II ЛИТОВСКИЙ МАТЕМАТИЧЕСКИЙ СБОРНИК
LIETUVOS MATEMATIKOS RINKINYS 1

1962

НЕКОТОРЫЕ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ОБРАЗОВАНИЯ СЛОВ ИЗ СЛОГОВ И СЛОГОВ ИЗ БУКВ ДЛЯ ЛИТОВСКОГО ЯЗЫКА

Р. МЕРКИТЕ

В работе проводится исследование процесса образования слов из слогов. С этой целью рассматриваются распределения относительных частот слогов в словах для различных текстов нескольких литовских авторов. Проводится сравнение этих распределений с распределениями, полученными из текстов литовских писателей, творивших в разные периоды времени. Рассматривается распределение числа слогов в слове, характерное для всего литовского языка, и это распределение сравнивается с аналогичными распределениями других девяти языков. Проверяется выведенная Фуксом формула, описывающая с некоторым приближением распределение частот слогов в словах для восьми языков, для данных литовского языка. Во второй части работы проводится аналогичное исследование процесса образования слогов из букв.

* * *

В настоящее время многие вопросы лингвистики исследуются математическими методами. Одной из важных статистических характеристик того или иного уклада речи является распределение слов по числу входящих в них слогов. Советский учёный С. Г. Чебанов [1] в 1947 г. установил, что распределение слов по числу слогов в различных языках подчиняется некоторой общей закономерности, а немецкий математик В. Фукс [2] в конце 50-х годов создал математическую теорию образования лингвистических элементов из их компонентов. При создании своей теории Фукс все лингвистические элементы делит на два класса: 1) элементы, подчиняющиеся условию „хотя бы один компонент" (всякое слово состоит хотя бы из одного слога) и 2) элементы, не подчиняющиеся условию „хотя бы один компонент" (не во всяком слоге имеется согласная буква). В данной работе речь пойдет только об элементах первого класса. Фуксом выведена формула, описывающая с некоторым приближением распределение частот слогов в словах для восьми языков. Начато подробное изучение вопроса, является ли этот закон универсальным, описывающим основные свойства словообразования для всех языков, или имеются языки или даже группы языков, в которых слова образуются из слогов по правилам, совершенно отличным от выведенного уравнения. Данная работа имеет целью провести соответствующие обследования литовского языка, конкретней, исследуется

процесс образования слов из слогов и слогов из букв. Сначала исследуем распределение относительных частот слогов в словах для различных текстов литовского языка, найдем распределение числа слогов в слове для всего обследованного литовского текста и проверим, описывается ли это распределение выведенной Фуксом формулой.

Были обследованы тексты таких современных литовских авторов, как И. Балтушис, И. Матуленис и К. Мешкаускас, а также тексты писателей, творивших в недавнем прошлом: И. Жимантене-Жемайте (1845–1921) и И. Тумаса-Вайжгантаса (1869–1933).

Обратим внимание, что наряду с художественными текстами Балтушиса, Жемайте, Вайжгантаса обследовались также математический текст Матуленица и экономико-политический текст Мешкаускаса. Это сделано затем, чтобы подобранные тексты лучше отражали общий характер всего языка.

Тексты обследовались не полностью. Из каждого произведения взято 10–15 тыс. слов. Такое количество слов является достаточным для нахождения важнейших характеристик данного текста; с увеличением числа слов обследованного текста данного произведения найденные характеристики меняются незначительно. Чтобы обследованная часть текста представляла все произведение, текст брался не подряд, а небольшими отрезками в 100–200 слов с различных случайно подобранных страниц данного произведения (номера страниц брались из таблиц случайных чисел).

Различные авторы, пишущие на одном языке, обладают индивидуальными различиями в распределении числа слогов в словах, которое можно подсчитать следующим образом.

Будем рассматривать определенный текст. Число слогов в слове обозначим i, число слов рассматриваемого текста обозначим K. Если число i-сложных (состоящих из i слогов) слов обозначим z_i, то доля p_i i-сложных слов в тексте выразится так:

$$p_i = \frac{z_i}{K}, \qquad \sum_{i=1}^{K} p_i = 1.$$

Для текстов рассматриваемых пяти авторов получим следующие значения p_i, приведенные в табл. 1.

Таблица 1

	Балтушис	Жемайте	Вайжгантас	Мешкаускас	Матуленис
p_1	0,3186	0,2751	0,3219	0,1392	0,2275
p_2	0,3348	0,3635	0,3436	0,2913	0,2918
p_3	0,2192	0,2423	0,2022	0,3112	0,2750
p_4	0,1069	0,0965	0,1023	0,1635	0,1457
p_5	0,0182	0,0200	0,0252	0,0663	0,0455
p_6	0,0021	0,0024	0,0043	0,0212	0,0118
p_7	0,0002	0,0003	0,0005	0,0045	0,0027
p_8	–	–	–	0,0027	–
p_9	–	–	–	0,0001	–

Для более полной характеристики стиля высчитаны еще следующие характеристики, называемые характеристиками стиля первого порядка.

