УДК–519.2

К ЗАКОНУ ОБРАЗОВАНИЯ ЛИНГВИСТИЧЕСКИХ ЭЛЕМЕНТОВ В. ФУКСА

Р. МЕРКИТЕ, В. КАЛИНКА

1. В математической лингвистике большой интерес представляет исследование образования элементов речи из составляющих компонентов, например, образование предложений из слов; слов из слогов, букв, звуков; слогов из букв и т. д.

Изучением распределения частот компонентов элементов речи занимались С. Г. Чебанов [1] и В. Фукс [2]. В указанной статье В. Фукс описал статистическую модель образования лингвистических элементов из определенных компонентов и вывел формулу, описывающую распределение этих компонентов в элементах. Рассмотрим случай образования таких элементов, которые всегда содержат хотя бы один компонент (каждое слово имеет хотя бы один звук, каждое предложение – хотя бы одно слово и т. д.). Мы пользуемся слегка модифицированной моделью В. Фукса, так как считаем, что его модель не безупречна с математической точки зрения. Внесенная нами модификация, на наш взгляд, позволяет строить модель на достаточно строгой математической основе.

Приступаем к описанию модели. Для простоты изложения будем считать, что рассматривается образование слов из слогов.

Пусть имеется большое количество шаров (слогов). Будем их распределять по большому числу ящиков (слов), каждый из которых вмещает сколь угодно большое число шаров.

В каждый ящик поместим по одному шару (каждое слово содержит хотя бы один слог). Дальнейшее распределение будем осуществлять стохастически в два этапа.

В первом этапе шары распределим по дискретному закону

$$P\{\tilde{\xi}=l-1\}=\alpha_l,\ l=1,\ \ldots,\ n,\ \sum_{l=1}^{n}\alpha_l=1,$$

где $\tilde{\xi}$ – случайная величина, означающая число шаров в ящике и принимающая n целых значений от 0 до $n-1$. Если сюда добавить ранее помещенный в каждый ящик шар и обозначить число шаров в ящике ξ, то

$$P\{\xi=l\}=\alpha_l,\ l=1,\ \ldots,\ n,\ \sum_{l=1}^{n}\alpha_l=1. \tag{1}$$

Во втором этапе шары будем распределять по закону Пуассона с параметром λ

$$P\{\eta=m\}=\frac{e^{-\lambda}\lambda^m}{m!}, \qquad m=0,\ 1,\ 2,\ \ldots \tag{2}$$

Здесь λ означает среднее число шаров, попавших в один ящик во втором этапе.

Появление в модели двух разных стохастических законов распределения легко понять на конкретном примере. Интуитивно ясно, что слоги, составляющие основу слов, и остальные слоги должны быть распределены по разным законам.

Обозначив общее количество шаров в ящике (число слогов, составляющих слово) ζ, получаем

$$\zeta=\xi+\eta,$$

где ξ и η — независимые случайные величины, распределенные по законам (1) и (2). Случайная величина ζ будет распределена по закону, который является композицией законов (1) и (2). Таким образом, характеристическая функция ζ будет

$$Me^{it\zeta}=Me^{it\xi}\cdot Me^{it\eta}=\sum_{\nu=1}^{n}\alpha_\nu e^{it\nu}\exp\{\lambda(e^{it}-1)\}.$$

Тогда распределение компонентов в элементах описывается законом

$$p_k=P\{\zeta=k\}=e^{-\lambda}\sum_{m=1}^{\min(k,\,n)}\frac{\alpha_m\lambda^{k-m}}{(k-m)!}=$$

$$=e^{-\lambda}\sum_{m=\max(0,\,k-n)}^{k-1}\frac{\alpha_{k-m}\lambda^m}{m!}, \qquad k\geqslant 1, \qquad \sum_{m=1}^{n}\alpha_m=1. \tag{3}$$

Соотношение (3) доказывается ниже.

