Особое место занимает ритмическое решение в частях Мессы А.Безенсон. В основе всей ритмической организации лежит ритмика древнейшего григорианского хорала, основанная на нерегулярном чередовании длинных и коротких длительностей. Особая роль принадлежит выдержанному органному пункту, который может содержать как один звук, так и созвучие («Credo»). Сильное воздействие имеют сочетание различных ритмических фигур с отдельными фрагментами текста («Credo»), а также тончайшие ритмические нюансы (небольшое удлинение или уменьшение длительностей, легкие акценты внутри групп из кратких звуков).

В. В. Нешитой,
доктор технических наук, профессор,
профессор кафедры информационных ресурсов

ДВОЙСТВЕННАЯ ПРИРОДА РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

Ранговым называют всякое статистическое распределение, в котором однородные элементы упорядочены по убыванию (точнее, по невозрастанию) частоты их встречаемости в некоторой выборке. Так, для слов частотного словаря это ранговое их распределение в списке по убывающим частотам – абсолютным или относительным. Порядковый номер слова в этом списке называется его рангом, обозначается символом r. Сумма относительных частот всех разных слов в некотором тексте всегда равна единице:

$$\sum_{r=1}^{n} p_r = 1.$$

Накопленная относительная частота r первых слов представляет собой функцию распределения

$$F(r) = P(R \le r),$$

т. е. сумму относительных частот слов от 1 до r включительно. Она показывает полноту покрытия текста словарем объемом r. Это очень важная характеристика. Чтобы читать текст (например, художественную прозу на иностранном языке) и понимать

содержание без частого обращения к словарю, необходимо помнить перевод не менее 16–20 тысяч наиболее частых слов. Накопленная относительная частота этих слов будет в пределах 0,93–0,95, хотя словарь генеральной совокупности текстов художественной прозы составляет около 100 тысяч разных слов (для русского языка) [1].

Кривая роста новых слов в связном тексте

Под новым словом будем понимать любое из разных слов (частотного словаря большого объема) при первом его употреблении в тексте.

Кривая роста новых слов в связном тексте несколько отличается от кривой роста разных слов в случайно составленной выборке прежде всего меньшей скоростью пополнения словаря, поскольку в связном тексте на сочетаемость между словами накладываются семантические и лексические ограничения.

Для описания ранговых распределений слов частотного словаря автором данной статьи был предложен трехпараметрический закон Вейбулла с функцией и плотностью распределения [2],

$$F(z) = 1 - \frac{1}{e^{\alpha[(z+1)^{\beta} - e^{-cz}]}}; \qquad (1)$$

$$p(z) = \frac{\dfrac{1}{(z+1)^{1-\beta}} + \dfrac{\alpha c}{e^{cz}}}{e^{\alpha[(z+1)^{\beta} - e^{-cz}]}}, \qquad (2)$$

где z – ранг слова в частотном словаре, а параметр c=0,01 – 0,1. Приведенные формулы при достаточно больших значениях z преобразуются к виду двухпараметрического распределения Вейбулла:

$$F(z) = 1 - \frac{1}{e^{\alpha z^{\beta}}}; \quad p(z) = \frac{\alpha\beta}{z^{1-\beta} e \alpha^{\alpha z^{\beta}}}.$$

Опытная проверка показывает, что это распределение справедливо уже при z > 100–500 слов.

Двухпараметрическое распределение Вейбулла впервые было применено Г. Белоноговым [3] для описания статистического рангового распределения слов в русской письменной речи.

На основании известного закона распределения вероятностей разных слов частотного словаря, который задан плотностью (2), и формулы автора [2]

$$z = \int_0^{\infty} \left(1 - e^{-xp(z)}\right) dz \qquad (3)$$

при заданных значениях параметров α, β, c автором была рассчитана таблица зависимости z=f(x).

Так как расчеты по формуле (3) сложны, были найдены простые и весьма точные выражения для описания зависимости объема словаря y=Y-1 от объема выборки x=X-1. Наиболее подходящим оказалось уравнение [2]

$$Y = X^{\frac{1}{\sqrt{1+\alpha \ln X}}}, \qquad (4)$$

где $\alpha = \alpha_0 + k \ln X;\ X = x + 1;\ Y = y + 1.$

Из (4) можно выразить величину α через Y и X:

$$\alpha = \frac{1}{\mathrm{LnX}} \left[\left(\frac{LnX}{LnY} \right)^2 - 1 \right]. \qquad (5)$$

