XII LIETUVOS MATEMATIKOS RINKINYS
ЛИТОВСКИЙ МАТЕМАТИЧЕСКИЙ СБОРНИК 4

1972

УДК 519.21

ЗАКОН, ОПИСЫВАЮЩИЙ РАСПРЕДЕЛЕНИЕ СЛОГОВ В СЛОВАХ СЛОВАРЕЙ

Р. Ю. Мерките

Цель данной работы – выявить закон, описывающий образование слов из слогов.

На число слогов в слове можно смотреть как на случайную величину Z, принимающую значения k, $k=1, 2, \ldots, K$, где K – число слогов в самом длинном слове (число слогов в слове равно числу гласных или дифтонгов в этом слове). Пусть имеется выборка в N слов. По этой выборке можно вычислить выборочное распределение случайной величны Z. Наша задача – по выборочному распределению слогов в словах найти такое теоретическое распределение, которое было бы близким к распределению случайной величины Z.

Следует отметить, что распределения слогов в словах текстов изучались многими исследователями, которые считали, что закон, описывающий эти распределения, описывает словообразование в целом. По нашему мнению, закон, описывающий распределение слогов в словах текстов, должен учитывать многочисленные повторения слов, поэтому закон словообразования гораздо рациональней выявлять, исследуя распределение слогов в словах словарей.

По морфологическому строению выделяются синтетические языки, выражающие отношение между словами в речи посредством форм самих слов; в таких языках для передачи различных отношений между словами широко используются аффиксы (префиксы, суффиксы, окончания). Синтетическим языкам противопоставляются аналитические языки – языки, в которых отношения между словами выражаются не посредством форм слов, а посредством служебных слов, порядка слов в предложении, интонации и пр. К синтетическим языкам относятся, напр., литовский, латинский, немецкий, русский и др., к аналитическим – романские, болгарский, датский, английский и др. Такая классификация не отличается строгостью, так как во многих языках встречаются черты, характерные как для синтетических, так и аналитических языков.

В образовании слов условно выделим два этапа: 1) образование корней слов, 2) присоединение к корням аффиксов. Очевидно, образование слов в синтетических и аналитических языках будет различаться: в аналитических языках фактически будет наблюдаться лишь первый этап.

Исследуем распределение слогов в корнях слов.

Так как граница корня не совпадает с границей слога, принадлежность слога к корню будем определять по принадлежности к корню слогообразующей гласной. Число слогов в корне есть дискретная случайная величина U, принимающая значения i, $i=1, 2, \ldots, I$. По выборке объема N вычислим выборочное распределение случайной величины U; по этому распределению будем искать такое теоретическое распределение, с которым согласуется полученное выборочное распределение. Поскольку в результате исследований В. Фукса (см. [1]) распространено мнение, что образование слов из слогов в языках достаточно точно описывается пуассоновским законом, в первую очередь проверим, пригоден ли в этом случае закон Пуассона

$$P\{X=i\}=\frac{\lambda^i e^{-\lambda}}{i!}, \qquad X=U-1, \qquad i=0, 1, \ldots, I-1; \tag{1}$$

в качестве оценки параметра λ берется выборочное среднее случайной величины X

$$\hat{\lambda}=\sum_{i=0}^{I-1} ip_i. \tag{2}$$

Л. Н. Большев предлагает для проверки гипотезы о том, что выборка получена, наблюдая пуассоновские величины, применять критерий „пуассоновости" (см. [2]). Суть этого критерия состоит в следующем: показано, что при фиксированной сумме

$$K=\sum_{n=1}^{N} X_n$$

система случайных величин $X_1, X_2, \ldots, X_n \ldots, X_N$ имеет равномерное полиномиальное распределение, т. е.

$$P\{X_i=N_i;\ i=1, \ldots, N\}=N!\left(\frac{1}{N}\right)^N \prod_{i=1}^{n} \frac{1}{N_i!}.$$

