ИЗДАТЕЛЬСТВО «НАУКА»
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ

СТАТИСТИКА РЕЧИ

АКАДЕМИЯ НАУК СССР

СТАТИСТИКА РЕЧИ

АКАДЕМИЯ НАУК СССР
НАУЧНЫЙ СОВЕТ ПО КИБЕРНЕТИКЕ
СЕКЦИЯ СЕМИОТИКИ

СТАТИСТИКА РЕЧИ

ИЗДАТЕЛЬСТВО «НАУКА»
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ
ЛЕНИНГРАД · 1968

Сборник представляет собой коллективную монографию, посвященную статистическому и теоретико-информационному описанию русского, английского, немецкого, французского и некоторых других европейских языков на буквенном, лексическом и фразеологическом уровнях. Представленные в сборнике статистические материалы и результаты исследований могут быть использованы для целей машинной переработки языковой информации, в построении теории разборчивости речи, при построении математически обоснованной методики преподавания языков и для решения некоторых лингвистических задач.

Сборник рассчитан на научных работников — математиков, специалистов по теории связи, языковедов, интересующихся применением кибернетических и вероятностно-статистических методов к исследованию естественных языков. Сборник может быть использован также в качестве учебного пособия при преподавании математического языкознания и некоторых лингвистических дисциплин.

7-1-1

192—67 (I пол.)

ВВЕДЕНИЕ

В настоящий сборник включены выполненные в период с 1960 по 1965 г. работы членов группы «Статистика речи», а также работы, подготовленные по тематике группы другими авторами.

Возникшая в 1959 г. группа «Статистика речи» в настоящее время объединяет пять коллективов — ленинградский, белорусский, дагестанский, среднеазиатский и молдавский, в работе которых участвует более ста языковедов, математиков и инженеров.

Группа работает в трех направлениях: во-первых, над выяснением кодовых (энтропийных) свойств языка, во-вторых, над определением статистических характеристик словаря, грамматики, фразеологии, в-третьих, над автоматическим анализом текста. Эти исследования ведутся на материале научно-технической, публицистической и разговорной речи.

Внутреннее единство всех исследований, осуществляемых в группе, определяется их основной целенаправленностью. Она заключается в том, чтобы получить такие информационно-статистические описания речи, на основе которых можно было бы создать работающие программы по массовому автоматическому анализу и синтезу научно-технических, публицистических и разговорных текстов по основным европейским языкам.

В рамках этой общей целенаправленности выделяются более частные задачи.

Первая задача состоит в том, чтобы получить оценки таких общих характеристик, как энтропия и избыточность языка, а также определить долю лексико-грамматических связей в общей контекстной обусловленности лингвистических единиц. Этим вопросам посвящены статьи в первой части сборника. Общие принципы и методика извлечения кодовых характеристик языка изложены в статьях Н. В. Петровой и Г. П. Богуславской.

Вторая задача заключается в том, чтобы дать конкретное статистическое описание разных аспектов языка — лексики, фразеологии, грамматики, а также фонемно-графемного уровня. Этой задаче посвящены работы, объединенные во второй части сборника, причем основной упор здесь делается на статистическое моделирование лексики. В статьях П. М. Алексеева, Е. А. Калининой и В. М. Калинина описывается методика построения лексико-статистических моделей на основе частотных списков слов (словоформ), излагаются некоторые принципы их лингвистического и

математического анализа, а также раскрываются перспективы их использования при построении объективной методики изучения языков. В статье А. В. Зубова, К. Ф. Лукьяненкова, Р. Г. Пиотровского и Э. Н. Хотяшова представлены некоторые результаты и ближайшие перспективы статистического описания текста с помощью электронно-вычислительных машин. В работах Л. И. Ешана, И. А. Исенина, Л. А. Кочетковой и Л. М. Скрелиной, Л. А. Турко, Л. А. Турыгиной и др. даются частотные списки словоформ и слов для конкретных подъязыков.

Осуществляя моделирование лексики с помощью частотных словарей, исследователь не учитывает вероятностных связей между лексическими единицами текста. Но без учета этих связей невозможно осуществить автоматическую переработку текста, в связи с чем возникает необходимость разработать приемы для статистического описания сочетаемости слов в тексте. Поиски в этом направлении отражены в коллективной статье М. В. Данейко, Л. Е. Машкиной, О. А. Нехай, В. А. Соркиной, А. Н. Шаранды, а также в работе Л. Г. Кравца. В первой статье предлагается чисто формальная процедура для выделения трехсловных сочетаний. В статье Л. Г. Кравца выделение именных словосочетаний осуществляется исходя из лингвистической интуиции носителя языка. Хотя последние две работы не всегда дают достаточно достоверные статистические сведения, они включены в сборник с целью привлечь внимание языковедов и математиков к совершенно еще не разработанной проблеме марковских связей внутри лингвистического текста.

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ[1]

H — энтропия.
I — информация.
R — избыточность.
p — вероятность.
n — номер буквы в тексте.
S — длина алфавита.
i — ранг (порядковый номер в частотном списке) словоформы.
N — объем выборки в словоупотреблениях.
F — абсолютная частота встречаемости словоформы.
f — относительная частота встречаемости словоформы.
F^* — накопленная абсолютная частота по N.
f^* — накопленная относительная частота по N.
L — количество разных словоформ, составляющих словарь (объем словника).
m — количество словоформ, имеющих данную частоту.
δ — относительная ошибка.
k, γ, ϱ — параметры закона Ципфа—Мандельброта.
Δ — пробел.

[1] Все отклонения от приведенных выше обозначений, а также вновь вводимые символы специально оговариваются.

Часть I. ***ИНФОРМАЦИОННЫЕ ИЗМЕРЕНИЯ ЯЗЫКА И ТЕКСТА***

Н. В. Петрова

КОДОВЫЕ ХАРАКТЕРИСТИКИ ПИСЬМЕННОГО ТЕКСТА

Глава I. МЕТОДЫ ИССЛЕДОВАНИЯ

§ 1. Задачи работы

Целью настоящей работы является описание некоторых экспериментов по определению основных кодовых характеристик письменного текста. Теоретические положения работы рассматриваются на материале трех основных разновидностей французского литературного языка (стилей). Это:

1) литературно-разговорная речь в ее письменной фиксации (в дальнейшем — разговорная речь);

2) беллетристический стиль;

3) научно-деловая речь (в дальнейшем — научно-деловой стиль). Исходные понятия и основные приемы для определения подобных статистико-информационных характеристик речи и языка были предложены еще в классических работах К. Шеннона.[1]

§ 2. Определение верхней и нижней границ энтропии письменного текста по методу Шеннона

Энтропия языка определялась Шенноном как предел, к которому стремится ряд последовательных приближений $H_0H_1 \ldots H_n$, каждое из которых учитывает все более далекие статистические связи языка, т. е.

$$H_\infty = \lim_{n\to\infty} H_n, \tag{1}$$

где H_n — среднее значение условной энтропии буквы, если известны предшествующие $n-1$ букв:

$$H_n = -\sum_{b_1^{n-1}} p\left(b_1^{n-1}\right) \sum_{j} p\left(j/b_1^{n-1}\right) \log p\left(j/b_1^{n-1}\right), \tag{2}$$

[1] К. Шеннон. 1) Математическая теория связи. Работы по теории информации и кибернетике. М., 1963, стр. 243—332; 2) Предсказание и энтропия английского печатного текста. Там же, стр. 669—686.

где b_i^{n-1} — некоторое сочетание из $n-1$ букв; $p(b_i^{n-1})$ — вероятность появления сочетания b_i^{n-1}; j — буква, следующая за b_i^{n-1}; $p(j/b_i^{n-1})$ — условная вероятность появления буквы j после b_i^{n-1}.

Для оценки энтропии письменного текста К. Шенноном был разработан экспериментальный метод. Основной предпосылкой этого метода является предположение, что при проведении эксперимента по угадыванию букв неизвестного текста испытуемый, являющийся носителем данного языка, должен наиболее рационально предсказывать появление той или иной буквы исходя из предыдущего контекста. Высказывая такое предположение, мы допускаем, что в языковом сознании каждого человека заданы не только единицы его родного языка и их структурные связи, но также и вероятностно-статистические характеристики этих единиц.

При этом вся совокупность предсказаний испытуемого рассматривается как определенная система лингво-психологических реакций, которая обнаруживает вероятностно-статистические связи, объективно существующие в речи.

Существует по крайней мере три варианта метода угадывания.

Первый вариант заключается в том, что испытуемому, говорящему на данном языке, дается неизвестный текст и предлагается последовательно отгадывать буквы этого текста. Если буква отгадана правильно, переходят к угадыванию следующей буквы. В случае ошибки испытуемому сообщается об этом и предлагается отгадывать снова. Так продолжается до тех пор, пока не будет найдена правильная буква. При этом фиксируется число попыток, необходимых для угадывания той или иной буквы.

Покажем, что при использовании этого приема можно определить верхнюю $\bar{H}_n''$ и нижнюю $\underline{H}_n$ оценки энтропии H_n и что эти оценки удовлетворяют неравенству (3)[2]

$$\underline{H}_n = \sum_{k=1}^{S} k\,(q_k^n - q_{k+1}^n)\log k \leqslant H_n \leqslant -\sum_{k=1}^{S} q_k^n \log q_k^n = \bar{H}_n''. \qquad (3)$$

где q_k^n — вероятность правильно угадать n-ную букву, с k-той попытки.

Предположим, что угадыватель идеальный, это значит, что

$$q_k^n = \sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1}), \qquad (4)$$

где

$$p(j_{i,1}/b_i^{n-1}) \geqslant p(j_{i,2}/b_i^{n-1}) \geqslant \ldots \geqslant p(j_{i,k}/b_i^{n-1}) \geqslant p(j_{i,k+1}/b_i^{n-1}) \geqslant \ldots \geqslant p(j_{i,S}/b_i^{n-1})$$

[2] А. П. Савчук. Об оценках энтропии языка по Шеннону. «Теория вероятностей и ее применения», IX, 1, 1964, стр. 154—157.

для любого b_i^{n-1}. Отсюда $q_k^n \geqslant q_{k+1}^n$, т. е. угадывание происходит в порядке убывания условных вероятностей.

Для получения верхней оценки применим неравенство Иенсена

$$p_1 f(x_1) + p_2 f(x_2) + \ldots + p_n f(x_n) \leqslant f(p_1 x_1 + \ldots + p_n x_n), \qquad (5)$$

где $f(x)$ — функция, выпуклая[3] в интервале (a, b), и $\sum_{k=1}^{n} p_k = 1$ $(p_k \geqslant 0)$; тогда получим (см. формулу 4)

$$H_n = -\sum_{j,\, b_i^{n-1}} p(b_i^{n-1})\,p(j/b_i^{n-1}) \log p(j/b_i^{n-1}) =$$

$$= -\sum_{k=1}^{S} \sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1}) \log p(j_{i,k}/b_i^{n-1}) \leqslant$$

$$\leqslant -\sum_{k=1}^{S} \left[\sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1})\right] \log \sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1}) =$$

$$= -\sum_{k=1}^{S} q_k^n \log q_k^n = \bar{H}_n''. \qquad (6)$$

Верхняя оценка достигается тогда и только тогда, когда

$$\sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1}) \log p(j_{i,k}/b_i^{n-1}) = \left[\sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1})\right] \times$$

$$\times \log \sum_{b_i^{n-1}} p(b_i^{n-1})\,p(j_{i,k}/b_i^{n-1}), \quad k = 1, 2 \ldots S,$$

т. е. когда при каждом k неравенство (5) превращается в равенство. А это при $p_i < 1$ происходит тогда и только тогда, когда все x_i (здесь $= p(j_{i,k}/b_i^{n-1})$) равны между собой. Т. е. $p(j_{i,k}/b_i^{n-1})$ не зависит от b_i^{n-1} (из условия $p(b_i^1) < 1$ следует, что и $p(b_i^{n-1}) = p_i < 1$), а это значит, что если для каждого b_i^{n-1} упорядочить множество букв языка по убыванию вероятностей их появления после b_i^{n-1}, т. е. $j_{i,1}, j_{i,2}, \ldots j_{i,s}$ (6) и принять во внимание, что $p(j_{i,1}/b_i^{n-1}) \geqslant p(j_{i,2}/b_i^{n-1}) \geqslant \ldots \geqslant p(j_{i,k}/b_i^{n-1}) \geqslant p(j_{i,k+1}/b_i^{n-1}) \geqslant \ldots$ $\ldots \geqslant p(j_{i,s}/b_i^{n-1})$, то $p(j_{i,k}/b_i^{n-1})$ не зависит от i, а зависит только от k и n. Однако расположение букв в последовательности (6) зависит от b_i^{n-1}.

Таким образом, в реальном языке верхняя оценка не достигается.

Для получения нижней оценки применим вспомогательное неравенство

$$-\sum_{k=1}^{n} p_k \log p_k \geqslant \sum_{k=1}^{n} k\,(p_k - p_{k+1}) \log k, \qquad (7)$$

[3] Функция $f(x)$ считается выпуклой на отрезке (a, b), если вторая производная этой функции отрицательна на этом отрезке.

где

$$p_1 \geqslant p_2 \geqslant \ldots \geqslant p_k \geqslant p_{k+1} \geqslant \ldots \geqslant p_n;$$

$$\sum_{k=1}^{n} p_k = 1; \quad 0 \leqslant p_k \leqslant 1; \quad p_{n+1} = 0;$$

тогда получим

$$H_n = -\sum_{j,\, b_i^{n-1}} p(b_i^{n-1})\, p(j/b_i^{n-1}) \log p(j/b_i^{n-1}) =$$

$$= -\sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} p(j_{i,k}/b_i^{n-1}) \log p(j_{i,k}/b_i^{n-1}) \geqslant$$

$$\geqslant \sum_{b_i^{n-1}} p(b_i^{n-1}) \sum_{k=1}^{S} k\,[p(j_{i,k}/b_i^{n-1}) - p(j_{i,k+1}/b_i^{n-1})] \log k =$$

$$= \sum_{k=1}^{S} k \log k \left[\sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k}/b_i^{n-1}) - \sum_{b_i^{n-1}} p(b_i^{n-1})\, p(j_{i,k+1}/b_i^{n-1}) \right] =$$

$$= \sum_{k=1}^{S} k\,(q_k^n - q_{k+1}^n) \log k = H_n.$$

Нижняя оценка достигается тогда и только тогда, когда для каждого b_i^{n-1} такого, что $p(b_i^{n-1}) > 0$,

$$\sum_{k=1}^{S} p(j_{i,k}/b_i^{n-1}) \log p(j_{i,k}/b_i^{n-1}) = \sum_{k=1}^{S} k\,[p(j_{i,k}/b_i^{n-1}) - p(j_{i,k+1}/b_i^{n-1})] \log k,$$

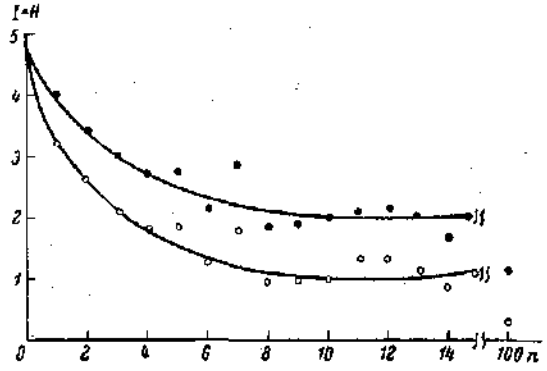


Рис. 1. Верхняя и нижняя оценки энтропии английского текста (по Шеннону).

По оси абсцисс — число букв, по оси ординат — энтропия (количество информации), в дв. ед.

т. е. неравенство (7) переходит в равенство. А это происходит тогда и только тогда, когда $p_k = \frac{1}{m}$ при $k < m$; $p_k = 0$ при $k \geqslant$

$\geqslant m$, где $m = 1, 2, \ldots S$ (здесь $p_k = p(j_{i,k}/b_i^{n-1})$), т. е. неравенство переходит в равенство, когда

$$p(j_{i,k}/b_i^{n-1}) = \begin{cases} \frac{1}{m} & \text{при } k \leqslant m, \\ 0 & \text{при } k > m. \end{cases}$$

Это значит, что после любого сочетания из $n-1$ букв, имеющего ненулевую вероятность появления, некоторые m букв равновероятны, а вероятности появления остальных $n-m$ букв равны нулю.

Так как в реальном языке это условие не выполняется, то нижняя оценка также не достигается.

Пользуясь этим методом, Шеннон определил значения верхней и нижней границ английской письменной речи вплоть до значения $n = 15$ и, кроме того, при $n = 100$ (рис. 1).

§ 3. Определение дальности действия лингво-статистических связей

Для оценки дальности действия лингво-статистических связей, т. е. для определения номера n, при котором значение энтропии H_n практически приближается к значению $H_\infty \left(H = \lim_{n\to\infty} H_n\right)$, Н. Бар-

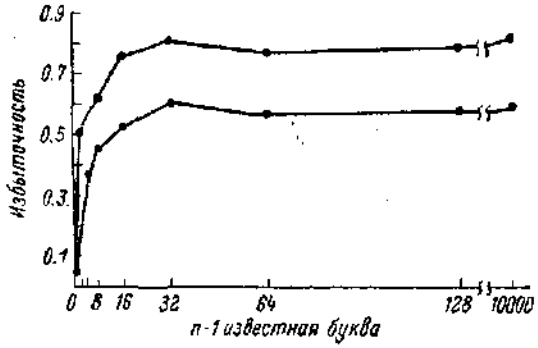


Рис. 2. Верхняя и нижняя оценки энтропии английского текста (по Бартону и Ликлайдеру).

тон и Дж. Ликлайдер,[4] пользуясь методом Шеннона, определили зависимость избыточности от числа известных букв текста (вплоть до значения $n \approx 10\,000$).

[4] N. C. Burton, J. C. R. Licklider. Long-range constraints in the statistical structure of printed English. «The American Journal of Psychology», v. 68, № 4, 1955, pp. 650—653.

При этом из 10 различных источников было сделано 10 выборок следующей длительности: $n-1=0,\ 1,\ 2,\ 4,\ 8,\ 16,\ 32,\ 64,\ 128$ и $\sim 10\,000$ букв. Для всех этих отрывков были определены верхняя и нижняя границы энтропии соответствующей буквы и верхняя и нижняя границы избыточности. Результаты исследования этих авторов приведены на рис. 2.

Как видно из графика рис. 2, по мере увеличения длины текста сверх определенной (около 30—32-й буквы) избыточность практически не увеличивается.[5] Таким образом, авторы приходят к выводу, что избыточность растет приблизительно до 32-й буквы текста, оставаясь в дальнейшем постоянной и равной примерно 65%.

Приведенные выше оценки нижней границы энтропии английского печатного текста находят косвенные подтверждения в эксперименте Дж. Миллера и Э. Фридман.[6]

Авторы исследовали способность человека исправлять искажения, умышленно вводимые в печатный английский текст. Было показано, что человек средних лингвистических способностей в определенно заданное время (10 мин.) не в состоянии декодировать текст, если в нем искажено более 10% букв. Работа усложняется при хаотической замене знаков. Отдельные более развитые испытуемые при неограниченном времени декодирования были в состоянии восстанавливать до 50% текста в случае пропуска чередующихся знаков или в случае пропуска гласных и пробелов. Эти данные соответствуют нижней границе избыточности английского печатного текста $\sim$60—70%, полученной К. Шенноном, а также Н. Бартоном и Дж. Ликлайдером.

§ 4. Сужение интервала между верхней и нижней границами энтропии. Учет «нулей информации»

Как уже говорилось, истинное значение энтропии находится в интервале, заключенном между верхней и нижней границами, полученными в результате применения метода Шеннона. Этот интервал достаточно широк (в экспериментах Шеннона, а также Бартона-Ликлайдера он достигает одной двоичной единицы). Иными словами, погрешность при определении истинного значения энтропии достигает здесь $\pm 33\%$.

Более точные оценки истинного значения энтропии можно получить путем сужения интервала между верхней и нижней границами. Такое сужение можно получить, если учитывать случаи, когда появление буквы текста полностью предопределено предшествующей ей последовательностью букв данного слова или лексическим контекстом,[7] иначе говоря, те случаи, когда энтропия буквы равна нулю.

Ясно, что учет подобных случаев достоверных продолжений (так называемых нулей информации) неизбежно должен сказываться на конечных результатах эксперимента. Он приводит к снижению верхней границы энтропии и, следовательно, к интересующему нас сужению интервала, в котором заключено истинное значение энтропии.

Это впервые было показано Р. Г. Пиотровским [8] с соавторами на материале русского языка. В данном случае авторы использовали формулу Шеннона (3) для определения верхней границы энтропии, в которую А. Н. Колмогоровым был введен поправочный член, учитывающий влияние «нулей информации». При этом приводятся следующие рассуждения.

Вероятность недостоверных продолжений текста (т. е. угадывания с одной и более попыток) равна $1-p_0$. Положив, что вероятность угадывания с k попыток относительно общего числа недостоверных продолжений равна $p'_k=\frac{p_k}{1-p_0}$ при $k\geqslant 0$, получаем верхнюю границу интервала в виде

$$\bar{H}_n=(1-p_0)(-p'_1\log_2 p'_1-p'_2\log_2 p'_2-\ldots-p'_S\log_2 p'_S),$$

где выражение $(-p'_1\log_2 p'_1-p'_2\log_2 p'_2-\ldots-p'_S\log_2 p'_S)$ представляет собой оценку верхнего интервала энтропии по Шеннону, а множитель $(1-p_0)$ указывает на вероятность этой оценки. Эта оценка, записанная в вероятностях p_k, получает вид:

$$\bar{H}_n=(1-p_0)\left(-\frac{p_1}{1-p_0}\log_2\frac{p_1}{1-p_0}-\right.$$

$$\left.-\frac{p_2}{1-p_0}\log_2\frac{p_2}{1-p_0}-\ldots-\frac{p_S}{1-p_0}\log_2\frac{p_S}{1-p_0}\right).$$

Все выражение после преобразований принимает вид:

$$\bar{H}_n=-p_1[\log_2 p_1-\log_2 p_2(1-p_0)]-p_2[\log_2 p_2-\log_2(1-p_0)]-\ldots$$
$$\ldots-p_S[\log_2 p_S-\log_2(1-p_0)]=-p_1\log_2 p_1-p_2\log_2 p_2-\ldots$$
$$\ldots-p_S\log_2 p_S+(p_1+p_2+\ldots+p_S)\log_2(1-p_0),$$

а, поскольку $p_1+p_2+\ldots+p_S=1-p_0$, все выражение преобразуется в следующую формулу:

$$\bar{H}_n=(1-p_0)\log_2(1-p_0)-p_1\log_2 p_1-p_2\log_2 p_2-\ldots-p_S\log_2 p_S. \quad (8)$$

[5] См. также: А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин. Энтропия русского языка. ВЯ, XI, 6, 1962, стр. 115—130.

[6] G. Miller, E. Friedman. The reconstruction of multilated english texts. «Information and Control», v. I, № 1, 1957, pp. 38—55.

[7] В качестве примера достоверного продолжения можно охарактеризовать конечное *t* и идущий за ним пробел в словосочетании *il me visitait*.

[8] А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин, ук. соч., стр. 115—130.

§ 5. Упрощенный вариант метода Шеннона

Для оценки верхней границы энтропии может быть также использован упрощенный вариант метода Шеннона, который отличается от описанного выше тем, что в случае, когда информант дает неправильный ответ, экспериментатор сообщает ему истинную букву текста, фиксируя в протоколе, что испытуемому понадобилось *не менее двух попыток для отгадывания данной буквы.*

При обработке результатов сокращенной программы угадывания верхняя оценка энтропии ($\bar{H}'_n$) представляет собой сумму из двух членов.[9] Первый член указывает на полученную испытуемым с вероятностью $(1 - p_0)$ информацию о том, правильно его первое угадывание или неправильно. Эта информация численно равна

$$-p'_1 \log_2 p'_1 - (1 - p'_1) \log_2 (1 - p'_1).$$

Второй член представляет собой информацию, получаемую испытуемым с вероятностью $(1-p_0-p_1)$ при сообщении ему экспериментатором правильной буквы в том случае, когда не было достоверного продолжения и угадывание с первой попытки не удалось. Эта информация всегда меньше информации, получаемой после правильного угадывания буквы текста при условии, что испытуемому *были известны две предшествующие буквы (начиная со* второй буквы текста испытуемому известны по крайней мере две буквы, предшествующие угадываемой букве). Последняя информация и равна количественно H_{III}. В результате всех этих рассуждений получаем

$$\bar{H}'_n = H_{III}(1 - p_0 - p_1) - p_1 \log_2 p_1 + (1 - p_0) \log_2 (1 - p_0) - \\ - (1 - p_0 - p_1) \log_2 (1 - p_0 - p_1). \qquad (9)$$

§ 6. Метод «коллективного» эксперимента

Для определения верхней границы энтропии можно провести эксперимент, построенный на иных принципах организации и обработки данных.

Группе испытуемых предлагалось последовательно угадывать буквы в определенных контрольных словах, употребляемых либо в тексте, либо вне контекста.

Испытуемым сообщался предшествующий слову текст (в случае, если слово бралось в контексте) или первая буква отдельно взятого слова, после чего каждый испытуемый записывал наиболее вероятную с его точки зрения следующую букву. Все предложения испытуемых фиксировались экспериментатором в протоколе, и им сообщалась правильная буква. Затем испытуемые снова давали свои продолжения для следующей за отгаданной буквой.

[9] Рассматриваемый вывод формулы (9) также был дан академиком А. Н. Колмогоровым.

В результате экспериментатор получал распределения вероятностей для каждой буквенной позиции словоформы (p_n^k) при условии, что испытуемому был известен предшествующий данной позиции буквенный или лексико-грамматический контекст.

По этим данным можно определить верхнюю границу энтропии по формуле

$$\bar{H}''_n = -\sum_{k=1}^{s} p_k^n \log_2 p_k^n. \qquad (10)$$

Используя описанный метод, И. В. Матковский [10] получил оценку энтропии молдавского печатного текста.

§ 7. Определение состоятельной оценки энтропии по методу А. Н. Колмогорова

В последние годы акад. А. Н. Колмогоровым и его сотрудниками [11] был разработан метод, который позволяет определить состоятельную оценку истинной энтропии, а не оценки верхней и нижней ее границ. Идеальная схема эксперимента выглядит следующим образом: выбирается произвольный текст длиной в N букв и предполагается распределение вероятностей появления букв алфавита *для* t-й буквы $[p_k(t);\ k=1, 2, \ldots S]$ при условии, что известны первые $t-1$ букв. При этом t принимает последовательно значения от 1 до N. В предположении оптимальной стратегии угадывания и теоретической схемы речи как стационарного случайного процесса с затухающими связями для энтропии языка предлагается оценка

$$\hat{H} = -\frac{\sum_{t=1}^{N} \log p_{x(t)}(t)}{N}, \qquad (11)$$

где $x(t)$ — буква, стоящая в тексте на t-м шагу; $p_{x(t)}(t)$ — вероятность буквы $x(t)$ в предположенном на t-м шагу распределении вероятностей. Пусть $q_k(t)$ $(k=1, 2, \ldots S)$ — распределение вероятностей появления букв на t-м шагу $(t=1, 2, \ldots N)$ нашего текста, которое имеет место в действительности.

Тогда математическое ожидание величины $\hat{H}$

$$M\hat{H} = -\frac{\sum_{t=1}^{N} M \log p_{x(t)}(t)}{N} = -\frac{\sum_{t=1}^{N} \sum_{k=1}^{S} q_k(t) \log p_k(t)}{N}.$$

[10] Координационное совещание по сравнительному и типологическому изучению романских языков. Л., 1964, стр. 29.

[11] Описание данного метода приводится по материалу А. П. Савчук, любезно предоставленному ею в наше распоряжение. См. также: А. М. Jaglom, J. M. Jaglom. Wahrscheinlichkeit und Information. Berlin, 1965, SS. 216—218.

Всегда имеет место неравенство

$$-\sum_{k=1}^{n} q_k \log p_k \geqslant -\sum_{k=1}^{n} q_k \log q_k,$$

так как $-\sum_{k=1}^{n} q_k \log p_k = -\sum_{k=1}^{n} q_k \log \frac{p_k}{q_k} - \sum_{k=1}^{n} q_k \log q_k,$

и $\sum_{k=1}^{n} q_k \log \frac{p_k}{q_k} \geqslant 0.$

Таким образом,

$$M\hat{H} \geqslant -\frac{\sum_{t=1}^{N}\sum_{k=1}^{S} q_{k(t)} \log q_{k(t)}}{N}.$$

Величина $-\sum_{k=1}^{S} q_{k(t)} \log q_{k(t)} = H_{b_t^{t-1}} = \tilde{H}_t$ есть не что иное, как условная энтропия буквы, вычисленная при условии, что известно $N-1$ букв нашего текста.

Следовательно,

$$M\hat{H} \geqslant \frac{\tilde{H}_{\mathrm{I}} + \tilde{H}_{\mathrm{II}} + \ldots + \tilde{H}_N}{N}$$

или

$$M\hat{H} \geqslant M\left(\frac{\tilde{H}_{\mathrm{I}} + \tilde{H}_{\mathrm{II}} + \ldots + \tilde{H}_N}{N}\right) = \frac{M\tilde{H}_{\mathrm{I}} + M\tilde{H}_{\mathrm{II}} + \ldots + M\tilde{H}_N}{N}.$$

Но

$$M\tilde{H}_t = \sum_{b_t} p\left(b_t^{t-1}\right) H_{b_t^{t-1}} = H_t,$$

или

$$M\hat{H} > \frac{H_{\mathrm{I}} + H_{\mathrm{II}} + \ldots + H_N}{N} = H_N^{\text{удельн.}},$$

но

$$\lim_{N\to\infty} H_N^{\text{удельн.}} = \lim_{N\to\infty} H_N = H$$

и $M\hat{H} = H$.

Предположив, что $p\{|H - M\hat{H}| < \delta\} \geqslant 1-\varepsilon$ $\varepsilon, \delta \geqslant 0$, можно считать $\hat{H}$ верхней оценкой для H.

Однако практически невозможно дать распределение вероятностей букв. Значения величин $\{p_k(t),\ k=1,\ 2,\ \ldots\ S;\ t=1,\ 2,\ \ldots,\ N\}$ можно определить из результатов эксперимента по угадыванию букв на t-м шагу, если известны предшествующие $t-1$ букв $(t=1,\ 2,\ \ldots,\ N)$.

В ходе эксперимента испытуемому предлагается угадать последовательно 50 знаков текста (предшествующий контекст ему известен).

В этой ситуации различаются следующие типы угадывания:

1) угадывание с большой степенью уверенности,
2) угадывание с меньшей степенью уверенности,
3) предлагаются 2 буквы с равной вероятностью,
4) предлагаются 2 буквы, из которых одна угадывается с большей уверенностью, чем вторая,
5) отказ от угадывания.

Рассмотрим каждый из этих типов в отдельности.

1. Буква $k_0(t)$ угадывается с большой степенью уверенности,

$$p_{k_0(t)}(t) = 1 - \alpha_0,$$

тогда $p_{k(t)}(t) = \alpha_0 \frac{\bar{p}_k}{1-\bar{p}_{k_0}}$ для $k \neq k_0$, где $\bar{p}_k$ — значение из таблицы безусловного распределения вероятностей букв для вероятности буквы с номером k $(1 \leqslant k \leqslant S)$. При этом, если мы угадываем правильно (пусть число исходов равно β_1), то $\varkappa(t) = k_0(t)$; и $-\log p_{\varkappa(t)}(t) = -\log(1-\alpha_0)$, если не угадываем (число исходов β_2), то

$$-\log p_{\varkappa(t)}(t) = -\log \varkappa_{a0} - \log \frac{\bar{p}_\varkappa}{1-\bar{p}_{k_0}}.$$

2. Буква $k_0(t)$ угадывается с меньшей степенью уверенности,

$$p_{k_0(t)}(t) = 1 - \alpha_1,$$

тогда $p_{k(t)}(t) = \alpha_1 \frac{\bar{p}_k}{1-\bar{p}_{k_0}}$ для $k \neq k_0$.

Здесь при правильном угадывании (число исходов β_3)

$$\varkappa(t) = k_0(t), \text{ и } -\log p_{\varkappa(t)}(t) = -\log(1-\alpha_1);$$

при неправильном угадывании (число исходов β_4)

$$-\log p_{\varkappa(t)}(t) = -\log \alpha_1 - \log \frac{\bar{p}_\varkappa}{1-\bar{p}_{k_0}}.$$

3. Угадываются две буквы $k_0(t)$ и $k_1(t)$ с равной вероятностью, $p_{k_0(t)}(t) = p_{k_1(t)}(t) = \frac{\alpha_2}{2}$, тогда $p_{k(t)}(t) = (1-\alpha_2)\frac{\bar{p}_k}{1-\bar{p}_{k_0}-\bar{p}_{k_1}}$. При этом при правильном угадывании (число исходов β_5)

$$\varkappa(t) = \begin{cases} k_0(t) \\ k_1(t) \end{cases} \text{ и } -\log p_{\varkappa(t)}(t) = -\log \frac{\alpha_2}{2},$$

при ошибочном угадывании (число исходов β_6)

$$-\log p_{\varkappa(t)}(t) = -\log(1-\alpha_2) - \log \frac{\bar{p}_\varkappa}{1-\bar{p}_{k_0}-\bar{p}_{k_1}}.$$

4. Угадываются две буквы $k_0(t)$ и $k_1(t)$, причем $k_0(t)$ с бо́льшей уверенностью, чем $k_1(t)$.

$$p_{k_0(t)}(t)=\alpha_3,\quad p_{k_1(t)}(t)=\alpha_4,$$

тогда $p_{k(t)}(t)=(1-\alpha_3-\alpha_4)\frac{\bar p_k}{1-\bar p_{k_0}-\bar p_{k_1}}$. Если угадываем правильно (число исходов β_7 и β_8), то

$$x(t)=k_0(t),\ \text{и}\ -\log p_{x(t)}(t)=-\log\alpha_3;$$
$$x_1(t)=k_1(t),\ \text{и}\ -\log p_{x(t)}(t)=-\log\alpha_4.$$

При ошибочном угадывании (число исходов β_9)

$$-\log p_{x(t)}(t)=-\log(1-\alpha_3-\alpha_4)-\log\frac{\bar p_x}{1-\bar p_{k_0}-\bar p_{k_1}}.$$

5. Отказ от угадывания дает

$$p_{k(t)}(t)=\bar p_k.$$

В этом случае (число отказов β_{10})

$$-\log p_{x(t)}(t)=-\log\bar p_x.$$

При выполнении всех этих условий формула (11) примет вид

$$\hat H=\hat H(\alpha_1;\ \alpha_2;\ \alpha_3;\ \alpha_4;\ \alpha_0)=-\frac{1}{N}[\beta_1\log(1-\alpha_0)+\beta_2\log\alpha_0+\beta_3\log(1-\alpha_1)+$$
$$+\beta_4\log\alpha_1+\beta_5\log\frac{\alpha_2}{2}+\beta_6\log(1-\alpha_2)+\beta_7\log\alpha_3+\beta_8\log\alpha_4+$$
$$+\beta_9\log(1-\alpha_3-\alpha_4)]+\sum_{\beta_2;\,\beta_4}\log\frac{\bar p_x}{1-\bar p_{k_0}}+$$
$$+\sum_{\beta_6;\,\beta_9}\log\frac{\bar p_x}{1-\bar p_{k_0}-\bar p_{k_1}}+\sum_{\beta_{10}}\log\bar p_x;$$

Обозначим $\sum_{\beta_2;\,\beta_4}\log\frac{\bar p_x}{1-\bar p_{k_0}}=\sum$ ошибок 1-го класса,

$\sum_{\beta_6;\,\beta_9}\log\frac{\bar p_x}{1-\bar p_{k_0}-\bar p_{k_1}}=\sum$ ошибок 2-го класса, а $\sum_{\beta_{10}}\log\bar p_x=\sum_{\text{отк.}}$ и $\frac{\Sigma_{\text{ош. 1 кл.}}+\Sigma_{\text{ош. 2 кл.}}+\Sigma_{\text{отк.}}}{N}=\sum$. Параметры α_i $(i=0,\ 1,\ \ldots,\ 4)$ определяются дополнительным условием: $\hat H=\min_{\alpha_0,\ \ldots,\ \alpha_4}\hat H(\alpha_0;\ \alpha_1,\ \ldots,\ \alpha_4)$; таким образом, для определения параметров α имеем уравнение

$$\frac{\partial\hat H}{\partial\alpha_i}=0;\ i=0,\ 1,\ \ldots,\ 4,$$

решение которого приводит к следующим значениям:

$$\alpha_0=\frac{\beta_2}{\beta_1+\beta_2};\quad \alpha_1=\frac{\beta_4}{\beta_3+\beta_4};\quad \alpha_2=\frac{\beta_5}{\beta_5+\beta_6};\quad \alpha_3=\frac{\beta_7}{\beta_7+\beta_8+\beta_9};$$
$$\alpha_4=\frac{\beta_8}{\beta_7+\beta_8+\beta_9}.$$

Внося эти значения в формулу (11), получим окончательно

$$\hat H=-\frac{\gamma_3}{N}[\alpha_0\log\alpha_0+(1-\alpha_0)\log(1-\alpha_0)]-\frac{\gamma_6}{N}[\alpha_1\log\alpha_1+$$
$$+(1-\alpha_1)\log(1-\alpha_1)]-\frac{\gamma_5}{N}[\alpha_2\log\alpha_2+(1-\alpha_2)\log(1-\alpha_2)]+\frac{\gamma_5}{n}\alpha_2\log 2-$$
$$-\frac{\gamma_7}{N}[\alpha_3\log\alpha_3+\alpha_4\log\alpha_4+(1-\alpha_3-\alpha_4)\log(1-\alpha_3-\alpha_4)]-\sum,\quad(12)$$

где $\gamma_3=\beta_1+\beta_2$; $\gamma_6=\beta_3+\beta_4$; $\gamma_5=\beta_5+\beta_6$; $\gamma_7=\beta_7+\beta_8+\beta_9$; α_0 — вероятность не угадать в случае (1), α_1 — в случае (2); $1-\alpha_2$ — в случае (3); α_3 — вероятность угадать букву k_0 в случае (4); α_4 — вероятность угадать букву k_1 в случае (5).

Пользуясь только что описанным методом, А. П. Савчук получила значение энтропии русского языка, равное 1.7 дв. ед.

§ 8. Недостатки изложенных методов

Как метод Колмогорова, так и все варианты метода Шеннона обладают одним общим недостатком: получаемые результаты зависят в известной степени от языкового «чутья», образования, степени усталости и даже настроения испытуемых. Для того чтобы получить результаты, характеризующие язык в целом, а не степень владения им данным испытуемым, приходится проводить дополнительные контрольные эксперименты и сопоставлять их результаты с данными основного эксперимента (см. гл. II, § 5).

§ 9. Объективные приемы определения энтропии

Существуют приемы определения энтропии письменного текста, свободные от недостатка, присущего методам угадывания. Этими приемами, которые мы называем объективными, являются: а) метод, основанный на статистической обработке частотных словарей [12] и б) метод Ньюмена и Герстмана.[13]

---

[12] См.: К. Шеннон, ук. соч., стр. 672; G. A. Barnard. Statistical calculation of word entropies for four western languages. «IRE, Transactions of Information Theory», v. 1, № 1, 1955, pp. 49—53; K. Küpfmüller. Die Entropie der deutschen Sprache. FTZ, H. 6, 1954, SS. 262—272.

[13] См.: E. B. Newman and L. I. Gerstman. A New Method for Analising Printed English. «Journal of the Experimental Psychology», XLIV, 2, 1952, pp. 114—125.

§ 10. Заключение

Итак, выделяются следующие четыре основных метода для определения энтропии письменной речи:

1) тестовый метод Шеннона в нескольких вариантах; этот метод дает возможность определить верхнюю и нижнюю границы интервала, в котором заключено истинное значение энтропии;

2) тестовый метод А. Н. Колмогорова, позволяющий получить состоятельную оценку энтропии;

3) метод, основанный на статистической обработке частотных словарей;

4) метод, предложенный Ньюменом и Герстманом.

Как уже указывалось, задачей настоящей работы является определение не только таких общих информационных характеристик французского печатного текста, как энтропия (количество информации), избыточность, но и их распределения в различных участках связного текста и отдельно взятого слова. Поэтому мы должны использовать в работе тот метод, который позволяет вскрыть это распределение на синтагматической оси.

Единственным методом, позволяющим решить указанную задачу, является метод Шеннона и его варианты. Тот факт, что метод Шеннона не дает возможности получить состоятельную оценку энтропии, а дает лишь приближения к границам интервала, в которых заключено истинное значение энтропии данной единицы текста (в нашем случае — буквы), не должен нас смущать. Для языкознания больший интерес представляет соотношение между количественными характеристиками отдельных лингвистических единиц, чем сами эти характеристики.

Использование в настоящей работе первого и упрощенного вариантов метода Шеннона дает возможность сопоставить наши данные по энтропии французского языка с аналогичными результатами, полученными тем же путем Р. Г. Пиотровским и его соавторами относительно русского и румынского языков,[14] а также с данными об энтропии английского языка, которые были получены Г. П. Богуславской (см. ее статью в настоящем сборнике). Другие методы (в первую очередь метод Колмогорова) привлекаются в работе для проверки результатов основного эксперимента.

[14] А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин, ук. соч., стр. 115—130; L. Novac, R. Piotrowski, Experimentul de predicţie şi entropia limbii române. «Studii şi cercetări lingvistice», 1968 (в печати).

Глава II. ЭНТРОПИЯ ФРАНЦУЗСКОГО ПИСЬМЕННОГО ТЕКСТА

§ 1. Таблицы двух- и трехбуквенных сочетаний

Определение H_1, H_2 и H_3

Как было показано выше (см. гл. I, § 2), для определения энтропии речи следует взять ряд последовательных приближений H_0, H_1. . .H_n к H_∞ как к пределу.

При этом H_n для малых значений n (H_1, H_2, H_3) может быть вычислена из таблиц частотностей отдельных букв,[15] двухбуквенных и трехбуквенных сочетаний.

Подобные таблицы были составлены нами на материале 30 000 букв, взятых из различных стилей современного французского языка — беллетристического, разговорного, научно-делового стилей речи и поэзии (табл. 1—5).

Таблица 1

Вероятности французских букв

Буква	Собственные данные	Данные Моро *	Буква	Собственные данные	Данные Моро *
Пробел	0,190	0,160	*m*	0,020	0,029
e	0,134	0,145	*r*	0,015	0,012
a	0,073	0,069	*q*	0,010	0,004
i	0,067	0,068	*f*	0,009	0,012
s	0,065	0,060	*j*	0,008	0,003
t	0,062	0,059	*g*	0,007	0,011
n	0,056	0,066	*h*	0,007	0,007
u	0,054	0,038	*b*	0,007	0,007
r	0,052	0,074	*x*	0,003	0,003
l	0,044	0,032	*y*	0,001	0,002
o	0,041	0,048	*z*	0,001	0,001
d	0,026	0,033	*k*	0,000	0,000
c	0,026	0,033	*w*	0,000	0,000
p	0,023	0,024			

* См.: R. Moreau, ук. соч.

[15] Как было показано Моро (см.: R. Moreau, Linguistique et télécommuniction, «L'Onde Electrique», XLII, 1962, pp. 731—737), частотность первых букв различна в различных текстах. Однако значение H_1 остается примерно постоянным и лежит в пределах 3.865—3.951 дв. ед. Частотность первых букв, полученная Моро на статистическом материале около 400 000 букв, также приведена в табл. 1.

Частотность двухбуквенных сочета

Таблица 2

ний французского языка ($p \cdot 10^4$)

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
a	—	12.0	32.0	7.7	—	4.7	12.3	2.0	166.0	1.7	—	25.6
b	11.0	—	—	—	6.3	—	—	—	11.0	—	—	17.0
c	18.0	—	2.3	—	50.0	—	—	38.0	16.0	—	—	7.7
d	23.0	—	—	—	126.0	—	—	—	30.0	—	—	—
e	13.0	4.0	37.0	7.3	17.0	5.3	9.0	3.0	11.0	2.0	—	29.0
f	15.0	—	—	—	11.0	13.0	—	—	17.0	—	—	7.3
g	10.0	—	—	—	31.0	—	—	—	4.3	—	—	3.0
h	18.3	—	—	—	36.0	—	—	—	4.3	—	—	—
i	9.3	7.3	11.0	8.7	63.0	6.3	13.0	—	—	—	—	95.0
j	27.0	—	—	—	27.0	—	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	—	—	—	—
l	72.0	10.0	—	—	130.0	0.3	17.0	0.3	22.0	—	—	29.0
m	47.0	7.7	—	—	68.0	—	—	—	22.0	—	—	—
n	24.0	—	26.0	28.0	80.0	3.7	13.0	2.7	20.0	—	—	—
o	—	3.7	9.7	1.7	6.0	3.0	1.3	1.0	61.0	—	—	9.7
p	49.0	—	0.3	—	42.0	—	—	3.0	7.3	—	—	19.3
q	—	—	—	—	—	—	—	—	—	—	—	—
r	59.0	1.7	8.7	14.0	178.0	—	3.0	—	35.0	—	—	2.0
s	42.0	—	5.3	—	86.0	0.3	—	—	34.0	—	—	—
t	50.0	—	0.3	—	96.0	—	—	2.3	37.0	—	—	—
u	8.3	3.3	7.0	4.7	73.0	6.3	1.0	—	50.0	2.7	—	18.3
v	35.0	—	—	—	49.0	—	—	—	26.0	—	—	1.7
w	—	—	—	—	—	—	—	—	—	—	—	—
x	2.0	—	1.0	—	2.0	—	—	—	0.3	—	—	—
y	1.7	0.3	0.3	—	4.7	—	—	—	—	—	—	0.3
z	1.0	—	—	—	1.0	—	—	—	—	—	—	—
Пробел	190.0	26.0	124.0	190.0	158.0	53.0	20.0	17.0	94.0	72.0	—	173.0

m	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*x*	*y*	*z*	Пробел
12.3	99.0	0.67	17.0	1.7	40.0	37.0	19.0	44.0	34.0	—	1.3	—	154.0
—	—	6.6	—	—	11.0	2.0	—	3.6	—	—	—	—	—
—	—	62.0	—	19.3	11.0	—	4.7	6.0	—	—	—	—	19.3
0.7	—	14.0	—	—	10.0	2.7	—	16.0	—	—	—	—	34.0
47.0	153.0	1.3	17.0	—	91.0	166.0	109.0	71.0	21.0	7.3	—	6.0	516.0
—	—	12.0	—	—	12.0	—	—	3.3	—	—	—	—	3.3
—	6.0	2.7	—	—	7.3	—	2.7	4.3	—	—	—	—	1.3
—	—	8.3	—	—	0.3	—	—	2.3	—	—	0.7	—	7.0
10.0	71.0	18.0	2.3	6.0	44.0	100.0	123.0	0.7	8.0	5.7	—	0.7	73.0
—	—	6.7	—	—	—	—	—	3.7	—	—	—	—	12.0
—	—	—	—	—	—	—	—	—		—	—	—	—
0.7	—	17.0	0.3	4.7	—	9.0	1.3	32.0	—	—	0.7	—	111.0
15.0	—	19.0	17.0	—	—	—	—	5.0	—	—	—	—	4.0
0.3	13.0	20.0	—	2.3	0.3	50.0	136.0	7.3	3.0	—	—	0.7	129.0
26.0	104.0	—	5.3	0.7	30.0	10.0	9.7	120.0	0.7	—	2.7	—	1.0
—	—	31.0	9.7	—	41.0	3.7	1.3	14.0	—	—	—	—	6.7
—	—	—	—	—	—	—	—	100.0	—	—	—	—	1.3
5.7	6.3	39.0	1.3	2.7	14.0	23.0	25.0	7.7	2.3	—	—	—	97.0
0.3	—	47.0	50.0	6.3	—	30.0	32.0	24.0	—	—	1.0	—	333.0
—	—	53.0	—	—	51.0	11.0	21.0	21.0	—	—	0.3	—	276.0
2.3	57.0	1.7	10.0	—	86.0	44.0	47.0	—	16.0	19.0	—	0.7	78.0
—	—	27.0	—	—	9.7	—	—	1.7	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	2.3	—	—	—	0.3	—	—	—	—	—	23.0
0.3	—	1.0	0.3	—	0.3	0.7	—	—	—	—	—	—	4.3
—	—	0.3	—	—	—	—	—	—	—	—	—	—	6.0
88.0	51.0	27.0	140.0	57.0	61.0	166.0	88.0	52.0	65.0	—	6.7	—	—

Таблица 3

Частотность ($p \cdot 10^4$) трехбуквенных сочетаний в современном французском языке

abi	1.0	arm	1.0	bul	1.0
abl	5.7	arq	1.3	but	1.0
abr	2.3	arr	4.3		
ace	3.0	art	3.3	cab	1.0
ach	5.3	ar(-)	7.0	cac	1.0
acq	19.3	ase	1.0	cai	1.0
ade	2.7	ass	6.7	cal	1.3
adi	2.7	ast	1.0	cam	1.0
aff	2.7	as(-)	27.3	car	1.0
afi	1.0	ati	9.3	cat	1.7
age	9.3	atr	1.0	cau	2.7
agn	1.0	att	4.7	ca(-)	4.3
aid	1.3	at(-)	1.7	cel	2.0
aie	8.3	auc	2.7	cen	1.3
aig	2.7	auf	1.0	cer	2.0
ail	3.0	auj	1.0	ces	6.3
aim	2.0	aul	2.0	cet	9.0
ain	10.7	aur	2.3	ceu	1.7
air	12.7	aus	8.3	cev	1.3
ais	43.6	aut	8.0	ce(-)	25.3
ait	68.6	auv	1.0	cha	12.3
ai(-)	12.7	aux	3.7	che	17.0
ajo	1.3	au(-)	13.7	chi	2.7
ala	1.7	ava	17.7	cho	4.7
alb	1.0	ave	10.7	ch(-)	1.0
ale	3.0	avi	1.7	cid	1.7
ali	5.7	avo	4.0	cie	3.7
all	5.3	aya	0.7	cil	2.3
alo	3.3			cin	2.0
al(-)	3.3	bac	0.7	cip	1.0
ama	3.0	bai	1.7	cit	1.7
amb	1.7	bal	1.3	ci(-)	1.7
ame	1.7	ban	0.7	cla	3.3
ami	4.0	bar	1.7	cle	2.3
anc	8.9	bas	2.3	cli	1.0
and	7.0	bat	1.3	clo	1.0
ang	5.3	bea	1.0	coe	1.3
anh	2.7	bes	1.7	col	3.3
ani	5.0	be	1.3	com	14.3
anu	2.0	bia	0.7	con	18.0
ans	17.7	bie	7.3	cor	6.0
ant	47.3	bil	1.0	cou	17.0
ape	2.3	bit	1.3	cqu	19.3
apl	2.0	bla	4.7	cre	2.7
app	7.3	ble	11.3	cri	3.7
apr	4.3	bon	1.7	cro	3.0
aqu	1.7	bor	1.0	cru	1.3
ara	5.3	bou	3.7	cta	1.0
arc	2.3	bra	2.3	cte	1.0
ard	6.3	bre	4.0	cti	1.3
are	1.3	bri	1.3	cul	1.3
arg	1.7	bro	3.0	cun	1.7
ari	1.3	bru	1.7		
arl	1.7	bse	1.3	dai	3.7

dan	17.7	eja	1.7	ett	14.3
deb	2.0	ela	2.3	et(-)	38.0
dec	5.0	ele	5.0	euf	1.3
dee	1.7	eli	1.3	eul	4.7
def	1.0	ell	8.3	eur	27.0
deg	1.0	elq	4.3	eus	4.3
dej	2.0	el(-)	6.0	eut	6.7
del	1.0	ema	3.3	eux	16.0
dem	5.3	emb	4.3	eu(-)	7.7
den	2.0	eme	28.3	eva	6.0
dep	1.7	emi	2.3	eve	7.7
der	4.0	emm	1.3	evi	2.7
des	16.3	emp	5.3	evo	1.3
det	1.7	emu	1.0	evr	3.0
deu	3.0	ena	6.3	exa	1.3
dev	5.0	enc	11.0	exe	1.0
de(-)	72.0	end	13.3	exp	2.7
die	2.3	ene	4.7	ez(-)	5.7
dig	2.0	enf	1.7		
dim	1.3	eni	4.0	fac	1.0
din	1.3	eno	1.3	fai	8.3
dir	3.3	ens	11.7	fal	1.7
dis	5.7	ent	62.6	fau	1.7
dit	10.7	enu	3.0	fec	1.0
doi	2.0	env	1.3	fen	2.3
don	4.7	en(-)	30.0	fer	3.3
dor	1.3	eoi	1.0	fes	1.7
dou	5.3	epa	4.0	feu	1.3
dra	2.3	epe	1.3	fe(-)	1.3
dre	7.0	epl	1.0	ffa	1.0
dui	1.0	epo	2.3	ffe	3.3
du(-)	12.7	epr	5.0	ffi	1.3
		epu	1.3	ffl	2.7
eai	2.8	era	8.7	ffr	4.0
eau	6.3	erc	3.0	fig	1.0
ea(-)	1.3	ere	11.3	fil	1.7
eba	1.0	erg	1.3	fin	3.0
ebo	1.7	eri	3.3	fit	7.3
ech	7.0	erm	2.0	fix	1.0
eci	3.0	ern	1.3	fla	2.0
ecl	1.3	err	5.0	fle	3.0
eco	8.0	ers	9.3	flo	1.7
ecr	4.7	ert	4.0	foi	5.3
ect	3.3	erv	2.0	fon	1.7
ec(-)	8.0	er(-)	36.3	for	3.0
ede	3.7	ese	1.3	fou	1.0
edr	1.3	esi	2.3	fra	4.7
eed	2.7	esp	1.7	fre	4.7
ees	4.7	esq	2.0	fri	1.0
ee(-)	8.7	ess	10.3	fro	1.7
eff	1.7	est	24.7	fut	2.3
efl	1.3	es(-)	121.0		
ega	4.7	eta	17.7	gag	1.0
ege	2.3	ete	14.3	gar	4.7
eh(-)	2.0	eti	6.3	gau	1.0
eil	7.3	eto	4.7	gea	5.0
ein	3.0	etr	11.7	gen	3.7

ger	3.3	iff	2.7	jeu	1.0
ges	5.7	if(-)	2.0	je(-)	23.3
ge(-)	11.7	ige	1.3	jou	5.7
gis	2.0	igl	1.0	jus	2.7
gle	1.7	ign	5.0		
gna	2.0	igr	1.7	la(-)	50.0
gne	2.7	igt	1.0	lac	2.7
goi	1.0	igu	2.0	lad	2.0
gou	1.3	ile	5.0	lai	11.0
gra	3.0	ili	1.3	lam	1.3
gre	2.3	ill	12.3	lan	5.7
gri	1.3	ils	7.7	lec	2.3
gto	1.0	il(-)	95.3	lee	2.3
gt(-)	1.7	ime	3.0	leg	2.0
gue	2.3	imi	1.7	lei	1.3
		imp	3.3	lem	5.0
hai	2.0	ina	3.7	len	3.0
ham	1.0	inc	2.7	ler	3.7
han	3.7	ind	4.3	les	27.7
hap	1.3	ine	22.7	let	4.7
har	2.6	ing	2.0	leu	13.0
hau	3.0	ini	1.7	lev	7.0
ha(-)	2.0	ino	1.3	le(-)	62.0
hee	5.0	inq	1.7	lia	1.0
hen	2.3	ins	5.7	lib	1.7
her	5.0	int	12.3	lic	5.0
hes	3.0	inu	1.7	lie	4.7
het	2.0	in(-)	10.0	lin	1.0
heu	7.0	ion	17.3	liq	2.7
hez	1.3	iqu	6.0	lir	1.3
he(-)	7.0	ira	2.7	lis	1.3
hir	1.3	ire	19.7	lit	1.3
hom	1.3	ir(-)	19.0	liv	1.0
hos	4.0	isa	7.3	lla	3.7
hui	1.3	isc	1.3	lle	21.0
		ise	7.0	lli	1.7
iab	2.3	isi	5.0	llu	2.0
iai	2.3	iso	3.7	loc	1.0
ian	1.7	iss	7.3	loi	2.3
ia(-)	1.7	ist	3.0	lon	4.3
ibl	4.3	is(-)	100.0	lor	5.0
ibr	1.7	ita	5.3	lot	1.7
ica	1.7	ite	11.0	lou	1.0
ich	2.0	ito	1.7	lqu	4.6
ici	2.7	itr	1.3	ls(-)	8.3
ico	3.3	its	3.0	lui	18.7
ide	7.0	itu	2.7	lum	1.7
idi	1.0	it(-)	123.0	lun	1.0
ied	1.3	ive	4.0	lus	7.0
iei	1.3	ivr	123.0	lut	1.3
iel	5.0	ix(-)	4.7	lu(-)	1.3
ien	25.0				
ier	7.3	jac	18.3	mai	27.3
ies	1.3	jal	3.7	mal	3.0
ieu	7.7	jam	3.0	man	5.7
ie(-)	11.3	ja(-)	1.7	mar	3.7
ife	1.0	jet	2.3	mat	2.0

mau	1.3	nem	1.0	off	2.3
ma(-)	3.0	ner	6.0	oh(-)	1.0
mbl	3.0	nes	2.0	oig	3.3
mbr	2.3	net	4.3	oin	2.0
mem	7.0	neu	1.3	oir	10.3
men	28.6	ne(-)	60.6	ois	13.3
mer	2.3	nfi	1.7	oit	1.7
mes	1.7	nge	7.7	oix	3.7
met	3.0	ngl	1.0	oi(-)	8.7
meu	1.3	ngo	1.0	ole	3.7
me(-)	22.3	ngt	1.7	oli	2.0
mid	1.7	nhe	2.7	oll	1.3
mie	2.3	nia	1.3	olo	1.0
mig	1.0	nib	1.0	oma	1.3
mil	1.0	nic	1.3	omb	1.3
min	3.0	nie	5.0	ome	3.3
mir	2.0	nif	1.3	omm	12.0
mis	4.0	nir	3.0	omp	7.3
mit	3.3	nis	1.0	one	3.7
mi(-)	3.0	nit	1.7	ond	4.3
mme	12.7	ni(-)	3.3	onf	2.0
moi	6.0	nna	3.3	ong	4.7
mom	1.3	nne	9.3	onn	10.0
mon	4.3	noc	1.3	ono	1.7
mor	1.7	nom	1.3	ons	13.0
mot	2.6	non	6.9	ont	12.7
mou	2.0	nor	1.0	on(-)	48.6
mpe	1.7	nou	8.3	opi	1.3
mpo	2.0	nqu	1.0	op(-)	3.0
mpr	6.7	nq(-)	1.3	orc	2.0
mps	3.0	nsa	1.7	ord	4.0
mpu	1.4	nse	4.6	ore	4.3
mun	1.7	nsi	4.6	orm	1.7
mur	2.0	nst	1.7	orr	2.7
		nsu	1.0	ors	3.7
nac	1.0	ns(-)	35.0	ort	10.0
nad	1.0	nta	4.7	osa	1.3
nai	7.0	nte	16.3	ose	6.3
nan	3.7	nti	7.0	os(-)	1.7
nap	1.7	nto	18.0	ote	2.3
nat	2.7	ntr	8.0	ots	1.7
nav	1.3	nts	4.7	ot(-)	4.7
na(-)	4.3	nt(-)	75.3	oub	2.3
nca	2.6	nue	1.3	ouc	4.7
nce	10.3	nui	2.0	ouf	3.0
nch	2.0	nu(-)	2.7	oui	4.3
nci	2.7	nvo	1.0	ouj	1.3
nco	6.0			oul	5.0
nc(-)	1.0	oba	1.0	oup	7.0
nda	4.7	obs	1.3	our	31.6
nde	3.0	oca	1.0	ous	18.7
ndi	4.3	oci	4.7	out	21.0
ndr	7.0	oce	1.7	ouv	14.0
nds	15.3	ode	1.0	ou(-)	4.7
ndu	3.0	oei	1.7	oya	1.0
nd(-)	4.7	oel	1.7	oye	1.3
nel	2.7	oeu	1.3		

Таблица 3 (*продолжение*)

pai	1.3	rap	3.7	rqu	2.7
pal	1.7	ras	4.7	rra	2.3
pan	1.3	rat	1.7	rre	6.7
par	21.0	rav	1.7	rri	2.0
pas	20.3	ra(-)	7.7	rro	3.3
pau	2.0	rbe	1.0	rse	1.7
pea	1.0	rce	3.7	rso	1.3
pec	3.0	rch	2.0	rs(-)	18.0
pei	1.3	rci	1.7	rta	4.3
pel	1.3	rda	1.3	rte	7.3
pen	10.0	rde	2.0	rti	3.3
per	6.0	rdi	3.0	rt(-)	7.3
pet	5.7	rd(-)	6.0	rue	1.7
peu	7.7	rea	1.7	rui	1.0
pe(-)	3.7	reb	1.0	rus	1.3
phe	1.0	rec	5.0	rut	1.7
pir	1.7	red	2.0	rve	1.0
pit	2.0	ree	1.3	rvi	1.0
pla	3.0	ref	2.3		
ple	4.0	reg	5.0	sag	3.0
pli	2.7	reh	1.3	sai	11.7
plu	9.3	rei	1.7	sal	1.0
poc	1.0	rel	1.0	sau	9.7
poe	2.7	rem	7.7	sau	1.0
poi	2.0	ren	16.0	sav	1.0
pol	1.0	rep	10.0	sa(-)	10.3
pon	2.0	rer	5.7	sce	1.7
por	7.0	res	32.3	sci	1.3
pos	2.7	ret	10.0	sec	2.3
pou	11.7	reu	2.3	sei	1.3
ppa	1.0	rev	4.3	sem	5.3
ppe	2.7	re(-)	67.3	sen	5.7
ppo	2.3	rge	2.3	sep	1.7
ppr	3.3	ria	2.3	ser	13.0
pre	22.3	ric	1.0	ses	12.0
pri	8.7	rie	7.0	seu	5.0
pro	9.3	rin	1.3	se(-)	35.6
ps(-)	3.7	rir	3.0	sib	2.0
pui	5.3	ris	8.3	sie	3.7
pul	1.0	rit	5.0	sif	1.3
pur	1.0	riv	1.7	sig	1.0
pu(-)	5.0	rle	1.0	sil	2.3
		rma	1.3	sio	6.0
qua	6.0	rme	1.0	sit	5.7
que	62.3	rmi	1.0	siv	1.0
qui	11.7	rna	2.3	si(-)	8.3
quo	2.3	rne	3.3	soi	4.7
qu(-)	18.0	roh	1.3	sol	1.7
		roc	3.0	som	1.0
rab	1.0	roi	4.7	son	24.3
rac	1.3	rol	1.7	sor	1.0
rad	1.0	rom	0.7	sou	14.0
rag	3.0	ron	4.0	spe	3.0
rai	19.3	rop	3.0	squ	3.3
ral	1.0	ros	1.3	ssa	7.7
ran	9.0	rou	8.3	sse	12.0
		rpr	1.0	ssi	9.7

Таблица 3 (*продолжение*)

sta	4.3	ubl	2.3	uvi	1.0
ste	7.3	uce	1.7	uvr	1.3
sti	3.7	uch	3.0	uv(-)	16.0
st(-)	15.7	ucu	1.3		
sui	5.7	uda	1.0	vai	20.0
sul	1.0	ude	2.7	van	10.3
sur	13.0	uel	8.3	va(-)	1.0
		uer	4.3	vec	6.7
tab	1.7	ues	27.0	vee	1.7
tac	2.3	ue(-)	30.3	vei	1.3
tai	25.0	uff	4.7	vel	1.3
tal	3.0	uf(-)	1.3	vem	1.7
tan	5.3	uie	3.0	ven	11.0
tat	1.3	uil	2.7	ver	11.0
ta(-)	8.3	uis	11.0	veu	3.3
tee	1.0	uit	4.3	vez	1.3
tei	1.3	uiv	1.0	vid	1.3
tem	5.7	ui(-)	27.3	vie	7.0
ten	11.3	ujo	2.3	vil	2.0
ter	12.3	ula	2.0	vin	3.3
tes	5.7	ule	9.0	vir	1.0
tet	4.7	ulo	1.3	vis	6.7
teu	2.7	uls	1.0	vit	1.3
te(-)	49.0	ult	1.3	viv	1.0
tho	1.0	uls	1.0	vla	1.3
tie	4.3	ul(-)	2.0	voc	1.0
tig	1.3	ume	1.3	voi	12.3
til	1.0	une	27.0	von	1.3
tim	2.7	uni	3.0	vou	10.3
tin	3.7	un(-)	26.6	voy	1.0
tio	10.3	uoi	1.7	vra	2.7
tiq	1.7	upe	2.7	vre	6.7
tir	2.3	upi	1.0	vue	1.3
tit	5.7	up(-)	3.7		
tiv	1.0	ura	6.0	xam	1.0
toi	19.3	nre	1.3	xe(-)	1.0
ton	3.3	urd	2.3	xpl	1.7
tot	3.0	ure	12.3		
tou	25.0	uri	5.0	yeu	3.0
tra	8.0	urn	4.3	you	1.0
tre	31.6	urp	1.0	yse	0.6
tri	1.0	urq	1.0		
tro	10.3	urs	8.0	zan	0.7
ts(-)	10.7	urt	5.7	ze(-)	0.7
tta	1.7	ur(-)	36.6		
tte	13.3	usa	1.0		
ttr	5.3	use	8.0		
tud	2.0	usi	1.0		
tue	2.3	usq	3.0		
tun	1.0	uss	5.3		
tur	2.7	us(-)	24.3		
tu(-)	12.7	uta	4.0		
		ute	9.3		
uai	3.0	utr	4.3		
uau	2.3	ut(-)	27.6		
un(-)	2.3	uva	3.3		
		uve	9.7		

Таблица 4

Частотность первых букв слова

Буква	$p \cdot 10^2$	Буква	$p \cdot 10^2$	Буква	$p \cdot 10^2$
a	10.00	*m*	4.45	*o*	1.44
d	9.83	*j*	3.67	*g*	1.05
l	9.17	*v*	3.41	*h*	0.88
s	8.49	*r*	3.22	*y*	0.35
c	8.33	*q*	3.02	*k*	0.04
p	7.36	*u*	2.74	*w*	0
c	6 03	*f*	2.69	*x*	0
i	4.96	*n*	2.67	*y*	0
t	4.66	*b*	1.49		

Относительная частота первых

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*l*
a	—	0 16	0.16	0.07	—	0.16	0.05	0.12	0.95	0.07	0.40
b	0.04	—	—	—	0.14	—	—	—	0.56	—	0.12
c	0.56	—	—	—	1.63	—	—	0.65	0.21	—	0.14
d	0.76	—	—	—	5.50	—	—	—	0.98	—	—
e	0.04	0.04	0.58	—	—	0.09	—	0.11	—	—	0.18
f	0.63	—	—	—	0.30	—	—	—	0.54	—	0.12
g	0.19	—	—	—	0.28	—	—	—	0.14	—	0.04
h	0.25	—	—	—	0.44	—	—	—	0.02	—	—
i	—	—	0.07	0.09	—	—	0.02	—	—	—	3.69
j	1.37	—	—	—	1.34	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	—	—	—
l	2.53	—	—	—	3.62	—	—	—	0.28	—	—
m	1.86	—	—	—	0.97	—	—	—	0.42	—	—
n	0.18	—	—	—	0.83	—	—	—	0.14	—	—
o	—	0.07	0.07	0.04	0.11	0.07	—	0.05	0.02	—	0.02
p	2.27	—	—	—	1.41	—	—	0.11	0.16	—	0.65
q	—	—	—	—	—	—	—	—	—	—	—
r	0.42	—	—	—	2.11	—	—	—	0.28	—	—
s	1.20	—	0.07	—	2.34	—	—	—	0.53	—	—
t	0.39	—	—	—	0.69	—	—	0.04	0.19	—	—
u	—	—	—	—	—	—	—	—	—	—	—
v	0.30	—	—	—	0.93	—	—	—	0.93	—	0.07
y	—	—	—	—	0.14	—	—	—	—	—	—

На основе данных этих таблиц были получены следующие значения: H_1=3.94 дв. ед., H_2=3.17 дв. ед., H_3=2.83 дв. ед.[16] Также было определено значение H_I по первым буквам, которое составляет 3.90 дв. ед. $H_0=\log_2 27=4.76$ дв. ед.

