Математические заметки

том 80 выпуск 5 ноябрь 2006

УДК 519

О ЗАКОНЕ ЦИПФА И РАНГОВЫХ РАСПРЕДЕЛЕНИЯХ В ЛИНГВИСТИКЕ И СЕМИОТИКЕ

В. П. Маслов, Т. В. Маслова

В статье уточняется ряд формул лингвистической статистики. Вводится понятие реальной и виртуальной мощности знака. Показывается, что формула, уточняющая закон Ципфа для частот встречаемости в частотных словарях, может быть распространена на произвольные знаковые объекты, т.е. на семиотические системы.

Библиография: 11 названий.

1. Синтагматическая (упорядоченная) последовательность слов и маркеры писателей. А. Н. Колмогоров, Р. Л. Добрушин, С. М. Гусейн-Заде и многие другие математики и лингвисты интересовались математическими аспектами устройства и функционирования языка. У Колмогорова есть работы по статистическим законам стихосложения. Гусейн-Заде, основываясь на идеях Колмогорова и Шеннона, рассматривал частоту букв алфавита как маркер языка [1]. Естественно, возникает вопрос, нельзя ли рассматривать частоту встречаемости слов как маркер языка отдельных писателей.

Вопрос об атрибуции текстов, особенно текстов Шекспира (за что была предложена денежная премия в размере более миллиона фунтов стерлингов [2]), привлекал внимание исследователей не одно столетие. Известный Н. А. Морозов и его последователи пытались вручную вычислять частоты служебных слов, причем работа Морозова [3] была опровергнута знаменитым математиком А. А. Марковым [4]. Следовательно, и сам Марков интересовался этой проблемой.

Мы сомневаемся в возможности определить авторство по частотам встречаемости слов, однако удивительное совпадение статистики частот слов в произведениях разных авторов с приводимыми ниже формулами позволяет определить параметр, зависящий от длины текста, а значит, кривую зависимости этого параметра от длины текста, являющуюся однозначной характристикой данного текста.

Статистические закономерности текстов и закон Ципфа исследовались в работе М. В. Арапова, Е. Н. Ефимовой и Ю. А. Шрейдера [5]. Эти авторы обратили внимание на то, что на сравнительно небольших текстах наблюдается хорошее согласие с законом Ципфа, в то время как на слишком длинных текстах, состоящих из большого числа относительно самостоятельных замкнутых частей, закон нарушается. Естественно предположить, что при порождении текста автор учитывает ‘текст в

Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 05-01-00824.

целом', а не только написанную часть. Возникает ситуация, когда процесс порождения зависит не только от прошлого, но и от будущего – от той части текста, которая еще не написана. В устных дискуссиях М. В. Арапов высказал мнение, что закон Ципфа позволяет определить только жанр текста, но не стиль отдельного писателя.

Интересно исследовать статистические зависимости в языке (в частотном словаре) на базе нового подхода, предлагаемого в данной и других недавних работах одного из авторов, и с учетом возможностей, которые предоставляют современные электронные технологии. В работах [6], [7], [8] выведены формулы, которые значительно точней, чем закон Ципфа, описывают соотношения между частотой и рангом слов в словаре. Представляется, что зависимости между частотой встречаемости слов и другими параметрами словаря могут служить если не маркером, то существенной характеристикой языка писателя.

В частотном словаре каждому слову сопоставлена частота его встречаемости в исходном корпусе текстов. Некоторым словам может соответствовать одна и та же частота встречаемости.

Анализ частотных словарей показывает, что слова в них разбиваются на категории:

1) 'вспомогательные' (сверхчастотные, так называемые стоп-слова);
2) часто встречающиеся слова;
3) редко встречающиеся слова;
4) очень редко встречающиеся слова (конденсат).

К первой категории относятся высокочастотные служебные слова и местоимения. В информатике их называют 'стоп-словами' и не принимают в рассмотрение. Частоты в этой части словаре могут следвать с большим отрывом друг от друга.

Ко второй категории частотного словаря относятся слова с достаточно высокой частотой встречаемости. Для этих частот характерна лакунарность, т.е. в спектре частот представлены не все частоты подряд. Из таких слов обычно состоят адаптированные тексты, в которых редкие слова заменяются более частыми синонимами или родовыми терминами.

В третьей категории частотного словаря сосредоточены среднечастотные и редкие слова. На этом участке словаря одной частоте соответствует достаточно много слов, и все частоты представлены без лакун.

К четвертой категории относятся слова с самыми низкими частотами – ниже заданной пороговой частоты.

Закон Ципфа обычно рассматривается в логарифмических координатах:

$$\ln r + \frac{1}{D} \ln \omega_r = \text{const}, \tag{1}$$

где r – ранг слова, т.е. его номер в частотном списке по убыванию частоты, ω_r – частота встречаемости этого слова, т.е. число употреблений слова в тексте, D – константа, которая для словарей, как правило, равна 1. Эта формула обозначает, что произведение номера слова на его частоту встречаемости есть (приблизительно) постоянная величина.

На графике эта зависимость изображается прямой.

Формула Ципфа (1) в логарифмических координатах слишком огрубляет соотношение между частотой и рангом: в логарифмических переменных она асимптотически верна, но в переменных без логарифмов неверна.

В физической литературе не встречается сравнение экспериментальных данных с теоретическими в логарифмических координатах, как представляется, по следующей причине. Если проводить ‘осредненную’ кривую, аппроксимируя набор экспериментальных точек, то она является как бы арифметическим средним относительно соединенных между собой отрезками ломаной соседними точками. Площадь сегментов над кривой и под кривой совпадает.

