МАТЕМАТИЧЕСКИЕ МОДЕЛИ РОСТА СЛОВАРЯ И ИНФОРМАЦИОННЫХ ПОТОКОВ

В.В.Нешитой

При решении различных задач прогнозирования необходимо знать характер зависимости между изучаемыми величинами, например, между числом названий книг и брошюр или заявок на изобретения и временем; числом разных дескрипторов и числом заиндексированных документов; числом разных информационных запросов и числом абонементо-запросов (т.е. с учетом их повторяемости); числом разных слов частотного словаря и объемом выборки и т.д.

Целью настоящей статьи является отыскание уравнений, описывающих различного рода кривые роста из областей информатики и математической лингвистики и решение на их основе некоторых задач.

В качестве моделей кривых роста могут быть использованы различные формулы. Во-первых, это уравнения, найденные для описания кривых роста н о в ы х событий (под "н о в ы м" понимается любое из n разных событий, составляющих полную группу, при первом его появлении от начала испытаний). Во-вторых, это обобщенные плотности и функции распределения, задающие системы непрерывных распределений (при условии снятия ограничений, наложенных на параметры кривых распределения). Все эти модели содержат не более четырех параметров. Использование общих моделей значительно облегчает поиск выравнивающей кривой, хотя и не освобождает исследователя от глубокого анализа статистических данных и знания свойств указанных кривых. Как правило, частная модель содержит меньше параметров, чем их имеется в общей модели. Рассмотрим четыре основные системы кривых роста.

I. Система I кривых роста задается формулой (см.Нешитой В.В., 1986):

$$y = \int_0^n \left(1 - e^{-xp(t)}\right) dt , \qquad (1)$$

где y – математическое ожидание числа разных событий (разных слов), наступающих при x испытаниях (появляющихся в выборке объемом x словоупотреблений); $p(t)$ – непрерывная плотность распределения, аппроксимирующая вероятности p_k (k=I,2,...,

n) разных событий, составляющих полную группу.

При заданной плотности $p(t)$ по формуле (I) может быть рассчитана кривая роста новых событий. Все такие кривые удовлетворяют условиям

$$\left.\begin{array}{l} \frac{dy}{dx} = 1 \quad при \quad x \to 0 \\ \frac{dy}{dx} \to 0 \quad при \quad x \to \infty \end{array}\right\} \qquad (2)$$

В качестве плотности распределения $p(t)$ в формулу (I) может входить любая плотность, заданная на интервале $0 < t \leq n$, где число разных событий n может быть как конечным, так и бесконечным. Например, это могут быть обобщенные плотности

$$p(t) = N t^{\gamma-1}\left(1 - \alpha u t^{\beta}\right)^{\frac{1}{u}-1}, \qquad (3)$$

$$p(y) = \frac{N}{y}(\ln y)^{\gamma-1}\left(1 - \alpha u \ln^{\beta} y\right)^{\frac{1}{u}-1}. \qquad (4)$$

Метод нахождения оценок параметров α, β, γ, u обобщенных плотностей (3), (4) изложен в работе автора (Нешитой В.В., I985).

Недостатком этой системы кривых роста является то, что интеграл (I), как правило, не выражается конечным числом элементарных функций. Кроме того, необходимо знать закон распределения вероятностей разных событий, составляющих полную группу.

2. Система П (а, б) кривых роста.

Система Па кривых роста в общем виде задается уравнением

$$\frac{dy}{dx} = 1 - \bar{F}(y) = 1 - \bar{F}(x), \qquad (5)$$

где $\bar{F}(y), \bar{F}(x)$ - функции распределения вероятностей новых событий.

Пусть $\bar{F}(y), \bar{F}(x)$ задаются формулами (Нешитой В.В., I986)

$$\bar{F}(y) = 1 - \left(1 - \alpha u y\right)^{\frac{1}{u}}, \qquad (6)$$

$$\bar{F}(x) = 1 - \left[1 - \alpha(u-1)x\right]^{\frac{1}{u-1}}. \qquad (7)$$

Тогда система Па кривых роста на основании (5) и (6), (7) будет описываться уравнением

$$y = \frac{1}{\alpha u}\left[1 - (1-\alpha(u-1)x)^{\frac{u}{u-1}}\right], \qquad (8)$$

которое удовлетворяет условиям (2).

Кривые роста, заданные общим уравнением (8), разделяются на типы. К I типу относятся кривые при $0 < u < \infty$; ко II типу – при $u \to 0$; к III типу – при $-\infty < u < 0$.

Оценки параметров α, u кривых I типа рассчитываются по формулам

$$\alpha = \frac{1}{x^2}\sum_{m\geq 1} m^2 y_m - \frac{1}{x}, \quad u = \frac{1}{\alpha n},$$

где $x = \sum_{m\geq 1} m y_m$, $n = \sum_{m\geq 0} y_m$, y_m – число разных событий, наступающих при x испытаниях ровно m раз; n – число разных событий, составляющих полную группу.

Кривая роста относится к I типу, если $y_{m=0} > 0$.

