УДК 801

РАСПРЕДЕЛЕНИЕ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ ПО ДЛИНЕ

В. В. Нешитой

Цель настоящей статьи — установить общий закон распределения по длине однородных лингвистических единиц: слов или словоформ словаря, фраз, моделей терминов, словосочетаний.

Длина лингвистической единицы, измеряемая числом букв или слов, — T — является дискретной случайной величиной. Однако для выравнивания такого рода статистических распределений с большим успехом могут использоваться непрерывные распределения.

Примем в качестве выравнивающего распределение, заданное обобщенной плотностью [1]:

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

где N — нормирующий множитель; α, β, γ, u — параметры.

Для установления типа выравнивающей кривой и нахождения оценок параметров достаточно по данным статистического распределения вычислить значения: ν_1^*, S_1^*, B_1^*, H_3^*. Расчетные формулы приведены в работе [2, с. 22]. Поскольку в данном случае из опыта известны абсолютные частоты m_i случайной величины T при $t_i = 1, 2, \ldots$, то эти формулы удобнее записать в виде:

$$\left.\begin{aligned} &\nu_1^* = \overline{\ln t} = \sum_{i\geqslant 1} \ln t_i \frac{m_i}{M}, \quad S_1^* = \sum_{i\geqslant 1} t_i \left(\frac{m_i}{M}\right)^2, \\ &S_3^* = \sum_{i\geqslant 1} t_i^3 \left(\frac{m_i}{M}\right)^4, \quad H_3^* = S_3^*/(S_1^*)^3, \\ &B_1^* = \sum_{i\geqslant 1} t_i (\ln t_i)\left(\frac{m_i}{M}\right)^2 - \nu_1^* S_1^*, \end{aligned}\right\} \tag{2}$$

где $m_i/M = p_i$ — эмпирическая относительная частота случайной величины T при данных значениях t_i; $M = \sum_{i\geqslant 1} m_i$ — объем выборки.

Тип выравнивающей кривой и оценки параметров $k = \gamma/\beta$, u устанавливаются с помощью номограммы [1, с. 18] приравниванием статистических показателей B_1^*, H_3^* соответствующим теоретическим. Оценки параметров β, α (или произведения αu) находятся по величинам S_1^*, ν_1^*. Расчетные формулы в зависимости от типа кривой имеют вид:

типы I—V, I′, II′ — $\beta = S_1^*/S_1^{(z)}$, (3)

типы I, I′ — $\alpha u = e^{\pm\left(\nu_1^{(z)} - \beta\nu_1^*\right)}$, (4)

типы II, II′ — $\alpha = e^{\pm\left(\nu_1^{(z)} - \beta\nu_1^*\right)}$, (5)

типы III—V — $\alpha u = -e^{\nu_1^{(z)} - \beta\nu_1^*}$ (6)

(знак «минус» перед скобкой здесь и ниже относится к распределениям I и II′ типов).

Величины $S_1^{(z)}$, $\nu_1^{(z)}$ в зависимости от типа кривой вычисляются по формулам, приведенным в работе [1], либо по следующим более простым формулам:

Типы I, I′:

$$S_1^{(z)} = \frac{2\left(k+\frac{1}{u}\right)-1}{2\sqrt{\pi}\left(\frac{2}{u}-1\right)} \cdot \frac{g(k)\, g\left(\frac{1}{u}\right)}{g\left(k+\frac{1}{u}\right)}, \tag{7}$$

$$\nu_1^{(z)} = \pm\left[\psi(k) - \psi\left(k+\frac{1}{u}\right)\right]. \tag{8}$$

Типы II, II′:

$$S_1^{(z)} = \frac{g(k)}{2\sqrt{\pi}}, \tag{9}$$

$$\nu_1^{(z)} = \pm\psi(k) = \pm\left(\sum_{s=1}^{k-1}\frac{1}{s} - C\right). \tag{10}$$

Типы III—V:

$$S_1^{(z)} = \frac{1}{2\sqrt{\pi}} \cdot \frac{g(k)\, g\left(1-\frac{1}{u}-k\right)}{g\left(1-\frac{1}{u}\right)}, \tag{11}$$

$$\nu_1^{(z)} = \psi(k) - \psi\left(1-\frac{1}{u}-k\right). \tag{12}$$

Величины $\psi(x)$, $g(x)$, а также $\ln\Gamma(x)$ приближенно могут быть вычислены по формулам:

$$\psi(x)\approx-\left(\frac{1}{x}+\frac{1}{x+1}\right)+\ln(x+2)-$$
$$-\frac{1}{2(x+2)}\left[1+\frac{1}{6(x+2)}\right]; \qquad (13)$$

$$g(x)=\frac{\Gamma\left(x+\frac{1}{2}\right)}{\Gamma(x)}\approx\frac{x(x+0{,}875)}{(x+0{,}5)\sqrt{x+1}}\approx$$
$$\approx\frac{x(x+1)(x+1{,}875)}{(x+0{,}5)(x+1{,}5)\sqrt{x+2}}; \qquad (14)$$

$$\ln\Gamma(x)\approx 0{,}91894-\ln[x(x+1)]+$$
$$+(x+1{,}5)\ln(x+2)-(x+2)+\frac{1}{12(x+2)}. \qquad (15)$$

Точность этих формул тем выше, чем больше x.