Для дальнейшего удобно определить α_m для всех $m>n$ следующим образом: $\alpha_{n+1}=\alpha_{n+2}=\ \ldots\ =0$ и ввести новые параметры β_m так, что $\beta_1=1$, $\alpha_m=\beta_m-\beta_{m+1}$, $m\geqslant 1$. В. Фукс указал, что из формул первых четырех центральных моментов можно найти первые β_m и λ. Однако удобные для вычислений аналитические выражения β_m и λ не были найдены, что затрудняло практическое применение закона (1). В. Фукс ограничился нахождением $\beta_1=1$, β_2 и λ. Мы не встретили ни одной работы, где оценивается большее число параметров, хотя наши исследования показали, что с увеличением числа параметров β_m достигается большее соответствие между теоретическими и практическими данными.

Вероятностное исследование закона (1) дало возможность авторам найти аналитические выражения для оценки β_m и λ. Цель данной работы — отыскание этих аналитических выражений и иллюстрация их применения на примере литовского языка.

В первом разделе приведена формула, дающая закон образования лингвистических элементов из компонентов. Во втором разделе приводи-

тся вероятностное исследование этого закона, в третьем — выводятся формулы для оценки параметров из (3). В четвертом, последнем, разделе приводятся удобные для вычислений формулы и рассматривается пример.

2. Для вывода (3) и формул моментов s-го порядка M_s распределения (3) рассмотрим производящую функцию, полученную из характеристической функции путем замены $e^{it}=z$

$$f(z)=\sum_{\nu} \alpha_\nu z^\nu \exp\{\lambda(z-1)\}. \tag{4}$$

Производная i-го порядка (4) функции

$$f^{(i)}(z)=e^{\lambda(z-1)} i! \sum_{m=0}^{i} A_m(z) \frac{\lambda^{i-m}}{(i-m)!}, \tag{5}$$

где

$$A_m(z)=\sum_{\nu=m}^{n} \binom{\nu}{m} \alpha_\nu z^{\nu-m}, \qquad A_0(z)=\sum_{\nu} \alpha_\nu z^\nu. \tag{5а}$$

Как известно, $p_k=\frac{1}{k!} f^{(i)}(0)$, и из (5) получаем (3).

Найдем формулу моментов M_s. Производные производящей функции дают факториальные моменты распределения (3):

$$f^{(i)}(1)=\sum_{k} k(k-1) \cdots (k-i+1) p_k;$$

s-ые моменты распределения (3) выражаются через факториальные формулой

$$M_s=\sum_{i=1}^{s} b_{si} f^{(i)}(1), \tag{6}$$

где b_{si} — коэффициенты разложения

$$k^s \equiv \sum_{i=1}^{s} b_{si} \frac{k!}{(k-i)!}. \tag{6а}$$

Как известно, для них действительна рекуррентная формула

$$b_{s+1,\,i+1}=b_{si}+(i+1)\,b_{s,\,i+1}, \quad 1 \leqslant i \leqslant s-1, \quad b_{s1}=b_{ss}=1. \tag{6б}$$

В нашем случае из (5)

$$f^{(i)}(1)=\sum_{m=0}^{i} \frac{i!}{(i-m)!} A_m \lambda^{i-m}, \tag{7}$$

где

$$A_m=A_m(1)=\sum_{\nu=m}^{n} \binom{\nu}{m} \alpha_\nu. \tag{7а}$$

Из (6) и (7) имеем:

$$M_s=\sum_{i=1}^{s} i!\, b_{si} \sum_{m=0}^{i} A_m \frac{\lambda^{i-m}}{(i-m)!}. \tag{8}$$

Первые четыре момента распределения (3), найденные по (8), будут иметь следующий вид:

$$M_1=\lambda+A_1,$$
$$M_2=\lambda^2+(2A_1+1)\lambda+2A_2+A_1,$$
$$M_3=\lambda^3+(3A_1+3)\lambda^2+(6A_2+6A_1+1)\lambda+6A_3+6A_2+A_1, \tag{9}$$
$$M_4=\lambda^4+(4A_1+6)\lambda^3+(12A_2+18A_1+7)\lambda^2+(24A_3+36A_2+14A_1+1)\lambda+$$
$$+24A_4+36A_3+14A_2+A_1.$$

3. Оценим параметры распределения (3).