[16] Полученное нами значение H_3 может вызывать некоторые сомнения вследствие малого объема статистического материала. Однако, так как наши данные носят скорее качественно-оценочный, нежели количественный характер, значение это вполне может быть принято. Кроме того, хорошее совпадение данных, полученных по полной и сокращенной программам угадывания, позволяет считать это значение достаточно надежным.

§ 2. Экспериментальное определение границ энтропий высших порядков по основному и упрощенному вариантам метода Шеннона

Для оценки H_n при больших n были использованы основной и упрощенный варианты экспериментального метода Шеннона по угадыванию букв текста. Расскажем несколько подробнее о методике проведения эксперимента.

В ходе эксперимента испытуемый восстанавливал неизвестный ему текст, последовательно отгадывая все его буквы начиная с первой. Пробел между словами, апостроф и черточка (дефис) считались конечной буквой слова. Знаки препинания не учитывались. Угадывание было осуществлено как по полной, так и по сокращенной программам эксперимента: по полной программе угадывались

Таблица 5

двух букв слов ($p \cdot 10^2$)

m	*n*	*o*	*p*	*r*	*s*	*t*	*u*	*v*	*x*	*y*	Пробел
0.18	1.02	—	0.51	0.23	0.32	0.19	1.37	1.51	—	—	2.57
—	—	0.26	—	0.28	—	—	0.09	—	—	—	—
—	—	2.14	—	0.26	—	—	0.04	—	—	0.02	0.40
—	—	0.69	—	0.05	—	—	0.58	—	—	—	1.27
0.21	1.62	—	0.19	0.02	0.98	3.39	0.40	0.12	0.37	—	—
—	—	0 54	—	0.44	—	—	0.14	—	—	—	—
—	—	0.11	—	0.28	—	—	0.02	—	—	—	—
—	—	0.11	—	—	—	—	0.07	—	—	—	—
0.19	0.74	0.02	—	0.05	—	0.07	—	0.02	—	—	—
—	—	0.16	—	—	—	—	0.19	—	—	—	0.63
—	—	—	—	—	—	—	—	—	—	—	—
—	—	0.30	—	—	—	—	0.88	—	—	0.02	1.49
—	—	0.86	—	—	—		0.14		—	—	0.19
—	—	0.81	—	—	—	—	0.16	—	—	—	0.56
—	0.44	—	0.02	0.05	0.05	—	0.44	—	—	—	—
—	—	1.23	—	1.04	—	—	0.49	—	—	—	—
—	—	—	—	—	—	—	3.13	—	—	—	—
—	—	0.26	—	—	—	—	0.14	—	—	—	—
—	—	2.13	0 05	—	—	0.05	1.09	—	—	0.05	0.98
—	—	1.51	—	0.81	—	—	0.81	—	—	0.02	0.21
—	2.74	—	—	—	—	—	—	—	—	—	—
—	—	1.07	—	0.07	—	—	0.04	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	0.21

5-я, 10-я, 15-я, 20-я, 30-я, 40-я, 50-я, 75-я, 100-я буквы текста,[17] остальные буквы угадывались по сокращенной программе.

Такое комбинированное применение полной и сокращенной программ не только сокращает объем работы,[18] но и позволяет получить два взаимоконтролируемых ряда числовых данных.

При осуществлении описываемого эксперимента достоверные продолжения определялись самим испытуемым в ходе угадывания. Кроме того, все случаи «нулевой информации» проверялись и дополнительно выявлялись в ходе корректировочной обработки результатов эксперимента (см. стр. 32).

В качестве экспериментального материала было использовано 116 текстов длиной в 100—115 букв каждый (обычно это простые распространенные или сложные предложения средней длины), в которых угадывалось 100 первых букв.[19] Экспериментальные тексты принадлежат к четырем жанрово-стилистическим разновидностям современного французского языка.

1. Разговорная речь. Было отгадано 33 текста, из которых 28 заимствовано из разговорных текстов, приводимых в работе: G. Gougenheim, R. Michéa, P. Rivenc, A. Sauvageot. L'élaboration du français élémentaire, Paris, 1956; 5 текстов взято из книги Р. Кено «Зази в метро» (R. Queneau. Zazie dans le métro. Paris, 1958). По своей структуре эти тексты близки к нормам современной речи.

2. Беллетристический стиль, охватывающий 33 текста из произведений Р. Роллана, Р. М. дю Гара, А. Стиля, М. Понса (R. Rolland. Pierre et Luce. M., 1953; R. M. du Gard. Les Thibault. M., 1960; A. Stil. Nous nous aimerons demain. Paris, 1960; M. Pons. Le passager de la nuit. Paris, 1960).

3. Деловой стиль, включающий 33 общественно-политических и научно-специальных текста из статей и сообщений, опубликованных в газетах «Юманите» и «Авангард» за 1961—1962 гг., и из лингвистических работ (M. Cohen. Grammaire et style. Paris, 1954; A. Dauzat. Historie de la langue française. Paris, 1959).

---

[17] Целесообразность подобной выборки обусловлена тем, что функция H_n лишь для начальных n претерпевает резкие изменения, оставаясь примерно постоянной для $n \geqslant 30$ (см.: N. G. Burton, J. C. R. Licklider, ук. соч.; А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин, ук. соч.).

[18] Необходимость такого сокращения вызвана трудностями проведения эксперимента с испытуемыми французами.

[19] Поскольку значения H_1, H_2 и H_3 были нами определены непосредственными расчетами, угадывание начиналось с 4-й буквы. При этом мы исходили из следующих соображений: так как каждый текст начинается каким-либо целым словом, то значения H_1, H_2 и H_3 необходимо рассчитывать по первым буквам слов. Однако было показано, что значения H_I и H_{II} по первым буквам близки к значениям частных энтропий H_1 и H_2 (H_I=3.90 дв. ед.; H_1=3.94 дв. ед.; H_{II}=3.07 дв. ед.; H_2=3.17 дв. ед.). По аналогии мы приняли значение $H_3 = H_{III}$=2.83 дв. ед.

4. Поэзия. Сюда входят 17 текстов, взятых из произведений Л. Арагона, В. Гюго, А. де Мюссе, Ж. Превера, Ф. Тамма и др.

Ясно, что результаты угадывания в значительной мере зависят от лингвистической культуры и языкового чутья испытуемого. Сказываются здесь и такие факторы, как усталость, внимательность и т. д. Для того чтобы в возможной степени ослабить действие этих субъективных факторов и получить результаты, наилучшим образом отражающие распределение энтропии, были предприняты следующие меры.

I. Для проведения эксперимента среди нескольких информантов был выбран испытуемый, обнаруживший наилучшее «чутье» языка.

II. В ходе эксперимента был использован справочно-вспомогательный материал:

1) специально составленные таблицы частотности начальных букв и двухбуквенных сочетаний французского слова, а также таблицы частотности отдельных букв, двухбуквенных и трехбуквенных сочетаний независимо от их позиции в слове; таблица частотности начальных букв использовалась при угадывании первой буквы слова, остальные таблицы применялись при угадывании второй и третьей букв;

2) энциклопедические и двуязычные словари: а) Dictionnaire de la langue française. E. Littre, Paris, 1959; б) Dictionnaire de la langue française, Azed, Paris, 1959; в) Larousse élémentaire illustré, Paris, 1959; г) Французско-русский политехнический словарь. Ред. Л. Д. Белкинд. М., 1948.

Словари использовались при угадывании четвертой и последующих букв, а также при определении «нулей информации». Применение статистических таблиц и словарей не только облегчает и ускоряет процесс угадывания, но в значительной степени нивелирует отклонения в результатах эксперимента у разных испытуемых. Об этом можно судить, в частности, по небольшим отклонениям контрольного эксперимента.

§ 3. Вероятностно-лингвистическая корректировка

Результаты основного опыта по угадыванию букв после его завершения были подвергнуты вероятностно-лингвистической корректировке. Эта корректировка, проводившаяся экспериментатором совместно с испытуемыми при помощи указанных выше таблиц и словарей, преследовала следующие цели:

1) проверить, насколько объективно испытуемый выделил достоверные продолжения;

2) выявить те «нули информации», которые не были учтены в процессе эксперимента;

3) устранить «лишние попытки» при угадывании отдельных букв. О том, как проводилась вероятностно-лингвистическая корректировка, можно судить по данным табл. 6.

Таблица 6

Номер букв	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Текст	*c*	*'*	*é*	*t*	*a*	*i*	*t*		*u*	*n*		*f*	*a*	*n*	*a*	*t*	*i*	*q*	*u*	*e*		*b*	*o*	*r*	*n*	*é*		*a*	*v*	*e*	*c*		*q*	*u*	*i*
Результат угадывания в количестве попыток	2	1	1	1	0	0	0	0	2	1	2	2	2	2	1	1	0	2	0	0	0	2	2	0	0	0	0	2	2	1	1	0	2	0	1
То же после корректировки	2	1	1	1	0	0	0	0	2	1	2	2	2	2	1	1	0	1	0	0	0	2	2	1	1	1	1	2	2	2	1	0	2	0	1

Испытуемый определил буквы 25 и 27 (ср. графу 3) как достоверные продолжения (в рабочем протоколе отмечены «нули информации»). Однако в ходе корректировки выяснилось, что в обоих случаях есть и другие возможные продолжения. Так, например, если предположить, что после слова fanatique стоит точка, то после 24-й буквы наряду с *n* возможно продолжение *g* (ср. borgne), а после 26-й буквы возможен не только пробел, но и буква *r* (ср. borner ses désirs). Поэтому в протоколе в обоих случаях вместо нуля отмечена одна попытка.

Проверка по словарю возможных вариантов буквы 18, угаданной со второй попытки, показала, что в данном случае имеется только два возможных варианта: fanati(s)me — fanati(q)ue. Таким образом, уже после первой попытки угадать данную букву испытуемый получил полную информацию о следующей за *i* букве (если не *s*, то *q*). Вторая попытка, отмеченная в рабочем протоколе, была, следовательно, избыточной, и в окончательном протоколе вместо первоначальных двух попыток отмечена одна попытка.

§ 4. Статистическая обработка экспериментальных результатов

После завершения эксперимента и корректировочной работы результаты опыта были подвергнуты статистической обработке.

Следует отметить, что в итоге этой обработки были получены данные, характеризующие энтропию не конкретной буквы в конкретном тексте (уже угаданная буква не содержит энтропии), но энтропию 1-й, 2-й, 3-й. . . 100-й букв в различных выборках, относящихся к четырем жанрово-стилистическим разновидностям современного французского языка (см. стр. 30).

Поэтические тексты в особую выборку не выделялись, поскольку их угадывание требует особой организации эксперимента (специальный выбор испытуемого, использование словарей рифм, таблиц частот вариантов ритма для данного метра и т. п.). В данном же опыте использовалась обычная методика угадывания, что

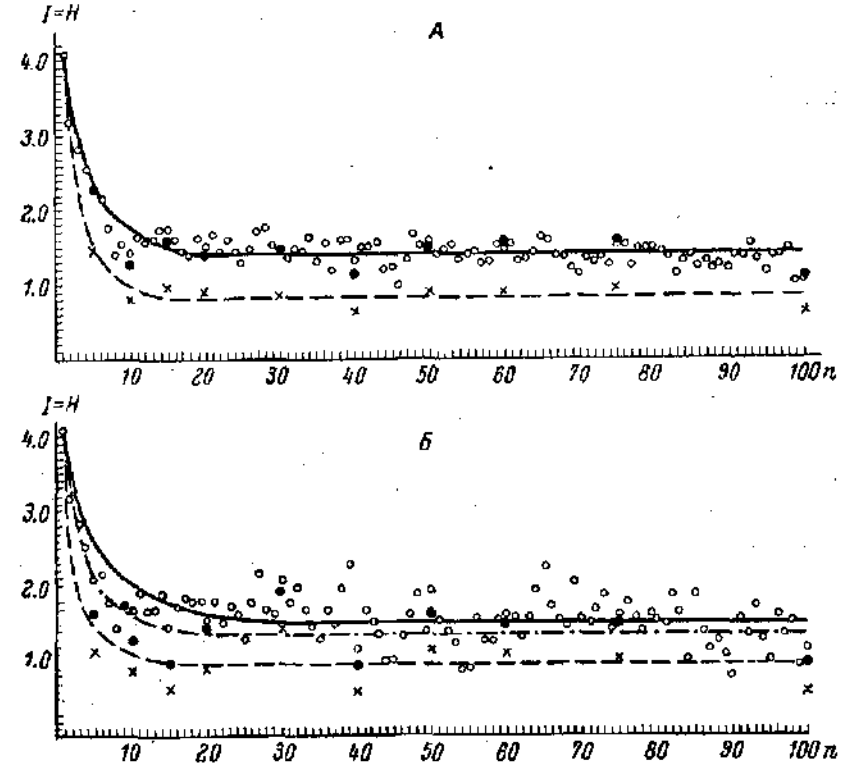


Рис. 3. График распределения энтропии.

По оси абсцисс — число букв, *по оси ординат* — энтропия (количество информации), в дв. ед. *А* — язык в целом; *Б* — разговорная речь.

несколько нарушает строгость эксперимента. Однако отсутствие специальной методики при угадывании поэтических текстов не может существенно сказаться на конечных результатах.

Как было показано выше, существующие методы расчета не дают возможности точно определить то количество информации (*энтропии*), *которое несет n-я буква* в исследуемой выборке текстов, они позволяют определить границы (точнее — приближения к ним) интервала, в котором заключено истинное значение энтропии данной буквы.

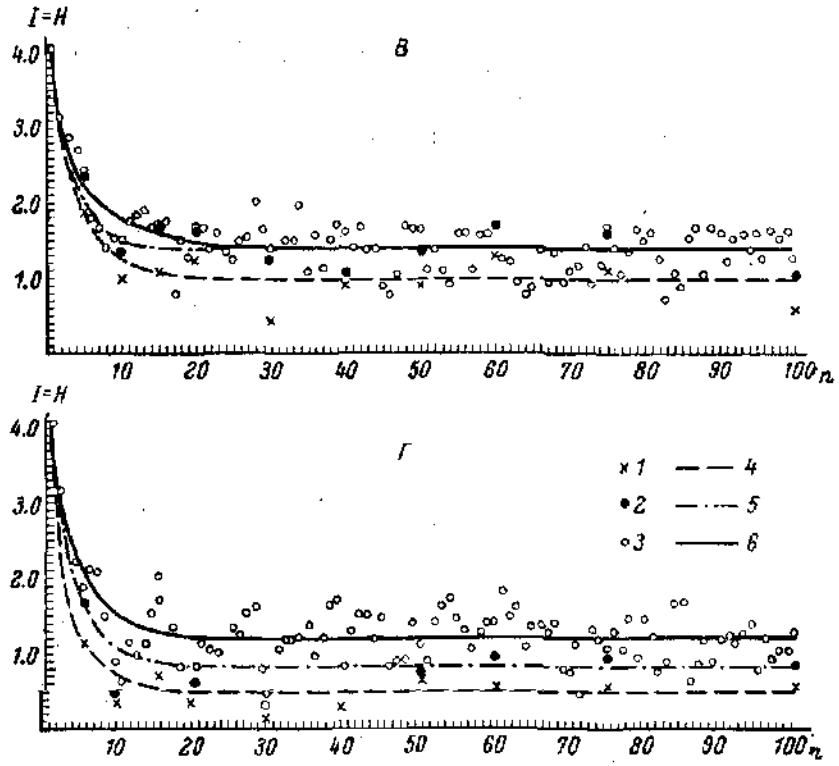


Рис. 3 *(продолжение)*.

В — беллетристический стиль; *Г* — деловой стиль.
1 — позиция H_n; *2* — позиция H_n; *3* — позиция H_n'; *4* — нижняя граница энтропии; *5* — верхняя граница энтропии (полная программа угадывания); *6* — *то же* (сокращенная программа угадывания).

Нижняя граница этого интервала определяется путем обработки данных, полученных в результате угадывания по полной программе. Эти данные рассчитывались по данным неравенства (3).

Формула (8) использовалась для расчета данных, полученных путем проведения полной программы угадывания, формула (9) — для обработки данных, полученных по сокращенной программе эксперимента.

Данные об энтропии французского языка и его стилей приведены в табл. 7, а зависимости, существующие между величинами энтропии и номерами букв в исследованных выборках текстов, графически представлены на рис. 3.

Сравнение данных, полученных при обработке материала с помощью формул (8) и (9), говорит о том, что в результате применения сокращенной программы угадывания мы получаем достаточно надежные данные в верхней границе энтропии. Поэтому в дальнейшем, говоря о верхней границе энтропии, мы будем пользоваться величинами H' и их производными.

§ 5. Контрольные эксперименты

Для того чтобы считать полученные данные репрезентативными информационными характеристиками французской письменной речи, необходимо провести контрольную проверку этих данных.

Эта проверка была осуществлена двумя путями.

Во-первых, объективность результатов угадывания была проверена с помощью выборочного угадывания того же материала на другом информанте. Проверялись результаты угадывания 20-й, 40-й, 80-й и 100-й букв по всем использованным текстам. Отклонения (в основном в худшую сторону) результатов второго информанта по сравнению с рабочим экспериментом не превышают $\pm 10\%$.

Во-вторых, французские тексты были исследованы с помощью тестового эксперимента Колмогорова (ср. стр. 13—17). Результаты этого исследования были сопоставлены с общими оценками верхней и нижней границ энтропии, полученными в итоге проведения основного эксперимента.

Расскажем несколько подробнее об организации этого эксперимента. Было взято 12 отрывков по 50 знаков каждый. Из них 6 отрывков — из газеты «Юманите», 6 — из романа Р. М. дю Гара «Семья Тибо». Таким образом, общее количество угадывавшихся знаков составило 600 (включая пробелы), т. е. — по 300 знаков на каждый из стилей (деловой и беллетристический).

Испытуемый читал все начало текста до первого отрывка, угадывал первый отрывок с первой буквы до пятидесятой, затем читал текст, отделяющий первый отрывок от второго, угадывал второй отрывок и т. д.

Угадывая очередной отрывок, испытуемый заполняет специально составленный бланк протокола (табл. 8). В бланк внесены последние строки текста. Под ними размещаются 50 столбцов (по 5 клеток в каждом столбце). В столбцах и фиксируются результаты угадывания.

При заполнении каждого столбца испытуемый, основываясь на известном ему предыдущем контексте, должен поступить следующим образом.

1. В случае большой уверенности в правильности предполагаемой следующей буквы записать эту букву в верхней (первой) клетке.

2. В случае меньшей уверенности сделать запись буквы во второй клетке.

3. Если испытуемый предлагает две буквы на выбор с равной вероятностью, то он записывает их в третьей клетке.

4. Если предлагаются две буквы, из которых одна более вероятна, чем другая, запись делается в четвертой клетке.

5. В случае отказа от угадывания испытуемый ставит вопросительный знак в первой (верхней) клетке.

После того как соответствующий возможный вариант угадывания выбран испытуемым и зафиксирован в одной из граф протокола, экспериментатор сообщает истинный знак, который записывается в пятой (нижней) клетке столбика. Пробелы в нижней строке (в истинном тексте) остаются пустыми, а при угадывании обозначаются двоеточием.

Все сказанное выше поясним примером (табл. 8). Здесь первые 12 букв угадываются довольно свободно (хотя 1-я и 5-я с малой уверенностью), так как на этом участке текста испытуемому, очевидно, ясны смысловые связи с предыдущим контекстом.

При угадывании 13-й буквы испытуемый предложил 2 знака, из которых «пробел» — с большей уверенностью, чем *е*.

Следующая (14-я) буква была угадана правильно опять-таки из-за очевидной связи с предыдущим контекстом, но испытуемый счел необходимым поставить этот знак в класс меньшей уверенности, так как начальная буква слова может вызывать некоторые колебания.

Дальнейший текст до 28-й буквы был угадан с большой уверенностью, однако с двумя отказами на двух начальных буквах

Таблица 9

Стиль	№ протокола	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	$\bar{H}$	H' (по основному эксперименту)	R (по основному эксперименту)
Деловой стиль	1	36	0	3	2	2	0	2	0	1	4	0.79		
	2	41	1	1	2	3	0	0	0	0	2	0.67		
	3	38	0	6	1	1	0	1	0	0	3	0.41		
	4	34	0	2	4	2	0	4	1	0	3	0.82		
	5	41	0	4	1	0	0	2	0	0	2	0.34		
	6	43	0	3	0	0	0	0	0	0	4	0.40		
	Сводный	233	1	19	10	8	0	9	1	1	18	0.61	1.20	0.48
Беллетристический стиль	1	34	0	2	1	3	0	3	1	0	6	0.85		
	2	28	1	6	1	1	1	2	1	0	9	1.39		
	3	33	2	7	1	1	0	2	0	1	3	1.00		
	4	27	1	8	1	1	0	4	2	0	6	1.06		
	5	30	0	7	4	3	0	0	0	0	6	1.13		
	6	30	1	11	2	0	0	1	1	0	4	1.06		
	Сводный	182	5	41	10	9	1	12	5	1	34	1.11	1.38	0.94

Таблица 7

Вид выборки	Французский язык: суммарные вероятностные зависимости в тексте: предельная энтропия, в дв. ед., для границ: нижней $\underline{H}_\infty$	Французский язык: суммарные вероятностные зависимости в тексте: предельная энтропия, в дв. ед., для границ: верхней (по полн. прогр.) $\bar{H}_\infty$	Французский язык: суммарные вероятностные зависимости в тексте: предельная энтропия, в дв. ед., для границ: верхней (по сокращ. прогр.) $\bar{H}'_\infty$	Французский язык: суммарные вероятностные зависимости в тексте: контекстные коэффициенты для границ: нижней $\underline{s}$	Французский язык: суммарные вероятностные зависимости в тексте: контекстные коэффициенты для границ: верхней (по сокращ. прогр.) $\bar{s}$	Французский язык: суммарные вероятностные зависимости в тексте: среднее квадратическое отклонение для границ: нижней $\underline{\sigma}$	Французский язык: суммарные вероятностные зависимости в тексте: среднее квадратическое отклонение для границ: верхней (по сокращ. прогр.) $\bar{\sigma}'$	Французский язык: суммарные вероятностные зависимости в тексте: избыточность, в %, для границ: нижней (по сокращ. прогр.) $\underline{R}$	Французский язык: суммарные вероятностные зависимости в тексте: избыточность, в %, для границ: верхней $\bar{R}$	Французский язык: лексические вероятностные зависимости по сокращ. пр: предельная обусловленность. Верхняя граница по сокращ. прогр. $\bar{K}'_\infty$	Французский язык: лексические вероятностные зависимости по сокращ. пр: предельная энтропия, в дв. ед. H^L_∞	Французский язык: лексические вероятностные зависимости по сокращ. пр: лексические коэффициенты, l	Французский язык: лексические вероятностные зависимости по сокращ. пр: среднее квадратическое о клонени σ	я граница: $L'_\infty - L_\infty$, дв. ед.	я граница: $\frac{L_\infty}{\bar{K}'_\infty}$, в %	Русский язык (суммарные вероятностные связи): предельная энтропия, в дв. ед., для границ: нижней (по полн. прогр.) $\underline{H}_\infty$	Русский язык (суммарные вероятностные связи): предельная энтропия, в дв. ед., для границ: верхней (по полной прогр.) $\bar{H}'_\infty$	Русский язык (суммарные вероятностные связи): контекстные коэффициенты для границ: нижней, $\underline{s}$	Русский язык (суммарные вероятностные связи): контекстные коэффициенты для границ: верхней (по полной прогр.) $\bar{s}$	Русский язык (суммарные вероятностные связи): среднее квадратическое отклонение для границ: нижней, $\underline{\sigma}$	Русский язык (суммарные вероятностные связи): среднее квадратическое отклонение для границ: верхней (по полн. прогр.) $\bar{\sigma}$	Русский язык (суммарные вероятностные связи): избыточность, в %, для границ: нижней (по полн. прогр.) $\underline{R}$	Русский язык (суммарные вероятностные связи): избыточность, в %, для границ: верхней, $\bar{R}$	Русский язык (суммарные вероятностные связи): предельная контекстная обусловленность (верхняя граница по полн. прогр.) $\bar{K}'_\infty$
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Язык в целом	0.82	1.39	1.40	0.30	0.28	0.17	0.15	71	83	3.36	2.45	0.43	0.13	2.19	35	0.87	1.37	0.31	0.19	0.198	0.235	72.1	82.6	3.63
Разговорная речь	0.95	1.36	1.47	0.28	0.24	0.30	0.28	70	80	3.29	2.55	0.57	0.18	2.22	32	1.00	1.40	0.31	0.22	0.401	0.450	72.0	79.6	3.60
Беллетристический стиль	0.96	1.37	1.38	0.27	0.26	0.22	0.27	70	80	3.38	2.40	0.34	0.15	2.16	36	0.87	1.19	0.29	0.21	0.237	0.282	76.3	82.6	3.81
Деловой стиль	0.50	0.83	1.20	0.33	0.31	0.24	0.26	75	89	3.56	2.13	0.48	0.20	2.07	42	0.59	0.83	0.32	0.24	0.212	0.288	83.4	88.1	4.17

Примечание. Данные по русскому языку заимствованы из работы: А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разжи стр. 128. Данные по английскому языку приводятся в статье Г. П. Богуславской (стр. 60 в настоящем сборнике).

Таблица 8

Образец протокола проверочного эксперимента по Колмогорова

Контрольному тексту в таблице предшествует отрывок *Quand nous aurons m etes, ce que nous sommes, et qu' on n'a . . .*

Возможные исходы опыта	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Угадывание с большой уверенностью		*a*	*s*	:		*e*	*s*	*o*	*i*	*n*	:	*d*			*u*	*x*	:	?		:	?		
Угадывание с малой уверенностью	*p*				*b*									*e*					*u*			*e*	*r*
2 знака, оба с равной вероятностью																							
2 знака, один из которых — с большей вероятностью													: > *e*										
Текст	*p*	*a*	*s*		*b*	*e*	*s*	*o*	*i*	*n*		*d*	,	*e*	*u*	*x*		*t*	*u*		*v*	*e*	*r*

Возможные исходы опыта	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
Угадывание с большой уверенностью	*m*		*e*	*n*	?	*l*	*s*	:			*u*	*s*	:	?				*e*	*c*	*t*	*e*	
Угадывание с малой уверенностью									*n*	*o*					*e*	*s*	*p*					
2 знака, оба с равной вероятностью																						
2 знака, один из которых — с большей вероятностью		*m* > *b*																				
Текст	*m*	*m*	*e*		*i*	*l*	*s*		*n*	*o*	*u*	*s*		*r*	*e*	*s*	*p*	*e*	*c*	*t*	*e*	*ront*

слов tu и verras. Эти буквы действительно трудно предугадать исходя из контекста.

28-я буква (тоже начальная буква слова) угадана ошибочно (предполагалось que), что повысило оценку энтропии на 0.11 дв. ед., а следующая, 29-я буква, также угаданная ошибочно, составляла 0.09 дв. ед.

Большой рост оценки энтропии данного отрывка вызвало угадывание 33-й буквы (предполагалось comment). Такое ошибочное угадывание вполне допустимо, однако нельзя было относить его в класс большой уверенности. Такая ошибка увеличила оценку энтропии на 0.18 дв. ед.

Угадывание остальных 17 букв не вызывает необходимости в специальных пояснениях.

Все 12 отрывков угадывались по тому же принципу. Результаты угадывания по всем отрывкам даны в табл. 9.

Произведя расчеты экспериментальных данных с помощью формулы (12) и используя данные табл. 1, мы получили значение оценки энтропии французской письменной речи по методу А. Н. Колмогорова:

$H''=0.61$ дв. ед. для делового стиля,
$H''=1.11$ дв. ед. для беллетристического стиля.

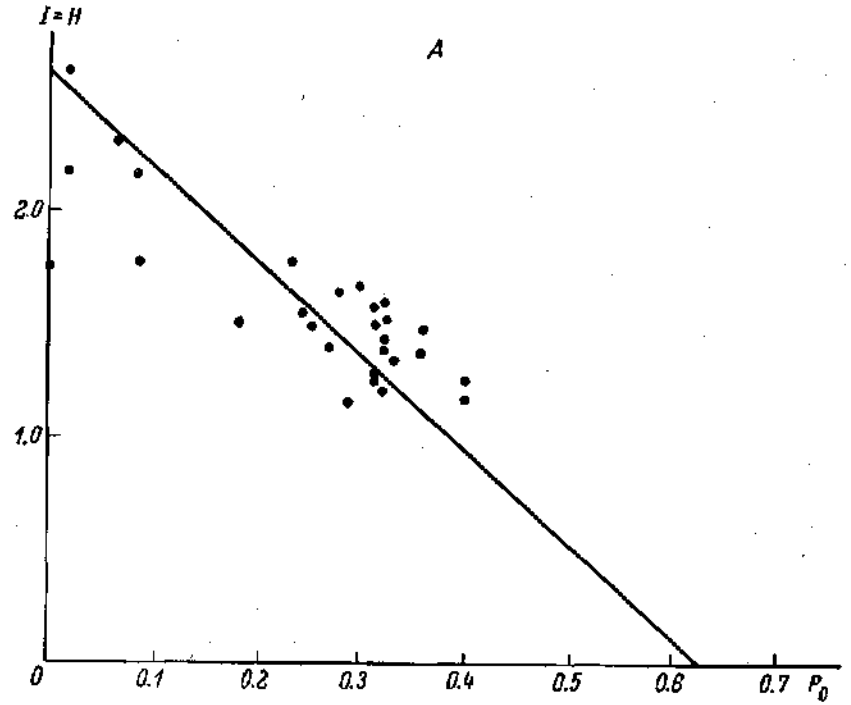
Значения этих оценок находятся в хорошем согласии с данными, полученными по основному и упрощенному вариантам метода Шеннона (табл. 7, графы 1—3). Это дает нам право считать результаты основного эксперимента и их производные в качестве репрезентативных информационных характеристик французской письменной речи.

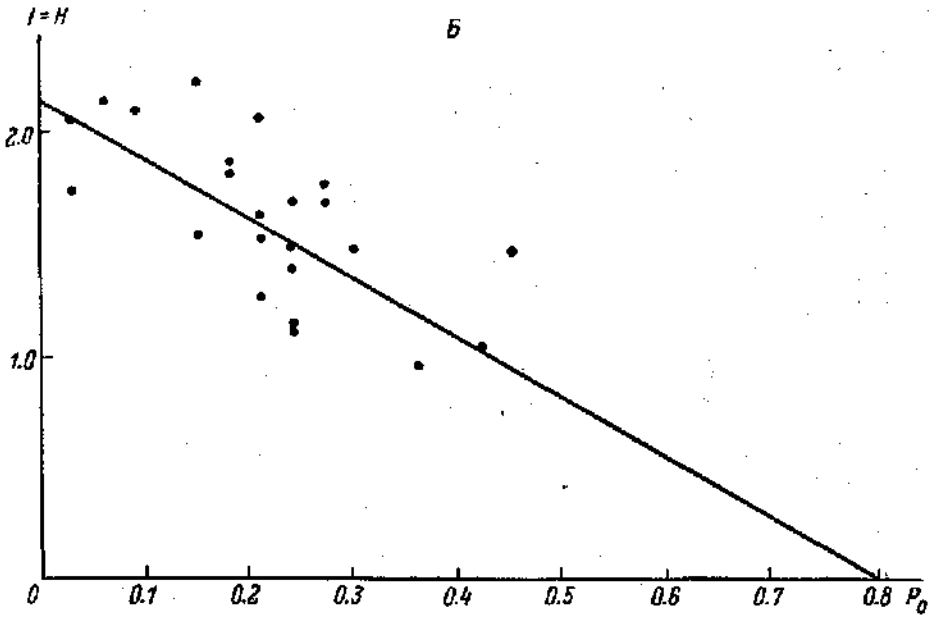
§ 6. Лингвистическое обсуждение полученных результатов

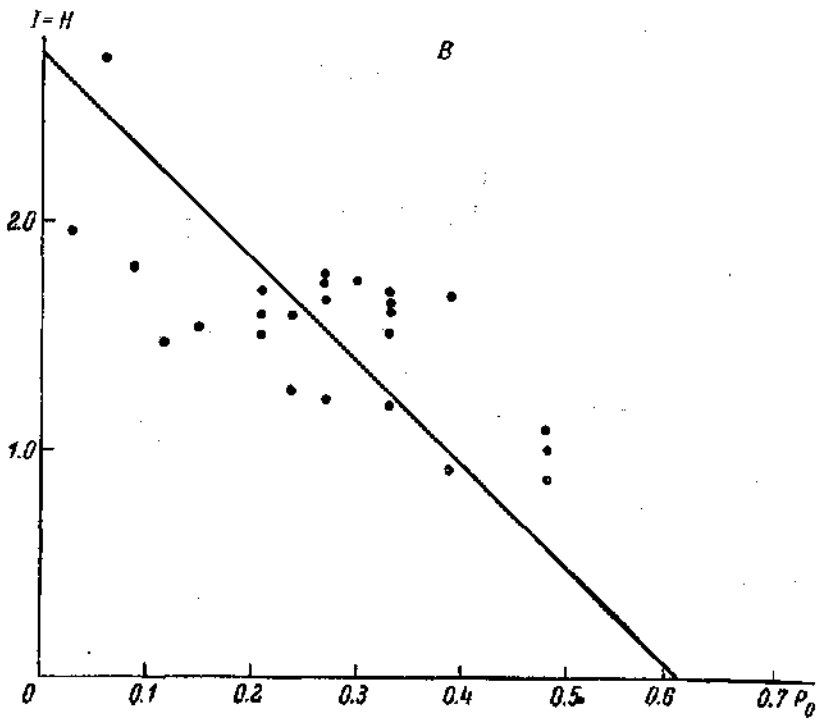
Наша задача заключается в том, чтобы, с одной стороны, дать лингвистическую интерпретацию математических величин, получаемых в итоге расчета экспериментальных данных, а с другой — количественно оценить через эти величины некоторые лингвистические категории.

Для этого проанализируем распределение верхних и нижних оценок энтропии в различных участках модели стобуквенного текста.

С этой целью будем рассматривать верхнюю и нижнюю границы энтропии не как последовательности дискретных значений, но как непрерывные кривые (точнее — как непрерывные функции непрерывного аргумента). Если отвлечься от непериодических колебаний, относя их за счет разброса, то верхние и нижние границы энтропии в каждой из наших моделей могут быть в первом







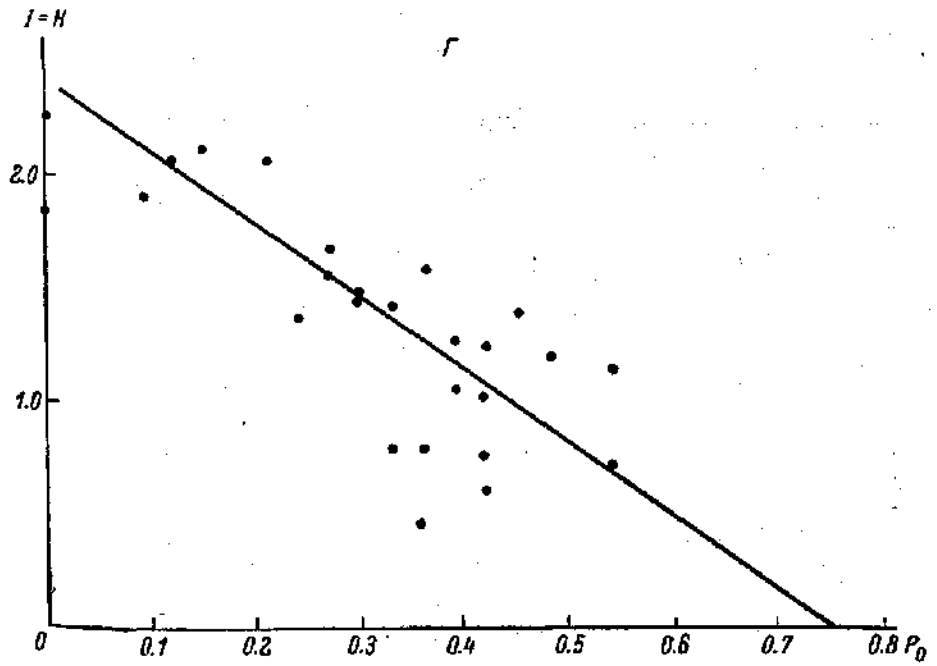


Рис. 4. График зависимости $H' = H_a - gq_0^n$.

По оси абсцисс — вероятность достоверных продолжений, *по оси ординат* — энтропия (количество информации), в дв. ед. *А* — язык в целом; *Б* — разговорная речь; *В* — беллетристический стиль; *Г* — деловой стиль.

приближении аппроксимированы теоретической кривой экспоненциального типа, которая имеет вид

$$H(t) = (H_0 - H_\infty)\, e^{-st} + H_\infty, \qquad (13)$$

где $H(t)$ — верхний или нижний теоретический предел энтропии; t — непрерывный аргумент функции, заменяющий дискретные величины; H_0 — энтропия алфавита французского языка; e — основание натуральных логарифмов; s — специальный рассчитанный для каждой кривой контекстный коэффициент; H_∞ — предельная энтропия, или иначе — энтропия языка. Значения s, H_∞ и среднее квадратическое отклонение в распределении оценок энтропии (σ) даны в табл. 7.

Для того чтобы выявить лингвистический смысл этой функции, попытаемся найти зависимость между величинами H' и q_0^n. Построенные на основании данных о вероятностях «нулей информации» графики (рис. 4) показывают, что во всех выборках текстов между величинами H' и q_0^n существует зависимость, выражаемая двучленом первой степени вида

$$H'_n = H_a - g q_0^n, \qquad (14)$$

где q_0^n — вероятность достоверных продолжений для данного номера буквы, а H_a и g — константы формулы (14). Их значения, а также среднее квадратическое отклонение даны в табл. 10.

Таблица 10

Вид выборки	H_a	g	s
Язык в целом	2.60	4.19	0.272
Разговорная речь	2.15	2.65	0.352
Беллетристический стиль	2.73	4.46	0.341
Деловой стиль	2.48	3.24	0.387

Выражение (14) нужно рассматривать, разумеется, как интерполяционную эмпирическую формулу. Однако эта формула с достаточной ясностью говорит о том, что величина энтропии буквы убывает в зависимости от роста числа случаев, когда появление n-й буквы полностью определено предшествующей последовательностью из $n-1$ букв.

Отсюда следует, что зависимость (13), графическим выражением которой является показательная кривая, имеет следующий лингвистический смысл: постепенное опускание кривой по направлению к оси абсцисс (точнее, к некоторой параллельной ей прямой H'_∞) отражает нарастание контекстных связей. Таким образом, появляется возможность дать количественную оценку лингвистическому понятию «контекстные связи», используя для этого понятие контекстной обусловленности каждой буквенной

позиции (номера буквы), которое легко может быть оценено в величинах энтропии и информации.

Действительно, если бы в языке не существовало парадигматических и синтагматических (вероятностных) связей, энтропия каждой буквенной позиции была бы равна энтропии алфавита H_0. Однако в действительности энтропия каждой буквенной позиции меньше величины H_0. Причины этого — разного вида статистико-лингвистические ограничения, которые накладываются на синтагматику языка (в том числе и на его буквенный код). Вся сумма этих ограничений, которую мы будем обозначать термином «контекстная обусловленность буквы» (K_n), может быть представлена как

$$K'_n = H_0 - H'_n. \qquad (15)$$

Заменяя величину H'_n выражением (13), а дискретные величины n непрерывным аргументом t, получаем зависимость

$$K'(t) = H_0 - (H_0 - H'_\infty)\, e^{-st} - H'_\infty,$$

которая после преобразования принимает вид

$$K'(t) = (H_0 - H'_\infty)(1 - e^{-st}). \qquad (16)$$

Эта формула, показывающая контекстную обусловленность n-й буквы в зависимости от n, служит характеристикой нарастания информационных связей между единицами текста по мере удаления от его начала; график этой зависимости относительно величин H'_n для французского языка и его трех стилей представлен на рис. 5. Ход зависимости K'_n (а также H'_n) характеризуется коэффициентом s. Чем больше абсолютная величина s, тем выше темп нарастания контекстной обусловленности букв (а также и других единиц текста).

Использование понятия «предельная энтропия» (H'_∞) предусматривает существование понятия «предельная контекстная обусловленность» (K'_∞),

$$K'_\infty = H_0 - H'_\infty. \qquad (17)$$

Значения K'_∞, характеризующие французский язык в целом и его стили, даны в табл. 7 (графа 10).

Для языкознания интерес представляют не только суммарные оценки вероятностных связей текста (величины H_∞, s и K_∞), но и числовые характеристики каждой из указанных выше лингвостатистических связей или по крайней мере комбинаций из двух-трех таких зависимостей. Но чтобы получить эти характеристики, необходимо найти приемы обработки экспериментального материала, которые позволили бы выделять интересующие нас зависимости из общей суммы вероятностно-лингвистических связей в тексте.

Попробуем получить количественную оценку связей, существующих между отдельными словами текста. Сумму этих связей, охватывающую в основном лексические и лишь частично грамматические зависимости, мы будем называть лексической обусловленностью данного участка текста. С этой целью, последовательно

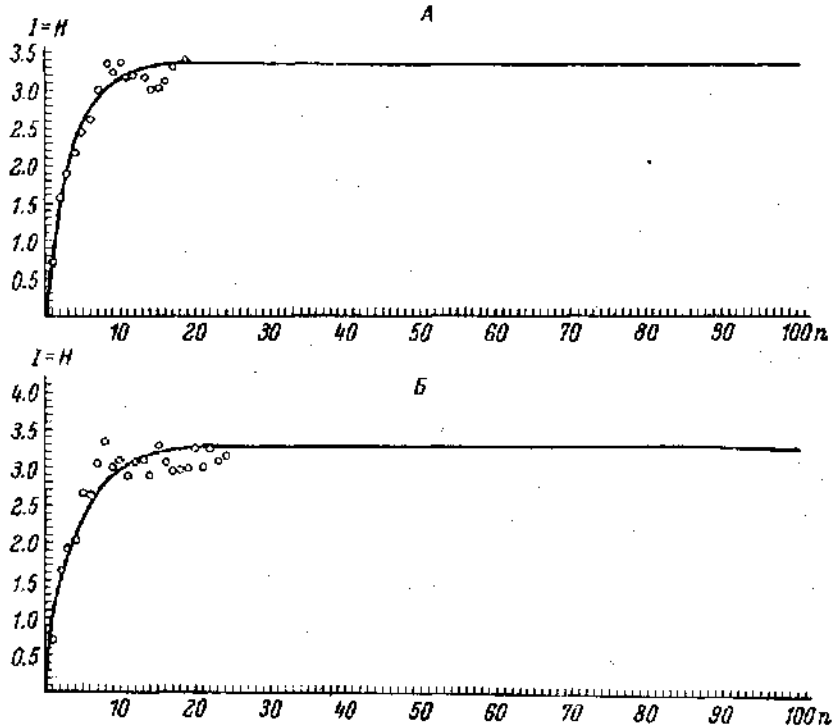


Рис. 5. График роста контекстной обусловленности в зависимости от номера буквы в выборке.

По оси абсцисс — количество букв, *по оси ординат* — энтропия (количество информации), в дв. ед.
А — язык в целом; *Б* — разговорная речь.

отсекая первое, второе, третье и т. д. слова во всех обследованных текстах, рассчитаем H' двух начальных букв второго, третьего, четвертого и т. д. слов для каждой выборки. Значения энтропии начальных букв первого слова (H_{I} и H_{II}) были получены при исходном расчете; они составляют 3.9 и 3.07 дв. ед. соответственно. Располагая средние арифметические значения сумм энтропии двух начальных букв второго, третьего и т. д. слов на расстоянии средней длины слова, получаем график убывания энтропии под влиянием лексической обусловленности слова (рис. 6). В первом

приближении этот график можно снова рассматривать как показательную кривую вида

$$H^{(л)}(t) = \left(\frac{\bar{H}'_1 + \bar{H}'_2}{2} - H_\infty^{(л)}\right) e^{-lt} + H_\infty^{(л)}, \qquad (18)$$

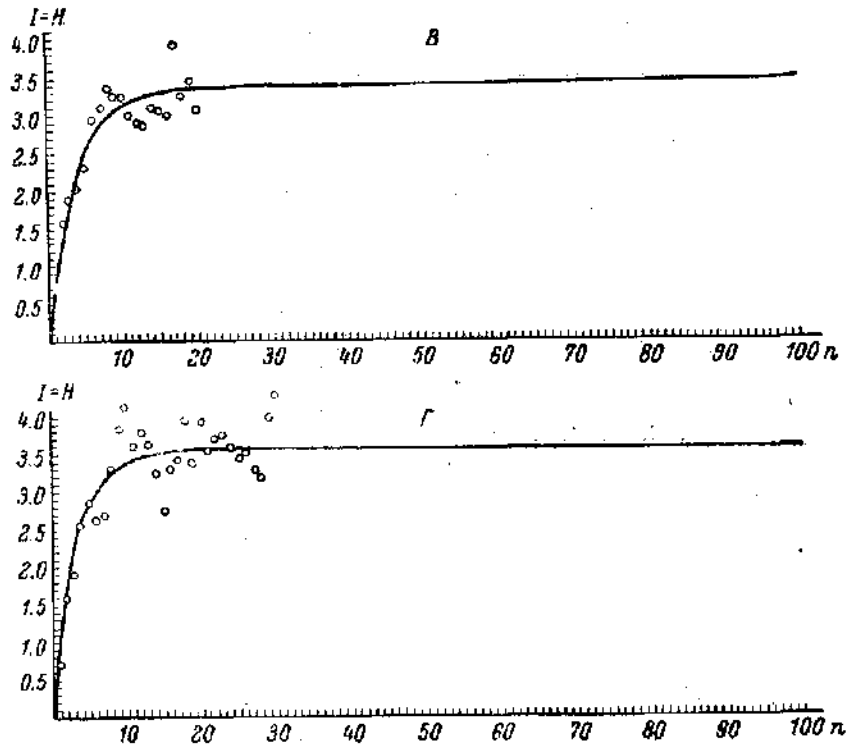


Рис. 5 *(продолжение)*.

В — беллетристический стиль; *Г* — деловой стиль.

где $H^{(л)}(t)$ — энтропия данного участка сообщения (буквы), получаемая при учете лексических (частично — грамматических) связей между словами; $H_\infty^{(л)}$ — предельная энтропия, получаемая при тех же условиях; l — лексический коэффициент, аналогичный коэффициенту s в выражении (13). Значения остальных символов те же, что и в формуле (13). Величины $H_\infty^{(л)}$ и l даны в табл. 7 (графа 11).

Применяя рассуждения, приведенные на стр. 41, получаем

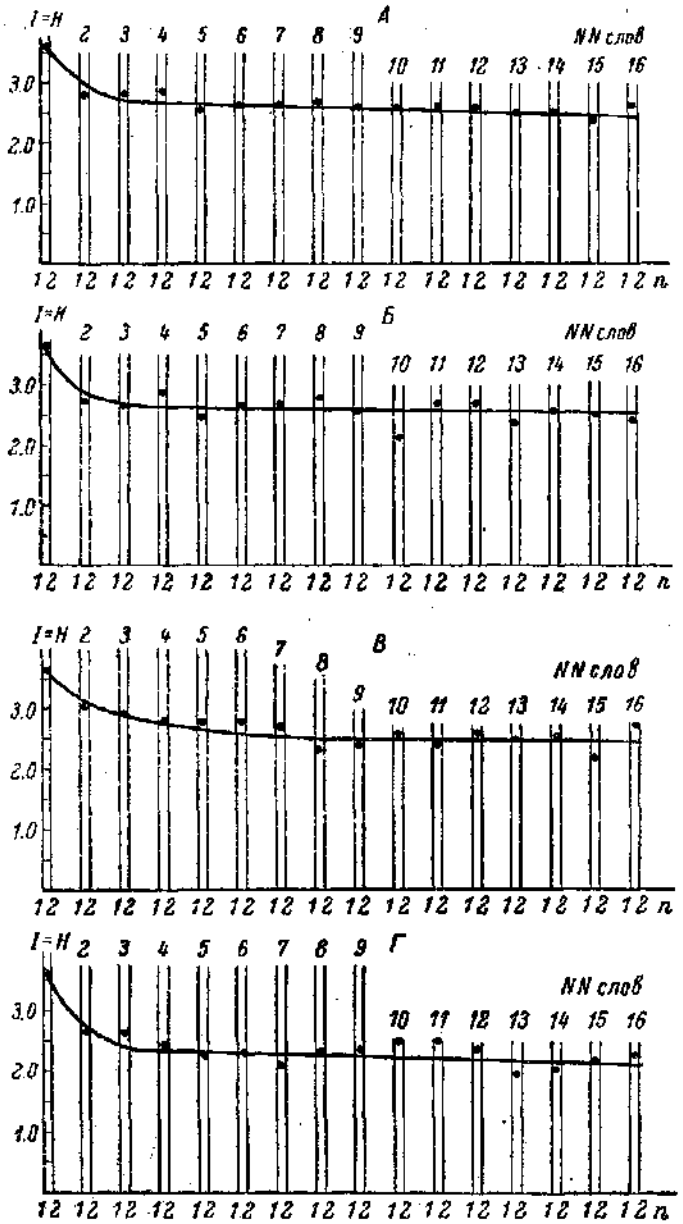


Рис. 6. График убывания энтропии буквы в зависимости от ее лексической обусловленности.

По оси абсцисс — начальные буквы слова, *по оси ординат* — энтропия (количество информации), в дв. ед.
А — язык в целом; *Б* — разговорная речь; *В* — беллетристический стиль; *Г* — деловой стиль.

количественную оценку лексической обусловленности данного участка текста, которая имеет следующий вид:

$$L(t) = \left(\frac{H'_1 + H'_2}{2} - H_\infty^{(л)}\right)(1 - e^{-t}). \qquad (19)$$

График зависимости (19) представлен на рис. 7.

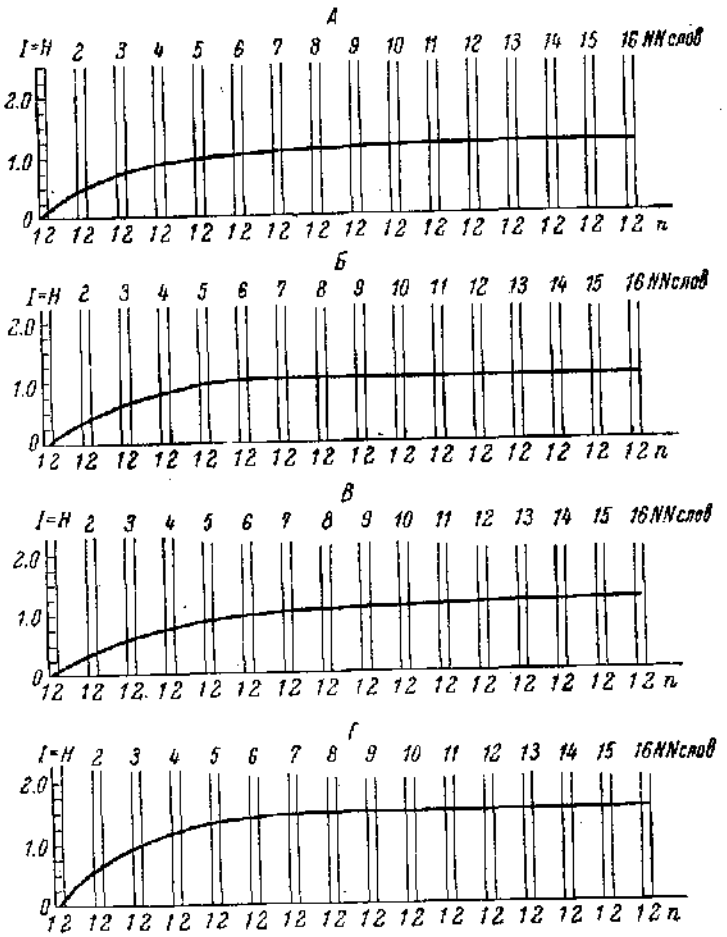


Рис. 7. График роста лексической обусловленности.
Обозначения те же, что и на рис. 6.