Среднее значение распределения

$$\bar{i} = \sum_i i p_i. \tag{1}$$

В нашем случае $\bar{i}$ есть среднее число слогов, приходящееся на одно слово.

Центральные моменты высших порядков

$$\mu_r = \sum_i (i - \bar{i})^r p_i \tag{2}$$

и среднее квадратическое уклонение

$$\sigma = \sqrt{\mu_2}, \tag{3}$$

которое в нашем случае представляет собой среднее уклонение числа слогов в слове от $\bar{i}$ – среднего числа слогов в слове.

Коэффициент асимметрии

$$\sum_k = \frac{\mu_3}{\sigma^3}, \tag{4}$$

который характеризует симметричность распределения относительно числа слогов в слове m, имеющего наибольшую вероятность (для большинства литовских авторов $m=2$). Если распределение симметрично, то $\Sigma_k = 0$. В нашем случае $\Sigma_k > 0$. Это означает, что распределение не симметрично и более „длинная" часть распределения лежит справа от m.

Коэффициент эксцесса

$$\varepsilon_k = \frac{\mu_4}{\sigma^4} - 3, \tag{5}$$

который служит для сравнения распределения случайной величины числа слогов в слове с так называемым нормальным распределением случайной величины. Для нормального распределения $\varepsilon_k = 0$; $\varepsilon_k > 0$ указывает на больший подъём графика распределения числа слогов в слове по сравнению с графиком нормального распределения, $\varepsilon_k < 0$ – на меньший подъём.

Энтропия

$$H = -\sum_i p_i \log p_i, \tag{6}$$

в нашем случае логарифм десятичный.

В табл. 2 и приведены характеристики стиля рассмотренных текстов пяти авторов.

Таблица 2

Автор	$\bar{i}$	σ	μ_3	Σ_k	μ_4	ε_k	H
Балтушис	2,1784	1,0658	0,7929	0,6549	3,6012	−0,2093	0,6037
Жемайте	2,2310	1,0361	0,7018	0,6310	2,2618	−1,0371	0,6022
Вайжгантас . . .	2,1802	1,0995	1,0574	0,7955	4,6166	0,1588	0,6118
Мешкаускас . . .	2,8221	1,2366	1,3569	0,6838	9,0938	0,6471	0,6931
Матуленис	2,5361	1,2098	1,0262	0,5795	6,4204	−1,0032	0,6692

Чтобы наглядней сравнить распределения частот числа слогов в слове для текстов этих авторов, изобразим эти распределения графически, откладывая на оси абсцисс i, а на оси ординат p_i (рис. 1).

Сравнивая полученные кривые между собой, замечаем, что из всех выделяется кривая, характеризующая распределение частот текстов Мешкаускаса, которая как бы выпадает из общей картины. Это объясняется тем,

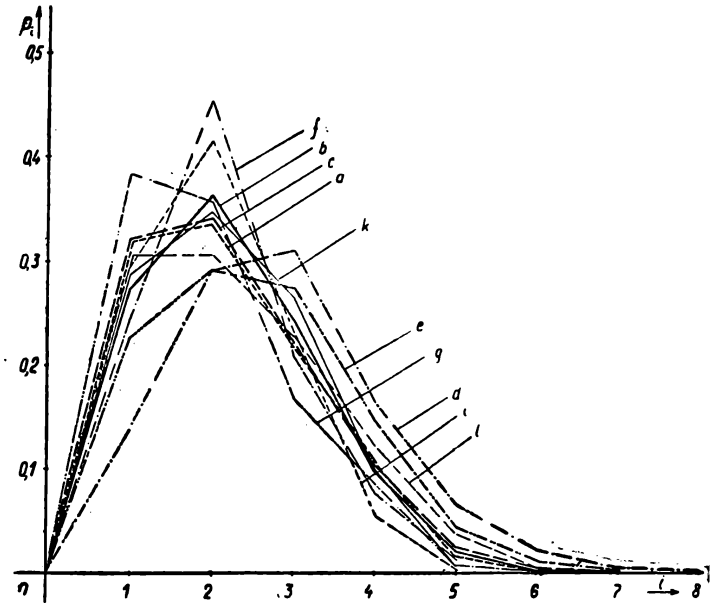

Рис. 1. Распределение относительных частот числа слогов в слове для текстов литовских авторов. Буквами $a-e$ помечены распределения соответственно табл. 2, буквами $f-l$ — соответственно табл. 4

что политический текст насыщен международными словами, которые в общем являются более длинными, чем литовские слова. Остальные кривые разнятся меньше. Это наводит на мысль, что распределение частот числа слогов в слове для текстов рассматриваемых авторов имеет какой-то общий характер, что обуславливается, очевидно, особенностями литовского языка.

Для наглядности на рис. 2 сравниваются кривая распределения частот слогов в слове в тексте „Parduotos vasaros" Балтушиса и „Отелло" Шекспира.

