

LINGVISTICA CANTITATIVĂ

Statistica cuvintelor

TONI HERNÁNDEZ ȘI RAMÓN FERRER I CANCHO

Traducere de Maria Cristina Montanari

CUPRINS

Capitolul 0

Știința în norii de cuvinte	7
Legea lui Zipf	11
<i>Legile potențiale sunt drepte pe scara dublu logaritmică</i>	14
Alte legi zipfiene	16
<i>Constantele A și B: de la Euler la Riemann</i>	18
<i>O derivație de la lege între numărul de semnificați și frecvență</i>	21

Capitolul 1

De la Zipf la teoria matematică a comunicației	23
Legi probabilistice	25
<i>Provocările Hapax Legomena (HL) și de traducere</i>	
<i>Hapax Legomena (THL)</i>	26
<i>Constantele a și b: ne întâlnim din nou cu Euler și Riemann</i>	28
Principiul minimului efort: unificare și diversificare	31
Teoria informației a lui Shannon (și Weaver)	33
<i>Valoarea minimă și valoarea maximă a entropiei lui Shannon</i>	36
<i>Contribuția nulă la entropia elementelor cu probabilitate zero</i>	40
<i>Combinând forțele de unificare și diversificare</i>	42
<i>Rata de entropie</i>	45

Capitolul 2

Brevitate, principiu de compresie și Chiquito de la Calzada	47
Universalitatea legii prescurtării și contraexemplul lui Chiquito	50
Principiul de compresie al teoriei informației	53
<i>Predicție a legii lui Zipf pentru frecvență și a legii brevității</i>	60

Capitolul 3

Mai mult decât întâmplare	65
Critica adusă lui Zipf	65
Actul 1: Dactilografia aleatorie și legile statistice ale limbajului	68
<i>Distribuția probabilității în dactilografia aleatorie</i>	70
Actul 2: Efortul dactilografiei aleatorii	72
Actul 3: Pe maimuțe reale	76

Capitolul 4

Vocabulare și limbi: masele formelor lingvistice	79
<i>O relație între legea lui Zipf și legea lui Herdan</i>	82
De la cuvinte la un copac al limbilor	83
Revoluția digitală	89
<i>Relație între exponenții lui Zipf de ranguri și frecvențe</i>	93
Proprietăți dinamice ale limbajului	94

Capitolul 5

Parcursuri stimulante	99
<i>Spanische Lautdauer</i> : de la fonetică la genetică	104
De la Herder la lingvistica cantitativă a lui Menzerath	111
<i>Legea Weber-Fechner</i>	118

Capitolul 6

Viitorul lingvisticii cantitative: mașinile?	123
Învățarea statistică	131
Înspre tehnologiile limbajului	135
<i>Modele de recunoaștere vocală</i>	136
Evaluare și provocări ale lingvisticii cantitative	137
Bibliografie recomandată	143

Capitolul 0

Știința în norii de cuvinte

Acum circa zece ani, pe când studiam lingvistica la Universitatea din Berlin, mi-a venit ideea că ar fi de bun augur să studiez limbajul ca un fenomen natural, mai mult sau mai puțin ca un fiziolog care ar studia bătaia inimii, sau un entomolog tropismele unei insecte, sau un ornitolog obiceiurile de cuibărire ale păsărilor. Cu alte cuvinte, limbajul ar trebui să fie considerat o formă specială de comportament, a unei insolite specii existente; ar trebui să fie studiat cum fac științele exacte, prin aplicarea principiilor statistice fenomenelor obiective ale limbajului.

Prefață la *Psihobiologia limbajului*, 19 decembrie 1934,

George Kingsley Zipf (1902-1950)

Al Doilea Război Mondial era pe sfârșite și la Newton, un orașel din Massachusetts, într-o cameră la etajul al treilea al unui edificiu, transformată în birou, o lumină rămânea aprinsă până la primele ore ale dimineții, atrăgând atenția vecinilor. Era camera lui **George Kingsley Zipf** (1902-1950)), profesor de germană la Universitatea Harvard care își dedica nopțile concretizării unui vis: să transforme lingvistica într-o „știință exactă”, așa cum este, spre exemplu, fizica.

