УДК 519.22

О ВЗАИМОСВЯЗИ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ СО СПЕКТРОВЫМИ

В. В. Нешитой

ВВЕДЕНИЕ

Пусть имеется полная группа несовместных событий $A_1, A_2, \ldots, A_k, \ldots, A_n$, вероятности которых заданы и, соответственно, равны $p_1, p_2, \ldots, p_k, \ldots, p_n$. Пусть далее производится x независимых испытаний, в каждом из которых может наступить любое из n разных событий, составляющих полную группу. Обозначим через $y (0<y\leqslant n)$ количество всех разных событий, наступивших хотя бы один раз при x независимых испытаниях, а через y_m — количество разных событий, наступивших ровно m раз при x независимых испытаниях ($m=1, 2, \ldots$). При этом должны выполняться равенства

$$\sum_{m\geqslant 1} y_m = y, \quad \sum_{m\geqslant 0} y_m = n, \quad \sum_{m\geqslant 1} m y_m = x. \qquad (1)$$

Для выявления статистических закономерностей и решения практических задач эти данные должны быть каким-то образом упорядочены (ранжированы). Например, все наступившие события могут быть упорядочены по убыванию (невозрастанию) частоты их появления в данной выборке (т. е. при x испытаниях). Если упорядочить таким образом все y разных событий (например, y разных слов, которые встретились в выборке объемом x словоупотреблений) и каждому событию (слову) приписать порядковый номер (ранг) r ($r=1,2,\ldots,y$), а также частоту m_r, которая изменяется от наибольшего значения до единицы, то мы получим так называемый частотный список (словарь). Таблица же значений r, m_r задает известное *ранговое распределение.*

По данным рангового распределения легко составить таблицу значений m, y_m, где m — частота, y_m — количество разных событий с данной частотой. Упорядочив значения y_m по возрастанию частоты m, получим так называемое *спектровое распределение,* которое всегда является дискретным.

Из сказанного ясно, что *одни и те же статистические данные могут быть представлены в виде двух распределений — рангового и спектрового.* Если в частотном списке (т. е. в случае рангового распределения) на первом месте стоит событие с наибольшей частотой, а в конце — событие с частотой $m=1$, то в спектровом распределении на первом месте указывается количество разных событий, имеющих одинаковую частоту $m=1$ (в некоторых случаях $m=0$), а на последнем месте — количество разных событий (как правило, не более одного) с наибольшей частотой.

Для нахождения выравнивающего распределения достаточно знать лишь количественные характеристики статистического распределения, без перечисления самих событий, например, слов. В таком случае статистические данные можно свести в таблицу, задающую одновременно ранговое и спектровое распределения, правда, последнее будет ранжировано по убыванию частоты m, т. е. в обратном порядке:

Табличная форма задания рангового и спектрального распределений

Интервал рангов (с r... по r)	Частота m (по убыванию)	Количество событий с данной частотой y_m	Количество испытаний my_m
1	2	3	4

В первом столбце указываются не ранги, а интервалы рангов событий, имеющих одну и ту же частоту m. Первые два столбца задают ранговое распределение; второй и третий столбцы задают спектровое распределение *(статистическую структуру выборки),* которое следует читать снизу вверх, так как в спектровых распределениях, заданных, например, системой дискретных распределений [1], величина m возрастает. Данные третьего и четвертого столбцов должны удовлетворять равенствам (1).

Таким образом, ранговое и спектровое распределения описывают разные стороны одного и того же процесса, причем из одного распределения может быть получено другое. Оба распределения дополняют друг друга и позволяют более полно исследовать и описать интересующее нас явление.

В настоящей статье ставятся и решаются следующие задачи:

— исследуется взаимосвязь ранговых распределений со спектровыми;

— строится система спектровых распределений на основе системы непрерывных, в том числе ранговых, распределений;

— отыскивается закон распределения вероятностей разных слов, заданный непрерывной плотностью распределения;

— исследуется статистическая структура выборок различного объема.