Для того чтобы охарактеризовать предел, к которому стремится лексическая обусловленность в данном языке и его стилях, введем понятие предельной лексической обусловленности (L_∞),

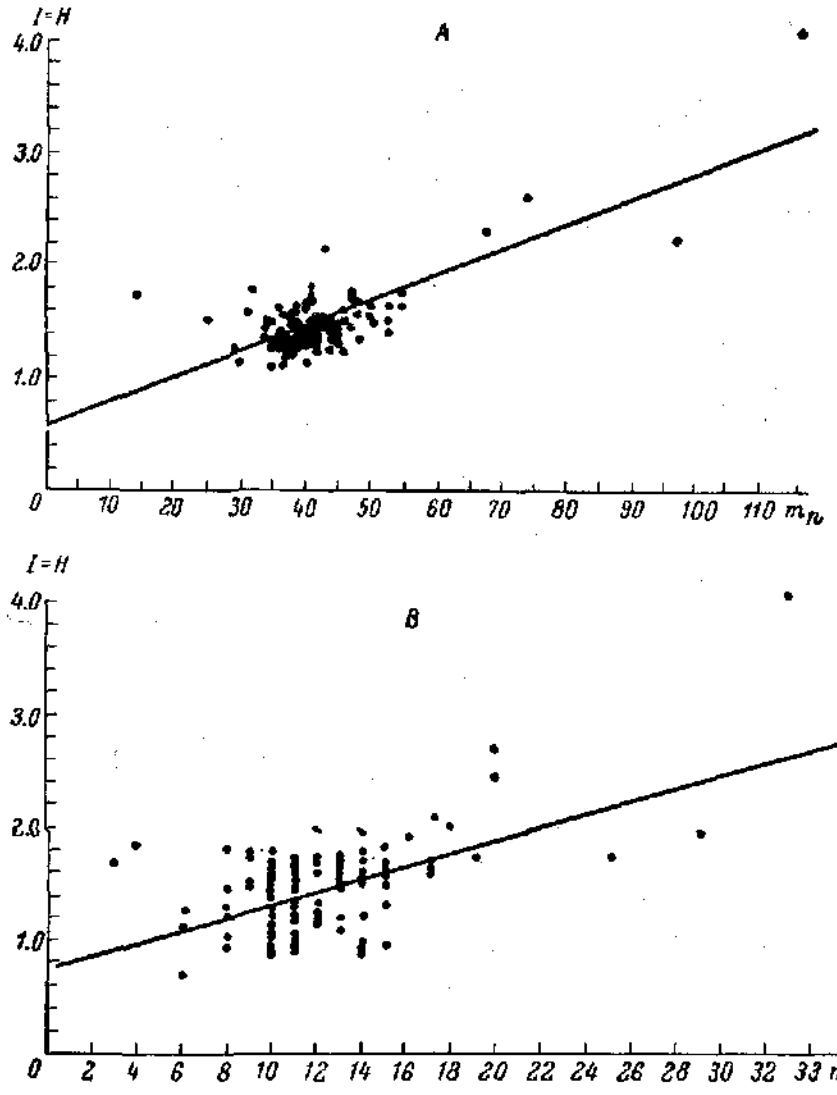


Рис. 8. График зависимости между величиной $\widetilde{H}'_n$ и количеством двух первых букв, падающих на буквенную позицию n.

По оси абсцисс — количество двух первых букв, падающих на буквенную позицию n, *по оси ординат* — энтропия (количество информации), в дв. ед.
А — язык в целом; *В* — беллетристический стиль.

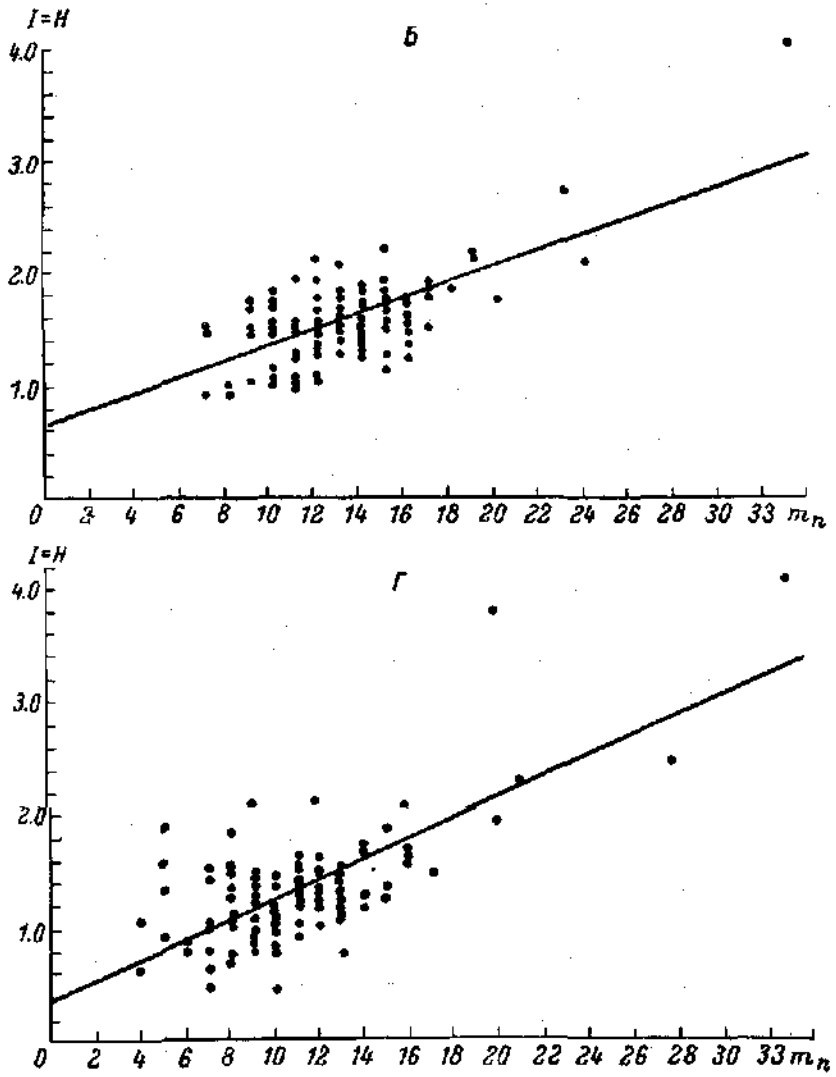


Рис. 8 *(продолжение)*.

Б — разговорная речь; Г — деловой стиль.

аналогичное понятию «предельная контекстная обусловленность» (см. выражение (17)). При этом

$$L_\infty = \frac{\bar{H}'_1 + \bar{H}'_2}{2} - H^{(n)}_\infty.$$

Значения L_∞ *даны в табл. 7 (графа 14).*

Нетрудно видеть, что предельная лексическая обусловленность единицы текста составляет от 32 до 42% от ее суммарной обусловленности (ср. табл. 7, графа 16).

Количественные оценки величин H_∞, s, K_∞, L_∞, l, *а также* величина избыточности языка, рассчитываемая по известной формуле вида $R = \frac{H_0 - H_\infty}{H_0}$, могут быть использованы для сопоставления языков и стилей.

Данные табл. 7 (графы 1—3, 8, 9, 17, 18, 23, 24, 26, 27) показывают, что значения предельной энтропии и избыточности во французском языке близки к значениям этих величин в русском и английском языках.

Сопоставление указанных величин по стилям французского языка показало, что наибольшей избыточностью и контекстной обусловленностью (соответственно наименьшей энтропией) обладает деловой стиль. Большая избыточность в этой выборке является следствием высокой лексической обусловленности единиц в деловом тексте (табл. 7, графа 16).

Такая высокая лексическая обусловленность, которая составляет 42% суммарной контекстной обусловленности буквы в тексте, определяется следующими факторами: 1) использованием большого количества устойчивых словосочетаний, связанных с той или иной тематикой; 2) сравнительно ограниченным кругом лексики, значительную часть которой образует терминология данной специальности; 3) логическим и строго нормализованным построением предложений.

Более низкая избыточность беллетристического стиля является результатом большей по сравнению с деловой речью неопределенности в выборе языковых элементов. Хотя беллетристический стиль также использует строго нормализованный синтаксис, лексические связи здесь значительно слабее: языковые штампы применяются реже, вместе с тем используется большое количество неожиданных сочетаний слов (метафоры и другие стилистические приемы), а круг лексики гораздо шире, чем в деловой речи.

Несколько неожиданным оказывается то, что разговорная речь имеет самую низкую избыточность и соответственно — самую высокую энтропию. Ведь разговорная речь использует большое количество штампов, а словарь ее довольно ограничен. По-видимому, отсутствие строгой нормализации как в словоупотреблении, так и во фразеологии значительно снижает лексическую обусловленность элементов текста, которая составляет всего лишь 32% суммарной контекстной обусловленности.

Перейдем теперь к рассмотрению непериодических колебаний в распределении энтропии, которое мы относили до сих пор за счет разброса. Мы будем рассматривать эти колебания только в верхней границе энтропии.

Уже в ходе эксперимента стало ясно, что угадывание первых букв слова требует от информанта бо́льших усилий, чем угадывание середины и конца слова. В связи с этим можно предположить, что появление максимумов на кривой распределения энтропии является результатом кумуляции начал слов на соответствующих номерах букв в тексте. Чтобы проверить это предположение, найдем зависимость между величиной энтропии буквы $\bar{H}'_n$ и количеством первых и вторых букв слова, падающих на букву (точнее — на буквенную позицию) n. Эта зависимость аппроксимируется двучленом первой степени вида

$$\bar{H}'_n = am_n + b, \qquad (20)$$

где m_n — количество первых и вторых букв на буквенной позиции n; a и b — константы формулы.

Таблица 11

Вид выборки	a	b	ε
Язык в целом	0,081	0,55	0,185
Разговорная речь	0,073	0,61	0,246
Беллетристический стиль	0,057	0,76	0,233
Деловой стиль	0,091	0,85	0,20

Значение этих констант и квадратическое отклонение представлены в табл. 11. Графики зависимости (20), подтверждающие наше предположение, даны на рис. 8. Формула (20) не является, разумеется, математическим законом. Ее следует рассматривать как грубое, но достаточно разумное приближение к реальной действительности.

Г. П. Богуславская

НОВЫЙ ЭКСПЕРИМЕНТ ПО ОПРЕДЕЛЕНИЮ ЭНТРОПИИ АНГЛИЙСКОГО ЯЗЫКА

В 1964 г. в Минском педагогическом институте иностранных языков был проведен новый эксперимент по определению энтропии английского языка и его разновидностей (стилей).

Необходимость такого эксперимента диктовалась следующими соображениями.

1. Осуществленные до настоящего времени исследования энтропии английского языка проводились либо на стилистически однородных текстах, либо на текстах, имевших незначительную стилистическую дифференциацию. Так, например, К. Шеннон использовал в своем эксперименте один текст — роман Д. Малона «Виргинец Джефферсон». [1] Н. Бартон и Дж. Ликлайдер провели эксперимент на текстах, взятых из десяти романов, принадлежащих современным американским авторам, — произведений, обладающих примерно одинаковой степенью трудности языка и отвлеченности изложения.[2] Э. Ньюмен и Н. Вог осуществили эксперимент относительно трех жанрово-стилистических разновидностей английского языка (религиозные и философские тексты, а также публицистика), однако в каждой из этих разновидностей были выбраны тексты, принадлежащие одному источнику (Библия, произведения В. Джеймса, журнал «Атлантик Монтли»).[3]

Выбор стилистически однородных текстов вполне понятен со статистической точки зрения. Этим путем исследователи стремились сократить статистический «разброс» и избежать произвола при определении пропорций жанров и выборе самих текстов. Использование стилистически однородных текстов оправдано с точки зрения основной задачи указанных работ. Эта задача состояла в том, чтобы показать возможность определить энтропию языка, а также в том, чтобы разработать лингвистическую и математическую методики исследования.

Вместе с тем для решения задач инженерной лингвистики необходимы сведения об энтропии не столько относительно отдельного произведения или подстиля (например, подстиля религиозных текстов), сколько относительно языка в целом и его основных разновидностей. Ведь машина должна быть готова к анализу любого текста, написанного на данном языке или подъязыке. Отсюда следует, что наряду с исследованиями стилистически однородных текстов необходимо провести эксперимент по определению энтропии английского языка в целом и его главных стилистических разновидностей. При этом нужно помнить, что в самой задаче заложена возможность заметного увеличения ошибки наблюдения.

2. Все исследователи, проводившие указанный эксперимент, пользовались при обработке его результатов такой методикой, которая либо давала завышенные значения энтропии,[4] либо показывала очень широкий интервал между верхней и нижней границами той области, в которой заключено истинное значение энтропии.[5]

В настоящее время разработана методика, позволяющая с большей точностью измерять энтропию текста в различных его участках. [6] Поэтому представляется целесообразным произвести на основе этой новой методики измерение основных информационных характеристик современного английского текста.

При проведении опыта по угадыванию букв неизвестного текста было взято 100 связных английских текстов длиной в 200 букв каждый. Тексты представляют следующие четыре стиля современного английского литературного языка.

1. Беллетристический стиль представлен 33 текстами из следующих произведений английских, американских и австралийских писателей: J. Aldridge. The Last Inch; D. M. Lessing. The Old Chief Mshlanga; G. Greene. Special Duties; T. Warner. Emil; [7] M. Quin. Survival of the Finkiest; M. Gold. The Damned Agitator; E. Hemingway. Old Man at the Bridge; J. Steinbeck. The

---

[1] См.: К. Шеннон. Предсказание и энтропия печатного английского текста. «Работы по теории информации и кибернетике», М., 1963, стр. 669—686.

[2] N. G. Burton and J. C. R. Licklider. Long-range Constraints in statistical Structure of Printed English. «American Journal of Psychology», LXVIII, 4, 1955, pp. 652, 653.

[3] E. B. Newman and N. C. Waugh. The Redundancy of Texts in Three Languages. «Information and Control», III, 2, 1960, pp. 141—153.

[4] См.: N. G. Burton and J. C. R. Licklider, ук. соч.; E. B. Newman and N. G. Waugh, ук. соч.; D. H. Carson. Letter Constraints within Words in Printed English. «Kybernetik», I, 1, 1961, pp. 46—54; E. B. Newman and L. J. Gerstman. A New Method for Analysing Printed English. «Journal of Experimental Psychology», XLIV, 2, 1952.

[5] К. Шеннон, ук. соч.; N. G. Burton and J. C. R. Licklider, ук. соч.

[6] См.: А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин. Энтропия русского языка. ВЯ, № 6, 1962.

[7] Первые четыре рассказа взяты из сб. «Modern English Short Stories», Foreign Languages Publishing House, M., 1961.

Chrysanthemums; S. V. Benet. Freedom's a Hard-bought Thing; [8] P. S. Buck. Dragon Seed (Philadelphia, 1946); R. Aldington. Death of a Hero (Foreign Languages Publishing House, M., 1958); K. Mansfield. Her First Ball, The Fly (Foreign Languages Publishing House, M., 1959); A. G. Cronin. Hatter's Castle (Foreign Languages Publishing House, M., 1963); D. Cusack. Say No to Death (Foreign Languages Publishing House, M., 1961) и др.

2. Публицистический стиль представлен 33 текстами из следующих английских и американских газет за 1963—1964 гг.: «The People», «The New York Times», «Sunday Mirror», «The Times», «The Financial Times», «Huddersfield Weekly Examiner» и др. Для отгадывания использовались статьи преимущественно на политические и экономические темы.

3. Поэзия представлена 10 текстами из стихотворений современных английских и американских поэтов: [9] A. E. Houseman. Soldier from the Wars Returning; W. B. Yeats. When you are Old; E. V. Milley. Justice Denied; W. C. Williams. Dawn; R. Frost. The Road Not Taken.

4. Для анализа устно-разговорной речи использовались магнитофонные записи непринужденных бесед с английскими и американскими туристами на бытовые темы.

Образец текста разговорной речи:

Oh, it is partly, to a certain degree, not completely. Some people like it, and others don't. Personally I don't understand it. But I'm told you are not supposed to understand it, so I even really don't know. (Запись устной речи. 16, IV, 1964. 200 букв).

Эксперимент проводился следующим образом.

Испытуемый последовательно угадывал, начиная со 2-й по 100-ю, буквы неизвестного ему текста (первая буква ему давалась). Затем экспериментатор сообщал продолжение текста от 101-й до 149-й буквы. 150-ю букву испытуемый отгадывал сам. Текст от 151-й до 199-й буквы снова сообщался испытуемому, и, наконец, 200-ю букву он угадывал сам.

Буквы 5, 10, 20, 30, 40, 50, 75, 100, 150, 200 отгадывались по полной программе, остальные — по сокращенной. При угадывании учитывались достоверные продолжения. Напоминаем, что под достоверным продолжением понимается единственно возможное с точки зрения орфографических норм и предшествующего контекста появление буквы.[10]

[8] Рассказы с 5-го по 9-й напечатаны в сб. «Modern American Short Stories», Foreign Languages Publishing House, M., 1960.

[9] Указанные поэтические тексты были взяты из сб. «An Anthology of Modern English and American Poetry», Госучпедгиз, Л., 1963.

[10] Описание полной и сокращенной программ угадывания с учетом достоверных продолжений см. в работе: А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин, ук. соч., стр. 116, 119, а также в статье Н. В. Петровой (часть I настоящего сборника).

В качестве испытуемого выступал студент-филолог (американец, родной язык — английский). Ему был предоставлен необходимый справочный материал: специально составленная таблица частотности начальных букв английских слов, таблицы частотности английских букв для отгадывания букв в середине слов и, наконец, следующие словари и грамматики: A. S. Hornby, E. V. Gatenby, H. W. Wakefield. The Advanced Learner's Dictionary of Current English. London, 1958; The Concise Oxford Dictionary of Current English. Oxford, 1956; D. Jones. English Pronouncing Dictionary. London, 1924, 1964; Scholastic Dictionary of American English. New York, 1962; Webster's Third New International Dictionary (unabridged). London, 1961; The Oxford English Dictionary. Oxford, 1961, vv. I—XII; Thorndike English Dictionary. London, 1948; Encyclopædia Britannica. London, 1962; В. К. Мюллер. Англо-русский словарь. М., 1960; А. В. Кунин. Англо-русский фразеологический словарь. М., 1955; R. W. Zandvoort. A Handbook of English Grammar. Bristol, 1960; G. Scheurweghs. Present-Day English Syntax. Bristol, 1959. Этот же материал служил вспомогательным средством в процессе вероятностно-лингвистической корректировки, которой были подвергнуты результаты нашего эксперимента (см. статью Н. В. Петровой, гл. II, § 3).

Вероятностно-лингвистическая корректировка осуществлялась самим экспериментатором (автором настоящей статьи) вместе с испытуемым. Для обсуждения сомнительных случаев привлекались лучшие знатоки современного английского языка из числа преподавателей МГПИИЯ (мы будем называть их впредь экспертами). Особо редкие, архаичные, диалектные и жаргонные слова при этом не учитывались.

В качестве примера такой корректировки рассмотрим один из протоколов угадывания букв в газетном тексте (табл. 1). Если руководствоваться данными словарей и современными лексико-грамматическими нормами английского языка, то полная достоверность появления соответствующих букв в позициях 12, 38, 39, 40, 46, 47, 61, 62, 71, 72, 80, 91, 92, 200 не вызывает сомнений. Буква 55 дает на первый взгляд четыре варианта продолжений предшествующего контекста. Эти варианты представлены в табл. 2. 2-й, 3-й и 4-й варианты признаны экспертами недопустимыми для данного контекста. Отсюда следует, что мы имеем здесь единственно возможное продолжение предшествующего контекста. Это означает, что буква 150 несет нуль информации.

Аналогичным путем были выявлены нули информации для букв 57, 69, 97.

Помимо определения достоверных продолжений, наша вероятностно-лингвистическая корректировка должна была выявить и устранить «лишние» попытки при угадывании отдельных букв. Например, при отгадывании буквы 37 экспериментатором были

Таблица 1

Образец протокола эксперимента по угадыванию букв с последующей вероятностно-лингвистической корректировкой

№№ букв	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Текст	*l*	*t*	Δ	*i*	*s*	Δ	*a*	*r*	*g*	*u*	*e*	*d*	Δ	*t*	*h*	*a*	*t*	Δ	*a*	Δ	*f*	*i*	*r*	*m*	Δ
Результаты угадывания в количестве попыток		2	1	2	1	1	2	2	2	1	2	1	1	1	1	1	1	1	2	1	2	2	2	2	2
То же после корректировки		2	1	2	1	1	2	2	2	1	2	0	0	1	1	1	1	1	2	1	2	2	2	2	2

№№ букв	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Текст	*f*	*i*	*n*	*d*	*i*	*n*	*g*	Δ	*i*	*t*	*s*	*e*	*l*	*f*	Δ	*e*	*a*	*r*	*n*	*i*	*n*	*g*	Δ	*s*	*u*
Результаты угадывания в количестве попыток	2	1	2	1	3	1	1	1	2	2	2	2	1	1	1	2	2	2	1	1	1	1	1	2	14
То же после корректировки	2	1	2	1	3	0	0	0	2	2	2	1	0	0	0	2	2	2	1	1	0	0	1	2	14

№№ букв	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
Текст	*b*	*s*	*t*	*a*	*n*	*t*	*i*	*a*	*l*	*l*	*y*	Δ	*i*	*n*	*c*	*r*	*e*	*a*	*s*	*e*	*d*	Δ	*p*	*r*	*o*
Результаты угадывания в количестве попыток	2	1	1	1	1	1	1	1	1	2	1	1	2	1	2	2	1	1	1	1	1	1	2	2	1
То же после корректировки	2	1	1	1	0	1	0	1	1	2	0	0	2	1	2	2	1	1	0	0	1	0	2	2	1

№№ букв	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Текст	*f*	*i*	*t*	*s*	Δ	*i*	*n*	Δ	*a*	Δ	*e*	*e*	*r*	*t*	*a*	*i*	*n*	Δ	*y*	*e*	*a*	*r*	Δ	*s*	*h*
Результаты угадывания в количестве попыток	1	1	1	1	1	2	1	1	2	1	2	2	2	1	1	1	1	1	2	1	1	1	1	2	1
То же после корректировки	1	1	1	1	0	2	1	1	2	1	2	2	2	1	1	0	0	2	2	1	1	0	1	2	1

Таблица 1 (*продолжение*)

№№ букв	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
Текст	*o*	*u*	*l*	*d*	Δ	*l*	*o*	*w*	*e*	*r*	Δ	*p*	*r*	*i*	*c*	*e*	*s*	Δ	*b*	*u*	*t*	Δ	*w*	*i*	*l*
Результаты угадывания в количестве попыток																									
То же после корректировки																									

№№ букв	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
Текст	*l*	Δ	*t*	*h*	*a*	*t*	Δ	*f*	*i*	*r*	*m*	Δ	*t*	*h*	*e*	*n*	Δ	*b*	*e*	Δ	*a*	*b*	*l*	*e*	Δ
Результаты угадывания в количестве попыток																									1
То же после корректировки																									1

№№ букв	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
Текст	*t*	*o*	Δ	*r*	*a*	*i*	*s*	*e*	Δ	*i*	*t*	*s*	Δ	*p*	*r*	*i*	*c*	*e*	*s*	Δ	*a*	*g*	*a*	*i*	*n*
Результаты угадывания в количестве попыток																									
То же после корректировки																									

№№ букв	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
Текст	Δ	*a*	*n*	*d*	Δ	*r*	*e*	*s*	*t*	*o*	*r*	*e*	Δ	*i*	*t*	*s*	Δ	*p*	*r*	*o*	*f*	*i*	*t*	*s*	Δ
Результаты угадывания в количестве попыток																									1
То же после корректировки																									6

Таблица 2

	Буквы														Значение *
	49	50	51	52	53	54	55	56	57	58	59	60	61	62	in a substantial manner
1	s	u	b	s	t	a	n	t	i	a	t	t	y	Δ	
2	s	u	b	s	t	a	g	e	Δ						(*n*) 1) a subdivision of a stage and esp. of a geological stage 2) an attachement to a microscope by means of which accessories are held in place beneath the stage of the instrument
3	s	u	b	s	t	a	l	a	g	m	i	t	e	Δ	(*n*) compact monocrystalline deposit of calcium carbonate (*adj*)
4	s s	u u	b b	s s	t t	a a	l i	a o	g n	m Δ	i	t	i	e	(*n*) a station subordinate or subsidiary to another station

* Значения слов взяты из словаря Уэбстера (Webster's Third New International Dictionary (unabridged), London, 1961).

зафиксированы две попытки (табл. 1). Однако проверка по словарям возможных продолжений показала, что вероятны только два выбора (табл. 3).

Таблица 3

	Буквы							Значение
	34	35	36	37	38	39	40	
1	i	t	s	e	l	f	Δ	(*pron.*) emphatic and reflexive form corresp. to it
2	i	t	s	Δ				(*adj.*) of or belonging to it or itself as possessor

Нетрудно заметить, что независимо от того, отгадал или не отгадал испытуемый с первой попытки правильное продолжение, ответ экспериментатора дает исчерпывающую информацию о том, какая буква стоит в рассматриваемых участках текста (если не *e*, то «пробел», и если не «пробел», то *e*). Это дает нам право отметить в протоколе для буквы 37 одну попытку независимо от того, угадал или не угадал испытуемый соответствующую букву.

После завершения вероятностно-лингвистической корректировки был осуществлен расчет энтропии для тех букв, которые угадывались по полной программе.

Нижняя граница энтропии в экспериментальных текстах определялась по формуле Шеннона

$$\underline{H}_n = 2(p_2 - p_3) + 3(p_3 - p_4)\log_2 3 + \ldots + (S-1)(p_{S-1} - p_S) \times \log_2(S-1) + S p_S \log_2 S \qquad (1)$$

(ср. выражение (1) в статье Н. В. Петровой),

где p_2, $p_3 \cdots p_{S-1}$, p_S — вероятности угадывания n-й буквы исходя из $n-1$ предшествующих букв.

В нашем опыте $S \leqslant 27$, поскольку, помимо 26 букв, составляющих алфавит английского языка, включались еще два символа: пробел и дефис (один символ) и апостроф (один символ).

Верхняя граница определялась по модифицированной формуле Шеннона [11]

$$H_n = (1 - p_0)\log_2(1 - p_0) - p_1 \log_2 p_1 - p_2 \log p_2 - \ldots - p_N \log_2 N. \qquad (2)$$

Применение формулы (2) заметно сокращает по сравнению с результатами К. Шеннона интервал между $\bar{H}$ и $\underline{H}$. Об этом можно судить по данным табл. 4.

Таблица 4

Ширина интервала, в котором заключено истинное значение энтропии

	Собственные данные (беллетристический стиль)			Данные К. Шеннона (роман Д. Малона «Виргинец Джефферсон»)		
№№ букв . .	5	10	100	5	10	100
$\bar{H} - \underline{H}$, дв. ед.	0.69	0.41	0.36	1.0	1.1	0.7

Эксперимент Бартона и Ликлайдера [12] также дает широкий интервал между $\bar{H}$ и $\underline{H}$ (этот интервал $\approx$1.0 дв. ед.).

Оценка энтропии определялась относительно следующих четырех выборок (табл. 5—9):

1) «язык в целом», представленный 100 текстами различных жанров;

2) беллетристический стиль (33 текста);

3) публицистический стиль (33 текста);

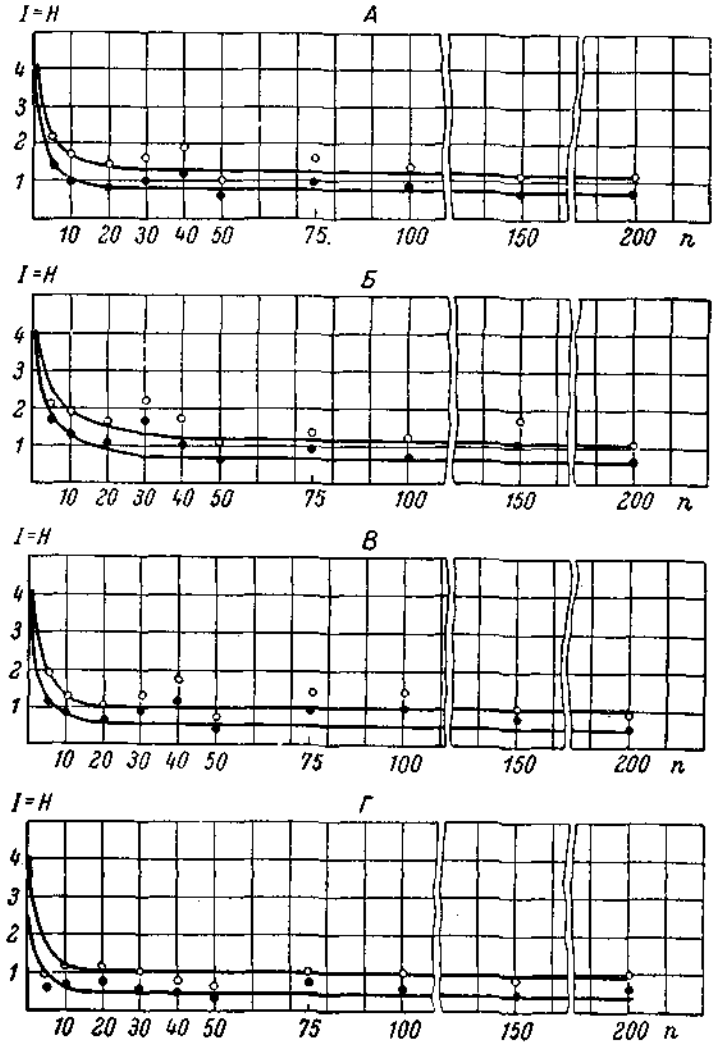
4) разговорная речь (33 текста).

Выборка 1 («язык в целом») составлена (см. стр. 51) из 30 беллетристических, 30 публицистических, 30 разговорных текстов, входящих одновременно в выборки 2, 3, 4. Кроме того, в выборку 1 включено 10 поэтических текстов, не образующих

[11] См.: А. А. Потровская, Р. Г. Потровский, К. А. Разживин, ук. соч., стр. 119.

[12] См.: N. G. Burton and J. C. R. Licklider, ук. соч.

самостоятельной выборки. Соотношение энтропии по стилям аналогично соответствующим данным в русском языке. Наибольшей



Графики распределения энтропии в английских печатных текстах.

По оси абсцисс — номера букв, *по оси ординат* — энтропия (количество информации), в дв. ед.
А — язык в целом; *Б* — разговорная речь; *В* — беллетристический стиль; *Г* — публицистический стиль.

энтропией обладает разговорная речь, неопределенность в публицистическом стиле является самой низкой (табл. 10 и рисунок).

Таблица 5

Распределение энтропии в английском 200-буквенном тексте (язык в целом)

№№ букв (n)	H_n	$\bar{H}_n$	№№ букв (n)	H_n	$\bar{H}_n$
5	1,32	2,07	50	0,54	1,05
10	0,97	1,67	75	0,95	1,58
20	0,80	1,43	100	0,82	1,39
30	0,97	1,60	150	0,69	1,13
40	1,15	1,83	200	0,63	1,15

Таблица 6

Распределение энтропии в английском 200-буквенном тексте (беллетристический стиль)

№№ букв (n)	H_n	$\bar{H}_n$	№№ букв (n)	H_n	$\bar{H}_n$
5	1,16	1,85	50	0,41	0,76
10	0,80	1,21	75	0,93	1,42
20	0,56	1,00	100	1,05	1,41
30	0,85	1,31	150	0,76	0,95
40	1,14	1,69	200	0,46	0,84

Таблица 7

Распределение энтропии в английском 200-буквенном тексте (публицистический стиль)

№№ букв (n)	H_n	$\bar{H}_n$	№№ букв (n)	H_n	$\bar{H}_n$
5	0,56	0,98	50	0,27	0,57
10	0,59	1,14	75	0,69	1,09
20	0,61	1,12	100	0,53	0,98
30	0,53	0,99	150	0,43	0,79
40	0,48	0,76	200	0,59	0,99

Таблица 8

Распределение энтропии в английском 200-буквенном тексте (разговорная речь)

№№ букв (n)	H_n	$\bar{H}_n$	№№ букв (n)	H_n	$\bar{H}_n$
5	1,65	2,06	50	0,60	1,08
10	1,23	1,76	75	0,83	1,35
20	1,03	1,65	100	0,69	1,21
30	1,60	2,22	150	1,12	1,72
40	1,08	1,71	200	0,53	1,02

Таблица 9

Вероятности «нулей информации» и верхние границы энтропии в английском 200-буквенном тексте (язык в целом)

№№ букв	Язык в целом			№№ букв	Язык в целом		
	p_0	H_n	H'^{*}_n		p_0	H_n	H'^{*}_n
5	0,19	2,07	2,25	75	0,34	1,58	1,78
10	0,23	1,67	1,82	100	0,24	1,39	1,55
20	0,18	1,43	1,52	150	0,35	1,13	1,34
30	0,27	1,60	1,78	200	0,32	1,15	1,30
40	0,23	1,83	2,02				
50	0,29	1,05	1,15	Средняя величина энтропии на букву в данной выборке букв		1,49	1,65

* $\bar{H}'$ — оценка верхней границы энтропии без учета нулей информации по формуле К. Шеннона, которую можно записать в виде

$$\bar{H}' = -(p_0 + p_1)\log_2(p_0 + p_1) - p_2\log_2 p_2 - \cdots p_S\log_2 p_S.$$

Таблица 10

Энтропия и избыточность английского языка и его жанрово-стилистических разновидностей

Английский язык, собственные данные					Избыточность (в %) английского языка, по данным других исследователей	
Энтропия, в дв. ед.			избыточность в %			
	H	$\bar{H}$	R	$\bar{R}$	R	$\bar{R}$
Язык в целом . . .	0.82	1.39	71.1	83.0	70.0 * 67.0 **	81.0 80.0 **
Разговорная речь . .	0.92	1.47	69.4	81.0		
Беллетристический стиль	0.80	1.20	75.0	83.4		
Публицистический стиль	0.50	0.88	81.8	89.6		

* C. Shannon, ук. соч., стр. 58—84.
** N. G. Burton, J. C. R. Licklider. Long-range Constraints in Statistical Structure of Printed English. «American Journal of Psychology», LXVIII, 4, 1955, pp. 652—653.

Поскольку величина энтропии зависит от числа букв в алфавите данного языка и не может служить общей мерой для количественного сопоставления языков с разным числом символов в алфавите, удобнее пользоваться при сравнении языков оценками избыточности (R). Эти оценки практически мало зависят от числа букв в алфавите.

Избыточность определялась по известной формуле

$$R = \frac{H_0 - H_\infty}{H_0}. \qquad (3)$$

Сопоставление наших данных об избыточности и энтропии английских печатных текстов с данными К. Шеннона и Н. Бартона и Дж. Ликлайдера показывает, во-первых, общее совпадение результатов всех трех экспериментов. Это говорит о том, что наши данные обладают достаточно высокой степенью надежности.

Вместе с тем использование новых приемов для определения верхней границы энтропии дало возможность получить более узкий интервал между верхней и нижней границами той области, в которой заключено истинное значение энтропии. Иными словами, мы получили более точные количественные оценки для энтропии и избыточности английского языка и его разновидностей. Несмотря на различие грамматических и лексических структур английского, русского и французского языков, значения избыточности и отчасти энтропии в этих языках очень близки (см. табл. 7 в статье Н. В. Петровой и табл. 10 настоящей статьи).

Часть II. СТАТИСТИЧЕСКАЯ СТРУКТУРА ТЕКСТА

Н. М. Алексеев

ЧАСТОТНЫЕ СЛОВАРИ И ПРИЕМЫ ИХ СОСТАВЛЕНИЯ

Частотные словари отличаются от любых других словарей тем, что в них лексические единицы регистрируются вместе с частотами их употребления. Составленные к настоящему времени частотные словари и списки можно классифицировать по следующим особенностям формального и содержательного характера.

1. По расположению словарного материала. Лексические единицы частотных словарей могут быть упорядочены либо по алфавиту, либо по убыванию их частот.

2. По объему словника. К полным следует отнести те частотные словари, в которых фиксируются все лексические единицы, обнаруженные при анализе данного текста или совокупности текстов, а к неполным — те словари, в которых приводится лишь часть единиц (обычно самых употребительных).

3. По объему обследованного материала. В зависимости от объема обследованного речевого материала условно выделяются «большие», «средние» и «малые» частотные словари.[1]

4. По жанрово-тематической или авторской принадлежности обследованного материала. Следует различать частотные словари устной и письменной форм речи, общие и отраслевые словари, словари отдельных произведений или авторов.

5. По лексическим единицам словника. Частотные словари могут регистрировать словоформы, исходные формы слов, основы (в терминах машинного перевода), словосочетания.

6. По способу частотной квалификации единиц. Указывается либо абсолютная частота употребления каждой единицы, либо какая-то другая частотная характеристика, например количество источников, в которых встретилась данная единица (так называемые распределительные словари).

7. По способу анализа речевого материала. В зависимости от методики обработки материала можно различать частотные словари, составленные в результате «сплошного» расписывания текста

[1] Ср.: Л. Н. Засорина. О статистике и автоматизации в лексикографии. НДВШ, сер. «Филологические науки», 1963, № 4, стр. 97.

(например, словоуказатель к произведениям А. С. Пушкина [2] или индекс к «Улиссу» Дж. Джойса [3]), и словари, полученные путем выборочного анализа лексики. Последний подход является наиболее распространенным при составлении частотных словарей.

Достоверность частотного словаря обычно оценивается в зависимости от объема обследованного материала (количественная характеристика) и от содержания этого материала (качественная характеристика). Для количественной оценки используются приемы математической статистики. Качественная надежность словаря обусловливается подбором текстов для анализа при его составлении.

В группе «Статистика речи» при составлении частотных словарей вручную используется в общем единообразная методика.

1. Анализируются отрезки связного текста, относящиеся к определенной жанрово-тематической совокупности — «подъязыку» (ср. подъязыки электроники, устной речи, газетных текстов и т. д.).

2. Минимальным отрезком текста для статистических подсчетов является текст или группа текстов общей длиной в 1000 словоупотреблений.

3. Под словоупотреблением понимается одна из всех словоформ текста или любая последовательность букв, ограниченная двумя пробелами. При этом дефис считается буквой, а цифры и формулы исключаются из подсчетов.

4. Лексической единицей основного частотного списка считается одна из разных словоформ текста.

5. Среди словоформ, которые группируются под тем или иным словом, выделяется исходная форма слова, выступающая одновременно репрезентантом данного слова. Исходные формы слов являются единицами дополнительного частотного списка.

6. В зависимости от задач, которые ставит перед собой составитель, учитываются один или более из трех возможных видов омонимии (омографии) — лексической, лексико-грамматической или грамматической.

7. Материал для анализа подбирается по специально разработанной для каждого подъязыка схеме дозировки текстов.

8. Объем материала для каждого частотного словаря устанавливается равным 200 000 словоупотреблений.

9. В целях выявления и описания закономерностей, которым подчиняется словарь текста, весь материал разбивается на четыре части. Иными словами, составляются частотные словари на основе выборок в 50 000, 100 000, 150 000 и 200 000 словоупотреблений.

10. В окончательном виде словарь представлен в двух основных списках — частотном и алфавитном списках словоформ — и дополнительных, например в виде частотного списка исходных форм слов, списка словоформ без какого-либо различения омографов и. т. п.

Технические приемы анализа речевого материала заключаются в регистрации каждой обнаруженной словоформы на карточке с указанием ее частоты и количества источников, грамматического класса и категории. Некоторые составители используют в качестве первого этапа работы регистрацию словоформ в списках (т. е. изготовление алфавитно-частотного словаря каждого отрезка текста в 1000 словоупотреблений) с последующим переносом их на карточки.

---

[2] Словарь языка Пушкина, тт. I—IV, М., 1956—1961.

[3] M. H a n l e y. Word Index to James Joyce's «Ulysses». Madison, 1951.

Е. А. Калинина

ИЗУЧЕНИЕ ЛЕКСИКО-СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ НА ОСНОВЕ ВЕРОЯТНОСТНОЙ МОДЕЛИ

§ 1. Задачи работы

Данная работа посвящается решению следующих теоретических и практических задач.

1. Найти количественную меру эффективности и достоверности частотных словарей и количественный метод сравнения частотных словарей.

2. Исследовать и описать статистические закономерности, управляющие лексикой русских текстов по электронике.

3. Составить частотный словарь современных русских текстов по электронике.

Теоретической основой работы является вероятностная модель текста.

Экспериментальную основу работы образовали восемь алфавитных словарей, составленных нами по выборкам различной длины.[1]

По основной выборке были составлены частотный словарь словоформ в частотном порядке и частотный словарь слов в алфавитном и частотном порядке. Для проверки теоретических выводов привлекались также некоторые другие частотные словари.

Мы считаем известными такие основные понятия и положения математической статистики и статистической обработки лексики, как выборка, частота в выборке, частотный словарь, ранг, математическое ожидание и дисперсия, связь частоты и вероятности,

[1] Чтобы облегчить ссылки в тексте, присвоим некоторым выборкам номера: 15 620 словоформ — I выборка, 9418 словоформ — II, 50 000 словоформ — III, 100 000 словоформ — IV, 150 000 словоформ — V, 200 894 словоформы — VI (основная). I и II выборки входят составной частью в III, *выборка III в IV, выборка IV в V и V — в основную.*

соотношение математического ожидания, случайного и выборочного среднего значения случайной величины. Эти понятия в работе не определяются.

§ 2. Вероятностная модель

Основное положение вероятностной модели — это представление текста (речи) как случайного процесса, а единицы текста (речи) как случайного события. Такими единицами могут быть элементы различных уровней иерархической структуры (фонемы, буквы, морфемы, слоги и т. д.). При исследовании лексики основными единицами являются словоформа и слово.

Случайный процесс предполагает некоторую производящую его систему, а случайное событие — некоторый комплекс условий, при выполнении которых происходит или не происходит данное событие. Производящей системой текста является язык.[2] Комплекс условий, определяющих набор возможных случайных событий, представлен темой, жанром, близлежащими словоформами и т. д. Этот комплекс может быть определен с той или иной степенью детализации, диктуемой поставленными задачами.

В нашей работе мы ограничиваемся текстами по электронике на русском языке.

В вероятностном подходе существенна принципиальная возможность воспроизведения комплекса условий. Относительная частота словоформы при неограниченном числе воспроизведений (практически — при достаточно большом) совпадает с вероятностью словоформы. Ясно, что действительные условия, в которых были написаны определенные тексты, в точности невоспроизводимы и неповторимы. Однако возможно приближенное воспроизведение этих условий.

При моделировании текста мы отказываемся от описания одного ряда особенностей текста с тем, чтобы получить возможность описать другой ряд его свойств.

Так, мы считаем все тексты выбранного типа реализацией некоторой одной производящей системы, так сказать, «среднего языка».[3]

Основное условие, накладываемое на производящую систему, — требование того, чтобы вероятность словоформ была близка к их частотам в достаточно большом числе действительных текстов.

Идеализируя свойства реального текста, будем считать, что в моделирующем тексте вероятность употребления словоформы не зависит от соседних словоформ. Такая идеализация дает возмож-

[2] Р. М. Фрумкина. От статистического описания речи к статистическим моделям языка. Тез. докл. межвузовск. конфер. на тему «Язык и речь», М., 1962.

[3] И. М. Яглом, Р. Л. Добрушин и А. М. Яглом. Язык и *теория информации.* ВЯ, IX, 1, 1960.

ность описывать такие закономерности, как покрываемость словарем нового текста, точность в определении частот и т. д. Но в качестве платы за возможность описывать данной моделью эти закономерности мы должны отказаться от описания, например, избыточности в языке, синтаксических связей, корреляционных связей внутри модели.

Для описания этих закономерностей нужно обращаться к другим моделям, например к марковским цепям.

Из сказанного ясно, что мы принимаем два упрощающих допущения.

Во-первых, следуя Мандельброту,[4] предполагаем моделирующий текст стационарным случайным процессом, т. е. предполагаем неизменность во времени производящей системы и комплекса условий. Реальные тексты являются реализациями нестационарного случайного процесса, так как тексты, написанные *разными авторами*, очевидно, будут отличаться друг от друга лексикой, как и тексты, написанные одним и тем же автором на разные темы. Если снова привлечь терминологию теории вероятностей, то можно сказать, что в реальных текстах распределение вероятностей *употребления словоформ не постоянно*, а зависит от места в тексте, от автора, от развиваемой темы, жанра, стиля и т. д.

Во-вторых, мы *считаем моделирующий текст последовательностью* независимых испытаний, т. е. не учитываем вероятностных связей между элементами. Таким образом, мы считаем вероятностную модель полностью заданной, если нам даны вероятности употребления изучаемых единиц (у нас — слов и словоформ).

Все единицы можно перенумеровать в порядке убывания их вероятностей.

Список словоформ даст вероятностный словарь словоформ, список слов — вероятностный словарь слов. Номер единицы в словаре даст ранг.

Очевидно, что оба предположения делают моделирующий текст в некоторых отношениях условным и не совсем похожим на реальный текст, прежде всего в отношении передаваемой информации и осмысленности.

Допущение упрощающих предположений дает нам возможность применять точные методы. Здесь мы сталкиваемся с противоречием применения точных методов в лингвистике: чтобы их применять, мы вынуждены заранее идти на принятие моделей, лишь приближенно адекватных изучаемым объектам. Те точные выводы, которые делаются, точны лишь для модели, для реального текста они становятся приближенными. Поэтому их справедливость, *практическая применимость* должны проверяться на фактическом материале. Мы увидим, что часто выводы для модели вполне удовлетворительно выполняются и на реальных текстах, но в каждом отдельном случае проверка обязательна.

[4] Б. Мандельброт. Закон Ципфа и определение «ударения». Сб. «Теория передачи сообщений». Под ред. В. И. Сифорова. М., 1957.

Заметим здесь, что совершенно аналогичное положение имеет место и в других лингвистических моделях, *широко развиваемых* в последнее время. Эффективность и ценность модели должна проверяться на *реальном материале*, только *тогда свойства модели*, теоретически исследуемые легче и полнее, чем свойства действительных *текстов*, *могут считаться описанием свойств* самого текста.

§ 3. Достоверность частотного словаря

Частотный словарь представляет собой словник некоторой выборки из текстов, причем у словарных единиц указывается частота их в выборке и порядковый номер (ранг) в ряду словарных единиц, если словарные единицы поставлены по порядку убывания частот. Под словарной единицей мы понимаем слово или словоформу.

Частотный словарь по своей природе случаен; если взять новую выборку из аналогичных текстов той же длины, то второй частотный словарь будет отличаться от первого: во-первых, те же словарные единицы будут иметь иной ранг и частоту, во-вторых, какое-то количество словарных единиц будет новым, а какое-то совсем не войдет во второй словарь. Под достоверностью частотного словаря мы понимаем точность в определении частот, вероятностей и рангов словарных единиц, т. е. близость частотного словаря к вероятностному. Вероятностная модель и теория вероятностей позволяют нам считать, что при достаточно большой выборке частотный словарь будет совпадать с вероятностным словарем. Числовые характеристики вероятностного словаря являются математическими ожиданиями аналогичных характеристик частотного словаря.

Обозначим буквой i ранг словарной единицы в вероятностном словаре, буквой p_i — ее вероятность и F_i — ее частоту в выборке длины N, *по которой составляется частотный словарь*. Прежде всего заметим, что математическое ожидание частоты F_i (среднее значение по достаточно большому количеству выборок длины N) равно Np_i. Это значит, что относительная ошибка в определении абсолютной частоты i-единицы словаря и ее вероятности одинаковы, так как математическое ожидание абсолютной частоты и вероятность пропорциональны. Найдем ошибку в определении вероятности. Мы можем судить о величине вероятности p_i только по относительной частоте $f=\frac{F_i}{N}$ i-той единицы в выборке. В теории вероятностей задача обобщается: нужно оценить точность определения вероятности некоторого события по его относительной частоте.

Если требовать абсолютно достоверного суждения о вероятности p_i, то нельзя сказать ничего, кроме тривиального соображения, что определяемая вероятность заключается в пределах от 0 до 1. Всякое более содержательное суждение о вероятности по относительной частоте сопряжено с риском ошибки.

В теории вероятностей определена вероятность отклонения относительной частоты события от его вероятности на данную величину, а именно: если p — вероятность события, $f=\frac{F}{N}$ его относительная частота в N опытах, то при некоторых допущениях вероятность неравенства

$$\left|p-\frac{F}{N}\right| \leqslant Z_\rho \sqrt{\frac{p_0(1-p)}{N}}$$

равна ρ, где коэффициенты Z_ρ и ρ связаны табулированной зависимостью (табл. 1).

Таблица 1

ρ	Z_ρ
0.90	1.64
0.95	1.96
0.98	2.33
0.99	2.58

В нашем случае событие — это появление в выборке словарной единицы с номером i, вероятность $p=p_i$, и мы можем сказать, что вероятность неравенства

$$\left|p_i-\frac{F_i}{N}\right| \leqslant Z_\rho \sqrt{\frac{p_i(1-p_i)}{N}} \qquad (1)$$

равна ρ. Для слов и словоформ вероятности p_i много меньше единицы. Поэтому можно считать, что $1-p_i \approx 1$. По формуле (1) мы можем написать неравенство для относительной ошибки δ_i в определении вероятности p_i по относительной частоте $\frac{F_i}{N}$, разделив неравенство (1) на p_i и заменив $1-p_i$ на 1:

$$\delta_i=\frac{\left|p_i-\frac{F_i}{N}\right|}{p_i} \leqslant \frac{Z_\rho}{\sqrt{Np_i}} \qquad (2)$$

с вероятностью ρ. Число $\frac{Z_\rho}{\sqrt{Np_i}}$ играет роль максимальной относительной ошибки в определении вероятности и частоты i-единицы словаря.

Если нас интересует отклонение частоты F_i i-единицы от математического ожидания, то, умножив неравенство (1) на N и заменив $1-p_i \approx 1$, найдем, что вероятность неравенства

$$|Np_i-F_i| \leqslant Z_\rho \sqrt{Np_i} \qquad (3)$$

равна ρ, другими словами 100 · ρ% всех словарных единиц должны иметь частоту, удовлетворяющую неравенству (3).

Интересующая нас ошибка δ_i в определении вероятности p_i по относительной частоте $f_i=\frac{F_i}{N}$ с вероятностью ρ не превосходит величины $\frac{Z_\rho}{\sqrt{Np_i}}$. Чтобы пользоваться этой формулой, необходимо было бы знать p_i. Мы можем воспользоваться значением p_i, которое дает закон Ципфа:

$$p_i=\frac{K}{i}, \qquad (4)$$

где K — константа, для которой Ципф дал значение 0.1. Как будет показано ниже, для русских текстов по электронике закон Ципфа для словоформ выполняется хорошо на рангах от $i=300$ до $i=4200$ с $K=0.145$.

Для этих рангов оценку относительной ошибки δ_i найдем, подставив выражение (4) в неравенство (2),

$$\delta_i \leqslant Z_\rho \sqrt{\frac{i}{NK}},$$

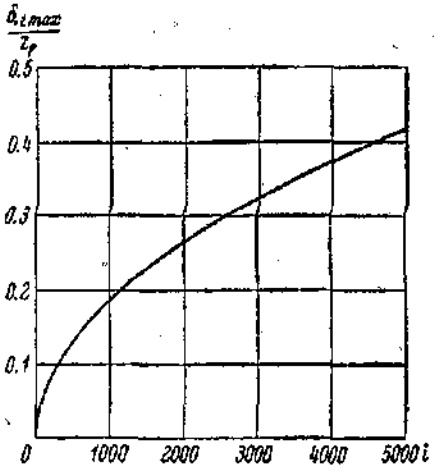


Рис. 1.

т. е. зависимость максимальной ошибки от ранга имеет вид параболы. На рис. 1 приводится зависимость максимальной относительной ошибки $\delta_{i\,max}=\frac{Z_\rho}{\sqrt{NK}}\sqrt{i}$ от ранга. Чтобы не связывать себя заранее выбором ρ и Z_ρ, мы откладываем на оси ординат $\frac{\delta_{i\,max}}{Z_\rho}$. Если выбор ρ сделан, то достаточно умножить данные графика на Z_ρ, чтобы получить максимальную ошибку при взятом ρ.

Но, как известно, на больших рангах закон Ципфа выполняется плохо, и здесь априорного значения p_i мы не имеем. Выход из этого указывает теория вероятностей: мы можем считать, что $Np_i \approx F_i$, если только F_i не слишком мало. Окончательно находим по формуле (2): относительная ошибка δ_i при замене вероятности i-единицы словаря ее относительной частотой с вероятностью ρ не превосходит величины $\frac{Z_\rho}{\sqrt{F_i}}$.

Соответственно по формуле (3) получаем, что частота F_i будет удовлетворять неравенству

$$|Np_i-F_i| \leqslant Z_\rho \sqrt{F_i} \qquad (5)$$

в среднем для 100 · ρ% словарных единиц.

Вероятность ρ выступает как степень достоверности суждения $\delta_i \leqslant \frac{Z_\rho}{\sqrt{F_i}}$, как уровень надежности этого суждения и называется достоверным уровнем.

Если мы будем высказывать о вероятности p_i, зная только относительную частоту $\frac{F_i}{N}$, лишь очень достоверные суждения (ρ близко к единице), то мы должны указывать для p_i широкий интервал, если же мы хотим указать для p_i узкий интервал, то наше суждение будет менее достоверно (ρ уменьшается).

Когда мы говорим, что вероятность определяется относительной частотой с ошибкой, не большей $\frac{Z_\rho}{\sqrt{F_i}}$, то мы в среднем ошибаемся в $(1-\rho)$ 100% случаев.

Выбор доверительного уровня ρ остается произвольным и диктуется требуемой надежностью выводов. Нас не должны соблазнять слишком большие значения ρ (ρ как вероятность не может превышать единицы, поэтому понятие «большие ρ» означает ρ, близкие к единице): требуя от своих суждений большой степени достоверности, мы рискуем высказывать только тривиальные, общеизвестные истины.[5] В нашем случае, потребовав 100%-й достоверности, мы вынуждены были бы сказать: искомая вероятность p_i находится между 0 и 1.

Чаще всего в прикладных задачах берут доверительный уровень $\rho=0.95$ или $\rho=0.99$. В тех задачах статистики, которые требуют исключительной надежности, даже доверительный уровень $\rho=0.99$ может быть недопустимо мал. В математической лингвистике столь высокой достоверности в ее прикладных задачах не требуется. В известных нам работах берут $\rho=0.95$.

Полученное выражение для максимальной ошибки может быть привлечено к выделению практически достоверной части словаря.

Разберем сначала случай, когда закон Ципфа выполняется удовлетворительно. Как мы получили в этом случае, максимальная относительная ошибка

$$\delta_{i\,\max} = Z_\rho \sqrt{\frac{i}{NK}}. \qquad (6)$$

Эта формула позволяет для имеющегося частотного словаря определить ту его часть, которую можно считать практически достоверной. «Достоверная часть» определяется не только словарем, но и его практическим применением: для одного применения может быть нужна бо́льшая точность, чем для другого.

Чтобы определить достоверную часть словаря, нужно задать максимальную ошибку $\delta_{i\max}$ и требуемую достоверность ρ. Тогда

[5] Эти рассуждения предполагают, что объем выборки N фиксирован.

достоверную часть словаря составляют словарные единицы до ранга $i_{\max}$, удовлетворяющего равенству

$$i_{\max} = \frac{NK}{Z_\rho^2}\delta_{i\max}^2. \qquad (7)$$

Например, для $\rho=0.90$ ($Z_\rho=1.64$), $N=200894$ (основная выборка), $\delta_{i\max}=0.5$, $K=0.145$ достоверны первые 2680 словоформ.

Если закон Ципфа выполняется неудовлетворительно, то максимальная ошибка равна

$$\delta_{i\,\max} = \frac{Z_\rho}{\sqrt{F_i}}.$$

В этом случае сначала по заданной максимальной ошибке $\delta_{i\max}$ и доверительному уровню ρ находим наименьшую частоту, считающуюся достоверной,

$$F_{i\,\min} = \frac{Z_\rho^2}{\delta_{i\,\max}^2}. \qquad (8)$$

Словоформы, имеющие меньшую частоту, опускаются в данном приложении.

Например, если в частотном словаре словоформ исключить все частоты ниже 31, то это будет соответствовать при доверительном уровне $\rho=0.95$ тому, что не учитываются все словоформы, вероятность которых определена с ошибкой, возможно достигающей 36% и более.

По этим же формулам определяется необходимый объем выборки, чтобы гарантировать достаточную точность.

Все выводы этого параграфа делались на основе вероятностной модели и, следовательно, строго справедливы лишь для моделирующего текста. Необходимо проверить, как они выполняются для реального текста, и здесь мы сталкиваемся с новым источником недостоверности частотного словаря, именно с нестационарностью, которой обладают реальные тексты как реализации случайного речевого процесса. Все же выводы с применением точных вероятностных формул не учитывают нестационарность условий, в которых создаются реальные тексты. Это приводит прежде всего к тому, что достоверность оценок для реального текста заметно ниже априорно принимаемой (и выполняющейся в модели). Был произведен следующий опыт: наугад были выбраны 100 словоформ с большими частотами, их частоты F_{1i} в первой половине основной выборки сравнивались с частотами F_{2i} во второй половине, причем принимались меры к тому, чтобы тематически обе половины были близки. Можно написать в соответствии с формулой (5), что в 100 $(2\rho-1)$ случаях должны выполняться неравенства

$$|F_{1i}-F_{2i}| = |(Np_i-F_{2i})-(Np_i-F_{1i})| \leqslant$$
$$\leqslant |Np_i-F_{2i}|+|Np_i-F_{1i}| \leqslant Z_\rho(\sqrt{F_{1i}}+\sqrt{F_{2i}}).$$

Здесь мы учли, что абсолютное значение суммы не больше суммы абсолютных значений слагаемых.

Для доверительного уровня $\rho=0.95$ удовлетворяли этому неравенству 77 словоформ вместо 90, а для $\rho=0.98$ удовлетворяли 87 вместо 96.

Таким образом, уровень надежности наших суждений о достоверности частотного словаря несколько ниже предполагаемого. Описанный опыт в силу его случайного характера не доказывает, разумеется, этого утверждения, а лишь его иллюстрирует. Здесь мы впервые сталкиваемся с действием указанного в § 2 противоречия: мы применяем точные методы к модели, а для реальных текстов точные выводы уже не точны. Об этом нужно помнить, чтобы точность методов не заслоняла неточности результатов.

§ 4. Эффективность частотного словаря

Под эффективностью частотного словаря мы понимаем степень покрываемости словарем нового текста, не вошедшего в выборку, по которой составлен частотный словарь. При этом подразумевается, что новый текст аналогичен вошедшим в выборку (в нашем случае новый текст должен принадлежать к текстам по электронике).

Таблица 2

r	$100P(r)$ %	
	словарь словоформ	словарь слов
50	26.3	33.4
100	32.5	43.5
500	51.5	72.2
1000	62	83.8

Обычно используемой характеристикой эффективности служит накопленная относительная частота $P(r)$,[6] определяемая как сумма относительных частот r первых по частотности единиц словаря,

$$P(r)=\frac{1}{N}\sum_{i=1}^{r}F_i. \qquad (9)$$

На рис. 2 изображена накопленная относительная частота $P(r)$ для основной выборки. На оси абсцисс выбран логарифмический масштаб, а на оси ординат — равномерный. Причины такого выбора будут объяснены ниже.

В табл. 2 представлены значения накопленной частоты для некоторых рангов.

Половину текста покрывают 447 словоформ и 148 слов, 75% текста покрывают 2420 словоформ и 587 слов.

Для тех рангов, где закон Ципфа выполняется с удовлетворительной точностью, может быть выведено теоретическое выражение для накопленной вероятности $P(r)$. Для русских текстов по электронике закон Ципфа (4)

$$p_i=\frac{K}{i}$$

[6] В других работах сборника эта величина обозначается символом f^*,

выполняется на рангах от $i=300$ до $i=4200$ с $K=0.145$, причем среднеквадратическая ошибка в определении постоянной K равна 4.2%.

В соответствии с определением $P(r)$ можем написать

$$P(r)=\sum_{i=1}^{r}p_i=\sum_{i=1}^{300}p_i+K\sum_{i=301}^{r}\frac{1}{i}=$$

$$=P(300)+K\left[\left(1+\frac{1}{2}+\ldots+\frac{1}{r}\right)-\left(1+\frac{1}{2}+\ldots+\frac{1}{300}\right)\right]=$$

$$=P(300)+K\left[(0.577+\ln r)-(0.577+\ln 300)\right]=$$

$$=P(300)-K\ln 300+K\ln r. \qquad (10)$$

При выводе мы использовали равенство

$$1+\frac{1}{2}+\ldots+\frac{1}{r}\approx\ln r+0.577.$$

Из данных частотного словаря получим, что 300 первых слов покрывают 44.7% текста, т. е. $P(300)=0.447$, по таблице натуральных логарифмов находим $\ln 300=5.704$. Подставим найденные константы в формулу (10) и перейдем от натуральных *логарифмов к десятичным по формуле*

$$\ln r=2.3026\cdot\lg r.$$

Окончательно получаем

$$P(r)=0.334\lg r-0.38. \qquad (11)$$

В табл. 3 приводим значения $P(r)$, вычисленные по формуле (11), и значения накопленной частоты для словаря словоформ.