Пусть имеется достаточно большая выборка x_k, $k=1, \ldots, N$ определенных лингвистических элементов, которые распределены по закону (3). Определим эмпирические моменты M_s, $s=1, 2, \ldots, n$. Заменяя в (8) теоретические моменты эмпирическими, получим систему уравнений, решив которую, найдем параметры распределения (3) λ и α_ν, $\nu=1, \ldots, n$. Итак, имеем систему

$$\sum_{i=1}^{s} i!\ b_{si} \sum_{m=0}^{i} A_m \frac{\lambda^{i-m}}{(i-m)!}=M_s, \qquad s=1, 2, \ldots, n \tag{8a}$$

с неизвестными λ, α_ν, которые входят в A_m, причем $\sum_\nu \alpha_\nu=1$. Для облегчения решения системы определим функции

$$V_m=\sum_{\nu\geqslant m}\binom{\nu}{m}\alpha_{\nu+1}, \qquad m\geqslant 1. \tag{10}$$

Легко проверить, что

$$A_m=V_m+V_{m-1}, \qquad m\geqslant 1, \qquad A_n=V_{n-1}, \qquad A_0=1. \tag{11}$$

Введем новые неизвестные $\lambda=\lambda$ и V_m, $m=2, \ldots, n-1$. Тогда

$$A_1=M_1-\lambda,$$
$$A_2=V_2-\lambda+M_1-1.$$

Остальные A_m, $m=3, \ldots, n$ определим по (11).

Тогда в уравнение M_2 войдут неизвестные λ и V_2, и при увеличении s в уравнениях M_s будут появляться новые V_s. Только в уравнении n-го момента новые неизвестные не появятся. Последовательно выражая V_m, получим уравнение n-го момента, содержащее неизвестные лишь в степени n и более низких степеней. Из этого уравнения найдем λ и, подставляя его значение в другие уравнения, найдем V_m.

Вернемся к примеру, когда $n=4$. Решая систему (9) после введения новых неизвестных, получим:

$$2V_2-\lambda^2+2(M_1-1)\lambda+3M_1-2=M_2,$$
$$6V_3-2\lambda^3+3(M_1-3)\lambda^2+12(M_1-1)\lambda+6V_2(\lambda+2)+7M_1-6=M_3,$$
$$-3\lambda^4+4(M_1-6)\lambda^3+5(6M_1-11)\lambda^2+50(M_1-1)\lambda+$$
$$+2V_2(6\lambda^2+30\lambda+25)+24V_3(\lambda+60)+15M_1-14=M_4.$$

Выражая V_m, имеем:

$$2V_2=\lambda^2-2(M_1-1)\lambda+(M_2-3M_1+2),$$

$$6V_3=-\lambda^3+3(M_1-1)\lambda^2-3(M_2-3M_1+2)\lambda+(M_3-6M_2+11M_1-6),$$

$$0=24V_4=\lambda^4-4(M_1-1)\lambda^3+6(M_2-3M_1+2)\lambda^2-$$

$$-4(M_3-6M_2+11M_1-6)\lambda+(M_4-10M_3+35M_2-50M_1+24).$$

В образовании коэффициентов вышезаписанных уравнений замечаем закономерность, которую сформулируем в виде следующей леммы.

Лемма. *Функции* V_m, *определенные по формуле* (10), *выражаются через* λ *следующим образом*:

$$V_m=V_m(\lambda)=\sum_{i=0}^{m}\frac{(-1)^{m-i}a_i\lambda^{m-i}}{(m-i)!}, \tag{12}$$