1. СИСТЕМА СПЕКТРОВЫХ РАСПРЕДЕЛЕНИЙ

Пусть, по-прежнему, производятся независимые испытания, в каждом из которых может наступить любое из n разных событий, составляющих полную группу, причем вероятности всех n событий заданы. Найдем вероятность того, что какое-нибудь k-е событие при x испытаниях появится ровно m раз. Как известно, эта вероятность определяется по формуле Бернулли

$$P_k(m, x) = C_x^m p_k^m (1-p_k)^{x-m}, \qquad (2)$$

где

$$C_x^m = \frac{x!}{m!\,(x-m)!}. \qquad (3)$$

В случае полной группы несовместных событий вероятность $P_k(m, x)$ совпадает с математическим ожиданием числа появлений отдельного k-го события ровно m раз при x испытаниях. Поэтому математическое ожидание числа всех тех разных событий, которые появятся ровно m раз при x испытаниях, будет равно сумме

$$M[y_{m,x}] = \sum_{k=1}^{n} P_k(m, x) = C_x^m \sum_{k=1}^{n} p_k^m (1-p_k)^{x-m}. \qquad (4)$$

Формула (4) позволяет устанавливать статистическую структуру выборки по известному закону распределения вероятностей разных событий, составляющих полную группу.

Используя приближенные равенства, преобразуем формулу (4) к виду, более удобному для практических расчетов. Так, при больших x и малых m вместо выражения (3) можем записать

$$C_x^m \approx \frac{x^m}{m!}. \qquad (5)$$

Например, при $x=100$, $m=2$ формула (5) дает $C_x^m \approx \approx 4950$ вместо точного значения 5000.

В лингвистических исследованиях объемы выборок (текстов), как правило, составляют десятки и сотни тысяч, поэтому переход к приближенным формулам в этих условиях необходим.

Далее, если количество разных событий n велико и вероятности отдельных событий p_k малы, то справедливо приближенное равенство

$$(1-p_k)^{x-m} \approx e^{-xp_k}. \qquad (6)$$

С учетом равенств (5) и (6) формула (4) примет вид

$$M[y_{m,x}] \approx \frac{1}{m!} \sum_{k=1}^{n} \frac{(xp_k)^m}{e^{xp_k}}. \qquad (7)$$

Установим с помощью формулы (7) статистическую структуру выборки объемом x в случае, когда события, составляющие полную группу, имеют равные вероятности $p_k = 1/n = \alpha$. Из формулы (7) находим

$$M[y_{m,x}] = \frac{n}{m!} \cdot \frac{(\alpha x)^m}{e^{\alpha x}}, \qquad (8)$$

где αx—средняя частота появления разных событий при x испытаниях. Если разделить обе части равенства (8) на n, то получим долю событий с частотой появления, равной m раз при x испытаниях или, другими словами, вероятность $p_{m,x}$

$$\frac{M[y_{m,x}]}{n} = p_{m,x} = \frac{(\alpha x)^m}{m! e^{\alpha x}}. \qquad (9)$$

Выражение (9) представляет собой известный закон Пуассона.

Здесь следует подчеркнуть, что полученные *формулы (7)—(9) оказываются справедливыми не только в случае n разных несовместных событий, составляющих полную группу, но и при рассмотрении наступлений одного и того же события в n разных подвыборках.*

Если закон распределения вероятностей разных событий задан непрерывной плотностью $p(t)$, то по аналогии с формулой (7) можем записать

$$M[y_{m,x}] = \frac{1}{m!} \int_0^n \frac{[xp(t)]^m}{e^{xp(t)}}\, dt. \qquad (10)$$

Формулы (7) и (10) задают систему дискретных (спектровых) распределений.

Для описания распределений непрерывных случайных величин автором построены системы непрерывных распределений. В случае неотрицательных случайных величин можно использовать обобщенные распределения, заданные плотностями (2):

$$p(t) = Nt^{\gamma-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}, \qquad (11)$$

$$P(Y) = \frac{N(\ln Y)^{\gamma-1}}{Y}[1-\alpha u(\ln Y)^{\beta}]^{\frac{1}{u}-1}. \qquad (12)$$

В зависимости от значений параметров u, α (при $\beta, \gamma > 0$) распределения (11), (12) разделяются на пять типов. К I типу относятся распределения с параметрами $u>0$, $\alpha>0$; ко II типу — $u\to 0$, $\alpha>0$; к III типу — $-\infty < u < 0$, $\alpha>0$; к IV типу — $u\to-\infty$, $\alpha\to 0$ ($\alpha u \neq 0$); к V типу — $u>1$, $\alpha<0$.

Поскольку непрерывная плотность распределения вероятностей разных событий $p(t)$ (или $p(Y)$) входит в формулу (10), задающую систему дискретных распределений, распространим принятую классификацию на дискретные распределения.

Отметим, что в случае равномерной плотности $p(t)=\alpha$, которая следует из плотности (11) при $\beta=\gamma=u=1$ (тип I), обобщенное дискретное распределение (10) дает тот же закон Пуассона (8), который и следует отнести к I типу дискретных распределений. В общем же случае интеграл (10) не выражается конечным числом элементарных функций. Однако это не является помехой для восстановления спектрового распределения по заданной плотности (11) и (12). Необходимые расчеты легко осуществляются численным интегрированием выражения (10) на программируемой микро-ЭВМ типа МК-56 или БЗ-34.