Таблица 3

r	$P(r)$		r	$P(r)$	
	теоретическое значение	экспериментальное значение		теоретическое значение	экспериментальное значение
300	0.447	0.447	2000	0.723	0.720
400	0.489	0.485	2200	0.736	0.734
500	0.521	0.511	2400	0.749	0.747
600	0.548	0.541	2600	0.761	0.760
700	0.570	0.569	2800	0.771	0.771
800	0.590	0.585	3000	0.781	0.781
900	0.607	0.601	3400	0.795	0.799
1000	0.622	0.617	3800	0.816	0.815
1200	0.648	0.643	4200	0.830	0.830
1400	0.671	0.667	4600	0.843	0.842
1600	0.690	0.687	5000	0.855	0.853
1800	0.707	0.704	6000	0.882	0.876

Вернемся к вопросу о выборе масштаба на графиках. Масштаб может быть равномерным и неравномерным. Выбор единиц равномерного масштаба *диктуется диапазоном изменения переменных* и желаемым размером графика. Выбор неравномерного масштаба

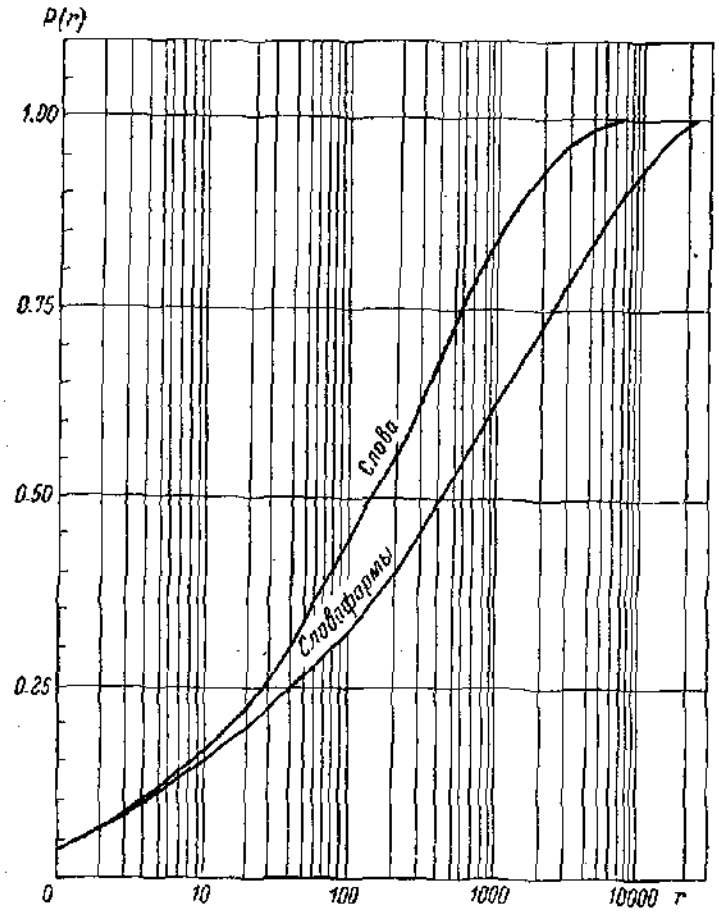


Рис. 2.

может вызываться разными причинами. Одна из основных причин — желание получить простое графическое изображение изучаемой зависимости, на котором легче выявить степень совпадения экспериментальных и теоретических значений. Чаще всего стремятся к тому, чтобы теоретическая зависимость изображалась в виде прямой линии. Например, закон Ципфа при логарифмическом масштабе по обеим осям имеет вид прямой линии.

Чтобы накопленные частоты изображались на графике прямой линией, нужно, как видно из формулы (11), взять на оси абсцисс логарифмический масштаб, а на оси ординат — равномерный. На рис. 2 видно, что в области рангов от $i=300$ до $i=4200$ точки $P(r)$ действительно в среднем ложатся на прямую. На остальных рангах формула (11) выполняется хуже. Другой причиной выбора масштаба может явиться желание особо выделить наиболее интересную часть диапазона изменения переменных. Например, логарифмический масштаб предоставляет на графике $P(r)$ рангам от $i=1$ до $i=100$ столько же места, сколько рангам от $i=101$ до $i=10\,000$. Это отвечает тому, что в отношении накопленной частоты малые ранги существеннее больших.

Строго говоря, накопленная частота $P(r)$ показывает заполнение первыми по частотности r-единицами словаря выборки, положенной в основу при составлении частотного словаря, а не нового текста. Но для достоверной (т. е. удовлетворяющей нас по точности) части словаря можно считать, что заполнение нового текста идет так же, как и заполнение основной выборки. Конец же кривой $P(r)$ недостоверен и дает завышенное значение эффективности словаря. Например, значение накопленной относительной частоты, сосчитанное по всему словарю, всегда дает значение 1, т. е. 100%-е заполнение, как бы мала ни была исходная выборка N.

Поставим задачу оценить эффективность всего частотного словаря. Принятая нами вероятностная модель позволяет решить эту задачу.

Пусть для составления частотного словаря словоформ взята выборка из N словоформ, в которой по одному разу встретилось m_1 словоформ. Мы берем новый текст из N_0 словоформ и хотим знать, какое число его словоформ не входит в словарь. Считаем, что словарь составлен по достаточно большой выборке N, много большей N_0.

Так как выборка N велика, то в соответствии с вероятностной моделью считаем, что не вошедшие в нее словоформы достаточно редкие и тем более они будут редкими для текста N_0. Если же они в текст N_0 войдут, то чаще всего по одному разу. Это значит, что новые словоформы однократные. В согласии с вероятностной моделью мы считаем однократные словоформы разбросанными в среднем равномерно по моделирующему тексту. Отсюда следует вывод о том, что прирост новых однократных словоформ в тексте N_0 пропорционален уже накопленным однократным словоформам. Обозначив буквой q число новых словоформ в испытуемом тексте N_0, можем записать пропорцию и получить окончательно

$$q = m_1 \frac{N_0}{N}. \qquad (12)$$

Если нас интересует не количество новых словоформ, а доля заполненного текста, для которой можно предложить название коэффициент заполнения ζ, то очевидно

$$\zeta = 1 - \frac{q}{N_0} = 1 - \frac{m_1}{N}. \qquad (13)$$

В табл. 4 приводятся вычисленные по этой формуле коэффициенты заполнения для словарей по разным выборкам.

Таблица 4

Выборка	N	m_1	ζ
III	50000	4774	0.905
IV	100000	6366	0.936
V	150000	7734	0.948
VI (основная)	200894	9581	0.952

Аналогичное рассуждение можно привести и для слов. Коэффициент заполнения для словаря слов (основная выборка) оказывается равным 0.99, т. е. значительно выше, чем для словаря словоформ. Это объясняется тем обстоятельством, что новые словоформы чаще всего являются редкими словоформами слов, имеющих и частые словоформы.

Вывод математического выражения для q и ζ был сделан для моделирующего текста. Здесь мы сталкиваемся с необходимостью проверить, как выполняются полученные соотношения на реальных текстах. С этой целью были взяты две выборки по 1000 словоформ (именно такие по длине отрывки составили основную выборку). В табл. 5 приводятся вычисленные по формуле (12) и наблюденные значения числа новых словоформ q в этих текстах относительно частотных словарей по III, IV, V и основной выборкам.

Таблица 5

№ текста	Число новых словоформ q относительно выборок			
	III	IV	V	VI
1	80	58	46	39
2	93	55	44	36
Теоретическое значение	95	64	52	48

В табл. 6 приводятся значения коэффициента заполнения ζ для 1-го и 2-го пробных текстов и теоретические значения.

Для словаря слов (основная выборка) теоретическое значение числа новых слов в новом тексте длиной 1000 словоформ по формуле (12) $q=10$.

В первом пробном тексте новых слов 10, во втором 5.

Приведенные данные показывают удовлетворительное совпадение теоретических и экспериментальных значений. Новых словоформ оказалось несколько меньше, чем было теоретически вычислено. Это объясняется тем, что словарь составлен по многим авторам и темам, поэтому он богаче словаря одного автора, пишущего на одну тему.

Ясно, что текст, составленный из отрывков разных авторов и по разным темам, в среднем лексически разнообразнее текста, принадлежащего одному автору. Чтобы определить количественно разницу, были взяты еще 4 текста по 1000 словоформ (3-й — из произведений 7 авторов, 4-й—10 авторов, 5-й — 20 авторов, 6-й — из статьи одного автора, правда, эта статья дает обзор литературы и потому тематически достаточно богата). В табл. 7 приводятся числа q для этих текстов относительно III—VI выборок.

Таблица 6

№ текста	ζ			
	III	IV	V	VI
1	0.861	0.894	0.917	0.937
2	0.897	0.939	0.951	0.959
Теоретическое значение	0.905	0.936	0.948	0.952

Данные табл. 7 свидетельствуют о значительном увеличении словарного богатства текстов, составленных из работ разных авторов. Ограниченность собственного активного запаса слов автора и необходимость придерживаться одной темы существенно уменьшают лексическое богатство текстов. Если при составлении частотного словаря ставится задача охватить больше лексики, то необходимо предельно увеличивать число авторов. Это очевидное качественное положение иллюстрируется табл. 7 с количественной

Таблица 7

№ текста	Число авторов	Число различных словоформ	Число новых словоформ q относительно выборок				Число новых слов относительно VI выборки
			III	IV	V	VI	
3	7	612	119	87	68	50	9
4	10	796	150	108	83	73	14
5	20	740	191	143	117	100	25
6	1	587	135	102	80	71	14

стороны. Однако при этом нужно помнить, что одновременно будут накапливаться случайные слова, нехарактерные для той литературы, по которой составляется частотный словарь.

Описанные опыты позволяют количественно описать еще одну особенность реальных текстов, отличающую их от моделирующего текста, построенного в соответствии с вероятностной моделью. В выводе коэффициента заполнения и числа новых словоформ (слов) q утверждалось, что новые лексические единицы все однократные. Для моделирующего текста это утверждение справедливо. Проверим его на пробных выборках.

В 6 текстах общей длиной 6000 словоформ оказалось в общей сложности 369 новых словоформ относительно основной выборки,

из них более одного раза встретились 23 словоформы, т. е. хотя число новых неоднократных словоформ и не нуль, как в моделирующем тексте, практически мы можем считать его малым по сравнению с общим количеством новых словоформ. Общее число новых слов на 6000 равно 77, из них 7 встретилось более одного раза.

Выведенная формула (12) для числа новых единиц текста может быть использована для получения еще одной интересной характеристики частотных словарей и текста. По формуле (12) среднее число новых словоформ в тексте N_0 равно

$$q_{c\phi} = m_{1c\phi}\frac{N_0}{N}, \tag{14}$$

а среднее число новых слов в этом же тексте

$$q_c = m_{1c}\frac{N_0}{N}. \tag{15}$$

Здесь $m_{1c\phi}$ — число словоформ в исходной для составления частотного словаря выборке N, которые встретились ровно по одному разу, m_{1c} — число однократных слов в выборке N.

Отношение η чисел $q_{c\phi}$ и q_c показывает, как увеличивается покрываемость новых текстов при переходе от словаря словоформ к словарю слов. Разделив равенство (14) на (15), найдем, что η равно отношению числа однократных словоформ к числу однократных слов в выборке, по которой составлен частотный словарь,

$$\eta = \frac{m_{1c\phi}}{m_{1c}}. \tag{16}$$

Для нашей основной выборки $m_{1c\phi} = 9581$, $m_{1c} = 2009$. Поэтому $\eta = \frac{9581}{2009} = 4.79$.

Наши 6 пробных текстов позволяют сравнить теоретическое значение $\eta = 4.79$ с экспериментальным, которое вычисляется по данным табл. 5 и 7.

Таблица 8

№ текста	Число новых словоформ $q_{c\phi}$	Число новых слов q_c	$\eta = \frac{q_{c\phi}}{q_c}$
1	39	10	3.9
2	36	5	7.2
3	50	9	5.56
4	73	14	5.21
5	100	25	4
6	71	14	5.07
1—6	369	77	4.79

Экспериментальные значения η разбросаны вокруг среднего теоретического значения η, что подтверждает правильность выведенной формулы. Экспериментальное значение η для текста, составленного из всех шести пробных выборок, точно совпало с теоретическим. Однако столь идеальное совпадение является случайным. При проверке статистических равенств на случайных значениях мы всегда должны учитывать возможность некоторого разброса экспериментальных значений.

Значение η является важным при решении вопроса о том, какой словарь взять за основу в практических приложениях (например, при обучении иностранному языку или в машинном переводе): словарь словоформ или словарь слов. Преимущество словаря слов для русского языка заключается в значительном увеличении эффективности, заполняемости и существенном сокращении его объема (в частотном словаре по основной выборке 21 468 словоформ и только 6826 слов). Преимущество словаря словоформ в том, что он дает статистические веса различных грамматических форм, полнее отражает текстовой материал. Он описывает не только лексико-статистическую, но и статистико-морфологическую структуру текста и потому может быть использован для исследования большего круга вопросов, чем словарь слов. Кроме того, частотный словарь словоформ всегда можно преобразовать в словарь слов, но не наоборот.

§ 5. Количественный метод сравнения частотных словарей

Сравнение частотных словарей может быть необходимо при сравнительном анализе лексики двух текстов или двух групп текстов, для оценки достоверности частотного словаря и зависимости достоверности от объема выборки, для оценки качества частотного словаря, полученного каким-либо приближенным способом, и т. п.

Во всех этих случаях качественное сравнение словарей необходимо, но оно не может считаться достаточным, пока нет количественной меры, измеряющей качественную близость. Желательно уметь оценивать близость, родство словарей одним числом.

В последние годы, несмотря на возражения со стороны некоторых математиков, в статистической лингвистике [7] с этой целью стали применять методы ранговой корреляции.

Изложим кратко сущность ранговой корреляции.[8] Пусть создается некоторый комплекс условий, и при этом происходит n со-

[7] Р. М. Фрумкина. Статистические методы изучения лексики. М., 1964.

[8] G. K. Zipf. Human behaviour and the principle of least effort. Cambridge, 1949; Б. Л. Ван дер Варден. Математическая статистика. М., 1960.

бытий. Пусть события могут характеризоваться двумя признаками и по убыванию степени выраженности признаков могут быть расставлены в два ряда и перенумерованы. Ставится задача определить связь двух изучаемых признаков. Обозначим номер события в первом ряду i и во втором — j_i. Номера событий называются их рангами. В качестве меры связи признаков можно брать любую из двух величин:
либо сумму квадратов разностей рангов в двух рядах

$$\sum_{i=1}^{n} (i - j_i)^2,$$

либо сумму абсолютных значений разностей рангов

$$\sum_{i=1}^{n} |i - j_i|.$$

Математики обычно останавливаются из несущественных для нашей задачи соображений на сумме квадратов. Нам же будет удобнее выбрать сумму абсолютных значений разностей с целью облегчить вычисления. Обозначим эту сумму буквой $\hat{S}$:

$$\hat{S} = \sum_{i=1}^{n} |i - j_i|.$$

Если связь абсолютная и прямая, то номера событий в обоих рядах совпадают, $i = j_i$ и $\hat{S} = 0$. Если связь статистическая, то чем она слабее, тем больше сумма $\hat{S}$, и можно рассчитывать, что $\hat{S}$ будет мерой связи.

Найдем значение суммы в крайнем случае, т. е. когда всякая связь отсутствует. Мы будем проводить вывод точно по аналогии с выводом для того случая, когда мера связи измеряется суммой квадратов [9].

Если между двумя рядами нет никакой связи, то i-событие во втором ряду может иметь с равной вероятностью $\frac{1}{n}$ любой из n номеров; если i-событие первого ряда имеет во втором ряду номер 1, то разность номеров равна $(i - 1)$; если номер во втором ряду 2, то разность равна $(i - 2)$ и т. д. Записываем табличку:

j_i	1	2	. . .	i	$i+1$	. . .	n
$\lvert i - j_i \rvert$	$i-1$	$i-2$	. . .	0	1	. . .	$n-i$

[9] А. А. Чупров. Основные проблемы теории корреляции. М., 1960.

Все разности равновероятны, поэтому среднее значение (т. е. математическое ожидание разности) мы найдем, сложив все возможные разности и разделив сумму на их общее число n. Замечаем, что сумма разностей распадается на две арифметические прогрессии. Воспользовавшись известной формулой для арифметической прогрессии, найдем s_i — среднее отклонение $|i - j_i|$,

$$s_i = \frac{(i-1)\,i + (n-i)\,(n-i+1)}{2n}. \tag{17}$$

Чтобы найти среднее значение разности $\hat{S} = \sum_{i=1}^{n} |i - j_i|$ по всем номерам i, нужно просуммировать s_i по всем номерам. При этом снова воспользуемся формулой для арифметической прогрессии и формулой для суммы квадратов натурального ряда чисел

$$\sum_{i=1}^{n} i^2 = \frac{n\,(n+1)\,(2n+1)}{6}.$$

Окончательно получаем следующее выражение для среднего значения (точнее, для математического ожидания) $\hat{S}$ в случае отсутствия связи между признаками и рядами:

$$\hat{S} = \hat{S}_0 = \frac{n^2 - 1}{3}.$$

Если связь абсолютная и прямая ($i = j_i$), то $\hat{S} = 0$. Если связь отсутствует, то $\hat{S} = \hat{S}_0 = \frac{n^2 - 1}{3}$. Если связь промежуточная, или, как говорят, статистическая, то $\hat{S}$ тем ближе к 0, чем теснее связь, и тем ближе к $\hat{S}_0$, чем слабее связь. Удобно вместо $\hat{S}$ ввести эквивалентную величину α по формуле

$$\alpha = 1 - \frac{\hat{S}}{\hat{S}_0} = 1 - \frac{3\sum_{i=1}^{n} |i - j_i|}{n^2 - 1}. \tag{18}$$

Если связь абсолютная, то $\alpha = 1$.
Если связь отсутствует, то $\alpha = 0$.

В случае статистической связи α может быть ее мерой.

Описанную вероятностную схему применим к оценке близости двух частотных словарей. Здесь набор событий — набор словарных единиц, первый ряд — один частотный словарь, второй ряд — второй словарь, признак, по убыванию степени которого нумеруются словарные единицы, есть частота, n — число словарных единиц. Коэффициент α можно назвать коэффициентом близости словарей.

Проиллюстрируем схему ранговой корреляции на буквенном уровне. По первой выборке, в которую вошло 15 620 словоформ (100 000 букв), были подсчитаны частоты букв русского алфавита. В табл. 9 приводятся результаты подсчета, могущие представлять и самостоятельный интерес. Средняя длина словоформы в русских текстах по электронике равна 6.4 буквы (не считая пробела). Вероятность пробела 0.135.

Таблица 9

Ранг в электронике, i	Буква	Частота на 100 000 букв в электронике	Относительная частота в литературном тексте	Ранг в литературном тексте, i_1
1	о	11376	0.110	1
2	е	8907	0.087	2
3	и	7852	0.075	4
4	т	7338	0.065	5
5	а	7020	0.075	3
6	н	6889	0.065	6
7	р	5498	0.048	8
8	с	5116	0.055	7
9	л	4227	0.042	10
10	в	4104	0.046	9
11	к	3358	0.034	11
12	п	3072	0.028	14
13	м	3047	0.031	12
14	д	2641	0.030	13
15	я	2302	0.022	16
16	ы	1919	0.019	17
17	у	1915	0.025	15
18	ч	1752	0.015	22
19	з	1563	0.018	18
20	ь, ъ	1364	0.017	19
21	г	1256	0.016	21
22	б	1210	0.017	20
23	х	1200	0.011	24
24	й	1032	0.012	23
25	э	789	0.003	30
26	ж	753	0.009	25
27	ю	692	0.007	26
28	ц	477	0.005	28
29	щ	460	0.004	29
30	ф	449	0.002	31
31	ш	422	0.007	27

Этот материал интересует нас здесь как иллюстрация. В табл. 9 приводятся также частоты и ранги букв в русской литературе.[10] Мы можем оценить связь распределения букв в текстах по электронике и в русской литературе нашим коэффициентом α. Подсчет дает $\hat{S}=38$, $n=31$, $\alpha=0.881$.

[10] А. М. Яглом, И. М. Яглом. Вероятность и информация. М., 1960; А. А. Харкевич. *Очерки общей теории связи*. М., 1955.

Чтобы определить, как влияет объем выборки на рост достоверности, по мере роста выборки пять раз подсчитывалось распределение букв и вычислялся коэффициент связи с первой выборкой, которая предполагалась достаточно большой, чтобы частотный список букв по ней считать практически совпадающим с вероятностным списком. Из табл. 10 можно видеть, как нарастает связь списков.

Таблица 10

N	α	N	α
14912	0.625	62423	0.894
31399	*0.823*	*78469*	*0.919*
44988	0.838		

При подсчете первая выборка была предварительно расписана по алфавиту. Поэтому α для связного текста растет быстрее, чем это видно из табл. 10.

Приведенный пример показывает, что коэффициент близости словарей может быть мерой достоверности частотного словаря. Было бы желательно найти математическое ожидание α для частотного и вероятностного словарей хотя бы при условии выполнения закона Ципфа. Однако решение этой сложной задачи выходит за рамки настоящей статьи.

Коэффициент α можно использовать также для количественного измерения связи различных уровней иерархической структуры текста. Сделаем это для буквенного и лексического уровней.

Исходим из вероятностной модели. Вероятность словоформы можно определить как произведение вероятностей образующих ее букв в соответствии с теоремой о вероятности произведения независимых событий. Были взяты первые 100 словоформ, сосчитаны по данным табл. 9 их вероятности, и словоформы расставлены по убыванию вероятностей. После этого вычислен коэффициент α. В опыте оказалась сумма разностей рангов $\hat{S}=2309$, $n=100$, $\alpha=0.31$, т. е. связь довольно велика, хотя и ближе к случайной, чем к абсолютной. Если учитывать не только распределение первого порядка (частоты букв), но и распределение второго порядка (частоты пар букв), то связь уровней будет еще теснее. Если иметь распределение достаточно высокого порядка для букв, то частотный словарь словоформ из него получится достоверно. Вернемся к оценке близости частотных словарей. Применение коэффициента α имеет некоторые особенности.

1. В двух сравниваемых словарях может быть несколько различная лексика: часть словарных единиц одного словаря не входит в другой (имеет в нем частоту 0). Перед вычислением α можно приписать снизу недостающие словарные единицы в словарях

в случайном порядке относительно их расположения в другом словаре, чтобы не внести добавочную связь. Такая процедура естественна, так как, чем больше таких слов, тем будет меньше a, но это соответствует нашему качественному представлению о родстве словарей, требующему прежде всего одинаковости словарных единиц.

2. Можно проводить сравнение не всего словаря, а его части, например только знаменательных слов или только существительных и т. д.

В этом случае нужно перенумеровать только сравниваемые части. Сравнение части словарей, например достоверной части либо условного количества словарных единиц, можно рекомендовать также для того, чтобы избежать слишком громоздких вычислений, но следует помнить, что коэффициент близости словарей будет зависеть от того, сколько словарных единиц мы возьмем *для сравнения*.

Мы избрали второй путь (сравнение условного количества словарных единиц), так как понятие «достоверная часть словаря» определяется прежде всего требуемой точностью в конкретном приложении.

Оказалось, что уже первые 100 словарных единиц, хотя и содержат много служебных слов, общих в самых различных текстах, достаточно характерны и можно ограничиться ими.

Для этого случая нужно преобразовать основную расчетную формулу.

Как было выведено (формула 17), среднее абсолютное значение разности $|i - j_i|$ равно

$$s_i = \frac{i(i-1) + (n-i+1)(n-1)}{2n} = \frac{n}{2} - i + \frac{i^2}{n} - \frac{i}{n} + \frac{1}{2}.$$

Считаем, что объем словаря много больше 100 единиц: $n \gg 100$. Если мы ограничиваемся первыми 100 единицами, то $i \leqslant 100 \ll n$, и в предыдущей формуле можно для этих рангов пренебречь всеми слагаемыми, кроме первого

$$s_i = \frac{n}{2}.$$

Сумма отклонений для 100 единиц равна

$$\hat{S}_0 = 50n.$$

Окончательно можем написать выражение для искомого коэффициента родства словарей

$$a = 1 - \frac{\sum_{i=1}^{100} |i - j_i|}{50n}. \qquad (19)$$

Таблица 11

i	Выборка II	j_i	Выборка III	j_i	Выборка IX	j_i
1	в	1	в	1	в	1
2	и	2	и	2	и	2
3	при	3	на	4	при	3
4	как	12	при	3	на	4
5	на	4	для	5	с	6
6	с	6	с	6	для	5
7	по	8	по	8	от	7
8	рис.	10	что	9	напряжения	19
9	от	7	как	12	рис.	10
10	что	9	от	7	из	11
11	из	11	рис.	10	поля	25
12	к	14	к	14	что	9
13	можно	17	из	11	по	8
14	случае	22	не	13	до	23
15	не	13	а	15	так	20
16	а	15	электронов	21	можно	17
17	бария	484	ток	18	может	26
18	системы	53	тока	16	не	13
19	выход	422	случае	22	как	12
20	напряжения	19	напряжения	19	или	24
21	электронов	21	можно	17	между	27
22	волны	169	так	20	времени	30
23	длины	272	лампы	50	а	15
24	зависимость	71	или	24	к	14
25	где	38	до	23	если	28
26	разряда	40	то	34	тока	16
27	тока	16	же	35	через	29
28	длина	461	напряжение	32	разряда	40
29	поверхности	49	схемы	51	этом	31
30	же	35	будет	45	области	39
31	замедляющей	696	этом	31	время	37
32	работы	62	быть	43	это	33
33	времени	30	между	27	катода	36
34	или	24	может	26	чем	41
35	имеет	57	где	38	напряжение	32
36	образом	59	если	28	мм	56
37	чем	41	это	33	где	38
38	более	47	катода	36	случае	22
39	заряда	125	через	29	ток	18
40	между	27	образом	59	ионов	65
41	это	33	является	42	то	34
42	быть	43	более	47	системы	53
43	видно	124	триода	91	же	35
44	концентрации	148	также	44	эмиссии	46
45	может	26	частоты	66	величины	54
46	порядка	87	однако	52	линзы	96
47	характеристики	48	ее	75	уже	156
48	этом	31	поверхности	49	характеристики	48
49	двух	106	сопротивление	74	является	42
50	следует	84	величина	55	также	44
51	тем	94	имеет	57	мишени	97

Таблица 11 (*продолжение*)

i	Выборка II	j_i	Выборка III	j_i	Выборка IX	j_i
52	типа	104	ею	58	поверхности	49
53	то	34	времени	30	во	63
54	время	37	время	37	более	47
55	высокой	247	ламп	123	мксек	103
56	необходимо	116	сопротивления	93	электронов	21
57	рекомбинации	467	области	39	этого	70
58	так	20	тем	94	быть	43
59	величина	55	эмиссии	46	поле	73
60	вольфраматов	1930	характеристики	48	за	67
61	значения	78	цепи	60	оси	113
62	кривые	253	работы	62	в.	68
63	кривых	302	чем	41	величина	55
64	однако	52	волны	169	однако	52
65	отсоса	1982	порядка	87	т. е.	61
66	систем	504	т. е.	61	их	64
67	ток	18	эмиттера	100	температуры	95
68	частоты	66	их	64	электродов	115
69	*является*	42	*типа*	*104*	*электронных*	*230*
70	величину	160	фиг.	294	будет	45
71	выше	98	о	101	значения	78
72	если	28	поэтому	82	цепи	60
73	их	64	следует	84	имеет	57
74	ловушек	2114	величины	54	импульсов	83
75	но	92	выше	98	схема	69
76	области	39	сетки	182	его	58
77	определения	384	схема	69	зависимость	71
78	определяется	157	зависимость	71	значение	81
79	оси	113	больше	72	работы	62
80	температура	287	генератора	85	после	80
81	будет	45	поля	25	больше	72
82	возникновения	578	только	77	трубки	173
83	длине	1022	импульса	76	энергии	88
84	изменении	329	таким	79	импульса	76
85	после	80	мгц	303	все	89
86	характеристика	353	но	92	таким	79
87	эв.	586	того	111	только	77
88	эмиссии	46	бария	484	когда	86
89	барьера	633	двух	106	меньше	99
90	до	23	за	67	пучка	122
91	за	67	в.	68	потенциал	90
92	когда	86	колебаний	201	работе	121
93	которых	126	необходимо	116	генератора	85
94	малых	311	во	63	было	105
95	относительно	188	длины	272	изменения	119
96	пленки	560	когда	86	поэтому	82
97	приведены	366	несколько	112	но	92
98	результаты	266	после	80	образом	59
99	случая	409	потенциал	90	потенциала	131
100	см.	110	см.	110	следует	84

Здесь n означает общее количество различных словарных единиц в двух словарях. Если нам известен объем лишь одного словаря n, то

$$\alpha \approx 1 - \frac{\sum_{i=1}^{100} |i - j_i|}{50n}. \qquad (20)$$

В табл. 11 *приводятся первые* 100 *словоформ по второй* (9418 словоформ) и третьей (50 000 словоформ) выборкам. Для проверки устойчивости α при повторении опыта были выделены первые 100 словоформ для еще одной — IX выборки в 50 000 словоформ (дополняющей III выборку до IV). В табл. 11 приводятся также эти словоформы.

Таблица 12

Выборка	N	$\hat{S}$	α
II	9418	16157	0.985
III	50000	2427	0.998
IX	50000	1498	0.999

В табл. 12 даются значения $\hat{S}$ и α при сравнении словарей перечисленных выборок со словарем по основной выборке. Можно видеть, что с ростом выборки возрастает близость словарей и что для III и IX выборок α оказался почти одинаков.

Величина n=21468 (*объем словаря словоформ по основной* выборке).

Интересно определить степень близости частотного словаря современных русских текстов по электронике и частотного словаря современного русского литературного языка. Материал для этого нам дает наш словарь слов (см. табл. 11) и словарь Э. А. Штейнфельдт. [11] В табл. 13 напечатаны первые 100 слов частотного словаря литературного языка с указанием их рангов i, а также рангов j_i этих слов в частотном словаре электроники. Из этих 100 слов в наших текстах по электронике (в основной выборке) ни разу не встретилось 14 слов. Считаем, что ранги их в нашем словаре превосходят 6826 (ранг последнего слова). Сумма абсолютных разностей рангов превышает 15 6158. По формуле (20) находим $\alpha \approx 0.541$.

При составлении табл. 13 было учтено некоторое расхождение в определении слова у нас и Э. А. Штейнфельдт.

Мы не рассматриваем коэффициент ранговой корреляции как критерий для проверки статистической гипотезы о близости выборок или родстве частотных словарей. Дело в том, что этот коэффициент *предложен не исходя из его распределения, а на основе* интуитивных соображений: чем «родственнее» словари, тем меньше

[11] Э. А. Штейнфельдт. Частотный словарь современного русского языка. Таллин, 1963.

Таблица 13

i	Слово	j_i	i	Слово	j_i	i	Слово	j_i
1	и	2	34	наш	50	68	если	55
2	он (она, оно, они)	11	35	только	155	69	два	77
			36	еще	524	70	мой	—
3	в (во)	1	37	от	12	71	жизнь	1060
4	на	4	38	такой	36	72	до	37
5	не	19	39	мочь	30	73	где	70
6	*я*	—	40	*говорить*	723	74	*каждый*	217
7	что	9	41	бы (б)	375	75	хотеть	—
8	с (со)	5	42	для	6	76	здесь	292
9	этот (это)	8	43	уже	246	77	надо	1321
10	быть (есть)	13	44	знать	1084	78	теперь	800
11	а	27	45	да	—	79	дом	—
12	весь (все)	49	46	какой	1287	80	пойти	2697
13	тот (то)	16	47	*когда*	167	81	*рай*	386
14	как	18	48	другой	104	82	товарищ	—
15	мы	195	49	первый	121	83	ни	31
16	к (ко)	21	50	ребята	—	84	или	43
17	у	308	51	день	3069	85	ведь	4830
18	ты	—	52	год	1572	86	советский	—
19	за	135	53	кто	4817	87	работать	312
20	*но*	108	54	*себя*	277	88	город	—
21	вы	—	55	дело	883	89	там	2122
22	по	14	56	нет	886	90	слово	878
23	из (изо)	17	57	рука	4814	91	глаз	2752
24	о (об, обо)	160	58	очень	269	92	потом	4481
25	свой	251	59	большой	48	93	видеть	152
26	же (ж)	64	60	ну	—	94	под	239
27	*сказать*	769	61	*новый*	501	95	даже	530
28	так	35	62	стать	862	96	думать	602
29	один	62	63	школа	6775	97	хорошо	827
30	вот	—	64	работа	41	98	можно	33
31	сам (самый)	256	65	сейчас	3545	99	тут	—
32	который	26	66	время	22	100	тысяча	1238
33	человек (люди)	1840	67	идти (шел)	977			

различаются в них ранги (номера) одних и тех же лексических единиц. Поэтому мы хорошо отдаем себе отчет в том, что применяемые в статье понятия «связь», «статистическая связь» и «отсутствие связи» лишены строгого математического смысла.

Все же мы надеемся, что осторожное применение ранговой корреляции [12] в качестве эмпирической оценки результатов современных лингво-статистических исследований окажется вполне допустимым.

[12] Может быть, основной недостаток этого подхода состоит в отсутствии подходящего веса в сумме $\hat{S}$.

§ 6. Анализ статистических закономерностей, управляющих лексикой русских текстов по электронике

Основу для изучения статистики лексики дает распределение выборки — таблица встретившихся в выборке частот с указанием числа словарных единиц, имеющих данную частоту. Распределение основной выборки для словоформ представлено в табл. 14, для слов — в табл. 15. Распределение III, IV и V выборок для словоформ представлено в табл. 16—18.

Выборка	N	n	m_1
III	50000	9464	4774
IV	100000	14062	6366
V	150000	17263	7734
VI	200894	21468	9581

Ниже представлены основные данные по выборкам. Обозначения следующие: объем выборки N, число различных словоформ n, число однократных словоформ m_1.

В табл. 15—18 указываются ранг i, частота F_i и количество словарных единиц m, имеющих частоту F_i. В таблицах для основной выборки указывается также общее число единиц текста, покрытое словарными единицами с частотой F_i, и накопленная абсолютная частота $\sum_1^i F_i m$. Структура всех распределений очень сходна. Большие частоты все однократны, убывают большими скачками, промежуточные частоты не встречаются, ранги определяются однозначно и растут как натуральный ряд. Этой части таблиц соответствует наиболее достоверная часть словарей.

В средней части словаря пропущенных частот все меньше, в графе m среди единиц начинают появляться другие числа, постепенно вытесняющие единицы.

В нижней части словаря и таблиц пропущенных частот нет, m растет большими скачками. Это наиболее недостоверная часть словаря. Ранг в средней и нижней части словаря определяется неоднозначно: все m словарных единиц имеют одинаковую частоту F_i и все «делят» между собой ранги общим числом m от минимального $i_{\min}$ до максимального $i_{\max}$.

В § 3 мы нашли максимальную ошибку в определении ранга i-той словарной единицы. Здесь мы можем указать минимальную ошибку в определении ранга. Абсолютная минимальная ошибка, очевидно, равна $\frac{m}{2}$, а относительная $\frac{m}{2i}$.

Условно однозначность рангов мы получаем, расположив все m словарных единиц в алфавитном порядке.

Чтобы получить наглядное представление о распределении выборки, обычно на графике в логарифмическом масштабе по обеим

осям откладывают зависимость частот от рангов. Закон Ципфа (4) изображается на таком графике прямой линией. Закон Мандельброта

$$p_i = \frac{K}{(B+i)^{\gamma}}$$

изображается также прямой линией, кроме начального участка. Распределение VI и IV выборок представлено на рис. 3 и 4.

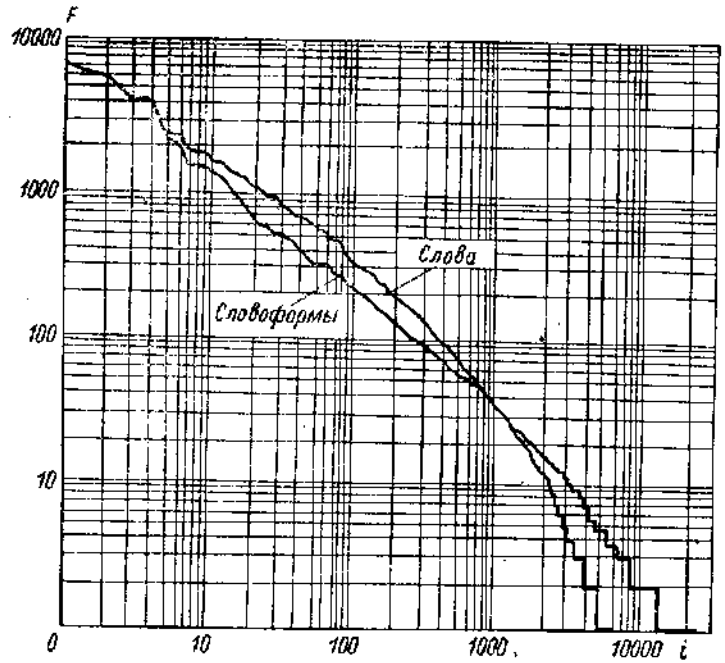


Рис. 3.

О логарифмическом масштабе следует сделать еще несколько замечаний. В этом масштабе диапазон изменения переменной очень сильно сжимается. Например, если частоты меняются от 10 до 10 000, т. е. в 1000 раз, то логарифм частоты меняется от 1 до 4, т. е. только в 4 раза. Логарифмический масштаб на оси рангов предоставляет много места малым рангам и сильно сжимает большие: первые 100 словарных единиц занимают столько же места, сколько и 9900 следующих единиц, что не для всех прикладных задач желательно.

Если мы будем, например, определять константы K, B и γ закона Мандельброта методом наименьших квадратов, разделив для этого ось рангов на равные участки в логарифмическом масштабе, то полученные таким образом параметры и построенная

по ним прямая Мандельброта будут на графике довольно хорошо совпадать с распределением выборки. Для IV выборки мы нашли $K=0.52$; $B=1.90$; $\gamma=0.84$. На рис. 4 прямая соответствует этим величинам. Но определенные таким способом константы ни в коем случае нельзя приписывать всем рангам: главным образом они характеризуют распределение частых словоформ.

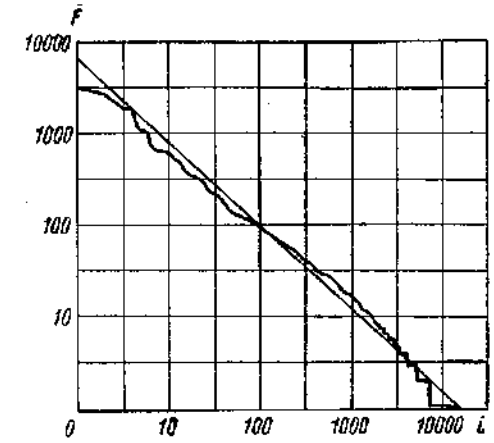


Рис. 4.

Следует обратить внимание на величину γ, оказавшуюся меньше единицы для русского языка электроники.

Сказанное заставляет сделать вывод о необходимости изучать распределение не графически, а аналитически — по таблицам.

В последнее время в математической лингвистике была поставлена и решена задача выявления функциональной зависимости параметров закона Ципфа и Мандельброта от ранга.[13]

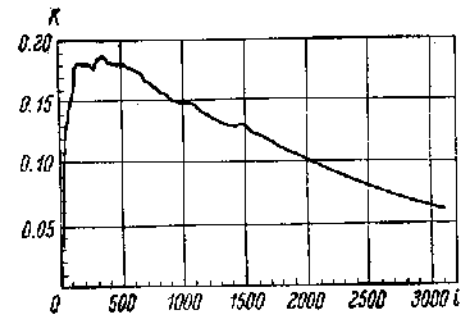


Рис. 5.

Мы воспользуемся двумя результатами этих работ, которые сформулируем в обозначениях, принятых в данной работе.

1. Параметр закона Ципфа K можно приближенно определить по частоте F_i, минимальному рангу $i_{\min}$ и длине выборки N по формуле

$$K = \frac{i_{\min} F_i}{N}.$$

2. Параметр закона Мандельброта γ можно приближенно определить по частоте F_i, числу m и максимальному рангу $i_{\max}$ по формуле

$$\gamma = \frac{i_{\max}}{F_i m}.$$

[13] Р. М. Фрумкина. К вопросу о так называемом «законе Ципфа». ВЯ, X, 2, 1961; Д. М. Сегал. Некоторые уточнения вероятностей модели Ципфа. «Машинный перевод и прикладная лингвистика», № 5, 1961; В. М. Калинин. О статистике литературного текста. ВЯ, XIII, 1, 1964.

В табл. 15 и 16 приведены результаты вычислений по этим формулам для основной выборки, а на рис. 5, 6, 7 и 8 эти вычисления представлены графически.

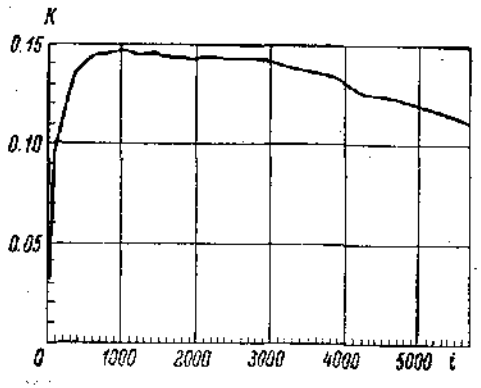


Рис. 6.

По графикам можно сделать выводы о выполнимости закона Ципфа для текстов по электронике. Для слов закон Ципфа выполняется неудовлетворительно. Для словоформ закон Ципфа выполняется удовлетворительно на средних рангах. Для русских текстов по электронике на рангах от $i=300$ до $i=4200$ среднее значение постоянной $K=0.145$ со среднеквадратической ошибкой 4.2%.

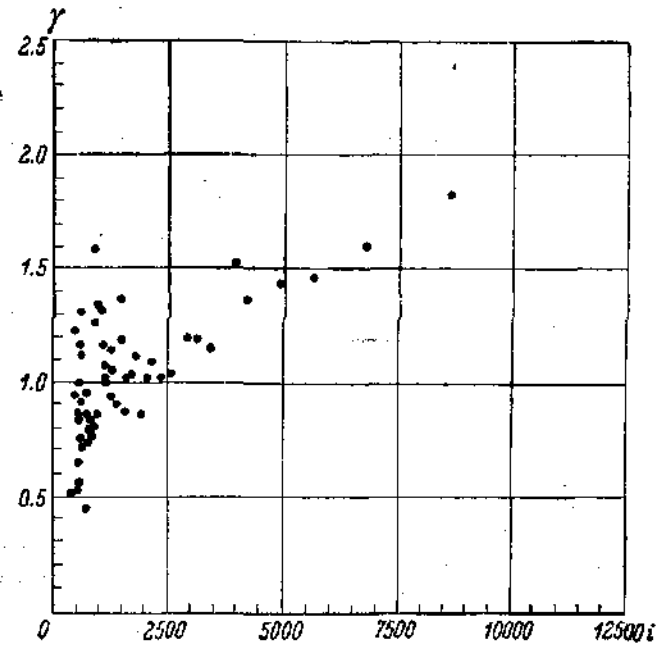


Рис. 7.

Таблица 14

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
6418	1	1	6418	6418	0.032	
5742	1	2	5742	12160	0.057	
3924	1	3	3924	16084	0.059	
3914	1	4	3914	19998	0.078	
2218	1	5	2218	22216	0.055	
2060	1	6	2060	24276	0.061	
1460	1	7	1460	25736	0.051	
1368	1	8	1368	27104	0.054	
1324	1	9	1324	28428	0.059	
1308	1	10	1308	29736	0.065	
1218	1	11	1218	30954	0.067	
1166	1	12	1166	32120	0.070	
1066	1	13	1066	33186	0.069	
1032	1	14	1032	34218	0.072	
914	1	15	914	35122	0.068	
784	1	16	784	35916	0.062	
738	1	17	738	36654	0.062	
720	1	18	720	37374	0.064	
716	2	19—20	1432	38806	0.068	
698	1	21	698	39504	0.073	
692	1	22	692	40196	0.076	
680	1	23	680	40876	0.078	
634	1	24	634	41510	0.076	
612	1	25	612	42122	0.076	
578	1	26	578	42700	0.075	
552	1	27	552	43252	0.074	
506	1	28	506	43758	0.071	
500	1	29	500	44258	0.072	
498	2	30—31	996	45254	0.074	
466	1	32	466	45720	0.074	
464	1	33	464	46184	0.076	
460	1	34	460	46644	0.078	
446	1	35	446	47090	0.078	
442	1	36	442	47532	0.079	
432	3	37—39	1296	48828	0.080	
390	1	40	390	49218	0.078	
386	1	41	386	49604	0.079	
362	1	42	362	49966	0.076	
360	1	43	360	50326	0.077	
352	1	44	352	50678	0.077	
348	1	45	348	51026	0.078	
342	1	46	342	51368	0.078	
338	1	47	338	51706	0.079	
332	1	48	332	52038	0.079	
330	1	49	330	52368	0.073	
316	2	50—51	632	53000	0.079	
312	1	52	312	53312	0.081	
306	1	53	306	53618	0.081	
304	1	54	304	53922	0.082	
300	1	55	300	54222	0.082	
290	1	56	290	54512	0.081	
288	1	57	288	54800	0.082	

Таблица 14 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
284	2	58—59	568	55368	0.082	
278	1	60	278	55646	0.083	
272	1	61	272	55918	0.083	
270	1	62	270	56188	0.083	
268	2	63—64	536	56724	0.084	
266	1	65	266	56990	0.086	
264	1	66	264	57254	0.087	
256	1	67	256	57510	0.085	
252	1	68	252	57762	0.085	
250	2	69—70	500	58262	0.086	
246	1	71	246	58508	0.087	
240	1	72	240	58748	0.086	
238	1	73	238	58986	0.086	
236	1	74	236	59222	0.087	
234	3	75—77	702	59924	0.087	
232	2	78—79	464	60388	0.090	
230	1	80	230	60618	0.092	
228	2	81—82	456	61074	0.092	
226	2	83—84	452	61526	0.093	
224	1	85	224	61750	0.095	
220	3	86—88	660	62410	0.094	
216	4	89—92	864	63274	0.096	
214	2	93—94	428	63702	0.099	
212	1	95	212	63914	0.100	
208	3	96—98	624	64538	0.099	
204	1	99	204	64742	0.100	
202	1	100	202	64944	0.101	
200	1	101	200	65144	0.101	
198	1	102	198	65342	0.100	
196	1	103	196	65538	0.100	
194	1	104	194	65732	0.100	
190	3	105—107	570	66302	0.099	
188	3	108—110	564	66866	0.101	
186	1	111	186	67052	0.103	
184	4	112—115	736	67788	0.103	
182	2	116—117	364	68152	0.105	
180	1	118	180	68332	0.106	
174	3	119—121	522	68854	0.103	
172	1	122	172	69026	0.104	
170	1	123	170	69196	0.104	
168	4	124—127	672	69868	0.104	
166	4	128—131	664	70532	0.106	
164	1	132	164	70696	0.108	
163	1	133	163	70859	0.108	
162	2	134—135	664	71183	0.108	
160	4	136—139	640	71823	0.108	
158	1	140	158	71981	0.110	
157	1	141	157	72138	0.110	
156	2	142—143	312	72450	0.110	
152	2	144—145	304	72754	0.109	
150	5	146—150	750	73504	0.109	
149	1	151	149	73653	0.112	

Таблица 14 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
148	1	152	148	73801	0.112	
146	3	153—155	438	74239	0.111	
145	2	156—157	290	74529	0.113	
144	1	158	144	74673	0.113	
143	1	159	143	74816	0.113	
139	1	160	139	74955	0.111	
137	1	161	137	75092	0.110	
136	1	162	136	75228	0.110	
135	2	163—164	270	75498	0.110	
134	1	165	134	75632	0.110	
133	2	166—167	266	75898	0.110	
132	1	168	132	76030	0.110	
131	1	169	131	76161	0.110	
130	1	170	130	76291	0.110	
129	4	171—174	516	76807	0.110	
128	2	175—176	256	77063	0.110	
126	5	177—181	630	77693	0.110	
125	1	182	125	77818	0.113	
124	1	183	124	77942	0.113	
123	2	184—185	246	78188	0.113	
121	2	186—187	242	78430	0.112	
120	2	188—189	240	78670	0.112	
119	8	190—197	952	79622	0.112	
116	3	198—200	348	79970	0.114	
115	1	201	115	80085	0.115	
114	3	202—204	342	80427	0.114	
112	2	205—206	224	80651	0.114	
110	2	207—208	220	80871	0.113	
109	3	209—211	327	81198	0.113	
108	3	212—214	324	81522	0.114	
107	1	215	107	81629	0.114	
106	2	216—217	212	81841	0.114	
105	3	218—220	315	82156	0.114	
104	5	221—225	520	82676	0.114	
103	5	226—230	515	83191	0.116	
102	4	231—234	408	83599	0.117	
101	6	235—240	606	84205	0.118	
100	4	241—244	400	84605	0.120	
99	8	245—252	792	85397	0.121	
98	5	253—257	490	85887	0.123	
97	1	258	97	85984	0.124	
96	5	259—263	480	86464	0.124	
95	4	264—268	380	86844	0.125	
94	3	269—270	282	87126	0.126	
93	3	271—273	279	87405	0.125	
92	3	274—276	276	87681	0.125	
91	1	277	91	87772	0.125	
90	5	278—282	450	88222	0.124	
89	7	283—289	623	88845	0.125	
88	5	290—294	440	89285	0.127	
87	6	295—300	522	89807	0.128	
86	6	301—306	516	90323	0.129	

Таблица 14 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
85	3	307—309	255	90578	0.130	
84	4	310—313	336	90914	0.129	
83	6	314—319	498	91412	0.130	
82	6	320—325	492	91904	0.130	
81	6	326—331	486	92390	0.131	
80	4	332—335	320	92710	0.132	
79	5	336—340	395	93105	0.132	
78	7	341—347	546	93651	0.132	
77	4	348—351	308	93959	0.133	
76	2	352—353	152	94111	0.133	
75	8	354—361	600	94711	0.132	
74	5	362—366	370	95081	0.133	
73	10	367—376	730	95811	0.133	
72	6	377—382	432	96243	0.135	
71	2	383—384	142	96385	0.135	
70	7	385—391	490	96875	0.134	
69	4	392—395	276	97151	0.134	
68	4	396—399	272	97423	0.134	
67	5	400—404	335	97758	0.133	
66	6	405—410	396	98154	0.133	
65	11	411—421	715	98869	0.133	
64	7	422—428	448	99317	0.134	
63	4	429—432	252	95569	0.134	
62	4	435—436	248	99817	0.133	
61	13	437—449	793	100610	0.133	
60	11	450—460	660	101270	0.134	
59	11	461—471	649	101919	0.135	
58	12	472—483	696	102615	0.136	
57	7	484—490	399	103014	0.137	1.23
56	16	491—506	896	103100	0.137	0.51
55	17	507—523	935	104845	0.139	0.56
54	19	524—542	1026	105871	0.141	0.53
53	12	543—554	636	106507	0.143	0.87
52	12	555—566	624	107131	0.143	0.91
51	11	567—577	561	107692	0.144	1.03
50	9	578—586	450	108142	0.144	1.30
49	13	587—599	637	108779	0.143	0.94
48	16	600—615	768	109547	0.143	0.67
47	16	616—631	752	109299	0.144	0.84
46	12	632—643	552	109851	0.144	1.16
45	13	644—656	585	111436	0.144	1.12
44	20	657—676	880	112316	0.144	0.77
43	17	677—693	731	113047	0.145	0.95
42	24	694—717	1008	114055	0.145	0.72
41	15	718—732	615	114670	0.146	0.45
40	20	733—752	800	115470	0.146	0.94
39	23	753—775	897	116370	0.146	0.84
38	13	776—788	494	116861	0.147	1.66
37	28	789—816	1036	117897	0.145	0.79
36	30	817—846	1080	118977	0.146	0.78
35	30	847—876	1050	120027	0.147	0.83
34	20	877—896	680	120707	0.148	1.32

Таблица 14 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
33	22	897—918	726	121433	0.147	1.26
32	37	919—955	1184	122617	0.146	0.81
31	37	956—992	1147	123764	0.147	0.86
30	26	993—1018	780	124544	0.148	1.31
29	38	1019—1056	1102	125646	0.147	1.16
28	36	1057—1092	1008	126654	0.147	1.08
27	42	1093—1134	1134	127788	0.147	1.00
26	44	1135—1178	1144	128932	0.147	1.03
25	43	1179—1221	1075	130007	0.147	1.14
24	50	1222—1271	1200	131207	0.146	1.06
23	62	1272—1333	1426	132633	0.145	0.93
22	71	1334—1404	1562	134195	0.146	0.90
21	51	1405—1455	1071	135266	0.147	1.36
20	65	1456—1520	1300	136566	0.145	1.17
19	96	1521—1616	1824	138390	0.144	0.89
18	93	1617—1709	1674	140064	0.145	1.02
17	103	1710—1812	1751	141815	0.145	1.03
16	102	1813—1914	1632	143447	0.144	1.11
15	158	1915—2072	2370	145817	0.143	0.87
14	158	2073—2230	2212	148029	0.144	1.01
13	169	2231—2399	2197	150226	0.144	1.09
12	212	2400—2611	2544	152770	0.143	1.03
11	251	2612—2862	2761	155531	0.143	1.04
10	260	2863—3122	2600	158131	0.142	1.20
9	321	3123—3443	2889	161020	0.140	1.19
8	421	3444—3864	3368	164388	0.137	1.15
7	401	3865—4265	2807	167195	0.134	1.52
6	597	4266—4862	3582	170777	0.127	1.36
5	788	4863—5650	3940	174717	0.121	1.43
4	1158	5651—6808	4632	179349	0.112	1.47
3	1806	6809—8614	5418	184767	0.102	1.59
2	3273	8615—11887	6546	191313	0.857	1.82
1	9581	11888—21468	9581	200894	0.591	2.24

Таблица 15

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	$\bar{s}$ γ
6686	1	1	6686	6686	0.033	
5742	1	2	5742	12428	0.057	
3924	1	3	3924	16352	0.059	
3914	1	4	3914	20266	0.078	
2228	1	5	2228	22494	0.056	
2218	1	6	2218	24712	0.066	
1867	1	7	1867	26579	0.065	
1864	1	8	1864	28443	0.074	
1773	1	9	1773	30216	0.080	
1493	1	10	1493	31709	0.074	
1460	1	11	1460	33169	0.080	
1427	1	12	1427	34496	0.085	
1395	1	13	1395	35991	0.091	

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
1368	1	14	1368	37359	0.096	
1308	1	15	1308	38667	0.098	
1218	1	16	1218	39885	0.097	
1202	1	17	1202	41087	0.102	
1166	1	18	1166	42253	0.105	
1159	1	19	1159	43412	0.110	
1146	1	20	1146	44558	0.114	
1041	1	21	1041	45599	0.109	
1003	1	22	1003	46602	0.110	
978	1	23	978	47580	0.112	
968	1	24	968	48548	0.116	
958	1	25	958	49506	0.120	
915	1	26	915	50421	0.119	
914	1	27	914	51335	0.123	
911	1	28	911	52246	0.127	
816	1	29	816	53062	0.118	
785	1	30	785	53847	0.117	
753	1	31	753	54600	0.116	
746	1	32	746	55346	0.119	
738	1	33	738	56084	0.122	
731	1	34	731	56815	0.124	
716	1	35	716	57531	0.125	
685	1	36	685	58216	0.123	
680	1	37	680	58896	0.126	
678	2	38—39	1356	60252	0.129	
651	2	40—41	1302	61554	0.130	
635	1	42	635	62189	0.133	
634	1	43	634	62823	0.136	
622	1	44	622	63445	0.137	
606	1	45	606	64051	0.136	
589	1	46	589	64640	0.135	
570	1	47	570	65210	0.134	
566	1	48	566	65776	0.136	
562	1	49	562	66338	0.137	
554	1	50	554	66892	0.138	
552	1	51	552	67444	0.140	
534	1	52	534	67978	0.139	
517	1	53	517	68495	0.137	
511	1	54	511	69006	0.138	
506	1	55	506	69512	0.139	
500	1	56	500	70012	0.140	
496	1	57	496	70508	0.141	
483	1	58	483	70991	0.140	
472	1	59	472	71463	0.139	
465	1	60	465	71928	0.139	
462	1	61	462	72390	0.141	
459	1	62	459	72849	0.142	
455	1	63	455	73304	0.143	
446	1	64	446	73750	0.142	
439	1	65	439	74189	0.142	
438	1	66	438	74627	0.144	
437	2	67—68	874	75501	0.146	
433	1	69	433	75934	0.149	

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
432	2	70—71	864	76798	0.151	
430	1	72	430	77228	0.155	
419	1	73	419	77647	0.153	
417	1	74	417	78064	0.154	
404	1	75	404	78468	0.151	
392	1	76	392	78860	0.149	
390	1	77	390	79250	0.150	
382	1	78	382	79632	0.149	
374	1	79	374	80006	0.147	
363	1	80	363	80369	0.145	
358	1	81	358	80727	0.145	
353	1	82	353	81080	0.144	
352	1	83	352	81432	0.146	
348	1	84	348	81780	0.146	
346	2	85—86	692	82472	0.147	
345	1	87	345	82817	0.150	
343	1	88	343	83160	0.151	
338	1	89	338	83498	0.150	
335	1	90	335	83833	0.150	
333	1	91	333	84166	0.151	
329	1	92	329	84495	0.151	
324	2	93—94	648	85143	0.150	
323	1	95	323	85466	0.153	
321	1	96	321	85787	0.154	
317	2	97—98	634	86421	0.153	
316	1	99	316	86737	0.156	
312	2	100—101	624	87361	0.156	
310	1	102	310	87671	0.158	
308	1	103	308	87979	0.158	
305	1	104	305	88284	0.158	
300	1	105	300	88584	0.157	
298	3	106—108	894	89478	0.158	
297	1	109	297	89775	0.162	
296	1	110	296	90071	0.162	
295	1	111	295	90366	0.163	
294	1	112	294	90660	0.164	
291	1	113	291	90951	0.164	
290	1	114	290	91241	0.165	
287	1	115	287	91528	0.165	
286	2	116—117	572	92100	0.160	
284	1	118	284	92384	0.167	
283	2	119—120	566	92950	0.168	
280	1	121	280	93230	0.169	
279	2	122—123	558	93788	0.170	
277	1	124	277	94065	0.171	
276	1	125	276	94341	0.172	
275	1	126	275	94616	0.173	
273	2	127—128	546	95162	0.173	
272	1	129	272	95434	0.175	
271	1	130	271	95702	0.176	
266	1	131	266	95971	0.174	
265	1	132	265	96236	0.175	
264	1	133	264	96500	0.175	

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
257	1	134	257	96757	0.172	
256	1	135	256	97013	0.172	
254	1	136	254	97256	0.172	
253	2	137—138	506	97773	0.173	
252	1	139	252	98025	0.175	
251	2	140—141	502	98527	0.175	
250	2	142—143	500	99027	0.177	
247	1	144	247	99274	0.177	
245	1	145	245	99519	0.177	
244	2	146—147	488	100007	0.178	
242	1	148	242	100249	0.179	
241	2	149—150	482	100731	0.180	
240	1	151	240	100971	0.181	
237	3	152—154	711	101682	0.180	
234	1	155	234	101916	0.181	
233	1	156	233	102149	0.181	
231	1	157	231	102380	0.181	
230	1	158	230	102610	0.181	
228	1	159	228	102838	0.181	
227	1	160	227	103065	0.181	
222	4	161—164	888	103953	0.177	
221	2	165—166	442	104395	0.182	
220	1	167	220	104615	0.183	
216	1	168	216	104831	0.181	
215	1	169	215	105046	0.181	
213	1	170	213	105259	0.181	
212	1	171	212	105471	0.181	
211	1	172	211	105682	0.181	
210	1	173	210	105892	0.181	
209	2	174—175	418	106310	0.181	
208	1	176	208	106518	0.183	
207	1	177	207	106725	0.183	
205	3	178—180	615	107440	0.182	
204	1	181	204	107544	0.184	
203	3	182—184	609	108153	0.184	
200	3	185—187	600	108753	0.185	
199	1	188	199	108952	0.187	
198	1	189	198	109150	0.187	
196	2	190—191	392	109542	0.186	
194	1	192	194	109736	0.186	
190	1	193	190	109926	0.183	
188	2	194—195	376	110302	0.182	
184	2	196—197	368	110670	0.180	
182	1	198	182	110852	0.180	
181	1	199	181	111033	0.180	
180	3	200—202	540	111573	0.180	
179	1	203	179	111752	0.181	
176	1	204	176	111928	0.179	
175	1	205	175	112103	0.179	
174	2	206—207	348	112451	0.179	
173	3	208—210	519	112970	0.180	
172	4	211—214	688	113658	0.181	
171	1	215	171	113829	0.183	

