БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ * ACADÉMIE BULGARE DES SCIENCES
ИЗВЕСТИЯ НА МАТЕМАТИЧЕСКИЯ ИНСТИТУТ
BULLETIN DE L'INSTITUT DE MATHÉMATIQUES
Том (Tome) IX

ВЪРХУ ЕДИН НАЧИН ЗА ОПРЕДЕЛЯНЕ НА „ЦЕНИ" НА СЪЧЕТАНИЯ ОТ БУКВИ НА ПИСМЕНИЯ ЕЗИК. I

Емануил Симеонов

Настоящата работа съдържа един метод за определяне на „цени" на съчетания от букви на писмения език, илюстриран върху извадки от български текстове.

Идеята за „цена" на дума от даден текст води началото си още от работите на Ципф [8], който, изследвайки честотите на думите в писмен текст с достатъчно голям обем, установява по емпиричен път, че между ранга r (който една дума от разглеждания текст заема при подреждането на всички думи от извадката по намаляваща честота) и честотата на думата p_r съществува връзката

$$p_r = Pr^{-B}, \tag{1}$$

където P и B са константи. Тази връзка освен за думите, които се срещат най-често, е в сила при произволен език.

За релацията (1) и начините на извеждането ѝ в [5] на стр. 192 Манделброт пише: „Едни от най-фрапиращите особености на тези закони, които свързват честотата с ранга, се отнасят повече до метода на извеждането, отколкото до съдържанието им. Те могат да бъдат обяснени, като се следват установените начини на представяне на нещата с физически модели. При този начин на разглеждане се правят допускания, които имат следните характерни особености:

а) моделите да не са нито абсурдни, нито невероятни, но да опростяват така нещата, че да могат да бъдат изследвани аналитично, и

б) моделите да имат далеч по-малко съдържание, отколкото фактическите свойства на речта, но все пак дотолкова да са лишени от съдържание, та да имат смисъл при всички случаи... Излизайки от този начин на разглеждане, може да се стигне до предсказване на свойства на речта, които съвсем не са очевидни. За съжаление познати са много малко примери на приложения на този метод в лингвистиката."

В основата на теоретичните работи на Манделброт [5], Белевич [1], Готуссо [3] и др. стои връзката

$$-\ln p = \beta C, \tag{2}$$

където p е честотата на думата в извадката, C е „цената" на думата, а β е константа, която зависи от основата на логаритмите.

§ 1. Манделброт [5] допуска, че писмената реч може да бъде представена с модел, подобен на модела на Болцман, за разпределението на броя на елементарните системи, принадлежащи на една макросистема; при това всяка елементарна система може да взема краен брой дискретни състояния, определени от количеството енергия ε, което елементарната система притежава, когато макросистемата е в термично равновесие. Резултатите на Манделброт са получени при следните хипотези:

1. Ако азбуката на езика има M букви, в езика съществуват като думи всички възможни съчетания, които могат да бъдат образувани от n ($n=1, 2, 3, \ldots$) букви.

2. Дължината на думите е неограничена.

3. „Цената" на думите (съчетанията) в езика зависи от броя на буквите, от които се състои думата (съчетанието).

Несъгласията, които съществуват между теоретичните изводи и емпиричните резултати от досегашните изследвания, може да се дължат на доста грубите допускания за строежа на думите, а, от друга страна, изглежда, че отъждествяването на „цената" на думата с нейната дължина по букви не е подходящо.

При азбука от 31 букви (30 обикновени букви + шпация), както е в българския език, най-голямата дължина на думите не надминава 20 букви. Освен това очевидно не всички съчетания от n букви, които могат да бъдат образувани с M букви, са думи на езика. Броят на съчетанията, които представляват думи на езика, е далеч по-малък.

„Цената" на дадена дума или съчетание от букви в даден текст би могла да се отъждестви с броя на буквите от азбуката на езика, ако отделните букви в текста имат една и съща „цена" и ако съчетанията в текста по две, три и повече букви имаха съответно еднакви „цени". Както отбелязва и Манделброт [5], това не е така: всяка буква в даден текст има собствена „цена". Различните съчетания по две букви, по три и т. н. съответно имат различни „цени".

Не бива да се пропуска и това, че писменият език е един доста несъвършен код на говоримия език, и трябва да се има пред вид обстоятелството, че например едни съчетания с дадена дължина по букви се изговарят по-лесно от други; това не може да не се отрази върху честотата на съчетанията, която би трябвало да бъде мярка за „цената" им, както това се приема във формула (2). Например съчетанието ГА в българския език се изговаря много по-лесно от АГ и „цената" на ГА би трябвало да бъде различна от „цената" на АГ. И наистина в текст, който съдържа 68 251 съчетания по две букви, съчетанието ГА се среща 202 пъти, а съчетанието АГ само 58 пъти. В същата извадка съчетанието НА се среща 1303 пъти, а съчетанието АН се среща 287 пъти и т. н.[1] Разликите в честотите на отделните съчетания от две букви са чувствителни и не може да се вземе дължината по букви за „цена" на съчетанията. Ако броят на буквите определяше цената на дадено съчетание, би трябвало според формула (2) всички съчетания от две букви да имат еднаква „цена" и следователно да имат еднакви честоти. Изглежда, че опростя-

[1] Числените данни, използвани в настоящата работа, са взети от непубликувани материали по статистика на писмения български език на Б. Пенков, Б. Сендов, А. Обретенов, Т. Джуканов и Т. Кирпикова.