Zipf s-a născut pe 7 ianuarie 1902 într-o familie de origine germană. Bunicul său emigrase din Tauberbischofsheim (Germania) în Statele Unite în jurul lui 1850. Profitând de originile sale germane, Zipf a îmbinat munca sa de doctorand în limba germană cu cercetarea lingvistică. După ce a obținut titlul de doctor în 1929, cu teza *Frecvența relativă ca determinant al variantelor fonetice*, Zipf și-a urmat cercetările până la a publica în 1935 prima sa carte importantă, *Psihobiologia limbajului*, chiar dacă pentru mulți oameni de știință, opera sa cea mai cunoscută, și cea mai



George Kingsley Zipf.

citată, este *Comportamentul uman și principiul minimului efort*, publicată în 1949, cu un an înainte de moartea sa neașteptată, din 1950. Mulți îl citează pe Zipf, dar sunt puțini cei care realmente cunosc opera sa și relevanța ei istorică, ca părinte al lingvisticii cantitative și al științei lingvistice actuale.

În ciuda faptului că limbajul ne însoțește încă din originile noastre ancestrale, ne-au trebuit mii de ani – de fapt chiar e un pas recent – să înțelegem necesitatea explorării limbilor și unităților lor de bază de studiu (ca, de exemplu, fonemele, silabele, cuvintele sau frazele) făcând recurs la metoda științifică. Pe de altă parte, o astfel de întârziere nu trebuie să surprindă, dacă ne gândim că limbajul e un fenomen biologic și că studiul originii biologice a speciei noastre nu a luat avânt realmente până în 1859 cu *Originea speciilor* de **Charles Darwin** (1809-1882). Încă în zilele noastre limbajul este considerat caracteristică exclusiv umană care ne distinge de restul speciilor care locuiesc pe planetă, chiar dacă suntem incapabili să segmentăm clar semnalele de comunicare complexă ale celorlalte specii de ființe cu care împărțim pământul.

De fapt, tumultul darwinist din secolul al XIX-lea a fost atât de puternic încât societățile principale de lingviști ale epocii au interzis cu desăvârșire să se trateze problema originii limbajului, obstaculând astfel dezvoltarea unuia dintre domeniile de studiu cele mai importante din lingvistică, poate în ideea de a menține supremația noastră în fața tuturor ființelor de pe planetă, deoarece ce s-ar întâmpla dacă am descoperi o specie cu o comunicare sofisticată? Ar trebui să aibă drepturi „umane”? Mitul Babel ar începe să scârțâie, stârnind panica unora.

Pentru a-și concretiza visul, Zipf a cules în cursul anilor o serie lungă de reguli statistice ale limbajului, pe care le-a numit legi, și a încercat să construiască o teorie unificată care să cuprindă nu doar limbajul, ci și alte dimensiuni ale comportamentului uman, precum sexul, societățile sau economia. Zipf avea foarte clar în minte că baza oricărei științe este dovada empirică. Putem afirma că Zipf este destul



Figura 1. Nor de cuvinte din tag-urile cele mai utilizate în tweet-urile unui ansamblu de utilizatori ibero-americieni după Coronel-Salas și Mier Sanmatin (2016). Pentru mai multe detalii vezi <http://www.revistalatinacs.org/071/paper/1115/35es.html>

de cunoscut pentru patrimoniul său teoretic, un pic mai celebru pentru legile pe care le-a cules și faimos pentru o lege deosebită, pe care o cunoaștem ca legea lui Zipf, și pe care totuși nu a fost el primul care a enunțat-o.

Înainte de a prezenta legea lui Zipf, ne vom permite un salt în timp de la deceniul anilor '40 din secolul al XX-lea la era internetului. Norii de cuvinte (sau *tag clouds*, literalmente „nori de etichete”) sunt reprezentări pictografice ale conținutului unei surse de informații în care cuvintele apar cu o dimensiune care indică importanța lor într-un text (figura 1). Găsim adesea norii de cuvinte asociate paginilor de internet.

De obicei, norii de cuvinte arată doar așa-zisele „cuvinte conținut” (substantive, verbe, adjective și adverbe), cele care prin ele însele posedă o semnificație mai mare, în timp ce cuvintele funcții (articole, pronume, prepoziții, conjuncții), pe care le folosim de obicei drept conectori, sunt omise. Norii de cuvinte din multe texte, ca în figura 1, arată o confuzie care are sens mai mult sau mai puțin în funcție de afinitatea textelor în care au fost incluse (dacă sunt scrise în aceeași limbă, dacă sunt de tematică asemănătoare...), dar se folosesc și pentru a rezuma vizual conținutul unui text specific, în mod special dacă numărăm exact datele în care fiecare cuvânt apare într-un text, merită să îi spunem frecvență.



Figura 2. Nor cu cele 50 de cuvinte conținut cele mai frecvente din versiunea spaniolă a Declarației Universale a Drepturilor Omului.