Таким образом, *чтобы восстановить статистическую структуру выборки объемом x, необходимо и достаточно знать закон распределения вероятностей разных событий, составляющих полную группу,* заданный, например, непрерывной плотностью (11) или (12). Порядок отыскания этого закона распределения рассмотрим на примере «Частотного словаря русского языка» [3].

2. УСТАНОВЛЕНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ РАЗНЫХ СЛОВ

Для подбора выравнивающей кривой распределения, заданной плотностью (11), ее целесообразно привести к форме

$$p(x) = Ne^{\gamma x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}, \qquad (13)$$

где $p(x)=tp(t)$; $x=\ln t$. Величина t соответствует рангу статистического распределения.

Если выравнивающее распределение задано плотностью (12), ее также можно привести к форме (13), но тогда приведенное к этой же форме статистическое распределение будет иметь сильную левостороннюю асимметрию. Чтобы получить распределение, близкое к симметричному, приведем плотность (12) к форме (11), для чего представим ее в виде

$$Yp(Y) = N(\ln Y)^{\gamma-1}[1-\alpha u(\ln Y)^{\beta}]^{\frac{1}{u}-1}, \qquad (14)$$

где $Yp(Y)=p(t)$; $\ln Y = t$; $Y = y+1$ (величина y соответствует рангу r статистического распределения).

Предполагая, что выравнивающее распределение задается плотностью (11), построим по Таблице распределения частот, приведенной в [3, с. 895—915] (объем выборки $x=1\,056\,382$ словоупотребления, объем словаря $y=39268$ разных слов), график зависимости

$$rp_r = f(\ln r), \tag{15}$$

т. е. приведем ранговое распределение к форме (13). Путем такого же преобразования приводится к форме (13) и теоретическое распределение (11): $tp(t)=f(\ln t)$.

При построении такого графика для выявления особенностей статистического распределения и достижения наибольшей наглядности и точности графического изображения необходимо соблюдать следующие правила:

— начальный участок графика (для слов с рангами $1\leqslant r<50$) строится в соответствии с формулой $(r-0{,}5)p_r=f[\ln(r-0{,}5)]$, которая учитывает дискретность статистического распределения слов по частоте их употребления в выборке и непрерывность выравнивающего распределения (здесь считается, что относительные частоты «сосредоточены» в серединах интервалов 0—1, 1—2, ..., $(r-1)$—r);

— конечный участок графика (для слов с частотами $1\leqslant m<50$) строится в соответствии с формулой

$$r\frac{m_r+m_{r+1}}{2x}=f(\ln r), \tag{16}$$

где частоты двух соседних слов с рангами r и $r+1$, как правило, различаются на единицу, при этом учитывается средняя их частота $\bar{m}=(m_r+m_{r+1})/2$. Из формулы (16) следует, что последняя справа точка по построению имеет ординату

$$rp_r=\frac{y}{2x}, \tag{17}$$

поскольку в этом случае $r=y$, $m_r=1$, $m_{r+1}=0$;

— средняя часть графика строится в соответствии с формулой (15);

— для рангов $r=1\div10$ строятся все точки подряд; в остальных случаях точки строятся по рангам, возрастающим в геометрической прогрессии со знаменателем $1{,}25\div1{,}5$;

— полученные точки соединяются отрезками прямых.

Статистическая кривая распределения, построенная по правилам, рассмотренным выше (рис. 1), имеет четыре ярко выраженные вершины, что свидетельствует о неоднородности лексического состава частотного словаря русского языка. Так и должно быть, поскольку в частотном словаре представлены как полнозначные, так и служебные слова. Построенный график позволяет выделить неоднородную часть: последняя впадина

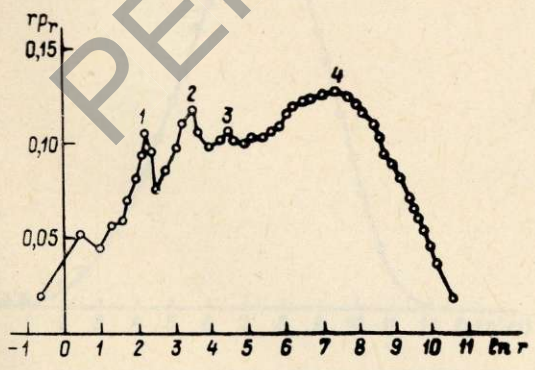

Рис. 1. Ранговое распределение слов по данным «Частотного словаря русского языка»

(между вершинами 3 и 4) имеет абсциссу $\ln r\approx4{,}6\div4{,}9$, что соответствует рангам $r\approx100\div135$. Таким образом, около 100 первых слов частотного словаря представляют собой неоднородную часть (в основном это служебные слова). Всю остальную лексику можно считать однородной.