ването, което се прави, като се приема, че „цената" на съчетанието е пропорционална на дължината му по букви, води до модел, който е твърде различен от това, което се наблюдава.

§ 2. В теоретичната физика [6], [7] се прави хипотезата, че съвкупността от микросъстояния на една физическа система образува дискретно множество. Всяко макросъстояние на системата обхваща напълно определен брой G нейни микросъстояния. Това число представлява от себе си термодинамическата вероятност или статистическото тегло на макросистемата.

Статистическото тегло G се явява като мярка на ентропията S,

$$S = k \ln G, \tag{3}$$

на физическата система, намираща се в дадено макростояние.

Пресмятането на ентропията въз основа на (3) може да стане в случаите, когато дадена физическа система представлява съвкупност от голям брой съвършено еднакви, съвсем изолирани една от друга системи, на брой N, които се наричат „елементарни системи". Всяка от тези елементарни системи може да взема една дискретна редица от състояния, на брой R. На всяко състояние на елементарната система съответствува определена енергия ε_i, $i = 1, 2, \ldots, n \leq R$[1], и обратно — на всяка енергия съответствува определено състояние на елементарната система.

При тази хипотеза всички микросъстояния имат еднакви вероятности. Физическият смисъл на това положение е следният: дадена физическа система, подчинена на определени постоянни макроскопични условия, може да се намира във всички възможни за тези макроусловия микросъстояния и в даден момент вероятностите за кое да е от тези микросъстояния са равни.

Числото $\ln G$ се дава с израза

$$\ln G = -N \sum_{1}^{n} p_i \ln p_i, \tag{4}$$

където $p_i = N_i/N$ ($i = 1, 2, \ldots, n$ и N_i е броят на елементарните системи в състояние i) е честотата на елементарните системи, които съдържат енергия ε_i. Броят на различните видове енергии, на които се разпада тоталната енергия E на физическата система, е $n \leq R$. Този брой показва същевременно и броя на различните състояния, в които може да се срещне една микросистема, когато макросистемата е в термодинамично равновесие. Горната граница на n е R. В сила са още връзките

$$\sum_{i=1}^{n} N_i = N, \quad \sum_{i=1}^{n} \varepsilon_i N_i = E.$$

Болцман търси израз за p_i, когато разглежданата система е в термодинамично равновесие, или, което е все едно, когато $\ln G$ е най-голямо,

[1] С n R ще бележим броя на различните микросъстояния, които могат да се срещнат в макросистемата, понеже, когато макросистемата е в дадено състояние, изглежда, че не всички състояния на брой R са застъпени от микросистемите.

За да намери тези p_l, за които $\ln G$ става максимум, Болцман максимизира (4) при ограничителните условия

$$(5)\quad \text{a)}\quad \sum_{l=1}^{n} p_l = 1,$$

$$\text{b)}\quad \sum_{l=1}^{n} \varepsilon_l p_l = \bar{\varepsilon} = \frac{E}{N}$$

и получава за p_l израза [6], [7]

$$(6)\quad p_l = \frac{e^{-\beta\varepsilon_l}}{\sum_{i=1}^{n} e^{-\beta\varepsilon_i}},$$

който има смисъл, когато физическата система е затворена и има постоянна енергия E.

Условие а) в (5) означава, че в термодинамичния процес участвуват всички микросистеми, а условие b) не е нищо друго освен универсалния принцип за съхранение на енергията, приложен към топлинните процеси.

Условията а) и b) от (5) са много общи и могат да бъдат отнесени не само за равновесно състояние на системата, но и за всяко състояние, в което тя въобще би могла да бъде. Ограничителните условия от (5) не са типични за състояние на термодинамично равновесие.

При по-нататъшното разглеждане на задачата за разпределението на елементарните енергии ε_l по елементарните системи и техните количествени оценки ще въведем едно ново ограничение, което отразява условието, че системата е в равновесие, и което гласи: когато една физическа система е в термодинамично равновесие, дисперсията σ^2_{ε} на ε_l около $\bar{\varepsilon}$ е постоянна и не зависи от времето.

Това ограничение може да се разбира така. Нека е дадена една физическа система, термодинамичното равновесие на която в момента t_0 е нарушено, като е прибавено или отнето от системата известно количество топлина. В един следващ момент $t_1 > t_0$ системата отново ще бъде практически в равновесие. В интервала от време $[t_0, t_1]$ с дължина $t_1 - t_0$ порциите енергии ε_l, на които се разпада тоталната енергия E, а следователно и честотите p_l на елементарните системи с енергии ε_l, ще се променят по някакъв начин, като се стремят да вземат онези стойности, които отговарят на термодинамичното равновесие на системата при новите условия. Промените на p_l и ε_l с времето ще затихват все повече и повече и от момента, в който системата практически ще бъде отново в равновесие, разпределението на тези промени ще бъде практически стационарно.