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
169	1	216	169	113998	0.182	
168	1	217	168	114166	0.182	
166	2	218—219	332	114498	0.181	
165	1	220	165	114663	0.181	
162	1	221	162	114825	0.179	
161	2	222—223	322	115147	0.178	
160	2	224—225	320	115467	0.179	
159	1	226	159	115626	0.179	
158	1	227	158	115784	0.179	
157	1	228	157	115941	0.179	
156	2	229—230	312	116253	0.178	
155	3	231—233	465	116718	0.179	
154	2	234—235	308	117026	0.180	
153	1	236	153	117179	0.180	
152	1	237	152	117331	0.180	
151	1	238	151	117482	0.179	
150	1	239	150	117632	0.179	
147	2	240—241	294	117926	0.176	
146	3	242—244	438	118364	0.176	
145	2	245—246	290	118654	0.177	
144	2	247—248	288	118942	0.178	
143	3	249—251	429	119371	0.178	
141	2	252—253	282	119653	0.177	
140	2	254—255	280	119933	0.178	
139	1	256	139	120072	0.178	
138	2	257—258	276	120348	0.177	
137	2	259—260	274	120622	0.177	
136	2	261—262	272	120894	0.177	
135	4	263—266	540	121434	0.177	
134	1	267	134	121568	0.179	
133	1	268	133	121701	0.178	
132	1	269	132	121833	0.177	
131	1	270	131	121964	0.177	
130	4	271—274	520	122484	0.176	
129	3	275—277	387	122871	0.177	
128	2	278—279	256	123127	0.178	
127	1	280	127	123254	0.177	
126	3	281—283	378	123632	0.177	
125	5	284—288	625	124257	0.177	
124	3	289—291	372	124629	0.179	
123	4	292—295	492	125121	0.179	
122	5	296—300	610	125731	0.180	
121	4	301—304	484	126215	0.182	
120	5	305—309	600	126815	0.183	
119	4	310—313	476	127291	0.184	
118	2	314—315	236	127527	0.185	
117	2	316—317	234	127761	0.184	
116	3	318—320	348	128109	0.184	
114	5	321—325	570	128679	0.183	
113	4	326—329	452	129131	0.184	
112	2	330—331	224	129355	0.184	
111	1	332	111	129466	0.184	
110	4	333—336	440	129906	0.183	
109	4	337—340	436	130342	0.183	

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
108	1	341	108	130450	0.184	
107	3	342—344	321	130771	0.183	
106	7	345—351	742	131513	0.182	
105	1	352	105	131618	0.184	
104	3	353—355	312	131930	0.183	
103	4	356—359	412	132342	0.183	
102	7	360—366	714	133056	0.183	
101	4	367—370	404	133460	0.185	
100	4	371—374	400	133860	0.185	
99	4	375—378	396	134256	0.185	
98	2	379—380	196	134452	0.185	
97	2	381—382	194	134646	0.184	
96	1	383	96	134742	0.183	
95	3	384—386	285	135027	0.182	
94	3	387—389	282	135309	0.182	
93	2	390—391	186	135495	0.181	
92	5	392—396	460	135955	0.180	
91	7	397—403	637	136592	0.180	
90	3	404—406	270	136862	0.181	
89	11	407—417	979	137841	0.181	
88	5	418—422	440	138281	0.184	
87	6	423—428	522	138803	0.184	
86	4	429—432	364	139167	0.184	
85	6	433—438	510	139677	0.184	
84	4	439—442	336	140013	0.184	
83	3	443—445	249	140262	0.183	
82	3	446—448	246	140508	0.182	
81	4	449—452	324	140832	0.181	
80	5	453—457	400	141232	0.181	
79	5	458—462	395	141627	0.181	
78	4	463—466	312	141939	0.180	
77	11	467—477	847	142786	0.179	
76	3	478—480	228	143014	0.181	
75	5	481—485	375	143389	0.180	
74	6	486—491	444	143833	0.179	
73	8	492—499	584	144417	0.179	
72	5	500—504	360	144777	0.180	
71	8	505—512	568	145345	0.179	
70	6	513—518	420	145765	0.179	
69	5	519—523	345	146110	0.179	
68	5	524—528	340	146450	0.178	
67	10	529—538	670	147120	0.177	
66	6	539—544	396	147516	0.177	
65	9	545—553	585	148101	0.177	
64	6	554—559	384	148485	0.177	
63	9	560—568	567	149052	0.176	
62	6	569—574	372	149424	0.176	
61	7	575—581	427	149851	0.175	
60	8	582—589	480	150331	0.174	1.23
59	14	590—603	826	151157	0.174	0.73
58	6	604—609	348	151505	0.175	1.75
57	8	610—617	456	151961	0.173	1.35
56	9	618—626	504	152465	0.173	1.24
55	19	627—645	1045	153510	0.172	0.62

Таблица 15 (*продолжение*)

F_i	m	i	$F_i m$	$\sum_1^i F_i m$	K	γ
54	7	646—652	378	153888	0.174	1.72
53	6	653—658	318	154206	0.173	2.07
52	10	659—668	520	154726	0.171	1.28
51	10	669—678	510	155236	0.170	1.33
50	6	679—684	300	155536	0.169	2.28
49	13	685—697	637	156173	0.167	1.09
48	11	698—708	528	156701	0.167	1.34
47	10	709—718	470	157171	0.166	1.53
46	18	719—736	828	157999	0.165	0.89
45	13	737—749	585	158584	0.165	1.28
44	15	750—764	660	159244	0.165	1.16
43	12	765—776	516	159760	0.164	1.50
42	15	777—791	630	160390	0.163	1.26
41	11	792—802	451	160841	0.162	1.78
40	10	803—812	400	161241	0.160	2.03
39	15	813—827	585	161826	0.158	1.41
38	17	828—844	646	162472	0.157	1.29
37	21	845—865	777	163249	0.156	1.11
36	15	866—880	540	163789	0.156	1.63
35	24	881—902	840	164629	0.154	1.07
34	23	903—927	782	165411	0.153	1.19
33	18	928—945	594	166005	0.153	1.59
32	24	946—969	768	166637	0.151	1.26
31	28	970—997	868	167641	0.150	1.15
30	29	998—1026	870	168511	0.149	1.18
29	29	1027—1055	841	169352	0.149	1.25
28	22	1056—1077	616	169968	0.148	1.75
27	41	1078—1118	1107	171075	0.145	1.01
26	27	1119—1145	702	171777	0.145	1.68
25	28	1146—1173	700	172477	0.143	1.66
24	35	1174—1208	840	173317	0.141	1.44
23	35	1209—1243	805	174122	0.139	1.54
22	37	1244—1280	814	174936	0.137	1.57
21	40	1281—1320	840	175776	0.134	1.57
20	50	1321—1370	1000	176776	0.132	1.37
19	60	1371—1430	1140	177916	0.130	1.25
18	61	1431—1491	1098	179014	0.129	1.36
17	66	1492—1557	1122	180136	0.132	1.39
16	68	1558—1625	1088	181224	0.124	1.49
15	75	1626—1700	1125	182349	0.122	1.51
14	65	1701—1765	910	183259	0.119	1.94
13	75	1766—1840	975	184234	0.115	1.89
12	90	1841—1930	1080	185314	0.110	1.79
11	100	1931—2030	1100	186414	0.106	1.85
10	112	2031—2142	1120	187534	0.101	1.91
9	138	2143—2280	1242	188776	0.096	1.84
8	135	2281—2415	1080	189856	0.091	2.24
7	151	2416—2566	1057	190913	0.084	2.43
6	221	2567—2787	1326	192239	0.077	2.10
5	282	2788—3069	1410	193649	0.070	2.18
4	349	3070—3418	1396	195045	0.061	2.45
3	536	3419—3954	1608	196653	0.051	2.46
2	863	3955—4817	1726	198379	0.039	2.79
1	2009	4818—6826	2009	200388	0.024	3.40

Таблица 16

F_i	m	i
1647	1	1
1435	1	2
928	1	3
918	1	4
633	1	5
551	1	6
444	1	7
417	1	8
403	1	9
399	1	10
392	1	11
352	1	12
350	2	13—14
286	1	15
270	1	16
248	1	17
237	1	18
231	1	19
185	1	20
171	1	21
153	1	22
150	1	23
139	1	24
123	1	25
121	1	26
116	1	27
110	1	28
107	1	29
104	2	30—31
102	2	32—33
101	1	34
99	1	35
98	1	36
97	1	37
96	1	38
95	1	39
88	2	40—41
87	2	42—43
85	2	44—45
83	1	46
80	1	47
79	2	48—49
77	2	50—51
76	3	52—54
74	2	55—56
73	3	57—59
72	1	60
71	1	61
69	1	62
68	1	63
67	1	64
65	1	65
63	2	66—67
62	2	68—69
61	1	70
59	3	71—73
58	4	74—77
57	1	78
56	4	79—82
55	2	83—84
54	3	85—87
52	2	88—89
51	4	90—93
50	8	94—101
48	5	102—106
47	3	107—109
46	8	110—117
45	4	118—121
44	2	122—123
43	3	124—126
42	7	127—133
41	6	134—139
40	2	140—141
39	4	142—145
38	7	146—152
37	3	153—155
36	3	156—158
35	3	159—161
34	10	162—171
33	11	172—182
32	6	183—188
31	10	189—198
30	10	199—208
29	13	209—221
28	13	222—234
27	17	235—251
26	15	252—266
25	14	267—280
24	10	281—290
23	17	291—307
22	20	308—327
21	25	328—352
20	30	353—382
19	34	383—416
18	32	417—448
17	29	449—477
16	30	478—507
15	42	508—549
14	41	550—590
13	54	591—644
12	62	645—706
11	77	707—783
10	87	784—870
9	100	871—970
8	145	971—1115
7	157	1116—1272
6	229	1273—1501
5	324	1502—1825
4	489	1826—2314
3	803	2315—3117
2	1573	3118—4690
1	4774	4691—9464

Таблица 17

F_i	m	i
3209	1	1
2871	1	2
1962	1	3
1957	1	4
1109	1	5
1030	1	6
730	1	7
684	1	8
662	1	9
654	1	10
609	1	11
583	1	12
533	1	13
516	1	14
457	1	15
392	1	16
369	1	17
360	1	18
358	2	19—20
349	1	21
348	1	22

Таблица 17 (*продолжение*)

F_i	m	i
340	1	23
317	1	24
306	1	25
289	1	26
276	2	27—28
253	1	29
250	1	30
249	2	31—32
233	1	33
232	1	34
230	1	35
221	1	36
216	3	37—39
207	1	40
195	1	41
193	1	42
181	1	43
180	1	44
174	1	45
173	1	46
172	1	47
166	1	48
165	1	49
158	2	50—51
156	1	52
154	1	53
153	1	54
150	1	55
144	1	56
143	1	57
142	2	58—59
139	1	60
136	1	61
135	1	62
134	2	63—64
133	1	65
132	1	66
128	1	67
126	1	68
125	2	69—70
123	2	71—72
120	1	73
119	1	74
118	1	75
117	1	76
116	2	77—78
115	1	79
114	3	80—82
113	2	83—84
112	1	85
110	1	86
108	4	87—90
107	2	91—92
104	3	93—95
103	1	96
102	1	97
101	1	98
100	1	99
99	1	100
98	2	101—102
95	4	103—106
94	5	107—111
92	3	112—114
91	1	115
90	1	116
87	2	117—118
86	2	119—120
85	2	121—122
84	4	123—126
83	3	127—129
82	4	130—133
81	1	134
80	3	135—137
79	1	138
78	2	139—140
77	1	141
76	2	142—143
75	3	144—146
74	1	147
73	3	148—150
72	7	151—157
71	2	158—159
70	3	160—162
69	1	163
68	1	164
67	2	165—166
66	1	167
65	3	168—170
64	8	171—178
63	4	179—182
62	1	183
61	5	184—188
60	2	189—190
59	6	191—196
58	5	197—201
57	6	202—207
56	7	208—214
55	2	215—216
54	6	217—222
53	10	223—232
52	8	233—240
51	15	241—255
50	8	256—263
49	9	264—272
48	6	273—278
47	4	279—282
46	6	283—288
45	13	289—301
44	12	302—313
43	8	314—321
42	12	322—333
41	9	334—342
40	11	343—353
39	9	354—362
38	15	363—377
37	12	378—389
36	19	390—408
35	8	409—416
34	15	417—431
33	18	432—449
32	16	450—465
31	16	466—481
30	15	482—496
29	18	497—514
28	25	515—539
27	27	540—566
26	25	567—591
25	36	592—627
24	32	628—659
23	23	660—682
22	35	683—717
21	42	718—759
20	36	760—795
19	50	796—845
18	62	846—907
17	64	908—971
16	65	972—1036
15	78	1037—1114
14	77	1115—1191
13	92	1192—1283
12	95	1284—1378
11	126	1379—1504
10	151	1505—1655
9	225	1656—1880
8	260	1881—2140
7	315	2141—2455
6	392	2456—2847
5	505	2848—3352
4	765	3353—4117
3	1192	4118—5309
2	2387	5310—7696
1	6366	7697—14062

Таблица 18

F_i	m	i	F_i	m	i	F_i	m	i
4813	1	1	208	1	60	112	4	146—149
4306	1	2	204	1	61	111	2	150—151
2943	1	3	202	1	62	109	4	152—155
2935	1	4	201	1	63	108	2	156—157
1663	1	5	199	1	64	107	4	158—161
1545	1	6	198	1	65	106	1	162
1090	1	7	190	1	66	105	1	163
1026	1	8	189	1	67	103	2	164—165
993	1	9	187	2	68—69	100	5	166—170
981	1	10	184	1	70	98	1	171
913	1	11	180	1	71	97	3	172—174
874	1	12	178	1	72	96	3	175—177
799	1	13	177	1	73	95	1	178
724	1	14	175	3	74—76	94	1	179
685	1	15	174	2	77—78	93	5	180—184
588	1	16	172	1	79	92	3	185—187
553	1	17	171	2	80—81	91	4	188—191
540	1	18	169	2	82—83	90	2	192—193
537	2	19—20	168	1	84	88	3	194—196
523	1	21	165	3	85—87	87	7	197—203
519	1	22	162	4	88—91	86	4	204—207
510	1	23	161	1	92	85	4	208—211
475	1	24	160	2	93—94	84	4	212—215
459	1	25	159	1	95	83	5	216—220
433	1	26	156	3	96—98	82	6	221—226
414	1	27	153	1	99	81	1	227
379	1	28	151	1	100	80	4	228—231
375	2	29—30	150	1	101	79	3	232—234
373	1	31	148	1	102	78	3	235—237
349	1	32	147	1	103	77	3	238—240
348	1	33	145	1	104	76	3	241—243
345	1	34	142	3	105—107	75	6	244—249
334	1	35	141	3	108—110	74	8	248—255
331	1	36	139	1	111	73	4	256—259
324	3	37—39	138	4	112—115	72	4	260—263
292	1	40	137	1	116	71	6	264—269
289	1	41	136	1	117	70	2	270—271
271	1	42	135	1	118	69	8	272—279
270	1	43	132	1	119	68	6	280—285
264	1	44	130	2	120—121	67	10	286—295
261	1	45	129	2	122—123	66	7	296—302
256	1	46	126	4	124—127	65	7	303—309
253	1	47	124	4	128—131	64	5	310—314
249	1	48	123	1	132	63	6	315—320
247	1	49	122	1	133	62	6	321—326
237	1	50	121	2	134—135	61	5	327—331
234	2	51—52	120	3	136—138	60	5	332—336
231	1	53	118	1	139	59	7	337—343
225	2	54—55	117	2	140—141	58	10	344—353
217	1	56	115	2	142—143	57	8	354—361
215	1	57	114	1	144	56	11	362—372
213	2	58—59	113	1	145	55	11	373—383

Таблица 18 (*продолжение*)

F_i	m	i	F_i	m	i	F_i	m	i
54	12	384—395	36	11	625—635	18	69	1244—1312
53	9	396—404	35	21	636—656	17	78	1313—1390
52	7	405—411	34	20	657—676	16	75	1391—1465
51	13	412—424	33	30	677—706	15	94	1466—1559
50	12	425—436	32	31	707—737	14	114	1560—1673
49	5	437—441	31	22	738—759	13	132	1674—1805
48	13	442—454	30	23	760—782	12	162	1806—1967
47	14	455—468	29	29	783—811	11	176	1968—2143
46	14	469—482	28	29	812—840	10	217	2144—2360
45	9	483—491	27	33	841—873	9	268	2361—2628
44	10	492—501	26	26	874—899	8	307	2629—2935
43	17	502—518	25	32	900—931	7	364	2936—3329
42	14	519—532	24	45	932—976	6	480	3330—3779
41	18	533—550	23	49	977—1025	5	654	3780—4433
40	16	551—566	22	46	1026—1071	4	854	4434—5287
39	17	567—583	21	61	1072—1132	3	1533	5288—6820
38	18	584—601	20	50	1133—1182	2	2709	6821—9532
37	23	602—624	19	61	1183—1243	1	7734	9533—17263

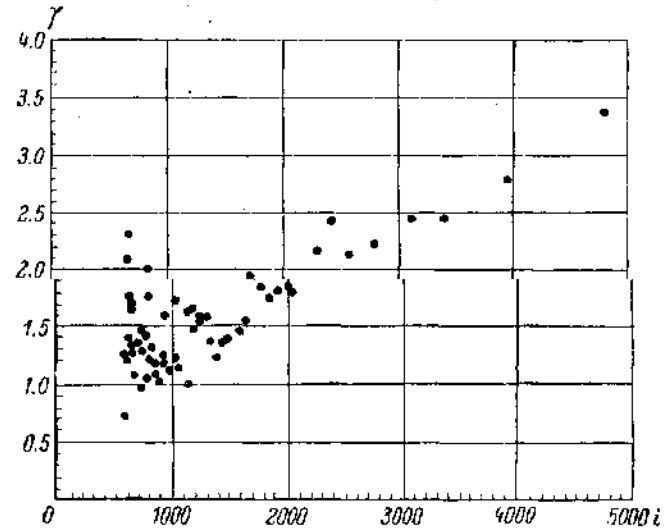


Рис. 8.

А. В. Зубов, К. Ф. Лукьяненков,
Р. Г. Пиотровский, Э. Н. Хотяшов

ЛЕКСИКО-СТАТИСТИЧЕСКОЕ ОПИСАНИЕ ТЕКСТА НА ЭЛЕКТРОННО-ВЫЧИСЛИТЕЛЬНЫХ МАШИНАХ

§ 1. Введение

Действующие алгоритмы автоматического анализа и синтеза письменного текста могут быть разработаны лишь на основе логико-комбинаторного и статистического описания соответствующего языка или его подъязыка (т. е. совокупности текстов определенной тематики).

Чтобы получить достоверное статистическое описание, необходимо обработать большие массивы текста, измеряемые сотнями страниц и сотнями тысяч словоформ (ср. статьи П. М. Алексеева, Е. А. Калининой, а также статью М. В. Данейко и др., напечатанные в настоящем сборнике).

Статистическую обработку целесообразно производить с помощью электронно-вычислительной машины (ЭВМ). В этом случае автоматизация освобождает исследователя от утомительной и малопроизводительной механической работы по извлечению, сортировке и подсчету лингвистических единиц текста, представляя ему одновременно возможность сосредоточиться на исследовательской стороне статистического описания текста.

Статистическое описание лексики текста с помощью ЭВМ предусматривает машинное решение двух различных по своему характеру общих задач.

Сущность первой общей задачи состоит в поиске и упорядочении по частоте и алфавиту лексических единиц (слов или словоформ) и их комбинаций. Решение этой задачи связано с машинным анализом больших массивов информации.

Вторая задача имеет вычислительный характер. Она предусматривает выявление некоторых общих лингво-статистических и информационных закономерностей текста (ср., например, определение параметров закона Эсту—Ципфа—Мандельброта, стр. 89—90), установление доверительных границ частот, расчет среднего количества информации, приходящегося на слова, и т. д. Здесь ЭВМ

обрабатывает только цифровую информацию небольшого объема. Поэтому решение этой задачи особых трудностей не представляет, она может быть решена на любой ЭВМ.[1]

Напротив, необходимость обработки большого объема лингвистической информации при решении первой задачи накладывает ряд ограничений на выбор вычислительной машины. Прежде чем говорить об этих ограничениях, рассмотрим лингвистические аспекты статистического описания лексики текста.

§ 2. Лингвистические аспекты статистического описания текста

Первая общая задача статистического описания лексики текста распадается на семь более простых задач, которые мы будем называть элементарными.

Первая элементарная задача состоит в составлении частотно-алфавитного списка словоформ, который рассматривается в качестве модели вероятностного распределения лексики исследуемого подъязыка (ср. стр. 65).

Вторая элементарная задача заключается в том, чтобы определить характер распределения частот словоформ в сериях (обычно используются серии длиной в 1000, 2000, 4000, 8000, 16 000 словоупотреблений), составляющих исследуемую выборку.

Третья элементарная задача состоит в выборе и упорядочении по частоте и алфавиту[2] всех трехсловных сочетаний (будем называть их триадами) текста (структура триады описывается в статье М. В. Данейко и др.). Так, например, при решении этой задачи относительно предложения

Δ' *Вершинами графов служат сами операторы* Δ'

будут выделены следующие триады:

1) Δ' *вершинами графов*
2) *вершинами графов служат*
3) *графов служат сами*
4) *служат сами операторы*
5) *сами операторы* Δ'.

---

[1] Подробное описание программы для определения этих характеристик текста, а также для оценок достоверности частотных словарей дано в статье: А. В. Зубов и Э. Н. Хотяшов. Статистический анализ текста с помощью электронно-вычислительной машины. Сб. «Энтропия языка и статистика речи», Минск, 1966.

[2] В частотных списках словоформы и словосочетания располагаются, как уже говорилось (см. стр. 66), в порядке убывающих частот. Алфавитное упорядочение используется для тех словоформ и словосочетаний, которые имеют одну и ту же частоту.

Каждая из них имеет частоту 1. Получаемый список триад может стать основой для автоматического разрешения омонимии, а иногда и полисемии словоформ. Кроме того, он может быть использован для статистического описания идиоматики, морфологии и отчасти синтаксиса текста.

Богатые комбинаторные возможности лексических единиц порождают исключительно большое количество разных триад в тексте. Общий частотный список триад включает в десятки раз больше единиц по сравнению с составленным по той же выборке частотным списком словоформ. Чтобы получить достаточно надежное статистическое описание триад, необходимо обработать огромные массивы текстов. Поэтому оказывается более выгодным и экономным составлять не общий список триад, а выбирать наиболее частые триады текста. Этот выбор можно осуществить, используя предположение о том, что наиболее частыми являются такие триады, которые включают в себя некоторые частотные словоформы, которые мы будем называть опорными словоформами.

Распознавание по опорному слову и выделение из текста частых триад осуществляется в рамках четвертой, пятой, шестой и седьмой элементарных задач.

Четвертая элементарная задача состоит в том, чтобы выбрать из текста и упорядочить по частоте и алфавиту все триады, в которых на втором месте стоит одна из наиболее частых словоформ частотного списка соответствующего подъязыка (обычно отбирается 120—150 первых единиц списка, ср. стр. 202).

Пятая, шестая и седьмая элементарные задачи заключаются в том, чтобы отобрать и упорядочить триады, опирающиеся соответственно на наиболее частые существительные, глаголы и прилагательные. Следует при этом иметь в виду, что триады строятся каждый раз с таким расчетом, чтобы полнее охватить лингвистические связи опорного слова с другими компонентами триады. Так, например, русские и английские именные триады выбираются по схеме: опорное существительное и две словоформы слева (ср. *эти экспериментальные результаты* или *the experimental results*); для французского языка, предпочитающего постпозицию эпитета, более уместным будет такое построение триады, при котором опорное существительное займет центральное положение (ср. *les résultats expérimentaux*). Подробнее об этом см. в статье М. В. Данейко и других авторов.

Первая, вторая и третья элементарные задачи могут быть решены независимо друг от друга. При реализации последующих задач всегда необходимо использовать результаты решения первой задачи (ср. выбор опорных словоформ из частотного списка).

На вычислительной системе, состоящей только из одной ЭВМ последовательного действия, одновременно можно реализовать

лишь одну элементарную задачу. Однако даже при решении одной элементарной задачи на выбор ЭВМ накладывается ряд ограничений.

Первое ограничение состоит в том, что каждая задача может быть выполнена лишь на тех машинах, которые имеют буквенный ввод и вывод: русский — при исследовании текстов, написанных русским алфавитом, латинский — для текстов, использующих латиницу.

Вторым ограничением, причем наиболее серьезным, является объем памяти машины, на которой предполагается осуществить эксперимент.

Рассмотрим с этой точки зрения реализацию нашей первой элементарной задачи — составление частотно-алфавитного списка словоформ текста.

§ 3. Принципиальная схема решения первой элементарной задачи на ЭВМ и расчет памяти машины

Опыт показывает, что за один раз удобнее всего обрабатывать массивы текста объемом в 50 000 словоупотреблений. Принципиальная схема этой обработки включает пять блоков — Б1, Б2, Б3, Б4, Б5 (схема 1). Рассмотрим работу машины на каждой из этих ступеней, определяя при этом, какие объемы оперативной и внешней памяти машины понадобятся для выполнения промежуточных операций каждого блока.

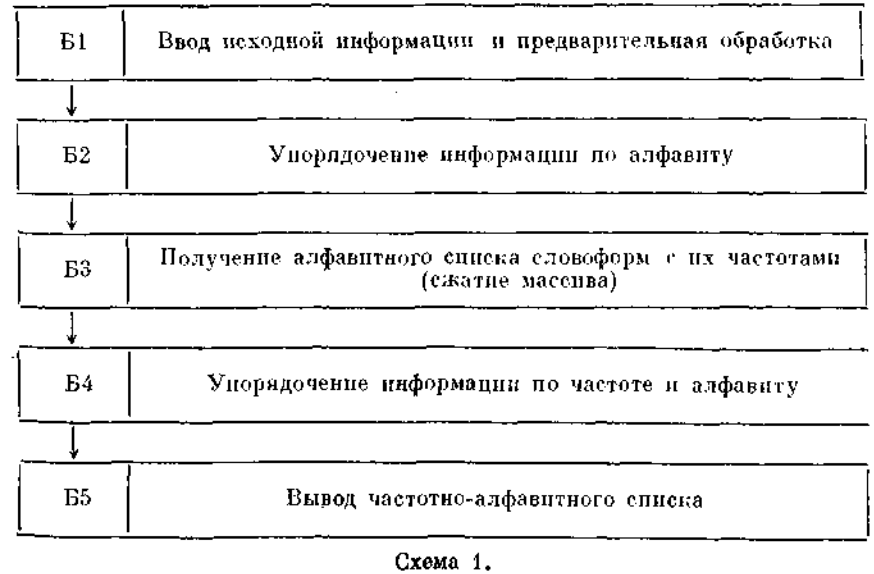


Схема 1.

Текст, кодированный на перфораторе, вводится в оперативную память электронной машины порциями по 1000 словоупотреблений. Если принять условно, что длина словоформы в исследуемом языке равна в среднем 6 буквам, то для порции в 1000 словоформ необходимо 1000 машинных слов (ячеек) по 6 символов в каждом.

В процессе предварительной обработки в Б1 каждая словоформа переписывается заново. Для ее записи отводятся теперь 4 ячейки, содержащие 24 символа (дело в том, что в исследуемых европейских языках довольно часто встречаются слова длиной в 15—20 букв). Кроме того, еще одна ячейка предназначается для записи частоты словоформы в данной выборке. Универсальная программа сортировки больших массивов информации,[3] используемая в Б2 и в Б4, предусматривает только что описанный вид представления словоформы.

Из всего сказанного следует, что для выполнения операций блока Б1 необходимо иметь дополнительно около 5000 ячеек оперативной памяти, в которых накапливаются разные словоформы текста, представленные в указанном выше виде. По мере заполнения этих 5000 ячеек информация из них переносится на магнитную ленту (МЛ) внешней памяти машины. Затем вводится новая порция в 1000 словоупотреблений, и весь цикл обработки повторяется заново. Во внешней памяти накапливается неупорядоченный массив словоформ текста, каждая из которых записана в 5 ячейках.

Сама программа, охватывающая все ступени обработки текста, занимает более 2000 машинных слов. Простой подсчет показывает, что для полного решения первой элементарной задачи необходимо около 8000 ячеек оперативной памяти по 6 символов в каждой.

В итоге предварительной обработки массива с каждой порции в 1000 словоупотреблений снимается примерно 400 разных словоформ. Поскольку, как уже было сказано, каждая словоформа записывается в 5 ячейках, на МЛ внешней памяти окажутся занятыми примерно $400 \times 5 \times 50 = 100\,000$ машинных слов.

Полученная на МЛ информация упорядочивается в Б2 по алфавиту. Для этого необходимо еще 5 магнитных лент, каждая объемом не менее 100 000 машинных слов.

Таким образом, машина, на которой происходит решение первой элементарной задачи, должна обладать внешней памятью не менее чем в 600 000 машинных слов.

В Б3 осуществляется сжатие упорядоченной по алфавиту информации: упорядоченная информация просматривается заново с целью обнаружить серии рядом стоящих одинаковых словоформ, при этом вместо серии одинаковых словоформ на МЛ заново за-

[3] См.: Э. Н. Хотяшов. Универсальная программа сортировки для ЭВМ «Минск-22». Минск, 1965.

i	F		F*
1	38028	THE	38028
2	16630	OF	54658
3	11025	AND	65683
4	9494	TO	75177
5	8010	IN	84177
6	6764	IS	92941
7	8322	A	101263
8	7151	I	108414
9	5290	Z	113704
10	4208	BE	117912
11	3705	FOR	121617
12	3675	ARE	125292
13	3203	BY	128495
14	3067	WITH	131562
15	2644	THAT	134226
16	2603	AT	136829
17	2573	AS	139402
18	2566	THIS	141968
19	2526	ON	144494
20	2426	WHICH	146920
21	2414	IT	149334
22	2158	FROM	151490
23	2080	STEAM	153570
24	1947	ENGINE	155517
25	1920	OR	157437
26	1703	TURBINE	159140
27	1699	PRESSURE	160839
. . . . .			
12951	1	X-PANEL	
12952	1	X-POUND-PER-SQUARE-INCH	
12953	1	X-PSIC	
12954	1	X-ROD	
12955	1	X-RON	
12956	1	X-ROWS	
12957	1	X-SECOND	
12958	1	X-SECTION	
12959	1	X-SHAPED	
12960	1	X-TORNERS	
12961	1	X-WAY	
12962	1	X-WINDING	
12963	1	YEARS'	
12964	1	YORK	
12965	1	Z-COMBINED	
12966	1	Z-CYLINDER	
12967	1	Z-DIAMETER	
12968	1	Z-RUN	
12969	1	Z-STROKE	
12970	1	Z-TURBO-CHARGING	
12971	1	Z-TYPE	
12972	1	Z-YEAR	
12973	1	ZEAL'	
12974	1	ZERO-ANGLE	
12975	1	ZEUNERS	

Начало и конец машинного частотного словаря английских текстов по судовым механизмам.

писывается одна словоформа вместе с суммой частот всей серии. Обычно из прошедшего через Б3 текста в 50 000 словоупотреблений на МЛ извлекается 5—7 тыс. разных словоформ, которые занимают 25—35 тыс. машинных слов внешней памяти.

В результате проведения операций в Б4 словоформы упорядочиваются по частоте их употребления в тексте. Если имеются серии разных словоформ, обладающих одинаковой частотой, то эти словоформы расставляются внутри серии по алфавиту.

Информация, полученная в результате осуществления операций в Б3 и Б4, записывается на уже использованные в Б1 и Б2 МЛ внешней памяти.

Упорядоченную в Б4 информацию необходимо отпечатать на бумаге. Для этой цели можно использовать либо выводной перфоратор с телетайпом, либо специальные алфавитно-цифровые устройства печати (АЦПУ). Вывод на АЦПУ осуществляется во много раз быстрее, чем вывод на перфоратор с телетайпом. Образец полученного на машине частотного списка словоформ дан на рисунке.

Из всего сказанного выше следует, что составление частотно-алфавитных списков (выборок) текстов длиной 50 000 словоформ целесообразно осуществлять на ЭВМ, имеющих оперативную память объемом около 8000 машинных слов, внешнюю память объемом не менее 600 000 машинных слов по шесть символов в каждом, буквенный ввод и алфавитно-цифровое устройство печати. Этим условиям отвечают такие ЭВМ, как «Минск-22», «Минск-23», «Минск 32», БЭСМ-4, БЭСМ-6, «Стрела», «Урал-4», «Урал-1», М-220.[4]

Описанная ситуация в принципе сохраняется и при решении любой другой из вышеперечисленных элементарных задач.

Если при решении первой элементарной задачи для текста длиной в 50 000 словоформ уходит от трех до четырех с половиной часов, то при увеличении выборки пропорционально возрастает и машинное время. Опыт показывает, что расход машинного времени растет здесь за счет вывода (в случае использования перфоратора и телетайпа) и особенно за счет операций ввода и предварительной обработки (Б1).

§ 4. От электронно-вычислительной машины к системе машин

Объем машинного времени, который затрачивается на выполнение той или иной задачи, всегда остается важнейшим крите-

[4] В тех случаях, когда ввод текста в оперативную память машины осуществляется мелкими порциями (до 500 словоупотреблений) и одновременно используется иная программа сортировки, чем та, которая применяется в нашем случае (см. стр. 112, примеч. 3), для решения первой элементарной задачи могут быть использованы и такие ЭВМ, которые обладают более ограниченной оперативной и внешней памятью.

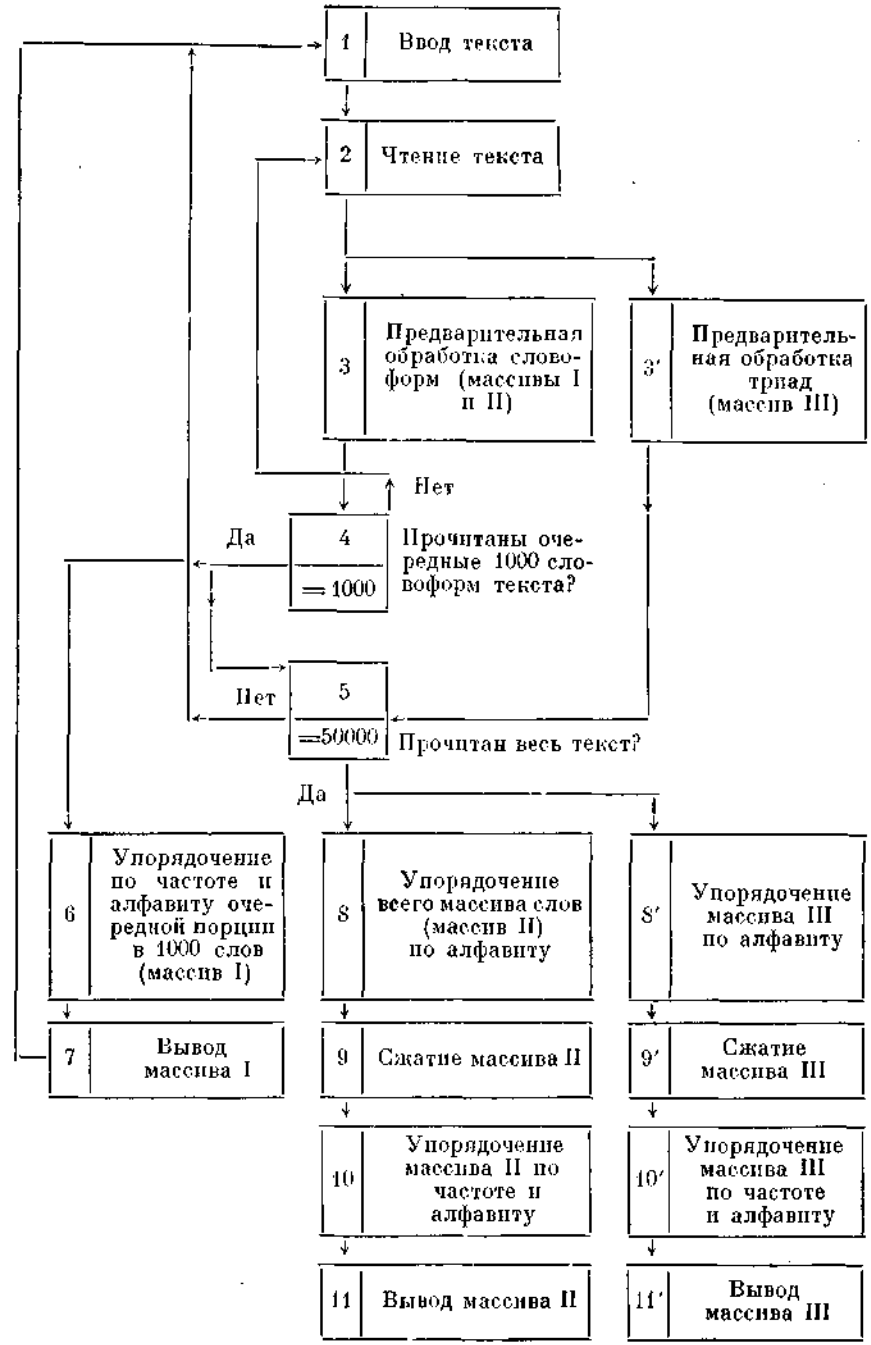


Схема 2.

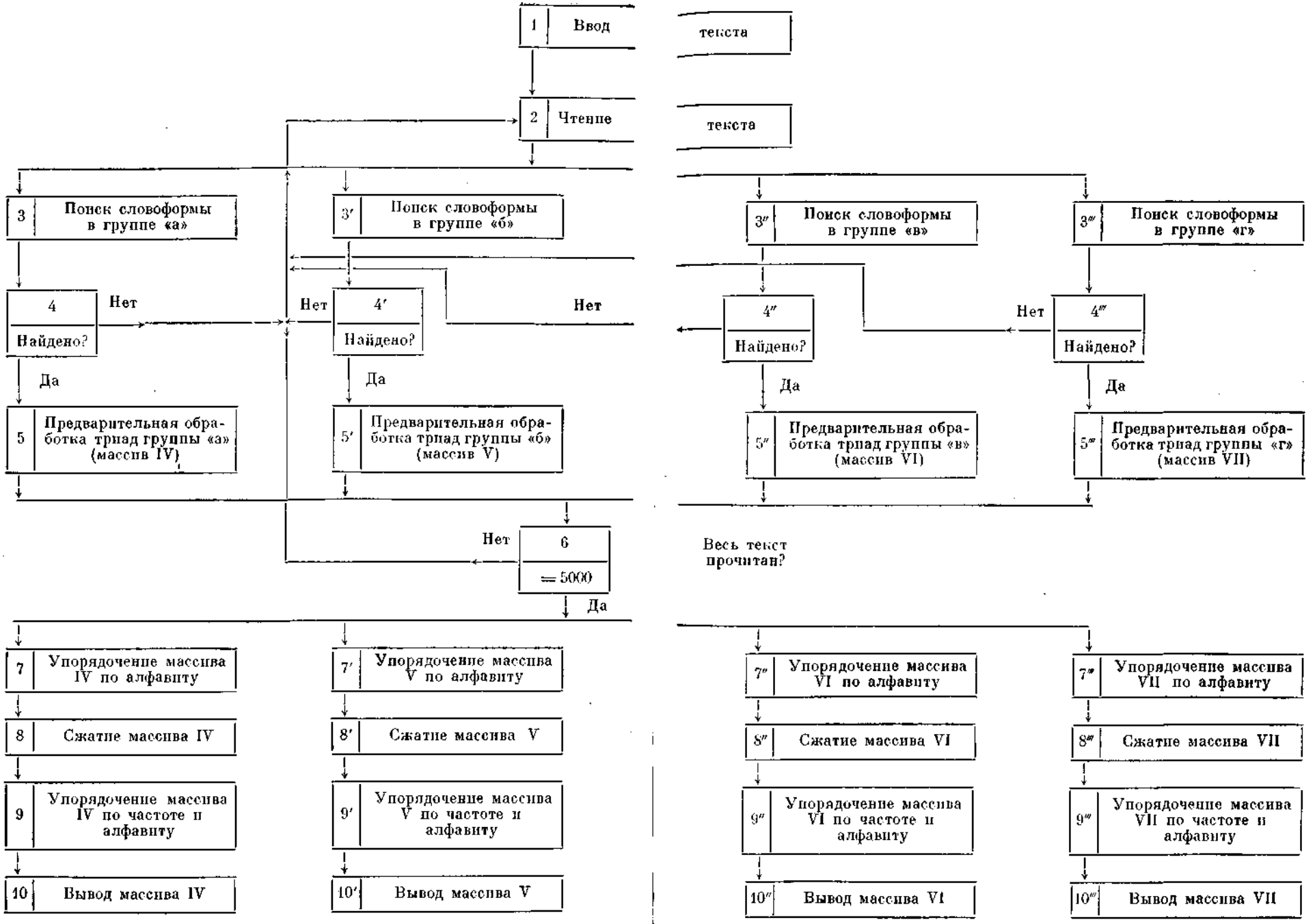

1 Ввод текста
2 Чтение текста
3 Поиск словоформы в группе «а»
3′ Поиск словоформы в группе «б»
3″ Поиск словоформы в группе «в»
3‴ Поиск словоформы в группе «г»
4 Найдено?
4′ Найдено?
4″ Найдено?
4‴ Найдено?
Нет
Да
5 Предварительная обработка триад группы «а» (массив IV)
5′ Предварительная обработка триад группы «б» (массив V)
5″ Предварительная обработка триад группы «в» (массив VI)
5‴ Предварительная обработка триад группы «г» (массив VII)
6 Весь текст прочитан? = 5000
7 Упорядочение массива IV по алфавиту
7′ Упорядочение массива V по алфавиту
7″ Упорядочение массива VI по алфавиту
7‴ Упорядочение массива VII по алфавиту
8 Сжатие массива IV
8′ Сжатие массива V
8″ Сжатие массива VI
8‴ Сжатие массива VII
9 Упорядочение массива IV по частоте и алфавиту
9′ Упорядочение массива V по частоте и алфавиту
9″ Упорядочение массива VI по частоте и алфавиту
9‴ Упорядочение массива VII по частоте и алфавиту
10 Вывод массива IV
10′ Вывод массива V
10″ Вывод массива VI
10‴ Вывод массива VII

Схема 3.

рием, определяющим целесообразность применения ЭВМ для решения этой задачи. В связи с этим рассмотрим распределение машинного времени при решении первой элементарной задачи на ЭВМ «Минск-22».[5]

При обработке текста длиной в 50 000 словоупотреблений (около 125 страниц) имеет место следующее распределение машинного времени по блокам.

Б1 (ввод исходной информации и предварительная обработка)	1 час
Б2 (упорядочение информации по алфавиту)	50 мин.
Б3 сжатие массива информации	10 мин.
Б4 (упорядочение информации по частоте и алфавиту)	35 мин.
Б5 (вывод частотно-алфавитного списка на перфоратор)	2 часа
Б5 (вывод частотно-алфавитного списка на АЦПУ-128)	20 мин.

Особенно велики и непроизводительны затраты машинного времени на операции в Б1 при последовательном решении всех элементарных задач, образующих полную программу лексико-статистического описания текста. В этом случае, переходя к следующей задаче, машина вынуждена заново читать и анализировать один и тот же текст. Так, например, в ходе решения одной элементарной задачи для ввода и предварительной обработки текста длиной в 400 000 словоупотреблений (8 порций по 50 000 словоформ), что соответствует примерно тысяче страниц, машина «Минск-22» затрачивает более 8 час. При полной первичной обработке текста указанного объема на одной машине последовательного действия это время соответственно возрастет в 6—7 раз. Общее время, необходимое для автоматической обработки текста, составит в этом случае 150—250 час.[6] Естественно, что при таком расходе машинного времени целесообразность использования ЭВМ для лексико-статистического описания текста становится сомнительной.

Время полного описания текста может быть значительно сокращено путем использования для этих целей универсальной вычислительной системы (УВС), состоящей из нескольких машин. УВС дает не только увеличение объема памяти. В этом случае в системе может параллельно работать несколько алгоритмов, что позволит решать при однократном чтении текста несколько элементарных задач.

Рассмотрим теперь примерную схему первичной статистической обработки текста с помощью вычислительной системы. Ниже приводятся две принципиальные блок-схемы обработки исходной информации, по которым может быть осуществлено параллельное решение нескольких элементарных задач.

Одна блок-схема (схема 2) предусматривает параллельное решение первых трех элементарных задач при однократном чтении текста длиной в 50 000 словоупотреблений.

Другая блок-схема (схема 3) описывает параллельное выполнение последних четырех элементарных задач.

Следует помнить, что решение последних задач возможно лишь при условии, что имеется частотный словарь исследуемого текста (ср. первую элементарную задачу), из которого вручную или с помощью ЭВМ выбраны: а) наиболее частые словоформы; б) наиболее частые существительные; в) наиболее частые глаголы; г) наиболее частые прилагательные.

Эти группы словоформ (в схеме 3 они для краткости обозначаются как «группа а», «группа б» и т. д.) вместе с исследуемым текстом помещаются в память системы, которая при втором однократном чтении машины решает оставшиеся четыре элементарные задачи.

Приведенная методика использования УВС позволяет значительно сократить машинное время, необходимое для автоматического описания текста. Основной выигрыш во времени получается за счет одновременной работы нескольких алгоритмов. Значительно сокращается и время, необходимое для ввода и чтения текста, ибо при использовании системы текст читается не семь раз, а только дважды.

При использовании вычислительных систем, состоящих из большего числа машин, можно расширить объем некоторых уже описанных элементарных задач и ввести новые элементарные задачи. Так, например, вторая элементарная задача может быть распространена на разные виды триад; кроме того, можно реализовать задачи выбора сочетаний более чем в три словоформы или выделять из триад двухсловные сочетания.

---

[5] ЭВМ «Минск-22» имеет следующие основные технические характеристики:

1) объем оперативной (внутренней) памяти — 8192 машинных слова по 6 символов в каждом;

2) объем внешней памяти — до 1 600 000 машинных слов;

3) быстродействие 5—6 тыс. операций в секунду;

4) вывод результатов на перфоратор осуществляется со скоростью 20 строк в секунду;

5) вывод результатов на УПч (АЦПУ-128) осуществляется со скоростью 400 строк в минуту.

[6] При решении каждой из элементарных задач используются универсальные программы сортировки и вывода на печать.

П. М. Алексеев

ЛЕКСИЧЕСКАЯ И МОРФОЛОГИЧЕСКАЯ СТАТИСТИКА АНГЛИЙСКОГО ПОДЪЯЗЫКА ЭЛЕКТРОНИКИ

§ 1. Методика

Объектом исследования в настоящей работе являются 200 научно-технических текстов по электронике на английском языке. Длина каждого текста равна 1000 словоупотреблений, а общая длина всех текстов равна 200 000 словоупотреблений.

Термин «электроника» понимается автором как «отдел науки и техники, занимающийся изучением прохождения электрического тока в вакууме, газах и полупроводниках и вопросами применения приборов, основанных на этом явлении».[1]

Подбор текстов производился в соответствии с этим определением по специально разработанной схеме (табл. 1).

Для каждого из 200 текстов составлялся алфавитный список всех обнаруженных в нем словоформ. При каждой словоформе указывалась частота ее употребления в данном тексте.

Этот прием позволил обойтись без картотеки при самом трудоемком этапе работы — анализе текста — и максимально «автоматизировал» сам процесс анализа.

После расписывания текстов, т. е. составления 200 алфавитно-частотных списков, словоформы были перенесены на карточки, разделенные на 200 клеток, соответствующих общему числу текстов. В клетках проставлялась частота употребления словоформы в тексте. В случаях омографии на карточке указывался индекс грамматического класса и категории данной словоформы.

При анализе текстов лексико-грамматические (типа «light» — существительное и прилагательное) и грамматические («reading» — герундий и причастие) омографы учитывались как отдельные словоформы.

[1] Л. Т. Эгер. Основы электроники. Пер. с англ. под ред. проф. Б. П. Козырева. Л., 1959, стр. 7 (сноска).

Таблица 1

Распределение текстов по разделам электроники *

Раздел	Количество текстов	% от общего количества текстов
Теоретические основы электроники		
Электрический ток в газах и вакууме	30	15
Электронная и ионная эмиссия	30	15
Электрический ток в полупроводниковых и фотоэлектрических материалах	50	25
Основные электронные и ионные приборы		
Электронные и ионные приборы	20	10
Полупроводниковые и фотоэлектрические приборы	20	10
Электроника в схемах и устройствах		
Радиотехнические устройства	10	5
Вычислительная техника и различные кибернетические устройства	20	10
Связь, телевидение, радиолокация	8	4
Автоматика и телемеханика	12	6
Итого	200	100

* Схема составлена совместно с инж. К. А. Быковым (кафедра биофизики ЛГУ) и мл. научн. сотр. Б. Т. Туссевым (Институт полупроводников АН СССР), которым автор выражает самую искреннюю признательность.

§ 2. Некоторые данные о лексико-статистической структуре английского подъязыка электроники

В текстах длиной 200 000 словоупотреблений встретились 10 582 разные словоформы. Более половины (5503) этих словоформ имеют частоту 1 и 2. Первая в частотном списке словоформа («the») занимает 9.6% всех словоупотреблений, первые 10 словоформ — 28%, 100 словоформ — 50%, 500 словоформ — 70%, 1000 словоформ — 80%, 2000 словоформ — 89%, 5000 словоформ — 98%.

В каждом тексте длиной 1000 словоупотреблений встретилось около 250 разных словоформ. После первых 10 текстов прирост словаря стал резко падать и к последним из 200 текстов составлял 20—25 новых словоформ на 1000 словоупотреблений. В табл. 2 приведены цифры, показывающие темп прироста словника. Из табл. 3 видны соотношения между длиной текста и его словником.

Таблица 2

Появление новых словоформ на каждые последующие 10 000 словоупотреблений

Выборки по 10 000 словоупотреблений	Количество новых словоформ на каждые последующие 10 000 словоупотреблений
1	2043
2	905
3	1092
4	796
5	563
6	469
7	629
8	516
9	411
10	429
11	348
12	*362*
13	298
14	245
15	255
16	294
17	243
18	236
19	225
20	223

Таблица 3

Зависимость объема словаря от длины выборки

Длина выборки	Всего разных словоформ
10 000	2043
20 000	2948
30 000	4040
40 000	4836
50 000	5399
60 000	5968
70 000	6497
80 000	7013
90 000	7424
100 000	7853
110 000	8201
120 000	*8563*
130 000	8861
140 000	9106
150 000	9361
160 000	9655
170 000	9898
180 000	10134
190 000	10359
200 000	10582

§ 3. Оценка надежности частотного словаря английского подъязыка электроники

Как планирование, так и результаты исследования определяются исходными условиями лингвистического и математического характера.

В лингвистическом плане ценность результатов зависит в первую очередь от подбора текстов,[2] а также от лингвистической методики их анализа (ср. сказанное выше о принципах выделения лексических единиц).

Выбор математических постулатов не в меньшей степени влияет на результаты лексико-статистического исследования. Основным математическим условием при составлении частотного словаря является допущение, что лексические единицы располагаются в мозгу человека или другом запоминающем устройстве в порядке роста их длины («стоимости»). На это допущение опирается доказательство известного закона распределения вероятностей слов (словоформ) в зависимости от их номера (ранга) в списке, построенном в порядке убывания их вероятностей (закон Эсту—Ципфа—Мандельброта).[3]

[2] Ср.: Л. Е ш а н, П. А л е к с е е в. Рец. на кн.: Э. А. Штейнфельдт. Частотный словарь современного русского литературного языка. «Język Rosyjski», 1964, № 3.

«Канонический закон» распределения вероятностей слов дает возможность получить предварительную оценку еще не составленного словаря, т. е. определить объем выборки, необходимый для обеспечения надежности заданного количества самых частых единиц словаря.

Опыт предшественников показал, что частотный словарь можно считать удовлетворительным, если он покрывает до $^3/_4$ объема любого текста данной жанровой тематической совокупности (подъязыка).

В английской письменной речи, по данным Г. Дьюи [4], $^3/_4$ объема текста занимают словоформы с относительными частотами, не меньшими чем 0.00019.

Пользуясь уравнением

$$\delta = \frac{Z_\rho}{\sqrt{Nf}}, \qquad (1)$$

представляющим собой упрощение формулы [5]

$$|f - p| \leq g\sqrt{\frac{pq}{N}}, \qquad (2)$$

определяем объем выборки, необходимый для обеспечения надежности той части словаря, нижним пределом которой служит частота 0.00019. В равенстве (3), преобразованном из (1),

$$N = \frac{Z_\rho^2}{\delta^2 f}, \qquad (3)$$

N — объем выборки в словоупотреблениях, Z_ρ — константа (у нас она равна 1.96), δ — относительная ошибка наблюдения и f — относительная частота словоформы. При заданной величине ошибки $\delta = 0.33$ и $f = 0.00019$ объем выборки должен быть равным 182 857 (округленно 200 000) словоупотреблений. Таковы исходные лингво-математические условия нашего исследования.

Теперь необходимо оценить надежность результатов исследования с точки зрения и синтагматики, и парадигматики языка. В этих целях используются приемы:

[3] Б. М а н д е л ь б р о т. О рекуррентном кодировании, ограничивающем влияние помех. Сб. «Теория передачи сообщений», М., 1957, стр. 139, 140, 148—153.

[4] G. D e w e y. Relativ Frequency of English Speech Sounds. Cambridge, 1923.

[5] Б. Л. В а н дер В а р д е н. Математическая статистика. М., 1960, стр. 45. Формула (1) приведена в интерпретации Р. М. Фрумкиной.

1) эмпирическая оценка — проверка наличия в частотном списке словоформ, встретившихся во взятом наугад тексте;

2) теоретическая оценка — определение доверительных интервалов для неизвестных вероятностей словоформ.

Эмпирическая проверка показала хорошую покрываемость нашим словарем взятых наугад текстов по электронике (табл. 4).

Таблица 4

Эффективность частотного словаря английского подъязыка электроники, в %

Текст	Первые 100 словоформ	Первая 1000 словоформ	Весь словарь — 10582 словоформы
Произвольный	49.9	80.4	97.3
Текст-база для словаря	49.7	79.6	100.0

Доверительные границы вероятностей словоформ частотного словаря вычислялись по упрощенным равенствам [6]

$$\left.\begin{aligned} p_1 &= \frac{F + 1.92 - 1.96\sqrt{F + 0.96}}{n}, \\ p_2 &= \frac{F + 1.92 + 1.96\sqrt{F + 0.96}}{n}, \end{aligned}\right\} \qquad (4)$$

где p_1 и p_2 — крайние точки доверительных интервалов, F — абсолютная частота словоформы, n — объем выборки в словоупотреблениях.

Относительная ошибка ε определения вероятностей по доверительным границам вычислялась по выражению

$$\left.\begin{aligned} \varepsilon_1 &= \frac{f - p_1}{p_1}, \\ \varepsilon_2 &= \frac{p_2 - f}{p_2}. \end{aligned}\right\} \qquad (5)$$

Некоторые численные значения ε_1, ε_2, p_1, p_2 и f приведены в табл. 5.

Из табл. 5 видно, что предварительно заданный нами порог (0.00019) достигается в районе 765-й словоформы. 765 словоформ покрывают около $^3/_4$ объема текста, и это соответствует предварительным условиям.

[6] Подробнее см.: Л. И. Е ш а н, П. М. А л е к с е е в. Рец. на кн.: Э. А. Штейнфельдт. Частотный словарь современного русского литературного языка. ВЯ, 1964, № 6.

Таблица 5

Доверительные границы для вероятностей словоформ частотного словаря английского подъязыка электроники

Номер словоформы, i	Относительная частота, f	Доверительные границы		Границы относительной ошибки	
		нижняя, p_1	верхняя, p_2	максимум, ε_1	минимум, ε_2
10	0.00942	0.00901	0.00986	0.04551	0.04451
500	0.00028	0.00022	0.00037	0.27273	0.024324
765	0.00019	0.00014	0.00026	0.35714	0.26923
1000	0.00014	0.00010	0.00020	0.40000	0.30000

Относительная ошибка для 765-й словоформы лежит в интервале между 0.36 и 0.27, что также согласуется с заданным требованием.

Совпадение данных словаря с их предварительными оценками позволяет считать, что лингво-математические условия исследования правильно сформулированы и достаточно корректно реализованы.

§ 4. Статистическое распределение словоформ по морфологическим классам в английском подъязыке электроники

Частотный словарь, единицам которого придана информация об их соотнесенности с определенными морфологическими классами, позволяет получить четкие количественные и качественные характеристики морфологической структуры текстов, на базе которых составлен словарь.[7]

Среди особенностей подъязыка технической литературы, отличающих его от подъязыка других жанровых совокупностей, отмечены более высокая употребительность имен существительных, преобладание форм страдательного залога и причастий над другими глагольными формами.[8]

Из известных нам частотных словарей английского языка лишь один дает сводку распределения словоупотреблений и разных словоформ по частям речи. Это словарь телефонных разговоров, составленный Френчем, Купером и Кенигом.

В табл. 6 представлены для сравнения данные по двум подъязыкам — электроники и телефонных разговоров. Для каждого подъязыка таблица имеет две вертикальные графы: первая показывает

[7] Такие сведения дают, например, работы: 1) Э. А. Ш т е й н ф е л ь д т. Частотный словарь современного русского литературного языка, Таллин, 1964; 2) T. Jakubaite, D. Kristovska, V. Ozola, R. Prūse, N. Sika, Latviešu valodas biežuma vārdnīca. Rīgā, 1966—68.

[8] С. И. К а у ф м а н. Об именном характере технического стиля (на материале английской литературы). ВЯ, 1961, № 5, стр. 105.

Таблица 6

Грамматические классы в английском подъязыке электроники и разговорном стиле *

Классы	Удельный вес класса, в %			
	в подъязыке электроники		в разговорном стиле	
	в тексте	в словаре	в речи	в словаре
Имена существительные	32.6	50.7	14.7	45.9
Имена прилагательные и наречия	12.8	17.7	12.4	28.3
Глаголы (знаменательные)	9.6	29.5	15.8	20.4
Вспомогательные и модальные глаголы	6.9	0.3	11 9	1.7
Местоимения	4.5	0.6	22.6	2.0
Предлоги и союзы	18.9	0.6	15.6	1.6
Артикли	12.8	0.03	7.0	0.1
Имена числительные	0.7	0.5		
Частицы	1.2	0.07		
Итого	100.0	100.0	100.0	100.0

* N. R. French, C. W. Carter and W. Koenig. The Words and Sounds of Telephone Conversations. «The Bell System Technical Journal», v. 9, № 2, 1930.

Таблица 7

Глагольные формы в английском подъязыке электроники

Подкласс глагола и форма	Удельный вес, в %	
	в тексте	в словаре
Знаменательные глаголы		
Наст. время (кроме 3-го л. ед. числа)	4.24	8.32
3-е л. ед. числа	5.57	14.17
Прошедшее время	0.72	2.23
Инфинитив	6.35	12.51
Герундий	3.51	13.03
Причастие I	6.64	16.36
Причастие II	31.06	32.42
Вспомогательные глаголы и связки		
Наст. время (кроме 3-го л. ед. числа) и инфинитив	15.71	0.32
3-е л. ед. числа	15.05	0.16
Прошедшее время	4.68	0.11
Причастие I	0.52	0.03
Причастие II	1.32	0.06
Модальные глаголы		
Настоящее время	4.14	0.22
Прошедшее время	0.49	0.06
Итого	100.00	100.00

отношение количества случаев употребления форм данного класса к общему количеству всех словоупотреблений в тексте (речи) и вторая — отношение количества разных словоформ данного класса к общему количеству разных словоформ в словаре.

Таблица 8

Подклассы глаголов в английском подъязыке электроники

Глаголы	Удельный вес, в %	
	в тексте	в словаре
Знаменательные	58.50	99.13
Вспомогательные и связки	36.84	0.61
Модальные	4.66	0.26
Итого	100.00	100.00

Таблица 9

Формы имени существительного в английском подъязыке электроники

Форма	Удельный вес, в %	
	в тексте	в словаре
Ед. число, общий падеж	78.97	74.31
Мн. число, общий падеж	20.68	23.24
Ед. число, притяж. падеж	0.34	2.41
Мн. число, притяж. падеж	0.01	0.04
Итого	100.00	100.00

В табл. 7, 8, 9, 10, 11, 12 приведены данные о распределении в тексте и словаре подклассов и форм глагола, существительного и артикля.