1. РАСПРЕДЕЛЕНИЕ МОДЕЛЕЙ ФИЛОСОФСКИХ ТЕРМИНОВ ПО ДЛИНЕ

Рассмотрим статистические данные [3, с. 4], приведенные в табл. 1, столбцы 1 и 2. Здесь $M=824$ (объем выборки). Вычислим относительные частоты $p_i=m_i/M$

Таблица 1

Распределение моделей философских терминов по длине

Число слов в модели, t_i	Число моделей с данным количеством слов, m_i	$p_i=\frac{m_i}{M}$	$p(t)$
1	1	0,0012	0,0012
2	19	0,0231	0,0213
3	78	0,0947	0,0985
4	181	0,2197	0,2152
5	209	0,2536	0,2612
6	172	0,2087	0,1998
7	86	0,1044	0,1115
8	45	0,0546	0,0520
9	19	0,0231	0,0224
10	8	0,0097	0,0095
11	3	0,0036	0,0041
12	2	0,0024	0,0018
13	1	0,0012	0,0008

и запишем их в столбец 3. Установим тип выравнивающего распределения и найдем оценки параметров. Для этого по данным табл. 1 и выражениям (2) вычислим статистические показатели ν_1^*, S_1^*, B_1^*, H_3^*. Они оказались равными: $\nu_1^*=\overline{\ln t}=1{,}614266$; $S_1^*=\overline{t\,p(t)}=0{,}910119$; $H_3^*=1{,}491060$; $B_1^*=0{,}026025$. По рис. 1, приведенному в работе [1], находим, что $u\approx-1/3$, $k\approx 1{,}4$. Выравнивающее распределение относится к третьему типу (для которого $-\infty<u<0$). Далее по выражениям (11), (12) при известных значениях k, u находим: $\nu_1^{(z)}=-0{,}8124$, $S_1^{(z)}=0{,}2425$. Тогда параметры выравнивающего распределения в соответствии с (3), (6) будут равны: $\beta=3{,}7530$; $\gamma=k\beta=5{,}2542$; $\alpha u=-1/963{,}78$. Нормирующий множитель для распределений третьего типа задается формулой

$$N=\frac{\beta(-\alpha u)^k\Gamma\left(1-\frac{1}{u}\right)}{\Gamma(k)\,\Gamma\left(1-\frac{1}{u}-k\right)}$$

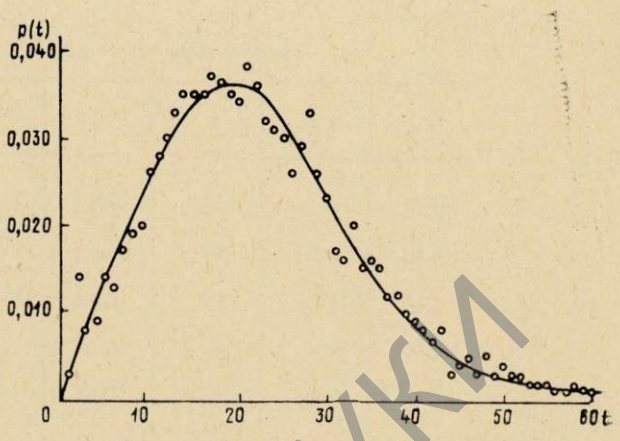

Рис. 1. Распределение фраз английских текстов по электронике по длине

и равен $N=1/847{,}842$. Выравнивающее распределение имеет вид:

$$p(t)=\frac{t^{4{,}2542}}{847{,}842\left(1+\frac{t^{3{,}753}}{963{,}78}\right)^4},\quad 0<t<\infty.$$

В табл. 1 приведены расчетные значения $p(t)$ (столбец 4).

Для проверки степени согласия выравнивающего распределения с эмпирическим вычислим критерий «хи-квадрат» Пирсона:

$$\chi^2=\sum_{i=1}^{13}\frac{[m_i-M\cdot p(t)]^2}{M\cdot p(t)}=1{,}62.$$

По статистическим таблицам находим, что такому значению χ^2 при числе степеней свободы $r=13-1-4=8$ соответствует $P(1{,}62)=0{,}99$, т. е. наблюдаемое расхождение могло появиться за счет случайных причин с вероятностью 0,99. Другими словами, совпадение обоих распределений весьма хорошее.