Поскольку *выравнивающее распределение может описывать только одновершинные кривые распределения,* удалим из частотного словаря первые $r_0=100$ слов, а с ними — лишние вершины на рис. 1. Накопленные абсолютная и относительная частоты этих слов, соответственно, равны: $x_0=436940$, $F(r_0)=x_0/x=0{,}413619$. Тогда объем выборки x', приходящийся на оставшиеся $y'=y-100=39168$ слов, будет равен $x'=x-x_0=x[1-F(r_0)]=619442$ словоупотреблениям. Присвоим далее словам с рангами $r>100$ новые ранги $r'=r-r_0$. Если теперь снова построить график зависимости $r'p_{r'}=f(\ln r')$, где $p_{r'}=m_{r'}/x'$, то на этот раз получим одновершинную кривую распределения, что будет свидетельствовать об однородности лексического состава частотного словаря (без первых 100 слов). Хотя в усеченном словаре еще имеются служебные слова, особенно в его начале, но они рассеяны по словарю и поэтому заметно не искажают закономерной формы статистической кривой распределения, что весьма важно для отыскания закона распределения вероятностей разных слов. Такую кривую можно попытаться описать непрерывной плотностью (11). Однако расчеты показывают, что точка с координатами (B_1*, H_3*) (порядок вычисления этих статистических показателей изложен в работе [2]) ложится на белое поле рис. 1, приведенного там же. Следовательно, плотность (11) не может быть использована в нашем примере для описания рангового закона распределения слов.

Примем в качестве выравнивающего распределение, заданное плотностью (12). Приведем статистическое распределение к форме (11), построив график зависимости

$$(r'+1)\,p_{r'}=f[\ln(r'+1)]. \tag{18}$$

Теоретическое распределение (12) приводится к той же форме аналогичным преобразованием $Yp(Y)=f(\ln Y)$ (см. формулу (14)), где $Y=y+1$, а величина y соответствует рангу r' статистического распределения.

Прибавление единицы к рангу r' (чего требует плотность (14)) несколько меняет форму начала кривой распределения, при этом распределение перемещается в начало координат. При достаточно больших рангах ($r'>100$) данный график не отличается от распределения (15).

Для построения графика зависимости (18) по данным частотного словаря составим табл. 1 (столбцы 1—4).

Построенная в соответствии с правилами, изложенными выше, статистическая кривая распределения изображена на рис. 2. Ордината крайней справа точки в соответствии с формулой (17) равна 0,0316 и составляет 0,154 от наибольшей высоты кривой распределения ($0{,}0316/0{,}2049=0{,}154$). Чем меньше эта величина, тем надежнее может быть установлен закон распределения и точнее оценены его параметры.

Величина, определяемая формулой (17), а также ее отношение к наибольшей высоте кривой распределения, уменьшается с ростом объема выборки и, следовательно, характеризует ее размер.

Для нахождения оценок параметров выравнивающего распределения (12) разобьем построенную кривую распределения на интервалы шириной $\Delta\ln(r'+1)=\Delta\ln Y=\Delta t=0{,}5\div1$ и снимем с графика значения ординат $(r'+1)p_{r'}=Yp(Y)=p(t)$ в серединах интервалов. При этом должно выполняться условие $\sum_{i=1}^{n}p(t_i)(\Delta t)_i=1$. По

Таблица 1

Статистические и расчетные данные для построения кривых распределения (рис. 2)