Това ограничение е в духа на принципа на Онзагер за микроскопична обратимост, който може да бъде изказан така [2], [4]: „В условията на равновесие всеки молекулярен процес и процес, обратен на дадения, протичат средно с еднаква скорост."

Понеже дисперсията σ^2_{ε} е равна на

(7) $$\sigma_\varepsilon^2 = \sum_{i=1}^{n} \varepsilon_i^2 p_i - \bar{\varepsilon}^2,$$

допълнителното ограничение, което ще отразява условието, че системата е в термодинамично равновесие, може да бъде дадено с условието

(8) $$c) \quad \sum_{i=1}^{n} \varepsilon_i^2 p_i = q.$$

Изразът за p_i, който ще търсим при тази постановка на задачата, се получава, като се максимизира $\ln G$ при условията

(9)
$$a) \quad \varphi_1 = \sum_{i=1}^{n} p_i = 1,$$
$$b) \quad \varphi_2 = \sum_{i=1}^{n} \varepsilon_i p_i = \bar{\varepsilon} = \frac{E}{N},$$
$$c) \quad \varphi_3 = \sum_{i=1}^{n} \varepsilon_i^2 p_i = q.$$

Максимумът на S при условия a), b) и c) се дава от релациите

(10) $$V = S + \lambda\varphi_1 + \lambda_1\varphi_2 + \lambda_2\varphi_3,$$
$$\delta V = 0.$$

По-нататък имаме

(9') $$V = \sum_{i=1}^{n} \{-p_i \ln p_i + \lambda p_i + \lambda_1 \varepsilon_i p_i + \lambda_2 \varepsilon_i^2 p_i\}$$

и

(10') $$\delta V = \sum_{i=1}^{n} \{-\ln p_i - 1 + \lambda + \lambda_1 \varepsilon_i + \lambda_2 \varepsilon_i^2\} \delta p_i = 0,$$

откъдето се получават връзките

(11) $$\ln p_i = \lambda_0 + \lambda_1 \varepsilon_i + \lambda_2 \varepsilon_i^2, \quad \lambda_0 = \lambda - 1.$$

Изразът $\lambda_0 + \lambda_1 \varepsilon_i + \lambda_2 \varepsilon_i^2$ може да се напише още така:

$$\lambda_0 + \lambda_1 \varepsilon_i + \lambda_2 \varepsilon_i^2 = \lambda_2 \left(\varepsilon_i + \frac{\lambda_1}{2\lambda_2}\right)^2 + \lambda_0 - \frac{\lambda_1^2}{4\lambda_2};$$

като се положи $\frac{\lambda_1}{2\lambda_2} = -\mu$, $\lambda_0 - \mu^2\lambda_2 = \varrho$, за p_i може да се напише формулата

(12) $$p_i = e^{\varrho} e^{\lambda_2(\varepsilon_i - \mu)^2}$$

От условие a) следва

(13) $$e^{-\varrho} = \sum_{i=1}^{n} e^{\lambda_2(\varepsilon_i - \mu)^2}$$

От (12) следва, че $\lambda_2<0$, понеже p_i е честота и $0<p_i\leq 1$. Окончателният вид на честотата p_i е

$$p_i = \frac{e^{-\beta^2(\varepsilon_i-\mu)^2}}{\sum_{i=1}^{n} e^{-\beta^2(\varepsilon_i-\mu)^2}}, \tag{14}$$

където $\lambda_2=-\beta^2$, и ентропията S на системата в равновесие е

$$S=\beta^2\sigma_\mu^2+\ln\sum_{i=1}^{u} e^{-\beta^2(\varepsilon_i-\mu)^2} \tag{15}$$

§ 3. Аналогичният модел в статистическата лингвистика изглежда така. Разглежда се достатъчно обемиста извадка от писмен текст. Счита се, че когато е писан текстът, от който се прави извадката, езикът е достигнал определено развитие, което ще се схваща като състояние на езика по това време. Това състояние ще бъде състояние на равновесие, стига езикът да е формиран отдавна и да е имал дълго развитие.

Елементарните системи са съчетанията от k последователни букви и са на брой N, колкото такива съчетания се съдържат в разглежданата извадка от текст. Като пример ще се разгледат съчетания от 2 букви и за по-нататъшното изложение ще имаме $k=2$.

Всяко съчетание от две букви представлява отделно състояние на елементарната система. Например българската азбука има 31 знака (букви заедно с шпацията), от които могат да се образуват $31\times 31=961$ съчетания от две букви. Всяко от тези съчетания представлява отделно състояние на елементарната система от две букви и съдържа някаква енергия.[1]

Това са всички възможни състояния, в които може да се срещне една елементарна система от две букви в българския език, и броят им е $R=961$.