Если же рассматривть данные в логарифмических координатах, то среднее арифметическое

$$\frac{1}{n}\sum_{i=1}^{n}\ln x_i = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

совпадает со средним геометрическим в обычных координатах. А такое осреднение по точкам экспериментатор не может воспринять.

Помимо того, что закон Ципфа сильно огрубляет картину, он не описывает часть словаря с редко встречающимися словами.

Предлагаемый нами подход отличается от принятого в лингвистической статистике и заключается в следующем. Считается, что частота встречаемости слова – это вероятность встречаемости слова в тексте. Мы рассматриваем задачу с противоположной точки зрения. Пусть имеется алфавитный словарь, в котором указаны частоты встречаемости каждого слова. Если выбирать в нем слова случайным образом, какова вероятность попасть на слово с заданной частотой или какова вероятность ‘наткнуться’ в словаре на данную частоту? Допустим, мы выбираем слово из алфавитного списка; какова вероятность того, что ему отвечает такая-то частота? Эта вероятность равна числу слов, отвечающих данной частоте, деленному на общее число слов в словаре (объем словаря N). Действительно, вероятность случайно наткнуться в словаре на слово с самой высокой частотой, например, на слово u, чрезвычайно мала, она равна $1/N$, тогда как вероятность встретить случайно слово с частотой 1 будет самой большой, так как такие слова составляют в словаре самую большую долю.

В нашем рассмотрении частота выступает в качестве случайной величины, а число слов – в качестве числа выпадений этой случайной величины. Таким образом, мы рассматриваем в ‘перевернутом’ виде и сам частотный словарь (упорядоченный по возрастанию частот), и соотношение случайной величины и числа ее выпадений. Мы говорим о вероятности частоты встречаемости в словаре заданной частоты (т.е. частоты встречаемости слова в массиве текстов).

Итак, мы будем рассматривать частоты (число встречаемости слова в тексте) и число слов, отвечающих этим частотам. Прежде всего заметим, что слова с одним и тем же числом встречаемости в словаре можно переставлять как угодно: их можно располагать в алфавитном порядке или в порядке обратном алфавитному, – от этого соотношение между частотами и числом слов, разумеется, не изменится. Это наводит на мысль об аналогии с расположением бозе-частиц по уровням энергии ε_i.

Как было доказано одним из авторов [6], распределение Бозе получается, если предположить, что все варианты $\{n_i\}$ при условии

$$\sum_{i=1}^{s} n_i\varepsilon_i \leqslant E,$$

где n_i – число частиц на уровне ε_i, а E – заданная общая энергия всех частиц, равновероятны (равнозначны).

В нашем случае, казалось бы, энергии всех частиц отвечает объем массива текстов, из которых составлен словарь (обрезанный нами словарь).

Однако это не так. Как мы покажем ниже из лингвистических соображений, нужно учитывать другой, виртуальный, массив текстов.

Действительно, каждому реальному тексту какого-либо произведения соответствует более широкий, более подробный виртуальный текст.

Язык предоставляет возможность существенно экономить свои материальные ресурсы. Хорошо изучены, по крайней мере, два таких механизма: замена полнозначных элементов текста (слов, фраз и даже предложений) ограниченным количеством слов-заместителей – местоимений – и эллипсис – пропуск в тексте какого-либо легко подразумеваемого слова. Например, в предложении "Там росли ивы, а здесь березы." за местоимениями *там* и *здесь* скрываются фразы, которые встретились ранее в тексте, а во второй части предложения эллиптировано сказумое *росли*. Таким образом, в данном коротком предложении 'скрыто' более длинное, полное предложение. Можно легко себе представить, как увеличился бы текст, если бы эти языковые механизмы не использовались.

На примере математического текста можно также продемонстрировать, как экономятся лингвистические средства и сокращается длина текста. Вводя в первый раз какой-либо термин, автор объясняет, определяет его и вводит для него обозначение в виде буквы. В дальнейшем в тексте используется этот символ, что существенно сокращает длину текста, так как повторного обстоятельного описания не требуется.

В энциклопедиях сокращают физический объем текста, обозначая повторяющиеся слова их начальными буквами, аналогично математическому символу, независимо от того, в каком падеже они употреблены. Известны и другие примеры экономии языковых усилий.

Текст, в котором механизмы экономии задействованы в минимальной степени, может быть очень полезным иностранцу, недостаточно хорошо владеющему данным языком. В компьютерных технологиях процедуры снятия анафорики, т.е. восстановление заменяемых местоимениями слов и восстановление эллипсиса, являются необходимым промежуточным этапом анализа текста. Без такого этапа нельзя добиться построения правильной семантической структуры текста, так как компьютер не умеет однозначно устанавливать связи между местоимением и заменяемым словом. Например, фрагмент текста "В комнату вошла молодая дама с белой сумкой. Она была высокого роста." должен быть переписан как "В комнату вошла молодая дама с белой сумкой. Молодая дама была высокого роста.", в силу того, что и два других существительных в первом предложении потенциально могли бы служить антецедентами местоимения *она*.

Компьютер может точно подсчитать длину такого 'восстановленного' текста и сравнить реальные и 'восстановленные' частоты одних и тех же слов.

Известно, что за любым текстом, помимо 'сэкономленных' элементов, стоят не выраженные явно фоновые знания, общие для писателя и его потенциальных читателей. За счет этого 'скрытого' содержания возможны недоговорки, подтекст, аллюзии и т.п. Лаконичность стиля писателя высоко оценивается литературоведами.