Табл. 13 показывает «удельные веса» морфологических классов в различных частотных зонах словаря электроники. По вертикали расположены частотные зоны, в горизонтальном ряду размещены выделенные нами 12 морфологических классов: имя существительное — S, имя прилагательное — A, наречие — B, глагол — V, вспомогательный глагол — $V_{вс}$, модальный глагол — $V_м$, имя числительное — Q, местоимение — H, артикль — a, предлог — F, союз — J, частица — X.

Таблица 10

Формы множественного числа имени существительного в английском подъязыке электроники по способу образования

Способ образования	Удельный вес, в %	
	в тексте	в словаре
Регулярный	98.30	98.46
Нерегулярный «исконный»	0.29	0.24
Нерегулярный заимствованный	0.41	1.30
Итого	100.00	100.00

Среди первых, наиболее частых десяти словоформ отмечены только пять классов, из них 40% падает на предлог. Эта же про-

порция сохраняется и во втором десятке. Затем доля предлога в числе словарных единиц постепенно убывает и практически сходит на нет. Подобная картина и с другими служебными классами.

Класс наречий тоже имеет тенденцию к убыванию, и лишь три класса — имя существительное, имя прилагательное и глагол — показывают определенную стабильность. Можно ожидать, что при дальнейшем увеличении объема выборки прирост словаря будет происходить за счет этих трех классов.

Таблица 11

Подклассы имен существительных в английском подъязыке электроники

Подкласс	Удельный вес, в %	
	в тексте	в словаре
Собственные . .	3.08	17.15
Нарицательные .	96.92	82.85
Итого	100.00	100.00

Обнаруженная зависимость (пусть она описана вербально и не оформлена аналитически) должна представить несомненный интерес. Она показывает, за счет каких морфологических классов происходит прирост словаря подъязыка электроники. При составлении оптимальных систем машинной переработки накопления и передачи информации этим классам будет уделена бóльшая часть объема памяти, бóльшая часть операций, бóльшая часть машинного времени и т. д.

Для методики обучения языку зависимость между весом определенного морфологического класса и ростом словаря также имеет свой смысл: рост объема лексического материала, подлежащего усвоению за пределами какого-то минимума (табл. 13), происходит за счет имен существительных (50%), глаголов (30%) и имен прилагательных (12—14%).

Таблица 12

Формы артикля в английском подъязыке электроники

Артикль	Удельный вес, в %	
	в тексте	в словаре
the	76.3	33.(3)
a	19.3	33.(3)
an	4.4	33.(3)
Итого	100.00	100.00

В результате анализа табл. 6—13 можно сделать следующие выводы.

1. Преобладающей особенностью морфологической структуры подъязыка электроники является ее именной характер.

2. Словоформы имен существительных составляют более 30% всех словоупотреблений в текстах по электронике и более 50% словаря.

3. Классы наречий, модальных глаголов, имен числительных, местоимений можно отнести к группе «второстепенных» классов вместе со служебными классами, поскольку и те и другие пред-

Таблица 13

Распределение грамматических классов в словаре английского подъязыка электроники

Зоны словаря		% от общего количества словоформ в зоне												Итого, %
порядковые номера словоформ	частоты	S	A	B	V	$V_{ПС}$	$V_{Х}$	Q	H	a	F	J	X	
1—10	19158—8868	—	—	—	—	21	—	—	—	20	40	10	10	100
11—20	1722—1144	—	—	—	—	10	—	—	20	10	40	20	—	100
21—30	1122—558	20	—	—	—	30	10	—	10	—	10	10	10	100
31—40	550—450	60	—	10	—	—	10	—	10	—	—	10	—	100
41—50	444—358	30	—	—	20	20	—	10	—	—	10	10	—	100
51—100	344—222	40	8	8	6	4	2	—	16	—	6	4	—	100
101—200	222—125	48	14	8	10	2	—	2	8	—	6	2	—	100
201—300	125—99	60	17	9	8	1	1	1	2	—	1	—	—	100
301—400	89—70	58	14	6	18	—	—	1	1	—	1	—	—	100
401—500	70—56	63	11	5	16	1	1	—	2	—	1	—	—	100
501—1000	56—28	49.3	19.5	9.6	18.4	0.2	—	0.2	1.4	—	0.6	0.6	0.2	100
1001—2000	27—11	46.5	14.1	6.5	30.8	—	—	0.4	0.7	—	0.7	0.3	—	100
2001—3000	11—6	48.6	13.1	5.7	31.1	—	—	0.6	0.5	—	0.2	0.2	—	100
3001—10582	6—1	51.3	11.9	4.4	31.2	0.1	0.05	0.4	0.2	—	0.2	0.05	0.2	100

ставляют вполне ограниченный набор слов (словоформ), занимающих незначительную часть словаря и в большинстве своем не зависящих от жанровых различий.

4. Средне- и низкочастотную часть словаря, т. е. ту часть, которая характеризует «богатство» словаря,[9] составляют на 50% словоформы имен существительных, на 30% глагольные формы и около 12% занимают словоформы имен прилагательных.

5. Прирост словаря происходит за счет низкочастотных словоформ имен существительных и глаголов.

6. На словоформы, относящиеся к классам имени существительного и знаменательного глагола, ложится основная тематическая нагрузка, выделяющая подъязык электроники в системе английского языка в целом.

7. Выявление статистико-морфологической структуры текстов данной речевой совокупности должно иметь определенное значение для ряда теоретических и прикладных задач.

§ 5. К вопросу об аналитизме и омонимии в английском языке

Статистика английской письменной речи дает возможность обнаружить по крайней мере еще две структурные особенности английского языка в целом.

Во-первых, определяется количественная мера аналитизма и языка, рассматриваемая как частное от деления числа лексем на число словоформ частотного списка (табл. 14).

Таблица 14

Степень аналитизма английского языка по данным различных частотных словарей

Подъязык	Объем выборки	Разных лексем (слов)	Разных словоформ	Степень аналитизма
Электроника	100 000	5197	7853	0.66
Электроника	200 000	7160	10 582	0.67
Личная и деловая переписка	309 400	6682	10 107	0.66
Деловая переписка*	10 830	1058	1576	0.68
Разговорный стиль**	79 300	2240	2822	0.79

* Данные взяты из работы: E. Horn. A Basic Writing Vocabulary. Iowa City, 1926.
** N. R. French, C. W. Carter and W. Koenig, ук. соч.

Во-вторых, дается оценка степени омонимичности, полученная в результате сравнения двух частотных списков, составленных с учетом и без учета омонимии (омографии). Рассмотренную зависимость можно описать эмпирической формулой

$$K_h = 1 - \frac{L_0}{L}, \qquad (6)$$

где K_h — коэффициент омонимичности, L_0 — количество словоформ без учета омонимии (омографии) и L — число словоформ с учетом омографии. Нетрудно видеть, что $0 \leqslant K_h \leqslant 1$, т. е. при широком использовании омографии в языке K_h будет стремиться к единице: при бедной омографии K_h приближается к нулю.

Для наших частотных списков величина коэффициента омонимичности равна

$$K_h = 1 - \frac{9782}{10582} = 0.076.$$

Анализ частотных списков показывает, что в английской письменной речи (на материале текстов по электронике) лексико-грамматическая и грамматическая омонимия используются весьма умеренно (за исключением небольшого числа высокоупотребительных служебных словоформ).

Таким образом, развитая система омонимии грамматических форм и классов слов в английском языке не мешает его нормальному функционированию.

[9] Ср.: О. С. Ахманова и др. О точных методах исследования языка. М., 1961, стр. 85.

В. М. Калинин

МАТЕМАТИЧЕСКИЕ АСПЕКТЫ ВОСПРИЯТИЯ ИНОЯЗЫЧНОГО ТЕКСТА

§ 1. Статистические характеристики текста

Одна из целей работ, включенных в настоящий сборник, — собрать статистические сведения по различным языкам, чтобы облегчить изучение этих языков. Задача данной статьи — показать, как можно использовать значение статистической структуры речи для описания процесса изучения иностранного языка и построения оптимальной в определенном смысле конкретной методики изучения языка.

Мы будем пользоваться результатами ранее опубликованных работ автора по статистике речи[1], придерживаясь введенных в этих работах обозначений:

N — длина текста, измеряемая числом лексических единиц, которые мы будем условно называть словами. Для математического описания достаточно предположения, что единица текста строго определена, причем текст (речь) — последовательность таких единиц; считаем также определенным понятие различных слов.

$n(N)$ — число различных слов текста N.

$n(m, N)$ — число различных слов текста N, каждое из которых употреблено ровно m раз.

$r_m(N)$ — число различных слов текста N, каждое из которых употреблено чаще, чем m раз.

p_i — априорные вероятности словоупотреблений.

L — число слов, вероятности которых отличны от 0.

Между этими величинами справедливы следующие соотношения:[2]

$$n(m, N) = \sum_{i=1}^{L} \frac{(Np_i)^m}{m!} e^{-Np_i}, \tag{1}$$

[1] Проблемы кибернетики, вып. 11, 1964; Тр. матем. инст. им. В. А. Стеклова АН СССР, т. 79, 1965.

[2] Математические ожидания случайных величин обозначены так же, как и сами случайные величины.

$$n(N) = \sum_{i=1}^{L} (1 - e^{-Np_i}), \tag{2}$$

$$r_m(N) = \frac{1}{m!} \sum_{i=1}^{L} \Gamma(m+1, Np_i), \tag{3}$$

где $\Gamma(x, y) = \int_0^y t^{x-1}e^{-t}dt$ — неполная гамма-функция. Из равенств (1), (2) и (3) следует, что можно считать функции $n(m, N)$, $n(N)$ и $r_m(N)$ бесконечно дифференцируемыми функциями непрерывного аргумента $N \geqslant 0$.

Из (1) и (2) легко получаем равенства

$$\frac{dn(m, N)}{dN} = \frac{mn(m, N) - (m+1)n(m+1, N)}{N}, \tag{4}$$

$$\frac{d^m n(N)}{dN^m} = (-1)^{m+1} \frac{m!}{N^m} n(m, N). \tag{5}$$

Равенство (5) приводит к тождествам

$$n(N) = n(N_0) - \sum_{j=1}^{\infty} \left(1 - \frac{N}{N_0}\right)^j n(j, N_0), \tag{6}$$

$$n(m, N) = \sum_{j=m}^{\infty} C_j^m \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{j-m} n(j, N_0). \tag{7}$$

Формулы (6) и (7) позволяют рассчитывать распределение текста N по распределению текста N_0. Точность расчета зависит от соотношения N и N_0. При экстраполяции ($N > N_0$) суммирование по Чезаро заметно увеличивает точность, позволяя при этом пользоваться не всей таблицей $n(j, N_0)$, $(j = 1, 2\ldots)$.

Для методики изучения иностранного языка интересны следующие две характеристики: $s(N, N_0)$ — число различных слов, одновременно входящих в тексты N и N_0, и $q(N, N_0)$ — число различных слов текста N, которые не встретились в тексте N_0.

Можно показать, что

$$q(N, N_0) = n(N + N_0) - n(N_0) = -\sum_{j=1}^{\infty} \left(-\frac{N}{N_0}\right)^j n(j, N_0),$$

$$s(N, N_0) = n(N) + n(N_0) - n(N + N_0).$$

Наиболее просто по распределению текста N найти ожидаемое число различных слов, которые войдут в новую выборку той же длины N:

$$s(N, N) = 2 \sum_{j=1}^{\infty} n(2j, N).$$

Удвоенная сумма по нечетным частотам дает ожидаемое число различных слов в выборке двойной длины:

$$n(2N) = 2\sum_{j=1}^{\infty} n(2j-1,\ N).$$

В частности, эта величина показывает, насколько был бы богаче частотный словарь, если бы для его составления была взята выборка двойной длины.

Если с приемлемой точностью справедлив закон Ципфа $p_i = \frac{K}{i}$, где K — константа, то можно пользоваться следующими приближенными выражениями для распределения текста N:

$$n(m,\ N) = \frac{NK}{m(m-1)} - \frac{NK}{m!}\Gamma(m-1,\ y) =$$

$$= \frac{NK}{m(m-1)} e^{-y} \sum_{j=0}^{m-2} \frac{1}{j!} y^j, \quad m \geqslant 2, \tag{8}$$

$$n(1,\ N) = -NK \cdot \mathrm{Ei}(-y),$$

$$n(N) = L(1 - e^{-y}) - NK \cdot \mathrm{Ei}(-y),$$

$$r_m(N) = \frac{NK}{m} + \frac{L}{m!}[m\Gamma(m,\ y) - y(m-1)\Gamma(m-1,\ y)], \quad m \geqslant 2,$$

$$r_1(N) = L(1 - e^{-y}),$$

$$r_0(N) = n(N),$$

где $y = \frac{NK}{L}$, $\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t}\,dt$ — интегральная показательная функция.

При достаточно большом m или достаточно малом y можно пользоваться более грубыми приближениями

$$n(m,\ N) \approx \frac{NK}{m(m-1)}, \quad r_m(N) \approx \frac{NK}{m}.$$

§ 2. Скорость чтения и число выученных слов

При изучении иностранного языка наиболее интересными характеристиками эффективности методики представляются $T(N)$ — время, затраченное на обработку текста N; $\frac{dT(N)}{dN}$ — время на единицу текста после обработки текста N; $W(N)$ — число различных слов, запомнившихся учащемуся после работы с текстом N; $\frac{dW(N)}{dN}$ — скорость прироста новых слов в памяти учащегося.

Все эти величины зависят как от избранной методики обучения, так и от способностей учащегося. Считаем, что имеют смысл следующие величины:

τ_j — среднее время, необходимое для понимания слова, встретившегося j-тый раз; это время будем называть реакцией на j-тую встречу;

$A_j = \sum_{k=1}^{j} \tau_k$ — средняя реакция на j встреч;

π_j — вероятность запомнить слово при j-той встрече с ним;

$\Pi_j = \sum_{k=1}^{j} \pi_k$ — вероятность запомнить слово при одной из первых j встреч с ним.

Предполагаем, что τ_j и π_j не зависят от слова, т. е. считаем их средними характеристиками времени узнавания слова и вероятности его запомнить. Их численные значения может дать только опыт.

Очевидно, что

$$T(N) = \sum_{j=1}^{\infty} \tau_j r_{j-1}(N) = \sum_{j=1}^{\infty} A_j n(j,\ N). \tag{9}$$

Использовав (4), найдем время на единицу текста после обработки текста N:

$$\frac{dT(N)}{dN} = \frac{1}{N}\sum_{j=1}^{\infty} j n(j,\ N)\,\tau_j. \tag{10}$$

Аналогично для числа запомнившихся слов имеем

$$W(N) = \sum_{j=1}^{\infty} \pi_j r_{j-1}(N) = \sum_{j=1}^{\infty} \Pi_j n(j,\ N) \tag{11}$$

и для скорости прироста их

$$\frac{dW(N)}{dN} = \frac{1}{N}\sum_{j=1}^{\infty} j n(j,\ N)\,\pi_j. \tag{12}$$

Если справедлив закон Ципфа, то из равенств (8), (10) и (12) получаем

$$\frac{dT(N)}{dN} = \frac{T(N)}{N} - \frac{Le^{-y}}{N}\sum_{j=1}^{\infty} A_j \frac{y^j}{j!}, \tag{13}$$

$$\frac{dW(N)}{dN} = \frac{W(N)}{N} - \frac{Le^{-y}}{N}\sum_{j=1}^{\infty} \Pi_j \frac{y^j}{j!}. \tag{14}$$

Величина $\frac{Le^{-y}}{N}\sum_{j=1}^{\infty}A_j\frac{y^j}{j!}$ показывает выигрыш в затрате времени на единицу текста по сравнению со средним временем $\frac{T(N)}{N}$ при изучении текста N (N измеряет текст, прочитанный с начала изучения).

При изучении языка, особенно на первой стадии, эффективно перечитывание текста. Время, необходимое для этого, а также развитую в результате скорость чтения и число запомнившихся слов можно выразить через уже введенные статистические характеристики текста, способностей учащегося и избранного им метода.

Пусть

$t_k(N)$ — время, необходимое на k-тое перечитывание текста N,

$T_r(N)$ — время всех r чтений текста N. Тогда

$$t_k(N)=\sum_{j=1}^{\infty}n(j,\ N)\sum_{m=(k-1)j+1}^{kj}\tau_m=\sum_{j=1}^{\infty}n(j,\ N)\,[A_{kj}-A_{k(j-1)}],$$

$$T_r(N)=\sum_{k=1}^{r}t_k(N)=\sum_{j=1}^{\infty}n(j,\ N)\sum_{m=1}^{rj}\tau_m=\sum_{j=1}^{\infty}n(j,\ N)\,A_{rj}.$$

Число новых слов, которые учащийся запомнит после r-кратного прочтения текста N, равно

$$W_r(N)=\sum_{j=1}^{\infty}n(j,\ N)\sum_{m=1}^{rj}\pi_m=\sum_{j=1}^{\infty}n(j,\ N)\,\Pi_{rj}, \tag{15}$$

а скорость прироста новых слов в памяти

$$\frac{dW_r(N)}{dN}=\frac{1}{N}\sum_{j=1}^{\infty}jn(j,\ N)\sum_{m=r(j-1)+1}^{rj}\pi_m=\frac{1}{N}\sum_{j=1}^{\infty}jn(j,\ N)\,[\Pi_{rj}-\Pi_{r(j-1)}].$$

Аналогичное выражение получаем для времени, затрачиваемого на r-кратное чтение единицы текста после работы с текстом N,

$$\frac{dT_r(N)}{dN}=\frac{1}{N}\sum_{j=1}^{\infty}jn(j,\ N)\sum_{m=r(j-1)+1}^{rj}\tau_m=\frac{1}{N}\sum_{j=1}^{\infty}jn(j,\ N)\,[A_{rj}-A_{r(j-1)}].$$

Оценить эффективность r-кратного чтения численно, а также найти оптимальное число перечитываний в зависимости от N можно лишь после экспериментального определения чисел τ_j и π_j.

В качестве простого примера рассмотрим идеальную методику, в которой предусматривается абсолютное заучивание новых слов:

$\pi_1=1$, $\pi_j=0$ ($j=2, 3, \ldots$), τ_1 — время, необходимое на узнавание значения слова (например, среднее время поиска слова в словаре) и выучивание его до такой степени, чтобы при новой встрече это слово требовало столько же времени, сколько и его эквивалент на родном языке. Обозначим это последнее время τ. Очевидно, равенство $\pi_1=1$ невыполнимо из-за естественного процесса забывания, особенно при механическом заучивании, а второе условие невыполнимо, даже если допустить возможность выучить иностранное слово столь хорошо, из-за того что время реакции для слов родного языка значительно сокращается благодаря контекстным связям, которые при заучивании отдельных слов никак не фиксируются.

По формулам (9), (10) и (11) находим

$$T(N)=(\tau_1-\tau)\,n(N)+\tau N,$$

$$\frac{dT(N)}{dN}=\tau+\frac{n(1, N)}{N}(\tau_1-\tau),$$

$$W(N)=n(N).$$

§ 3. Требования, предъявляемые теорией информации к методике изучения иностранного языка

Изучение иностранных языков — чрезвычайно трудоемкая задача, и разумная методика может сэкономить годы, тогда как негодная делает цель недостижимой. Теория информации К. Шеннона позволяет сформулировать основные требования, которым должен отвечать оптимальный способ.

Установлено, в чем можно убедиться и из работ настоящего сборника, что различные языковые элементы (слова, словосочетания, грамматические формы и конструкции и т. д.) имеют различную частотность. Чем чаще элемент встречается, тем нужнее его знание. Поэтому слова, фразеология, грамматические правила должны изучаться в порядке уменьшения частоты их употребления в тексте (речи). Загромождение учебников случайной лексикой, подчеркнутое выпячивание редких конструкций замедляет изучение языка.

Еще важнее, чем правильная последовательность заучивания новых элементов, оптимальная степень их заучивания. Известно, что канал связи пропускает максимальное количество информации, если он согласован с кодом, а именно если время, необходимое для передачи кодового знака, пропорционально логарифму его вероятности. В нашем случае учащийся будет читать с максимальной скоростью, если время реакции на слово будет пропорционально логарифму его вероятности. Смысл этого требования в том, что, чем чаще слово встречается, тем меньше времени на него должно быть затрачено при чтении.

Эффективность оптимального кодирования можно продемонстрировать на таком иллюстративном примере. Пусть справедлив закон Ципфа и I — средняя информация в слове, если априорные вероятности получаются как

$$p_i = \frac{K}{i}.$$

Предположим, что время реакции на слово пропорционально содержащейся в нем информации:

C_1 — время реакции на двоичную единицу в случае, если выполнено условие оптимального согласования;

C_2 — время реакции на двоичную единицу в случае, если все L слов априорно равновероятны (учились с затратами времени, в среднем независимыми от ранга i).

Тогда один и тот же текст будет читаться за одинаковое время, если C_1 и C_2 связаны условием

$$C_1 I = C_2 \log_2 L.$$

Найдем ранг слова, время реакции на которое равно времени реакции $C_2 \log_2 L$ на слова, заученные одинаково:

$$C_1 \log_2 \frac{i}{K} = C_2 \log_2 L = C_1 I,$$

$$i = K 2^I.$$

Если взять данные, приводимые в литературе [3] для английского языка, $K=0.1$, $L=8727$, $I=11.82$, то $i=362$.

Понятно, сколько напрасного труда нужно вложить, чтобы время реакции на все $L=8727$ слов было таким же, как и время реакции на слово с номером 362.

Если переход к пониманию устной речи осуществляется от понимания письменной (как у большинства занимающихся языком самостоятельно), то необходимое условие понимания устной речи состоит в том, что реакция на слово должна быть быстрее, чем время его произнесения, притом для всех слов, иначе будут происходить постоянно нарушения понимания, не позволяющие понять и те слова, которые были бы поняты, если бы не следовали за плохо выученным словом. Невыполнение этого условия приводит к знакомому многим положению, когда непонятный устный текст в письменном виде оказывается чрезвычайно простым. Если время реакции на слово больше времени его произнесения (в котором входит также пауза между словами), то слух и мозг заняты осознанием этого слова, когда произносятся уже следующие слова, остающиеся невоспринятыми.

[3] Л. Бриллюэн. Наука и теория информации. М, 1960, стр. 85.

Еще одно требование теории информации состоит в том, что в интуитивном сознании носителя языка должны быть отражены не только слова, но и вероятностные связи между ними, уменьшающие время реакции на слово в контексте и облегчающие выбор нужного продолжения в устной речи. Изучить язык — значит достичь средней для иностранца избыточности, а само изучение должно осуществлять предельный переход теоремы Шеннона к оптимальному кодированию, обеспечивающему максимальную пропускную способность носителя языка как звена в канале связи. Нельзя признать оптимальной никакую методику, в которой не предусмотрен механизм согласования кода с каналом. Речь человека, в сознании которого при изучении иностранного языка не отразились реальные корреляционные связи, должна быть полна грамматических и стилистических ошибок. Такой странной речью говорят, например, люди, изучавшие язык по стихам или детективным романам. Источник примитивной речи или, наоборот, чрезмерно изысканных конструкций часто лежит в том, что при изучении не перешло в интуицию учащегося умение оценивать вероятностный вес различных языковых элементов.

Методика изучения языка может отвечать сформулированным выше требованиям только тогда, когда она предусматривает обработку учащимся большого устного или письменного текста, неизмеримо большего, чем мизерные тексты школьных учебников или «домашнего чтения» в неязыковых институтах. Общепринятая методика обучения языку, очевидно, не удовлетворяет высказанным условиям, и в этом, может быть, основная причина ее малой эффективности. Для большинства учащихся по этой методике оказывается непреодолимым барьер между учебным эрзацем и неадаптированным текстом, между письменной и устной речью. Препятствием к увеличению обрабатываемых текстов при обычной методике является необходимость разыскивать значения слов в словаре, что отнимает массу времени и сил. После отыскания значений слов остается весьма трудная задача выбора одного из синонимов, отсеивания неподходящих омонимов и увязывания слов в единое осмысленное предложение. Построение оптимальной методики, следовательно, заставляет искать иной способ выяснения смысла иноязычной фразы, более экономный, требующий меньше механической работы. Разумеется, наиболее благоприятствует изучению языка постоянное общение с говорящими на этом языке, когда непонятное слово или предложение всегда может быть объяснено.

В следующем параграфе описывается другой и, что главное, более доступный способ, также отвечающий информационным требованиям.

§ 4. Метод параллельного чтения

Метод параллельного чтения состоит в том, что берется для изучения иноязычный текст, имеющий перевод на родной или известный учащемуся язык, все непонятные места выясняются в переводном тексте. Предполагается, что делается активная попытка разобраться в неясности, прежде чем учащийся обратится ко второму тексту. Процесс обучения происходит прежде всего на таких полезных усилиях. Несовпадение оригинала и перевода не только не препятствует сопоставлению, как может показаться с первого взгляда, но даже оказывается положительным фактором, заставляя переводить не словами, а фразами с сохранением единства фразеологических сочетаний. Практика показывает, что индивидуальность художественного перевода, о которой так много говорят, если и мешает однозначно находить значение всех языковых элементов по переводу, то лишь на начальной стадии, когда знаний еще недостаточно. При росте знаний, на первом этапе очень быстром, в переводном тексте почти всегда есть доступная информация, позволяющая снять всю неясность. При непосредственном сопоставлении выявляется, что для целей излагаемого метода распространенное мнение об определенной независимости перевода от оригинала оказывается сильно преувеличенным. На начальном же этапе обучения следует выбирать достаточно простые тексты.

Положительным качеством описываемого метода представляется отсутствие временно́го интервала между совершением ошибки и ее исправлением.

При параллельном чтении раньше запоминаются те языковые элементы, которые чаще встречаются; таким образом, в среднем они заучиваются в оптимальной последовательности. Чем чаще встречается, например, некоторое слово, тем лучше оно будет выучено, тем меньше будет время реакции на него. Простота метода, сокращение до минимума механической работы со словарем позволяет обрабатывать большие тексты. Вероятностные распределения высших порядков, корреляционные связи автоматически отражаются в сознании учащегося вместе с отдельными словами. К тому же учащийся прочитывает большой текст, и сам интерес к чтению поддерживает метод.

Особенности статистической структуры текста приводят к равномерности заучивания новых слов. Пусть в среднем нужно ν раз встретить слово, чтобы его запомнить. По закону Ципфа после прочтения текста N будет выучено в среднем $i=\frac{NK}{\nu}$ слов, т. е. при прочтении каждых $\frac{\nu}{K}$ слов текста в среднем запоминается одно новое слово.

Однородность метода дает надежду на удовлетворительную точность формул § 2. Пусть память учащегося можно охарактеризовать числом p — вероятностью запомнить слово при встрече с ним. Нужно считать этот параметр средней величиной, так как вероятность запомнить слово зависит также от слова, его этимологии, корневого родства, графики и т. д. Вероятность не запомнить слово $q=1-p$. Введенная в § 2 вероятность запомнить слово при j-той встрече с ним оказывается равной

$$\pi_j = pq^{j-1},$$

а вероятность запомнить слово при одной из j первых встреч

$$\Pi_j = 1 - q^j.$$

По формулам (6) и (11) отсюда находим

$$W(N) = \sum_{j=1}^{\infty} n(j,\ N)(1-q^j) = n(pN). \tag{16}$$

Последнее равенство позволяет высказать утверждение об определенной эквивалентности способностей и трудолюбия: если параметр памяти p у одного учащегося в α раз меньше, чем у другого, — $p_1=\alpha p_2$, то второй должен прочитать в α раз больший текст, чтобы запомнить то же число различных слов. Нетривиальность этого утверждения только в том, что оно отвечает конкретной весьма условной характеристике способностей.

При r-кратном чтении число запомнившихся слов $W_r(N)$ выражается также через функцию $n(N)$:

$$W_r(N) = n(N(1-q^r)).$$

Допустим, что реакция на слово при j-той встрече может быть аппроксимирована гармонически:

$$\tau_j = \tau + \frac{D}{j},$$

где τ и D — константы.

Постоянную τ можно приближенно считать равной времени, затрачиваемому на слово родного языка, отвлекаясь от разницы между языками в этом отношении. Тогда можно получить простое выражение для времени на единицу текста после работы с текстом N:

$$\frac{dT(N)}{dN} = \tau + \frac{D \cdot n(N)}{N}. \tag{17}$$

В частности, если справедлив закон Ципфа,

$$\frac{dT(N)}{dN} = \tau + K\left[\frac{1-e^{-y}}{y} - E_i(-y)\right]. \tag{18}$$

Гармоническая аппроксимация τ_j дает возможность построить простой расчет необходимых затрат труда и времени на изуче-

ние языка по начальному этапу для достижения конкретного результата — заданной скорости чтения. Вычисление сводится к определению констант по измеренной скорости чтения и экстраполяции. По формуле (17) видно, что для не слишком малого N можно пользоваться приближением

$$\frac{dT(N)}{dN} \approx \tau + \frac{C}{N}, \qquad (19)$$

где $C = DL = \text{const}$. (Здесь можно измерять длину текста N числом страниц, строк...).

Будем отсчитывать N от некоторого начального значения N_0, эквивалентного начальным знаниям:

$$\frac{dT(N)}{dN} \approx \tau + \frac{C}{N + N_0}.$$

Тогда задача сводится к определению чисел C и N_0. Это можно сделать, если известна скорость чтения для N_1 и N_2:

$$\frac{dT(N_1)}{dN} = T_1', \quad \frac{dT(N_2)}{dN} = T_2'.$$

Решая два уравнения с двумя неизвестными, находим

$$N_0 = \frac{(T_2' - \tau) N_2 - (T_1' - \tau) N_1}{T_1' - T_2'},$$

$$C = (N_1 + N_0)(T_1' - \tau).$$

После этого формула (19) служит для экстраполяции на большие N. Понятно, что оценивая N_0 и C по случайным значениям T_1' и T_2', мы можем надеяться на разумный результат, если разность пар случайных величин T_1, T_2 и $N_1 T_1'$, $N_2 T_2'$ заметно больше их средних квадратических уклонений, т. е. когда метод уже дал ощутимые результаты.

Примерный объем текста, который необходимо обработать для достижения заданной величины $\frac{dT(N)}{dN} = T'$, равен

$$N \approx \frac{C}{T' - \tau} - N_0,$$

и для этого понадобится затратить время

$$T \approx \tau N + C \ln \frac{N + N_0}{N_0},$$

в результате чего учащийся будет читать текст со скоростью

$$\frac{dT(N)}{dN} \approx \tau + \frac{C}{N + N_0}.$$

Метод параллельного чтения не может быть пригоден для всех. Он предполагает привычку к систематической самостоятельной умственной работе и известный минимум начальных знаний. В особенности его можно рекомендовать студентам и выпускникам технических вузов, которые к тому же легко смогут по приведенным формулам оценить на первом этапе обучения пригодность этого метода для них. Начальные грамматические знания могут быть небольшими: следует знать спряжение и склонение, порядок слов, вспомогательные глаголы и т. п. Начальный запас слов может ограничиваться первыми сотнями слов частотного словаря. Если ставится задача научиться читать тексты по радиоэлектронике, то для этой цели можно воспользоваться соответствующим частотным словарем настоящего сборника. В этот минимум, разумеется, входят все основные служебные слова.

Известные языковые элементы образуют для учащегося «сетку» в тексте, позволяющую ему отождествлять оригинал с переводом.

Совершенное знание языка предполагает, что его знает глаз (письменная речь), ухо и язык (устная речь) и что человек способен мыслить на данном языке. Излагаемый метод нацелен главным образом на обучение глаза, но может быть трансформирован и на обучение уха и языка. Понятно, что многие затронутые здесь вопросы требуют гораздо более обстоятельного изучения, как теоретического, так и практического, и хочется надеяться, это кем-нибудь будет сделано.

Е. А. Калинина

ЧАСТОТНЫЙ СЛОВАРЬ РУССКОГО ПОДЪЯЗЫКА ЭЛЕКТРОНИКИ

i	Словоформа	F
1	**в**	6418
2	**и**	5742
3	при	3924
4	на	3914
5	для	2218
6	с	2060
7	от	1460
8	по	1368
9	что	1324
10	рис.	1308
11	из	1218
12	как	1166
13	не	1066
14	к	1032
15	а	914
16	тока	784
17	можно	738
18	ток	720
19—20	напряжения, так	716
21	электронов	698
22	случае	692
23	до	680
24	или	634
25	поля	612
26	может	578
27	между	552
28	если	506
29	через	500
30—31	времени, этом	498
32	напряжение	466
33	это	464
34	то	460
35	же	446
36	катода	442
37—39	время, где, области	432
40	разряда	390

Продолжение

i	Словоформа	F
41	чем	386
42	является	362
43	быть	360
44	также	352
45	будет	348
46	эмиссии	342
47	более	338
48	характеристики	332
49	поверхности	330
50—51	лампы, схемы	316
52	однако	312
53	системы	306
54	величины	304
55	величина	300
56	мм	290
57	имеет	288
58—59	его, образом	284
60	цепи	278
61	т. е.	272
62	работы	270
63—64	во, их	268
65	ионов	266
66	частоты	264
67	за	256
68	в.	252
69—70	схема, этого	250
71	зависимость	246
72	больше	240
73	поле	238
74	сопротивление	236
75—77	ее, импульса, только	234
78—79	значения, таким	232
80	после	230
81—82	значение, поэтому	228
83—84	импульсов, следует	226
85	генератора	224
86—88	когда, порядка, энергии	220
89—92	все, потенциал, триода, но	216
93—94	сопротивления, тем	214
95	температуры	212
96—98	линзы, мишени, выше	208
99	меньше	204
100	эмиттера	202
101	о	200
102	зависимости	198
103	мксек	196
104	типа	194
105—107	было, двух, коэффициент	190
108—110	зависит, значительно, см.	188
111	того	186
112—115	несколько, оси, различных, электродов	184
116—117	необходимо, распределение	182

Продолжение

i	Словоформа	F
118	мы	180
119—121	изменения, перехода, работе	174
122	пучка	172
123	ламп	170
124—127	видно, заряда, которых, со	168
128—131	измерения, например, питания, потенциала	166
132	могут	164
133	кривая	163
134—135	обычно, этой	162
136—139	газа, проводимости, следовательно, характеристик	160
140	этих	158
141	усилителя	157
142—143	всех, мощности	156
144—145	анода, изменение	152
146	базы	150
147—150	была, концентрации, связи, электроны	150
151	коэффициента	149
152	вторичной	148
153—155	виде, под, схеме	146
156—157	уже, определяется	145
158	достаточно	144
159	сигнала	143
160	величину	139
161	носителей	137
162	электронной	136
163—164	диода, лини	135
165	число	134
166—167	скорости, усиления	133
168	очень	132
169	волны	131
170	вид	130
171—174	вследствие, переход, трубки, условиях	129
175—176	которого, путем	128
177—181	имеют, кроме, практически, числа, чтобы	126
182	сетки	125
183	выхода	124
184—185	здесь, которые	123
186—187	катодом, соответствует	121
188—189	относительно, у	120
190—197	больших, дырок, импульс, коллектора, которой, причем, происходит, элементов	119
198—200	качестве, постоянной, распределения	116
201	колебаний	115
202—204	весьма, кристалла, электрона	114
205—206	катод, результате	112
207—208	плотности, устройства	110
209—211	без, получить, точки	109
212—214	приводит, реле, считать	108
215	такой	107
216—217	тогда, часть	106
218—220	иметь, помощи, увеличение	105
221—225	область, параметров, режиме, счет, электронного	104

Продолжение

i	Словоформа	F
226—230	оказывается, поскольку, том, электронами, электронным	103
231—234	выходе, ниже, позволяет, цепь	102
235—240	значений, они, пределах, представляет, частот, частоте	101
241—244	будут, легко, роль, токов	100
245—252	бы, второй, высокой, два, импульсы, одной, показано, являются	99
253—257	кривые, металла, слоя, частиц, частота	98
258	одного	97
259—263	возможность, выражения, кривой, уравнения, устройство	96
264—267	давления, емкости, результаты, этот	95
268—270	ионизации, теории, части	94
271—273	большой, длины, луча	93
274—276	анодного, таких, триодов	92
277	случаях	91
278—282	всего, количество, равна, скорость, током	90
283—289	который, линии, наиболее, потенциалов, температура, электрического, эти	89
290—294	влияние, увеличением, уменьшается, уравнение, фиг.	88
295—300	возрастает, металлов, мощность, наличие, смещения, температуре	87
301—306	емкость, кривых, мгц, напряжений, обратной, электрод	86
307—309	других, котором, разряд	85
310—313	атомов, малых, получения, прибора	84
314—319	величине, выражение, выходного, диодов, равно, собой	83
320—325	будем, вход, место, плотность, поток, условия	82
326—331	зажигания, заряд, затем, изменении, примерно, рассмотрим	81
332—335	были, действия, одновременно, очевидно	80
336—340	дает, раз, режим, электрода, электродами	79
341—347	всегда, он, плоскости, получается, проводимость, свойства, соответственно	78
348—351	них, около, сравнению, тех	77
352—353	получим, характеристика	76
354—361	источника, мере, переходе, полупроводника, полупроводников, света, система, энергия	75
362—366	вторичных, метод, нагрузки, помощью, приведены	74
367—376	вблизи, величин, возможно, излучения, положение, становится, стороны, функции, характер, чувствительности	73
377—382	анод, возникает, параметры, режима, состоянии, частицы	72
383—384	германия, определения	71
385—391	измерений, либо, нулю, одни, сильно, система, слой	70
392—395	был, сетка, схем, электрон	69
396—399	еще, приборов, применение, сигнал	68
400—404	даже, диапазоне, должны, момент, течение	67
405—410	анодом, диод, должна, методом, случая, управления	66
411—421	второго, выходной, другой, ионы, коллектор, которая, направлении, отношению, постоянного, почти, усилитель	65
422—428	выход, напряжением, получаем, согласно, сравнительно, увеличении, элемента	64
429—432	возбуждения, особенно, сигналов, электронных	63

Продолжение

i	Словоформа	F
433—436	напряжении, полупроводниковых, состоит, уменьшения	62
437—449	значениях, изменяется, катодов, нескольких, преобразователя, свойств, сек., сопротивлением, составляет, ст., степени, схему, такого	61
450—460	входной, высоких, зоны, изображения, которое, обратного, общей, отметить, процесс, решетки, формуле	60
461—471	длина, должен, перед, положительный, положительных, преобразования, рекомбинации, сторону, туннельного, уменьшение, часто	59
472—483	данных, ком., метода, непосредственно, образцов, она, памяти, первого, полученные, процесса, следующим, токи	58
484—490	бария, должно, меняется, некоторых, ряд, усиление, формы	57
491—506	благодаря, вдоль, задержки, концентрация, магнитного, по-видимому, полностью, прибор, применения процессе, равен, резко, связь, систем, слое, частотой	56
507—523	величиной, действием, записи, известно, используется, используются, много, наблюдается, отношение, передачи, переменного, подается, работа, рт., существенно, хотя, чувствительность	55
524—542	большим, две, диоды, из-за, называется, насыщения, определить, осуществляется, отдельных, плазмы, последнее, потенциального, промежутка, процессов, сеткой, увеличения, фотокатода, частоту, чего	54
543—554	амплитуды, катодного, максимума, находится, первой, поверхность, показывает, потока, разность, такие, типов, эмиттер	53
555—566	вольт-амперной, диффузии, колебания, объясняется, основном, пленки, работу, расстояния, средней, условий, частотах, ячеек	52
567—577	амплитуда, большое, интенсивности, материала, первичных, первом, полем, постоянная, повышает, разряде, расчета	51
578—586	возникновения, высокого, газов, мка., отрицательный, производится, току, фильтра, эв.	50
587—599	действие, канала, конденсатора, основных, промежутке, ростом, толщины, точке, увеличивается, формулы, фронта, экрана, эффект	49
600—615	большие, возможности, действительно, зарядов, ионами, малой, машины, нами, остается, соответствии, соответствующих, сопротивлении, состояния, такое, энергию, эта	48
616—631	включения, волн, входе, входного, диодах, использования, использовать, найти, него, некоторые, работ, равной, решения, считается, функция, эмиссия	47
632—643	а/см^2, барьера, влияния, внешней, данной, данные, достигает, источник, процессы, прямой, условие, этим	46
644—656	вероятность, длительности, имеется, мало, нужно, различными, расстояние, случай, составляющей, сравнения, той, чисел, шумов	45

Продолжение

i	Словоформа	F
657—676	дальнейшем, данном, исследования, каждый, каскада, коллекторного, лампе, ом, оно, отличается, показана, пространственного, ртути, свечения, соответствующей, спектра, схемах, температурах, триод, усилителей	44
677—693	аноде, конденсатор, лампа, напряженности, настоящее, обработки, падает, приведена, равным, разности, соответствующие, соотношение, сопротивлений, такая, т. д., уменьшением, ширина	43
694—717	генератор, друг, замедляющей, изображение, молекулы, напряженность, образуется, основной, отрицательного, отсутствие, положительного, последовательно, представлены, применяются, проводимостью, разрядного, результатов, сетке, сетку, случаев, соотношения, существует, точностью, трех	42
718—732	внешнего, использовании, напряжению, низких, нити, обладает, обмотки, обоих, одна, промежуток, пучок, силы, создания, теперь, элемент	41
733—752	анодный, внутри, задачи, наличии, напряжениях, ней, нет, отрицательной, паров, плазме, погрешность, приблизительно, пусть, разрядов, растет, сравнение, участка, электрическое, энергией, энергий	40
753—775	вакуума, другие, зоне, изготовления, изменением, источником, количества, контакта, контура, линий, ма., мишень, одним, одном, осциллографа, поверхностей, подаче, табл., тому, условии, формирования, хорошо, эмиттеров	39
776—788	аналогично, вместе, вниз., другими, коэффициентом, магнитных, обладают, падение, первый, рода, следующие, требуется, уровень	38
789—816	анодной, вещества, вопрос, данного, двумя, закону, интервале, каждого, модуляции, молекул, отличие, отсюда, пока, полного, потенциальный, представлена, пренебречь, примесей, прожектора, пятна, разных, расчет, решение, систему, скоростью, стенок, электронной, элементы	37
817—846	атомы, базе, большого, вообще, движение, длительность, код, команд, которую, максимум, меди, называют, начала, обеспечивает, обратный, окиси, операции, основании, отрицательных, полярности, потери, пространстве, проходит, развертки, сделать, состояние, стабилизации, устройств, учитывать, ход	36
847—876	блок, быстро, взаимодействия, втором, дырки, зависят, иногда, ионного, используя, кода, лежит, обстоятельство, определяются, отношении, очередь, параллельно, покрытия, полупроводнике, последнего, приборе, различие, свободных, сеточного, системой, точность, уровни, форму, числу, экране, эмиттером	35
877—896	а., анализ, атома, группы, значению, зрения, играет, именно, интенсивность, общем, определяет, показаны, положения, постоянным, появляется, прозрачности, пути, служит, уровней, цезия	34

Продолжение

i	Словоформа	F
897—918	выпрямителя, давление, движения, короны, момента, отклонение, переходом, повышение, полей, примеси, пространства, температур, теплового, уровня, форма, характеристике, частности, электронная. электронные, эмиссию, эффекта, явлений	33
919—955	большую, воздуха, всей, всем, выпрямителей, выходное, гц, диаметром, дырочной, заключается, замедление, измерение, знаки, интерес, использование, к. п. д., малым, менее, накала, нельзя, нём, положении, положительной, полосы, появление, приложенного, расстоянии, сердечника, точек, трансформатора, увеличению, установки, характера, э. д. с., эмиттерного, явление, ячейки	32
956—992	анализа, большая, большей, вполне, говоря, границы, далее, десятков, диапазона, диэлектрика. достигается, знак, импульсами, конструкции, концентрацию. луч, магнитной, макс., нее, некоторое, образования, основе, отсутствии, первичного, правило, приборы, распространения, рассматривать, результат, слоев, тиратрона, токе, три, характеристику, экран, экспериментальные, явления	31

Распределение словоформ по абсолютным частотам при $F \leqslant 30$ см. в табл. 14 статьи Е. А. Калининой «Изучение лексико-статических закономерностей на основе вероятностей модели».

П. М. Алексеев

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКОГО ПОДЪЯЗЫКА ЭЛЕКТРОНИКИ

При статистическом анализе английских текстов по электронике длиной в 200000 словоупотреблений обнаружены 10 582 разные словоформы (лексические омографы не выделялись). Здесь приводятся 2240 словоформ с частотами от 19 158 по 10 включительно.[1] Тексты подбирались по схеме, приведенной на стр. 121.

i	Словоформа	F
1	the	19158
2	of	8868
3	and	5050
4	a	4936
5	is	4570
6	in	4432
7	to	4162
8	be	2006
9	for	1952
10	with	1722
11	by	1648
12	are	1610
13	as	1560
14	that	1478
15	this	1378
16	from	1294
17	at	1222
18	an	1160
19	it	1144
20	which	1122
21	was	970
22	on	936
23	current	736

[1] В приводимом здесь списке омографы (лексико-грамматические, грамматические и лексические) не регистрируются как отдельные словоформы. Таким образом, полный частотный список без учета какой бы то ни было омографии включает в себя уже не 10 582, а 9782 словоформы.

Продолжение

i	Словоформа	F
24	temperature	630
25	will	604
26	or	594
27	can	564
28	were	562
29	if	560
30	not	558
31	voltage	548
32	have	532
33	fig.	520
34	these	514
35	cathode	480
36	energy	460
37—38	may, when	452
39—40	electrons, field	450
41	than	444
42	has	442
43	used	428
44	one	424
45	been	422
46	electron	398
47	two	388
48	between	386
49	value	380
50	shown	362
51	anode	358
52	potential	345
53—54	function, region	344
55	made	343
56	tube	338
57	ions	336
58	work	324
59	only	321
60	surface	320
61	emission	316
62	output	314
63	more	310
64	given	303
65	where	298
66—67	high, time	294
68	but	290
69	resistance	284
70—71	also, must	282
72	frequency	278
73	junction	276
74—75	all, its	274
76—78	about, other, such	272
79—80	positive, results	270
81	into	264
82	however	258
83	circuit	256
84	through	250

Продолжение

i	Словоформа	F
85	some	248
86	obtained	247
87—88	measurements, thus	244
89	would	242
90—91	constant, since	241
92	values	240
93—96	case, material, power, we	238
97	system	234
98	gas	233
99	each	230
100	small	224
101—103	equation, low, no	222
104—105	both, very	219
106	ion	217
107	number	214
108	first	213
109—110	density, effect	210
111	heat	209
112	figure	207
113—116	operation, same, so, then	206
117	data	204
118	applied	201
119	range	195
120	plasma	194
121	order	191
122	due	189
123—124	crystal, per	184
125	their	183
126	should	179
127	computer	178
128	point	177
129	collector	176
130	charge	174
131—134	because, cm, process, theory	173
135—136	first, large	172
137	possible	171
138—139	discharge, observed	169
140	conditions	168
141—142	present, they	167
143	diode	165
144	over	163
145—146	method, there	161
147—148	electric, zero	160
149—151	eq., negative, use	159
152	form	156
153—155	state, transistor, using	155
156—157	any, ionization	154
158	therefore	153
159	result	152
160—163	breakdown, impurity, less, under	151
164	experimental	150
165—167	after, higher, pulse	149

Продолжение

i	Словоформа	F
168	required	148
169—171	beam, distribution, type	146
172	shows	145
173	within	142
174	line	141
175	crystals	140
176—178	being, during, measured	138
179	magnetic	136
180	those	135
181	velocity	134
182	temperatures	133
183—187	above, base, found, second, up	131
188—189	necessary, out	130
190—192	diffusion, maximum, terms	129
193—195	determined, effective, electrode	128
196—197	radiation, rate	127
198—201	different, states, term, total	126
202—205	functions, light, pressure, signal	125
206—208	emitter, solution, source	124
209	change	123
210	section	121
211	control	120
212—214	effects, flow, thermal	119
215—216	curve, space	118
217—220	could, difference, greater, unit	117
221—223	either, increase, well	116
224	three	115
225—226	device, wave	114
227	important	113
228—229	level, several	112
230—233	amplifier, here, now, secondary	111
234—235	characteristic, normal	110
236—239	many, reverse, sample, single	109
240—245	characteristics, described, experiments, general, vacuum, various	108
246	particles	107
247—248	equal, information	106
249—250	does, following	105
251—253	conduction, distance, set	104
254—256	approximately, band, resistivity	103
257—258	properties, similar	102
259—260	heating, problem	101
261—262	across, becomes	100
263—264	series, while	99
265—267	carriers, machine, means	98
268—270	coefficient, curves, increases	97
271—274	area, grid, much, transistors	96
275—278	below, devices, input, target	95
279	along	94
280—287	compared, condition, discussed, electrical, minimum, see, simple, voltages	93
288—292	carrier, concentration, limit, materials, noise	92

Продолжение

i	Словоформа	F
293—294	before, cross	91
295—296	average, germanium	90
297—302	lower, mass, near, part, phase, produced	89
303—306	analysis, calculated, example, taken	88
307—313	approximation, direction, further, increased, linear, operating, usually	87
314—319	angle, atoms, known, obtain, position, ratio	86
320—324	considered, design, equations, parallel, space-charge	85
325—333	available, currents, dependence, electronic, free, load model, produce, samples	84
334—338	fact, gives, interaction, thermionic, times	83
339—342	absorption, corresponding, metal, table	82
343—346	although, increasing, initial, intensity	81
347—356	efficiency, equilibrium, equipment, factor, fixed, i. e. transition, valve, volts, way	80
357—358	assumed, optical	79
359—360	elements, neglecting	78
361—363	end, hence, proportional	77
364—365	pulses, structure	76
366—369	even, magnitude, response, upon	75
370—378	defined, frequencies, give, good, long, particular, side, study, without	74
379—382	axis, energies, fall, new	73
383—385	behavior, layer, seen	72
386—395	diodes, emitted, ground, having, independent, length, mm, show, silicon, treatment	71
396—401	electrodes, factors, introduction, particle, respectively, tunnel	70
402—405	collision, our, signals, tubes	69
406—409	expression, lattice, lines, reported	68
410—414	addition, applications, cases, developed, production	67
415—421	another, cathodes, contact, hydrogen, leads, read, test	66
422—427	changes, find, human, occurs, oxide, peak	65
428—434	conductivity, oscillator, relatively, silver, spectrum, speed, tungsten	64
435—440	generator, plane, radius, scattering, switch, thickness	63
441—442	parameters, theoretical	62
443—455	air, circuits, depends, excited, internal, levels, loss, mean, presence, rather, semiconductor, smaller, thermoelectric	61
456—462	account, associated, forward, indicated, mechanism, resonance, switching	60
463—476	cannot, close, consider, dc (d-c), dipole, equivalent, junctions, just, make, regions, resulting, saturation, variation, velocities	59
477—489	atomic, calculations, components, connected, edge, element, ev., mode, optimum, parameter, plate, rise, studies	58
490—494	calculation, physical, relaxation, them, typical	57
495—504	barrier, computers, do, Fermi, follows, interest, reduced, shall, techniques, until	56
505—511	arc, cell, discussion, expected, experiment, methods, waves	55

Продолжение

i	Словоформа	F
512—521	capacitance, center, direct, directly, few, meter, mobility, reference, transmitter, valves	54
522—534	application, bias, caused, certain, error, glow, hot, measuring, motion, processing, quantity, relation, respect	53
535—546	course, derived, essentially, filament, flux, force, glass, limited, provided, studied, surfaces, what	52
547—565	according, antenna, appears, carried, derived, determine, gap, integration, occur, path, place, p-n, probe, relative, room, stability, supply, variable, vary	51
566—573	donor, later, lifetime, previously, problems, short, types, uniform	50
574—581	account, again, decreases, excitation, external, photon, still, thin	49
582—597	always, cent, charged, complex, coupling, decrease, difficult, done, generally, holes, performance, points, quite, rapidly, special, Zener (zener)	48
598—608	accuracy, amp, classical, he, minority, overlap, primary, quantity, solutions, unity, whose	47
609—630	amplifiers, appear, approach, appropriate, discharges, drop, employed, eqs., introduced, involved, kinetic, larger, little, move, next, relationship, semiconductors, systems, transitions, work, written, yield	46
631—644	arrangement, chosen, comparison, conventional, critical, cutoff, development, film, generated, latter, pure, recombination, sufficient, wide	45
645—657	activation, formed, geometry, matrix, might, noted, presented, significant, spin, take, transmission, units, working	44
658—680	additional, agrement, cold, complete, completely, considerable, densities, fraction, his, indicate, negligible, neutral, operations, parts, photoelectric, plot, procedure, sufficiently, together, varies, vertical, volume, wire	43
681—698	automatic, basic, become, centers, contribution, controlled, delay, diameter, far, manner, microwave, provide, represents, shift, solid, square, storage, strength, useful	42
699—711	amount, assumption, bombardment, chamber, desired, down, enough, horizontal, multiplication, requires, stress, sulphide	41
712—727	around, cadmium, cells, collisions, consequently, except, heated, how, instrument, lens, memory, metals, numerical, open, purpose, usual	40
728—753	absolute, assume, beyond, bulk, called, common, correction, differential, drop, final, great, helium, hole, inside, itself, maintenance, measurement, nearly, note, n-type, previous, scale, tape, throughout, varying, wall	39
754—776	ac (d-c), back, based, basis, better, capacitor, component, copper, correct, dependent, detailed, digital, earlier, films, group, indicates, lead, Mc, practical, practice, probability, related, right	38
777—796	constants, counter, deformation, excess, finally, four, impedance, incident, including, integral, least, off, outside, radar, reading, reason, recent, screen, step, walls	37

Продолжение

i	Словоформа	F
797—822	analog, article, carbon, causes, containing, degree, gradient, Hamiltonian, laboratory, liquid, modes, normally, nuclear, numbers, probably, reach, recently, resistor, sec., slightly, technique, turn, us, variations, view, years	36
823—839	best, details, exponent, expressed, impact, impurities, included, instead, magnesium, solar, spectral, stage, threshold, too, true, TV, voltage-doubling	35
840—864	action, almost, appendix, coefficients, column, decay, description, dissociation, donors, evidence, fast, limits, main, moving, nature, nitrogen, ohms, particularly, phonons, plotted, reduced, retarding, secondaries, sensitive, suitable	34
865—882	adjusted, against, capacity, expansion, left, measure, observations, often, oxygen, potentials, reached, received, research, start, takes, vacancy, whether, zone	33
883—910	active, angular, anodes, appreciable, assuming, atom, cause, color, considerations, forces, intermediate, maintained, mechanical, nearest, placed, program, proper, radial, received, rf (RF), selector, sensitivity, setting, sign, somewhat, straight, sum, van	32
911—928	assembly, avalanche, contacts, depend, diagram, easily, exact, exist, heater, interesting, makes, making, pair, pressures, quantities, sections, size, slowly	31
929—946	acceleration, added, cm^2, cooling, concentrations, cylindrical, display, feet, gases, infrared, interference, peaks, phonon, product, p-type, sources, taking, transfer	30
947—975	actual, already, bath, cavity, chapter, consistent, consists, cycle, degenerate, desirable, determining, electrostatic, engineering, expressions, full, gate, greatly, hyperfine, investigation, neglected, operate, picture, readily, requirements, spherical, toward, trouble, water, wavelength	29
976—1005	advantage, aperture, apply, bandwidth, channel, clean, cubic, designed, early, half, hand, investigated, ionic, ionized, law, leakage, longer, mentioned, needed, neighbors, performed, perpendicular, produces, represent, represented, separation, specific, stable, suggested, unless	28
1006—1028	absence, applicable, construction, detail, distributions, established, evident, exists, knowledge, magnetron, mounted, object, ohm, passing, peaks, photovoltaic, plates, portion, remains, seems, shape, thermocouple, x-rays	27
1029—1066	accelerating, adding, argument, buildup, closed, concerned, condencer, considerably, consideration, contribute, detected, entire, estimated, every, finite, Fourier, front, generation, kV, lamp, layers, machines, mercury, oscillation, oscillations, period, perturbation, provided, ranges, rates, register, sequence, spacing, standard, targets, vs, weight, you	26
1067—1101	able, allow, among, arbitrary, biased, black, corrections, corresponds, Coulomb, decreased, dielectric, division, eliminate, etc, falls, five, giving, I, include, lowest, mainly, metastables, nickel, nucleus, passes, pointed, proton, rapid, referred, replaced, showed, trap, volt, write, yields	25

Продолжение

i	Словоформа	F
1102—1135	accelerated, accurate, allowed, apparatus, battery, central, check, clear, coil, compensated, configuration, controls, convenient, cut, depending, depth, did, differences, dissipation, estimate, exactly, experimentally, fed, feedback, frequently, gas-filled, metastable, momentum, neutrals, predicted, random, trigger, valid, x-ray	24
1136—1180	alloy, approaches, background, believed, capacitors, code, contains, cost, cylinder, describe, drift, drive, driving, duration, evaluated, evolution, extended, flows, height, interval, inversely, logic, ma (mA), mechanisms, modified, net, operator, paramagnetic, photons, possibility, principle, provides, radio, readings, reasons, satisfied, say, shield, stages, stated, tends, tests, third, top, video	23
1181—1232	actually, apparent, assumptions, calculate, carefully, character, clearly, combination, current-voltage, decreasing, determination, digit, emissivity, engineers, essential, exposure, furthermore, figs., held, help, identical, immediately, individual, inequality, inner, let, life, located, major, majority, mathematical, mev., observation, orbit, passed, purposes, reasonable, recorded, reflection, relay, run, slope, spectrometer, student, studied, supplied, ten, transient, transmitted, treated, validity, via	22
1233—1272	adequate, aluminum, argon, away, coating, computing, connection, converter, degeneracy, edges, eliminated, entirely, evaluation, extremely, first-order, flowing, generators, graph, highest, highly, illustrated, in., influence, keep, kev., limitations, limiting, molecules, moment, neon, operated, origin, pass, perhaps, producing, proposed, sides, strong, whereas, who	21
1273—1320	achieved, appreciably, approximate, arises, avoid, boundary, changed, channels, comparable, considering, constructed, continously, currently, detects, detection, dimensions, divergence, easy, evaporation, exposed, fairly, filter, go, going, hours, infinite, intrinsic, kept, last, lies, non-degenerate, others, outer, phenomena, reliability, remain, resistors, roughly, shielding, shifts, similarly, steady, twice, ultraviolet, understanding, upper, varied, whole	20
1321—1380	analyzer, angles, approximations, areas, attached, balance, calibration, compensation, contamination, contained, content, coordinates, detector, diffused, discrepancy, discuss, double, ends, explained, filled, formation, get, growth, hexagonal, homogeneous, initially, inputs, installation, instruction, investigations, involving, jump, leading, leaving, length, mantissa, metallic, narrow, nonlinear, opposite, outgassing, pairs, percent, perfect, polarization, preceding, published, quantitative, reactions, require, satisfactory, seem, selected, sets, situation, specimens, spot, strengths, strongly, subject, substituting, suppose, switches, symmetry, thick, valence, widely, width, yet	19

Продолжение

i	Словоформа	F
1390—1443	absorbed, accelerator, achieve, affected, apart, attention, cards, combined, complicated, compression, computation, consequence, contrast, conversion, diagrams, difficulties, dislocations, efficiencies, errors, figures, flat, floating, followed, gave, halides, instructions, key, marked, modulation, namely, need, obvious, original, oscilloscope, photoelectrons, preparation, prepared, presents, primarily, protons, question, reflected, satellite, service, sometimes, sputtering, subsequent, substitution, suggests, tend, transformer, varied, vector, uses	18
1444—1500	abrupt, acoustic, advantages, alpha, alternating, arm, bands, Bohr, card, compare, composed, computed, contain, cooled, correspond, covered, deflection, digits, directions, doped, drawn, especially, evaporated, exceeds, existence, exponential, faster, features, gaseous, goes, idea, ideal, iron, knife, listed, losses, lost, luminescent, normalized, onto, overall, phenomenon, piece, polarity, prevent, removed, representation, responsible, rises, separate, sharp, simply, specimen, straight-line, sweep, terminal, towards	17
1501—1574	accurately, adjust, agree, allows, analogous, animals, attempt, audio, author, avoided, became, characterized, closely, conclusions, continuous, count, created, describes, diamagnetic, dispersion, displacement, doppler (Doppler), e. g., elastic, electromagnetic, engineer, exceed, exchange, expect, extend, extreme, fluorescence, forms, greatest, inch, independently, industrial, integrals, interpreted, introduce, led, lie, limitation, local, luminescence, macroscopic, mark, m/sec, outputs, panel, parametric, permit, products, pumping, quench, recorder, registers, requirement, scattered, separated, sheet, showing, simplified, six, splitting, strain, subsequently, superconducting, time-base, transverse, vapor, waveform, weak, window	16
1575—1653	altered, amplified, annealing, applies, attributed, Boltzmann, bulb, camera, capable, cathode-anode, charges, charging, coated, communications, consisting, contributions, defect, department, depletion, destruction, difficulty, dislocation, domain, driven, earth, examined, exponentially, extends, Foldy, formula, fundamental, heavy, high-voltage, imperfections, improved, indeed, indicating, influence, inverse, ionizing, know, large-signal, literature, longitudinal, medium, neighboring, penetration, plots, punched, quenched, quotient, real, reduction, regarded, relays, reliable, remaining, resistivities, return, Richardson, role, sheath, small-signal, specified, spectra, switched, technical, tetrode, though, trapped, treatments, two-photon, understand, university, uranium, vacancies, versus, whenever, yoke	15
1654—1736	adjacent, altitude, apparently, backing, bombarding, boundary, bring, built, business, calibrated, careful, carry, changing, check, combining, comes, company, completed, conducting, continued, controlling, definition,	14

i	Словоформа	F
	differ, diffuse, dissipated, employing, enter, extracted, feature, flip-flop, focal, focus, geometrical, Hall, helpful, icebergs, implies, index, lenses, man, magnetore sistance, merit, microscope, milliamperes, minimize, molecule, multivibrator, objects, once, ones, outlined, overcome, own, perform, personnel, photovoltage, plus, potentiometer, principal, proceed, profile, programmer, property, radii, reaching, reaction, reasonably, rectifier, released, residual, respective, rest, review, short-circuit, significance, tap-effect, theories, thereby, vanishes, variables, watts, wheel, words	
1737—1827	Allen, analytical, answer, atmosphere, barriers, broad, broadening, bunch, calculating, cloud, come, counts, criterion, damping, Debye, derivation, determines, deviation, diam., disk, doping, echoes, electrically, electrometer, encountered, ensure, erase, extent, extraction, follow, former, future, gage, hard, Honeywell, IBM, illustrate, illustrates, infinity, intense, introducing, inversion, involves, ionize, maintain, mixture, mobilities, mounting, network, neutron, observe, obviously, occurring, partial, past, paths, pattern, poor, pre-breakdown, preliminary, programming, programs, put, quickly, radiometer, recording, refer, release, remained, repeated, representing, resultant, said, salts, saturated, screens, serious, silica, solved, sparkover, sputtered, stereo, stop, stored, streaming, tantalum, terminals, tracer, transducer, uniaxial, visual	13
1828—1961	absorb, activated, adjustment, agrees, aid, alkali, alone, amplification, appeared, arcs, arising, arithmetic, automatically, beams, behind, binary, block, bound, brought, centre, of, checked, checking, composite, composition, connecting, concentrated, concept, concerning, contaminated, core, corrected, cps, crystalline, denotes, derivative, detect, deuterium, discovered, distortion, divided, dominant, effectively, ejected, electronics, envelope, equally, et, examine, expanded, explain, explanation, express, extending, fabrication, failure, filling, foil, formulas, grounded, Hankel, illuminated, incidence, indication, industry, inherent, instance, instantaneous, insulators, integrated, interaction, investigate, involve, kind, Kramers, laboratories, leave, leaves, list, localized, logical, magnification, merely, mind, modern, modification, modulated, molybdenum, numerous, obtainable, occurred, odd, opposing, ordinary, otherwise, passage, permits, physics, plasmas, platinum, pnp, powers, precession, properly, qualitatively, quality, radiated, ratios, reaches, ref., regulator, remove, requiring, resistances, resolved, restricted, reversed, shock, shooter, situations, slight, soon, starts, steps, strip, trace, triggered, tunneling, utilized, vapour, voltmeter, volts/division, wavelengths, withstand	12

i	Словоформа	F
1962—2078	ability, acceptor, acts, alternative, anneal, arise, assumes, attain, auxiliary, backscattered, bakeout, bottom, brass, bremsstrahlung, Brooks-Herring, came, carrying, Cer-Alloy, chemical, circuitry, collected, collection, compounds, consist, cool, couple, cylinders, db (dB), denoted, differs, distances, dynamic, effort, eliminating, employ, emulsion, enclosed, entering, evaluate, examination, excessive, exhibit, exhibits, finds, forbidden, four-magnon, gradients, gradually, grown, him, hold, hr, image, includes, laser, likely, low-energy, low-temperature, matter, mesh, min., minute, minutes, molecular, monochromator, monophonic, multiplexer, neglect, notation, obtaining, organization, oscillograph, page, photoemission, photoemissive, phototube, Poisson's, portion, preset, prior, probable, proved, pump, quenching, radiator, rare, rays, relationships, record, replace, returned, scaler, scheme, sealed, secondly, seconds, shields, simplest, simplicity, singlet, slide, so-called, space-current, sparking, standard, static, striking, substitute, symmetrical, transformation, transit, ultimate, understood, vice, visible, Vleck, your	11
2079—2240	accomplishes, accordingly, act, activator, al, alternate, ambient, amp/cm^2, analyzed, anisotropy, answers, applying, appearance, approximated, attained, balanced, briefly, centres, choice, class, coaxial, cold-cathode, comparing, confined, connectors, continuum, cordinate, cycles, dark, date, deal, deduced, degrees, describing, distinguish, distributed, dividend, dry, enables, eqn., equals, escape, establish, event, experience, extensive, face, faces, family, fit, FM, focusing, foregoing, fourth, gaps, generate, getter, grids, harmonic, heavily, high-speed, hundred, increment, injection, insertion, instability, instruments, integrations, intensities, interpretation, intervals, irradiation, isolated, job, jumps, largest, ma/cm^2, manufacturers, match, Meissner, memories, microseconds, moves, m/sec, natural, never, nevertheless, non-uniform, nor, orbits, originally, parabolic, perturbations, photoconductivity, photograph, piezoresistance, plasmons, PME, practically, probes, procedures, p-states, Pyrex, punch, quartz, Raman, reads, receiving, recognized, refractive, regardless, relax, relations, remainder, reports, reproducible, resolution, resolving, right-hand, rising, root, rotation, Schottky, scope, s. c. r., self-diffusion, serves, shell, site, solving, spatial, species, speeds, started, starting, steady-state, steel, storage-grid, stream, substances, summarized, summation, surrounding, suspension, temperature-dependent, temperature-limited, themselves, thermocouples, tried, triode, tuning, uhf, valley, valleys, vanish, variety, vessel, vicinity, watt, white, why, work-function	10

В. К. Кочеткова и *Л. М. Скрелина*

ЧАСТОТНЫЙ СЛОВАРЬ ФРАНЦУЗСКОГО ПОДЪЯЗЫКА ЭЛЕКТРОНИКИ

Приводимые ниже частотные списки включают в себя наиболее употребительные слова и словоформы в текстах по электронике на французском языке. Общая длина обследованных текстов равна 100 000 словоупотреблений. Лексико-грамматические и грамматические омографы учитывались как отдельные единицы словаря.