В писмен текст, който съдържа $N=68\,251$ съчетания от две букви, са изброени само $m=585$ различни съчетания вместо 961. Всички 585 съчетания от две букви са разпределени между $n=219$ различни честоти, така че броят на различните състояния, които се срещат в извадката от 68 251 съчетания, е $n=219$.

Ако допуснем, че съществува аналогия между термодинамичните процеси и процесите, които се осъществяват при развитието на даден език, от (14) енергиите ε_i като функции на p_i могат да бъдат пресметнати еднозначно. Като положим $e^{\beta^2(\varepsilon_i-u)^2}=u_i$, за честотата p_i имаме израза

$$p_i=\frac{u_i}{\sum_{i=1}^{n} u_i}, \tag{16}$$

откъдето за u_i се получава хомогенната линейна система

$$p_iu_1+p_iu_2+\cdots+(p_i-1)u_i+\cdots+p_iu_n=0,\quad i=1,2,\ldots,m. \tag{17}$$

[1] Енергията, която съдържат съчетанията, може да се мери с усилието, с което става изговарянето им.

Системата (17) има единствено, с точност до постоянен множител, решение, понеже

$$A=\begin{vmatrix} p_1-1 & p_1 & p_1 \dots p_1 \\ p_2 & p_2-1 & p_2 \dots p_2 \\ \\ p_m & p_m & p_m \dots p_m-1 \end{vmatrix} =(-1)^m\left(1-\sum_{i=1}^{m} p_i\right)=0 \quad \left(\sum p_i=1\right)$$

и

$$A_{mm}=\begin{vmatrix} p_1-1 & p_1 & \dots p_1 \\ p_2 & p_2-1 & \dots p_2 \\ \\ p_{m-1} & p_{m-1} & \dots p_{m-1}-1 \end{vmatrix} =(-1)^{m-1}\left(1-\sum_{i=1}^{m-1} p_i\right)=(-1)^{m-1}p_m\neq 0.$$

Ако броят на различните честоти, или, което е все едно, на различните видове енергии ε_i, на които се разпада тоталната енергия E на системата, е $n<m$, което фактически се наблюдава в извадката от 68 251 съчетания от по две букви, тогава за онези две значения u_i и u_j, за които $p_i=p_j$, ще се получат и равни енергии, което следва от връзката (16).

Числеността пресмятане на енергиите ε_i, които отговарят на съчетания от две букви с честота p_i, е извършено по следния начин. Честотите p_i, $i=1, 2, \dots, n$, се нареждат по намаляващи стойности. От (14) за $\ln p_i$ се получава

$$\ln p_i=-\beta^2(\varepsilon_i-\mu)^2-\ln \sum_{i=1}^{n} e^{-\beta^2(\varepsilon_i-\mu)^2} \tag{18}$$

Формула (18) зависи от параметрите μ и β. От μ зависи изборът на началото на координатната система, а от β се определя мащабът по оста ε. За μ може да се избере стойност, равна на $\mu \quad \varepsilon_1$. Мащабът по оста ε може да остане 1:1, като в (14) се положи β^2-1.

От (18) следва

$$\ln p_i \quad -(\varepsilon_i \quad \varepsilon_1)^2-\ln \sum_{i=1}^{n} e^{-(\varepsilon_i-\varepsilon_1)} \tag{19}$$

и

$$\ln p_1=-\ln \sum_{i=1}^{n} e^{-(\varepsilon_i-\varepsilon_1)}. \tag{20}$$

От (19) и (20) се получава

$$\ln p_1 \quad \ln p_i=(\varepsilon_i-\varepsilon_1)^2. \tag{21}$$

Ако са дадени абсолютните честоти N_i, понеже $p_i=N_i/N$, от (21) се получава

$$\delta_i=\varepsilon_i-\varepsilon_1=+\sqrt{\ln N_1-\ln N_i}. \tag{22}$$

От (22) се вижда, че елементарните енергии ε_i могат да се пресметнат с разлика до една константа ε_1.

В приложената накрая таблица, колона 2, са дадени резултатите от изброяванията на съчетанията от две букви. В колона 3 на таблицата са дадени честотите N_i на съответните съчетания, получени от текст, който съдържа 68 251 съчетания по две букви, наредени по намаляваща стойност. В колона 4 са пресметнати разликите $\ln N_1 - \ln N_i$, които, от една страна, дават енергиите $\delta_i = \varepsilon_i - \varepsilon_1$ с разлика до константа, равна на ε_1, пресметнати по формулата на Болцман, а, от друга страна, дават $(\varepsilon_i - \mu)^2$, когато $\mu = \varepsilon_1$. В колона 5 са дадени енергиите $\delta_i = \varepsilon_i - \varepsilon_1$, пресметнати до константа ε_1 по формула (14).