$r'=r-100$	$m_{r'}$	$\ln(r'+1)$	$(r'+1)P_{r'}$	$Yp(Y)$
1	2	3	4	5
1	1084	0,405	0,0026	0,0001
2	1074	0,916	0,0043	0,0009
3	1074	1,253	0,0061	0,0023
4	1039	1,504	0,0075	0,0039
5	1038	1,705	0,0092	0,0057
7	1031	2,015	0,0125	0,0093
10	992	2,351	0,0168	0,0146
12	984	2,526	0,0199	0,0179
15	935	2,741	0,0234	0,0227
20	903	3,020	0,0299	0,0300
30	841	3,418	0,0414	0,0423
50	732	3,922	0,0597	0,0615
70	644	4,263	0,0738	0,0764
100	557	4,615	0,0908	0,0931
150	462	5,017	0,1126	0,1133
200	387	5,303	0,1256	0,1279
300	312	5,707	0,1516	0,1481
400	256	5,994	0,1657	0,1615
500	217	6,217	0,1755	0,1711
700	164	6,553	0,1856	0,1836
1 000	121	6,909	0,1955	0,1935
1 519	83,5	7,326	0,2049	0,1995
2 018	62,5	7,610	0,2037	0,1996
3 030	39,5	8,017	0,1933	0,1934
4 112	27,5	8,322	0,1826	0,1837
5 138	20,5	8,545	0,1701	0,1740
7 379	12,5	8,907	0,1489	0,1538
9 658	8,5	9,176	0,1325	0,1358
11 476	6,5	9,348	0,1204	0,1232
14 436	4,5	9,578	0,1049	0,1057
16 679	3,5	9,722	0,0942	0,0943
20 043	2,5	9,906	0,0809	0,0799
25 789	1,5	10,158	0,0625	0,0607
39 168	0,5	10,576	0,0316	0,0326
—	—	11,000	—	0,0121
139 038	—	11,843	—	0

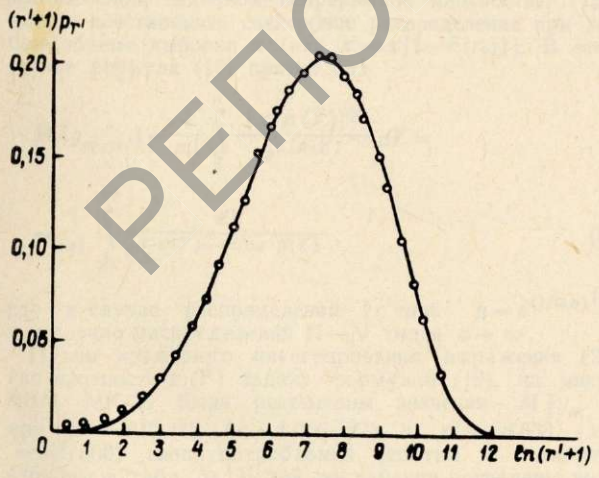

Рис. 2. Статистическая и выравнивающая кривые распределения слов по данным «Частотного словаря русского языка» (без первых 100 слов)

этим данным рассчитаем значения статистических показателей ν^*, S_1^*, S_3^*, H_3^*, B_1^* [2]:

$$\nu_1^* = \overline{\ln t} = \sum_{i\geqslant 1} \ln t_i p(t_i)(\Delta t)_i = 1{,}902637;$$

$$S_1^* = \sum_{i\geqslant 1} t_i [p(t_i)]^2 (\Delta t)_i = 1{,}058225;$$

$$S_3^* = \sum_{i\geqslant 1} t_i^3 [p(t_i)]^4 (\Delta t)_i = 1{,}768733;$$

$$H_3^* = S_3^*/(S_1^*)^3 = 1{,}492548;$$

$$B_1^* = \sum_{i\geqslant 1} t_i (\ln t_i)[p(t_i)]^2 (\Delta t)_i - \nu_1^* S_1^* = 0{,}096605.$$

Далее с помощью рис. 1, приведенного в работе [2], при известных показателях B_1^*, H_3^* находим, что выравнивающее распределение относится к I типу и имеет параметры $u=0{,}25$, $k=1{,}3$. По формулам, приведенным в той же работе, рассчитаем оценки остальных параметров: $\beta=3{,}0571$, $\gamma=k\beta=3{,}9743$, $\alpha u=1/1912{,}6$. Нормирующий множитель N равен

$$N = \frac{\beta\Gamma\left(k+\frac{1}{u}\right)}{\left(\frac{1}{\alpha u}\right)^k \Gamma(k)\,\Gamma\left(\frac{1}{u}\right)} = \frac{1}{853{,}69}.$$

Выравнивающее распределение (12) с учетом найденных оценок параметров имеет вид

$$p(Y) = \frac{(\ln Y)^{2{,}9743}}{853{,}69\,Y}\left[1-\frac{(\ln Y)^{3{,}0571}}{1912{,}6}\right]^3 \qquad (19)$$

и задано на интервале $1<Y<139\,038$.

Рассчитаем по формуле (19) теоретические значения произведения $Yp(Y)$ (пятый столбец в табл. 1). Расчетные и эмпирические данные в четвертом и пятом столбцах при $15\leqslant r<39\,168$ различаются между собой не более, чем на 4%, что оправдывает принятие распределения (12) в качестве выравнивающего. Но, к сожалению, начальный участок не описывается этим распределением, поскольку при $1<Y<17{,}4$ плотность (19) растет от нуля до наибольшего значения $P(Y)=0{,}00146$, затем медленно убывает до нуля при $Y\to 139\,038$, в то время как эмпирическое распределение является невозрастающим при всех рангах $0<r'<39\,168$.

