Ю. А. Шрейдер, О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа),
Пробл. передачи информ., 1967, том 3, выпуск 1, 57–63

ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том III *1967* *Вып. 1*

УДК 621. 391.194

О ВОЗМОЖНОСТИ ТЕОРЕТИЧЕСКОГО ВЫВОДА СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ ТЕКСТА (К ОБОСНОВАНИЮ ЗАКОНА ЦИПФА)

Ю. А. Шрейдер

Выясняются свойства модели порождения языка, из которых теоретически следует выполнение известного эмпирического закона Ципфа для распределения частот слов в текстах.

1. Известно, что статистическая структура лексики текстов на естественных языках * подчиняется закономерности, известной под названием закона Ципфа. Этот закон состоит в следующем. Пусть T — некоторый достаточно длинный текст, а S_T — его словник, т. е. перечень всех слов, участвующих в данном тексте. Обозначим через N_k количество вхождений k-го слова из словника S_T в текст T и занумеруем элементы словника в порядке убывания (невозрастания) величин N_k. Тогда выполняется следующая эмпирическая зависимость:

$$N_k \approx Ck^{-\gamma}, \tag{1}$$

где величина γ близка к единице ($\gamma \geqslant 1$). Эта зависимость и называется законом Ципфа.

Вместо величин N_k удобно ввести частоты появления k-го слова $f_k = N_k / N$, где N — общее количество слов в тексте T. Тогда соотношение (1) можно переписать в виде

$$f_k \approx ck^{-\gamma}, \tag{2}$$

где $c = C / N$.

Содержательному обсуждению этого закона, выражаемого соотношениями (1) и (2), посвящена вторая глава монографии [1].

Зависимость, выражаемая соотношениями (1) и (2), является экспериментальным фактом, установленным на основе анализа частотных словарей.

Так как $\gamma \approx 1$, то закон Ципфа может быть сформулирован в форме, впервые установленной Эсту: произведение частоты слова f_k на его порядковый номер k в частотном словаре приблизительно постоянно.

---

* Мы пользуемся термином «естественный», чтобы отличать обычные человеческие языки (русский, английский и т. п.) от искусственных или формализованных языков с формальными правилами употребления лексики и четко ограниченным синтаксисом. С этой точки зрения языки типа Эсперанто или различные Арго являются естественными языками. Было бы весьма любопытно проверить, не удовлетворяют ли тексты на достаточно богатых искусственных языках тем же статистическим закономерностям, что и естественные. С точки зрения положений данной статьи автору кажется весьма правдоподобным, что тексты, публикующиеся на языке алгол, также удовлетворяют закону Ципфа.

В самом деле из (2) следует:

$$f_k k \approx Ck^{1-\gamma} \approx \text{const}^*. \qquad (3)$$

Согласно закону Ципфа, подавляющая часть словника состоит из малочастотных слов.

Соотношения (1) и (2) хорошо выполняются только при $k > k_0$ **. Более того, соотношения (1) и (2) снижают точность при слишком больших k. Некоторые авторы (см. [1]) считают, что при больших k нужно использовать несколько увеличенные значения γ. С другой стороны, неточность соотношений (1) и (2) при больших k естественно объясняется тем фактом, что при очень малых частотах и, следовательно, малых величинах N_k статистические закономерности хуже выполняются.

Значение показателя γ зависит от типа речи (см. [1]). Так, этот показатель отличается от среднеязыкового значения (и притом в сторону увеличения), например, для речи душевнобольных, отчасти для детской речи.

Достаточно общая формулировка закона Ципфа состоит в следующем. Дан текст T, элементами которого служат слова, образующие словник текста S. В словнике S введено некоторое отношение эквивалентности, так что множество классов эквивалентности образует словарь S_0. Обозначим через N_k количество появлений в тексте T представителей k-го класса эквивалентности (k-го элемента словаря S_0). Пусть числа N_k упорядочены по убыванию, т. е. именно из этого условия выбрана нумерация элементов словаря S_0. Тогда мы будем говорить, что текст T с заданной факторизацией словника удовлетворяет закону Ципфа, если, начиная с некоторого k, $N_k \approx Ck^{-\gamma}$.