Когато макросистемата е в състояние на равновесие, броят n на различните състояния, които вземат елементарните системи, е далеч по-малък от броя R на състоянията, в които може въобще да се срещне една елементарна система. Това поражда идеята, че при равновесно състояние на макросистемата разпределението на тоталната енергия E на елементарни енергии ε_i не е съвсем произволно. Вероятно елементарните енергии ε_i имат едно определено разпределение, което е типично за една макросистема в състояние на равновесие.

Идеята, че елементарните енергии ε_i не са разпределени произволно, когато макросистемата е в равновесие, а също така разкриването на свойствата на $\delta_i = \varepsilon_i - \varepsilon_1$ (вж. таблицата, колона 5) са предмет на следващи изследвания.

Засега ще се отбележи само следното свойство: ако се образува една извадка от енергиите $\delta_i = \sqrt{\ln N_1 - \ln N_i}$, като всеки вид енергия участвува само един път, честотите им са разпределени нормално с параметри

$$\bar{\delta} = \frac{\sum \delta_i}{n}; \quad \sigma_{\delta_0} = \frac{\sum (\delta_i - \bar{\delta})^2}{n}.$$

Таблица 1

Съчетания от две букви в българския език, наредени по намаляваща честота, и на пресметнатите им енергия

№	Съчетания	Честоти	ε_i по Болцман	$\delta_i = \sqrt{\lg N_1 - \lg N_i}$
1	2	3	4	5
1	А-	3198		
2	Е-	2551	0,0982	0,3133
3	И-	2032	0,1970	0,4438
4	-С	1654	0,2863	0,5351
5	О-	1528	0,3208	0,5664
6	НА	1303	0,3899	0,6244
7	-Н	1238	0,4122	0,6420
8	-П	1016	0,4980	0,7057
9	-И	966	0,5199	0,7210
10	КА	932	0,5355	0,7318
11	-К	914	0,5439	0,7375
12	ТО	898	0,5516	0,7427
13	-Д	878	0,5614	0,7493
14	ТА	778	0,6139	0,7835
15	АТ	753	0,6281	0,7925
16	-В	740	0,6356	0,7972
17	-Т	663	0,6834	0,8267
18	РА	628	0,7069	0,8408
19	СЕ	620	0,7125	0,8441
20	-О	599	0,7275	0,8529
21	ДА	576	0,7445	0,8628
22	ВА	568	0,7505	0,8663
23	ІШЕ	567	0,7513	0,8668
24	-М	565	0,7528	0,8676
25	НИ	552	0,7629	0,8734
26	НЕ	544	0,7692	0,8770
27	ТЕ	539	0,7732	0,8793
28	-Б	529	0,7814	0,8840
29	СТ	515	0,7931	0,8906
30	Т-	510	0,7973	0,8929
31	ПО	508	0,7990	0,8939
32	КО	478	0,8255	0,9086
33	Я-	476	0,8273	0,9096
34	ОТ	465	0,8373	0,9150
35	-Г	450	0,8517	0,9229
36	ИТ	434	0,8674	0,9313
37	ЗА	428	0,8734	0,9346
38	ХА	418	0,8837	0,9401
39	-З	416	0,8858	0,9412
40	ЕН	412	0,8900	0,9434
41	РЕ	408	0,8942	0,9456
42	ЛИ	396	0,9072	0,9525
43	ЕД	394	0,9094	0,9536
44	АЛ	392	0,9116	0,9548
45	НО	385	0,9194	0,9589
46	ЛЕ	381	0,9240	0,9612
47	ПР, ЕТ	365	0,9326	0,9657
48	ЧЕ	355	0,9547	0,9771
49	ВИ	338	0,9760	0,9879
50	ЕШ	329	0,9877	0,9938
51	-Р	328	0,9890	0,9945
52	СИ	325	0,9930	0,9965