По опытным значениям Y и X легко вычисляется величина α, а после построения графика прямой $\alpha = \alpha_0 + \kappa \mathrm{LnX}$ определяются оценки параметров α_0 *и* κ, т. е. начальная ордината и угловой коэффициент прямой α =φ(LnX). Значения параметров α_0 *и* κ зависят от единицы подсчета количества разных слов (лексем или словоформ). На этом свойстве основан вывод формулы для определения коэффициента аналитичности языка, который не зависит от объема выборки (текста), в отличие от отношения $\Delta k_a = Y_л / Y_{сл}$, которое принято для оценки степени аналитичности языка. Из последнего отношения следует, что при близких значениях числа разных лексем и словоформ коэффициент аналитичности близок к единице. Кроме того, с изменением объемов текста и словаря этот коэффициент изменяется, что не дает возможности сравнивать степень аналитичности языков на текстах разных объемов. Коэффициент анали-

тичности, рассчитанный по нашей формуле, не зависит от объемов текста и словаря:

$$\Delta k_a = \frac{\alpha_л - \alpha_{сл}}{\mathrm{Ln}X} \qquad (6)$$

Этот коэффициент практически не зависит от длины текста (выборки) X. Чем ближе Δk_a к нулю, тем выше степень аналитичности языка.

Значения параметров $\alpha_0, к$ оказались различными для связного текста и для случайно составленной выборки. Примем обозначения:

– для связного текста:

$$\alpha_t = \alpha_{0t} + к_t \mathrm{Ln}X$$

– для случайной выборки:

$$\alpha_{вt} = \alpha_{0в} + к_в \mathrm{Ln}X.$$

Тогда угол между двумя прямыми приближенно будет равен

$$\Delta k = \frac{\alpha_{0в} - \alpha_{0t}}{\mathrm{Ln}X} \qquad (7)$$

Если для некоторой знаковой системы обнаружится, что $\Delta k > 0$, то это свидетельствует о том, что мы имеем дело со связным текстом. Для случайной последовательности знаков $\Delta k = 0$.

Полученные выше формулы позволяют описать преимущества частотного словаря перед обычным при изучении иностранного языка. Пусть Y – число разных слов, использованных в тексте длиной X словоупотреблений, Z – число наиболее частых слов, взятых от начала частотного списка. Найдем значения Yи Z при условии, что обе группы слов обеспечивают одинаковый процент покрытия текста, т. е. F(Y)=F(Z) или 1-F(Y)=1-F(Z)=dY/dX. На основании распределения Вейбулла и зависимости (4) была получена формула [2]:

$$Z = \left(\frac{1}{\alpha} Ln \frac{dX}{dY}\right)^{\frac{1}{\beta}}. \qquad (8)$$

где $dX/dY = 1/(dY/dX)$, $\quad dY/dX = \frac{Y}{X}\left(\frac{LnY}{LnX}\right)^3\left(1 + \frac{\alpha_0}{2} LnX\right)$, $\qquad (9)$

α, β – параметры распределения Вейбулла, α_0 – один из параметров текста.

Расчеты показали, что отношение Y/Z при условии соблюдения равенства F(Y)=F(Z) практически находятся в пределах 1,5–2. Это значит, что при изучении иностранного языка запоминание подряд всех новых слов, встретившихся в тексте, приводит к значительно большим затратам (примерно в 1,5–2 раза) по сравнению с тем случаем, когда учащийся запоминает лишь наиболее частые слова (при условии, что обе группы слов обеспечивают одинаковый процент покрытия текста).

Число усвоенных слов Y* после прочтения текста длиной X определяется по формуле

$$Y^* = (pX)^{\frac{1}{\sqrt{1+\alpha\, LnpX}}} \qquad (10)$$

где $\alpha = \alpha_0 + kLnpX$, p – вероятность запомнить слово при встрече с ним.

Скорость прироста усвоенных слов в зависимости от длины текста X равна

$$\frac{dY^*}{dX} = \frac{Y^*}{X}\left(\frac{LnY^*}{LnpX}\right)^3\left(1+\frac{\alpha_0}{2}LnpX\right).$$

1. Нешитой, В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие / В. В. Нешитой. – Минск : Беларус. гос. ун-т культуры и искусств, 2010. – 61 с.

2. Нешитой, В. В. Законы распределения слов в тексте и его лексическая параметризация : Дисс. ... канд. филол. Наук / Ин-т языкознания им. Я. Коласа АН БССР. – Минск, 1973. – 135 с.

3. Белоногов, Г. Г. О некоторых статистических закономерностях в русской письменной речи // Вопросы языкознания. – 1962. – № 1. – С. 100–101.