За некоторыми известными фразами или даже отдельными словами известных произведений для знатоков стоят целые ситуации, сцены, типажи, характеры, судьбы и т.п., и им достаточно обмениваться только этими отдельными фрагментами текста, чтобы воссоздать картину целиком.

Все это вместе объясняет наш тезис о том, что длина текста произведения может рассматриваться с реальной и виртуальной точек зрения. В виртуальной длине в добавление к реальной учитываются все случаи восстановленных замен и эллиптических конструкций, т.е. вместо местоимений употребляются их антецеденты и в эллиптических конструкциях восстановлены все пропущенные элементы.

Как уже говорилось, в дальнейшем рассмотрении мы не учитываем 'сверхчастотные' слова, т.е. слова вспомогательных частей речи, в том числе и местоимения. При этом следует принять во внимание тот факт, что чем чаще слово повторяется в тексте, тем чаще оно заменяется местоимением, и соответственно, тем в большей степени это сказывается на виртуальной частоте встречаемости такого слова.

Итак, в нашем виртуальном тексте частота встречаемости слов увеличивается тем больше, чем больше исходная частота. Насколько больше, мы не знаем, поэтому положим, что числу слов n_i отвечает большая, чем i, частота, а именно:

$$\widetilde{\omega}_i = i(1 + \alpha i^{\gamma}), \qquad \alpha > 0, \quad \gamma > 0. \tag{2}$$

Это дает монотонное уменьшение зависимости ω_i от n_i при возрастании i.

Мы будем полагать, что все наборы слов, отвечающие частотам (2) из обрезанного словаря такие, что массив текстов, составленный из них, не превышает некоторого виртуального массива $\mathscr{E}$, равноценны.

Обозначая через $\omega_i = 1, \ldots, k$ все различные частоты встречаемости в частотном словаре, а через n_i число слов с одной и той же частотой встречаемости ω_i, мы получим, что реальный объем текста M равен

$$\sum_{i=1}^{k} n_i \omega_i = M,$$

$N = \sum n_i$ – число слов нашего словаря.

Аксиома состоит в том, что все варианты $\{n_i\}$ при условии

$$\sum_{i=1}^{k} n_i \widetilde{\omega}_i = \mathscr{E},$$

где $\widetilde{\omega}_i$ – виртуальная частота встречаемости, а $\mathscr{E}$ – некоторый виртуальный объем текста, равноценны.

В этом случае мы можем опираться на теорему 1 работы [8]. Будем нумеровать слова, начиная от самой низкой частоты 'обрезанного' словаря, и называть номер слова рангом r.

Число *различных* частот 'обрезанного' словаря много меньше числа слов, отвечающих ему.

Рассмотрим известное в теории вероятностей понятие 'коммулятивной вероятностей'. Оно вводится в теории вероятностей, лишь когда устанавливается отношение порядка, например, в порядковой статистике, поскольку существенно зависит от отношения порядка.

Коммулятивная вероятность $\mathscr{P}_k$ есть сумма первых s вероятностей в последовательности ω_i:

$$\mathscr{P}_s = \frac{1}{N}\sum_{i=1}^{s} n_i,$$

где $s < k$.

Введем константу $s = \omega_0/N$ и будем менять длину текстов каждого писателя, например, обрывая текст на некоторой странице, а затем добавляя по k страниц. Построим графики для числа отброшенных слов с самой высокой и самой низкой частотой как функции N. Параметр s, вычисленный исходя из минимума квадратичного отклонения от логарифмического закона, также является функцией N. Эти графики, возможно, будут в известной степени маркерами данного произведения, данного автора или, по крайней мере, данного жанра.

Это значит, что $r_s = N\mathscr{P}_s$ есть ранг (номер) слова, если считать все слова, начиная от $\omega_{\min}$ подряд, безразлично в каком порядке на одной и той же частоте.

В силу нашей аксиомы мы можем применить для ранга формулу теоремы 1 из [6] (см. также [8])

$$r_l = \sum_{i=1}^{l} \frac{c}{e^{\beta\widetilde{\omega}_i} - 1}, \tag{3}$$

где r_l – ранг l-го слова. Поскольку $k \ll N$, то $\beta \ll 1$, отсюда при $\beta \to 0$, $\sigma = 0$

$$r_l = \frac{1}{\beta}\sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^{1+\gamma}} + O\left(\frac{1}{\beta^2}\right).$$

Поскольку в силу теоремы 1 из [6] $l > \varepsilon k$, где ε сколь угодно мало, но не зависит от $k \to \infty$, а $\Delta\omega_i = 1$, то при $N \to \infty$

$$r_l \cong \sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^2} = \int^{\omega} \frac{d\omega}{\omega + \alpha\omega^{1+\gamma}} = \ln\frac{\omega}{1+\alpha\omega^{\gamma}} + c,$$

а следовательно,

$$r_l = T\,\ln\varrho\frac{\omega_l}{1+\alpha\omega_l^{\gamma}}, \qquad T = \frac{1}{\beta}, \tag{4}$$

где $\varrho = \exp(c/T)$.

Две константы, участвующие в формуле теоремы 1 из [6], так же, как соответствующие константы в формуле для распределения Бозе–Эйнштейна, определяются через общее число частиц (в данном случае это соответствует объему словаря) и общую энергию (в данном случае это соответствует виртуальному объему текста).