По содержанию обследованные тексты распределены следующим образом (в % от общего числа текстов).

1. Связь, телевидение, телеметрия 32%
2. Прикладная электроника 13
3. Атомная энергия и электроника 11
4. Полупроводники 10
5. Радиотехнические устройства 9
6. Электронные схемы 6
7. Акустика, измерение 5
8. Транзисторы, спектральный метод 4
9. Аэронавтика, фотоэлектрические приборы и другая тематика, представленные каждая небольшим количеством текстов (1% на каждую тему от общего числа текстов) 10

Таблица 1

Частотный список словоформ

i	Словоформа	F
1	de	6465
2	la (*art.*)	3621
3	à	2515
4	l' (*art.*)	2464
5	le (*art.*)	2453
6	d'	2172
7	et	2014
8	les (*art.*)	1867
9	des	1696

Таблица 1 (*продолжение*)

i	Словоформа	F
10	est	1605
11	un	1552
12	en (*pron.*)	1424
13	une	1280
14	du	1253
15	par	1228
16	dans	977
17	que	897
18	pour	797
19	on	659
20	il	634
21	sur	616
22—23	au, qui	615
24	ce	511
25	sont	501
26	plus	430
27	fréquence (*s*)	408
28	ou	392
29	a	346
30	avec	331
31	se	319
32—33	figure (*s*), être (*v*)	317
34	ces	316
35—36	deux, tension	307
37	cette	298
38	peut	282
39	nous	278
40	qu'	264
41	aux	260
42	ne	249
43	pas (*nég.*)	224
44	courant (*s*)	227
45	s'	218
46—47	cas, signal	213
48	entré (*prép.*)	199
49	puissance	196
50	n'	195
51	soit	187
52	si	176
53	comme	169
54	mesure (*s*)	161
55	valeur	160
56—57	donc, été (*part.*)	157
58—59	circuit, système	153
60	mais	152
61	dont	151
62	ainsi	150
63	récepteur	146
64	fréquences	145
65	autre	142
66—67	son, temps	141
68	très	140

Таблица 1 (*продолжение*)

i	Словоформа	F
69	bande (*s*)	138
70	nombre	134
71	permet (*v*)	131
72—73	gain, type	122
74	chaque	121
75	même (*adj.*)	116
76	aussi	114
77	sortie	112
78—80	elle, fonction, ont	110
81	amplificateur	109
82—84	effet, grande, toutes (*adj.*)	107
85	où	106
86	doit	102
87	impulsions	101
88—90	étant, point (*s*), rapport	99
91	alors	98
92	sous	96
93—94	sa, part (*s*)	95
95	trois	94
96	ligne (*s*)	93
97—98	signaux, bien	92
99	possible	91
100	vitesse	89
101	cet	88
102—103	sans, tensions	87
104—106	c', émetteur, tubes	85
107—108	exemple, onde	83
109—111	entrée (*s*), récepteur, tube	82
112—115	alimentation, image (*s*), niveau, ordre	81
116—119	conditions, correspondant, faible, fonctionnement	80
120—121	circuits (*s*), non	79
122	radar	78
123—125	montage (*s*), base (*s*), contrôle (*s*)	77
126—128	après, éléments, environ	76
129—132	autres, peuvent, résistance, y	75
133—136	faut, intensité, moins, position	73
137	dispositif	72
138—140	énergie, température, sera	71
141—143	partie, région, transistors	70
144	quelques	69
145—146	façon, utilisation	68
147	caractéristiques (*s*)	67
148	fait (*s*)	66
149	forme (*s*)	65
150—151	répéteur, schéma	64
152	distance	62
153—154	mc/s, obtenir	61
155—159	commande (*s*), courbes, également, émission, jusqu'	60
160—162	c'est-à-dire, fois, lorsque	59
163—166	avons, principe, relais, variations	58
167—170	encore, faire, précision, tout (*adj.*)	57
171—172	couche (*s*), maximum (*adj.*)	56

Таблица 1 (*продолжение*)

i	Словоформа	F
173—178	celle, courbe (*s*), donné, ensemble (*s*), montre (*v*), télévision	55
179—180	générateur, lignes	54
181—184	avoir, électrique, inducteur, présente (*v*)	53
185—190	donne, élément, elles, largeur, nécéssaire, surface	52
191—194	continu, informations, plusieurs, suivant (*gér.*)	51
195—200	bruit, filtres, ils, premier, source, transmission	50
201—204	appareil, celui, impulsion, même (*adv.*)	49
205—211	amplitude, courants, fait (v), information, réglage, valeurs, vers	48
212—219	car, charge (*s*), impédance, leur (*pers.*), longueur, mémoire, sous-porteuse, transistor	47
220—231	ayant, cependant, contre, facteur, grand, lampe, lui, phase, quand, ses, utiliser, variable	46
232—235	diamètre, écart, laquelle, utilisé	45
236—242	amplificateurs, centimètre, différents, réaliser, sécurité, terme (*s*), trouve	44
243—252	antenne, assez, chauffage, diodes, échelle (*s*), emploi, maximum (*s*), peu, simple, variation	43
253—258	ceci, compteur, contact, points, représente (*v*), tous (*adj.*)	42
259—267	chromaticité, compte, état, méthode, or, première (*adj.*), puis, réseau, sens	41
268—283	aide (*s*), champ, devient, doivent, écran, haute, lecture, leur (*pos*), magnétique (*adj.*), mesures, mm, modulation, opérateur, oscillateur, voies, vue (*s*)	40
284—294	avant, balayage, beaucoup, bonne, collecteur, différentes, diffusion, ici, MHZ, noir, silicium	39
295—301	celle-ci, choix, directement, étage, ms, sensibilité, transfert	38
302—313	appel, automatique, couplage, essentiellement, luminesce (*s*), pièces, problème, radars, rayon, service, seulement, utile	37
314—324	acier, amplification, correspond, élevée, en (*prép.*), Hertz, lieu, manière, thermique, travail, toute (*adj.*)	36
325—341	afin, arsenic, autant, barres, but, comporte (*v*), comprend, condensateur, durée (*s*), était, filtre (*s*), procédé, pont, intermédiaire, particulier, toujours, utilise (*v*)	35
342—356	accoustique, blanc, calcul, diode, électronique (*adj.*), étude, faisceau, grille (*s*), jonction, intérêt, minimum (*adj.*), réponse, série, suite, utilisant	34
357—369	ailleurs, analyse (*s*), constante (*s*), côté, fabrication, grâce, invention, kHz, moteur, nécessaires, résultats, seul, seule	33
370—379	cours, éviter, mise (*s*), partir, pièce, porteur, répéteurs, technique (*s*), transformateur, zéro	32
380—393	alliage, câble, chaleur, certains, eau, enfin, ensuite, équation, étages, grandes, induction, mètre, pupitre, tableau	31
394—405	cellule (*s*), exploitation, opérations, pendant, portée (*s*), produit (*s*), relation, solution, sorte (*s*), tel, transmettre, va	30

Таблица 1 (*продолжение*)

i	Словоформа	F
406—422	capacité, chacune, commutateur, existe (*v*), installation, liaison, lumière, particulièrement, pratiquement, puisque, raison, soient, systèmes, trop, voir, voit, zone	29
423—439	avion (*s*), augmentation, bandes, certaine, coefficient, constant, couleurs, égale, formule (*s*), linéaire, location, permettent, près, quantité, réacteur, respectivement, séquence	28
440—464	appliquée, cathodique, chacun, classique (*adj.*), dernier, différence, élevé, évidemment, images, moyenne (*adj.*), moyens (*s*), nouveau, période, pompe (*s*), radiation, rapide, rayons, réduire, relativement, rendement, retard, seconde (*s*), semi-conducteur, seuil, simultanément	27
465—491	appareils, aucun, autour, bord, bornes, celui-ci, certaines, consiste, domaine, électriques, généralement, instant, mécanique (*adj.*), mêmes (*adj.*), obtenu, pilotage, pôles, problèmes, protection, régulation, sonore, souvent, spectre (*s*), tout (*pron.*), toutefois, utilisée	26
491—515	accord, actuellement, agit, alternatif, barre (*s*), constante (*adj.*), déjà, distances, dit (*part.*), donner, double (*adj.*) fil, général, obtient, parties, permettant, plan, puissances, quartz, rouge, seront, stabilité, transmise, travers, utilisés	25
516—547	arrêt, centre, changement, condition, constitué, dimensions, dispositifs, disque, données (*s*), égal, etc., extrêmement, grandeur, inférieure, kilomètres, le (*pronom.*), leurs (*pos.*), miroir, mis, moment, nouvelle (*adj.*), opération, outre, performances, pouvant, qualité, quatre, serait, structure, supérieure, trempe (*v*), unité	24
548—572	air, assurer, aucune, augmentation, avions (*s*), basse, canaux, cathode (*s*), conductibilité, cuivre (*s*), depuis, description, directe, donnés (*part.*), effectue, élevées, matière, minimum (*s*), moyen (*adj.*), présence, présente (*adj.*), potentiel (*s*), résistances, résulte, synchronisation,	23
573—601	abord, avantages, celles, chaînes, chambre, changements corps, fournit, facilement, définition, donnant, droite, électrons, erreur, générale, l' (*pron.*), lors, permettre, portée (part.), potentiomètre, pourrait, référence, régions, selon, semi-conducteurs, suivante, synchro, types, voie	22
602—628	applications, axe, certain, conséquent, contacts, dépend, électroniques, excitation, fixe (*adj.*), fusion, important, intérieur, machine, obtenue, parce, paramètres, polarisation, rapidement, rôle, rotation, sait, secondaire, self, techniques (*adj.*), trafic, transmis, volt	21
629—669	action, atteindre, apparait, augmenter, automatiquement, cela, ceux, codage, constitue, continue (*adj.*), critique (*adj.*), déplacement, détection, détermination, écarts, échos, électronique (*s*), en (*part.*), épaisseur, fixes (*adj.*), fluorescence, fraction, guichet, hauteur, importante, indique (*v*), inférieur, les (*pron.*), limite (*s*), lorsqu', maintenant (*adv.*), navire, nominale, possibilités, résistivités, somme (*s*), stabilisateur, suffisamment, surtout, termes, vu	20

Таблица 2

Распределение словоформ по рангам, а также по частотам абсолютной (*F*) и относительной накопленной (*f**)

i	F	m	f*
1	6465	1	0.06465
2	3621	1	0.10086
3	2515	1	0.12601
4	2464	1	0.15065
5	2453	1	0.17518
6	2172	1	0.19690
7	2014	1	0.21704
8	1867	1	0.23571
9	1696	1	0.25267
10	1605	1	0.26872
11	1552	1	0.28424
12	1424	1	0.29848
13	1280	1	0.31128
14	1253	1	0.32381
15	1228	1	0.33609
16	977	1	0.34586
17	897	1	0.35483
18	797	1	0.36280
19	659	1	0.36939
20	634	1	0.37573
21	616	1	0.38189
22—23	615	2	0.39419
24	511	1	0.39930
25	501	1	0.40431
26	430	1	0.40861
27	408	1	0.41269
28	392	1	0.41661
29	346	1	0.42007
30	331	1	0.42338
31	319	1	0.42657
32—33	317	2	0.43291
34	316	1	0.43607
35—36	307	2	0.44221
37	298	1	0.44519
38	282	1	0.44801
39	278	1	0.45079
40	264	1	0.45343
41	260	1	0.45603
42	249	1	0.45852
43	234	1	0.45886
44	227	1	0.46313
45	218	1	0.46531
46—47	213	2	0.46957
48	199	1	0.46156
49	196	1	0.47352
50	195	1	0.47547
51	187	1	0.47734
52	176	1	0.47910
53	169	1	0.48079
54	161	1	0.48240
55	160	1	0.48400
56—57	157	2	0.48714
58—59	153	2	0.49020
60	152	1	0.49172
61	151	1	0.49323
62	150	1	0.49473
63	146	1	0.49619
64	145	1	0.49764
65	142	1	0.49906
66—67	141	2	0.50188
68	140	1	0.50328
69	138	1	0.50466
70	134	1	0.50600
71	131	1	0.50731
72—73	122	2	0.50975
74	121	1	0.51096
75	116	1	0.51212
76	114	1	0.51326
77	112	1	0.51438
78—80	110	3	0.51168
81	109	1	0.51877
82—84	107	3	0.52198
85	106	1	0.52304
86	102	1	0.52406
87	101	1	0.52507
88—90	99	3	0.52804
91	98	1	0.52902
92	96	1	0.52998
93—94	95	2	0.53188
95	94	1	0.53282
96	93	1	0.53375
97—98	92	2	0.53559
99	91	1	0.53650
100	89	1	0.53739
101	88	1	0.53827
102—103	87	2	0.54001
104—106	85	3	0.54256
107—108	83	2	0.54422
109—111	82	3	0.54668
112—115	81	4	0.54992
116—119	80	4	0.55312
120—121	79	2	0.55470
122	78	1	0.55548
123—125	77	3	0.55779
126—128	76	3	0.56007
129—132	75	4	0.56307
133—136	73	4	0.56599
137	72	1	0.56671

Таблица 2 (*продолжение*)

i	*F*	*m*	*f**	*i*	*F*	*m*	*f**
138—140	71	3	0.56884	325—341	35	17	0.66094
141—143	70	3	0.57094	342—356	34	15	0.66603
144	69	1	0.57153	357—369	33	13	0.67032
145—146	68	2	0.57289	370—379	32	10	0.67352
147	67	1	0.57356	380—393	31	14	0.67786
148	66	1	0.57432	394—405	30	12	0.68146
149	65	1	0.57497	406—422	29	17	0.68639
150—151	64	2	0.57625	423—439	28	17	0.69115
152	62	1	0.57687	440—464	27	25	0.69790
153—154	61	2	0.57809	465—490	26	26	0.70466
155—159	60	5	0.58109	491—515	25	25	0.71091
160—162	59	3	0.58286	516—547	24	32	0.71859
163—166	58	4	0.58518	548—572	23	25	0.72434
167—170	57	4	0.58746	573—601	22	29	0.73072
171—172	56	2	0.58858	602—628	21	27	0.73639
173—178	55	6	0.59188	629—669	20	41	0.74459
179—180	54	2	0.59296	670—714	19	45	0.75314
181—184	53	4	0.59508	715—755	18	41	0.76052
185—190	52	6	0.59820	756—786	17	31	0.76579
191—194	51	4	0.60024	787—840	16	54	0.77443
195—200	50	6	0.60324	841—897	15	57	0.78298
201—204	49	4	0.60520	898—956	14	59	0.79124
205—211	48	7	0.6856	957—1023	13	67	0.79995
212—219	47	8	0.61232	1024—1111	12	88	0.81051
220—231	46	12	0.61784	1112—1199	11	88	0.82019
232—235	45	4	0.61964	1200—1318	10	119	0.83209
236—242	44	7	0.62272	1319—1460	9	142	0.84487
243—252	43	10	0.62702	1461—1610	8	150	0.85687
253—258	42	6	0.62954	1611—1819	7	209	0.87150
259—267	41	9	0.63323	1820—2048	6	230	0.88530
268—283	40	16	0.63963	2049—2393	5	345	0.90255
284—294	39	11	0.64392	2394—2852	4	459	0.92091
295—301	38	7	0.64658	2853—3565	3	713	0.94230
302—313	37	12	0.65102	3566—4792	2	1227	0.96684
314—324	36	11	0.65498	4793—8108	1	3316	1.00000

Таблица 3

Список 105 наиболее частых слов (по убывающей частоте) в текстах по электронике

i	Слово	*F*	*f**
1	le (*art.*) [1]	13916	0.1392
2	de (*prép.*)	11273	0.2519
3	à	3400	0.2859
4	un (*art.*)	3144	0.3173
5	être (*v*)	3086	0.3482
6	et	2014	0.3683
7	il	1521	0.3835
8	en (*pron.*)	1424	0.3978
9	par	1228	0.4101
10	ce (*adj. dèt.*)	1213	0.42222
11	que (*pron.-conj.*)	1161	0.4338
12	dans	977	0.4436
13	pour	797	0.4515
14	avoir (*v*)	685	0.4584
15	on	659	0.4649
16	sur	616	0.4711
17	qui (*pron.*)	615	0.4773
18	fréquence	552	0.4828
19	pouvoir (*v*)	465	0.4875
20	ne	444	0.4920
21	plus (*adv.*)	430	0.4962
22	tension	394	0.5001
23	ou	392	0.5041
24	son (*adj.*)	373	0.5078
25	avec	331	0.5111
26	figure (*n*)	330	0.5144
27	deux	307	0.5175
28	signal	305	0.5205
29	nous	278	0.5233
30	courant (*n*)	275	0.5261
31	utiliser	246	0.5285
32	permettre	244	0.5309
33	tout (*adj.*)	242	0.5334
34	pas (*nég.*)	234	0.5357
35	circuit	232	0.5380
36	devoir (*v*)	224	0.5403
37	puissance	221	0.5425
38	autre (*adj.*)	216	0.5446
39	donner	215	0.5468
40	cas	213	0.5489
41	valeur	208	0.5510
42	certain (*adj.*)	206	0.5531
43	mesure (*n*)	201	0.5551
44	entre (*prép.*)	199	0.5571
45	grand	188	0.5589
46	système	182	0.5608
47	récepteur	178	0.5625
48	comme (*adv.*)	169	0.5642
49	tube	167	0.5659
50	bande	166	0.5676
51	obtenir	160	0.5692
52	donc	157	0.5707
53	faire	155	0.5723
54	amplificateur	153	0.5738
55	mais	152	0.5753
56	dont	151	0.5768
57	ainsi	150	0.5798

[1] le-la, l',les, au, aux, du, des; общая частота форм артикля le, la, l', les равна 10405.

Таблица 3 (*продолжение*)

i	Слово	F	f*
58	impulsion	150	0.5798
59	ligne	147	0.5813
60	celui (*celle*)	146	0.5827
61	type	144	0.5842
62	temps	141	0.5870
63	point (*nég.*)	141	
64	nombre	140	0.5898
65	très	140	
66	même (*pron.*)	132	0.5911
67	fonction	129	0.5937
68	correspondre	129	
69	élément	128	0.5950
70	gain	124	0.5962
71	effet	123	0.5975
72	chaque (*adj.*)	121	0.5987
73	sortie	117	0.5998
74	courbe	115	0.6022
75	radar	115	
76	aussi	114	0.6033
77	présenter	110	0.6044
78	image	108	0.6055
79	transistor	107	0.6065
80	vitesse	106	0.6097
81	où	106	
82	certain (*adj.*)	106	
83	condition	104	0.6108
84	rapport	103	0.6128
85	élever	103	
86	premier (*adj.*)	102	0.6138
87	variation	101	0.6159
88	lequel	101	
89	voir	99	0.6198
90	faible (*adj.*)	99	
91	émetteur	99	
92	information	99	
93	alors	98	0.6236
94	réaliser	98	
95	résistance	98	
96	possible (*adj.*)	98	
97	dispositif (*adj.*)	96	0.6275
98	niveau	96	
99	sous (*prép.*)	96	
100	falloir	96	
101	part	95	0.6284
102	onde	94	0.6322
103	trois	94	
104	transmettre	94	
105	trouver	94	

Л. И. Ешан

ЧАСТОТНЫЙ СЛОВАРЬ РУМЫНСКОГО ПОДЪЯЗЫКА ЭЛЕКТРОНИКИ

Материалом для приводимого в табл. 1 частотного списка 1000 наиболее употребительных словоформ румынского подъязыка электроники послужили научно-технические тексты на румынском языке. Общая длина проанализированных текстов составила 200 000 словоупотреблений. В табл. 1 приводятся частоты по выборке в 200 тысяч словоупотреблений. Наш выбор текстов осуществлялся в соответствии со схемой, приведенной на стр. 121 настоящего сборника. При составлении частотного словаря лексические, лексико-грамматические и грамматические омографы учитывались отдельно. В табл. 2 приводится распределение количества словоформ по частоте 42 и ниже.

Таблица 1

Список распределения словоформ в порядке убывания абсолютной частоты, а также в порядке возрастания относительной накопленной частоты для выборки в 200 тыс. словоформ

i	Словоформа	F	f*
1	de	13434	0.06717
2	în	6764	0.10099
3	se	4356	0.12277
4	cu	3916	0.14235
5	şi (*conj.*)	3846	0.16158
6	la	3124	0.17720
7	a (*art. pos.*)	2748	0.19094
8	o (*art. nehot.*)	2580	0.20384
9	care	2364	0.21566
10	mai (*adv.*)	1864	0.22498
11	un (*art. nehot.*)	1798	0.23397
12	din	1678	0.24236
13	pentru	1636	0.25054

Таблица 1 (*продолжение*)

i	Словоформа	F	f*
14	pe	1580	0.25844
15	este (*vb. cop.*)	1552	0.26620
16	să	1310	0.27275
17	prin	1060	0.27805
18	că	956	0.28283
19	nu	924	0.28745
20	este (*vb. aux.*)	840	0.29165
21	al (*art. pos.*)	796	0.29563
22	poate	772	0.29949
23	sau	736	0.30317
24	ce (*pron. inv.*)	723	0.30679
25	mare (*adj.*)	691	0.31024
26	unui (*art. nehot.*)	651	0.31350
27	fi (*a*)	604	0.31652
28	acest	594	0.31949
29	ale	576	0.32237
30	a (*prep.*)	546	0.32510
31	două	529	0.32774
32	sînt (*vb. aux.*)	504	0.33026
33	după	500	0.33276
34	și (*adv.*)	494	0.33523
35	ca (*adv.*)	458	0.33752
36	dacă	444	0.33974
37	curentul	438	0.34193
38	electroni	425	0.34406
39	iar (*conj.*)	416	0.34614
40	sînt (*vb. cop.*)	402	0.34815
41	foarte	400	0.35015
42	tensiunea	399	0.35214
43	a (*vb. aux.*)	398	0.35413
44	unei (*art. nehot.*)	390	0.35608
45—46	această, între	368	0.36168
47	va	368	0.36168
48	trebuie	361	0.36349
49	curent (*s. m.*)	353	0.36705
50	cazul	353	0.36705
51	are	348	0.36879
52—53	cînd, mică (*adj.*)	340	0.37219
54	s-a	334	0.37385
55	pot	332	0.37552
56	figură	324	0.37714
57	numai	320	0.37874
58	curentului	319	0.38033
59—60	energie, tensiune	310	0.38343
61	mult (*adv.*)	308	0.38497
62	fost (*part.*)	304	0.38649
63	astfel (*adv.*)	300	0.38799
64	frecvență	297	0.38948
65	electronii	292	0.39094
66	electronilor	287	0.39237
67—68	ca (*conj.*), electric (*adj.*)	286	0.39523
69—70	aceste, timp	285	0.39808

Таблица 1 (*продолжение*)

i	Словоформа	F	f*
71—72	dintre, sub	280	0.40088
73—74	ajutorul, au (*vb. aux.*)	268	0.40356
75	au (*vb. pred.*)	260	0.40486
76	lor (*pron. pos.*)	253	0.40613
77	însă	252	0.40739
78	figura (*s. f.*)	251	0.40864
79	unor (*art. nehot.*)	247	0.40988
80	mari	244	0.41110
81	valoarea	242	0.41231
82	într-un	240	0.41356
83	deoarece	238	0.41470
84	tubului	232	0.41586
85—86	aceasta, temperatura	229	0.41815
87	cît (*adv.*)	228	0.41929
88	adică	226	0.42042
89—90	deci, pînă (*prep.*)	224	0.42266
91	atît (*adv.*)	218	0.42375
92	caz	214	0.42482
93	atunci	209	0.42586
94	suprafața	208	0.42690
95	catod	206	0.42793
96	regiunea	204	0.42895
97	tensiunii	203	0.42997
98	mod	202	0.43098
99	circuitul	200	0.43198
100—102	cum (*conj.*), rezistența, suprafață	198	0.43495
103—104	electronice, parte	194	0.43689
105—107	datorită (*prep.*), față, vor (*vb. aux.*)	192	0.43977
108—109	electrice, potențial (*s. n.*)	189	0.44166
110	cele (*art. adj.*)	187	0.44259
111	sarcină	185	0.44352
112	face	184	0.44444
113—114	trei, tub	182	0.44626
115	ele	181	0.44716
116	ceea (*pron. dem.*)	180	0.44806
117	tip	179	0.44896
118	fiind	176	0.44984
119	numărul	175	0.45071
120	siliciu	171	0.45157
121	energia	170	0.45242
122—123	există, mici	169	0.45411
124	cele (*adj. dem.*)	168	0.45495
125	decît (*prep.*)	166	0.45578
126	lui (*art. hot.*)	164	0.45660
127	acesta	163	0.45741
128—129	asupra, exemplu	162	0.45903
130	sus	161	0.45984
131	încît	160	0.46064
132—133	temperatură, variația	159	0.46223
134	bază	157	0.46301
135	timpul	153	0.46378
136	lucru	152	0.46454

Таблица 1 (*продолжение*)

i	Словоформа	F	i*
137	aceea	151	0.46529
138	cîmp	149	0.46604
139—141	anodic, diferite, electronic	148	0.46826
142	(să)fie (*vb. cop.*)	146	0.46899
143	forma (*s. f.*)	145	0.46971
144	creşterea	143	0.47043
145	transistorului	142	0.47114
146	reprezintă	141	0.47185
147—150	creşte, frecvenţa, spre, vedere	139	0.47462
151—152	calcul, valoare	138	0.47600
153	anod	135	0.47668
154	comandă	133	0.47734
155—156	chiar, sens	132	0.47866
157—158	valori, viteza	131	0.47997
159—160	loc, volţi	130	0.48127
161—162	apare, dar (*conj.*)	129	0.48256
163—165	funcţionare, grilă, multe	128	0.48448
166—168	cîmpul, el, obţine	127	0.48639
169—172	ea, electrică (*adj.*), electrozi, unde (*conj.*)	126	0.48891
173—177	apoi, ar, (*vb. aux.*) germaniu, serie, vom	125	0.49203
178—180	număr (s. n.), strat, toate (*adj.*)	124	0.49389
181	tubul	123	0.49451
182—183	amplificare, rezistenţa	122	0.49573
184—185	acestor, depinde	121	0.49694
186—187	difuzie, stratului	120	0.49814
188—190	acelaşi, avînd, rezultă	119	0.49992
191	descărcare	118	0.50051
192—193	tensiuni, undă	117	0.50168
194—196	acestui, capacitatea, funcţie	116	0.50342
197—198	arată, cîteva	115	0.50457
199—200	doi, s-au	114	0.50571
201—203	etc., transistor, trece	113	0.50741
204—206	primul, stratul, sute	112	0.50909
207—210	înaltă, semiconductor, tuburile, unde (*s. f.*)	111	0.51131
211—214	baza (*s. f.*), fiecare (*adj. nehot.*), sistemul, termică	110	0.51351
215—218	cel (*art. adj.*), patru, trecere, următoarele	109	0.51569
219	fel	108	0.51629
220	acestei	106	0.51676
221—223	circuit	105	0.51833
224—227	colectorului, goluri, mic, sistem	104	0.52041
228—229	către, emisie	103	0.52144
230—232	ieşire, prezintă, schema	102	0.52297
233—235	grade, joncţiunii, viteză	101	0.52449
236—240	efectul, punct, putere, temperaturi, termic (*adj.*)	100	0.52699
241—243	formă, suficient (*adv.*), trecerea	99	0.52847
244—248	continuu (*adj.*), decît (*conj.*), grila, intr-o, tuburilor	98	0.53092
249—252	egală, formează, intrare, metal	97	0.53286
253	produce	96	0.53334
254—255	fără, seleniu	95	0.53429

i	Словоформа	F	i*
256—260	aplică, electron, ioni, joncţiune, scade	94	0.53064
261—263	bine, grosimea, unui (*pron. nehot.*)	93	0.53804
264—270	alte, asemenea (*adv.*), bazei, emitorului, faptul, general (*s. n.*), potenţialul	92	0.54126
271—278	atomi, frecvenţe, invers (*adj.*), magnetic (*adj.*), placă (*s. f.*), una (*pron. nehot.*), urmează, variază	91	0.54490
279—283	filament, lumină, obicei, pozitivi, purtătorilor	89	0.54712
284—285	aceeaşi, cîmpului	88	0.54800
286—291	catodice, condiţii, conducţie, constantă (*adj.*), gaz, sarcina	87	0.55061
292—297	(al) doilea, mercur, negativă, pozitivă, spaţiul, vede	86	0.55319
298—305	cauza (*s. f.*), condiţiile, ori (*s. f.*), joncţiunea, liberi, maximă, permite, vid	85	0.55659
306—307	contact, (a) doua	84	0.55743
308—309	circuitului, reţea	83	0.55826
310—316	aici, avea (a), dă, efect, intensitatea, valorile, zero	82	0.56113
317—319	alternativ (*adj.*), fluorescente, punctul	81	0.56235
320—321	corespunzătoare, interiorul	80	0.56315
322—324	cea (*art. adj.*), concentraţia, lumina	79	0.56433
325—329	am (*vb. aux.*), aşa (*adv.*), ci, lămpile, luminoase	78	0.56628
330—334	creştere, lămpi, oarecare, regiune, vapori	77	0.56821
335—338	arătat (*part.*), cît (*conj.*), sarcini, variaţie	76	0.56973
339—341	banda, elemente, realizează	75	0.57085
342—351	alimentare, aproape, atomii, caracteristicile, expresia, impurităţi, le, lui (*pron. pos.*), prea, puterea	74	0.57455
352—361	altă, barierei, catodului, cea (*adj. dem.*), corespunde, devine, diferenţa, electronului, fie (*conj.*), metalului	73	0.57820
362—363	astfel (*adj. inv.*), peste	72	0.57892
364—372	anumită, determină, dintr-un, felul, încălzire, măsură, numeşte, obţine (a), orice	71	0.58212
373—375	cinci, distanţa, găseşte	70	0.58317
376—380	lungimea, necesară, putea (a), radio, raportul	69	0.58489
381—391	apar, cel (*adj. dem.*), decît (*adv.*), (să)fie (*vb. aux.*), potenţialului, raze, rezistenţei, seama, sensul, spaţiu, unele (*art. nehot.*)	68	0.58863
392—399	asemenea (*adj. inv.*), circa, douăzeci, fenomen, înainte, ordinul, relativ (*adv.*), temperaturii	67	0.59131
400—414	aproximativ (*adv.*), bună, celulelor, cinetică (*adj.*), constituie, densitatea, face (a), fapt, fenomenul, gaze, nou (*adj.*), pozitiv (*adj.*), prima, probleme, toate (*pron. nehot.*)	66	0.59626
415—423	cazuri, celule, ei (*pron. pos.*), ionii, limită, mii, sută, transistoare, undelor	65	0.59985
424—435	alt, aluminiu, catodul, cuarţ direct (*adj.*), egal (*adj.*), factorul, golurilor, singur (*adj.*), special (*adv.*), sticlă, undele	64	0.60303

Таблица 1 (*продолжение*)

i	Словоформа	F	f*
436—445	calculul, curenți, duce, intern (*adj.*), inversă, metale, ne, procesul, semnal, zece	63	0.60618
446—451	acum, capacitate, Fermi, lucrează, mașina, metoda	62	0.60804
452—461	emisia, mașini, măsurarea, necesar (*adj.*), o (*pron. pers.*), precum, semiconductorului, siliciului, straturi, urmare	61	0.61109
462—468	aer, constă, descărcarea, domeniul, folosese, lungime, montajul	60	0.61319
469—480	(să) aibă, Celsius, colector (*s. n.*), dată, electronică (*adj.*), evident (*adv.*), ionizare, recepție, rețelei, teoria, totuși, ușor (*adv.*)	59	0.61673
481—489	află, cum (*adv.*), dată (s. f.) (dăți), funcționarea, imaginii, nici (adv.), observă, studiul, tot (*adv.*)	58	0.61934
490—497	aplicată (*p. adj.*), căderea, circuite, destul (*adv.*), lămpilor, masă (mobilă), semnalului, un (*num.*)	57	0.62162
498—506	afară, anume, dau, intră, jurul, milimetri, presiunea, relația, suprafeței	56	0.62414
507—518	ai (*art. pos.*), avem, cauză, celor (*art. adj.*), condensator, condensatorul, determinarea, ecran, mărimea, rînd, structura, utilizarea	55	0.62744
519—528	amplificatorului, căci, colector (*adj.*), constată, legătură, prosibilitatea, presiune, rămîne, respectiv (*adv.*), zona	54	0.63014
529—538	acestea, bornele, circuitele, constant, construcția, cristal, formula (*s. f.*), ionilor, procente, transistorul	53	0.63279
539—549	conform, de-a, emisiei, inverse, luminos (*adj.*), montaj, necesare, realizarea, regim, speciale, șase	52	0.63565
550—561	acțiunea, anodul, apariția, cincizeci, curenților, normală (*adj.*), obținute (*p. adj.*), printr-un, structură, termice, tipuri, utilizează	51	0.63871
562—573	ajunge, anodică, atom, atomilor, celulele, coeficientul, condiția, electrozii, naștere, nivelul, obișnuit (*p. adj.*), partea	50	0.64171
574—584	aparate, emitor, joncțiuni, necesar (*adv.*), printr-o, purtători, rezistențe, rezolvarea, străpungere, suficientă, volum	49	0.64440
585—598	acestuia, are, grilei, joase, lungul, mărirea, minoritari (*adj.*), modul, obținerea, privește, raport, redresoare (*adj.*), sistemului, spațială	48	0.64776
599—611	anumite, complet (*adv.*), descărcării, dintr-o, durata, experimental (*adv.*) important (*adj.*), liber (*adj.*), locul, mașină, momentul, produc, rezultatele	47	0.65082
612—635	acestora, amplificarea, amplificatorul, bobină, curba, curenții, diode, energiei, explică, fac, folosirea, gazului, impulsuri, înseamnă, întîia,	46	0.65634

Таблица 1 (*продолжение*)

i	Словоформа	F	f*
	mărime, măsurat (*part.*), mișcare, putem, razele, semiconductoare, stare, unde (*adv.*), urma (*s. f.*)		
636—646	dispozitivul, filamentului, micșorarea, oscilant, paralel (*adv.*), posibilă, proces, redresoare (*s. n.*), rețeaua, unu, valență	45	0.65881
647—654	ales (*p. adj.*), ambele, distribuția, funcționează, internă, îl, mașinile, reglare	44	0.66057
655—670	comparație, condensatorului, dat (*part.*), dimensiunile, directă, distanță, fenomene, experimentale, exterior (*adj.*), început (*s. n.*), întrucît, poziția, produce (a), respectiv (*adj.*), știe, trec	43	0.66401
671—681	anodice, bobina, ecuația, ei (*pron. pers.*), este (*vb. pred.*), faptului, formarea, problema, practic (*adv.*), subțire, unitatea	42	0.66632
682—708	antenă, arcul, atomul, bandă, bobinei, capacității, caracteristica, cărui, conține, cristalului, diferență, direct (*adv.*), dispozitive, frecvenței, încă, parametrilor, pláci, precizie, proprietățile, redresoarele, rolul, sa, saturație, schemă, semiconductorul, simplă, trece (a)	41	0.67186
709—730	etajului, exterior (*s. n.*), față, făcut (*part.*), folosește, folosite (*p. adj.*), grosime, ieșirea, inițiala, lampă, legate (*p. adj.*), lui (*pron. pers.*), metalul, numesc, particule, pozitive, procesului, recombinare, singură, special, tensiunile, (a) treia	40	0.67626
731—747	aparatele, căror, dăt (*p. adj.*), dispozitiv, extrem (*adv.*), măsoară, necesită, negativ (*adj.*), noi (*adj.*), ore, proporțională, puternică, radiații, reprezentată (*part.*), s-ar, tipul, volumul	39	0.67957
748—767	amplificator (*s. n.*), aparatului, deosebit (*adv.*), hidrogen, linii, mașinii, metalic (*adj.*), micșorează, moment, oscilațiile, parametrii, pămînt, radiația, realiza (a), regiunii, scop, semnalul, tensiunilor, utilizate (*p. adj.*), viteze	38	0.68337
768—792	acești, cantitatea, coloana, conductor (s. n.), contactul, continuă (*adj.*), contrar (*adj.*), control, electricitate, funcționării, iluminat (*s. n.*), introduce, înalte, joasă, masa (-solid), metalice, pămîntului, radiofrecvență, receptor (*s. n.*), structurii, total (*adj.*), uneori, urmă, variabil, vedea (a)	37	0.68800
793—816	asigură, cei (*art. adj.*), celor (*adj. dem.*), cristalină, dată (*p. adj.*), diametrul, drumul, electromagnetice, electronică (*s. f.*), electronul, element, existența, factorului, filamentul, format (*p. adj.*), importanta, intensitate, placa, răcire, regimul, scăderea, substanțe, triode, wolfram	36	0.69232

Таблица 1 (*продолжение*)

i	Словоформа	F	f*
817—837	capătă, conductibilitatea, constanta (*s. f.*), elementele, emişi (*p. adj.*), exploziv, fiecare (*pron. nehot.*), ia, influenţa (*s. f.*), îi (*pron. pers.*), materialului, mărimi, oscilator (*s. n.*), puncte, rezonanţa, schimba, serie (*a*), scurte, stabilirea, ultimul, variabilă	35	0.69599
838—862	aplicarea, caracteristică (*s. f.*), catodic, da (*a*), datele, deşi (*conj.*), diodă, ecvivalent, ecranul, era (*vb. cop.*), etaj, etajul, jumătate, luminoasă, maşinilor, moleculele, nivele, operaţii, prezenţa, puternic (*adv.*), razelor, semnale, starea, totala, ţinînd	34	0.70024
863—883	aerul, antene, bun, bune, cantitate, catodică, celelalte, datoreşte, dioda, distanţe, electrod, importante, introducerea, îşi, obişnuite (*p. adj.*) provoacă, reacţie, ridicată (*p. adj.*), simplu (*adj.*), tocmai, variaţiei	33	0.70371
884—918	alta (*pron. nehot.*), bazează, calculat (*part.*), calitate, caracteristică (*adj.*), comanda (*s. f.*), condensatoarele, cristalul, deosebire, durată, ecuaţii, experienţă, factor, fluorescent, generator (*s. n.*), i, imediat (*adv.*), impurităţilor, întreaga, întrebuinţează, linie, material, mijlociu, oscilaţiilor, prezent (*s. n.*), rezistivitatea, rol, sarcinii, schimb (*s. n.*), secundare, transformatorului, triodă, variaţii, viaţă, vitezei	32	0.70931
919—948	automată, axa, cîtiva, corospunzător (*adj.*), corp, descărcări, faza, final (*adj.*), forţă, echilibru, echilibrul, electrozilor, emisiunea, intrarea, invers (*adv.*), încălzirea, joacă, medie (*adj.*), megahertzi, metode, montajului, oscilaţii, pînă (*conj.*), plăcii, principiul, regiuni, repede (*adv.*), rezultate (*s. n.*), sensibilitatea, variaţiile	31	0.71396
949—981	aparatelor, aprindere, condensatoare, cursul, despre, diametru, efectuează, emite, faţa, fenomenele, imaginea, impedanţa, începe, întotdeauna, lămpii, limitele, litere, metodă, metri, mişcarea, negative, nivel, noduri, oxid, paragraful, principiu, putut (*part.*), radiaţie, scurt (*adj.*), scurtă, sînt (*vb. pred.*), stabilit (*part.*), transistoarelor	30	0.71891
982—1013	audiofrecvenţă, caracteristici (*s. f.*), corpuri, diferită, diferitele, diodele, emitător (*s. n.*), etaje, figurile, găsesc, într-adevăr, lipsa, miez, milimetru, molecule, moleculelor, noi (*pron. pers.*), particulelor, receptorului, redresor (*s. n.*), reducerea, ridicate (*p. adj.*), situaţie, stabilitatea, totul, transfer (*s. n.*), transistoarele, trebui (*a*), ultrascurte, următoare, următoarea, următor	29	0.72355

Таблица 2

Распределение количества словоформ по частотам при F ⩽ 28

i	F	m	i	F	m
1014—1047	28	34	1876—1980	14	105
1048—1091	27	44	1981—2100	13	120
1092—1128	26	37	2101—2253	12	153
1129—1170	25	42	2254—2415	11	162
1171—1202	24	32	2416—2602	10	187
1203—1247	23	45	2603—2840	9	238
1248—1296	22	49	2841—3129	8	289
1297—1353	21	57	3130—3510	7	381
1354—1427	20	74	3511—3952	6	442
1428—1499	19	72	3953—4551	5	599
1500—1584	18	85	4552—5337	4	786
1585—1670	17	86	5338—6568	3	1231
1671—1770	16	100	6569—8755	2	2187
1771—1875	15	105	8756—14292	1	5537

Л. А. Турыгина

ЧАСТОТНЫЙ СЛОВАРЬ АНГЛИЙСКИХ И АМЕРИКАНСКИХ ГАЗЕТНЫХ ТЕКСТОВ

В настоящей работе представлены первые результаты по исследованию статистической структуры лексики газетных текстов на английском языке. Объем выборки 100 000 словоупотреблений. Выборка охватывает основные жанры, представленные в современной английской и американской прессе. Процентное распределение текстов по тематике дано ниже.

Тематика	%
Политические новости	18
Социальная жизнь	22
Наука	8
Экономика	14
Искусство	10
Спорт	14
Светские и полицейские новости	14

При статистической группировке словоформ учитывалась лексико-грамматическая и грамматическая омонимия.

Были обследованы тексты из газет 1961—1966 гг. следующих названий: «AFL — CIO News», «Daily Express», «Daily Herald», «Daily Mail», «Daily Mirror», «Daily Worker», «Tribune», «News of the World», «Sunday Mirror», «The Daily Telegraph and Morning Post», «The New York Herald Tribune», «The New York Times», «The Times», «The Worker», «Sunday Pictorial».

Ниже приводится 500 словоформ, наиболее употребительных в нашей выборке (они занимают 78% обследованного текста).

Для удобства сопоставления наших результатов с данными зарубежных частотных словарей, которые не регистрируют никакой омонимии, в этом списке лексико-грамматические омографы не выделяются.

Статистическое распределение первых по частоте словоформ в английских публицистических текстах

i	Словоформа	F	f*
1	the	7144	0.07144
2	of	3427	0.10571
3	to	2457	0.13028
4	in	2051	0.15079
5	and	1943	0.17022
6	a	1671	0.18693
7	for	928	0.19621
8	was	908	0.20529
9	is	870	0.21399
10	that	816	0.22215
11	on	787	0.23002
12	at	674	0.23676
13	he	644	0.24320
14	with	633	0.24953
15	by	628	0.25581
16	be	618	0.26199
17	it	579	0.26778
18	an	511	0.27289
19	as	497	0.27786
20	his	472	0.28258
21	mr.	461	0.28719
22	from	421	0.2914
23	have	405	0.29545
24	has	387	0.29932
25	will	384	0.30316
26	are	378	0.30694
27	but	371	0.31075
28	said	362	0.31427
29	had	362	0.31789
30	not	352	0.32141
31	this	349	0.32490
32	been	332	0.32882
33	which	323	0.33145
34	who	295	0.3344
35	were	294	0.33734
36	they	279	0.34013
37	last	253	0.34266
38	one	249	0.34515
39	I	237	0.34752
40	their	226	0.34978
41	new	224	0.35202
42	would	221	0.35423
43	all	219	0.35642
44	when	214	0.35856
45—46	up, yesterday	202	0.36258
47—48	or, there	199	0.36656
49	more	191	0.36847
50	first	188	0.37035
51	about	186	0.37221
52	its	174	0.37395
53	no	172	0.37567

Продолжение

i	Словоформа	F	f*
54	out	165	0.37732
55	we	163	0.37895
56	other	159	0.38054
57—58	two, some	156	0.38366
59—60	year, president	150	0.38666
61—62	only, after	145	0.38956
63	her	142	
64	well	140	0.39238
65	made	137	0.39375
66—67	over, united	136	0.39647
68	into	130	
69—71	if, she, so	128	0.40161
72—74	than, now, mrs.	127	0.40542
75	British	124	0.40666
76	today	122	0.40788
77	states	120	0.40908
78	years	119	0.41027
79	him	118	0.41145
80	before	117	0.41262
81	there	112	0.41374
82—83	between, against	109	0.41592
84—85	can, London	100	0.41792
86—87	what, them	98	0.41988
88	here	97	0.42085
89	time	96	0.42181
90	under	95	0.42276
91—92	could, also	94	0.42464
93	people	93	0.42557
94—95	night, committee	92	0.42649
96	should	90	0.42831
97—98	off, house	89	0.43009
99—101	many, most, American	88	0.43273
102	may	85	0.43358
103	you	84	0.43442
104—106	our, being, since	83	0.43691
107	week	82	0.43773
108—109	day, four	80	0.43933
110—112	make, union, world	79	0.44170
113	national	78	0.44248
114—115	minister, public	75	0.44398
116	meeting	74	0.44472
117	home	73	0.44545
118—120	any, these, like	71	0.44758
121—122	do, sir	70	0.44898
123—125	per, told, such	69	0.45105
126—128	six, party, down	68	0.45309
129—130	government, man	67	0.45443
131—133	even, because, must	66	0.45641
134	did	65	0.45706
135—139	work, dr., just, market, political	64	0.46026
140—145	country, five, very, Kennedy, much, south	63	0.46404
146—147	second, back	62	0.46528

Продолжение

i	Словоформа	F	f*
148	then	61	0.46589
149—150	good, state	60	0.46709
151—154	long, get, general, through	59	0.46945
155—161	during, same, company, few, me, put, where	58	0.47351
162—164	both, those, year-old	57	0.47408
165—166	cent, my	56	0.47634
167—170	called, great, part, take	55	0.47689
171—178	chairman, council, go, negro, next, set, until, way	54	0.48070
179—181	court, June, still	53	0.48445
182	every	52	0.48497
183—185	another, end, life	51	0.48548
186—191	city, report, Soviet, two, while, went	50	0.48900
192—195	action, John, men, own	49	0.49141
196—202	leader, left, little, nuclear, Washington, West, York	48	0.49429
203—206	came, international, school, secretary	47	0.49571
207—210	asked, Britain, Lord, without	46	0.49757
211—215	business, found, far, foreign, group	45	0.49894
216—218	how, months, play	44	0.50162
219—222	conference, however, leaders, workers	43	0.50249
223—228	area, better, common, congress, Miss, street	42	0.50504
229—234	ago, labour, major, May, never, young	41	0.50794
235—240	association, members, office, program, us, white	40	0.51116
241—249	away, board, cup, days, each, industrial, later, match, miles	39	0.51467
250—262	among, communist, countries, half, high, increase, local, million, military, right, show, statement, took	38	0.51961
263—269	cause, England, lost, number, old, see, taken	37	0.52220
270—276	best, given, government, might, police, times, U. S.	36	0.52472
277—289	able, again, almost, early, final, going, held, official, race, service, st., small, war	35	0.52927
290—297	althogh, Britain's, come, federal, issue, nearly, order, top	34	0.53199
298—307	America, big, control, give, large, morning, place, though, wife, won	33	0.53529
308—315	chief, making, member, minutes, policy, possible, power, prime	32	0.53785
316—325	announced, become, children, defence, July, free, league, plan, say, third	31	0.54095
326—340	aid, agreed, de, decision, does, East, eight, former, full, further, got, money, prices, university, your	30	0.54545
341—352	bill, brought, central, common-wealth, force, help, hour, itself, officials, park, position, used	29	0.54893
353—357	air, main, reported, sales, week	28	0.55033
358—370	already, began, church, club, economic, job, press, past, question, rather, round, several, whether	27	0.55385

Продолжение

i	Словоформа	F	f*
371—390	army, case, conditions, Cuba, European, field, gave, less, nations, production, Robert, Saturday, side, situation, think, total, want, western, whose, within	26	0.55905
391—402	added, agreement, department, French, important, once, parties, short, visit, whole, why, women	25	0.56205
403—413	following, declared, George, head, look, problems, rate, seen, sent, soon, view	24	0.56469
414—435	Chinese, development, ever, figures, hand, hard, including, interest, law, March, near, negotiations, north, organisation, recent, showed, special, stock, television, theatre, unions, weapons	23	0.56974
436—463	annual, appeared, arms, certain, country, effect, enough, expected, forward, higher, Indian, know, leading, Mayor, meet, met, move, month, Moscow, nine, p. c., premier, problem, result, says, shares, to-day, victory	22	0.57590
464—499	act, Africa, behind, captain, car, chance, civil, cost, democratic, died, earlier, election, future, health, heard, himself, information, latest, line, mother, nothing, price, progress, radio, received, royal, seems, session, strong, system, talks, test, things, together, trade, industry	21	0.58346

И. А. Есенин

О ЧАСТОТНОМ СЛОВАРЕ ПОДЪЯЗЫКА СОВРЕМЕННОЙ ФРАНЦУЗСКОЙ ПРЕССЫ

На кафедре французского языка Ивановского государственного педагогического института ведется работа по составлению частотного словаря подъязыка современной французской прессы.

Подсчитываются словоформы, употребляемые в текстах, подбор которых отражает основные стили и жанры, свойственные этому подъязыку.

В качестве источников привлечены следующие издания: «L'Humanité», «L'Humanité-Dimanche», «France-Nouvelle», «Libération», «Le Petit Varois» («La Marseillaise»), «Lettres Françaises», «Vaillant», «Heures Claires», «L'Europe», «Cahiers du Communisme», «L'Avant-Garde», «La Nouvelle revue internationale», «Jeunesse du monde», «France—URSS», «Le Théâtre et le cinéma», «Journal des jeunesses musicales».

Общая длина обследованных текстов составляет 120 000 словоупотреблений.

Ниже приводится частотный список 500 наиболее употребительных словоформ.

Частотный список словоформ [1]

i	Словоформа	F	f	f*
1	de (*prép.*)	6462	0.0538	0.0538
2	le (*art.*)	3272	0.0273	0.0811
3	la (*art.*)	2970	0.0247	0.1058
4	et	2441	0.0203	0.1261
5	les (*art.*)	2052	0.01715	0.14325
6	des	1662	0.01385	0.15710
7	est	1320	0.01100	0.16810
8	un (*art.*)	1240	0.01032	0.17842

[1] В подсчете словоформ принимали участие работники кафедры французского языка Ивановского педагогического института им. Д. А. Фурманова Ю. А. Титова, Т. В. Зотова, Л. Ф. Карточкина, В. А. Грязнова, Л. И. Тюрина, Т. Я. Камаргина, А. П. Сорокина и Т. В. Куликова.