Мода Y_c распределения (12) может быть вычислена по методу итераций по формуле

$$\ln Y_{i+1} = \gamma - 1 - \frac{\alpha u\beta\left(\frac{1}{u}-1\right)(\ln Y_i)^\beta}{1-\alpha u(\ln Y_i)^\beta},$$

откуда в частности следует, что $\ln Y_c<\gamma-1$, $Y_c<e^{\gamma-1}$. Следовательно, мода $y_c=Y_c-1$ будет равна нулю только при условии $\gamma=k\beta\leqslant 1$.

Таким же путем были найдены оценки параметров выравнивающих распределений, заданных обобщенной плотностью (12), для некоторых других частотных словарей (табл. 2). Большинство выравнивающих распределений относится к I типу ($u>0$), причем в среднем $Y_c<10$.

Отметим, что в нескольких случаях наряду с плотностью (12) в качестве выравнивающего распределения оказалась подходящей плотность (11).

Таблица 2

Параметры частотных словарей

Наименование частотного словаря (ЧС) или типа текста	x	y	r_0	$F(r_0)$	Параметры распределения (12), нормирующий множитель				
					u	k	β	$1/\alpha u$ ($1/\alpha$ при $u\to 0$)	$1/N$
1	2	3	4	5	6	7	8	9	10
ЧС русского языка [3]	1 056 382	39 268	100	0,413619	0,25	1,3	3,0571	1912,6	853,69
ЧС белорусского языка [4]:									
а) художественная проза	287 381	21 754	40	0,382464	0,40	10	0,3811	2,5068	90,654
б) публицистика	281 642	18 319	40	0,295009	0,15	1,15	3,3565	4652,7	511,61
в) устное народное творчество	300 000	20 903	40	0,352403	0,30	10	0,4577	3,0836	153,34
ЧС латышского языка [5]:									
а) техника	292 000	13 319	30	0,22	−0,25	0,44	7,0521	−9 526 015	177,05
б) газеты, журналы	282 689	16 103	30	0,23	−0,40	0,34	8,3063	$-1{,}4412\cdot10^8$	131,61
в) художественная литература	291 205	17 769	30	0,30	0,40	10	0,3507	2,3073	42,979
Русские тексты по электронике [6]:									
единица подсчета—словоформа	200 894	21 468	50	0,262248	0,35	1,9	2,0970	155,53	703,44
— ‖ — лексема	200 388	6 826	50	0,333812	0,10	0,66	5,1797	245 593	210,66
Русские тексты по радиотехнике [7]	385 094	28 863	40	0,232665	0,40	0,92	3,5512	4683,8	308,27
ЧС языка газеты («Правда», «Известия») [8]	172 595	14 765	40	0,292	0,10	1,1	3,6341	12 529	665,76
Тексты газет «Пионерская правда», «Пионер Востока» [9]	168 146	13427	10	0,170584	0,30	0,62	3,8364	9764,5	55,013
Все тексты А. С. Пушкина [10]	544 777	21 197	56	0,3999	0,25	0,66	5,0802	183998	330,39
Д. И. Писарев «Реалисты» [11]	48 354	6 348	0	0	0,30	0,9	2,2655	254,38	23,639
Древнерусские тексты («Мерило праведное») [12]	31 262	4 311	0	0	0,50	0,54	2,8725	589,41	13,118
ЧС дескрипторов (тезаурус по строительству в БелРАСНТИ)	340 310	10 035	6	0,083812	0,35	0,92	3,3066	1755 ,8	118,53
ЧС фразеологизмов французского языка [13]	50 022	3132	0	0	0	0,75	3,0162	174,03	19,468
ЧС английского подъязыка судовых механизмов [14]	403 932	12 971	40	0,437735	0,05	0,44	6,3794	10 885 607	106,06
ЧС однословных английских терминов [15]	65 093	4 602	0	0	−0,2	0,36	6,7218	3 139 051	43,955
Смешанные тексты на английском языке [16]	1 014 232	50 406	100	0,474305	0	0,76	4,9322	41 697	797,74
Смешанные тексты на немецком языке [17]	10 910 777	258 173	200	0,540232	0,1	1,5	2,9482	3774,5	2125,7
ЧС чешского языка [18]	1 623 527	54 486	100	0,402535	0	0,52	6,0533	608 894	287,01