В случае пуассоновского процесса ($u \to 1$) уравнение (8) примет более простой вид

$$y = \frac{1}{\alpha}\left(1 - \frac{1}{e^{\alpha x}}\right).$$

При $u \to 0$ из (8) имеем кривую роста II типа

$$y = \frac{1}{\alpha}\ln(1+\alpha x),$$

параметр α которой находится по методу итераций

$$\alpha_{i+1} = \frac{1}{y}\ln(1+\alpha_i x).$$

В случае кривых роста III типа, а также I и II типов, оценки параметров α, u могут быть найдены упрощенным методом по трем известным из опыта величинам: $x, y, y_{m=1}$, т.е. по объему выборки x, количеству наступивших разных событий y и количеству событий с частотой $m = 1$. При этом тип кривой устанавливается с помощью формулы

$$\frac{x}{y} = \frac{\frac{x}{y_{m=1}} - 1}{\ln\frac{x}{y_{m=1}}},$$

справедливой для кривых II типа. Если эмпирическое отношение $(x/y)^* = x/y$, то кривая роста относится ко II типу.

При $(x/y)^* < x/y$ – к I типу. При $(x/y)^* > x/y$ – к III типу.

Система IIб кривых роста получается из системы IIа введением замены $x = \ln X$, $y = \ln Y$. Тогда

$$\ln Y = \frac{1}{\alpha u}\left[1 - \left(1 - \alpha(u-1)\ln X\right)^{\frac{u}{u-1}}\right] \qquad (9)$$

или в дифференциальной форме

$$\frac{dY}{dX} = \frac{Y}{X}\left(1 - \alpha u \ln Y\right)^{\frac{1}{u}} = \frac{Y}{X}\left[1 - \alpha(u-1)\ln X\right]^{\frac{1}{u-1}} \qquad (10)$$

Дифференциальное уравнение (10) при различных значениях параметра u дает ряд формул для описания кривых роста новых событий (например, новых слов в тексте), однако удобными для практического использования являются те из них, которые позволяют в явном виде выражать параметр α через переменные X, Y. Это формулы при $u = I/2$, $u = -I$ (см. табл.I).

Чтобы иметь более широкий набор кривых роста (с равным интервалом $\Delta u = 0{,}75$), на основании формул с параметрами $u = I/2$, $u = -I$ были получены еще две формулы, которые соответствуют значениям $u \approx I{,}25$, $u \approx -0{,}25$. Они тоже приведены в табл.I.

Проверка кривых роста новых слов в текстах показала, что параметр α не является постоянной величиной, однако эмпирические точки в системе координат $(\ln X; \alpha)$ ложатся на прямую

$$\alpha = \alpha_0 + \kappa \ln X \,. \qquad (11)$$

Параметры α_0, κ являются параметрами текста и легко находятся по графику.

Эти же формулы оказались пригодными и для описания кривой роста новых слов в случайной выборке, но параметры выборки не совпадают с параметрами текста $(\alpha_{0в} < \alpha_{0т}, \kappa_в > \kappa_т)$. Примером такой кривой является кривая роста числа разных дескрипторов $Y = f(X)$, где $X = Dh$ – число использованных дескрипторов при индексировании D документов; h – среднее число дескрипторов в одном документе.

Система Пб кривых роста новых слов в выборке Таблица I

Параметр u	Уравнение кривой роста	$\alpha = \alpha_0 + \kappa \ln X$	Оценки параметров выборки α_0, κ
$u \approx 1{,}25$	$Y = X^{1/X^{\alpha/2}}$	$\alpha = \frac{2}{\ln X} \ln \frac{\ln X}{\ln Y}$	$\kappa = \frac{2}{\ln^2 X}\left(1 - \ln \frac{\ln X}{\ln Y} - \frac{Y_{m=1}}{Y}\frac{\ln X}{\ln Y}\right)$ $\alpha_0 = \alpha - \kappa \ln X$
$u = 0{,}5$	$Y = X^{1/(1+\frac{\alpha}{2}\ln X)}$	$\alpha = \frac{2}{\ln X}\left(\frac{\ln X}{\ln Y} - 1\right)$	$\kappa = \frac{2}{\ln^2 Y}\left[\left(\frac{\ln Y}{\ln X}\right)^2 - \frac{Y_{m=1}}{Y}\right]$ $\alpha_0 = \alpha - \kappa \ln X$
$u \approx -0{,}25$	$Y = X^{1/\sqrt{1+\alpha \ln X}}$	$\alpha = \frac{1}{\ln X}\left[\left(\frac{\ln X}{\ln Y}\right)^2 - 1\right]$	$\alpha_0 = \frac{2}{\ln X}\left[\frac{Y_{m=1}}{Y}\left(\frac{\ln X}{\ln Y}\right)^3 - 1\right]$ $\kappa = \frac{1}{\ln^2 Y} - \frac{1}{\ln^2 X} - \frac{\alpha_0}{\ln X}$
$u = -1$	$Y = e^{\frac{1}{\alpha}\left(\sqrt{1+2\alpha \ln X} - 1\right)}$	$\alpha = \frac{2}{\ln Y}\left(\frac{\ln X}{\ln Y} - 1\right)$	$\kappa = \frac{2}{\ln^2 Y}\left[1 - \frac{Y_{m=1}}{Y}\left(2\frac{\ln X}{\ln Y} - 1\right)\right]$ $\alpha_0 = \alpha - \kappa \ln X$

Оценки параметров выборки могут быть найдены по трем величинам: $X, Y, Y_{m=1}$. При этом используется формула В.М.Калинина (1964, с. 247)

$$\frac{d^m y}{dx^m} = (-1)^{m+1} \frac{m!}{x^m} y_m ,$$

которая при $m = 1$ дает

$$\frac{dy}{dx} = \frac{y_{m=1}}{x} . \qquad (12)$$

На основании (12) и соответствующих уравнений кривой роста с учетом равенства (11) были найдены оценки параметров выборки, которые приведены в табл. 1.

