicţionar uzual de neologisme*, Bucureşti: Ed. Saeculum I.O
19. O'Keeffe, A., M., McCarthy, J., and Carter, R. A. 2007. *From Corpus to Classroom*, Cambridge: Cambridge University Press.
20. Onu, L., 1992. Oportunitatea reformei ortografice, *Limba română*, nr. 4, pp. 199-207, Bucureşti.
21. Săvoiu, G., 2003. *Statistică generală. Argumente în favoarea formării gândirii statistice*, Piteşti: Editura Independenţa economică.
22. Săvoiu, G. (coord.), Asandei, M., Grigorescu, R., Manole, S. 2005. *Cercetări şi modelări de marketing. Metode cantitative în cercetarea pieţei,* Bucureşti: Editura Universitară.
23. Săvoiu, G., 2012. *Statistică generală cu aplicaţii în contabilitate,* Bucureşti: Editura Universitară.
24. Săvoiu, G., Crăciuneanu, V. Ţaicu, M. 2010. A New Method of Statistical Analysis of Markets' Concentration or Diversification. *Romanian Statistical Review*, vol. 58(2), pp. 15-27.
25. Săvoiu, G., et al. 2020. *Metode statistice şi crosdisciplinaritate*, Bucureşti: Editura Universitară.
26. Sesardić, N., 2018. *Când raţiunea pleacă în vacanţă. Filozofii în vacanţă***,** Bucureşti: Editura Humanitas.
27. *** *Biblia sau Sfânta Scriptură*. 2001. Ediţie jubiliară a Sfântului Sinod, versiune redactată și adnotată de Bartolomeu Valeriu Anania (ed.), Bucureşti: Editura Institutului Biblic şi de Misiune al Bisericii Ortodoxe Române, p. 22. Available on - line la: http://www.diacronia.ro/en/indexing/details/B347/pdf. Accessed on 30.04.2020.