где a_i – *линейные функции моментов*

$$a_i=\sum_{s=0}^{i}c_{is}M_s, \qquad a_0=1, \tag{12а}$$

а коэффициенты c_{is} *находятся из тождества*

$$\binom{k-1}{i}\equiv\sum_{s=0}^{i}c_{is}k^s. \tag{12б}$$

Другими словами, заменив k^s через M_s, получаем, что левая часть (12б) превращается в a_i:

$$k^s\to M_s, \qquad \binom{k-1}{i}\to a_i. \tag{12в}$$

Доказательство. По (11) имеем

$$A_m=V_m+V_{m-1}=\sum_{j=0}^{m}\frac{(-1)^{m-j}a_j\lambda^{m-j}}{(m-j)!}+\sum_{j=0}^{m-1}\frac{(-1)^{m-j-1}a_j\lambda^{m-j-1}}{(m-j-1)!}=$$

$$=\frac{(-1)^m\lambda^m}{m!}+\sum_{j=1}^{m}\frac{(-1)^{m-j}(a_j+a_{j-1})\lambda^{m-j}}{(m-j)!}.$$

Для доказательства значение A_m подставим в левую часть (8а) и, преобразуя, получим тождество $M_s=M_s$. Так,

$$\sum_{i=1}^{s}i!\,b_{si}\sum_{m=0}^{i}A_m\frac{\lambda^{i-m}}{(i-m)!}=\sum_{i=1}^{s}b_{si}\lambda^i+$$

$$+\sum_{i=1}^{s}i!\,b_{si}\sum_{m=1}^{i}A_m\frac{\lambda^{i-m}}{(i-m)!}=$$

$$=\sum_{i=1}^{s}b_{si}\lambda^i+\sum_{i=1}^{s}i!\,b_{si}\sum_{m=1}^{i}\frac{(-1)^m\lambda^i}{m!\,(i-m)!}+$$

$$+\sum_{i=1}^{s}i!\,b_{si}\sum_{m=1}^{i}\sum_{j=1}^{m}\frac{(-1)^{m-j}(a_j+a_{j-1})\lambda^{i-j}}{(m-j)!\,(i-m)!}.$$

Объединив первые два члена, получим выражение, равное нулю. Изменив в последнем члене порядок суммирования, имеем

$$\sum_{i=1}^{s} i!\ b_{si} \sum_{j=1}^{i} (a_j + a_{j-1})\lambda^{i-j} \sum_{m=j}^{i} \frac{(-1)^{m-j}}{(m-j)!\,(i-m)!} =$$

$$= \sum_{i=1}^{s} i!\ b_{si} \sum_{j=1}^{i} (a_j + a_{j-1})\lambda^{i-j} \sum_{m=0}^{i-j} \frac{(-1)^m}{m!\,(i-j-m)!}.$$

Полученное выражение обращается в нуль при $j \neq i$. Поэтому остается только член, когда $j=i$

$$\sum_{i=1}^{s} i!\ b_{si}(a_i + a_{i-1}).$$

Преобразуя полученное выражение, наконец, получаем

$$\sum_{i=0}^{s} i!\ b_{s+1,\,i+1} a_i = M_s, \tag{13}$$

которое легко доказывается. А именно, преобразуем (6а)

$$k^{s+1} \equiv \sum_{i=0}^{s} b_{s+1,\,i+1} \frac{k!}{(k-i-1)!}$$

или, сократив равенство на k,

$$k^s \equiv \sum_{i=0}^{s} b_{s+1,\,i+1} \frac{(k-1)!}{(k-i-1)!} \equiv \sum_{i=0}^{s} i!\ b_{s+1,\,i+1} \binom{k-1}{i}.$$

На основании (12в) последнее соотношение эквивалентно (13). Лемма доказана.

Для оценки λ приходится решать уравнение $V_n(\lambda)=0$. Поскольку α_m будем рассматривать как функции λ и моментов, из всех корней уравнения в качестве оценки λ выберем тот, который удовлетворяет условию $\alpha_m \geqslant 0$, $m=1, \ldots, n$. Этому условию удовлетворяет наименьший положительный корень уравнения.

Зависимость α_m от λ и моментов выражает

Теорема. *При известных* λ *и* M_s, $s=1, \ldots, n$, α_{m+1} *находятся по формуле*

$$\alpha_{m+1} = \sum_{i=m}^{n-1} (-1)^{m+i} \binom{i}{m} V_i(\lambda), \tag{14}$$

где V_i *определены в* (12).