Исходя из проведенных обследований, Чебанов утверждает, что распределение числа слогов в слове мало зависит от происходящих с течением времени изменений языка. Наиболее наглядно это следует из обследования германских языков, в которых, по данным Чебанова, распределение числа слогов в словах мало изменилось на протяжении 1200 лет. Правда, эти

выводы Чебанов делает на основе обследования текстов разной длины: распределение слогов в слове, полученное из обследования текста в 20308 слов, сравнивается с распределением, полученным из обследования текста в 1513 слов. Порой выводы делаются из обследования очень коротких текстов — в 521 слово и даже в 183 слова.

Для подобного обследования литовского языка были выбраны следующие пять писателей: М. Мажвидас (им в 1547 г. написана первая литов-

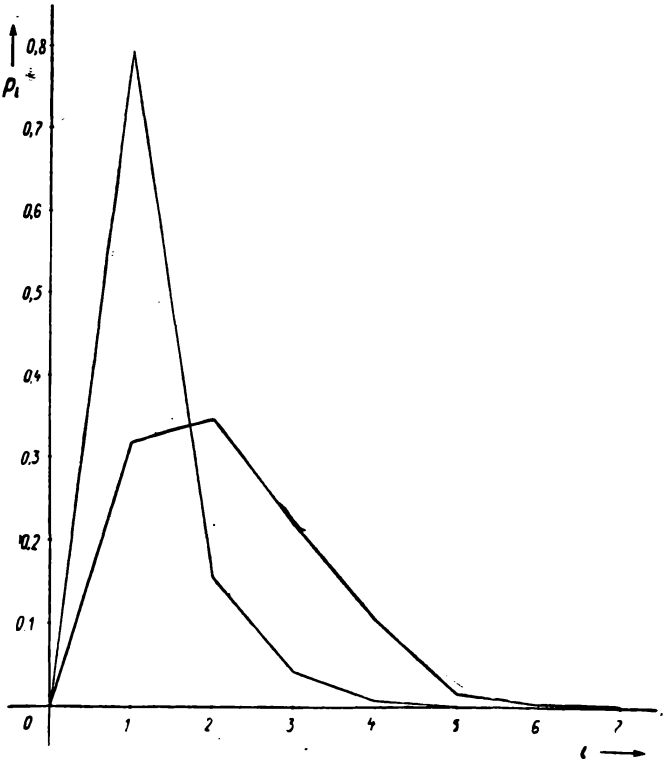

Рис. 2. Распределение относительных частот числа слогов в словах для текста Шекспира и Балтушиса (толстая линия)

ская книга) [8], К. Донелайтис (годы жизни 1714—1780) [9], Д. Пошка (годы жизни 1760—1831) [10], М. Валанчюс (обследованное произведение „Palangos Juzė" написано в 1866 г.) [11], В. Миколайтис-Путинас (обследованное произведение „Altorių šešėly" написано в 1933 г.) [12]. Тексты Валанчюса и Путинаса обследовались указанным ранее способом случайного подбора страниц, тексты остальных писателей обследовались, придерживаясь естественных делений текста — глав, границ произведений. Обследованный текст каждого автора содержит около 3000 слов. Такой объём выборки для нашей цели является достаточным, т. к. определенные по таким

выборкам важнейшие характеристики текста мало отличаются от характеристик, вычисленных по выборкам в 10 тыс. слов. Так, $\bar{i}$ для текстов Жемайте, Балтушиса и Вайжгантаса по выборке в 3 тыс. слов определяется с относительной ошибкой, не превышающей 0,007. Для наглядности в табл. 3 приводятся распределения относительных частот числа слогов в слове для ранее рассмотренных пяти авторов, полученные из текстов объёмом в 3

Таблица 3

	Балтушис	Жемайте	Вайжгантас	Матуленис
p_1	0,3137	0,2777	0,3283	0,2400
p_2	0,3347	0,3563	0,3420	0,2833
p_3	0,2240	0,2447	0,1990	0,2777
p_4	0,1053	0,0953	0,1027	0,1433
p_5	0,0183	0,0237	0,0230	0,0450
p_6	0,0033	0,0023	0,0043	0,0097
p_7	0,0007	–	0,0007	0,0010
p_8	–	–	–	–

тыс. слов. Данные этой таблицы не сильно расходятся с данными таблицы 1.

Полученные распределения и характеристики их стиля приведены в табл. 4 и 5, а соответствующие кривые даны для сравнения на рис. 1. Из

Таблица 4

	Мажвидас	Донелайтис	Пошка	Валанчюс	Миколайтис
p_1	0,2428	0,3823	0,2943	0,2830	0,3047
p_2	0,4572	0,3557	0,4153	0,3477	0,3023
p_3	0,2045	0,1687	0,2307	0,2627	0,2257
p_4	0,0793	0,0830	0,0574	0,0970	0,1203
p_5	0,0147	0,0097	0,0023	0,0093	0,0390
p_6	0,0013	0,0007	–	–	0,0073
p_7	0,0003	–	–	0,0003	0,0003
p_8	–	–	–	–	0,0003

Таблица 5

Автор	$\bar{i}$	σ	μ_3	Σ_k	μ_4	ε_k	H
Мажвидас	2,1713	0,9465	0,6482	0,7646	2,8243	0,6938	0,5647
Донелайтис . . .	1,9845	0,9906	0,8027	0,8258	2,8923	0,5022	0,5612
Пошка	2,0581	0,8786	0,3230	0,4762	1,5390	−0,4172	0,5391
Валанчюс	2,2031	0,9918	0,3366	0,3450	2,4387	−0,4796	0,5854
Миколайтис . . .	2,3109	1,1832	1,1692	0,7054	6,1962	0,1572	0,6435

результатов обследования нельзя сделать вывод о малой изменяемости распределений числа слогов в слове в связи с изменением языка.