Из приведенных в табл.1 формул лишь две последние ($u \approx -0{,}25$, $u = -1$) удовлетворяют условиям (2), при этом от параметра κ зависит наибольший объем словаря. Кривая роста с параметром $u \approx -0{,}25$ при $X \to \infty$ дает: $Y_{max} = e^{\sqrt{1/\kappa}}$. Для кривой с параметром $u = -1$ величина $Y_{max} = e^{\sqrt{2/\kappa}}$. При $\kappa \to 0$ в обоих случаях $Y_{max} \to \infty$. Следовательно, параметр κ может служить показателем лексического разнообразия (богатства) текста. При равных значениях κ бóльшим лексическим разнообразием обладает кривая с меньшим значением параметра α_0.

Первые две формулы ($u \approx 1{,}25$, $u = 0{,}5$) описывают кривые, которые вначале возрастают, затем убывают (при весьма больших значениях X). Следовательно, их можно использовать в качестве моделей кривых роста новых слов при ограниченных значениях X ($X < 10^6 \div 10^7$).

Формулы системы Пб кривых роста содержат всего два параметра, оценки которых легко находятся графическим методом либо рассчитываются по трем величинам $X, Y, Y_{m=1}$, при этом не требуется знание закона распределения разных слов по частоте их употребления в текстах.

3. Система Ш кривых роста задается формулами

$$y = NF(t), \qquad (13)$$

$$y = N[1 - F(t)], \qquad (14)$$

где N – некоторый параметр, причем, $N = y_{max}$. Свойства кривых роста этой системы полностью определяются свойствами функции распределения $F(t)$.

Формула (I3) может описывать, например, количество разных статей по определенной теме, опубликованных в первых t журналах, при условии, что последние упорядочены по убыванию количества таких статей, а также количество заболеваний при эпидемиях за время t от начала эпидемии.

4. Система IУ (а, б, в) кривых роста

Система IУ кривых роста строится на основе обобщенных распределений. Вводя другие обозначения переменных и освобождаясь от ограничений, накладываемых на кривые распределения, можем записать следующие уравнения для описания различного рода кривых роста

$$\text{IУ}a: \quad y = Nx^{\gamma-1}(1-\alpha u x^{\beta})^{\frac{1}{u}-1}, \tag{I5}$$

$$\text{IУ}б: \quad y = Ne^{\gamma x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}; \tag{I6}$$

$$\text{IУ}в: \quad \ln y = N(\ln x)^{\gamma-1}(1-\alpha u \ln^{\beta} x)^{\frac{1}{u}-1}. \tag{I7}$$

Система IУ кривых роста является наиболее широкой системой, включающей кривые самой разнообразной формы. Она включает также некоторые кривые, принадлежащие другим системам.

4.I. Система IУа кривых роста новых слов

Проверка показала, что общая формула (I5) может достаточно точно описывать кривые роста новых слов при $\gamma = 2$, $N = I$:

$$y = x(1-\alpha u x^{\beta})^{\frac{1}{u}-1}. \tag{I8}$$

Наиболее подходящей оказалась формула при $u = -I$ $(0 < \beta < I/2)$

$$y = \frac{x}{(1+\alpha x^{\beta})^2}, \tag{I9}$$

которая применима при $(I0^3 \div I0^4) < x < (I0^7 \div I0^8)$ словоупотреблений. Оценки параметров текста находятся из уравнения прямой

$$\ln\left(\sqrt{\frac{x}{y}} - 1\right) = \ln\alpha + \beta \ln x. \tag{20}$$

Оценки параметров выборки можно найти по трем величинам: $x, y, y_{m=1}$

$$\beta = \frac{1 - \frac{y_{m=1}}{y}}{2\left(1 - \sqrt{\frac{y}{x}}\right)} = \frac{\sqrt{\frac{x}{y}}\left(1 - \frac{y_{m=1}}{y}\right)}{2\left(\sqrt{\frac{x}{y}} - 1\right)}, \tag{21}$$

$$\alpha = \frac{1}{x^{\beta}}\left(\sqrt{\frac{x}{y}} - 1\right). \tag{22}$$

4.2. Система IУв кривых роста новых слов

Уравнение (I7) хорошо описывает кривую роста новых слов при $\gamma = 2$, $N = I$, $(-I \leq u < 0)$, т.е. система IУв кривых роста задается общей формулой

$$\ln Y = \ln X \left(1 - \alpha u \ln^{\beta} X\right)^{\frac{1}{u} - 1}. \tag{23}$$

Пусть $u \to 0$. Тогда из (23) получим уравнение

$$\ln Y = \frac{\ln X}{e^{\alpha \ln^{\beta} X}} \quad (X, Y \geq 1), \tag{24}$$

которое может быть приведено к прямой

$$\ln \ln \frac{\ln X}{\ln Y} = \ln \alpha + \beta \ln \ln X. \tag{25}$$

Параметры α, β связного текста находятся путем построения по опытным значениям X_i, Y_i графика зависимости (25).

