ДЛИНА ТЕКСТА И ОБЪЕМ СЛОВАРЯ. ПОКАЗАТЕЛИ ЛЕКСИЧЕСКОГО БОГАТСТВА ТЕКСТА

В. В. Нешитой

Для определения объема словаря при заданной длине текста необходимо знать закон распределения разных слов по частоте их употребления в тексте. Этот закон можно представить в виде вероятностного словаря, т. е. некоторого списка слов, упорядоченных по убыванию их вероятностей. Здесь и далее под текстом понимается некоторая случайная выборка из исследуемой совокупности текстов.

Объем словаря y в зависимости от длины текста x может быть вычислен по формуле [195]

$$y = \int_0^{\infty} \left(1 - \frac{1}{e^{xp(z)}}\right) dz, \qquad (1)$$

где $p(z)$ — плотность распределения вероятностей разных слов в тексте, z — порядковый номер слова в списке по убывающим вероятностям.

Г. Г. Белоногов [30] показал, что распределение разных слов в тексте, за исключением первой сотни наиболее частых слов, подчиняется закону Вейбулла, функция распределения и плотность вероятности которого определяются формулами

$$F(z) = 1 - \frac{1}{e^{az^b}}, \tag{2}$$

$$p(z) = \frac{ab}{z^{1-b}\, e^{az^b}}, \tag{3}$$

где a, b — параметры распределения: $a>0$; $b>0$; $0<z<\infty$; $F(z)$ — вероятность того, что взятое наудачу слово текста окажется среди первых z наиболее частых слов вероятностного словаря.

Выражение (2) можно привести к виду

$$\ln\ln\frac{1}{1-F(z)} = \ln a + b\ln z \tag{4}$$

или

$$\lg\lg\frac{1}{1-F(z)} = \lg a - 0{,}362 + b\lg z, \tag{4'}$$

т. е. получили уравнение прямой с начальной ординатой $\ln a$ (или $\lg a - 0{,}362$) и угловым коэффициентом b.

График зависимости (4), построенный по опытным значениям функции распределения $F^*(z)$, при $z>100$ представляет собой прямую, т. е. выполняется закон Вейбулла. При $z<100$ имеем кривую, т. е. для самых частых слов закон Вейбулла не выполняется. Здесь $F^*(z)$ есть накопленная относительная частота первых z слов частотного словаря.

Введем в формулы (2) и (3) третий параметр c, учитывающий особенность распределения в тексте первой сотни наиболее частых слов, и запишем выражение для функции распределения в виде [195]:

$$F(z) = 1 - \frac{1}{e^{a[(z+1)^b - e^{-cz}]}}, \tag{5}$$

тогда

$$p(z) = \frac{\dfrac{ab}{(z+1)^{1-b}} + \dfrac{ac}{e^{cz}}}{e^{a[(z+1)^b - e^{-cz}]}}. \tag{6}$$

При $z \to \infty$ (практически при $z > 100 \div 500$) приведенный закон распределения совпадает с распределением Вейбулла.

Для нахождения значений параметров распределения a и b необходимо по опытной функции распределения $F^*(z)$ при $z > 100$ построить график зависимости (4), которая представляет собой прямую, и определить начальную ординату $\ln a$ и угловой коэффициент b данной прямой. При известных параметрах a и b значение параметра c находится по формуле

$$c = -\frac{1}{z} \ln [(z+1)^b + \frac{1}{a} \ln (1 - F(z))], \quad (7)$$

которая следует из (5). При вычислении параметра c необходимо принимать $z = 1 \div 30$.

Если в формулу (1) вместо плотности распределения $p(z)$ подставить ее значение из (6), то получим интеграл, который не выражается через элементарные функции. В этом случае объем словаря y при заданной длине текста x можно рассчитать путем численного интегрирования, но такие расчеты громоздки. Поэтому на практике для вычисления объема словаря удобнее воспользоваться простыми приближенными формулами.

Зависимость между длиной текста X и объемом словаря Y приближенно может быть описана следующим дифференциальным уравнением [195]:

$$\frac{dY}{dX} = \frac{Y}{X} (1 - u \alpha \ln Y)^{\frac{1}{u}}. \quad (8)$$

где $Y = y + 1$; $X = x + 1$; $\alpha > 0$; u — некоторый параметр, могущий принимать как положительные, так и отрицательные целые и дробные значения.

Из общего уравнения (8) при различных значениях параметра u можно получить ряд формул для описания зависимости между длиной текста и объемом словаря. Однако наиболее подходящими являются формулы, полученные в следующих двух случаях.

Случай 1. $u = -1$.