Если нас интересуют общие законы, управляющие словообразованием, то индивидуальные различия одноязычных авторов для нас не представляют интереса. Их можно исключить, усредняя распределения частот для многих авторов, т. е. подсчитывая распределения

$$\bar{p}_i = \frac{1}{n} \sum_{\alpha=1}^{n} p_i^{(\alpha)}, \tag{7}$$

где $p_i^{(\alpha)}$ — распределение частот у автора, обозначенного индексом α, и где n — число различных авторов, тексты которых рассматриваются. Теоретически следует рассмотреть настолько большое число текстов n различных авторов, чтобы добавление других текстов не изменяло существенно среднего распределения $\bar{p}_i$. Такие усредненные распределения $\bar{p}_i$ подсчитаны для большого количества текстов многих авторов на девяти языках. То же проделано и для литовского языка. Значения $\bar{p}_i$ для десяти языков даны в табл. 6, здесь же даны значения $\bar{i}$ и H. Языки расположены в порядке возрастания $\bar{i}$, которое является простейшей числовой характеристикой языка.

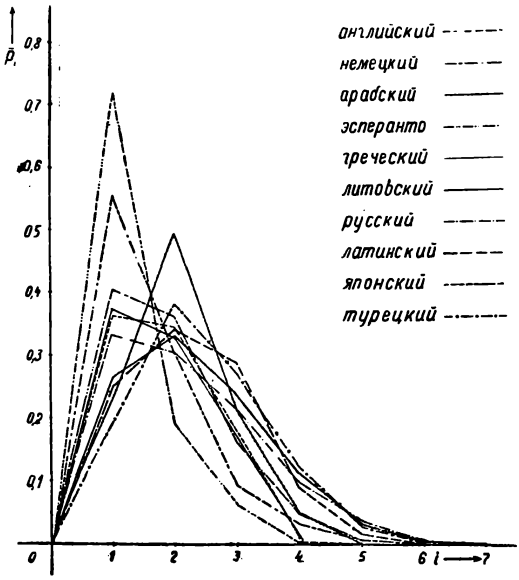

Рис. 3. Распределение относительных частот согласно табл. 6

Интересно отметить, какое большое влияние на усредненное распределение литовского языка оказывают тексты пяти старейших литовских авторов общим объемом всего в 15 тыс. слов. Так, если подсчитать среднее распределение $\bar{p}_i$ литовского языка только на основе более подробно изученных текстов Балтушиса, Жемайте, Вайжгантаса, Мешкаускаса и Матулиониса, то распределение $\bar{p}_i$ получается другим, а именно: $\bar{p}_1 = 0{,}2606$; $\bar{p}_2 = 0{,}3254$; $\bar{p}_3 = 0{,}2479$; $\bar{p}_4 = 0{,}1221$; $\bar{p}_5 = 0{,}0339$; $\bar{p}_6 = 0{,}0080$; $\bar{p}_7 = 0{,}0015$; $\bar{p}_8 = 0{,}0005$; $\bar{i} = 2{,}376$; $H = 0{,}645$, и литовский язык занимает другое место в ряду языков — между русским и латинским языками.

Чтобы наглядней показать общие закономерности усредненных распределений данных десяти языков, они изображены графически на рис. 3.

Таблица 6

	Английский (устный)	Немецкий	Эсперанто	Арабский	Греческий	Литовский	Японский	Русский	Латинский	Турецкий
$\bar{p}_1$	0,7152	0,5560	0,4040	0,2270	0,3760	0,2689	0,3620	0,3390	0,2420	0,1880
$\bar{p}_2$	0,1940	0,3080	0,3610	0,4970	0,3210	0,3354	0,3440	0,3030	0,3210	0,3784
$\bar{p}_3$	0,0680	0,0938	0,1770	0,2239	0,1680	0,2419	0,1780	0,2140	0,2870	0,2704
p_4	0,0160	0,0335	0,0476	0,0506	0,0889	0,1152	0,0868	0,0975	0,1168	0,1208
p_5	0,0056	0,0071	0,0082	0,0017	0,0346	0,0301	0,0232	0,0358	0,0282	0,0360
p_6	0,0012	0,0014	0,0011	–	0,0083	0,0067	0,0124	0,0101	0,0055	0,0056
p_7	–	0,0002	–	–	0,0007	0,0013	0,0040	0,0015	0,0007	0,0004
p_8	–	0,0001	–	–	–	0,0004	0,0004	0,0003	0,0002	0,0004
p_9	–	–	–	–	–	–	0,0004	–	–	–
$\bar{I}$	1,351	1,634	1,859	2,104	2,105	2,109	2,137	2,228	2,392	2,455
H	0,367	0,456	0,535	0,513	0,611	0,635	0,622	0,647	0,631	0,629