1	2	3	4	5
53	БЕ	312	1,0107	1,0055
54	АВ, РИ	308	1,0163	1,0080
55	ТИ	300	1,0278	1,0139
56	МА	299	1,0292	1,0144
57	ОВ	293	1,0380	1,0188
58	АШ	289	1,0440	1,0218
59	АН	287	1,0470	1,0232
60	-Ч	286	1,0485	1,0242
61	АХ	282	1,0546	1,0271
62	М-, ВЕ	271	1,0619	1,0305
63	ЪР	267	1,0784	1,0383
64	ИН	266	1,0800	1,0392
65	ДЕ	265	1,0816	1,0402
66	ЛА	264	1,0833	1,0407
67	ДО	261	1,0882	1,0431
68	АЗ	259	1,0916	1,0450
69	МЕ	255	1,0983	1,0479
70	-Е, ГО	254	1,1000	1,0488
71	АР	247	1,1122	1,0545
72	АК	246	1,1139	1,0555
73	Й-	230	1,1432	1,0691
74	-А	225	1,1527	1,0738
75	ДИ	224	1,1546	1,0747
76	АМ	223	1,1566	1,0756
77	С-	212	1,1785	1,0858
78	Н-	210	1,1827	1,0877
79	ОЛ	209	1,1847	1,0886
80	ОГ, МИ	206	1,1910	1,0913
81	ИЗ	205	1,1931	1,0922
82	ЕЛ	204	1,1952	1,0932
83	ГА	202	1,1995	1,0954
84	ОЙ	196	1,2126	1,1014
85	ОР, ВЬ	192	1,2216	1,1054
86	ТР	184	1,2401	1,1136
87	ДН, ОС	183	1,2424	1,1145
88	КЪ	181	1,2472	1,1167
89	ПА	180	1,2496	1,1180
90	ВО	179	1,2520	1,1189
91	ИЯ	178	1,2545	1,1198
92	КИ	175	1,2618	1,1234
93	В-	173	1,2668	1,1256
94	ЯТ	171	1,2719	1,1278
95	У-	169	1,2770	1,1300
96	К-	167	1,2822	1,1323
97	ГЛ, МО	165	1,2874	1,1345
98	АД	162	1,2954	1,1380
99	-Л	159	1,3035	1,1415
100	ИЦ, РО	151	1,3259	1,1515
101	ОД	150	1,3288	1,1528
102	ТЪ, СК	149	1,3317	1,1541
103	СЛ	148	1,3346	1,1554
104	ЯХ, ИЧ	145	1,3425	1,1593
105	ОЧ	143	1,3495	1,1619
106	ЖЕ, ЛО, АС	140	1,3588	1,1658
107	Д-, ЪЛ	137	1,3682	1,1696
108	СЪ, РЬ	135	1,3745	1,1726
109	НЯ, -Х	134	1,3778	1,1739
110	ЕГ	133	1,3810	1,1752
111	Л-	132	1,3843	1,1764

1	2	3	4	5
112	СА, ЛЯ	131	1,3876	1,1781
113	ЦА	130	1,3909	1,1794
114	ЕМ	128	1,3977	1,1824
115	МУ	126	1,4045	1,1853
116	ИМ	125	1,4080	1,1866
117	ИС	123	1,4150	1,1895
118	-Я, ИР, ЪТ	122	1,4185	1,1912
119	ЛК	120	1,4257	1,1942
120	-У	118	1,4330	1,1971
121	ОК	117	1,4367	1,1987
122	ИД	116	1,4404	1,2000
123	ЧИ, АП	115	1,4442	1,2017
124	ЕС	113	1,4518	1,2050
125	З-	112	1,4557	1,2066
126	ЕЗ	110	1,4635	1,2095
127	ИК, ЕР	109	1,4675	1,2112
128	ЪМ, ДР	108	1,4715	1,2128
129	ЩЕ, ВС	106	1,4796	1,2166
130	Х-, ЕК	105	1,4837	1,2182
131	ТЯ, РУ, АЙ	103	1,4920	1,2215
132	ОМ	102	1,4963	1,2231
133	БИ	100	1,5049	1,2268
134	ЗБ	98	1,5137	1,2304
135	ЕЧ	97	1,5181	1,2321
136	ЩО	96	1,5226	1,2341
137	ЧА	95	1,5272	1,2357
138	ИЛ, БЯ, ТУ	92	1,5411	1,2414
139	ЖД	91	1,5458	1,2434
140	БО, ОБ, ЯК	89	1,5555	1,2470
141	ЦИ, ОН	87	1,5654	1,2510
142	СН	86	1,5704	1,2530
143	ИВ, ДВ, ЪК	85	1,5755	1,2550
144	ТВ, ЧК	84	1,5806	1,2574
145	ГИ, ЗН, -Щ, ПЪ	83	1,5858	1,2594
146	РЯ	81	1,5964	1,2633
147	БР, ОЯ, МЪ	80	1,6018	1,2657
148	ПИ	79	1,6073	1,2677
149	Ш-	77	1,6184	1,2720
150	АБ	75	1,6298	1,2767
151	ОП	74	1,6356	1,2791
152	ОЖ, УС	73	1,6416	1,2814
153	ЦЕ, СП, ДЯ	72	1,6475	1,2837
154	РН	71	1,6536	1,2857
155	УМ	69	1,6660	1,2907
156	-Ж, ИЕ, ДЪ	68	1,6724	1,2931
157	ДУ, ВР, ОЩ	67	1,6788	1,2958
158	ЛН, БА	66	1,6853	1,2981
159	ЖА	65	1,6920	1,3008
160	ЗЕ, КУ, КР, ЪЖ, ИХ, УВ, ЯД	64	1,6987	1,3035
161	ЛУ, ОИ, ЪС, КВ	63	1,7055	1,3061
162	ГР, ЧУ, ЩА	62	1,7125	1,3084
163	УБ, ОЗ, ГН, ЕХ, СВ	61	1,7195	1,3115
164	АЯ, Р-, ХН	60	1,7267	1,3142
165	ПЕ, ВЯ, МН	59	1,7340	1,3168
166	УД, АГ, УК, ПЛ	58	1,7415	1,3195
167	СМ, ШИ, ЕЖ, ЯЛ	57	1,7490	1,3225
168	УГ	56	1,7567	1,3255
169	УЛ	55	1,7645	1,3285
170	ЕЩ, ЪЩ, ХО, ТН	53	1,7806	1,3345