Мы не знаем виртуального объема текста и не знаем, как точно обрезать объем словаря. Поэтому те четыре константы, которые содержит формула (4), мы должны определить по-другому.

Прежде всего заметим, что для словаря, как известно еще из исследований Ципфа, $\gamma \sim 1$. Поэтому положим вначале $\gamma = 1$. Далее выберем две точки: одну малого ранга, другую среднего ранга, – в которых теоретическая кривая точно совпадает с экспериментальными точками. Выбранные точки должны быть устойчивы в том смысле, что полученные значения теоретической кривой, взятой по точкам, близким к выбранным, не менялись бы в пределах нашей точности.

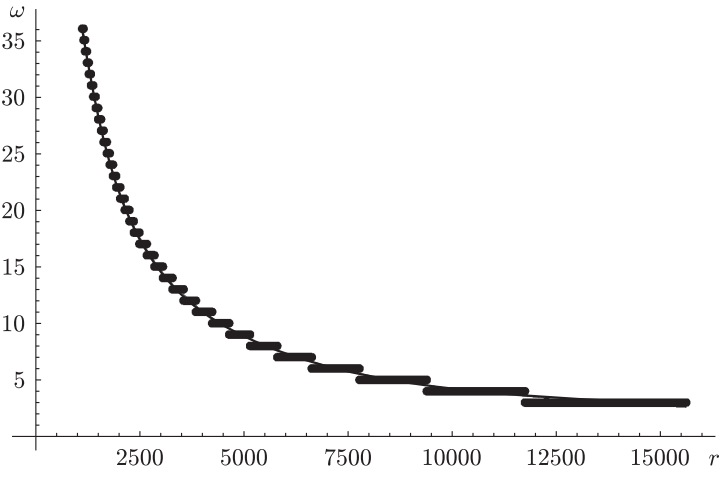

Рис. 1. r – номер слова в частотном словаре в порядке убывания частоты, ω – частота встречаемости слова

Последнюю константу нужно найти, вычисляя минимум отклонения теоретической кривой от экспериментальных точек. Это позволяет определить как длину обрезанного словаря N, так и длину виртуального текста $\mathscr{E}$.

Далее, полагая

$$r_c = N - r,$$

где r_c – есть ранг (номер) слова, отсчитываемый от самых высоких частот $\omega_{\max}$, мы придем к соотношению

$$\frac{r_c}{r_c^0} = \ln\left(1 + \alpha\left(\frac{1}{\omega}\right)^{\gamma}\right) + c_1, \tag{5}$$

где r_c^0 – нормировочная константа, которое при $\omega \to \infty$ переходит в закон Ципфа

$$\ln r_c + \gamma \ln \omega \sim \text{const}\,.$$

Отсюда по известным программам находим γ. Для частотных словарей, как правило, $\gamma = 1$. На рис. 1 и рис. 2 приведены графики зависимости ранга от частоты встречаемости слова для первого тома романа “Война и мир” Л. Н. Толстого. На рис. 1 низкочастотная часть словаря аппроксимирована формулой (5) (слова с частотами 3 и более; $r_0 = 43408.8$, $\alpha = 0.891995$, $c_1 = -189.321$); на рис. 2 для всего словаря приводится отклонение теоретических данных от данных словаря (в виде разности между частотой, зафиксированной в словаре, и частотой, вычисленной по формуле) в зависимости от ранга. Видно, что это отклонение для ранга, превосходящего 300, не превышает 1.5 слов.

После этого всю процедуру можно повторить и найти более точно виртуальный объем $\mathscr{E}$.

Вводя параметр l – длину текста определенного произведения от его начала – мы получаем кривые $\mathscr{E}(l)$, $N(l)$ и $\omega_{\min}(l)$, являющиеся маркерами данного произведения.

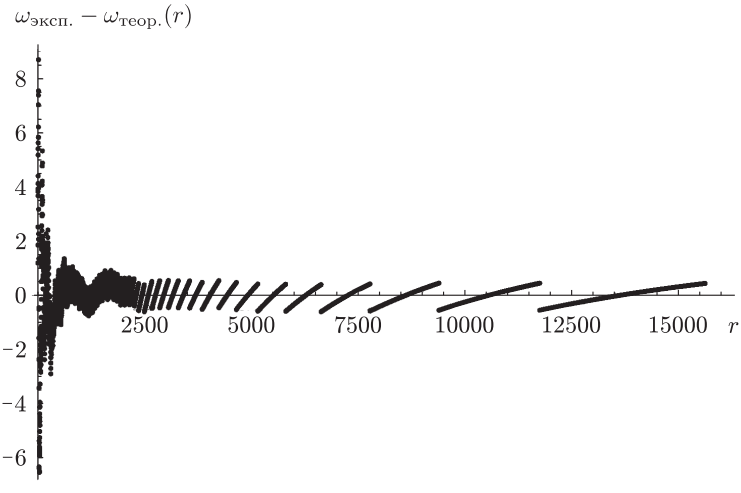

Рис. 2. $\omega_{\text{эксп.}}$ – реальная частота встречаемости слова в частотном словаре, составленном на материале первого тома романа “Война и мир”

2. Семиотика неупорядоченного набора знаков. Логарифмический закон, выведенный одним из авторов [6], [8] и существенно уточняющий и одновременно доказывающий закон Ципфа для частотных словарей, может быть распространен на широкий класс семиотических систем, например, на статистику распределения жителей по городам, которая в простейшем ее варианте закона типа Ципфа была замечена ранее. Законы Лотки и Бредфорда о документалистике, выведенные ранее эмпирически [9], также уточняются и укладываются в этот логарифмический закон.