Девять кривых образуют правильную картину с явным переходом от устного английского языка к турецкому. Выделяется только кривая арабского языка. По мнению Фукса, это объясняется тем, что распределение для арабского языка было подсчитано только из двух текстов одинакового типа (сказки), которые не могут считаться типичной подборкой арабских текстов.

Подробнее исследовать природу процесса образования слов из слогов можно, используя статистический прибор, который изображен на рис. 4. Он

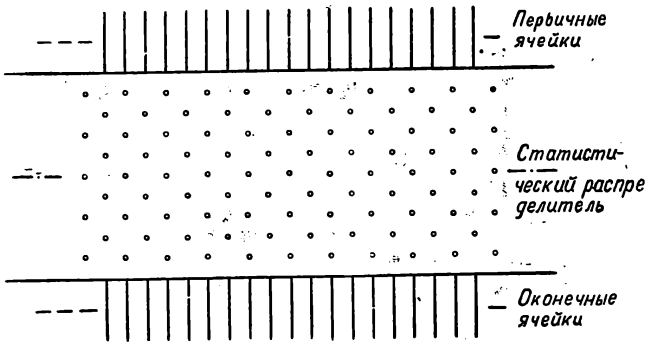

Рис. 4. Схема статистического прибора

представляет собой широкую доску, вращающуюся относительно горизонтальной оси, вдоль которой доска укреплена. По двум противоположным краям доски отверстиями к ее середине расположены ячейки. Статистический распределитель представляет собой несколько десятков рядов гвоздей, вбитых в доску.

Прибор используется следующим образом. Первичные ячейки заполняются стальными шариками в соответствии с функцией $f(x)$. Ячейки нумеруются числами 1, 2, 3, ..., x, ...,X; $f(x)$ — произвольная функция, ограничиваемая только конструктивными данными прибора. Затем доска наклоняется, и шарики из первичных ячеек направляются через статистический распределитель в оконечные ячейки. Математически этот процесс рассматривается как применение оператора U, преобразующего начальное распределение $f(x)$ в функцию $\varphi = Uf$, где $\varphi(y)$ есть частота шариков в оконечной ячейке y; $1 \leqslant y \leqslant Y$, где Y число оконечных ячеек.

Этот прибор применяется для решения различных проблем. Использовался он также и при исследовании распределения частот слогов в слове. Прежде всего, учитывая условие, что нет слов, содержащих менее одного слога, помещают по шарику в каждую оконечную ячейку перед началом процесса распределения. Затем, учитывая значение $\bar{i}$ определенного текста, размещают шарики в первичных ячейках. Предположим, исследование проводится для текста „Югуртинской войны" Саллюстия, в котором в среднем приходится 2,5 слога на слово. Тогда помещают один шарик в первую

первичную ячейку, два шарика — во вторую первичную ячейку, один шарик — в третью первичную ячейку, два шарика — в четвертую и т. д. попеременно. После этого пропускают шарики, помещенные в первичные ячейки, через статистический распределитель. В оконечных ячейках получится определенное распределение: z_1 ячеек будут содержать по одному шарику, z_2 ячеек — по два шарика и т. д.; в общем x_i ячеек будут содержать по i шариков.

Повторив этот процесс достаточно большое число раз, берут средние значения распределений относительных частот шариков в оконечных ячейках и по ним чертят кривую, как это делалось раньше для значений p_i. Кривая, полученная для упомянутого текста при помощи статистического прибора со 160 первичными ячейками, 160 оконечными ячейками и 30 рядами гвоздей в распределителе с учётом особенностей прибора, связанных с его конечными размерами, оказалась очень близкой кривой, полученной непосредственной обработкой текста. Это показывает, что механический прибор даже при его конечных размерах передает общую картину довольно хорошо. Аналитически определяя механические процессы, происходящие в приборе, Фукс вывел математическую формулу, описывающую процесс образования слов.