1	2	3	4	5
171	ЗИ, ИГ	52	1,7889	1,3375
172	ЙТ, ЗД, АЖ, РВ	51	1,7973	1,3405
173	ПУ, УШ	50	1,8059	1,3439
174	ЗП, ЧН	49	1,8147	1,3472
175	ЕВ	48	1,8236	1,3506
176	ЧО, НТ, ЗО	47	1,8328	1,3539
177	-Ш, БЪ	46	1,8421	1,3572
178	ОЕ, ХУ, ЪЙ, ЯМ, ЖИ	45	1,8517	1,3609
179	ГЪ, ИШ	43	1,8714	1,3678
180	ВН, ЪД, ХМ, ЪН, НЪ	42	1,8816	1,3719
181	ИЖ, ЪЛ	41	1,8921	1,3755
182	КТ	39	1,9138	1,3835
183	ОХ, ЕП	38	1,9251	1,3874
184	АЕ, УП, СЯ	37	1,9387	1,3925
185	ЛБ, МЯ, ХЧ	36	1,9486	1,3961
186	ЯВ, ЯС, ЙД, ХВ	35	1,9608	1,4004
187	ЯГ, ЕЯ, БУ, ЯН	34	1,9734	1,4046
188	ИЩ	33	1,9864	1,4093
189	СО, ША	32	1,9997	1,4142
190	ЕЙ, ЕБ, -Ц	31	2,0135	1,4192
191	ЗБ, АК, ЪЗ,	30	2,0278	1,4241
192	РШ, РЗ, АЦ	28	2,0577	1,4346
193	СР, КЕ, ЗЛ, УТ, Ж-, АЧ, ЗГ	27	2,0735	1,4401
194	ЪВ, АЩ, ЛЧ, УЦ	26	2,0899	1,4457
195	РД, ШН, ЩИ, ТТ, РТ, УХ	25	2,1069	1,4516
196	ЙН, ГЕ, ХР, ЛГ, КН	24	2,1247	1,4577
197	ЪБ, ЕЕ	23	2,1432	1,4639
198	ЛЪ, ВЛ, ТК, ЗК, ГУ, МР	22	2,1625	1,4704
199	ПН, РК, УН, ЗР, ЪЧ, ЪЦ	21	2,1827	1,4775
200	ШУ, ЪГ, ОШ, ВК, СУ, НД, ЯЗ, ЪП, УЙ	20	2,2038	1,4846
201	Г-, ШК, УЖ, ЯШ, ВЗ	19	2,2261	1,4920
202	ЙК, КЛ, ЗМ, ЗС, ГД, ЙС	18	2,2496	1,5000
203	РЛ, РЦ, РХ, ИП	17	2,2744	1,5080
204	ЯБ, РЧ, ЕЦ, МЛ	16	2,3008	1,5169
205	РС, ЗЪ	15	2,3288	1,5261
206	БВ, Щ-, ЯР, ЩЯ, ЛЮ, НС	14	2,3588	1,5359
207	Ч-, ПК, ТП, ШО, НУ	13	2,3909	1,5463
208	РП, -Ф, НЦ, ХИ, ЪХ, ПЯ	12	2,4257	1,5576
209	ФА, ИБ, ИИ, ВЦ, РЖ, ЗТ, ЖЪ, ЛП	11	2,4535	1,5662
210	ТГ, ЖК, ХЛ, ЕУ, БН, НН	10	2,5049	1,5827
211	ЖН, ЛЖ, ТС, ЧМ, ЛЗ, АИ, -Ю	9	2,5506	1,5972
212	КМ, УЩ, ШЛ, ИЙ, ФЕ, ВТ, ХР, ЛМ, ПЧ, АО, АС ЧВ, ГВ, ГЯ, ЮШ	8	2,6018	1,6131
213	ЛГ, НЧ, ТМ, МЧ, ДК, ТЛ, ДС, ОУ, ОО, П-, ТБ, АФ	7	2,6598	1,6310
214	ТЮ, ИО, РГ, КЮ, ЩЯ, ФИ, Ю-, ВЧ, ЗЯ, СБ, РБ	6	2,7267	1,6514
215	ТЗ, ТЧ, ХЕ, ЮЗ, ЮН, ВУ, РМ, НЗ, ХЪ, МФ, УЗ, ЗХ, ЙП, ДЧ, ПВ	5	2,8059	1,6751
216	ОФ, ЩУ, ХТ, ОА, СГ, СЮ, ЩН, ЬО, ЦЪ, ЮТ, ЛЦ, ДП, АУ, ЯЧ, МТ, ДЖ, НВ, ЕО, ПО, ФН, МБ	4	2,9028	1,7038
217	-Ъ, МЦ, Б-, ИА, МВ, ЛВ, БЩ, ЯЩ, ЪФ, НГ, ЩТ, ВД, ЙМ, ЗЧ, КД, ВП, Ц-, ПВ, ЮЧ	3	3,0278	1,7900
218	БС, ДБ, ЗУ, ГЧ, СЧ, ЧТ, ПМ, ЗЖ, ЛЛ, ФЛ, УЯ, МК, ЦВ, ЖУ, ОЦ, ПС, ТЬ, КЧ, УУ, КС, ЮС, ЩЬ, ПТ, ФД, ЙВ, ДГ, ЕФ, ИФ, НФ, ШЬ, ЪШ, ЙГ	2	3,2038	1,7900
219	ЕА, ЮВ, ЛА, ГЬ, ДЮ, ЮБ, ДЩ, ШМ, ФУ, ГД, ЯЯ, ЛД, МД, ЮЖ, ДХ, ГМ, ДМ, МЖ, ЗЦ, БЖ, НШ, ЙЦ, ВМ, ХС, КЦ, ЯЖ, ХЕ, ЙО, БТ, ЖТ, СД, УО, ВХ, НЮ, ЗЗ, ЙЧ, МС, ЖО, ДТ, МП, БЧ, ФО, ЮМ, ЪЕ, ЪО, СХ, МШ, ТХ, БМ, РР, ДЛ, СЖ	1	3,5049	1,8722