Параметры выборки можно рассчитать по трем величинам: $X, Y, Y_{m=1}$

$$\beta = \frac{1}{\ln \frac{\ln X}{\ln Y}}\left(1 - \frac{Y_{m=1}}{Y} \frac{\ln X}{\ln Y}\right), \tag{26}$$

$$\alpha = \frac{1}{\ln^{\beta} X} \ln \frac{\ln X}{\ln Y}. \tag{27}$$

Пусть далее $u = -I$. Тогда из (23) получим

$$\ln Y = \frac{\ln X}{\left(1 + \alpha \ln^{\beta} X\right)^2}, \tag{28}$$

откуда

$$\ln\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right) = \ln \alpha + \beta \ln \ln X. \tag{29}$$

Оценки параметров выборки равны

$$\beta = \frac{1 - \frac{Y_{m=1}}{Y}\frac{\ln X}{\ln Y}}{2\left(1 - \sqrt{\frac{\ln Y}{\ln X}}\right)}, \qquad (30)$$

$$\alpha = \frac{1}{\ln^{\beta} X}\left(\sqrt{\frac{\ln X}{\ln Y}} - 1\right). \qquad (31)$$

Достоинством формул (24), (28) является то, что они хорошо работают практически от начала координат до весьма больших значений X ($X = 10^7 \div 10^8$ словоупотреблений). Однако при дальнейшем увеличении X поведение этих кривых не соответствует поведению кривых роста новых событий.

Опыт показывает, что параметры выборки в приведенных выше формулах зависят как от типа текстов, на основе которых построена данная выборка, так и от ее объема. В связи с этим полученные формулы остаются справедливыми только в пределах заданной выборки, т.е. их нельзя использовать для экстраполяции кривой роста новых слов на выборку большего объема.

В то же время параметры текста не зависят от объема текста (при условии его лексической однородности), что позволяет прогнозировать объем словаря с ростом объема текста.

Проверим работу формул 10 и 1У (а, в) кривых роста новых слов в выборке. Восстановим кривую $y=f(x)$ на основе опытных данных по точным формулам, а также рассчитаем ее по приближенным формулам, при этом параметры выборки α_0, k , а также α, β определим по трем величинам: $X, Y, Y_{m=1}$.

Для восстановления кривой роста новых слов в выборке воспользуемся формулой В.М.Калинина (1964, с.247)

$$y = Y - \sum_{m \geq 1}\left(1 - \frac{x}{X}\right)^m Y_m , \qquad (32)$$

где Y – число разных слов в выборке объемом X словоупотреблений; Y_m – число слов с частотой m в выборке X ; y – ожидаемое среднее число разных слов в подвыборке произвольного объема x $(x < X)$.

Формулой (32) удобно пользоваться при $x/X \geq 0{,}1$.

При $x/X < 0{,}1$ целесообразно воспользоваться формулой (1), в которую вместо плотности $\rho(t)$ следует подставить относительные частоты слов $\rho^*_z = m^*_z / X$, где z – ранг слова в

частотном словаре. Тогда формула (I) примет вид

$$y = \int_0^n (1 - e^{-xp_z^{\mu}})\,dz. \qquad (I')$$

Интегрирование осуществляется численным методом по формуле прямоугольников.

Таблица 2

Параметры выборок, рассчитанные по данным двух частотных словарей (ЧС)

Параметры	Пб (табл. I			IУа	IУв	
	$\mu \approx$ I,25	μ =0,5	$\mu \approx$ -0,25	μ =- I	$\mu \to 0$	μ =-I
ЧС немецкого языка						
α_0	0,02029	0,0I622	0,00996			
κ	0,0007485	0,00I287	0,002017			
α				004223	0,005719	0,002363
β				0,3005	I,374I	I,46635
ЧС русского языка						
α_0	-0,00I633	-0,0I485	-0,03344			
κ	0,002935	0,004306	0,00615			
α				0,0I45I	0,00I262	0,000466
β				0,4084	2,04I8	2,I8324

В таблице 3 (столбец 2) приведены результаты расчетов по точным формулам (I') и (32) на основе опытных данных частотного словаря немецкого языка и по приближенным формулам систем кривых роста Пб, IУа, IУв (столбцы 3-8). Объем выборки здесь равен X =I09I077, объем словаря Y =258I73 и количество одноразовых слов $Y_{m=1}$ =I26862 (*Meier* H., I964). Параметры выборки для каждой аппроксимирующей кривой приведены в табл. 2.

Аналогичные расчеты выполнены по данным "Частотного словаря русского языка" (под ред.Л.Н.Засориной, I977). Здесь X =I056382, Y =39268, $Y_{m=1}$ =I3379.

Из табл. 3 видно, что в первом случае все формулы достаточно точно описывают кривую роста новых слов в выборке, при этом наименьшую точность показала формула при $\mu \approx$ I,25 (система Пб). Во втором случае менее точными оказались две формулы из той же системы кривых роста: при μ = 0,5, $\mu \approx$ -0,25.

