

2. Un fapt bine cunoscut celor ce se ocupă de lingvistica matematică este că în diversele limbi entropia are valori apropiate. Aceasta nu este totuși evident la prima vedere, deoarece limbile utilizează alfabetele cu lungimi variate, iar pe de altă parte distribuția statistică a literelor diferă de la o limbă la alta.

Pentru a explica acest fapt, autorul a încercat să constate dacă nu există unele legi care să acționeze asupra frecvenței de apariție a literelor. Evident, ne referim la texte scrise, pentru care se dispune de statistici suficiente de corecte.

Studiind un număr de 14 limbi, atât europene cât și asiatice (română, italiană, spaniolă, franceză, franceză vorbită*), rusă, germană, engleză, hindi, telegu, malayalam, marathi, kannada, tamil), autorul a reușit să arate că între frecvența de apariție f_i a unei litere și rangul i pe care îl are acea literă în mulțimea ordonată după frecvență a literelor, subsistă legea

$$f_i = A2^{-ki}. \quad (1)$$

Curbele ce ilustrează această lege au fost comunicate anterior [11]. În tabela alăturată se prezintă valorile obținute pentru A și k cu câteva limbi.

Tabela 1
Valorile lui A și ale altor mărimi conexe, pentru câteva limbi

Nr. crt.	Limba	A	k	$\log(A+1)$	$k \log 2$	%
1	Română	0,13	0,18	0,0531	0,0541	+1,8
2	Italiană	0,165	0,21	0,0663	0,0633	4,6
3	Spaniolă	0,17	0,21	0,062	0,0633	7,3
4	Franceza vorbită	0,10	0,13	0,0414	0,0392	5,3
5	" scrisă	0,165	0,21	0,0645	0,0602	6,6
6	Rusa	0,11	0,15	0,0453	0,0452	0,4
7	Germană	0,12	0,16	0,0492	0,042	2
8	Engleză	0,12	0,17	0,0492	0,0511	1,8
9	Hindi	0,10	0,16	0,0414	0,042	1,8
10	Marathi	0,075	0,096	0,0314	0,02	7,7
11	Telegu	0,11	0,13	0,0434	0,0392	9,7
12	Malayalam	0,11	0,13	0,0434	0,0392	9,7

Evident, trebuie ținut seama și de faptul că alfabetul respectiv are N litere, astfel încât trebuie să avem

$$\sum_{i=1}^N f_i = 1. \quad (2)$$

*) Ne referim la distribuția fonemelor.

Aceste două relații stabilesc în primul rând o legătură între A și k , cele două constante care caracterizează din punct de vedere statistic apariția literelor într-un text, în conformitate cu legea stabilită de noi.

Această relație se deduce imediat din (1) și (2)

$$A \sum_{i=1}^N 2^{-ik} = A \frac{1 - 2^{-Nk}}{1 - 2^{-k}} 2^{-k}. \quad (3)$$

Deoarece în general k este mic, dar $Nk \gg 1$, din (2) și (3) rezultă

$$\frac{A}{2^k - 1} \approx 1,$$

adică

$$A \approx 2^k - 1. \quad (4)$$

Acest rezultat pare surprinzător la prima vedere, deoarece el leagă pe A de k , fără a ține seama de lungimea alfabetului. Totuși, datele din tabela 1 arată că formula (4) este foarte bine verificată, în majoritatea cazurilor erorile fiind cu mult sub toleranțele inerente unor cercetări statistice în lingvistică, unde materialul este atât de variat.

Cele de mai sus arată că pentru a calcula entropia, nu este necesar să cunoaștem legea exactă după care sînt distribuite literele (N la număr), în ceea ce privește frecvența lor, în limba studiată. Este suficient să cunoaștem trei mărimi caracteristice, și anume k , N și A . Mai mult chiar, avînd în vedere relația care există între A și k , entropia depinde numai de doi parametri, fie ei k și N . Se poate deci scrie

$$H = H(k, N).$$

Dar, nici N și k nu sînt independente. Aceasta se poate justifica prin faptul că dacă alfabetul are un număr mare de litere, k trebuie să fie mic (evident că în acest caz și A va fi mic). Dacă alfabetul are puține litere, exponentul k va fi ceva mai mare.