2. РАСПРЕДЕЛЕНИЕ ФРАЗ ПО ДЛИНЕ

Построим по данным таблицы 4 из работы [4, с. 226] статистическое распределение по длине фраз английских текстов по электронике (рис. 1) и найдем выравнивающее распределение, заданное плотностью (1). Объем выборки $M=5000$ фраз.

В результате вычислений находим, что выравнивающее распределение, как и в предыдущем случае, относится к третьему типу, при этом его параметры равны: $u=-0{,}8$; $k=0{,}5$; $\beta=3{,}7748$; $\gamma=1{,}8874$; $\alpha u=-1/610067$; $N=1/297{,}503$.

На рис. 1 сплошная линия соответствует выравнивающей кривой распределения фраз по длине (длина фразы измеряется количеством словоупотреблений).

Таким же путем были найдены оценки параметров выравнивающих распределений по длине фраз других текстов (см. табл. 2). В пяти рассмотренных случаях выравнивающее распределение относится к третьему типу.

Таблица 2

Оценки параметров выравнивающих распределений фраз по длине

Текст, источник	u	κ	β	γ	$1/\alpha u$	$1/N$
1. Английские тексты по электронике [4, с. 226], M=5000 фраз	−0,8	0,5	3,7748	1,8874	−610067	297,503
2. Русские тексты по радиоэлектронике [5, с. 51—52], M=8000 фраз	−0,25	1,20	2,0066	2,4079	−1229,63	456,601
3. Болгарские технические тексты [5, с. 111], M=5000 Фраз	−0,40	0,78	2,4176	1,8857	−3857,05	145,459
4. Газетные тексты языка хауса, [5, с. 191], M=5000 Фраз	−0,30	0,60	2,7300	1,6380	52844,1	174,228
5. Болгарские агрономические тексты [5, с. 331], M=5400 Фраз	−1	0,48	3,5875	1,7220	−240398	174,606

3. РАСПРЕДЕЛЕНИЕ СЛОВОФОРМ И СЛОВ СЛОВАРЯ ПО ДЛИНЕ

Пример 1. Для установления закона распределения словоформ по длине обратимся к статистической таблице 16,8 из работы [6, с. 272]. Статистические показатели, вычисленные по данным этой таблицы, равны: $H_3^*=1{,}476493$; $B_1^*=0{,}022719$; $\nu_1^*=2{,}190308$; $S_1^*=0{,}774957$. Выравнивающее распределение имеет параметры: $u=-0{,}25$; $k=1{,}70$, т. е. относится к третьему типу. Вычисление остальных параметров дает: $\beta=2{,}8245$; $\gamma=4{,}8016$; $\alpha u=-1/1110{,}77$; $N=1/5413{,}65$.

Пример. 2. По данным таблицы 1 из работы [7, с. 20] (всего 17 086 слов), было найдено выравнивающее распределение с параметрами: $u=0{,}10$; $k=4$; $\beta=0{,}9466$; $\gamma=3{,}7865$; $\alpha u=1/33{,}5143$. Распределение относится к первому типу ($0<u<\infty$). Нормирующий множитель задается формулой

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

и в данном случае равен $N=1/465{,}988$

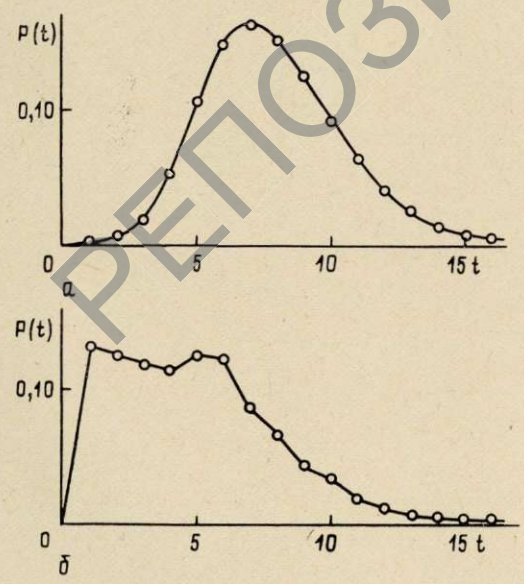

Рис. 2. Распределение словоформ словаря (*а*) и словоупотреблений текста (*б*) по длине

Пример 3. По статистическим данным о распределении словоформ по длине в словаре русского эпистолярного текста (всего 15 800 разных словоформ) [8, с. 142] установлены параметры выравнивающего распределения: $u=-0{,}05$; $k=3{,}5$; $\beta=1{,}8178$; $\gamma=6{,}3622$; $\alpha u=-1/214{,}643$; $N=1/9323{,}3$.