Таблица 3

Кривые роста новых слов, восстановленные по двум ЧС и рассчитанные по формулам* систем кривых роста Пб, IУа, IУв

Объем выборки X	Объем словаря по ЧС Y	IIб			IVа	IVв	
		$u\approx$I,25	$u=0,5$	$u=-0,25$	$u=-$I	$u\to 0$	$u=-$I
		ЧС немецкого языка					
I09I	625	600	6I9	646	603	626	636
2000	I0I9	979	I0I0	I054	I000	I0I9	I037
5000	2II8	2004	2066	2I56	2092	2082	2II6
I09II	3878	3609	37I6	3866	38I9	3787	3794
30000	–	7503	7699	7965	8009	7725	7828
I00000	I7800	I7II8	I7462	I79I2	I8207	I7485	I7663
300000	–	34735	352I5	358I5	36459	35220	35468
I09I078	77870	75440	75952	76545	77420	759I3	76I78
3273233	I408I0	I39063	I39377	I39670	I40230	I39293	I39474
I09I0777	258I73	258I67	258232	258I73	258I73	258083	258I77
		ЧС русского языка					
I000	625	650	7I3	–	645	646	659
3000	I5I0	I534	I674	–	I57I	I526	I556
I0000	3637	36I8	3889	4274	3790	3604	3667
30000	7384	7296	7682	8I93	7670	7275	7374
I05638	I48I5	I4703	I5095	I5569	I52I5	I4683	I479I
2II276	20700	20597	20890	2I227	2I028	20582	20666
528I9I	30495	30455	30544	30645	30600	30450	30477
I056382	39268	39268	39269	39268	39268	39269	39269

* IIб

$u\approx 1{,}25: \; Y = X^{1/X^{\alpha/2}}$

$u = 0{,}5: \; Y = X^{1/(1+\frac{\alpha}{2}\ln X)}$

$u\approx -0{,}25: \; Y = X^{1/\sqrt{1+\alpha\ln X}}$

$(\alpha = \alpha_0 + \kappa \ln X)$

IV

а) $u=-1: \quad y = \dfrac{x}{(1+\alpha x^{\beta})^2}$

в) $\begin{cases} u\to 0: \; \ln Y = \dfrac{\ln X}{e^{\alpha \ln^{\beta} X}} \\ u=-1: \; \ln Y = \dfrac{\ln X}{(1+\alpha \ln^{\beta} X)^2} \end{cases}$

Полученные результаты свидетельствуют о том, что параметр α изменяется от выборки к выборке. А это значит, что для более точного описания кривой роста новых слов в выборке необходимо использовать общие формулы (9), (I8), (23), каждая из которых (с учетом равенства (II) в случае формулы (9)) содержит три параметра. Для приближенных же расчетов удовлетворительные результаты дают рассмотренные выше формулы, имеющие всего по два параметра.

Возможность приведения этих формул к прямой оказалась весьма полезной для решения различных задач. Рассмотрим некоторые из них.

4.3. Оценка степени аналитичности языка

Показателем степени аналитичности языка принято считать коэффициент, равный отношению количества лексем к количеству словоформ частотного списка (Пиотровский Р.Г. и др., I962). Степень аналитичности языка тем выше, чем ближе этот показатель к единице.

Существенным недостатком этого показателя является его зависимость от объема текста. Знание аналитической зависимости между объемами словаря и текста позволяет по-другому измерять степень аналитичности языка.

Пусть кривая роста новых слов в тексте описывается формулой (I9). Величина параметра β зависит от выбора единицы подсчета количества разных слов, в качестве которой может быть принята словоформа (СЛ) или лексема (Л). При этом для одного и того же текста $\beta_{л} > \beta_{сл}$, в то время как $\alpha_{л} \approx \alpha_{сл}$. Последнее равенство позволяет ввести показатель степени аналитичности языка, который не зависит от объема текста:

$$\Delta\beta_{a} = \frac{v_{л} - v_{сл}}{\ln x}, \qquad (33)$$

где $v_{л}, v_{сл}$ рассчитываются по формуле

$$v = \ln\left(\sqrt{\frac{x}{y}} - 1\right). \qquad (34)$$

Чем ближе $\Delta\beta_{a}$ к нулю, тем выше степень аналитичности языка.

В табл.4 приведены значения $\Delta\beta_{a}$ для четырех языков, рассчитанные по формулам (33), (34) для текстов по электронике.

Полученные результаты позволяют осуществить переход от кривой роста новых словоформ в тексте к кривой роста новых лексем. Для этого достаточно воспользоваться формулами

$$\alpha_{л} = \alpha_{сл}\ ;\ \beta_{л} = \beta_{сл} + \Delta\beta_{a}\ .$$

Таблица 4

Степень аналитичности языков (тексты по электронике)

Язык, источник	Объем текста x	Объем словаря $y_{cл}$	Объем словаря $y_л$	$\Delta\beta_a$
Русский (Калинина,Е.А.,I968)	200894	2I468	6826	0,0627
Румынский (Ешан Л.И.,I966)	200000	I4292	5708	0,0479
Французский (Кочеткова В.К., Скрелина Л.М., I968)	I00000	8I08	4527	0,0336
Английский (Алексеев П.М., I968)	200000	I0582	7I60	0,0202

4.4. Показатель степени связности слов в лексически однородном тексте

Так как в связном тексте лексико-грамматические связи накладывают определенные ограничения на сочетаемость слов, то естественно предположить (и это подтверждается опытными данными), что число разных слов в отрезке сплошного текста в среднем будет меньше, чем в случайно составленной выборке равного объема, взятой из достаточно большой совокупности лексически однородных текстов.