Пример 1. Текст T — осмысленный текст естественного языка, словарь S — обычный словник текста, а факторизация не производится (S_0 совпадает с S или иначе, отношение эквивалентности есть графическое тождество).

Пример 2. В словник S вводится отношение эквивалентности — вхождение в общую парадигму (словарь S_0 состоит из списка слов в основных грамматических формах).

Пример 3. Отличается от примера 1 отношением эквивалентности, объединяющим в один класс слова с одинаковыми грамматическими формами. Словарь S_0 — это список грамматических форм, встречающихся в данном тексте. (Соответствующий подсчет был выполнен С. Якубович.)

Пример 4. Текст T — предметный указатель по некоторой области науки, элементы словаря S (слова текста) имеют вид: «автор и название статьи + название журнала + номер и том». Факторизация состоит в склеивании «слов» с одинаковыми названиями журналов. Словарь S_0 — есть список журналов, а закон Ципфа в данном случае есть закон рассеяния Брэдфорда (см. [3]).

Пример 5. Текст T — список всех жителей СССР с указанием города, где он проживает. Факторизация состоит в объединении в один класс жителей одного города. Словарь S_0 есть просто список городов Советского Союза в порядке убывания численности жителей. Обработка опубликованных статистических данных переписи показывает, что этот текст также удовлетворяет закону Ципфа, т. е. количество жителей в каждом городе $N_k = Ck^{-\gamma}$, причем закон верен вплоть до городов с населением 60 000.

Пример 6. Текст T — список книг, выдаваемых библиотекой за некоторый период времени. Каждый раз, когда книга выдается читателю, ее название вносится в список. В данном случае $S = S_0$ является перечнем всех книг данной библиотеки. А. И. Черный обратил внимание, что текст T, по-видимому, удовлетворяет закону Ципфа. В данном случае, число N_k есть количество выдач данной книги за фиксированный отрезок времени.

* Разумеется, при изменениях k в конечных пределах.

** По некоторым данным Мандельброта [2] $k_0 = 15$, а по данным Р. М. Фрумкиной [1] $k_0 = 50$, при этом величина $\sum_{k=1}^{k_0} f_k$ для фиксированного языка достаточно устойчива.

Пример 7. Пусть имеется некоторый социальный коллектив и некоторый член этого коллектива ведет список лиц, с которыми он последовательно вступал в контакт * в течение некоторого достаточно длительного отрезка времени (однако, такого, за который сам коллектив не претерпел существенных изменений). Этот список является текстом T, а словарь S есть перечень членов данного коллектива (за исключением одного). Представляется весьма вероятным, что и здесь имеет место нечто близкое к закону Ципфа. Факторизация в данном случае означала бы классификацию контактов по группам данного коллектива.

Ввиду такой универсальности закона Ципфа кажется очень естественным, что он должен теоретически выводиться из весьма общих свойств модели порождения текстов.

В работе Мандельброта [2] закон Ципфа выводится из соображений минимальной стоимости оптимального кода в предположении, что текст состоит из слов, разделенных пробелами. Предполагается, что слова образуются в φ-буквенном ($\varphi > 1$) алфавите с помощью марковского процесса, определяющего условные вероятности присоединения к готовой части слова очередной буквы или пробела. Интересно, что эта схема объясняет асимптотический характер закона Ципфа.

Такая модель порождения текста отнюдь не представляется адэкватной реальному речевому процессу, который, как замечено и в [5], вовсе не определяется свойством оптимального кодирования. Более того, процесс порождения речи вряд ли может описываться на уровне знаковой системы фиксированного уровня (ср. [6]).

2. Будем исходить из того, что существует некоторая модель типа исчисления или автомата, генерирующая все тексты данного языка **. Далее мы сформулируем некоторые общие свойства такой модели, на основе которых может быть получен закон Ципфа.

Дальнейший вывод будет основан на некоторых термодинамических аналогиях (в статистической теории информации такие аналогии являются вполне привычными, но в математической лингвистике они пока, по-видимому, не применялись) и одной важной идее, использованной в [2].