Доказательство. Для нахождения α_m, $m=1, 2, \ldots, n$ по (7а) придется решать систему

$$\sum_{\nu=m}^{n} \binom{\nu}{m} \alpha_\nu = A_m, \qquad m = 1, 2, \ldots, n. \tag{15}$$

Определитель системы равен 1. Поэтому система имеет одно решение:

$$\alpha_\nu = \sum_{i=\nu}^{n} (-1)^{i+\nu} \binom{i}{\nu} A_i. \tag{16}$$

Для доказательства подставим (16) в (15) и получим тождество. Рассмотрим левую часть (15)

$$\sum_{\nu=m}^{n} \frac{1}{(\nu-m)!\, m!} \sum_{i=\nu}^{n} \frac{(-1)^{i+\nu} i!\, A_i}{(i-\nu)!}.$$

Изменив порядок суммирования, имеем

$$\sum_{i=m}^{n} \frac{(-1)^{i+m} i!\, A_i}{m!} \sum_{\nu=0}^{i-m} \frac{(-1)^\nu}{\nu\,(i-\nu-m)!} = A_m.$$

В (16) подставим значение A_m из (11)

$$\alpha_\nu = \sum_{i=\nu}^{n-1} (-1)^{i+\nu} \binom{i}{\nu} V_i + \sum_{i=\nu}^{n} (-1)^{i+\nu} \binom{i}{\nu} V_{i-1}.$$

Преобразуя данное выражение, получаем

$$\alpha_\nu = \sum_{i=\nu-1}^{n-1} (-1)^{i+\nu+1} \binom{i}{\nu-1} V_i,$$

а это эквивалентно (14). Теорема доказана.

4. От введенных нами параметров α_m легко можно перейти к параметрам β_m

$$\beta_{m+1} = \sum_{i=m}^{n-1} \alpha_{i+1} = \sum_{i=m}^{n-1} \sum_{j=i}^{n-1} (-1)^{i+j} \binom{j}{i} V_j(\lambda).$$

Изменив порядок суммирования, будем иметь

$$\sum_{j=m}^{n-1} (-1)^j V_j(\lambda) \sum_{i=m}^{j} (-1)^i \binom{j}{i} =$$

$$= \sum_{j=m}^{n-1} (-1)^j V_j(\lambda) \sum_{i=0}^{m-1} (-1)^{i+1} \binom{j}{i}.$$

По известному тождеству

$$\sum_{i=0}^{m} (-1)^i \binom{j}{i} = \binom{j-1}{m} (-1)^m$$

получаем, что

$$\beta_{m+1} = \sum_{j=m}^{n-1} (-1)^{j+m} \binom{j-1}{m-1} V_j(\lambda), \tag{17}$$

$m \neq 0$, $\beta_1 = 1$.

При вычислениях удобно пользоваться следующими выражениями, которые просто получаются из (17):

$$n=1; \quad \beta_1=1, \quad \beta_2=0;$$

$$n=2; \quad \beta_1=1, \quad \beta_2=-\lambda+a_1, \quad \beta_3=0;$$

$$n=3; \quad \beta_1=1, \quad \beta_4=0,$$

$$\beta_2=-\frac{1}{2}\lambda^2+(a_1-1)\lambda-a_2+a_1,$$

$$\beta_3=\frac{1}{2}\lambda^2-a_1\lambda+a_2;$$

$$n=4; \quad \beta_1=1, \quad \beta_5=0,$$

$$\beta_2=-\frac{1}{6}\lambda^3+\frac{1}{2}(a_1-1)\lambda^2-(a_2-a_1+1)\lambda+a_3-a_2+a_1,$$

$$\beta_3=\frac{1}{3}\lambda^3-\frac{1}{2}(2a_1-1)\lambda^2+(2a_2-a_1)\lambda-2a_2+a_2,$$

$$\beta_4=-\frac{1}{6}\lambda^3+\frac{1}{2}a_1\lambda^2-a_2\lambda+a_3;$$

a_i определены по формуле (12а).