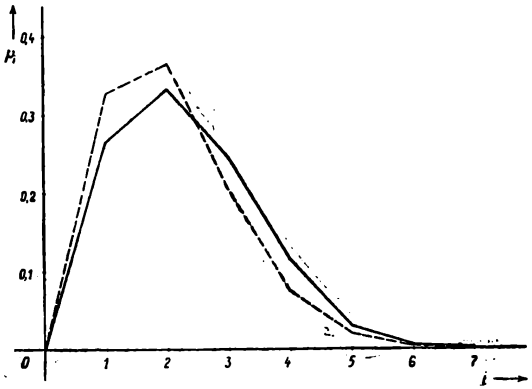

Рис. 5. Сравнение распределений слогов в слове, полученных из текста (сплошная линия) и по формуле для литовского языка

Так, предполагая, что в ячейки помещается предварительно по одному шарику (условие „хотя бы один элемент"), а потом по этим ячейкам распределяется большое число таких же шариков, Фукс получил следующую формулу, описывающую процесс образования слов.

$$p_i = \frac{e^{-(\bar{i}^*-1)}(\bar{i}^*-1)^{i-1}}{(i-1)!}, \qquad \bar{i} > 1, \tag{8}$$

т. е. закон Пауссона

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Если брать значениями $\bar{i}^*$ найденные путем подсчёта для различных языков значения $\bar{i}$, скажем, взять $\bar{i}=1{,}3$ (у. англ.); 1,6 (нем. яз.); 1,9 (эсперанто); ... 2,5 (турецк.), то получим распределение, близкое к распределению обследованных языков. Для литовского языка* это показано на рис. 5.

* Взятое значение $\bar{i}$ подсчитано из текстов десяти литовских авторов.

Это показывает, что распределение (8) описывает по крайней мере основные свойства процесса образования слов из слогов как для исследованных Фуксом живых и мертвых девяти языков, так и для литовского языка сравнительно точно.

Наконец, подсчитаем энтропию H для распределения (8) в зависимости от $\bar{i}$ по формуле

$$H = -\sum_{i=0}^{\infty} \frac{e^{-(\bar{i}-1)}(\bar{i}-1)^{i-1}}{(i-1)!} \log \frac{e^{-(\bar{i}-1)}(\bar{i}-1)^{i-1}}{(i-1)!}. \qquad (9)$$

Эта зависимость графически представлена сплошной линией на рис. 6. Сюда также нанесены точки, изображающие зависимость энтропии H от среднего значения $\bar{i}$ для десяти обследованных языков.

Полученные точки не разбросаны случайно по всей площади диаграммы, а явно тяготеют к теоретической кривой (исключение опять составляет арабский язык по уже указанной причине). Это еще раз подтверждает, что процесс образования слов из слогов достаточно хорошо описывается уравнением (8).

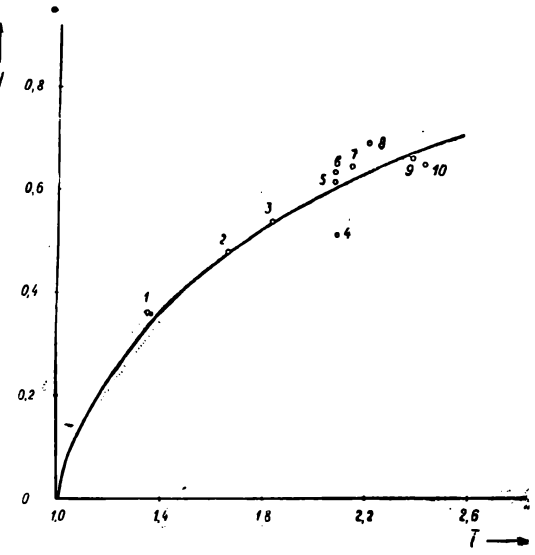

Рис. 6. Зависимость энтропии H от средних значений $\bar{i}$ для десяти языков. Сплошная кривая—теоретическая. Точки пронумерованы согласно табл. 6

* * *

Существуют другие лингвистические проблемы, в которых выполняются весьма близкие условия. Каждый слог содержит по крайней мере одну букву или один звук; более того, каждый слог в собственном смысле слова содержит по крайней мере один гласный; каждое предложение содержит по крайней мере одно слово и т. д. Можно думать, что теория образования всех этих элементов будет выражаться той же формулой, что и теория образования слов из слогов. В действительности же теория образования всех этих элементов из компонентов много сложнее и теория образования слов из слогов является лишь особо простым вариантом общей теории.

Для литовского языка более подробно исследовано образование слогов из букв. Обрабатывалась часть ранее указанного материала объемом примерно 10—15 тыс. слогов для каждого автора. Ввиду сходности данных обрабатывался текст лишь четырех авторов.

Распределение букв в слогах имеет следующий вид:

Таблица 7

	Балтушис	Жемайте	Мешкаускас	Матуленис	Общее для лит. языка
p_1	0,0278	0,0265	0,0393	0,0299	0,0316
p_2	0,5271	0,5567	0,5245	0,5616	0,5416
p_3	0,3481	0,3245	0,3248	0,2971	0,3227
p_4	0,0819	0,0812	0,1015	0,0919	0,0903
p_5	0,0136	0,0105	0,0096	0,0190	0,0131
p_6	0,0013	0,0005	0,0003	0,0003	0,0005
p_7	0,0002	0,0001	–	0,0002	0,0001

Характеристики этого распределения даны в таблице 8, а его графическое изображение – на рис. 7.