Să trecem efectiv la calcularea entropiei.

Prin definiție entropia este

$$H = - \sum f_i \lg_2 f_i.$$

Ținînd seama de (1) avem

$$\begin{aligned} H &= - \sum f_i \lg_2 f_i = - \sum f_i (\lg_2 A - ki) \\ &= - \lg_2 A \sum f_i + \sum k f_i i. \end{aligned}$$

Dar din (2) rezultă că

$$H = - \lg A + k \sum_{i=1}^N i f_i.$$

Ultimul termen al sumei cuprinde o progresie aritmetico-geometrică, ce se poate calcula cu o formulă cunoscută [16]. Ținînd seama și de (4) se obține

$$H = - \lg(2^k - 1) + k \frac{2^{-k}(1 - 2^{-Nk})}{(1 - 2^{-k})^2} - k \frac{N2^{-(N+1)k}}{1 - 2^{-k}}.$$

Deoarece însă atât A cât și k și N au în practică valori apropiate, din această formulă rezultă că valoarea lui H pentru diversele limbi nu va fi mult diferită de o valoare standard. Avem astfel o explicație a faptului că pentru diverse limbi, H are valori apropiate.

3. Este evident că nu este corect să se calculeze entropia unei limbi ținând seama numai de frecvența de apariție a literelor. După cum se știe, în cadrul limbilor acționează legi multiple, de fapt apariția literelor într-un text reprezentând evenimentele unui lanț cu legături. O problemă ce se pune pentru calculul curent al entropiei este aceea a determinării lungimii maxime la care se manifestă în mod practic legăturile lanțului. Pentru unele limbi s-a calculat frecvența de apariție a grupărilor de 2, 3 și mai multe litere [9], [15], [12].

O primă metodă pentru determinarea acestei lungimi maxime se poate face calculând entropia în diverse ipoteze. Astfel, în primul rând se admite că nu există legături și se calculează o valoare H_1 .

Se admite apoi că influența se manifestă numai asupra literelor vecine și se calculează o nouă entropie

$$H_2 = - \sum p(A_i, A_j) \log p(A_i, A_j).$$

La fel se continuă calculându-se pentru grupări de n litere. Determinarea lungimii maxime la care se manifestă legătura se face impunând o anumită diferență maximă între două entropii consecutive. Dacă se notează cu ϵ această diferență maximă, atunci spunem că s-a determinat lungimea maximă L la care se manifestă legătura, dacă este satisfăcută condiția

$$H_L - H_{L+1} \leq \epsilon.$$

Este evident că această metodă ne dă în același timp două rezultate. În primul rând ne permite să determinăm exact entropia unei limbi. În al doilea rând determinăm lungimea maximă de text la care s. face simțită influența unei litere.

Metoda are în schimb dezavantajul că necesită o foarte mare cantitate de calcule pentru stabilirea prealabilă a frecvențelor de apariție a diferitelor grupări de litere.

Autorul propune o metodă diferită pentru determinarea lui L . Metoda are avantajul că permite determinarea lungimii maxime la care se manifestă legătura între literele unui text, fără a face în prealabil determinarea frecvențelor de apariție a grupărilor de mai multe litere și nici calculul grupărilor entropiilor corespunzătoare. Această metodă se leagă de funcția de autocorelație.

Noțiunea de autocorelație nu s-a întâlnit pînă acum în problemele de lingvistică. Autorul propune utilizarea acestei metode în modul următor. În primul rând se acordă fiecărei litere o anumită valoare numerică. Se calculează apoi funcția de autocorelație a textului respectiv pentru 1, 2, ..., n deplasări.

Fie v_i valoarea numerică acordată literei i . Textul se va prezenta sub forma unui șir numeric

$$v_1 \dots v_i \dots$$

Îl putem transcrie sub forma

$$\prod_{i=1}^M v_i.$$

Funcția de autocorelație pentru τ deplasări se definește cu formula

$$A_\tau = \frac{1}{M - \tau} \sum_{i=1+\tau}^M v_i v_{i-\tau}.$$

Pentru o deplasare suficient de mare, legătura dintre litere dispare. Notînd cu p_i probabilitatea de apariție a literei v_i , se poate calcula funcția de autocorelație în raport cu valorile numerice acordate literelor și probabilitățile lor de apariție.