Рассмотрим некоторое счетное *** множество V, которое мы будем называть алфавитом. Элементы этого множества называются знаками (содержательно эти знаки могут интерпретироваться как слова, словоформы, морфемы, словосочетания, простые предложения и т. п.— допускается любой уровень знаковой системы).

Рассмотрим знаковую систему S, являющуюся некоторым множеством текстов $\{T\}$ над алфавитом V ****. Эта знаковая система S состоит, по определению, из всех текстов, порождаемых моделью $\mathfrak{M}$. Введем теперь следующие предположения.

1) Для знаков алфавита $x \in V$ определена положительная числовая функция $E(x)$, которая интерпретируется как «сложность» порождения знака x в модели $\mathfrak{M}$. В схеме Мандельброта [2] $E(\bar{x})$ есть число букв в словоформе x. Каждому тексту $T \in S$ приписывается «сложность» порождения $E(T)$, равная сумме «сложностей», входящих в текст знаков, т. е.

$$E(T) = \sum_{x \in T} E(x) = \sum_{x \in V} N_x E(x), \qquad (4)$$

---

* Естественно учитывать контакты, которые связаны с особенностями данного коллектива. Например, служебные, если коллектив есть учреждение.

** Кроме моделей Н. Хомского и С. К. Шаумяна (см. [7]), мы хотели бы обратить внимание на родственные модели М. И. Белецкого [8] и М. В. Ломковской [9], основанные на свойствах графа управления.

*** Существенным является объем словаря. Нам удобно рассматривать бесконечный словарь.

**** Общее определение текста см. в [6].

где N_x — число различных вхождений знака x в текст T. Очевидно, $\Sigma N_x = N$ есть объем текста T.

В силу термодинамических аналогий $E(T)$ можно было бы называть «энергией» порождения текста T.

2) Будем считать, что множество текстов $\{T\}$ данной знаковой системы образует статистический ансамбль, так что вероятность порождения данного текста T зависит и притом непрерывно только от его сложности $E(T)$ *. Тогда для каждого $x \in V$ имеет смысл говорить о математическом ожидании количества вхождений знака x в текст объема N. Эту величину будем обозначать как $m_{x,N}$. Если ансамбль текстов обладает свойством эргодичности, то для достаточно больших N $m_{x,N} \approx P_x N$, где P_x — вероятность вхождения знака x в фиксированную позицию текста.

3) Будем предполагать, что с вероятностью, сколь угодно близкой к единице, для достаточно больших N выполняется соотношение

$$\frac{E}{N} \approx \hat{E}, \tag{5}$$

где $\hat{E}$ — постоянная, характеризующая среднюю сложность текста. Соотношение (5) может быть получено из условия независимости появления в тексте знаков, имеющих сложность, различающуюся больше чем на фиксированную величину. Однако мы предпочитаем постулировать (5), нежели вводить даже самые слабые предположения о независимости знаков в тексте **.

4) Обозначим через $r(E)$ число знаков алфавита со сложностью, не превосходящей E. Иначе говоря, $r(E)$ есть число тех x, для которых $E(\bar{x}) \leqslant E$. Будем считать, что $r(E)$ асимптотически растет как экспонента, т. е.

$$r(E) \approx a^E. \tag{6}$$

Это предположение является естественным. Дело в том, что по существу все модели порождения связаны либо с выбором подграфа в некотором графе, либо с конструированием некоторого графа. Для таких графов существуют различные оценки сложности, и обычно оказывается, что количество графов с данной оценкой сложности растет экспоненциально (ср., в частности, оценки, полученные в работе [10]).

Напишем теперь формулу для вероятности того, что в тексте с объемом N будет иметься заданный набор $\{N_x\}$ чисел вхождений знаков x. Из соотношения (5) следует, что можно ограничиться текстами со значением $E \approx \hat{E}N$. Согласно условию 2), такие тексты будут почти равновероятны. Вероятность порождения одного текста обозначим через w. Число таких текстов равно

$$\frac{N!}{N_{x_1}!\,N_{x_2}!\ldots N_{x_h}!} \tag{7}$$

где среди величин N_{x_k} лишь конечное число отлично от нуля. Следовательно, в знаменателе (7) лишь конечное число факториалов отличается от единицы.