В качестве примера рассмотрим распределение слогов в словах литовской части „Литовско-французского словаря" И. Карсавиной, С. Кайрюкштите*. Объем выборки $N=25\,136$ слов; эмпирические моменты распределения следующие: $M_1=3{,}5303$, $M_2=13{,}5278$, $M_3=55{,}7952$, $M_4=246{,}0618$, $M_5=1154{,}3570$. Для $n=4$, $\lambda=0{,}5530$ находим из уравнения

$$\chi^4-10{,}1213\lambda^3+29{,}6210\lambda^2-29{,}8482\lambda+9{,}0659=0.$$

Тогда параметры β_m будут: $\beta_1=1$, $\beta_2=0{,}9926$, $\beta_3=0{,}7473$, $\beta_4=0{,}2373$, $\beta_5=\beta_6=\ldots=0$.

Эмпирические и теоретические данные сведены в таблицу. Проверка по критерию χ^2 подтверждает гипотезу о том, что слоги в словах данного текста распределены по закону (1).

	Эмпирическое распределение	$\lambda=2{,}5303$, $\beta_1=1$, $\beta_2=\beta_3=\cdots=0$.	$\lambda=0{,}7913$, $\beta_1=1$, $\beta_2=0{,}9598$, $\beta_3=0{,}7793$, $\beta_4=\beta_5=\cdots=0$.	$\lambda=0{,}5530$, $\beta_1=1$, $\beta_2=0{,}9926$, $\beta_3=0{,}7473$, $\beta_4=0{,}2373$, $\beta_5=\beta_6=\cdots=0$.
p_1	0,0042	0,0797	0,0182	0,0042
p_2	0,1432	0,2016	0,0962	0,1434
p_3	0,3738	0,2550	0,4236	0,3720
p_4	0,3170	0,2150	0,3066	0,3204
p_5	0,1269	0,1360	0,1176	0,1244
p_6	0,0294	0,0688	0,0306	0,0297
p_7	0,0047	0,0290	0,0060	0,0050
p_8	0,0006	0,0105	0,0009	0,0002
p_9	0,0002	0,0033	0,0001	0,0001

* J. Karsavina, S. Kairiūkštytė, Lietuvių-prancūzų kalbų žodynas, Vilnius, 1962.

В таблице во втором и третьем столбцах приведены теоретические распределения, полученные при $n=1$ и $n=3$. Эти распределения больше расходятся с эмпирическим. Улучшить оценку λ путем решения уравнения пятого порядка не удается, так как в интервале $0<\lambda<0{,}7913$ оно не имеет других корней. V_5 $(0{,}5530)=0{,}0009$, это указывает на достаточность оценки λ.

Институт физики и математики
Академии наук Литовской ССР

Поступило в редакцию
15.VIII.1967

ЛИТЕРАТУРА

1. С. Г. Чебанов, О подчинении речевых укладов „индоевропейской" группы закону Пуассона, ДАН СССР, новая сер., 55, № 2, 103–106 (1947).
2. В. Фукс, Математическая теория словообразования, Сб. статей „Теория передачи сообщений", 1957.
3. Р. Мерките, Некоторые статистические характеристики образования слов из слогов и слогов из букв для литовского языка, Лит. матем. сб., II, № 1 (1962).

APIE V. FUKSO DĖSNĮ KALBOS ELEMENTAMS TIRTI

R. Merkytė, V. Kalinka

(*Reziumė*)

Straipsnyje įvertinami V. Fukso dėsnio, aprašančio kalbos elementų sudarymą, parametrai. Tuo pasiekiamas didesnis teorinių ir praktinių duomenų atitikimas.

ON THE W. FUCKS'S DISTRIBUTION LAW FOR THE FORMATION OF LINGUISTICAL ELEMENTS

R. Merkytė, V. Kalinka

(*Summary*)

The practical method for evaluating the parameters of the Fucks's law for formation of linguistical elements is developed. This method enables to achieve the better conformity of theoretical and practical data.