3. ВОССТАНОВЛЕНИЕ СПЕКТРОВОГО РАСПРЕДЕЛЕНИЯ ПО РАНГОВОМУ

При известном законе распределения вероятностей разных слов, заданном непрерывной плотностью (12), можно восстановить спектровое распределение при любом объеме выборки x (или $x'=x[1-F(r_0)]$). В этом случае формула (10) примет вид

$$M[y_m, x'] = \frac{1}{m!}\int_0^n \frac{[x'p(Y)]^m}{e^{x'p(Y)}}\,dY =$$

$$= \frac{1}{m!}\int_0^n \frac{dY}{e^{x'p(Y)-m\ln x'p(Y)}}, \qquad (20)$$

где в случае распределений I типа $n=e^{(1/\alpha u)^{1/\beta}}$, а в случае распределений II—V типов $n\to\infty$.

Путем численного интегрирования выражения (20), где плотность $p(Y)$ задана формулой (19), на микро-ЭВМ МК-56 были рассчитаны значения $M[y_m, x']$ при $x'=619\,442$ ($x=1\,056\,382$) и $x'=58\,638$ ($x=100\,000$) словоупотреблений (третий и четвертый столбцы в табл. 3). В той же таблице приведены эмпирические значения количества разных слов с частотой m (второй столбец) для «Частотного словаря русского

Таблица 3

Спектровое распределение слов при $x=1\,056\,382$ и $x=100\,000$

m	эмп y_m при $x=1\,056\,382$ $y=39\,268$	$M[y_m, x']$ при	
		$x'=619\,442$ ($x=1\,056\,382$)	$x'=58\,638$ ($x=100\,000$)
1	13 379	12 853	7425
2	5 746	5636	2423
3	3 364	3337	2211
4	2 243	2 266	730
5	1 681	1 665	490
6	1 279	1 287	352
7	977	1 032	265
8	841	851	206
9	713	715	165
10	595	612	136
15	286	330	58
20	200	210	33
25	131	147	22
30	109	109	15
40	60	68	8
50	45	46	5
		$M[y_{x'}]=38\,917$	$M[y_{x'}]=14\,543$

языка». Данные второго и третьего столбцов достаточно близки между собой.

Проследим далее, как изменяется количество слов с частотами $m = 1, 2, 3$ и их доля, т. е. отношение к объему словаря $M[y_{m,x'}]/M[y_{x'}]$ с ростом объема выборки x'. Вычисляя значения $M[y_{m,x'}]$ по формуле (20), а значения $M[y_{x'}]$ — по формуле

$$M[y_{x'}] = \int_0^n (1 - e^{-x'p(Y)})\,dY, \qquad (21)$$

получим результаты, которые приведены в табл. 4.

Таблица 4

Зависимость объема словаря $M[y_{x'}]$ и количества слов с частотами $m=1, 2, 3$ от объема выборки x или $x'=x[1-F(r_0)]$

$x'=x[1-F(r_0)]$	x	$M[y_{x'}]$	$M[y_{m=1}]$	$M[y_{m=2}]$	$M[y_{m=3}]$
100	171	99	97	2	0
300	512	287	273	12	1
1 000	1 705	873	765	84	17
3 000	5 116	2 167	1 684	293	102
10 000	17 054	5 154	3 425	807	337
30 000	51 161	10 136	5 751	1681	790
100 000	170 538	18 877	8 812	3110	1623
300 000	511 613	30 068	11 480	4656	2634
1 000 000	1 705 376	45 238	13 504	6197	3773
3 000 000	5 116 127	60 507	14 071	7069	4550
10 000 000	17 053 760	77 122	13 303	7224	4887
30 000 000	51 161 270	90 912	11 690	6704	4701
100 000 000	170 537 600	103 681	9 476	5683	4107
∞	∞	139 038	0	0	0

Из табл. 4 видно, что количество слов с частотами $m = 1, 2, 3$, а также доля слов с частотами $m \geqslant 2$ с ростом объема выборки x' от нуля до бесконечности сначала растет от нуля до некоторого наибольшего значения, затем убывает до нуля при $x' \to \infty$. При этом максимум доли слов с частотой $m = 2$ приходится на $x' = 30\,000$ и составляет $1681/10\,136 = 0{,}1658$, а максимум доли слов с частотой $m = 3$ — на $x' = 300\,000$ и составляет $2634/30\,068 = 0{,}0876$. Доля же слов с частотой $m = 1$ с ростом объема выборки непрерывно убывает от 1 до 0.