---

* Поле элементарных событий состоит из всех текстов данного объема N, каждый из которых имеет вероятность, зависящую только от $E(T)$.

** Правда, предположение, что вероятность текста определяется аддитивной функцией от составляющих его знаков, есть уже некоторое косвенное предположение о какой-то независимости этих знаков.

Таким образом, вероятность порождения текста с заданными числами вхождения знаков алфавита V равна

$$\bar{p} = \frac{N!}{N_x! \dots N_{x_k}!} W.$$

Найдем максимум этой вероятности для фиксированных значений N и $E(T)$. Воспользовавшись методом Лагранжа, найдем условный максимум логарифма этой функции. Имеем

$$\ln p + \lambda \sum_x N_x E(x) + \mu \sum_x N_x = \max \tag{8}$$

или, применяя формулу Стирлинга,

$$\ln N! - \sum_x (N_x \ln N_x - N_x) + \ln W + \lambda \sum_x N_x E(x) + \mu \sum_x N_x = \max.$$

Дифференцируя по N_x, получаем $-\ln N_x + \lambda E(x) + \mu = 0$, откуда

$$N_x = \alpha e^{-\mu E(x)}, \tag{9}$$

где $\alpha = e^{\lambda}$.

Более точные оценки выражений типа (9) позволяют убедиться в том, что с вероятностью, сколь угодно близкой к единице, значения N_x отклоняются от правых частей (9) не более чем на величину порядка $\sqrt{N}$. Отсюда следует, что величины N_x / N сходятся к определенным пределам m_x. Очевидно, при больших N математическое ожидание числа появлений знака x в тексте длины N удовлетворяет приближенному равенству $m_{x,N} \approx m_x N$. Таким образом m_x можно интерпретировать как математическое ожидание частоты появления знака x в тексте.

Итак, частоты N_x / N появления каждого знака в тексте сходятся по вероятности * к величинам

$$m_x = \beta e^{-\mu E(x)}. \tag{10}$$

Упорядочим теперь знаки алфавита V по неубывающим сложностям $E(x)$ и воспользуемся предположением 4). Если $r(E)$ есть число знаков со сложностью, не превосходящей E, то сложность k-го по порядку слова выражается обратной $r(E)$ функцией. Следовательно, из того, что $r(E) \approx a^E$, следует, что

$$E_k = E(x_k) \approx \log_a k = \frac{\ln k}{\ln a}. \tag{11}$$

Подставляя (11) в (9), имеем

$$m_x \approx \beta e^{-\mu \frac{\ln k}{\ln a}} = \beta_k^{-\frac{\mu}{\ln a}} = \beta k^{-\nu}, \tag{12}$$

---

* Это доказательство почти совпадает с приведенным в [11] доказательством экспоненциального распределения энергий системы в термостате. Мы не делали доказательство более строгим, потому что нетрудно сформулировать такие уточнения предположений 1), 2), 3), при которых доказательство можно сделать математически безупречным.

что уже по существу совпадает с законом Ципфа (2), если учесть установленный выше факт, что частоты $f_k \approx m_{x_k}$.

Таким образом эмпирический закон Ципфа отражает тот факт, что математические ожидания частот появления слов в тексте распределены по закону (12) и что для больших текстов эти частоты мало отличаются от своих математических ожиданий. Для текстов, обладающих свойством эргодичности, можно было бы вместо математических ожиданий m_x говорить о вероятностях появления знака x в произвольной позиции данного текста. Так как условие (6) выполняется, вообще говоря, лишь асимптотически, то равенство (12) должно выполняться лишь начиная с некоторого номера k *.

Ввиду общности формулировки закона Ципфа имеет смысл проверить на опыте различные варианты статистик, рассматривая тексты различных языков в знаковых системах различных уровней, с учетом различных характеристик сложности.

В частности, любопытно изучить статистику появления в текстах пар слов, связанных управлениями или просто стоящих рядом в линейном тексте. Такая статистика связана с переходом от знаковой системы S, состоящей из текстов T (V, Γ) **, к двойственной знаковой системе S^*, тексты которой устроены следующим образом: для каждого текста $T(V, \Gamma) \in S$ берется двойственный граф Γ^*, вершины которого суть ребра графа Γ, а ребра — вершины графа Γ. В вершине графа Γ^* ставится упорядоченная пара знаков алфавита V, определяемая вершинами графа Γ, инцидентными соответствующему ребру. Полученные таким способом тексты T^* образуют знаковую систему S^*.