Решая дифференциальное уравнение (8) при $u = -1$ и используя начальные условия: $Y = 1$ при $X = 1$, найдем

$$X = Y^{1 + \frac{\alpha}{2} \ln Y},$$

откуда

$$Y = e^{\frac{1}{\alpha}\left(\sqrt{1+2\alpha \ln X}-1\right)}, \tag{9}$$

$$\alpha = \frac{2}{\ln Y}\left(\frac{\ln X}{\ln Y}-1\right). \tag{10}$$

Случай 2. $u \approx -\frac{1}{4}$.

$$Y = X^{\frac{1}{\sqrt{1+\alpha \ln X}}}, \tag{11}$$

$$\alpha = \frac{1}{\ln X}\left[\left(\frac{\ln X}{\ln Y}\right)^2 - 1\right]. \tag{12}$$

Здесь параметр α представляет собой среднее арифметическое из двух его значений, найденных из уравнения (8) при $u=-1$ и $u=\frac{1}{2}$.

Параметр α, вычисляемый по опытным значениям Y и X, не является постоянной величиной, однако он изменяется закономерно. Если построить график функции $\alpha = \varphi(\ln X)$, то получим прямую, уравнение которой имеет вид

$$\alpha = \alpha_0 + k \ln X, \tag{13}$$

или

$$\alpha = \alpha_0 + k' \lg X, \tag{13'}$$

где α_0 — начальная ордината; k, $k'=2,3k$ — угловые коэффициенты данной прямой.

Таким образом, если в формулах (9) и (11) считать α величиной переменной ($\alpha = \alpha_0 + k \ln X$), то они становятся практически точными и содержат два параметра: α_0 и k. Но пределы применимости этих формул ограничены. Опытная проверка показывает, что формула (9) ($u=-1$) справедлива в пределах:

$$10^3 < X < 10^8 \text{ при } 0{,}50 < b < 0{,}60.$$

При этом параметры α_0 и $k'=2,3k$, входящие в формулу (9), связаны с параметрами a и b распределения Вейбулла следующими соотношениями, найденными из опыта:

$$\alpha_0 = -0,\ 0163\, e^{2,83\, b}\,(1 - 3{,}62\, a e^{0,65\, b}); \tag{14}$$

$$\lg k' = -0{,}514 + 0{,}31 \lg a + 2{,}35 \lg b. \tag{15}$$

Пределы применимости для формулы (11) ($u \approx -\frac{1}{4}$):

$$10^4 < X < 10^8 \text{ при } 0{,}30 < b < 0{,}40;$$
$$10^3 < X < 10^6 \text{ при } b = 0{,}50.$$

Здесь

$$\alpha_0 = -0{,}00738\, e^{4{,}17\, b}(1 - 4{,}99\, a e^{0{,}92\, b}); \quad (16)$$
$$\lg k' = -0{,}732 + 0{,}22 \lg a + 2{,}36 \lg b. \quad (17)$$

Относительная погрешность формул (9) и (11) в указанных пределах не превышает 3÷5% по сравнению с теми данными, которые можно получить по формуле (1) на основе плотности распределения Вейбулла (3) или близкой к ней плотности распределения (6). При небольших значениях X ($X \leqslant 10^3 \div 10$) формулы (9) и (11) дают несколько завышенные значения Y, а при $X \geqslant 10^7 \div 10^8$ — заниженные значения Y.

Эмпирические формулы (14), (15) и (16), (17) позволяют находить значения параметров текста α_0 и k' по известным значениям параметров распределения a и b, а также решать обратную задачу. Если параметры a и b определены правильно, то должно выполняться равенство $\alpha_0 + k' \lg X \approx \alpha$, где величина α вычисляется по опытным значениям объема словаря Y и длины текста X по формулам (10) или (12).

Отметим, что формула (9) хорошо описывает прирост новых слов в текстах, которые являются наиболее бедными в лексическом отношении. Для таких текстов параметр $b = 0{,}50 \div 0{,}60$. Сюда относятся, например, тексты, составленные из ключевых слов, а также технические тексты, когда за единицу подсчета количества разных слов принимается лексема.

Формула (11) хорошо описывает прирост новых слов в текстах, для которых значения параметра b распределения Вейбулла находятся в пределах 0,30÷0,40.

Расчеты, произведенные по опытным данным, показали, что тексты, относящиеся к одному жанру, имеют примерно одинаковые значения параметра k. Это значит, что прямые $\alpha = \alpha_0 + k \ln X$, построенные для таких текстов, параллельны. Чем богаче в лексическом отношении текст, тем ниже располагается на графике прямая, характеризующая данный текст (т. е. меньше начальная ордината α_0). Следовательно, параметр α_0 может служить характеристикой лексического разнообразия или лексического богатства для текстов одного жанра.