Следовательно, проверка пуассоновости случайных величин $X_1, \ldots, X_N$ сводится к проверке полиномиальности их условного распределения при фиксированной сумме $K=\sum_{i=1}^{n} X_i$. Для проверки полиномиальности в случае, когда все вероятности одинаковы и равны $\frac{1}{N}$, а $K\cdot\frac{1}{N}$ невелико, в статье [2] рекомендуется использовать критерий (эквивалентный χ^2-критерию), основанный на статистике

$$L=X_1^2+ \ldots +X_N^2. \tag{3}$$

Если нулевая гипотеза верна и сумма K велика, то L имеет приближенно нормальное распределение, при этом

$$M\{L/K\}=K+\frac{K(K-1)}{N},$$

$$D\{L/K\}=2\frac{K(K-1)}{N}\left(1-\frac{1}{N}\right). \tag{4}$$

Следовательно, случайная величина

$$z=\frac{L-M\{L/K\}}{\sqrt{D\{L/K\}}} \tag{5}$$

имеет приближенно нормальное распределение с параметрами (0, 1). Л. Н. Большев указывает, что если альтернативы к „пуассоновости" таковы, что $DX<MX$ (соответственно $DX>MX$), то критической областью является $z<c_1$ (соответственно $z>c_2$); MX вычисляется по (2): здесь

$$MX=\frac{K}{N}; \tag{6}$$

дисперсия вычисляется так:

$$DX=\sum_{i=0}^{I-1}(i-MX)p_i, \tag{7}$$

в данном случае

$$DX=\frac{L}{N}-\left(\frac{K}{N}\right)^2.$$

Проверим „пуассоновость" распределения слогов в корнях слов литовского словаря по выборке в 700 слов; выборочное распределение приведено в табл. 1; r_k – число k-сложных корней в выборке. Вычисляем

$$K=\sum_{i=0}^{I=1} ir_i=247;\quad L=\sum_{i=0}^{I=1} i^2r_i=421,\quad MX=0{,}353;$$

$$DX=0{,}477;\qquad M\{L/K\}=86{,}80;\qquad D\{L/K\}=13{,}19;$$

$$z=25{,}35.$$

Как видим, гипотеза „пуассоновости" отвергается с высоким уровнем значимости. При этом в качестве альтернативы мы должны принять некоторое распределение, у которого $DX>MX$. Таким свойством обладает, например, одно из хорошо известных распределений дискретных случайных величин – отрицательное биномиальное распределение

$$P_i=P\{X=i\}=C_{i+m-1}^{m-1}pq^i,\quad q=1-p. \tag{8}$$

Допустим, что распределение слогов в корнях описывается отрицательным биномиальным распределением: по этой же выборке оценим параметры по методу моментов. Получаем:

$$\begin{cases} MX=m\dfrac{q}{p}, \\ DX=m\left(\dfrac{q}{p}\right)^2+m\dfrac{q}{p}; \end{cases}$$

поэтому

$$m\left(\frac{q}{p}\right)^2=DX-MX,\qquad \frac{q}{p}=\frac{DX-MX}{MX},$$

$$m=\frac{MX^2}{DX-MX}; \tag{9}$$

в нашем случае $m=1{,}005$.

Итак, при нашем допущении в качестве теоретической модели мы должны выбрать простейший случай отрицательного биноминального распределения $m=1$, т. е. просто геометрическое распределение

$$P_i=P\{X=i\}=pq^i. \tag{10}$$

Как известно, оценка параметра p в этом случае равна

$$\hat{p}=\frac{1}{MX}. \tag{11}$$

Для данного случая $\hat{p}=0{,}739$, $\hat{q}=0{,}261$. В табл. 1 приведены значения P_iN, где P_i вычислены по формуле (10), и значения P_i^*N, где P_i^* вычислены по закону Пуассона (1): в ней же приведены значения X^2, где

$$X^2=\sum_{i=1}^{I}\frac{(r_i-P_iN)^2}{P_iN}. \tag{12}$$

Таблица 1

i	r_i	P_iN	$\frac{(r_i-P_iN)^2}{P_iK}$	P_i^*N	$\frac{(r_i-P_i^*N)^2}{P_i^*N}$
1	525	518	0,095	493	2,077
2	116	135	2,674	173	18,780
3	48	34	5,765	30	10,800
4	9	9	0	4	12,250
5	2	2	0		
			$X^2=8{,}53$; $\nu=3$		$X^2=43{,}91$; $\nu=3$

Проведены аналогичные вычисления для английского языка (представителя аналитических языков). Выборочное распределение для этого языка находим у В. Фукса (см. [1]): он приводит частоты k-сложных слов, однако не указывает величину выборки, по которой они вычислены. Мы произвольно приняли $N=10\,000$. Выполнив вычисления, получаем $z=27{,}53$, т. е. опять гипотеза „пуассоновости" отвергается; при этом $m=1{,}05$, т. е. близко к единице. При сравнении эмпирических частот с теоретическими, вычисленными по геометрическому закону и закону Пуассона, получены значения X^2, соответственно равные 20,98 и 793,01.