Мы обобщим этот закон на семиотические объекты, а именно, рассматривая произвольные знаки и активность их использования. Из такого обобщения будет следовать, в частности, и уточнение формулировки законов Лотки и Бредфорда.

Основное свойство знака – его повторяемость в данном социуме. Символ, который никем, кроме автора, не используется, не есть знак. Пример ‘незнаков’ – придуманные Хлебниковым заумные слова.

Семиотические объекты – знаки – бывают разных типов.

В естественном языке знаками являются буквы, а также слова и иероглифы. Если мы рассматриваем произведение какого-либо писателя, мы имеем дело с единым текстом и с его словарем. Словарь – это множество разных слов, из которых построен текст. Текст в данном случае есть линейная последовательность синтагматически связанных слов. Если рассматривается массив разнородных документов, то текстом является не линейная последовательность, а набор слов.

Текст на естественном языке имеет реальную и виртуальную длину. Это связано с тем, что в языке действует принцип экономии. Например, повторяющиеся или известные адресату (читателю) элементы текста (слова и даже предложения) заменяются ограниченным количеством слов-заместителей – местоимений, а легко подразумеваемые слова или конструкции текста вообще опускаются (этот механизм называется эллипсисом). Таким образом, в коротком тексте, как правило, ‘скрыт’ более длинный текст. Насколько такой виртуальный текст длиннее реального, мож-

но легко себе представить, если попытаться восстановить все пропущенные, замененные или подразумеваемые элементы в этом тексте. Такая процедура, в частности, является необходимым этапом при автоматическом (компьютерном) анализе текста. Очевидно, что виртуальную длину текста существенно увеличивают также 'фоновые' знания, общие для автора и адресата (читателя), которые учитывает автор при построении текста.

Эту виртуальную длину, используя термин Гумбольта–Прието, мы будем называть 'энергéйей'[1] и обозначать $\mathscr{E}$. В современной лингвистике эта концепция все более актуализируется. Как мы покажем ниже, понятие виртуальной реальности и энергейи, применимо и к другим знаковым объектам.

Введем некоторые термины, которые мы будем использовать для обозначения семиотических объектов и их свойств. Прежде всего применим эти термины к естественному языку как наиболее изученному и понятному типу знаковых систем.

Слово естественного языка есть *знак*. Совокупность слов есть *словарь знаков*. Показателем активности знакоиспользования является число его употреблений. Этот показатель будем называть *реальной мощностью* знака. В естественном языке это число встречаемости слова в текстах. *Реальная мощность словаря знаков* – это суммарное число встречаемости всех знаков словаря (набор знаков). В языковых системах мощности словаря соответствует массив текстов, на базе которых составлен словарь (набор слов).

Знак обладает также *виртуальной мощностью*[2]. Реальная мощность слова – это реальное число его употреблений в заданном массиве текстов. Виртуальная мощность слова – это число использования данного слова во всех случаях, в том числе, когда слово не употреблено в тексте эксплицитно, но подразумевается и необходимо для правильного анализа и восприятия текста адресатом. Например, когда оно заменено местоимением, опущено (эллипсис) или заключено в подтексте. Виртуальная мощность слова больше реальной и растет с увеличением последней.

Некоторые знаки обладают одинаковой реальной мощностью (слова с одним числом встречаемости).

Аналогично, словарь знаков характеризуется и *виртуальной мощностью* – энергейей $\mathscr{E}$.

Забегая вперед, заметим, что нас будет интересовать число знаков с большей или равной мощностью, чем априорно заданная мощность, и с меньшей мощностью, чем априорно заданная мощность. Сумма этих двух чисел равна числу знаков в словаре. Для частотного словаря число слов с частотой большей, чем заданная, есть ранг слова, отвечающего этой частоте, отсчитываемый в порядке убывания частот, а число слов с частотой меньшей, чем заданная, соответственно ранг слова, отсчитываемый в порядке возрастания частот. Соответственно мы введем понятие ранга знака в словаре знаков.

Приведем примеры других знаковых систем.

---

[1]Греческим словом *энергейя* (творящая сила), вслед за знаменитым немецким языковедом В. фон Гумбольтом, в семиотике стали называть субъективную структуру мира литературного произведения, которая функционирует, сохраняется и воспроизводится, т.е. находится в 'движении' [10; с. 372–399].

[2]Дуис Скот в 13 веке первый, по-видимому (а за ним Суарес) обратил внимание на то, что к аристотелевским *actus* и *potentia* необходимо присовокупить *actus vertualis* как некоторое промежуточное важнейшее понятие.

В ситуации выборов выдвинутый кандидат или партия есть знак. Список кандидатов есть словарь знаков. Число голосов, которые получили кандидаты, соответствует реальной мощности знака. Виртуальная мощность – это число всех проголосовавших за данного кандидата плюс число людей, сочувствующих данному кандидату, но не голосовавших (рейтинг кандидата).

Город, отмеченный на карте жирно или тонко написанным названием и окружностью определенной толщины и диаметра, есть знак, указывающий на численность его населения. Число зарегистрированных жителей в этом городе есть реальная мощность этого знака. Общее число людей, находящихся в этом городе, есть его виртуальная мощность. Совокупность всех городов какого-либо государства есть словарь городов. Реальная мощность словаря знаков (городов) – это население данного государства; виртуальная мощность – общее число находящихся в нем людей, включая людей, временно работающих, туристов и т.д. – энергейя $\mathscr{E}$ государства.