По количеству слов с частотой $m = 1$ можно оценивать вероятность появления нового слова в выборке объемом x (или x'):

$$\frac{dM[y_x]}{dx} = \frac{M[y_{m=1,x}]}{x}, \qquad (22)$$

$$\frac{dM[y_{x'}]}{dx'} = \frac{M[y_{m=1,x}]}{x'} = \frac{M[y_{m=1,x}]}{x[1-F(r_0)]}. \qquad (23)$$

Из формул (22) и (23) следует, что *скорость роста новых слов (она же — вероятность появления нового слова) в полном тексте меньше скорости роста новых слов в тексте без первых r_0 наиболее частых слов в $1/[1-F(r_0)]$ раз*, что для частотного словаря русского языка составляет 1,7 раза.

Полнота частотного словаря (т. е. накопленная вероятность у разных слов, встретившихся в выборке объемом x словоупотреблений) определяется по формуле

$$F(y) = 1 - \frac{dy}{dx} \approx 1 - \frac{y_{m=1}}{x}$$

и для частотного словаря русского языка составляет 0,9873. По величине $F(y)$ (или $dy/dx \approx y_{m=1}/x$) можно судить о размерах выборки и сравнивать между собой различные частотные словари.

Отметим, что критерии (2) и (17), характеризующие объем выборки, при условии $y_{m=1}/y = 1/2$ совпадают.

На основании проведенного исследования можно сделать следующие выводы:

статистические ранговые распределения слов по частоте их употребления в выборках достаточно большого объема *в случае однородности лексического состава частотного словаря* с высокой точностью *описываются обобщенным логарифмическим распределением* (12) (за исключением одного — двух десятков наиболее частых слов);

знание выравнивающего распределения и оценок параметров позволяет находить спектровое распределение (т. е. восстанавливать статистическую структуру выборки) при любом объеме выборки x, *рассчитывать кривую роста разных слов, находить полноту словаря и вероятность появления нового слова в выборке заданного объема*, а также решать другие практические задачи.

ЛИТЕРАТУРА

1. Нешитой В. В. Построение и исследование системы дискретных распределений /БелНИИНТИ.— Минск, 1985. — 71 с. — Деп. в БелНИИНТИ 17.07.85, № 931.
2. Нешитой В. В. Исследование ранговых распределений // НТИ. Сер. 2.—1985.—№ 2.— С. 16—20.
3. Частотный словарь русского языка. /Под ред. Л. Н. Засориной. — М.: Русский язык, 1977.
4. Можейко Н. С., Супрун А. Е. Частотный словарь белорусского языка. — Минск: БГУ, 1976—1982.
5. Частотный словарь латышского языка. — Рига: Зинатне, 1966—1973.
6. Калинина Е. А. Изучение лексико-статистических закономерностей на основе вероятностной модели // // Статистика речи. Л.: Наука, 1968.
7. Межлумова А. Б. Лексический минимум по радиотехнической специальности. — Минск, 1972.
8. Полякова Г. П., Солганик Г. Я. Частотный словарь языка газеты. — М.: МГУ, 1971.
9. Ульяновская Р. П. Статистическое обследование лексики газет «Пионерская правда» и «Пионер Востока» // Вопросы методики преподавания русского языка в узбекской школе. Ташкент, 1962.
10. Фрумкина Р. М. Статистические методы изучения лексики. — М.: Наука, 1964.
11. Булахов М. Г. Материалы для частотного словаря русского языка Лексікалогія і грамматика // Минск; БГУ, 1969.
12. Вялкина Л. В., Лукина Г. Н. Материалы к частотному словарю древнерусских текстов // Лексикология и словообразование древнерусского языка. М.: Наука, 1966.
13. Частотный список фразеологизмов французского языка. / Сост. М. И. Берлин, Л. Н. Жолудева, З. Н. Левит и др. — Минск: Вышэйшая школа, 1979.
14. Лукьяненков К. Ф. Лексико-статистическое описание английского научно-технического текста с помощью электронно-вычислительной машины (подъязык судовых механизмов): Дис... канд. филол. наук — Минск, 1969.
15. Частотный англо-русский словарь-минимум по квантовым генераторам / Сост. Н. С. Манасян.— М.; Воениздат, 1983.

16. Kučera H, Francis W. N. Computational Analysis of Present — day American English.— Providence, 1967.
17. Meier H. Deutsche Sprachstatistik.— Hildesheim, 1964.
18. Jelínek J., Bečka J. V. Těšitelová M. Frekvence slov, slovnich druhů a tvarů v českém jazyce. — Praha, 1961.

Статья поступила в редакцию 3 октября 1985 г.