Эту разницу в объемах словарей можно использовать для оценки степени связности слов в тексте, причем, она оказывается независимой от объема текста.

Пусть для некоторого целого произведения из опыта известны объемы – всего текста X, словаря Y, одноразовых слов $Y_{m=1}$, а также несколько промежуточных точек $(x_i ; y_i)$ на кривой роста новых слов. По этим данным найдем параметры текста α_T, β_T путем построения графика зависимости (20). Для этого же произведения можно рассчитать параметры выборки $\alpha_в$, $\beta_в$ по формулам (2I), (22). Эти параметры будут относиться к такой кривой роста, которая получилась бы при случайном отборе словоупотреблений из данного произведения и подсчете количества разных слов.

Многочисленные расчеты показали, что параметры α_T и $\alpha_в$ находятся в отношении

$$\frac{\alpha_T}{\alpha_в} \approx 2, \qquad (35)$$

которое может быть использовано в качестве показателя степени связности слов в тексте.

Параметры $\beta_в$ и $\beta_т$ связаны соотношением

$$\beta_т = \beta_в - \frac{\ell g 2}{\ell g X} . \qquad (36)$$

формулы (35), (36) позволяют по параметрам выборки (которые легко вычислить по трем величинам $X, Y, Y_{m=1}$) найти параметры текста.

4.5. Оценка лексической близости двух связных текстов. Автоматическая классификация текстов

Пусть кривая роста новых слов в связном тексте описывается формулой (I9). Объединим два равных по объему текста с одинаковыми параметрами α, β и исследуем поведение обоих параметров этого объединенного текста. Рассмотрим два крайних случая.

Случай I. Параметры α, β объединенного текста равны соответствующим параметрам объединяемых текстов, а его словарь содержит такое же количество разных слов, сколько их было бы в каждом из объединяемых текстов при удвоенной его длине. В этом случае степень лексической близости двух текстов (будем измерять ее некоторым показателем ς) равна единице, а точка с координатами $(\ell n\, x_{12};\ v_{12\varsigma=1})$ лежит на прямой I (см.рис.I), поскольку

$$v_{12\varsigma=1} = \ell n\left(\sqrt{\frac{2x_1}{y_{12\varsigma=1}}} - 1\right) = \ell n\alpha + \beta \ell n 2x_1 = v_1 + \beta \ell n 2 .$$

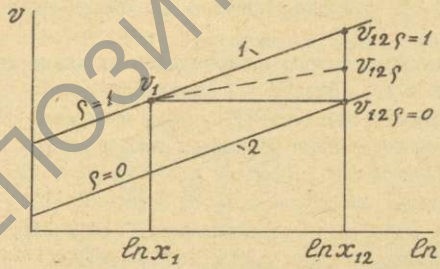

Рис. I. Графическое определение показателя ς

Случай 2. Словари обоих текстов не содержат общих слов. Это значит, что степень лексической близости таких текстов равна нулю ($\varsigma = 0$). В этом случае объем словаря объединенного текста равен удвоенному объему словаря одного из объединяемых текстов, а величина

$$v_{12\varsigma=0} = \ln\left(\sqrt{\frac{2x_1}{2y_1}} - 1\right) = v_1 = \ln\alpha + \beta \ln x_1,$$

т.е. прямая 2 параллельна прямой I и расположена ниже ее на расстоянии $\beta \ln 2$.

Из рис. I видно, что величина $v_{12\varsigma}$, характеризующая объем словаря реального объединенного текста, ограничена $v_{12\varsigma=0} < v_{12\varsigma} < v_{12\varsigma=1}$. Следовательно, показатель ς можно измерять отношением

$$\varsigma = \frac{v_{12\varsigma} - v_{12\varsigma=0}}{v_{12\varsigma=1} - v_{12\varsigma=0}} = \frac{v_{12\varsigma} - v_{12\varsigma=0}}{\beta \ln 2}. \qquad (37)$$

Из условия параллельности прямых I и 2 следует, что ς не зависит от объема текстов x_i, но оба они должны быть равными между собой.

Отметим, что для двух сравниваемых текстов отношение $y_{12\varsigma=0} / y_{12\varsigma=1}$ не может превосходить определенного значения, которое при $0 < x_1 < \infty$ заключено на интервале $1 < y_{12\varsigma=0} / y_{12\varsigma=1} < 4^{\beta}$. Например, для "Частотного словаря русского языка" при $\beta = 0{,}4084$ (табл.2, столбец 5) это отношение не должно превышать величины $4^{\beta} = 1{,}7615$. Действительно, при $x_1 = 52819$ по данным табл. 3 (столбец 2) имеем: $30495 \cdot 2/39268 = 1{,}5532$.

Параметр β, входящий в формулу (37), может быть надежно определен лишь по тексту достаточно большого объема. В случае коротких текстов показатель ς можно оценить проще, если объем объединенного текста принять равным $x_{12} = x_1/2 + x_2/2$. Тогда для вычисления показателя ς достаточно будет найти значения следующих величин:

$$y_{12\varsigma=0} = y_1(x_1/2) + y_2(x_2/2); \; y_{12\varsigma} = y(x_{12});$$

$$y_{12\varsigma=1} = \frac{1}{2}\left[y_1(x_1) + y_2(x_2)\right],$$

где $y_1(x_1/2)$ – количество разных слов среди половины словоупотреблений первого текста (например, стоящих только на четных местах); $y_2(x_2/2)$ – то же для второго текста; $y(x_{12})$ – количество разных слов среди половины словоупотреблений первого и второго текстов; $y_1(x_1)$, $y_2(x_2)$ – количество разных слов соответственно в первом и втором текстах.