ЛИТЕРАТУРА

1. Belevitch V., Langage des machines et langage humain, Of. coll. Labégue et nationale, Bruxelles, 1956.
2. Denbigh K. G., The Thermodynamics of the steady state, New York, 1951.
3. Gotusso L., Una dimostrazione elementare della legge sperimentale di Estoup-Zipf, Rend. di mat. Univ. di Roma, voll. XXII, fasc. 1-2, 1963.
4. Groct S. R., Thermodynamics of irreversible processes, Amsterdam, 1952.
5. Mandelbrot B., On the theory of word frequencies and on related marcovian models of discourse, Proceedings of the twelfth symposium in applied mathematics, Ed. R. Jacobson, New York, 1960.
6. Planck M., Einführung in die theoretische Physik, Bd. V, Leipzig, 1930.
7. Schrödinger E., Statistische Thermodynamik, Leipzig, 1952.
8. Zipf G. K., Human behaviour and the principle of least effort, Reading, Mass., 1949.

Постъпила на 7. XII. 1963 г.

ОБ ОДНОМ СПОСОБЕ ОПРЕДЕЛЕНИЯ „ЦЕНЫ" БУКВЕННЫХ СОЧЕТАНИЙ В ЯЗЫКЕ. I

Эмануил Симеонов

(Резюме)

В предлагаемой работе рассматривается развитие языка в аналогии с одним термодинамическим процессом болцмановского типа, причем каждое сочетание из двух букв принимается как элементарная система в определенном состоянии.

Как в термодинамике необратимых процессов, так и здесь средняя энтропия S элементарной системы выражается через

$$S = -\sum p_i \ln p_i,$$

где p_i представляет собой частоту элементарной системы в состоянии i.

Ограничительные условия Болцмана слишком общие и относятся к любому состоянию системы. В данном же случае энтропия S максимизируется при следующих ограничениях:

а) $$\sum p_i = 1,$$

б) $$\sum \varepsilon_i p_i = \bar{\varepsilon},$$

в) $$\sum \varepsilon_i^2 p_i = b.$$

Ограничение в) относится к состоянию равновесия в системе.

На таблице в болгарском тексте статьи даны вычисленные энергии сочетаний из двух последовательных букв. Эти энергии распределены нормально с параметрами

$$\delta = \frac{\sum \delta_i}{n}, \quad \sigma_\delta^2 = \frac{\sum (\delta_i - \bar{\delta})^2}{n}.$$

ON A METHOD OF DETERMINING "PRICES" OF COMBINATIONS OF LETTERS IN THE WRITTEN LANGUAGE. I.

Emanouil Simeonov

(Summary)

The paper examines the development of the language in analogy to a thermodynamic process of the Boltzmann type, each combination of two letters being considered as a microsystem in a given state.

In this case, as in the thermodynamics of the irreversible processes, the average entropy S of a microsystem is given by the expression

$$S = -\sum p_i \ln p_i ,$$

wherein p_i is the frequency of the microsystem in the i-th state.

Boltzmann's restricting conditions are of a very general character and they are related to any state of the system. Here the entropy S is maximized under the following restrictions:

a) $$\sum p_i = 1,$$

b) $$\sum \varepsilon_i p_i = \varepsilon ,$$

c) $$\sum \varepsilon_i^2 p_i = b.$$

Restriction c) is related to the equilibrium state of the system.

The table accompanying the text gives the calculated "prices" of the combinations of two consecutive letters. These "prices" are normally distributed with parameters

$$\delta = \frac{\sum \delta_i}{n}, \quad \sigma_\delta^2 = \frac{\sum (\delta_i - \delta)^2}{n} .$$