Имя человека – знак; реальная мощность имени – это фиксированное статистическими данными число людей, носящих это имя в некотором регионе. Виртуальная мощность имени – это число всех людей, проживающих (как зарегистрированных, так и не зарегистрированных) в данном регионе, носящих это имя. К виртуальной мощности может относиться также использование иностранцами привычных для местных жителей имен вместо своих экзотических национальных имен.

В библиотеке книга с определенным названием есть знак. Каталог книг есть словарь знаков. Число требований, по которым выдан данный экземпляр книги, есть реальная мощность знака (востребованность или популярность книги). Число требований, поступивших на все книги, есть реальная мощность словаря знаков. Виртуальная мощность книги как знака складывается из всех случаев обращения к данной книге, не только читателя, взявшего книгу в библиотеке, но и его друзей, соседей, родственников и т.д., кто тоже прочитал взятую им книгу.

Сайт в интеренете тоже есть знак. Каталог сайтов есть словарь знаков. Реальная мощность сайта – фиксируемое число посещений его разными пользователями. Виртуальная мощность сайта – это число его реальных посещений плюс использование его материалов без посещения сайта.

Икона, например, икона Николая Угодника – это знак, узнаваемый в разных воплощениях. Собрание икон в церкви есть словарь знаков-икон. Реальная мощность иконы как знака есть число обращений к ней прихожан, которое фиксируется числом свечек, поставленных перед иконой (считается, что каждый поставил не более одной свечи). Виртуальная мощность иконы – это все мысленные обращения к ней прихожан.

Церковь в свою очередь является знаком. Словарь знаков – множество церквей в городе (‘сорок сороков’). Реальная мощность церкви – это число свечек, поставленных ко всем иконам. Виртуальная мощность церкви – это число людей, которые пришли в церковь молиться. На этом примере продемонстрировано свойство иерархичности знаков.

Музей тоже можно рассматривать как знак. Словарь знаков – все музеи данного города. Число проданных билетов – его реальная мощность. Виртуальная мощность знака – число всех людей, которые пришли в данный музей, в том числе и не посетивших его коллекцию по причине высокой цены билетов, недостатка времени или ограничившись просмотром каталога и т.п.

Пусть организована рекламная акция на некоторый перечень товаров (словарь знаков). Реальная мощность рекламы отдельного товара как знака – это число людей, которые взяли рекламный проспект. Виртуальная мощность рекламы данного товара – общее число людей, которые этот рекламный проспект просмотрели.

Перейдем к основной формуле, обобщающей формулу, выведенную в работе [8] для частотных словарей, на более широкий класс семиотических объектов.

Рассмотрим некоторый абстрактный словарь знаков $\{s_i\}$, $i = 1, 2, \ldots, n$. Пусть реальная мощность знака s_k равна ω_i.

Мы введем две неизвестных константы α и γ и положим, что виртуальная мощность знака s_i

$$\omega_{\text{vir}} = \widetilde{\omega}_i = \omega_i(1 + \alpha\omega_i^{\gamma}).$$

Возможно, что если знаками выступают слова (команды) простейшего языка программирования, то для набора программ виртуальная длина равна реальной и $\gamma = 0$.

Наша парадигма базируется на определении А. Н. Колмогорова случайности как максимальной сложности (колмогоровской сложности) [11]. Это значит, что чем более длинный алгоритм придумывания линейной последовательности знаков (синтагматический набор) использовал автор, тем ближе он к общему положению большинства всех возможных вариантов таких последовательностей. Аналогично тому, как при игре в орла и решку, чем более длинный и сложный алгоритм использует игрок, тем ближе выпадения орлов и решек к ‘общему’ варианту, при котором в половине случаев выпадает орел, в половине – решка.

Если набор знаков (например, распределение населения по городам, посещение сайтов) происходил стихийным образом, т.е. согласно очень длинному алгоритму, если проследить за его детерминистической историей, то, как правило, он должен находиться в общем положении, т.е. вблизи точек, где расположено большинство вариантов.

Пусть n_i – число различных знаков одинаковой реальной мощности ω_i.

Поскольку

$$\sum_{i=1}^{k} n_i\widetilde{\omega}_i = \mathscr{E},$$

мы предположим, что число знаков n_i, отвечающих данной виртуальной мощности ω_i знака s_i, есть случайная величина с равновероятным распределением для любого набора $\{n_i\}$ такого, что

$$\sum_{i=1}^{k} n_i\widetilde{\omega}_i \leqslant \mathscr{E}.$$

Иначе говоря, *наборы* $\{n_i\}$ *есть элементарные события.* В этом состоит наша основная и единственная аксиома. Очевидно, что $\mathscr{E} \leqslant \widetilde{\omega}_{\max}N$, где N – длина словаря знаков.

Задача распадается на два случая:

1) $\mathscr{E} < \dfrac{\sum_{i=1}^{k}\widetilde{\omega}_i}{k}N$;

2) $\dfrac{N}{k}\displaystyle\sum_{i=1}^{k}\widetilde{\omega}_i \leqslant \mathscr{E} \leqslant \widetilde{\omega}_{\max}N.$

Эту аксиому нужно понимать так, что наша конкретная последовательность знаков 'энергейи' $\mathscr{E}$ есть одна из множества подобных же с энергейей, не превосходящей $\mathscr{E}$, и с тем же словарем знаков, и мы предполагаем, что по крайней мере в основной части знаков она находится в общем положении относительно всевозможных вариантов $\{n_i\}$, удовлетворяющих условиям 1) или 2).