Se poate arăta că dacă legăturile dintre litere dispar, atunci funcția de autocorelație devine*)

$$A_\tau = \left(\sum_i p_i v_i \right)^2.$$

Problema revine la a determina deplasarea minimă pentru care se întâlnește această valoare limită.

O a doua problemă care se pune este aceea de a se determina distribuția optimă de valori care, acordate diferitelor litere, face ca procesul să aibă viteza maximă de tindere către limita arătată, adică care conduce la acest rezultat pentru un număr suficient de mic de litere din textul respectiv.

Autorul nu a putut rezolva această ultimă problemă.

Se crede însă că o distribuție convenabilă este aceea care atribuie o valoare mare literelor foarte frecvente, și dimpotrivă o valoare mică literelor puțin frecvente.

*) Într-adevăr, din formarea produselor de autocorelație se obține un număr N_{ij} de termeni în care apare produsul $v_i v_j$. Din M produse, litera de valoare v_i apare ca prin termen de $N_i = M p_i$ ori, unde p_i este probabilitatea ei de apariție în text.

Dacă numărul τ al deplasărilor este suficient de mare spre a face să dispară legăturile statistice dintre litere, atunci după litera de indice i , putem întâlni oricare altă literă, cu aceeași probabilitate p_j cu care litera respectivă apare în text. Deci

$$N_{ij} = N_i p_j = M p_i p_j.$$

Rezultă imediat

$$\begin{aligned} A_\tau &= \frac{1}{M} \sum_{i,j} N_{ij} v_i v_j = \sum_{i,j} p_i v_i p_j v_j = \\ &= \sum_i p_i v_i \sum_j p_j v_j = \left(\sum_i p_i v_i \right)^2. \end{aligned}$$

O experiență a fost întreprinsă în acest sens de autor, împreună cu C. Sala. S-a lucrat pe poezia „Luceafărul” de M. Eminescu [6], pentru care C. Sala calculase în prealabil frecvențele de apariție a literelor. Valorile acordate literelor sînt indicate în tabela 2.

Tabela 2
Valori acordate literelor în primele încercări de autocorelare a textelor

a	1	a	-1
b	2	c	-2
d	3	e	-3
f	4	g	-4
h	5	i	-5
î	6	j	-6
l	7	m	-7
n	8	o	-8
p	9	r	-9
s	10	ș	-10
t	11	ț	-11
u	12	v	-12
z	13	x	-13
y	14	blanc	0

Ținînd seama de frecvențele de apariție a literelor, se găsește

$$\sum p_i v_i = 0,5358,$$

deci

$$(\sum p_i v_i)^2 = 0,28.$$

Pentru diferite deplasări date literelor, după 13 strofe (ceea ce corespunde la circa 1 200 semne) valorile obținute pentru funcțiile de corelație erau următoarele

$$A_2 = -1,35, \quad A_3 = +0,453,$$

$$A_4 = +1,28, \quad A_5 = +1,12.$$

Cum aceste valori determinate pe text sînt depărtate de valoarea calculată, rezultă că sîntem destul de departe de a ajunge la punctul urmărit. Aceasta se poate explica pe de o parte prin faptul că nu s-a considerat o lungime suficientă de text — dar, mult mai probabil, se datorește faptului că nu s-au ales valori convenabile pentru litere.

4. O altă problemă investigată de noi a fost legată de proprietățile statistice ale cuvintelor, problemă ce poate fi privită din puncte de vedere diferite. În cercetările sale Shannon s-a ocupat și de frecvența cuvintelor. Acest aspect este însă numai în parte interesant, pentru noi, deoarece se știe că există și posibilitatea calculării entropiei unei limbi, ținîndu-se seama de frecvența de apariție a cuvintelor. În acest sens se poate utiliza legea lui Estoup-Zipf-Mandelbrot.

Metoda clasică a fost arătată de C. Shannon.