Таблица 8

Автор	$\bar{i}$	σ	μ_3	Σ_k	μ_4	ε_k	H
Балтушис	2,5311	0,7547	0,3935	0,9153	1,3661	1,2102	0,4683
Жемайте	2,4944	0,7324	0,3592	0,9144	1,1344	0,9433	0,4534
Мешкаускас . . .	2,5185	0,7697	0,3149	0,6906	1,1614	0,3087	0,4822
Матуленис	2,5102	0,7843	0,4653	0,9646	1,5193	1,0157	0,4764

На рис. 8 дано общее распределение для четырех авторов, расчитанное по формуле (8), путем подставления найденного из текстов значения $\bar{i}$, и распределение, полученное обработкой текстов.

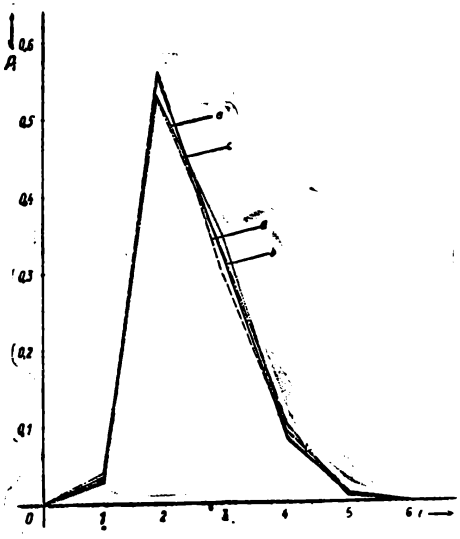

Рис. 7. Распределение относительных частот числа букв для текстов четырех литовских авторов. Балтушис — *a*, Жемайте — *b*, Мешкаускас — *c*, Матуленис — *d*

Явное несоответствие кривых показывает, что процесс образования слогов из букв нельзя представить этой формулой.

Формула, хорошо описывающая этот процесс, выведена Фуксом в предположении, что перед началом вероятностного распределения шариков по ячейкам в статистическом приборе в определенной части ячеек уже помещено определенное число шариков в следующем порядке. Часть $\beta_0 - \beta_1$ ячеек не содержит шариков, $\beta_1 - \beta_2$ ячеек содержат по одному

шарику и вообще $\beta_\nu - \beta_{\nu+1}$ ячеек содержат по ν шариков, причем $\beta_{\nu+1} \leqslant \beta_\nu$. Таким образом вводится последовательность числовых характеристик β, которая называется β-спектром лингвистического элемента.

Полученная формула имеет следующий вид:

$$V(i) = e^{-\left(\bar{i} - \sum_1^\infty \beta_\alpha\right)} \sum_{\nu=0}^{\infty} (\beta_\nu - \beta_{\nu+1}) \frac{\left(\bar{i} - \sum_1^\infty \beta_\alpha\right)^{i-\nu}}{(i-\nu)!}. \qquad (10)$$

Если положить $\beta_0 = 1$, $\beta_k = 0$, $k = 1, 2, \ldots$, то формула (10) превращается в формулу Пуассона.

Легко показать, что формула (8) является лишь частным случаем формулы (10).

Очевидно,

$$\sum_{\nu=0}^{\infty} \beta_\nu - \beta_{\nu+1} = 1. \qquad (11)$$

Поскольку мы рассматриваем образование таких элементов, которые подчинены условию „хотя бы один элемент", то

$$\beta_0 - \beta_1 = 0$$

и, следовательно,

$$\beta_0 = 1,$$
$$\beta_1 = 1.$$

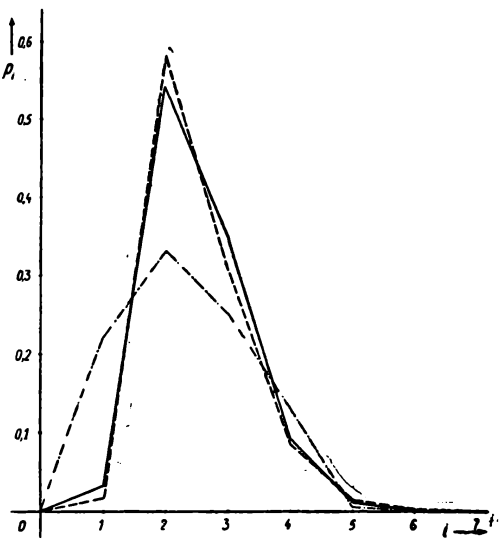

Рис. 8. Сравнение распределений букв в слогах, полученного из текста (сплошная кривая), по формуле (8) (штрихпунктирная) и по формуле (10) (пунктирная)

Если теперь положить $\beta_2 = \beta_3 = \ldots = 0$, то формула (10) переходит в в формулу (8).

При значениях $\beta_0 = \beta_1 = 1$, $\beta_2 \neq 0$ формула (10) хорошо описывает процесс образования слогов из звуков для различных языков*. β-спектр для каждого языка различен. Так, образование слов из звуков описывается формулой (10), если положить:

для латинского языка

$$\beta_1 = 1, \quad \beta_2 = 0{,}885, \quad \beta_3 = \beta_4 = \ldots = 0;$$

* Решение задачи о β-спектре подробно изложено в статье Фукса. β_2 может быть найдено из уравнения

$$\mu_2 = \bar{i} - \beta^2 + 2 \sum_{k=1}^{\infty} (k-1)\beta_k,$$

где $\beta = \sum_{k=1}^{\infty} \beta_k$. Значения μ_2 и $\bar{i}$ находятся из текста.