На рис. 2, *а* приведены статистическое (отдельные точки) и выравнивающее (сплошная линия) распределения словоформ словаря по длине. Оба распределения практически совпали. На рис. 2, *б* представлено эмпирическое распределение **словоупотреблений текста** (статистические данные взяты из того же источника [8, с, 142]), которое, в отличие от первого распределения, имеет неправильную форму.

Итак, статистические распределения по длине разных слов **(словоформ) в словаре** хорошо описываются плотностью (1), о чем свидетельствует рис. 2,*а*. В то же время статистические распределения словоупотреблений текста по длине не могут быть описаны плотностью (1), поскольку случайная величина T — длина словоупотреблений — не является однородной из-за высокой частоты служебных слов, имеющих, как правило, небольшую длину (рис. 2,*б*).

4. РАСПРЕДЕЛЕНИЕ СЛОВОСОЧЕТАНИЙ ПО ДЛИНЕ

Заимствованные из работы [9, с. 34] статистические данные о длине словосочетаний (табл. 3, столбцы 1, 2; объем выборки $M=1187$) отличаются от рассмотренных

Таблица 3

Распределение словосочетаний по длине

Количество слов в словосочетании, v_i	Количество словосочетаний, m_i	Относительная частота, p_i	$t_i=v_i-1$	$p(t)$
2	52	0,0438	1	0,0394
3	131	0,1104	2	0,1145
4	238	0,2005	3	0,1919
5	274	0,2308	4	0,2304
6	215	0,1811	5	0,2042
7	191	0,1609	6	0,1320
8	44	0,0371	7	0,0611
9	21	0,0177	8	0,0201
10	10	0,0084	9	0,0047
11	11	0,0093	10	0,0008

ранее примеров тем, что здесь случайная переменная $V \geqslant 2$, где V — количество слов в словосочетании.

Пусть по-прежнему выравнивающее распределение задано обобщенной плотностью (1). Тогда $T = V - 1$. Теперь по данным столбцов 2 и 4 табл. 3 нетрудно оценить параметры выравнивающего распределения: $u = -0{,}05$; $k = 0{,}80$; $\beta = 3{,}2503$; $\gamma = 2{,}6003$; $\alpha\ u = -1/4099{,}64$; $N = 1/25{,}2197$.

Расчетные значения плотности $p(t)$ при целых t близки к относительным частотам (см. столбцы 3 и 5 в табл. 3).

* * *

Рассмотренные примеры на выборках большого объема показали, что статистические распределения по длине однородных лингвистических единиц — слов или словоформ словаря, фраз, моделей терминов, словосочетаний — хорошо выравниваются обобщенной плотностью (1). При этом выравнивающие распределения относятся к I—III типам.

Случайная величина T — длина словоупотреблений текста — является неоднородной случайной величиной из-за высокой частоты употребления в тексте служебных слов, имеющих, как правило, небольшую длину. Поэтому ее распределение не может быть описано обобщенной плотностью (1).

ЛИТЕРАТУРА

1. Нешитой В. В. Исследование ранговых распределений // НТИ. Сер. 2. — 1985. — № 2. — С. 16—20.
2. Нешитой В. В. О взаимосвязи ранговых распределений со спектровыми // НТИ. Сер. 2. — 1986. — № 10. — С. 19—25.
3. Кобрин Р. Ю. О принципах терминологической работы при создании тезаурусов для информационно-поисковых систем // НТИ. Сер. 2. — 1979. — № 6. — С. 1—9.
4. Малаховский Л. В. Некоторые статистические характеристики английских текстов по электронике // Статистика речи. — Л.: Наука, 1968. — С. 222—227.
5. Статистико-комбинаторное моделирование языков/ Ин-т языкознания АН СССР/Под ред. Н. Д. Андреева. — М.; Л.: Наука, 1965. — 502 с.
6. Белоногов Г. Г., Богатырев В. И. Автоматизированные информационные системы. — М.: Советское радио, 1973. — 328 с.
7. Владимирова Е. В., Карпова Г. Д., Лескис Г. А., Уриновская П. Д., О размере слов в письменной речи // НТИ. Сер. 2. — 1986. — № 3. — С. 20—27.
8. Григорьева А. С. О лексико-морфологической статистике русской эпистолярной речи // Инженерная лингвистика и преподавание иностранных языков с помощью ТСО: Межвузовский сборник научных трудов/ЛГПИ им. А. И. Герцена. — Л., 1981. — С. 140—148.
9. Колтун А. Я., Пшеничная Л. Э. Использование терминов заглавия для автоматического реферирования текста научного документа // Автоматическая обработка текста: Препринт/Ин-т кибернетики АН УССР. — Киев, 1980. — С. 29—37.

Статья поступила в редакцию 6 июля 1987 г.