În legătură cu această chestiune, ținem să arătăm numai faptul că cercul de lingvistică matematică de la Institutul de lingvistică din București, sub conducerea acad. A. I. Rosetti, a calculat frecvența de apa-

riție a cuvintelor în limba noastră. Datele publicate de V. Suteu [19] au permis să arăt [10] că și pentru limba noastră se aplică legea amintită. Din cercetările întreprinse de autor, rezultă că în cazul limbii romîne între frecvența f_n de apariție a cuvintelor și rangul lor n subsistă legătura

$$f_n = P(m+n)^{-B},$$

unde

$$P = 0,03; \quad m = 1,5; \quad B = 1.$$

După părerea noastră, aceste studii [19] tranșează și problema fondului principal de cuvinte, deoarece de data aceasta se dispune de o inventariere statistică și amplă a acestui fond. Rezultatele publicate arată totodată justetea tezei lui B. P. Hașdeu, în ceea ce privește originea latină a celor mai multe cuvinte din fondul principal.

5. O altă serie de cercetări privind entropia limbii se referă la legătura dintre numărul de cuvinte avînd un anumit număr de silabe, lungimea medie a cuvintelor și alte mărimi.

Primele cercetări asupra numărului de silabe pe care le au diversele cuvinte sînt după părerea noastră cercetările specialistului sovietic S. G. Chebanov [2]. Acesta a arătat că pentru cazul limbilor indo-europene subsistă o distribuție Poisson.

Cercetările ulterioare, efectuate de Fucks [7], au stabilit o anumită legătură între entropia cuvintelor și lungimea medie a lor.

Ambii cercetători încearcă să stabilească o grupare a limbilor după lungimea medie a cuvintelor.

În ceea ce privește cercetările lui Fucks, reconsiderate de noi ele conduc la o concluzie destul de interesantă.

În teoria sa, Fucks consideră că fiecărei limbi îi corespunde un anumit punct în planul (n, y) , unde n este numărul mediu de silabe al cuvintelor, iar y entropia mulțimii cuvintelor clasificate după x .

Pentru a vedea în ce măsură aceste constatări ale lui Chebanov și Fucks se pot aplica și limbii romîne, autorul a studiat un număr de texte de limbă romînă, scrise, atît de literatură beletristică cît și de literatură științifică. Textele utilizate sînt indicate în bibliografie [1], [3], [5], [8], [17].

Investigațiile noastre au dus la concluzia că pentru textele literare, ambele teorii (Chebanov și Fucks) sînt verificate. Astfel, avem o distribuție Poisson pentru cuvintele textelor de literatură beletristică. De asemenea, avem o grupare strînsă a punctelor figurative în diagrama lui Fucks, dacă ne referim numai la textele literare menționate.

Dacă se trece însă la texte de literatură științifică, rezultatele sînt diferite.

În primul rînd este însă necesar să precizăm că pentru textele de literatură științifică nu s-a putut utiliza texte de matematică sau de fizică, deoarece în ele intervin foarte des formule matematice. Tocmai pentru a se evita această dificultate, s-au utilizat texte din gramatică și din științe juridice.

După părerea noastră, rezultatul obținut este deosebit de interesant.

Astfel, în primul rând lungimea medie a cuvintelor din textele de literatură științifică analizate este sensibil mai mare decât lungimea medie a cuvintelor din textele de literatură beletristică. Lungimea medie a cuvintelor din textele de literatură beletristică este de 1,6 silabe, în timp ce la textele de literatură științifică lungimea este circa 2,1 silabe. Diferența este foarte mare, deoarece în diagrama lui Fucks, la lungimea medie de 1,63 corespunde limba germană, iar la lungimea medie de 2,104 limba arabă. În general, la o creștere a lungimii medii de 1,35 (engleza) la 2,455 (turcă), se trece succesiv prin nouă limbi.

Acest fapt este ușor explicabil. În textele de literatură beletristică întâlnim cuvinte uzuale, or, este firesc ca în limbă, în mod natural cuvintele uzuale să fie foarte scurte, tocmai pentru a se putea exprima o cantitate cât mai mare de informații într-un timp cât mai scurt.

Pe de altă parte, în textele de literatură științifică se utilizează în mod obligatoriu un vocabular cu mult mai bogat decât în textele de literatură beletristică. Dar un vocabular mai bogat presupune în mod obligatoriu un număr de cuvinte lungi. Deoarece combinațiile de puține silabe au fost în general utilizate pentru cuvintele uzuale, rezultă că noțiunile mai complexe ale limbajului utilizat în literatura științifică nu pot să fie exprimate decât prin cuvinte cu o lungime din ce în ce mai mare.