для греческого языка

$$\beta_1 = 1, \quad \beta_2 = 0{,}826, \quad \beta_3 = \beta_4 = \ldots = 0;$$

для английского языка

$$\beta_1 = 1, \quad \beta_2 = 0{,}934, \quad \beta_3 = \beta_4 = \ldots = 0.$$

Образование слогов из букв для литовского языка описывает формулой (10), если положить

$$\beta_1 = 1, \quad \beta_2 = 0{,}966, \quad \beta_3 = \beta_4 = \ldots = 0.$$

Тогда формула (10) принимает следующий вид:

$$v(i) = e^{-(\bar{i}-1-\beta_2)}\left[(1-\beta_2)\frac{(i-1-\beta_2)^{i-1}}{(i-1)!} + \beta_2 \frac{(i-1-\beta_2)^{i-2}}{(i-2)!}\right]. \qquad (12)$$

Распределение (12) при указанных значениях β и найденном из текстов значении $\bar{i}$ изображено на рис. 8.

Хорошее соответствие наблюдаемой и расчетной кривых дает основание считать, что процесс образования слогов из букв в литовском языке описывается формулой (10) по крайней мере удовлетворительно.

Таким образом, данное исследование показывает, что словообразование и слогообразование как в девяти исследованных языках, так и в литовском языке хорошо описывается приведенными ранее формулами.

Подобным образом можно исследовать ряд других процессов. Отметим, что рассмотренные процессы являются лишь случаем общей проблемы образования лингвистических элементов из их компонентов.

В заключение пользуюсь случаем выразить сердечную благодарность И. П. Кубилюсу и В. А. Статулявичюсу за оказанную помощь при выборе темы и ценные указания.

Институт физики и математики
Академии наук Литовской ССР

Поступила в редакцию
29. XII. 1961

ЛИТЕРАТУРА

1. Чебанов С. Г. О подчинении речевых укладов „индоевропейской" группы закону Пуассона.
Доклады АН СССР, новая сер., 55, № 2, 103—106, 1947.
2. Фукс В. Математическая теория словообразования. Сборник статей. „Теория передачи сообщений", 1957.
3. Baltušis J. Parduotos vasaros, I d., Vilnius, 1958.
4. Matulionis J. Aukštoji matematika, I d., Vilnius, 1959.
5. Meškauskas K. Tarybų Lietuvos industrializavimas, Vilnius, 1960.
6. Žymantienė-Žemaitė J. Petras Kurmelis — Gera galva — Mieste, Kaunas, 1946.
7. Tumas-Vaižgantas J. Dėdės ir dėdienės, „Rinktiniai raštai", 1, Vilnius, 1957.
8. Mažvydas M. Catechismusa prasty szadei... Kaunas, 1947.
9. Donelaitis K. Metai, Kaunas, 1940.
10. Poška D. Raštai, Vilnius, 1959.
11. Valančius M. Palangos Juzė, Kaunas, 1947.
12. Mykolaitis-Putinas V. Altorių šešėly, Vilnius, 1954.

KAI KURIOS ŽODŽIŲ IŠ SKIEMENŲ IR SKIEMENŲ IŠ RAIDŽIŲ SUDARYMO STATISTINĖS CHARAKTERISTIKOS LIETUVIŲ KALBAI

R. MERKYTĖ

(Reziumė)

Darbe nagrinėjamas žodžių iš skiemenų sudarymo procesas. Tuo tikslu gaunami skiemenų žodžiuose pasiskirstymo dažnumai įvairiems lietuvių autorių tekstams, apskaičiuojamos kai kurios statistinės charakteristikos, kurios palyginamos su kitų kalbų charakteristikomis. Lietuvių kalbai patikrinamos Fukso formulė ir Čebanovo teorija. Analogiškai nagrinėjamas skiemenų iš raidžių sudarymas.

EINIGE STATISTISCHE CHARAKTERISTIKEN ZUR BILDUNG VON WÖRTERN AUS SILBEN UND SILBEN AUS BUCHSTABEN IN DER LITAUISCHEN SPRACHE

R. MERKYTE

(Zusammenfassung)

In dieser Arbeit wird der Prozess der Wortbildung aus Silben behandelt. Zu diesem Zweck wird die Häufigkeitsverteilung der Silben in Wörtern verschiedener Texte litauischer Autoren untersucht, es werden einige statistische Charakteristiken der litauischen Sprache berechnet und mit denen der anderen Sprachen verglichen. Die Formel von Fuchs und die Theore Tscheba nows wird in Bezug auf die litauische Sprache geprüft. Analogen Untersuchungen wurde die Silbenbildung aus Buchstaben unterzogen.