Из рассмотрения этих примеров мы можем сделать следующие выводы: во-первых, распределение слогов в корнях значительно лучше описывается геометрическим законом, чем законом Пуассона, во-вторых, геометрический закон является довольно точной моделью образования корней слов из слогов.

С присоединением к корням аффиксов обычно возникают и новые слоги: число слогов, которое прибавляется к слогам корня при аффиксации, будем рассматривать как независимую случайную величину V, принимающую значения $m=0,1,\ldots,n$.

Попытаемся установить закон, которому подчиняется случайная величина V. Во-первых, обратим внимание на то, что наша целочисленная случайная

величина принимает конечное число значений (число слогов в аффиксах конечно и не превышает n). Самым распространенным конечнозначным целочисленным распределением является биномиальный закон. Поэтому естественно начать проверку именно с него. Кроме того, в случае биномиального закона легко представить себе, что присоединение аффиксов к корням происходит следующим образом. В ящике имеются карточки; некоторые из них пустые, на остальных написаны аффиксы. Пустые карточки и те, аффиксы которых не содержат слогообразующей гласной, снабдим пометкой „0"; остальные карточки снабдим пометкой „1". Будем случайным образом извлекать из ящика n карточек. Как известно, количество слогов,которое мы получим при этом, подчиняется биномиальному закону. Пусть вероятность „нулевого" слога равна β, а вероятность наличия слога равна $\alpha=1-\beta$. Тогда наш биномиальный закон принимает вид

$$P_m=P\{V=m\}=C_n^m\alpha^m\beta^{n-m}, \qquad m=0,\ 1,\ \ldots,\ n. \tag{13}$$

При этих предположениях образование слов из слогов будет происходить по закону, представляющему собой композицию геометрического и биномиального законов. Следовательно, случайная величина $Z=U+V$ при заданном n распределена по закону

$$P_k=P\{Z=k\}=\begin{cases}\displaystyle\sum_{m=0}^{n} C_n^m\alpha^m\beta^{n-m}pq^{k-m-1} \text{ для } k\leqslant n,\\ \displaystyle\Big(1-\sum_{k=1}^{n} P_k\Big)pq^{k-n-1} \text{ для } k>n.\end{cases} \tag{14}$$

Для оценки параметров p и α воспользуемся методом наименьших квадратов. В качестве нулевых приближений можно взять оценки параметров, вычисленные по методу моментов. Как следует из свойств геометрического закона, подбором параметра p подгоняется „хвост" теоретического распределения к эмпирическому; подгонка начала распределения регулируется параметром α. Поэтому параметр p оценивается так:

$$\hat{p}=\frac{1}{M_1}, \qquad (\hat{q}=1-\hat{p}), \tag{15}$$

где M_1 — выборочное среднее случайной величины $(X-n)$ при условии, что рассматриваются только те слова, длина которых больше n.

Как известно, среднее значение случайной величины $Z=U+V$ равно

$$MZ=M_1+n\alpha_1;$$

отсюда

$$\hat{\alpha}=\frac{MZ-M_1}{n}, \qquad \hat{\beta}=1-\hat{\alpha}. \tag{16}$$

Выбор n зависит от характера словесного материала. В литовском языке слов, в которых при аффиксации присоединяется более трех слогов, относительно немного. Так, в выборке словаря в 750 слов, частота слов, имеющих четыре и более аффиксальных слога, равна 0,055, в отрезке текста П. Цвирки,

состоящем из 1000 слов, таких слов 16. Поэтому естественно принять n равным 3 или 4.