Astfel, norul de cuvinte al Declarației Universale a Drepturilor Omului ce apare în figura 2 este o variantă a figurii 1; cu cât mai frecvent este un cuvânt, cu atât mai mare este dimensiunea sa. După acest nor, cuvintele conținute în text cele mai importante sunt *derecho* („drept”) și *derechos* („drepturi”), *libertad* („libertate”), *persona* („persoana”). Aici cuvintele apar într-o ordine nedeterminată și lângă fiecare e indicată frecvența sa în paranteză. A studia matematic frecvența cuvintelor a fost punctul de plecare al tezei lui Zipf: i-a permis să observe legea care poartă numele său și să dea semnalul de deschidere lingvisticii cantitative.

Lingvistica cantitativă se naște ca o disciplină empirică, efemeră și ambițioasă. S-a născut cu mult înainte de era informatică și a *Big Data*, cu scopul de a înfrunta empiric problema limbajului și studiul limbilor umane prin analiza statistică. Face parte din lingvistica matematică și cu aceasta împarte teama formală, deoarece vrea să stabilească legi generale și universale care să ne poată explica cum funcționează limbile, motiv pentru care este într-un anume

sens sora celei care este cunoscută ca lingvistică computațională, dat fiind că astăzi se folosește potențialul enorm al computerelor și marea disponibilitate a datelor care, cum vom vedea, nu mai sunt doar simple texte și corpusuri lingvistice, ci și genomi sau semnale de origine animală sau necunoscută.

De aceea este ambițioasă, cum spuneam, deoarece metodele sale cantitative și conceptele sale lingvistice deschid o porțiță în alte câmpuri interdisciplinare; dar este și pentru un alt motiv: pretinde să se transforme în nucleul teoretic al lingvisticii. Acest lucru face din lingvistica cantitativă o ambițioasă disciplină de tranzit, fondatoare, la fel ca altele în procesul de instituire a științelor.

Legea lui Zipf

În figura 2 am reprezentat cuvintele din Declarația Universală a Drepturilor Omului ordonate alfabetic, dar ce s-ar fi întâmplat dacă nu am fi omis niciun cuvânt funcție și le-am fi ordonat în ordine descrescătoare după frecvență? În mod precis, acest criteriu de ordonare (plecând de la alte corpusuri) a fost cel care a dus la descoperirea în era preinformatică a legii lui Zipf, fără îndoială cea mai faimoasă dintre legile statistice ale limbajului.

Pentru a prezenta legea lui Zipf, să presupunem că ordonăm cuvintele din Declarație în versiunea spaniolă de la cea mai mare la cea mai mică frecvență. Cele zece cuvinte cele mai frecvente rezultă a fi:

rang	frecvență	cuvânt	rang	frecvență	cuvânt
1	138	de	6	48	el
2	96	la	7	42	derecho
3	94	y	8	40	los
4	86	a	9	32	que
5	49	en	10	31	persona

Zipf a numit „rang” poziția pe care fiecare cuvânt o ocupa în lista ordonată de frecvențe. După acest criteriu, cuvântul cel mai frecvent are rangul 1, al doilea cuvânt cel mai frecvent are rangul 2, și tot așa în succesiunea de poziții. Când se desenează relația dintre o frecvență (axa y) și rang (axa x), apare o curbă care se destramă imediat (figura 3).

Cel mai interesant lucru e că, desenând aceeași relație în scara dublu logaritmică, merită spus arătând logaritmul frecvenței (axa y) față de rang (axa x), apare o curbă cu o zonă destul de dreaptă (figura 4). Acesta este un indiciu al faptului că frecvența și rangul urmează o lege potențială, adică frecvența (f) este proporțională unei anume puteri a rangului (r), adică:

$$f = \frac{A}{r^\alpha}$$

unde tipic $\alpha = 1$ în engleză și A este o constantă corelată cu dimensiunea textului, așa încât Zipf a descoperit că aproximativ $A \approx T/10$, unde T e numărul total de cuvinte din text. Această relație potențială e cunoscută sub numele de legea lui Zipf deoarece el a fost cel care a făcut-o cunoscută, chiar dacă nu a fost primul care a postulat-o, cum vom vedea, și chiar dacă în practică, cum se întâmplă cu fiecare lege statistică, parametri pe care i-a inclus nu sunt constanți. Această formă de a exprima legea lui Zipf prezintă problema de atribuire arbitrară a rangului pentru cuvintele care apar de același număr de ori, lucru ce se poate constata grafic (figura 4) ca linii succesive de puncte pentru cuvintele cu frecvență joasă (și cu toate astea de rang elevat).