O concluzie importantă care se degajă din aceste cercetări este că, în general, teoria lui Fucks este justă în ceea ce privește legătura dintre entropie și lungimea medie a cuvintelor (măsurată în silabe), dar este greșită în ceea ce privește gruparea unei limbi într-o singură porțiune delimitată a diagramei mai sus-amintite. Din exemplele analizate rezultă clar că, de fapt, se diferențiază o limbă a literaturii beletristice și o limbă a literaturii științifice. Rămâne de examinat mai departe în ce măsură se poate întreprinde un studiu privind efectuarea unor calcule asemănătoare pentru textele în care apar formule matematice.

Desigur că problema ar putea fi extinsă într-un mod corespunzător și la limbile aglutinante. Aici însă intervin dificultăți analoge celor ce se întâlnesc în cazul formulelor matematice, ideogramele prezentând multe puncte comune cu simbolurile matematice.

După cum se vede, metodele teoriei informației [4] permit abordarea a numeroase aspecte noi, privind transmiterea informației prin intermediul limbajului [13], [14].

О ВЫЧИСЛЕНИИ ЭНТРОПИИ

РЕЗЮМЕ

В работе изучаются проблемы, связанные с вычислением энтропии текстов H . В первую очередь дается формула для H , как функции только от A и N , где N — максимальное число букв алфавита, употребляемого в соответствующем языке, а A — характеристическая константа данного языка.

Далее дается метод, который, используя технику автокорреляции текстов, разрешает определить максимальное расстояние, на котором проявляется — в среднем — влияние буквы в данном тексте.

Последние рассматриваемые вопросы относятся только к энтропии слов, группированных по числу их слогов. Статистические исследования, предпринятые в этом направлении для текстов на румынском языке позволяют сделать несколько выводов: 1) закон Фукса применим и для румынского языка; 2) с точки зрения диаграммы Фукса поэтический язык отличается от научного; 3) теория Кебанова применима и для румынского языка в литературных текстах, но не применима — в научных.

SUR LE CALCUL DE L'ENTROPIE

RÉSUMÉ

L'auteur étudie des problèmes qui sont en relation avec le calcul de l'entropie H des textes. Il donne d'abord une formule pour H , seulement en fonction de A et N , où N est le nombre maximum de lettres de l'alphabet utilisé dans la langue respective et A une constante caractéristique pour cette langue.

Il indique ensuite une méthode qui, utilisant le procédé d'autocorrélation des textes, permet de déterminer la distance maximum à laquelle se manifeste — en moyenne — l'influence d'une lettre d'un texte donné.

Les derniers problèmes étudiés se réfèrent à l'entropie des mots groupés d'après leur nombre de syllabes. Les recherches statistiques entreprises dans cette direction pour les textes en langue roumaine permettent de tirer plusieurs conclusions: 1) la loi de Fucks s'applique également au roumain; 2) au point de vue du diagramme de Fucks, on distingue une langue poétique d'une langue scientifique; 3) la théorie de Tchébanov s'applique aussi à la langue roumaine en ce qui concerne les textes littéraires mais non pas les textes scientifiques.

BIBLIOGRAPHIE

1. Arghezi T., *Versuri*. Biblioteca pentru toți, EȘPLA, București, 1960.
2. Chebanov S. G., *On conformity of language structures within the Indo-European Family to Poisson's Law*. C.R. (Doklady) de l'Acad. Sciences de l'U.R.S.S., LV, 2, 99—102 (1947).
3. * * * *Codul Civil*. Ed. științifică, București, 1958.
4. Constantinescu I., Condrea S., Nicolau Edm., *Teoria informației*. Edit. Tehnică, București, 1958.
5. Creangă I., *Povești și povestiri*. Biblioteca pentru toți, EȘPLA, București, 1954.
6. Eminescu M., *Poezii*. Biblioteca pentru toți, EȘPLA, București.
7. Fucks, *Mathematical Theory of Word Formation*. Information Theory, Third London Symposium, Edit. by Colin Cherry, Butterworths, London, 1956, 154—170.
8. * * * *Indreptar de punctuație*. Edit. Acad. R.P.R., București, 1956.