С другой стороны, при одинаковых значениях параметра α_0 лексически богаче тот текст, у которого параметр k меньше. Величина $\alpha = \alpha_0 + k \ln X$ является обобщенным показателем лексического богатства текста.

Величина параметра k зависит от выбора единицы подсчета количества разных слов (словоформа, лексема), при этом $k_{\text{лекс}} > k_{\text{сл}}$. Это позволяет ввести показатель степени аналитичности языка

$$\Delta k_a = \frac{\alpha_{\text{лекс}} - \alpha_{\text{сл}}}{\ln X}, \qquad (18)$$

который практически не зависит от длины текста.

Формула (18) позволяет находить значение Δk_a по трем известным величинам: X, $Y_{\text{лекс}}$, $Y_{\text{сл}}$. Чем ближе Δk_a к нулю, тем выше степень аналитичности языка. Опытная проверка показала, что коэффициент Δk_a принимает различные значения для разных функциональных стилей данного языка, т. е. он может служить показателем грамматического богатства стиля.

Для русского языка, по данным Л. Н. Засориной [98], были вычислены значения параметров α_0 и k'. Обобщенный показатель лексического богатства для смешанных текстов на русском языке оказался равным (при $u \approx -\frac{1}{4}$) — $\alpha_{\text{лекс}} = -0{,}0111 + 0{,}0105 \lg X$.

Тот же показатель для смешанных текстов на чешском языке [359] равен $\alpha_{\text{лекс}} = -0{,}0148 + 0{,}0105 \lg X$.

Таким образом, в обоих случаях $k' = 0{,}0105$.

Отметим, что показатели лексического богатства текста могут быть получены также на основе закона распределения разных слов, функция распределения и плотность вероятности которого определяются формулами (5) и (6). Запишем выражение для тангенса угла наклона $\gamma = \frac{d \ln p(z)}{d \ln z}$ касательной к кривой распределения $\lg p(z) = \varphi(\ln z)$, которое следует из (6):

$$\gamma = -\frac{b(1-b)e^{cz}\frac{z}{z+1} + c^2 z(z+1)^{1-b}}{be^{cz} + c(z+1)^{1-b}} - \frac{abz}{(z+1)^{1-b}} - \frac{acz}{e^{cz}}, \qquad (19)$$

График функции $\gamma = \varphi_1(\ln z)$ может пересекать прямую $\gamma = -1$ в одной или трех точках, а график функции

распределения $F(z) = \varphi_2(\ln z)$ имеет соответственно одну или три точки перегиба. С этой точки зрения все тексты можно разделить на две группы: 1 — тексты, характеризующиеся одной точкой, в которой $\gamma = -1$ (это тексты, наиболее бедные в лексическом отношении), и 2— тексты, характеризующиеся тремя точками, в которых $\gamma = -1$. Чем богаче в лексическом отношении текст, тем меньше кривизна кривой распределения $\ln p(z) = \varphi(\ln z)$, т. е. с увеличением числа разных слов, содержащихся в тексте неизменной длины, указанная кривая распрямляется.

В качестве показателя лексического богатства можно ввести еще одну величину, которую обозначим через z_A. Она находится по формуле

$$z_A = \left(\frac{\sqrt{9(1-b)^2 + 8} - (1-b)}{4ab} \right)^{\frac{1}{b}}, \qquad (20)$$

где a, b — параметры распределения Вейбулла. Величина $\ln z_A$ есть абсцисса некоторой точки А, расположенной на кривой распределения $\ln p(z) = \varphi(\ln z)$ в интервале, где $1 < |\gamma| < \infty$. В точке А кривизна этой кривой достигает максимального для данного интервала значения.

Рассмотрим случай, когда кривая распределения содержит три точки, в которых $\gamma = -1$. На участке слева от точки А величина γ близка к -1, а кривая распределения близка к прямой. Справа от точки А с ростом величины $\ln z$ значения параметра γ (а также величины $\ln p(z)$) по модулю неограниченно растут, т. е. правее точки А расположена зона низкочастотных слов. Следовательно, величина z_A определяет зону наиболее частых слов. Чем богаче в лексическом отношении текст, тем больше величина z_A. Для литературных текстов $z_A \approx 3000 \div 4000$.