Приступим к рассмотрению примеров. Проверим, как модель описывает распределение слогов в словах литовского словаря (исследование будем проводить по литовской части литовско-французского словаря). Односложные имеют частоту 0,004. Поскольку односложные занимают в языке особое положение (здесь это подробно не обсуждается), для их описания введем отдельный параметр. Здесь будем рассматривать сдвинутые распределения, т. е. распределения случайной величины $Z'=Z-1$, которая принимает значения $k=0, 1, \ldots, K-1$. Выборочное распределение случайной величины Z' приведено в табл. 2; $N=25036$. При $n=3$ имеем

$$\hat{\alpha}=0{,}4288, \qquad \hat{p}=0{,}7971;$$

$$\hat{\beta}=0{,}5712; \qquad \hat{q}=0{,}2029.$$

Теоретическое распределение, вычисленное по закону (14), приведено в табл. 2; значение $X^2=33{,}91$ при числе степеней свободы $\nu=3$. По критерию χ^2 нулевую гипотезу о том, что наша теоретическая модель в точности соответствует реальной модели, мы должны отвергнуть с высоким уровнем значимости. Но в нашей задаче (как и почти во всех практических задачах проверки гипотез)

Таблица 2

k	Выборочные данные	Теоретические данные
1	0,144	0,149
2	0,375	0,365
3	0,318	0,325
4	0,127	0,129
5	0,030	0,026
6	0,005	0,005
		$X^2=33{,}906, \quad \nu=3$

нас интересует не нулевая гипотеза о полном совпадении теоретической и реальной моделей, а гипотеза о некоторой их близости. Заметим, что как бы ни было мало „расстояние" между теоретической и реальной моделями, нулевая гипотеза будет отвергаться со сколь угодно большой вероятностью, если только число наблюдений достаточно большое. Мерой „расстояния" между моделями в нашей задаче служит величина

$$\alpha = N \sum_{k=1}^{K} \frac{(p_k - MP_k)^2}{MP_k},$$

где p_k — истинные вероятности, P_k — теоретические вероятности, вычисленные по (14). Дабы приемлемую модель не отвергнуть из-за большого числа наблюдений, Л. Н. Большев предлагает (см. [3]) вместо нулевой гипотезы

$H: \alpha=0$ проверять сложную гипотезу $H: \alpha \leqslant \alpha_0$ (число α_0 подбирается так, чтобы в рассматриваемой задаче значения $\alpha \leqslant \alpha_0$ были бы вполне приемлемыми). Для проверки гипотезы H критическая область выбирается по нецентральному χ^2-распределению (подробно см. в [3]).

Для нашей задачи получаем, что гипотеза $H: \alpha \leqslant 0{,}001N$ не отвергается при уровне значимости $P=0{,}05$. Таким образом, закон довольно точно описывает образование слов из слогов. Аналогичные вычисления проведены для немецкого и итальянского словарей, полученные данные подтверждают правильность гипотезы.

Поскольку предлагаемая модель отражает морфологическое строение языков и не требует сложных вычислений, она может найти применение в типологии языков.

Институт физики и математики
Академии наук Литовской ССР

Поступило в редакцию
10.II.1972

Литература

1. В. Фукс, Математическая теория словообразования, сб. „Теория передачи сообщений", М., ИЛ., 1957, 221–247.
2. Л. Н. Большев, О характеризации распределения Пуассона и ее статистических приложениях, Теор. вер. и ее прим., *X*, вып. 3 (1965), 488–499.
3. Л. Н. Большев, О построении доверительных пределов, Теор. вер. и ее прим., *X*, вып. 1 (1965), 187–192.

SKIEMENŲ PASISKIRSTYMO ŽODŽIUOSE IŠ ŽODYNŲ DĖSNIS

R. Merkytė

(*Reziumė*)

Darbe parodoma, kad skiemenų pasiskirstymą šaknyse galima aprašyti geometriniu dėsniu, o skiemenų pasiskirstymą žodžiuose – geometrinio ir binominio dėsnių kompozicija. Atlikti apskaičiavimai, imant lietuvių, anglų ir vokiečių kalbas, patvirtina, kad toks modelis atitinka empirinius duomenis.

A LAW DESCRIBING THE DISTRIBUTION OF SYLLABLES IN THE WORDS FROM DICTIONARIES

R. Merkytė

(*Summary*)

The paper shows that the distribution of syllables in the roots is described by a geometric law, while the distribution of syllables in the words by a composition of geometric and binomial laws. Calculations performed for the Lithuanian, Englisch and German languages confirm the agreement of the model to empirical data.