Пример 8.

$$T: x \to y \begin{matrix} \nearrow x \\ \searrow z \end{matrix},$$

$$T^*: xy \begin{matrix} \nearrow yx \\ \searrow yz \end{matrix}.$$

Если для знаков алфавита V существует функция расстояния $\rho(x, y)$, то для текстов T^* имеется естественная функция сложности

$$E(T^*) = \sum \rho(x, y),$$

где сумма берется по всем парам знаков, стоящих в вершинах текста T (ср. [12]).

Понятие «сложности» нам представляется вообще естественным для изучения языка. Обычно весьма трудно формулировать абсолютные ограничения на характеристики текстов (длина фразы, степень непроективности (см. [13]), глубина фразы и т. п.). Дело происходит так, как будто языку «трудно» порождать тексты со слишком большими значениями некоторых параметров. Целью данной работы является установление связи между характеристиками сложности и статистическими свойствами текстов. Некоторые принципиальные следствия из закона Ципфа указаны в [14].

---

* Так, даже в простейшей модели, когда появления букв в слове независимы, а сложность, согласно [2], есть число букв в слове, то $r(E) = (a^E - a) / (a - 1) \approx a^E$, где a — число букв. Таким образом, условие $r(E) \approx a^E$ выполняется лишь асимптотически.

** Здесь, следуя [6], под текстом понимается граф Γ с вершинами, помеченными знаками алфавита V.

ЛИТЕРАТУРА

1. Фрумкина Р. М. Статистические методы исследования лексики. М., «Наука», 1964, *3*, 114.
2. Mandelbrot B. On recurrent noise limiting coding. Laboratories d'Electronique et de Physique Appliquces. Paric, France, 1954. (Русск. пер.: Мандельброт Б. О рекуррентном кодировании, ограничивающим влияние помех. Сб. «Теория передачи сообщений». М., Изд-во иностр. литер., 1957, 139—157).
3. Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы научной информации. М., «Наука», 1965.
4. Фрумкина Р. М. Понимание текста в условиях ограниченного знания словаря. Сб. «Научно-техническая информация». М., ВИНИТИ, 1965, *4*, 44—48.
5. Miller G. A., Newman E. B. Tests of a statistical explanation of the rank-frequency relation for words in written English. American J. of Psychology, 1958, **71**, *1*, 209—218.
6. Шрейдер Ю. А. О формальном определении основных понятий семиотики. Сб. «Кибернетику на службу коммунизму». М., 1966, *3*.
7. Шаумян С. К. Структурная лингвистика. М., 1965.
8. Белецкий М. И. Модель русского языка, описывающая простые предложения без однородности. Сб. «Научно-техническая информация». М., ВИНИТИ, 1964, *7*, 37—42.
9. Ломковская М. В. Сб. «Научно-техническая информация». М., ВИНИТИ, 1965, *5*, *7*, 35—39.
10. Фитиалов С. Я., Цейтин Г. С. Оценка количества синтаксических структур при различных ограничениях. Кибернетика, в печати.
11. Шредингер Э. Статистическая термодинамика. М., Изд-во иностр. литер., 1948.
12. Шрейдер Ю. А. О вариационных принципах в лингвистике. Техническая кибернетика, 1966, *2*, 49—55.
13. Шрейдер Ю. А. О свойствах проективности языка. Сб. «Научно-техническая информация», М., ВИНИТИ, 1964, *5*.
14. Шрейдер Ю. А. Некоторые проблемы теории научной информации. Сб. «Научно-техническая информация», М., ВИНИТИ, 1966, *6*, 17—22.

Поступила в редакцию
15 января 1966 г.

Примечание. В недавно вышедшей работе Haigt, Frank A. Some Statistical Problems in Connection with Word Association Data. J. of Mathem. Psychol., 1966, 3, *1* приведены 11 примеров разнообразных ситуаций, где возникает распределение Ципфа и исследованы специфические методы статистической оценки параметров этого распределения.