При небольших размерах текстов ($x < 40000$) кривая роста новых слов достаточно точно описывается формулой

$$y = \frac{x}{1 + \alpha x^{\beta}},$$

которая следует из (I8) при $u \to -\infty, \alpha u > 0$. Ее можно привести к виду

$$v = \ln\left(\frac{x}{y} - 1\right) = \ln\alpha + \beta \ln x . \qquad (38)$$

Для определения показателя ζ (см. формулу (37)) необходимо по формуле (38) вычислить значения $v_{12\zeta}$ при заданных x_{12} и $y_{12\zeta}$. При этом получим оценку ζ, не зависящую от объема сравниваемых текстов (но оба текста должны быть равными между собой).

Все операции по вычислению ζ легко автоматизировать. Задавая пороговое значение этого показателя, можно классифицировать тексты (документы) по тематическим группам, включая в одну группу близкие по содержанию тексты.

4.6. Система IУа кривых роста информационных потоков

Одной из таких кривых, принадлежащих данной системе, является кривая роста (во времени x) числа названий книг и брошюр. Ее свойства отличаются от свойств кривых роста новых слов. При этом, как показывает анализ опытных данных, рост числа названий книг и брошюр, заявок на изобретения и т.д. может быть либо линейным, либо близким к экспоненте. Следовательно, общее уравнение для описания этих кривых роста может быть записано в виде (при $y = I$, $\alpha < 0$, $0 < u < I$)

$$y = y_0 \left(1 - \alpha u x^{\beta}\right)^{\frac{1}{u} - 1}, \qquad (39)$$

где y_0 – начальное значение y при $x = 0$.

Для нахождения параметров α, β, u преобразуем (39) к виду

$$\ln\left[\left(\frac{y}{y_0}\right)^{\frac{u}{1-u}} - 1\right] = \ln\alpha u + \beta \ln x . \qquad (40)$$

Значение параметра u должно быть подобрано таким, чтобы график зависимости (40), построенный по опытным значениям y, x, представлял собой прямую. Оценки параметров α, β могут быть определены по графику полученной прямой.

В случае, если ни при каких значениях параметра u ($0 < u < I$) прямая не получается, то это свидетельствует о том, что формула (39) в данном случае не работает и следует рассмотреть кривую другого типа, например,

$$y = y_0 e^{\alpha x^{\beta}} . \qquad (4I)$$

Оценки параметров α, β находятся на основании уравнения прямой

$$\ln\ln\frac{y}{y_0} = \ln\alpha + \beta\ln x \qquad (42)$$

при $\alpha > 0$, либо прямой

$$\ln\ln\frac{y_0}{y} = \ln\alpha + \beta\ln x \qquad (43)$$

при $\alpha < 0$. Исследования показывают, что кривая роста (41) при $\beta < 0$ имеет горизонтальную асимптоту $y = y_0$, которую также следует оценить по опытным данным. Для этого представим формулу (41) в виде

$$y_{nx} = a y_x^b , \qquad (44)$$

где y_{nx} может обозначать число наступивших разных событий в выборке объемом nx (за время nx). Параметры a, b находятся из уравнения прямой

$$\ln y_{nx} = \ln a + b\ln y_x . \qquad (45)$$

Тогда оценки параметров в формуле (41) будут равны

$$y_0 = a^{1/(1-b)} \quad (b \neq 1), \qquad (46)$$

$$\beta = \frac{\ln b}{\ln n} , \qquad (47)$$

$$\alpha = \frac{1}{s}\sum_{i=1}^{s}\frac{1}{x_i^\beta}\ln\frac{y_i}{y_0} . \qquad (48)$$

Если $b = 1$, то кривая роста задается формулой $y = \alpha x^\beta$.

Формулы (41)-(45) хорошо описывают кривые роста числа названий книг и брошюр, кривые роста производительности труда и произведенного национального дохода (Нешитой В.В., 1984) и некоторые другие кривые. Важной характеристикой этих кривых является темп роста

$$q_i = \frac{y_i}{y_{i-1}} = e^{\alpha(x_i^\beta - x_{i-1}^\beta)} . \qquad (49)$$

Из (49) следует, что при $\beta = 1$ $q_i = e^\alpha = const$, т.е. темп роста не зависит от времени x_i и равен темпу роста показательной функции $y = e^{\alpha x}$. При $\beta > 1$ величина q_i с ростом x_i растет, а при $\beta < 1$ – уменьшается, т.е. параметр β является показателем ускорения или замедления темпа роста кривой (41). При $x_1 = 1$,

$x_0 = 0$ из (49) имеем $q_1 = y_1/y_0 = e^{\alpha}$, откуда $\alpha = \ln(y_1/y_0) = \ln q_1$, т.е. величина α представляет собой натуральный логарифм темпа роста кривой на начальном отрезке времени, равном единице. Поскольку $q_1 = e^{\alpha} = 1 + \frac{\alpha}{1!} + \frac{\alpha^2}{2!} + \ldots$, то при малых α ($\alpha < 0{,}1$) имеем $q_1 \approx 1 + \alpha$, откуда $\alpha \approx q_1 - 1$.