Все приведенные выше формулы были получены для случайно составленной выборки, т. е. для такого условного текста, в котором разные слова появляются независимо и случайно. В таком тексте все разные слова могут сочетаться между собой без каких-либо ограничений. В реальном тексте грамматика и семантика накладывают определенные ограничения на сочетаемость между словами. Поэтому в отрезке сплошного текста содержится в среднем меньшее количество разных слов, чем в случайно составленной выборке равного объема. При постоянной

длине обследуемого текста, выступающего в роли генеральной совокупности, с увеличением длины выборки эта разница должна уменьшаться, а в предельном случае при длине выборки, равной длине обследуемого текста, она равна нулю.

Опытная проверка показала, что прирост новых слов в связном тексте описывается теми же формулами (9) и (11), что и для случайно составленной выборки, но с другими значениями параметров α_0 и k. Запишем уравнение прямой, характеризующей связный текст, в виде

$$\alpha_T = \alpha_{0T} + k_T \ln X \qquad (13'')$$

в отличие от прямой $\alpha = \alpha_0 + k \ln X$ для случайно составленной выборки. Здесь величина X обозначает одновременно длину текста и длину выборки.

Как показывают расчеты, параметр k не зависит от величины X, а параметр α_0 с ростом X уменьшается пропорционально $\ln X$. В то же время параметры α_{0T} и k_T не зависят от длины текста. При этом $k > k_T$. Величина $\Delta k = k - k_T$ зависит от степени связности слов в тексте (будем измерять ее показателем $h_{св}$), а также от степени неоднородности текста по лексическому составу ($\eta_{неодн}$). Чем выше показатели $\eta_{св}$ и $\eta_{неодн}$, тем больше величина Δk. Степень связности слов в тексте и степень неоднородности того же текста можно вычислять по формуле

$$\eta_{св} + \eta_{неодн} = \frac{\Delta k}{k} = 1 - \frac{k_T}{k}.$$

Если принять, что для лексически однородных текстов величина $\eta_{неодн} = 0$, то величина $\eta_{св}$, вычисленная для таких текстов на основе опытных данных, примерно равна 0,33 (при $u \approx -\frac{1}{4}$), а отношение $k/k_T \approx 1{,}5$.

Из формул (9) и (11) с учетом (13″) можно найти зависимость X от Y. В случае, когда $u = -1$, имеем

$$X = Y^{\frac{2 + \alpha_{0T} \ln Y}{2 - k_T \ln^2 Y}}.$$

Из последнего уравнения следует, что величина Y достигает максимального значения при $X = \infty$, при этом

$$Y_{max} = e^{\sqrt{\frac{2}{k_T}}}.$$

В случае 2, при $u \approx -\frac{1}{4}$

$$X = Y^{\frac{\alpha_{0T} \ln Y + \sqrt{(\alpha_{0T} \ln Y)^2 + 4[1 - k_T \ln^2 Y)}}{2(1 - k_T \ln^2 Y)}},$$

откуда

$$Y_{max} = e^{\sqrt{\frac{1}{k_T}}} \text{ при } X = \infty.$$

Таким образом, величина Y_{max} в обоих случаях зависит от одного параметра текста k_T. С уменьшением этого параметра значение Y_{max} увеличивается. При $k_T = 0$ $Y_{max} = \infty$.

Для определения параметров α_{0T} и k_T достаточно знать координаты двух точек на кривой роста новых слов. Для этого необходимо найти объем словаря Y_2 в тексте длиной X_2 словоупотреблений (например, при $X_2 = 50 \div 100$ тыс.) и среднее арифметическое значение числа разных слов Y_1, вычисленное по нескольким отрезкам того же текста длиной $X_1 = 5 \div 10$ тыс.).

Величина k_T находится по формуле

$$k_T = \frac{\alpha_{T2} - \alpha_{T_1}}{\ln X_2 - \ln X_1}.$$

Значения α_{T1} и α_{T2} вычисляются по формулам (10) или (12) при заданных значениях X_1, Y_1, X_2, Y_2.

Тогда параметр α_{0T} будет равен

$$\alpha_{0T} = \alpha_{T2} - k_T \ln X_2.$$

В заключение отметим, что параметры a и b распределения Вейбулла при $k - k_T > 0$ не являются постоянными. Расчеты показывают, что при условии, когда $k' = 2{,}3$ $k = 0{,}01$, $k'_T = 0{,}007$ ($u \approx -\frac{1}{4}$), при увеличении длины текста в 10 раз (от 10^5 до 10^6) параметр a уменьшается примерно на 0,014 (от 0,150 до 0,136), а параметр b незначительно увеличивается (примерно на 0,0034 — от 0,347 до 0,3504).

Если положить $b = \text{const}$, то при тех же начальных условиях параметр a уменьшится на 0,0138, а параметр выборки k' — на 0,0002.