Пронумеруем знаки словаря в порядке возрастания их мощности, начиная от минимальной мощности $\omega_{\min}$. В пределах одной мощности знаки могут располагаться в произвольном порядке. Номер упорядоченного таким образом знака будем называть его рангом и обозначать r. Если l – номер мощности ω_l (начиная от минимальной $\omega_{\min}$), то под r_l будем понимать число всех знаков с меньшей или равной мощностью ω_l.

В силу сформулированной аксиомы мы можем применить в случае 1) для ранга знака формулу теоремы 1 из [6] (см. также [8])

$$r_l = \sum_{i=1}^{l} \frac{c}{e^{\beta\widetilde{\omega}_i} - 1}, \tag{6}$$

где r_l – ранг l-го знака. Поскольку $k \ll N$, то $\beta \ll 1$, отсюда при $\beta \to 0$, $\sigma = 0$ мы получаем формулу (4) для ранга знака.

Аналогично доказывается, что в случае 2)

$$r_l = \sum_{i=1}^{l} \frac{c}{1 - e^{-\beta\omega_i}}, \tag{7}$$

что приводит к тому же соотношению (4) при $\beta \to 0$. Действительно,

$$r_l = \frac{1}{\beta} \sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^{1+\gamma}} + O\left(\frac{1}{\beta^2}\right).$$

Поскольку в силу указанной теоремы [6] $l > \varepsilon k$, где ε сколь угодно мало, но не зависит от $k \to \infty$, а $\Delta\omega_i = 1$, то при $N \to \infty$

$$r_l \cong \sum_{i=1}^{l} \frac{1}{\omega_i + \alpha\omega_i^2} = \int^{\omega} \frac{d\omega}{\omega + \alpha\omega^{1+\gamma}} = \ln \frac{\omega}{1 + \alpha\omega^{\gamma}} + c,$$

а следовательно,

$$r_l = T \ln \varrho \frac{\omega_l}{1 + \alpha\omega_l^{\gamma}}, \qquad T = \frac{1}{\beta}, \tag{8}$$

где $\varrho = \exp \frac{c}{T}$.

Две константы, участвующие в формуле теоремы 1 работы [6] так же, как соответствующие константы в формуле для распределения Бозе–Эйнштейна, определяются через общее число частиц (в данном случае это соответствует объему словаря знаков) и общую энергию (в данном случае это соответствует виртуальному объему словаря знаков).

Мы не знаем виртуального объема словаря знаков и не знаем, как точно обрезать объем словаря. Поэтому те четыре константы, которые содержит формула (4), мы должны определить по-другому.

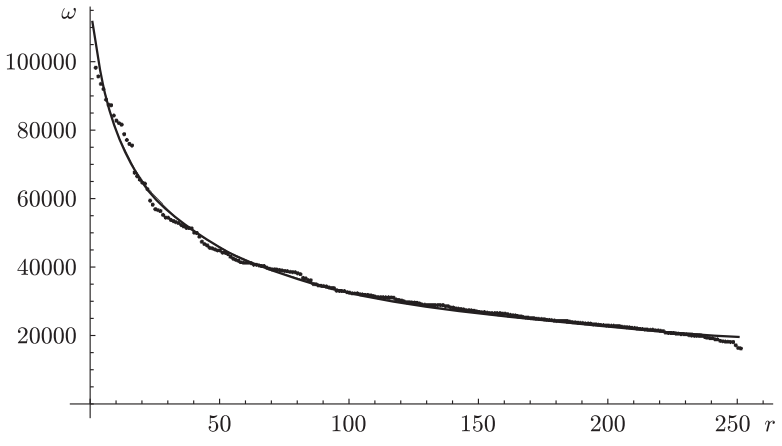

Рис. 3. Зависимость между марками американских автомобилей и их ценой

Прежде всего определим γ из закона Ципфа. Далее выберем две точки: одну малого ранга, другую среднего ранга так, чтобы в них теоретическая кривая точно совпадала с экспериментальными точками. Выбранные точки должны быть устойчивы в том смысле, что полученные значения теоретической кривой, взятой по точкам, близким к выбранным, не менялись бы в пределах нашей точности.

Последнюю константу нужно найти, вычисляя минимум отклонения теоретической кривой от экспериментальных точек. Это позволяет определить как длину обрезанного словаря N, так и длину виртуального текста $\mathscr{E}$.

Далее, полагая

$$r_c = N - r,$$

где r_c – есть ранг (номер) знака, отсчитываемый от самых высоких мощностей $\omega_{\max}$, мы придем к соотношению

$$\frac{r_c}{r_c^0} = \ln\left(1 + \left(\frac{\omega_0}{\omega}\right)^{\gamma}\right),$$

где ω_0, r_c^0 – нормировочные константы, которое при $\omega \to \infty$ переходит в закон Ципфа

$$\ln r_c + \gamma \ln \omega \sim \text{const}\,.$$

Отсюда по известным программам мы уточняем γ. После этого всю процедуру можно повторить и найти более точно виртуальный объем $\mathscr{E}$.

В заключение мы приведем два важных примера знаков и их мощностей, статистика которых достаточно точная.

Первый пример (см. рис. 3) – марки американских автомобилей (знаки) и их цены (мощность знака – цена данной марки).

Виртуальная мощность – это дополнительно к цене автомобиля шлейф неизбежных расходов: страховка, бензин, стоянка, а также услуги, которые учитывает покупатель автомобиля. Ранг на графике 3 – это номер марки в порядке убывания цены.