Проверим работу формул (41), (44) на практическом примере. Воспользовавшись опытными данными (Михайлов А.И. и др., 1976, с.212, 213), приведенными в табл.5, столбцы 1,3, найдем параметры выравнивающей кривой роста числа названий книг и брошюр, изданных в 16 странах за 1958-1972 г.г. Пусть эмпирическая кривая роста описывается формулой (41), которую представим в виде

$$\frac{y}{y_0} = e^{\alpha x^{\beta}}, \qquad (50)$$

где y_0 = 216696 (количество книг и брошюр на конец 1958 г.). Приведем ее к уравнению прямой

$$\left(\ln \frac{y}{y_0}\right)_{nx} = b\left(\ln \frac{y}{y_0}\right)x. \qquad (51)$$

Таблица 5

Число названий книг и брошюр (y), изданных в 1958-1972 г.г. в 16 странах (Михайлов А.И. и др.,1976)

Год	x	y	$\frac{y}{y_0}$	$\ln \frac{y}{y_0}$	$\left(\frac{y}{y_0}\right)^{расч.}$
1958	0	216696	1	0	1
1959	1	224201	1,0346	0,0341	1,0378
1960	2	237837	1,0976	0,0931	1,0824
1961	3	245382	1,1324	0,1243	1,1314
1962	4	257619	1,1889	0,1730	1,1843
1963	5	269942	1,2457	0,2197	1,2410
1964	6	282892	1,3055	0,2666	1,3016
1965	7	288422	1,3310	0,2859	1,3662
1966	8	323388	1,4924	0,4004	1,4350
1967	9	333848	1,5406	0,4322	1,5081
1968	10	336975	1,5551	0,4415	1,5858
1969	11	347699	1,6045	0,4728	1,6682
1970	12	385570	1,7793	0,5762	1,7557
1971	13	390247	1,8009	0,5883	1,8486
1972	14	396661	1,8305	0,6046	1,9471

Если построить по эмпирическим значениям величин $\ln(y/y_0)$ график зависимости (5I) при $n = 2$, т.е при всех возможных парах значений $(x, 2x)$: I,2; 2,4; 3,6;...; 7,I4, - то убедимся, что точки действительно рассеиваются вдоль прямой (5I), проходящей через начало координат. Ее угловой коэффициент $\beta = 2,I3537$. Следовательно, согласно (47), параметр $\beta = I,0945$. Параметр α, вычисленный по формуле (43), равен $\alpha = 0,03709$.

Таким образом, выравнивающая кривая роста может быть записана в виде

$$y = 216696\, e^{0,03709(x-1958)^{1,0945}}.$$

Поскольку параметр $\beta > I$, то темп роста числа названий книг и брошюр со временем увеличивается.

Л И Т Е Р А Т У Р А

Алексеев П.М. Частотный словарь английского подъязыка электроники // Статистика речи. - Л., I968. - с. I5I-I6I.

Ешан Л.И. Опыт статистического описания научно-технического стиля румынского языка (на материале текстов по радиоэлектронике): Автореф.дис... канд.филол.наук. - Л., I986. - I6 с.

Калинин В.М. Некоторые статистические законы математической лингвистики // Проблемы кибернетики. Вып. II. - М., I964 - с. 246-255.

Калинина Е.А. Изучение лексико-статистических закономерностей на основе вероятностной модели // Статистика речи. - Л.:Наука, I968. - с. 64-I07.

Кочеткова В.К., Скрелина Л.М. Частотный словарь французского подъязыка электроники // Статистика речи. - Л., I968. - с. I62-I70.

Михайлов А.И., Чёрный А.И., Гиляревский Р.С. Научные коммуникации и информатика. - М.: Наука, I976. - 436 с.

Нешитой В.В. Система непрерывных распределений как экономико-математическая модель// Проблемы макроэконометрического моделирования и прогнозирования: Тезисы докладов респ.конф. / АН Латв.ССР; ин-т эк-ки. - Рига: Зинатне,I984. - с. 237-239.

Нешитой В.В. Исследование ранговых распределений // НТИ. Сер.2. - I985. - № 2. - с. I6-20.

Нешитой В.В. Ранжирование слов по степени семантической нагрузки// НТИ. Сер.2. - I986. - № 4. - с. 20-25.

Пиотровский Р.Г., Алексеев П.М., Чернядьева Е.А. Статистика речи и закономерности языка// Тез.докл.межвуз.конф. на тему "Язык и речь" (27 ноября - I декабря)/ ЛГПИИ. - М.,I962. - с. 57-59.

Частотный словарь русского языка. / Под ред.Л.Н.Засориной. - М.: Русский язык, I977. - 936 с.

Meier H. Deutsche Sprachstatistik. - Hildesheim, 1964.

MATHEMATICAL MODELS OF THE DICTIONARY EXPANSION AND DATA FLOWS

V.V. Neshitoi

S u m m a r y

Four systems of the mathematical models for description of different growth curves are proposed, methods for parameter evaluation are given.

Indices of the analytical character of the language, word cohesion in a lexically uniform text and lexical similarity of two connected texts are introduced, these indices being independent of the text size.