Zipf și-a bazat o bună parte din analizele sale statistice pe corpul mastodontic al epocii, al lui **Friedrich Wilhelm Kaeding** (1843-1928), care în 1898 a completat un dicționar de frecvențe ale limbii germane, contabilizând 11 milioane de cuvinte (o infinitate pentru era preinformatică, obținută plătind „subordonați”). De fapt, sunt alte

Figura 3. Frecvența unui cuvânt în funcție de rangul său pentru textul Declarației Universale a Drepturilor Omului în spaniolă. Fiecare punct reprezintă un cuvânt.

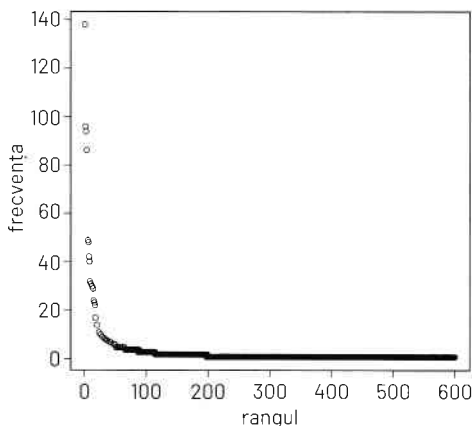
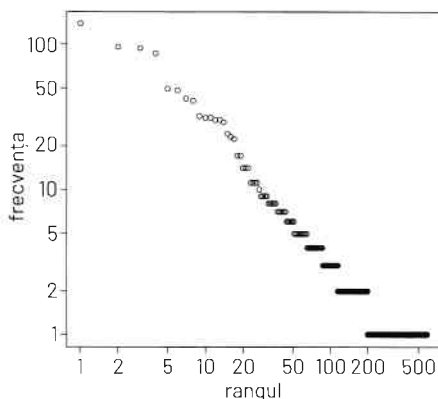


Figura 4. Frecvența unui cuvânt în funcție de rangul său pe scara dublu logaritmică (pe două axe) pentru textul Declarației Universale a Drepturilor Omului în spaniolă. Și în acest caz, fiecare punct reprezintă un cuvânt.



precedente de studii ale legilor potențiale, cum ar fi studiile lui Pareto în econometrie (1897) și alte contribuții anterioare lui Zipf în lingvistică, la începuturile secolului al XX-lea, orientate spre a determina vocabularul de bază și a realiza calcule ale cuvintelor celor mai frecvente în diverse limbi, pentru a facilita astfel învățarea lor, sau pentru simplificarea textelor scrise utilizând tahigrafia (sau stenografia). Acesta este cazul lucrărilor lui Estoup (1902), Eldridge (1911) – lucrare

LEGILE POTENȚIALE SUNT DREPTE PE SCARA DUBLU LOGARITMICĂ

Am văzut că desenând relația între frecvență (f) a unui cuvânt și rangul său (r) pe scara dublu logaritmică ne-a apărut aproape o dreaptă. De fapt, putem folosi un desen pe această scară pentru a scoate în evidență că o variabilă y este proporțională cu o anumite putere a unei variabile x , merită spus că

$$y = Ax^{-\alpha}.$$

Dacă apare o dreaptă, pe scara logaritmică, este relația. Motivul este ușor de explicat. Să presupunem că avem o relație liniară pe scara dublu logaritmică între două variabile x și y , și adică,

$$\log y = -\alpha \cdot \log x + c$$

unde α este panta dreptei. Atunci, ridicând la exponent cele două părți ale ecuației

$$e^{\log y} = e^{-\alpha \log x + c}$$

obținem

$$y = e^c (e^{\log x})^{-\alpha}$$

și, în fine

$$y = Ax^{-\alpha}$$

cu $A = e^c$. În cazul legii lui Zipf $f = Ar^{-\alpha}$, enunțată anterior, avem $y = f$, $x = r$ și de obicei $\alpha = 1$ în engleză.

pe care a folosit-o și Zipf în modelul său potențial –, Ayres (1915), Thorndike (1921) sau Dewey (1923), printre alții.

Jean-Baptiste Estoup (1868-1950), într-un text fundamental al literaturii stenografice din 1902, *Gammes sténographiques. Recueil*

de textes choisis pour l'acquisition méthodique de la vitesse, analizase distribuția frecvenței a 30 000 de cuvinte provenite din 75 de texte, întâlnind deja legea potențială pe care Zipf o formalizase, și că de aceea în literatură, uneori este numită și „legea Estoup-Zipf”.