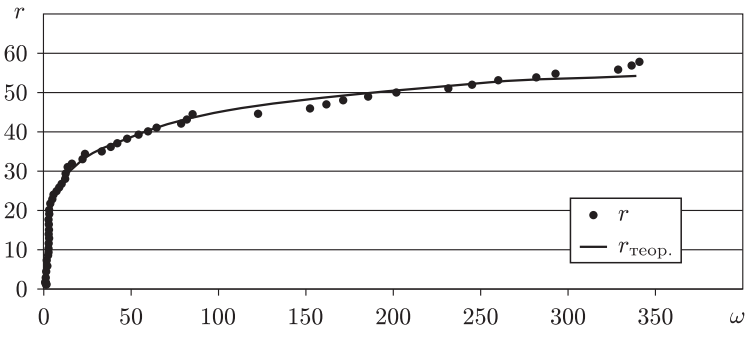

Рис. 4

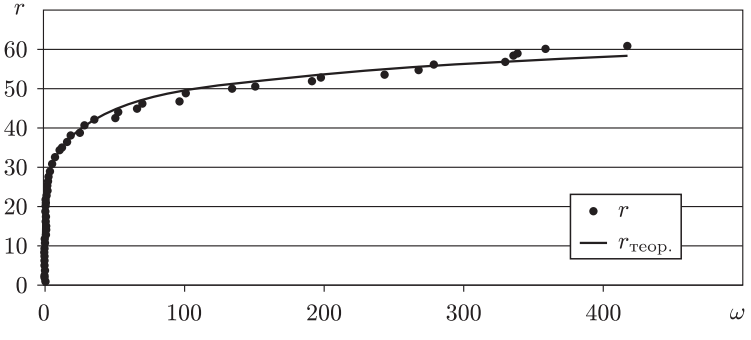

Рис. 5. Зависимость между порядковым номером японской свечи и ее частотой встречаемости: r – порядковый номер свечи, начиная от наименьшей, ω – частота встречаемости свечи

Колебание цены одного финансового инструмента (акции) на фондовой бирже удобно обозначать знаком так называемой ‘японской свечей’. ‘Японские свечи’, объединенные в классы с точностью до 0.2%, являются примером системы знаков. Мы нумеруем их в порядке убывания частоты повторяемости. Частота повторяемости знака ‘японской свечи’ есть его мощность. Его виртуальная мощность связана со сделками, которые совершаются брокерами вне биржи в электронных торговых сетях и близких по характеру финансовых инструментах, которые, незначительно отличаясь от данного финансового инструмента, могут для инвестора заменить его. В последнее время нарастает тенденция группировать акции в эквивалентные множества (аналогично тому, как слова группируются в дескрипторы) и для инвестора могут заменять друг друга, как местоимения или редко встречающиеся синонимы могут заменять повторяющиеся слова в тексте и, тем самым, менять статистику встречаемости слов. Инвестор избегает повторяемости еще в большей степени, чем писатель избегает повторяемости одних и тех же слов: чем более широкий диапазон покупок, тем меньший риск проиграть. Ранг на графиках 4 и 5 – это номер свечи, начиная от самой меньшей.

Авторы выражают глубокую благодарность М. Ю. Романовскому за предоставление данных по ценам на автомобили в США за 2005 год; А. В. Чуркину за помощь в

обработке этих данных и Н. В. Старченко за предоставление и помощь в обработке данных по финансовым инструментам.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

[1] С. М. Гусейн-Заде, "О распределении букв русского языка по частоте встречаемости", *Проблемы передачи информации*, **23**:4 (1988), 102–107.

[2] М. Б. Малютов, "Обзор методов и примеров атрибуции текстов", *Обозрение прикладной и промышленной математики*, **12**:1 (2005), 41–77.

[3] Н. А. Морозов, "Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд", *Известия отд. русского языка и словесности Имп. Акад. наук*, том XX, кн. 4 (1915).

[4] А. А. Марков, "Об одном применении статистического метода", *Известия Имп. Акад. наук*, *серия VI*, том X, № 4 (1916), 239.

[5] М. В. Арапов, Е. Н. Ефимова, Ю. А. Шрейдер, "О смысле ранговых распределений", *НТИ*, *сер.* 2, 1975, № 1, 9–20; http://kudrinbi.ru/public/442/index.htm.

[6] В. П. Маслов, "Об одной общей теореме теории множеств, приводящей к распределению Гиббса, Бозе–Эйнштейна, Парето и закону Ципфа–Мандельброта для фондового рынка", *Матем. заметки*, **78**:6 (2005), 870–877.

[7] V. P. Maslov, "Linguistic statistics", *Russ. J. Math. Phys.*, **13**:3 (2006), 315–325.

[8] В. П. Маслов, "Закон "отсутствия предпочтения" и соответствующие распределения в частотной теории вероятностей", *Матем. заметки*, **80**:2 (2006), 220–230.

[9] С. П. Капица, С. П. Курдюмов, Г. Г. Малинецкий, *Синергетика и прогнозы будущего*, Эдиториан УРСС, М., 2001.

[10] *Семиотика. Семиотика языка и литературы*, Радуга, М., 1983.

[11] А. Н. Колмогоров, "Теория передачи информации", *Труды сессии АН СССР по вопросам автоматизации производства*, М., 1957.

В. П. Маслов, Т. В. Маслова
Московский государственный университет
им. М. В. Ломоносова
E-mail: v.p.maslov@mail.ru

Поступило
27.09.2006