Mai puțin cunoscut este **Edward Uhler Condon** (1902-1974), profesor de fizică la Princeton, care în 1928 a publicat o scurtă notă „Statistics on vocabulary” în *Science* (vol. 67, p. 300), în care pentru prima dată a reprezentat pe axele logaritmice relația între frecvența cuvintelor și rangul lor (figura 5), plecând de la calcule de cuvinte din corpusurile lui Ayres (1915) și Dewey (1923). Va trebui să discutăm deci despre legea lui Condon? Istoria științei nu este corectă uneori, și în ciuda măreției lui Zipf, trebuie să recunoaștem că a fost Condon cel care a pus steagul relației potențiale între frecvența cuvintelor și rangul lor, chiar dacă, în lumina lucrărilor lui, a făcut-o cu siguranță dintr-un divertisment statistic, din moment ce nu s-a mai întors niciodată să aprofundeze studiul statistic al limbajului, cum ar fi făcut Zipf.

Meritul lui Zipf a fost că a formalizat și că a înțeles lucrările precedente despre statistica cuvintelor și că a explorat diferite limbi care mergeau de la sanscrită sau mandarină, la engleză și franceză, trecând prin limbile clasice, cum ar fi greaca și latina, printre multe altele; pe de altă parte, a studiat producția infantilă și pe cea a pacienților afectați de schizofrenie, drept cazuri particulare în învățarea limbilor și a patologiei. Scopul său era să adune suficiente dovezi care să demonstreze universalitatea legii potențiale pe care o formulase și să creeze o știință lingvistică.

Zipf era în favoarea abolirii separării între științe și litere, anticipând dezbaterea asupra celor două culturi pe care **Charles Percy Snow** (1905-1980) a creat-o la jumătatea secolului al XX-lea. Dacă această diviziune supraviețuiește și azi, dacă în studiile preuniversitare nu este inclusă lingvistica cantitativă, iar în studiile universitare de lingvistică se vorbește puțin sau se ignoră complet moștenirea lui Zipf, poate ar trebui să ne întrebăm ce a împiedicat acest lucru.

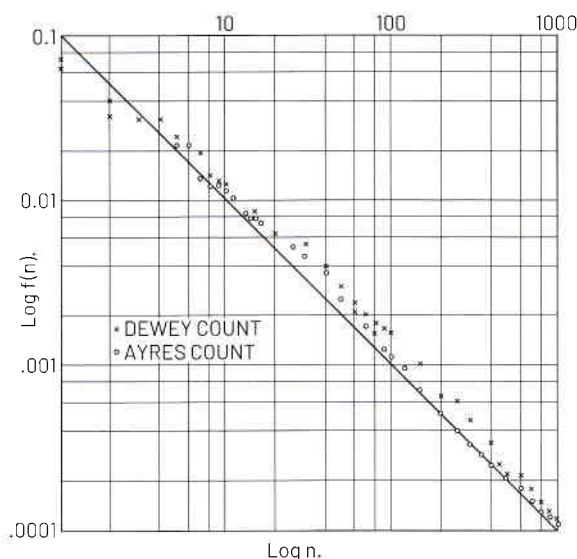


Figura 5. Prima apariție a relației potențiale între frecvența cuvintelor (f) și rangul lor (n) (sau ordinea de apariție într-un text), publicată în *Science* (vol. 67, p. 300) în studiul *Statistics on vocabulary* de Edward Condon (1928), bazat pe datele din lucrările lui Ayres (1915) și Dewey (1923). Trebuie subliniat că în graficul lui Condon n este rangul.

Alte legi zipfiene

Până acum am urmat curentul popular care duce înspre a defini legea lui Zipf plecând de la o ecuație specifică ce descrie relația între frecvența unui cuvânt și rangul acestuia, dar găsim multe alte legi în cartea sa cea mai celebră, *Comportamentul uman și principiul minimului efort* (1949). Să luăm în considerare câteva.

Legea lui Zipf care pune în relație frecvența unui cuvânt cu rangul lui are o soră: dacă regroupăm diferitele cuvinte care apar în același text de un același număr f de ori, așa încât, de exemplu, cuvintele cu $f = 1$ sunt toate cele care apar o singură dată (cele care în lexicografie se numesc *hapax legomena*), $f = 2$ sunt atunci cele care apar de două ori (*dis legomena*) și tot așa, se ajunge la concluzia că numărul de cuvinte diferite care apar de f ori este (aproximativ):