Продолжение

i	Словоформа	F	f	f*
9	une (*art.*)	1222	0.01018	0.18860
10	du	1217	0.01013	0.19873
11	que (*pron.*)	1213	0.01012	0.20885
12	dans	1115	0.00929	0.21814
13	il	1097	0.00913	0.22727
14	à	1072	0.00894	0.23621
15	en (*prép.*)	962	0.00800	0.24421
16	ne	907	0.00756	0,25177
17	on	859	0.00715	0.25892
18	qui	758	0.00632	0.26524
19	au	731	0.00626	0.27250
20	se	722	0.00602	0.27852
21	pour	635	0.00529	0.28381
22	ce (*pron.*)	598	0.00498	0.28879
23	par	579	0.00482	0.29361
24	ce (*adj.*)	542	0.00452	0.29813
25	nous	507	0.00422	0.30233
26	plus	493	0.00411	0.30644
27	faut	467	0.00389	0.31033
28	pas (*adv.*)	459	0.00382	0.31415
29	mais	457	0.00381	0.31796
30	en (*pron.*)	456	0.00380	0.32176
31	a	435	0.00363	0.32579
32	sur	434	0.00361	0.32940
33	l'(*art.*)	413	0.00344	0.33284
34	aussi	404	0.00337	0.33621
35	leur (*adj.*)	398	0.00331	0.33952
36	avec	390	0.00325	0.34277
37	son (*adj.*)	386	0.00322	0.34599
38	lui	381	0.00317	0.34916
39	je	374	0.00315	0.352275
40	aux	373	0.00311	0.355385
41	d'(*prép.*)	331	0.00276	0.358145
42	cette	325	0.00271	0.360855
43	sont	313	0.00261	0.363465
44	comme (*conj.*)	309	0.00257	0.366035
45	ils	287	0.00239	0.368425
46	bien (*adv.*)	273	0.00227	0.370695
47	elle	259	0.00216	0.372855
48	ces	250	0.00208	0.374935
49	sa	228	0.00190	0.376835
50	si (*conj.*)	224	0.001865	0.378700
51	le (*pron.*)	218	0.001815	0.380515
52	être	212	0.001778	0.382293
53	autre (*adj.*)	194	0.00162	0.383913
54	faire	186	0.00155	0.385463
55	tous (*adj.*)	183	0.001525	0.386988
56	tout (*pron.*)	180	0.00150	0.388488
57—58	contre, ont (*prép.*)	174	0.00145	0.391388
59	après	173	0.00144	0.392828
60	été (*v.*)	172	0.00143	0.394260
61	deux (*adj.*)	163	0.00136	0.395620

Продолжение

i	Словоформа	F	f	f*
62	avait	162	0.001350	0.396970
63	entre	161	0.001340	0.398310
64	où	159	0.001325	0.399635
65	même (*adj.*)	157	0.001308	0.400943
66	pays	152	0.001268	0.402211
67	vous	148	0.001234	0.403445
68—69	étais, encore	142	0.001182	0.405807
70	là	137	0.001141	0.406948
71	non	127	0.001058	0.408006
72	notre	125	0.001040	0.409046
73	était	121	0.001008	0.410054
74	y (*adv.*)	120	0.001000	0.411054
75—76	peut, toutes (*adj.*)	119	0.000992	0.413038
77	vie	117	0.000974	0.414032
78	gouvernement	116	0.000966	0.414998
79—81	grand (*adj.*), prix (*nom.*), sans	113	0.000942	0.417824
82—83	ceux (*pron.*) toujours	112	0.000933	0.419690
84	n'	109	0.000907	0.420597
85	temps	108	0.000900	0.421497
86—88	dit, socialiste (*adj.*), toute (*adj.*)	107	0.000892	0.424163
89	fait (*v.*)	105	0.000875	0.425038
90—92	s', bon (*adj.*), ses	103	0.000857	0.427609
93—95	alors, socialisme, sous	102	0.000850	0.430159
96—98	ans, me, monde	100	0.000833	0.432658
99	cela	99	0.000825	0.433483
100	fait (*nom.*)	98	0.000816	0.434299
101—102	aujourd'hui, avant	95	0.000792	0.435883
103—104	beaucoup, dont	94	0.000783	0.437449
105	avoir (*v.*)	90	0.000750	0.438199
106—107	ainsi, celui	89	0.000741	0.439681
108—113	devant, grande, guerre, homme, père, peuple	88	0.000733	0.444379
114	sûr	87	0.000725	0.445104
115	peu	85	0.000709	0.445813
116—118	c', pouvoir (*v.*), système	84	0.000700	0.447913
119	jour	83	0.000692	0.448605
120	jamais	81	0.000675	0.449280
121	mon	79	0.000658	0.449938
122—123	jusque, liberté	78	0.000650	0.451238
124—127	français (*adj.*), heures, moins, quand	76	0.000633	0.453770
128—131	car, déjà, fois, eux	73	0.000608	0.456202
132	hommes	72	0.000600	0.456802
133—134	fut, théâtre	70	0.000584	0.457970
135—136	grève, la (*pron.*)	69	0.000575	0.459120
137—139	cours, nouvelle (*adj.*), quelques (*adj.*)	68	0.000566	0.460818
140—142	comment, premier (*adj.*), vers	67	0.000558	0.462492
143—146	cas, film, jours, place	66	0.000550	0.464692
147—148	dire, œuvre	65	0.000542	0.465776
149—152	cet, chez, donc, politique (*nom.*)	64	0.000533	0.467908
153—156	aucun (*pron.*), petit (*adj.*), seulement, vieux (*adj.*)	63	0.000525	0.470008

Продолжение

i	Словоформа	F	f	f*
157—158	action, soit (*v.*)	62	0.000516	0.471040
159—160	an, nos	61	0.000508	0.472056
161—165	ailleurs, classe (*nom.*), dernier (*adj.*), pièce, vient	60	0.000500	0.474556
166—169	lutte (*nom.*), moi, nombre, Paris	59	0.000492	0.476524
170—171	économique, jeune	58	0.000483	0.477490
172	Etat	57	0.000475	0.477965
173—175	ici, journal, tant	56	0.000466	0.479363
176—178	cause (*nom.*), mouvement, partie (*nom.*)	55	0.000458	0.480737
179—183	affaires, état, française, ici, production	54	0.000450	0.482987
184	va	53	0.000442	0.483429
185—187	bonne (*adj.*), leurs, nationale	52	0.000433	0.484728
188—195	celle, chaque, communiste (*adj.*), cuisine, interne, près (de), selon, seule	51	0.000425	0.488128
196—198	mal (*adv.*), ordre, sera	50	0.000416	0.489376
199—200	maintenant, histoire	49	0.000408	0.489784
201—205	côté, effet, sommes, part (*nom.*), scène	48	0.000400	0.492192
206—210	certains, écrire, monsieur, pendant, union	47	0.000392	0.494152
211—216	années, elles, hier, plan, question, situation	46	0.000383	0.496449
217—225	ai, but, chose, fait (*v.*), général (*adj.*), grands, quatre (*adj.*), République, tour	45	0.000375	0.499824
226—233	année, conditions, depuis (*prép.*), développement, gaulliste (*adj.*), gens, fin, jeunes (*adj.*)	44	0.000366	0.502752
234—236	avons (*v.*), femme, même (*adv.*)	43	0.000358	0.503826
237—242	depuis (*adv.*), doit, mois, nom, petite (*adj.*), telle (*adj.*)	42	0.000350	0.505926
243—247	assez, bureau, ni, président, résulte	41	0.000342	0.507636
248—255	cinéma, cinq (*adj.*), femmes, loin (*adv.*), mesure, nouveau, organisation, pas (*nom.*)	40	0.000333	0.510300
256—262	auteur, ça, doute (*nom.*), exploitation, mes, millions, totale	39	0.000325	0.512575
263—269	droit (*nom.*), jeu, lorsque, ma, pourquoi, recherche, semaine	38	0.000317	0.514794
270—277	aller (*v.*), compte (*nom.*), dimanche, mauvais, moment, pourtant, première, soir	37	0.000308	0.517258
278—289	acheter, beau (*adj.*), commune (*adj.*), défense, haut, maison, mesures, mettre, ouvriers, parce que, prendre, ville	36	0.000300	0.520858

Продолжение

i	Словоформа	F	f	f*
290—299	ait, famille, font, demain, matin, ministre, ouvrière (*adj.*), problème, souvent, tête, titre	35	0.000292	0.524070
300—309	assemblée, autant (*nom.*), ici (*adv.*), cœur, mieux (*adv.*), national, nuit, passer, quelque (*adj.*), régime	34	0.000283	0.526900
310—316	cependant, commun, domaine, donner, eu, grandes, vite (*adv.*)	33	0.000275	0.528825
317—331	belle (*adj.*), besoin, chef, construction, enfin, étaient, heure, humanité, longtemps, paix, parfois, personnel (*adj.*), pris, sens, soviétiques (*adj.*)	32	0.000267	0.532830
332—343	américain (*adj.*), coupe, eau, journée, lieu, milliards, notamment, parmi, projet, téléphone, texte, vingt (*adj.*)	31	0.000258	0.535926
344—360	budget, centre, démocratie, Etats, juin, forces, majorité, mère, minute, moyen, niveau, novembre, peuples, raison, société, surtout, voyage	30	0.000250	0.540176
361—369	air, allemand (*adj.*), avaient, base, bourgeoisie, juste, maître, socialistes (*adj.*), terrain	29	0.000242	0.542354
370—388	autres (*adj.*), chambre, communistes (*nom.*), coup, enfants, esprit, francs (*nom.*), Gaulle(de), hôtel, laquelle, mondiale, nombreux, obtenir, passé (*nom.*), petits (*adj.*), pourrait, prochain, région, soviétique (*adj.*)	28	0.000233	0.546781
389—409	aucune (*adj.*), affaire, armée (*nom.*), avenir, commerce (*nom.*), face, façon, jeunesse, livre (*nom.*), mort (*nom.*), populaire (*adj.*), possible (*adj.*), période, roman (*nom.*), siècle, solidarité, tel (*adj.*), télévision, terre, visage, vue (*nom.*)	27	0.000225	0.551506
410—422	actuellement, combat, conseil, dernière (*adj.*), général (*nom.*), manifestation, moyens (*nom.*), page, parents, pacifique (*adj.*), quel, service, tenir	26	0.000217	0.554327
423—440	abord(d'), aide (*v.*), bout, camp, constitution, difficile, écrit (*v.*), idée, importance, lampion, malgré, parler, personne (*nom.*), rendre, six (*adj.*), sujet, unité, veut	25	0.000208	0.558071

Продолжение

i	Словоформа	F	f	f*
441—462	article, celles, certaines, deuxième, édition, force (*nom.*), gros (*adj.*), indépendance, jardin, marché (*nom.*), occasion, octobre, orchestre, outre, passage, programme, revendications, rue, serait, syndicats, tard, te	24	0.000200	0.562471
463—491	agit, Algérie, aurait, camarade, campagne, caractère, cent (*adj.*), condition, congrès, devoir (*nom.*), dix (*adj.*), écrivain, fille, vit, foi, Français, groupe, intérêts, libre (*adj.*), musique, nouveaux, parti (*nom.*), particulier (*adj.*), progrès, professeur, plusieurs (*adj.*), sort (*v.*), simplement, succès	23	0.000192	0.568039
492—500	accord, agricoles, ami, amies, amour, août, appartement, autour(de), ayant	22	0.000183	0.569686

Л. А. Турко

ЧАСТОТНЫЙ СЛОВАРЬ РУССКОЙ РАЗГОВОРНОЙ РЕЧИ

Частотный словарь современной русской разговорной речи составлен путем обработки записей на магнитофон непринужденной, необработанной, не рассчитанной на запись разговорной речи. Записана речь более чем 20 человек, большинство из которых имеет высшее образование.

Объем записи составляет 50 000 словоупотреблений.

Лексические, лексико-грамматические и грамматические омонимы (*омографы*) отмечались как разные словоформы.

В прилагаемом списке приведены 1172 словоформы, которые встретились не менее пяти раз во всех обследованных текстах.

Для слов, имеющих грамматические омонимы, в словаре даются следующие сокращения:

и, р, д, в, т, п — начальные буквы названия падежей.
м — мужской род.
ж — женский род.
ср. — средний род.
всп. — глагол, играющий роль вспомогательного.
вв — вводное слово.
прит. — притяжательное местоимение.
ч — частица.
н — наречие.
с — союз.

В таблице цифровые индексы при словоформах (например, у^2) указывают на номера омонимов по «Словарю русского языка» С. И. Ожегова (М., 1961).

В табл. 1 приведены цифры, показывающие распределение 1172 словоформ по частотам, а табл. 2 иллюстрирует зависимость между номером словоформы и ее накопленной относительной частотой.

Распределение 1172 наиболее частых словоформ

i	Словоформа	F
1	не	1723
2	я	1318
3	а[1]	885
4	в	840
5	у[2]	763
6	и[1]	674
7	ну	603
8	так	531
9	нет	507
10	он	495
11	что[2]	489
12	да[1]	488
13	на[1]	487
14	ты	481
15	вы	458
16	как	441
17	это *и*	409
18	она	393
19	с	371
20	что *и*	366
21	вот	359
22	мне *д*	330
23	там	289
24	меня *р*	225
25	очень	215
26	сейчас	206
27	надо[1]	205
28	мы	198
29—30	же[2], ничего	197
31	всё *и*	196
32	ещё	183
33	знаю	172
34—35	они, уже́	171
36	к	165
37	бы	164
38—39	сегодня, хорошо	161
40	если	155
41	и[2]	151
42—43	вам, но[1]	150
44—45	за, что *в*	146
46	когда	144
47	вас *р*	141
48—49	вообще, по	137
50	нас *р*	134
51	есть[2]	125
52	только	124
53	меня *в*	123
54	конечно	118
55	Вера	117
56	здесь	116

i	Словоформа	F
57	его *в м*	113
58	тебе *д*	112
59	говорит	111
60—62	даже, то *с*, это *в*	107
63	*а*[2]	105
64	просто	101
65	знаешь	100
66	можно	99
67—69	её *в*, сколько, тоже	94
70	вчера	93
71	Лиля	92
72—73	ему *д м*, потóм	91
74	это *ч*	90
75	всё *в*	89
76—77	или, наверно	87
78	ли	86
79	говорю	84
80	от	80
81—82	*было*, Фёдоровна	79
83—84	будет, Лидия	78
85—86	Александровна, до	77
87	будет *всп.*	76
88—89	вас *в*, Константин	75
90—92	где, Константинович, тебя *р*	74
93	все *и*	72
94—96	знаете, почему, такая	70
97	было *всп.*	69
98—99	ладно *ч*, ой	68
100	тогда	67
101—102	без, пожалуйста	66
103	понимаешь	65
104—105	нам, него *р м*	63
106—108	был, кто, потому	62
109—110	для, ни	60
111—112	много, такой *и*	59
113—114	два *в*, может *всп.*	58
115	могу *всп.*	57
116	чтоб	56
117—122	больше, ведь, из, них *р*, совсем, тут	54
123—124	лучше, Минна	53
125—126	нужно, сказать	51
127—128	Мироновна, себе *д*	50
129—132	неё *р*, общем *п ср.*, чего *р*, эта	49
133—134	значит *вв.*, спасибо	48
135—136	была, этот *и*	47
137—141	быть, ей *д*, куда, может, хоть	46
142—145	какой *и*, минут *р*, по-моему, туда	45
146—151	во, дело *и*, правда, прямо, равно́, сюда	44
152—159	буду *всп.*, время *в*, говорят, его *прит.*, раз[1] *в*, себя *в*, такое *и*, чтобы	43
160—162	о[1], плохо, после	42
163—168	всё-таки, давно, двадцать *и*, сказала, такие *и*, чего *н*	41

Таблица 1 (*продолжение*)

i	Словоформа	F
169	зачем	40
170—176	всегда, знает, значит, какая, со, теперь, три *в*	39
177—182	видела, завтра, их *в*, один *и*, сказал, совершенно	38
183—184	пока, человек *и*	37
185—189	делать, день *в*, ма́ло, одна, через	36
190—194	дома *н*, никогда, опять, хочу *всп.*, эту	35
195—196	придёт, сама	34
197—198	раньше, этого *р ср.*	33
199—203	была, господи, Люся, ним *т м*, скоро	32
204—206	два *и*, сам *и*, тебя *в*	31
207—210	ага́, подожди, понимаю, сразу	30
211—215	были, должна, люди, нравится, слушай	29
216—219	всё[1], мама, пять *в*, таку́ю	28
220—226	более, денег *р*, деньги *в*, есть[1], лет *р*, нету, чёрт *и*	27
227—234	был всп., видел, идти, интересно, кажется, нас *в*, откуда, уж[2]	26
235—242	думаю *ж*, здо́рово, какие *и*, как-то, сделать, часа́ *р*, часов[2] *р*	25
243—255	будем *всп.*, времени *р*, главное *и*, могу, мой *и*, нельзя, пойду, получается, работы *р*, раз[2], рублей *р*, то *и*, ужас *и*	24
256—259	вместе, дня *р*, пошла, что-нибудь *в*	23
260—267	года *р*, думала, его *р м*, наверное, немножко, никак, пусть, са́мое *и*	22
268—277	бывает, вышла, идёт, кого *р*, Люда, несколько, причём, три *и*, что-то *н*, этот *в м*	21
278—294	будете *всп.*, взять, говорить, домой *н* кому, месяца *р*, мной, наоборот, никто, под, порядке, пришёл, пришла, работает, скажу, хочется *всп.*, что-то *и*	20
295—307	десять *в*, какое *и*, какой-то *и*, Николаевич, перед, правильно, работать, работу, сорок *и*, ходить, хочу, хуже, этом *п м*	19
308—320	Арик *и*, *б*, Верочка, взяла, вся, давай, девочки *и*, девчонки *и*, день *и*, муж, приятно, того *р ср.*, этом *п ср.*	18
321—339	будешь *всп.*, быть *всп.*, говорили, говорите *изъяв.*, две *и*, ко, людей *р*, ней *т*, одну, писать, представляешь, разве, рано, случае, хорошая, хотела *всп.*, чем *т*, четыре *в*, ясно	17
340—362	были *всп.*, быстро, видишь, говорил, давайте *всп.*, думаешь, жалко, живёт, звонила, им *д*, копеек *р*. Лю (Люся), месте, неудобно, обязательно, посмотреть, пускай *ч*, пять *и*, стала *всп.*, сто́ит, стои́т, хватает[2], чем *с*	16
363—390	боже, вами, все *в*, говорила, говоришь, Григорий, действительно, её *прит.*, жена, каждый *в*, можете *всп.*, насчёт, неделю, некому, немного, нечего[1], о[2], оно, понимаете, при, ра́за[1] *р*, сами *и*, спать, точно, утром *н*, хватит[2], четыре *и*, эти *в*	15
391—411	ваш *и*, вообще-то, всего *р ср.*, должен, жить, здравствуйте, знала, какая-то, конце, Лилечка, Нина, нормально, около, поздно, представляете, свои *в*, слушаю, столько, такое *в*, хотя *с*, этим *т ср.*	14

Таблица 1 (*продолжение*)

i	Словоформа	F
412—433	вечером *н*, вижу, дай, друг[2], комнату, мать *и*, мои *и*, отдать, получилось, посмотри, поэтому, ребёнок, Серёжа, скажет, стол *в*, субботу, телефон *и*, то *в*, товарищи, тот *и*, хочешь, чём	13
434—477	большое *и*, будто, будьте *всп.*, весь *и*, вон[2], всего, всех *р*, вышел, далеко, забыла, идите, иногда, конца, купила, купить, Людка, между, мой, например, некогда, пальто, понравилось, почти, работа, раз[1] *р*, раз[3], роднуля, рубля *р*, самый *и*, сидит, скажи, сказали, слово *и*, спросить, сто *и*, такие *в*, тем *т ср.*, тобой, трудно, угодно, хороший *и*, хотите, час *в*, честное *и*	12
478—516	взял, видели, видимо *вв.*, видно, во-первых, всю, всяком *п м*, давай *всп.*, дальше, женщина, знать, иди, кого *в*, кто-то, люблю, Мария, месяц *в*, могла *всп.*, можешь *всп.*, никуда, Ниночка, обычно, один *в*, позвонить, посмотри́те, про, пятницу, скажите, смотреть, сто *в*, странно, таких *р*, ходит, хочется, чему, чуть, шесть *и*, шесть *в*, этой *р*	11
517—564	ах, ваша, видите, возьму, воскресенье *в*, вроде, глаза *в*, год *в*, голова, дайте, долго, зарядку, Иванович, какую-то, кем, кино, книга, кто-нибудь, куда-нибудь, куда-то, Лиль *и* (Лиля), любит, место *в*, могут *всп.*, мужчина, нашли[1], никаких *р*, ними, осталось, отпуск *в*, пахнет, пишет, погода, пор *р* (пора́), посмотрю, разу[1] *р*, сделала, сидеть, снова, сожалению, спросила, страшно, твоя, ужасно, хорошего *р ср.*, часто, чувствую, Юрка	10
565—644	абсолютно, безобразие *и*, болит, больша́я, будут *всп.*, Верка, возьми, волосы *и*, восьмого *р ср.*, выбросить, где-то, глаза *и*, говори, году *п*, голову, грязный *и*, дам, двадцать *в*, две *в*, должно, ездить, женщины *и*, живут, замуж, звонить, идём, известно, имею, институте, интересует, как-нибудь, какую, красиво, кроме, легче, летом *н*, Лидии *р*, Лилька, мамочка, мне *н*, над, найти[1], народу *р*, настроение *и*, нашла[1], нему *д м*, нужна, об, особенно, письмо *в*, пить, пойдём, пойдёшь, помню, понять, почему-то, прелесть *и*, приходи́те, пришли (прийти), прочным *т ср.*, пятнадцать *в*, работе *п*, ребята, свиданья *р*, себя *р*, семь *и*, Серёжка, сидела, смотри, смысле, собой, стал *всп.*, тридцать *в*, трубку, удовольствием, ума, ушла, ходила, хочешь *всп.*, эти *и*	9
645—729	Александр *и*, Арику *д*, больницу, ваше *и*, возьмите, Гена, Гофмана *р*, даёт, дали, даст, дают, двенадцать *и*, делает, десять *и*, дети *и*, довольна, довольно, должны, друга[2] *р*, думал, дурак, заниматься[1], звонил, знали, знают, идут, иметь, каком *п м*, которая, купили, Ленка, Марик, минутку, мол[2], молоде́ц, молодой *и*, Москве *п*, музыка, написано, недавно, нравятся, нужен, нужны, одно *в*, отец, оттуда, первый *и*, позвоню, понятно, похоже, пошёл, представляю, прийти, приехала, принёс, разница, рядом, *н*, Саша, свой *в*, сделали, сестра, слышно, смешно, смотри́те, смотришь, смотрю,	8

Таблица 1 (*продолжение*)

i	Словоформа	F
	сначала, столе, стыдно, счёт *в*, считается, сын, туфли *и*, тяжело, уехал, умер, устала, ушёл, Фёдоровны *р*, хотела, чего-то *н*, человек *р*, чёрту, что-нибудь *и*, шестьдесят *и*	
730—825	берёт, брать, будут, вдвоём, весело, весь *в*, во (вот), Вовка, возможно, восемь *в*, врач, всеми, всяких *р*, выхо́дит, Ге́рман, голос *и*, давайте, дал, дать, делается, делаешь, делаю, делают, деньги *и*, детей *р*, добрый *и*, дом *и*, доме, дорогая, другая, её *р*, Женя, заболел, заболела, знаем, пди *всп.*, из-за, институт *в*, искать, история, каждый *и*, какие-то *и*, квартиру, комната, который *и*, крайней *д*, красивая, Кутепов, Лиля *и вд.*, маме *д*, меньше, минуточку, моей *р*, можем *всп.*, Муза, наши *и*, него *в м*, нехорошо, обеда *р*, особенного *р ср.*, парень, первый *в*, платить, пойдёт, получает, приехал, приходи, приходила, простите *повел.*, пятнадцать *и*, рада, рождения *р*, ругаться, руку, сверху, своими, сделаю, сильно, синие *и*, следующий *в*, случай *в*, слышала, смотрела, специально, стало *всп.*, считает[1], считаю[1], театр *в*, телефон *в*, товарищ, том *и ср.*, туфли *в*, улице *п*, ходил, хотите *всп.*, целый *в*	7
826—954	алло, Анатольевна, Арика *р*, болеет, большой *и*, боюсь, будем, ваши *и*, вдруг, весна, взяли, возраста *р*, вопрос *и*, вполне, город *в*, двух *р*, девйца, дела *и*, делал, десяти *р*, директора *р*, директору, дней *р*, доволен, дрянь *и*, ездил, ерунда, ехать, жаль, женился, женой, Женька, жизнь *и*, забыл, занято,[2] идёшь, иду, извините *повел.*, именно, иначе, инженер, их *р*, их *прит.*, командировку, комнате *п*, концов, копейки *р*, мамаша, месяц *и*, мешает,[1] многие *и*, мог *всп.*, мужем, назад, написал, написать, неважно, ней *р*, некоторые *и*, неприятно, никого *р*, ноги *и*, ноги *в*, ночью *н*, одной *р*, одной *д*, оказывается, отдам, отдела *р*, отдыха *р*, отсюда, первая, пешком, платье *в*, поехать, позвони, пойдём *всп.*, пойду *всп.*, половина, пол-одиннадцатого *р м*, полтора *в*, полу́[1] *н*, получить, поняла, попросить, почитать,[2] придётся *всп.*, примерно, противная,[2] пятьдесят *в*, рассказывал, решила, рубль *и*, рубля *в*, рукой, сделаешь, Сергей, симпатичная, скажешь, скорей, слава, солнышко *и*, состояние *и*, стали *всп.*, старый *и*, стол *и*, стороны *р*, стоял, та, телевизор *в*, тем *с*, тётка, тихо, трамвае, тринадцать *и*, узнать, фамилия, ходили, хотел *всп.*, чего-то *р*, чёрные *и*, читала, читали, что-то *в*, шкафу́ *п*, шла, шуток *р*, этих *р*, Юра	6
955—1172	бегать, богу, бросьте, будете, буду, Васильевич, вернее, вероятно, Веры *р*, вещей, вещи *в*, вещь *и*, вид[1] *и*, водой, волосы *в*, восемь *и*, врачу, время *и*, всех *в*, вторник *в*, выйти, галстук *в*, гимнастику *в*, гладить, говоря, горячие *и*, грязные *и*, да[3], давать, дверь *в*, двести *и*, девять *и*, дело *в*, длинная, днём *н*, договорились, дом *в*, дорогие *и*, думает, ела, жду, жену, жизни *р*, жизни *п*, забрать, забываю, зависит, зайти, закрой, закры-	5

Таблица 1 (*продолжение*)

i	Словоформа	F
	вается, занят[2], заняты[2], зарядка, звонили, звонят, зря, зубы *в*, Игнат, института *р*, интересный *и*, какого *р м*, какое-то *и*, какой *п*, карточку, квартира, квартире *п*, ключи[1] *в*, книгу, комнат, коридоре легко, лежал, лёг, лю́бите, магазине, мальчишка, материал *и*, маю, менее, мере *д*, могли *всп.*, моего *р м*, мой *в*, молчит, мужчин *р*, наверх, надолго, нами, насколько, настолько, начинается, неё *в*, ней *д*, нему *д ср.*, ненормальный *и*, ниже, никакой *р*, никому, ножницы *и*, номер *в*, обе *и*, он-то, оставь, остаться, отдаю, отпуске, очередь *и*, ощущение *и*, папа, паразиты *и*, первого *р м*, передайте, платье *и*, плохая, подумать, поёт, пожалуй *вв.*, позже, пойти, пол[1] *и*, полная, получил, получка, помните *изъяв.*, понедельник *в*, понравится, пора́, поражаюсь, посмотри, поступила, по́том *т*, почерк *и*, почём, пошла *всп.*, права́, прекрасно, привыкли, приду, приедет, приезжал, приехать, принести, принесу, приходить, приятный, продала, продают, пройдет, противно[2], процентов, прошу, пятнадцати *р*, пятьдесят *и*, раз[1] *и*, ровно, рук *р*, руки *и*, руки *в*, садись, садитесь, само *и*, Саше *д*, свете[2], свои *и*, своим *т*, *м*, сдал, сдать, сделайте, сделал, семь *в*, серьёзно, сзади, сиди, сижу, сих *р*, сказано, следует, случилось, смеются, совесть *и*, согласна, солнце *и*, со́рок *в*, сперва, старая, старше, стирала[1], сто́ит *всп.*, стояли, сходим[1], такой *р*, такой *в*, твои *и*, те *в*, терпеть, убрать, уверена, удобно, умеете *всп.*, уходить, ухожу, философию, ходят, хорошие *и*, хотели *всп.*, хочет *всп.*, целая, чёрта *р*, четверти *р*, читать, чувствует, чувствуется, чудесно, шестого *р ср.*, шли, этого *в м*, этой *п*, эх	

Таблица 2

Относительная накопленная частота, в %

i	$f*100$	i	$f*100$	i	$f*100$
1	3	75	42	450	64
5	11	80	43	500	65
10	17	100	46	550	66
15	21	125	48	600	67
25	28	150	51	650	68
30	30	200	55	700	69
35	32	250	57	750	69.5
40	34	300	59	800	70
50	37	350	61	900	72
55	38	400	63	1000	73
60	39				

ОБРАЗЦЫ ЗАПИСЕЙ РАЗГОВОРНОЙ РЕЧИ

I. Разговор в трамвае

1-й. Мы опоздаем.

2-й. Ты думаешь?

1-й. Трамваи не идут.

2-й. Да, очень медленно.

1-й. Посмотри сколько времени.

2-й. Пятнадцать.

1-й. Ну еще ничего.

2-й. Если успеем, то впритирочку.

1-й. Опаздываешь, Борис?

2-й. Нет.

1-й. Что пишешь?

2-й. Тайна, покрытая мраком.

1-й. Погодка какая. Устойчивый антициклон. Если бы летом был такой, представляешь, месяц стояла бы жара. Летом такого не будет. Небо чистое, ветра нет, высокое давление. Что ты пишешь?

2-й. Слушай, это любопытство или любознательность? Как это классифицировать? . . . Читаешь сейчас что-нибудь?

1-й. Куда там. И в кино я уже забыл, когда был. Основной источник, ну, как бы тебе сказать, источник чувственной моей жизни — это радио и телевидение.

II. Разговор в учреждении

1-й. Ниночка, а какой вам цвет больше нравится?

2-й. Для свитера, конечно, потемнее: серый, бордовый. Со стоечкой. Шерстяной.

1-й. Вы знаете, я не знала, у нас был большой выбор.

3-й. У вас там хороший магазин, Минна Мироновна.

1-й. Я туда захожу после работы. А там были голубые. Люсин цвет.

2-й. Да, это люськин цвет.

3-й. Но Люсе будет нехорошо, там рукава короткие.

1-й. Главное, я иду с мужем, он крепко меня держит, чтобы я в магазин не зашла. . . Потом там был такой темно-зеленый цвет, темный-темный, полосатенький воротничок, наподобие этого.

III. Разговор в учреждении

1-й. «Удивительное рядом» видели? Это как раз замечательная вещь. О явлениях природы, которых мы не замечаем. О пингвинах там очень хорошо. Фильм исключительно хорошо сделан. Причем без журнала идет. Причем очень хорошая передача цвета, в чем я убедился, когда показывали выставку собак. Там и мухи, там и змеи, там черт-те кто, например китайские кузнечики вместо канареек.

2-й. Кузнечики?

1-й. Ну, стрекочет, как сверчок. Потом канареек показывают, как их обучают петь.

3-й. Вы о чем?

1-й. «Удивительное рядом». Замечательный фильм.

М. В. Данейко, Л. Е. Машкина, О. А. Нехай, В. А. Соркина, А. Н. Шаранда

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ЛЕКСИЧЕСКОЙ ДИСТРИБУЦИИ СЛОВОФОРМЫ

При описании лексики языка или его разновидности с помощью частотных словарей не учитывается по крайней мере около 20% информации о структуре языка — информации, связанной с сочетаемостью (дистрибуцией) слов в тексте. Восполнить этот пробел можно путем статистического исследования словосочетаний. Правда, полная комбинаторика сочетаний настолько велика, что трудно получить достоверные результаты на обозримом материале даже при условии использования машины.

Поэтому, обращаясь к статистике словосочетаний, необходимо ввести некоторые существенные ограничения.

Первое ограничение состоит в том, что статистическое исследование словосочетаний осуществляется в выборках текстов совершенно определенной узкой тематики (мы будем обозначать эти выборки термином «подъязык»).

В настоящей статье дается статистика словосочетаний в следующих подъязыках:

1) английском подъязыке радиоэлектроники;
2) немецком подъязыке радиоэлектроники;
3) немецком подъязыке публицистики.

Подъязыки радиоэлектроники охватывают тексты по следующим разделам:

А. Теоретические основы электроники:

1) электронные и ионные процессы в газах и вакууме;
2) электронная и ионная эмиссия;
3) ток в полупроводниковых и фотоэлектрических материалах;

Б. Основные электронные и ионные приборы:

1) электронные и ионные приборы;
2) полупроводниковые и фотоэлектрические приборы.

В. Электроника в схемах и устройствах:

1) радиотехнические устройства;
2) вычислительная техника и различные кибернетические устройства;
3) связь, телевидение, радиолокация;
4) автоматика и телемеханика

(ср. в этой связи статью П. М. Алексеева, напечатанную в настоящем сборнике, стр. 121).

Подъязык публицистики включает тексты по следующей тематике:

1) официальные сообщения;
2) заявления и письма глав правительств;
3) комментарии;
4) политические характеристики и обзоры;
5) политическую информацию типа «Последние известия» и «Коротко о политических событиях».

Второе ограничение состоит в том, что статистическими подсчетами охватываются не все словосочетания, но лишь те, в которые входят наиболее часто употребляющиеся словоформы того или иного класса. Эти последние мы будем называть о п о р н ы м и словоформами.

Процедура выделения трехсловных сочетаний относительно опорной словоформы строится с таким расчетом, чтобы захватить и перенести в частотный список оптимальное число синтаксических связей. Поэтому процедура подсчета зависит, во-первых, от характера опорных словоформ, а, во-вторых, от синтаксического строя языка.

В статье приводится статистика словосочетаний, для выделения которых использовались следующие три схемы.

1. В качестве опорной словоформы используется одна из наиболее частых словоформ частотного словаря; к этой опорной словоформе присоединяются одна словоформа слева и одна словоформа справа. Например:

1) the *energy* levels;
2) der *Deutschen* Demokratischen;
3) Als *Funktion* des.

По этой схеме исследовались все три указанные выше подъязыка.

2. В качестве опорной словоформы используется одно из наиболее частых существительных частотного словаря; к этой опорной словоформе присоединяются две словоформы слева. Например: the experimental *data*.

3. В качестве опорной словоформы используется один из наиболее частых глаголов частотного словаря. Этой опорной словоформой может быть смысловой глагол, модальный и вспомогательный. Для того чтобы наиболее полно охватить связи этих глаголов, к опорной словоформе (смысловой глагол) присоединя-

ются одна словоформа слева и одна словоформа справа. Например: are *shown* in.

Если опорная словоформа является модальным или вспомогательным глаголом, то к ней присоединяются одна словоформа слева и две словоформы справа: it *is* possible to, it *can* be seen.

По 2-й и 3-й схемам исследовался подъязык английской радиоэлектроники.

При выборе опорных словоформ для исследования трехсловных сочетаний в английском подъязыке радиоэлектроники использован частотный словарь текстов по радиоэлектронике, составленный П. М. Алексеевым (см. стр. 151 настоящего сборника).

Из этого словаря соответственно выбраны 120 наиболее частых словоформ, которые используются в качестве опорных в табл. 1 (см. стр. 207), 100 наиболее частых именных словоформ, которые выступают в качестве опорных в табл. 3 и 105 наиболее частых глагольных словоформ в табл. 2.

120 первых словоформ частотного списка покрывают в среднем 50% любого текста, покрываемость 100 наиболее частых существительных составляет около 8%, а покрываемость первых 105 глаголов примерно равна 10%.

Сложнее обстояло дело с составлением списка опорных словоформ по немецким подъязыкам радиоэлектроники и публицистики. Частотные списки словоформ по этим подъязыкам отсутствуют. Поэтому составлять списки пришлось путем применения лингвистической и статистической экстраполяции.

Относительно подъязыка публицистики экстраполяция осуществлялась следующим образом. За основу списка опорных словоформ (ядер) были взяты первые 110 самых частых слов словаря Ф. Кединга.[1] По этому списку был обработан текст длиною в 16 000 словоформ. Оказалось, что первые 110 самых частых слов словаря Ф. Кединга покрывают лишь около 40% текста.

18 словоформ списка в текстах не встречались совсем или встречались очень редко. Эти словоформы (ab, dar, geben, einzelnen, her, hin и др.) были исключены из списка.

Сокращенный список опорных словоформ (ядер), взятых из словаря Кединга (74 словоформы), был дополнен из частотного словаря Венглера.[2]

В итоге получен список, включающий 175 опорных словоформ. Этот список был снова проверен на покрываемость текста. Оказалось, что дополненный список покрывает политический текст приблизительно на 50%. Однако среди 175 словоформ списка были такие, которые во всех текстах встретились менее 7 раз. Все они были исключены из списка. Вместе с тем, учитывая специфическую тематику текстов, после дополнительного исследования мы туда включили такие 9 словоформ, как Völker (вместо Volk), Republik, Länder, Ländern и др.

В итоге этой работы в окончательном списке осталось 150 опорных словоформ, в среднем покрывающих 50% текста политической информации. Перечень этих словоформ, расположенных в порядке убывающей частотности, приводится в списке 1.

[1] F. W. K a e d i n g. Häufigkeitswörterbuch der deutschen Sprache, Steglitz, 1898.

[2] H. H. W ä n g l e r. Rangwörterbuch hochdeutscher Umgangsprache. Marburg, 1963.

Список 1

Опорные словоформы в трехсловных сочетаниях в немецких газетных текстах

1. der	34. Regierung	67. ihrer
2. die	35. aus	68. dieser
3. und	36. Republik	69. Frieden
4. in	37. er	70. nur
5. des	38. zum	71. ich
6. den	39. durch	72. Berlin
7. zu	40. nach	73. alle
8. das	41. haben	74. beiden
9. von	42. wird	75. man
10. für	43. werden	76. Sowjetunion
11. auf	44. um	77. oder
12. mit	45. zwischen	78. Partei
13. sich	46. sind	79. Welt
14. daß	47. wie	80. war
15. dem	48. Staaten	81. eines
16. sie	49. ihre	82. wenn
17. ist	50. einen	83. sein
18. im	51. wir	84. Völker
19. eine	52. einem	85. Westdeutschland
20. DDR	53. diese	86. neuen
21. an	54. wurde	87. ihnen
22. auch	55. demokratischen	88. seiner
23. deutschen	56. Politik	89. deutsche
24. es	57. bei	90. neue
25. nicht	58. unter	91. Bundesrepublik
26. ein	59. vom	92. heute
27. hat	60. aber	93. mehr
28. als	61. Bonner	94. ihren
29. einer	62. vor	95. Krieg
30. über	63. noch	96. Länder
31. am	64. so	97. seine
32. zur	65. Westberlin	98. uns
33. gegen	66. westdeutschem	99. immer

100. kann
101. Bevölkerung
102. sei
103. westdeutsche
104. Menschen
105. bis
106. Jahren
107. großen
108. schon
109. hatte
110. wurden
111. können
112. Deutschland
113. habe
114. worden
115. Zeit
116. selbst
117. damit
118. diesem
119. Ländern
120. was
121. anderen
122. demokratische
123. Leben
124. unsere
125. wieder
126. ersten
127. ihr
128. muß
129. beim
130. kiene
131. seit
132. waren
133. politischen
134. dieses
135. ohne
136. dann
137. Frage
138. jetzt
139. denn
140. ganz
141. Jahre
142. will
143. da
144. dabei
145. letzten
146. nichts
147. sehr
148. doch
149. lassen
150. soll
151. hier
152. gibt
153. Jahr

Несколько иным образом определялся список опорных словоформ для исследования немецкого подъязыка электроники. Из 110 наиболее частых словоформ словаря Кединга были исключены такие словоформы, которые, как это показала проверка, в текстах по радиоэлектронике практически не встречаются. Зато в список были введены немецкие термины, являющиеся эквивалентами английских терминов по радиоэлектронике, из частотного словаря, составленного П. А. Алексеевым.

Окончательный список содержит 155 опорных словоформ, покрывающих в среднем 50% любого немецкого текста по радиоэлектронике.

Список 2

Опорные словоформы в трехсловных сочетаниях в немецких текстах по электронике

1. Abb.
2. aber
3. Abhängigkeit
4. alle
5. als
6. also
7. am
8. an
9. Anode
10. auch
11. auf
12. aus
13. Ausgang
14. Bedeutung
15. bei
16. beiden
17. beim
18. Bereich
19. Betrieb
20. Bild
21. bis
22. da
23. dabei
24. damit
25. dann
26. das
27. daß
28. dem
29. den
30. denn
31. der
32. des
33. Dichte
34. die
35. diese
36. diesem
37. dieser
38. dieses
39. Dimensionierung

40. doch
41. durch
42. ein
43. eine
44. einem
45. einen
46. einer
47. eines
48. einzelnen
49. Elektron
50. Elektronen
51. Emission
52. Emitter
53. Energie
54. Entladung
55. er
56. es
57. etwa
58. Fall
59. Feld
60. Frequenz
61. Frequenzen
62. Funktion
63. für
64. Gas
65. Gases
66. Gebiet
67. gegen
68. gibt
69. Gleichung
70. großen
71. Größe
72. haben
73. hat
74. hoch
75. hohe
76. hohen
77. hier
78. ihre
79. im
80. in
81. Ionen
82. ist
83. kann
84. Katode
85. keine
86. kommt
87. läßt
88. Leistung
89. Linse
90. mm
91. man
92. macht
93. Material
94. Materials
95. mehr
96. Messung
97. mit
98. muß
99. nach
100. nicht
101. nimmt
102. noch
103. nun
104. nur
105. Oberfläche
106. oder
107. ohne
108. positiv
109. Röhre
110. Schaltung
111. Schon
112. sehr
113. sei
114. sein
115. seine
116. sich
117. sie
118. sind
119. so
120. soll
121. Spannung
122. System
123. Systems
124. Strom
125. Stroms
126. Temperatur
127. Triode
128. um
129. und
130. unter
131. über
132. Volt
133. vom
134. von
135. vor
136. was
137. war
138. wegen
139. weil
140. Weise
141. wenn
142. werden
143. Wert
144. Widerstand
145. wie
146. wieder
147. wird
148. Wirkung
149. wobei
150. wurde
151. Zeit
152. zu
153. zum
154. zur
155. zwischen

Во всех подъязыках при выделении трехсловных сочетаний (триад) соблюдались следующие условия.

1. Знаки препинания: точка, запятая, точка с запятой, скобка, тире, знак равенства, двоеточие — рассматриваются как граница между синтагмами. Впредь эти границы будут отмечаться термином «пробел» (Δ'). Они рассматриваются в рамках триады как самостоятельные словоформы.

2. В схемах и формулах сочетания не учитывались, а в подписях под схемами, рисунками и т. п. учитывались.

3. Словоформы, соединенные дефисом, рассматриваются как одна словоформа, например: the short-circuit *amplifier*.

4. Числительные в цифровой передаче учитывались как одна словоформа и обозначались буквой *f* для английского текста и буквой *F* для немецкого. Числительные в буквенном выражении считались как самостоятельные слова. Формулы передавались буквой *F*, а индексы *i*.

5. Сокращения считались одной словоформой, например: a. d. c. amplifier.

6. Имена собственные раздельно фиксировались только относительно политических текстов. В тексте по радиоэлектронике все имена собственные обозначались *NP*.

Кроме того, при выделении трехсловных комбинаций в немецком подъязыке публицистики дополнительно применялись следующие правила.

1. Грамматическая и лексическая омонимия не различались.

2. Заголовки учитывались.

3. Словоформы типа Arbeiter-und-Bauern-Staat считались разными словоформами, соединенными союзом, и выделялись как отдельные единицы.

4. Выше было отмечено, что скобка считалась пробелом, однако в таких случаях, как (West-) Deutschland, (West-) Berlin, т. е. там, где словоформа, стоящая в скобках, меняла значение слова за скобками, скобки и дефис во внимание не принимались, а считалось, что (West —) Deutschland = Westdeutschland.

5. В случаях типа Führer *der* KP Chinas после опорного слова учитывалось только KP, словоформа Chinas опускалась. В позиции перед опорной словоформой учитывалось только последнее Chinas *soll* heute.

6. В таких случаях, как Δ *die* neue и Δ *Die* neue, большие буквы во внимание не принимались.

7. При статистической обработке трехсловных сочетаний, включающих опорную именную форму, в английском подъязыке радиоэлектроники лексические омонимы фиксировались как отдельные словоформы. Например:

light (*a*) current (*a*),
light (*n*) current (*n*).

Точность соблюдения этого правила особенно важна тогда, когда опорные словоформы выступают в качестве определения, т. е. когда есть продолжение связи вправо. Например:

this *light* (emission) 'эмиссия света',
executive *and* light ('легкий') aircraft.

Таблица 1

Трехсловные сочетания с частотным опорным словом в английских текстах по радиоэлектронике. Выборка в 200 000 словосочетаний ($M \approx 400\,000$)

i	Трехсловные сочетания	F	$f \cdot 10^5$	$h \cdot 10^5$	$f^* \cdot 10^5$	$h^* \cdot 10^5$
1	f *and* f	1256	6280	3140	6280	3140
2	Δ *Fig.* f	761	3805	1905	10085	5045
3	Δ *it* is	595	2975	1485	13060	6530
4	in *Fig.* f	541	2705	1352	15765	7882
5	N *and* NP.	452	2260	1130	18025	9012
6	f *is* the	445	2225	1113	20250	10125
7	Δ *however* Δ′	437	2185	1092	22435	11217
8	Δ *and* the	397	1985	993	24420	12210
9	Δ *in* the	362	1810	905	26230	13115
10	shown *in* Fig. . . .	360	1800	900	28030	14015
11	f *to* f	318	1590	795	29620	14810
12	Δ *and* Δ	316	1580	790	31200	15600
13	Δ *and* f	294	1470	735	32670	16335
14	of *the* f	285	1425	712	34095	17047
15	Δ *if* the	260	1300	650	35395	17697

Примечание. В табл. 1—5 M — общая выборка; N — количество словосочетаний; f — относительные частоты по N; h — относительные частоты по M; f^* — накопленная частота относительных частот f; h^* — накопленная частота относительных частот h.

Таблица 2

Наиболее частые трехсловные словосочетания с опорной глагольной формой в английских текстах по радиоэлектронике ($N = 50\,000$, $M = 450\,000$)

i	Трехсловные сочетания	F	$f \cdot 10^5$	$h \cdot 10^5$	$f^* \cdot 10^5$	$h^* \cdot 10^5$
1	is *shown* in	162	3240	360	3240	360
2	is *given* by	131	2620	291	5860	651
3	F *shows* the	102	2040	227	7900	878
4	as *shown* in	99	1980	220	9880	1098
5	are *shown* in	97	1940	216	11820	1314
6	as *follows* Δ	81	1620	180	13440	1494
7	it *is* possible to . . .	69	1380	153	14820	1647
8	be *used* Δ	59	1180	131	16000	1778
9	be *used* to	55	1100	122	17100	1900
10	to *determine* the . . .	52	1040	116	18140	2016
11	was *found* to	52	1040	116	19180	2132
12	Δ *using* the	52	1040	116	20220	2258
13	be *seen* that	50	1000	111	21220	2369
14	is *applied* to	43	860	96	22080	2465
15	we *have* F	42	840	93	22920	2558
16	be *noted* that	42	840	93	23760	2651
17	F *shows* a	41	820	91	24580	2742
18	was *found* that . . .	40	800	89	25380	2831
19	to *obtain* the	40	800	89	26180	2920
20	is *used* to	40	800	89	26980	3009

Примечание. Условные обозначения те же, что и в табл. 1.

Таблица 3

Трехсловные словосочетания с опорным существительным в английских текстах по радиоэлектронике
($N=40\,000$ и $N=200\,000$)

i	Трехсловные сочетания	$N = 40\,000$			$M = 138\,115$		$N = 200\,000$		$M = 335\,618$	
		F	$f \cdot 10^6$	$h \cdot 10^6$	$f^* \cdot 10^6$	$h^* \cdot 10^6$	F	$f \cdot 10^6$	$h \cdot 10^6$	$f^* \cdot 10^6$	$h^* \cdot 10^6$
1	Δ The *results*	23	2300	166	2300	166	67	3350	262	3350	262
2	as a *function*	25	2500	181	4800	347	57	2850	223	6200	485
3	in the *case*	23	2300	166	7100	513	54	2700	211	8900	696
4	of the *amplifier*	24	2400	173	9500	686	46	2300	180	11200	876
5	Δ in *order*	13	1300	94	10800	780	44	2200	172	13400	1048
6	of the *order*	23	2300	166	13100	946	41	2050	160	15450	1208
7	of the *system*	20	2000	144	15100	1090	38	1900	148	17350	1356
8	Δ the *effect*	18	1800	130	16900	1220	37	1850	144	19200	1500
9	the ground *state*	16	1600	116	18500	1336	34	1700	133	20900	1633
10	as a *result*	18	1800	130	20300	1466	33	1650	129	22550	1762
11	Δ the *use*	20	2000	144	22300	1610	32	1600	125	24150	1887
12	in this *case*	11	1100	79	23400	1689	29	1450	113	25600	2000

Таблица 4

Трехсловные сочетания в немецких текстах по радиоэлектронике
($N=100\,000$, $M\approx200\,000$)

i	Трехсловные сочетания	F	$f \cdot 10^6$	$h \cdot 10^6$	$f^* \cdot 10^6$	$h^* \cdot 10^6$
1	F *und* F	1430	7065	3533	7065	3533
2	Δ *Bild* F	1251	6255	3127	13320	6660
3	Δ *Abb* F	604	3020	1510	16340	8170
4	Δ *daß* die	570	2850	1425	19190	9535
5	F *bis* F	386	1930	965	21120	10560
6	von *etwa* F	302	1510	755	22630	11315
7	Δ *so* daß	298	1490	745	24120	12060
8	Δ *in* der	273	1365	683	25485	12743
9	Δ *bei* der	265	1325	662	26810	13405
10	Δ *für* die	264	1320	660	28130	14065
11	N *und* N	252	1260	630	29390	14695
12	in *Bild* F	216	1080	540	30470	15235
13	Δ *daß* der	213	1065	533	31535	15768
14	in *Abb* F	210	1050	525	32585	16293
15	werden *kann* Δ′	191	955	477	33540	16770
16	Δ *da* die	189	945	473	34485	17243
17	Δ *es ist*	170	850	425	35335	17668
18	Δ *die* in	167	835	417	36170	18085
19	Δ *wenn* die	157	785	393	36953	18478
20	F *ist* Δ′	138	690	345	37645	18823

Таблица 5

Трехсловные словосочетания с опорным частотным словом
в немецких публицистических текстах ($N=200\,000$, $M\approx400\,000$)

i	Трехсловные сочетания	F	$f \cdot 10^6$	$h \cdot 10^6$	$f^* \cdot 10^6$	$h^* \cdot 10^6$
1	Δ *daß* die	817	4085	2042.5	4085	2042.5
2	Δ *in* der	315	1575	787.5	5660	2830
3	Deutschen *Demokratischen* Republik	266	1330	665	6990	3495
4	der *Deutschen* Demokratischen	255	1275	637.5	8265	4132.5
5	Δ *es* ist	219	1095	547.5	9360	4680
6	Δ *daß* der	208	1040	520	10400	5200
7	Δ *daß* sie	190	950	475	11350	5675
8	Δ *die* sich	170	850	425	12200	6100
9	beiden *deutschen* Staaten . .	164	820	410	13020	6510
10	Δ *für* die	157	785	392.5	13805	6902.5
11	Δ *daß* es	148	740	370	14545	7272.5
12	Δ *das* ist	142	710	355	15255	7627.5
13	Δ *in* dem	133	665	332.5	15920	7960
14	Δ *um* die	131	655	327.5	16575	8287.5
15	der *DDR* Δ	129	645	322.5	17220	8610
16	Δ *in* den	124	620	310	17840	8920
17	Δ *daß* sich	123	615	307.5	18455	9227.5
18	Δ *die* in	116	580	290	19035	9517.5
19	Deutsche *Demokratische* Republik	115	575	287.5	19610	9805
20	in *der* Welt	113	565	282.5	20175	10087

8. Продолжение связи (справа и слева) отмечалось знаком многоточия:

Δ The *time*. . .
. . . state of value.

9. Ложные или слабые связи отмечались знаком *. Например:
*groups different *temperatures*.

10. В случае, если алгоритм отсечения мог нарушить структуру и вызвать различное толкование, делалась пометка (в скобках). Например:

change (*inf.*) with *temperature*,
change (*imp.*) with *temperature*,
change (*n.*) with *temperature*,
rise (*n.*) in *temperature*.

Особую трудность представляют глаголы с послелогами, где алгоритм отсечения полностью нарушает смысловую связь и пометка о продолжении связи слева явно недостаточна. Например:

on the power,
(turn) on the power.

Представляется целесообразным в таких случаях считать глагол с послелогом как взаимосвязанное слово и делать соответствующую пометку: (turn)* on the power.

В табл. 1—5 приводятся начальные участки частотных списков трехсловных сочетаний, составленные относительно указанных выше подъязыков.

Л. Г. Кравец

НЕКОТОРЫЕ КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ АНГЛИЙСКИХ ИМЕННЫХ СЛОВОСОЧЕТАНИЙ

Именные (субстантивные) словосочетания с препозитивными определителями играют особо важную роль в процессе формирования и закрепления понятий, относящихся к различным областям знания. Достаточно отметить, что в разделе новых слов (Addenda) словаря Уэбстера [1] на долю таких словосочетаний, рассматриваемых в качестве самостоятельных единиц языка, приходится около 20% его состава. В отраслевых же англо-русских словарях упомянутые словосочетания занимают около 70% всего их объема.

Наиболее распространенной разновидностью именных словосочетаний являются двухкомпонентные сочетания типа «определение+ядро», составляющие 70% от общего числа именных словосочетаний.[2] Роль ядра в таких сочетаниях играет существительное, а в качестве определяющих слов чаще всего выступают существительные (task force), прилагательные (basic policy) и причастия (delayed action).[3]

Ниже излагаются отдельные количественные характеристики двухкомпонентных именных словосочетаний. Ограниченные размеры выборки не допускали проведения достаточно строгого математического анализа. Тем не менее полученные данные позволяют выделить некоторые количественные признаки исследуемых качественных явлений.

1. В интересах целого ряда прикладных задач (например, составление словников обычных отраслевых и особенно автомати-

[1] Webster's New International Dictionary of the English Language, 2nd. ed., London, 1955.

[2] Из общего количества 24 000 именных словосочетаний с препозитивными определениями, зафиксированных при обследовании специальных текстов военной тематики, двухкомпонентных словосочетаний оказалось примерно 19 000, а словосочетаний с тремя и более компонентами — около 5000.

[3] Проблема разграничения словосочетания и сложного слова в настоящей работе не затрагивается. Все именные комплексы слов, разделенных на письме пробелом, относятся к категории словосочетаний.

ческих словарей для машинного перевода, получение исходных данных при разработке проблем автоматического реферирования, индексация и построение информационно-поисковых языков) крайне желательно выделение активного (наиболее частотного) состава из числа именных словосочетаний, фиксируемых при обработке специальных текстов.[4] При этом в наше распоряжение поступают наиболее нужные словосочетания, потребность в которых ощущается особенно часто.

В рассматриваемом случае требовалось: а) установить частоту появления в специальных текстах различных двухкомпонентных словосочетаний и б) определить роль (или вес) активного состава именных словосочетаний, а также формирующих эти сочетания слов в общей совокупности лексического материала, составляющего специальный текст.

Частота появления в тексте того или иного словосочетания устанавливается делением количества случаев появления данного словосочетания на общее число зафиксированных словосочетаний. Например, если сочетание nuclear weapons 'ядерное оружие' зафиксировано 175 раз, а общее число двухкомпонентных словосочетаний равно 19 000, частота данного сочетания выражается отношением $\frac{175}{19\,000}=0{,}009$.

Для отграничения активного состава от остальной совокупности двухкомпонентных именных словосочетаний требуется установить порог частотности. Высота порога частотности устанавливается произвольно, исходя из практических потребностей. Естественно, что в интересах получения достаточно достоверных данных желательно, чтобы частота всех словосочетаний, расположившихся выше порога, вычислялась с относительной ошибкой хотя бы не более 0.5.

Расчет наименьшей допустимой частоты осуществлен по формуле

$$f=\frac{Z_p^2}{\delta^2 N},$$

где Z_p — константа, которая при условии нормального распределения слов в тексте и при уровне значимости 0.95 равна 1.96. При этих условиях f=0.00084.

Для получения более представительной выборки частотных словосочетаний было признано целесообразным установить также и второй, более низкий порог частотности, равный 0.00026. Относительная ошибка, с которой вычисляется частота самого редкого из выделяемых таким образом сочетаний, заметно возрастает, составляя 0.9.

[4] Специальными именуются тексты определенной тематики, в которых описываются явления и процессы, характерные для той или иной сферы человеческой деятельности (науки, техники, военного дела и т. д.).

С установлением первого (высокого) порога частотности из общей массы около 8100 различных двухкомпонентных комбинаций выделились всего 133 словосочетания с частотой 0.00084 и выше, т. е. около 1.7% от общего числа двухкомпонентных именных словосочетаний. Однако эти 133 сочетания, встретившись в тексте от 15 до многих десятков раз, составили выборку около 5000 сочетаний, или более 25% от общего числа словосочетаний, зафиксированных при обследовании текстов.

После установления второго частотного порога выделились уже 576 различных словосочетаний с частотой 0.00026 и выше, т. е. немногим более 7% от общего числа различных двухкомпонентных комбинаций. Эти сочетания, встретившись в тексте от 5 до многих десятков раз, дали выборку около 8500 сочетаний, что составляет уже около 45% от общего числа зафиксированных случаев появления двухкомпонентных словосочетаний.

В формировании 576 сочетаний приняли участие 262 определяющих слова и 212 ядер. Эти 474 слова играют важнейшую роль в анализируемых текстах, участвуя в образовании двухкомпонентных словосочетаний, вероятность встретить которые при чтении текстов приближается к 50%.

Между тем роль различных ядер и определяющих слов в формировании зафиксированной совокупности двухкомпонентных сочетаний далеко не одинакова.

Так, всего отмечено 1981 определение, принявшее участие в формировании около 8100 различных словосочетаний, встретившихся, как указывалось, примерно 19 000 раз. Из числа этих определений 889 единиц сочетаются всего с одним ядром и при этом — только по одному разу (т. е. они участвуют в создании 889 различных сочетаний, встретившихся в текстах только 889 раз). Следующие 830 определений сочетаются как с одним, так и с несколькими ядрами; однако частота совместной встречаемости ни с одним из ядер не достигает 0.00026. Эти 830 определений встречаются в тексте в общей сложности 4916 раз.

Наконец, имеется 262 определения, каждое из которых входит в одно или несколько сочетаний с частотой выше 0.00026. В общей же сложности эти определения встречаются в тексте (в препозиции к ядру) около 13 200 раз. Таким образом, упомянутые 262 определения, составляя всего 13% от общего числа определяющих слов, встречаются в 13 200 из 19 000 двухкомпонентных словосочетаний, т. е. на них приходится около 70% выборки!

Приведенные цифры наглядно показывают, насколько активную роль в обследованной выборке играют выделенные частотные словосочетания: хотя они составляют лишь 7% от общего числа сочетаний, на их долю приходится около половины случаев появления двухкомпонентных словосочетаний в анализируемых текстах, т. е. примерно столько же, сколько приходится на долю остальных 93% словосочетаний.

Не менее важное значение имеют и слова, выступающие в роли компонентов выделенных 576 словосочетаний. Их насчитывается всего 474. Если же учесть, что 57 определяющих слов и ядер представляют собой по сути дела одни и те же слова, точнее существительные, использующиеся как в роли ядер, так и в роли определяющих слов (типа battle, infantry и т. д.), общее число различных слов, участвующих в образовании частотных сочетаний, сократится до 417.

Таким образом, в формировании почти каждого второго словосочетания обследованного текста принимают участие всего около 400 наиболее активных слов!

Несомненно, что как словосочетания, входящие в активный состав, так и слова, участвующие в их формировании, заслуживают самого пристального внимания при решении прикладных лингвистических задач. Полученные данные, в частности, указывают на вполне реальную возможность разработки автоматических словарей для машинного перевода с достаточно широким охватом терминологических словосочетаний, вызывающих трудности при их межъязыковой интерпретации. Резкие колебания частот именных словосочетаний в специальных текстах позволяют разрабатывать компактные и вместе с тем достаточно эффективные словари оборотов, которые в значительной степени снимут необходимость алгоритмизации сложных семантических отношений между сочетающимися словами.

2. Известно, что одной из особенностей английского языка является возможность использовать существительные в роли препозитивных определений при других существительных. Эта возможность широко используется и при формировании словосочетаний в специальных текстах. Одновременно проявляется и обратная тенденция — субстантивация прилагательных и причастий, однако подобные явления встречаются значительно реже.

Из 262 слов, выполняющих определительную функцию в сочетаниях, входящих в активный состав, 93 слова по форме сходны с соответствующими существительными и могут также использоваться в роли ядер сочетаний.[5]

Часть из упомянутых 93 слов при формировании 576 сочетаний используется только в роли определений (однако это не исключает возможности их использования в роли ядер при формировании прочих сочетаний, не вошедших в число 576).

Имеются в виду следующие 36 слов:

air	denial	mass
assembly	drop	maximum
barrier	enemy	night
cold	escape	offensive
corps	field	phase
counterattack	key	repair
daisy	land	river
damage	major	road
shock	future	supply
chore	general	task
signal	guerrilla	theater
flank	infiltration	weather

[5] Так как размеры выборки ограничены, приводимые ниже данные по конкретным словам имеют силу лишь для рассматриваемой совокупности текстов. Однако результаты сопоставлений в целом могут представить интерес и в более широком плане.

Из числа оставшихся 57 слов 16 чаще используются в качестве определений:

battle	landing
chemical	police
combat	rear
communication	staff
deception	tank
engineer	transport
ground	warning
infantry	world

Особый интерес, естественно, представляют случаи, когда преимущественное использование слов в роли определений проявляется особенно отчетливо. Это относится к выделенным словам (например, для слова *combat* частоты использования в роли определения и ядра составляют соответственно 447 и 113, для *ground* — 103 и 19, для *rear* — 109 и 19 и т. д.).

Три слова (battlefield, division, maneuver) одинаково часто встречаются в каждой из двух позиций.

Наконец, 38 слов чаще используются в роли ядра сочетания:

advance	defense	replacement
aggressor	delivery	reserve
area	fire	security
army	force	service
artillery	intelligence	support
assault	missile	target
attack	mission	terrain
aviation	mobility	traffic
brigade	objective	training
command	observation	troop
commander ('s)	operations	war
component	personnel	weapons
control	reconnaissance	

В ряде случаев преимущественное употребление в роли ядра проявляется особенно отчетливо (например, для слова force частоты использования в роли ядра и определения составляют соответственно 1501 и 29, для support — 557 и 55, для area — 576 и 76, для operation — 625 и 30 для war — 258 и 16, для weapon — 277 и 13 и т. д.).

Учитывая, что рассмотренные слова в своем большинстве являются исконными существительными, можно заключить, что у некоторой части существительных намечается тенденция преимущественного функционирования в роли определений, т. е. наблюдается функциональное сближение этих существительных с прилагательными (по крайне мере в пределах специальных текстов данной тематики).

Факты функционального сближения отдельных существительных с прилагательными могут быть использованы при разработке правил формализованного выявления структуры многокомпонентных именных словосочетаний при машинном переводе.

Известно, например, что анализ структуры трехкомпонентного именного словосочетания, условно обозначаемого комплексом XYZ, существенно усложняется, когда в роли компонента Y выступает существительное, способное выполнять функцию как препозитивного определения к компоненту Z, так и ядра двухкомпонентного сочетания XY, которое в целом определяет компонент Z.[6]

Процедура анализа структуры словосочетания может быть значительно упрощена путем применения вероятностных оценок в тех случаях, когда в позиции Y находится существительное с достаточно ярко выраженной тенденцией употребления в одной из названных выше функций.

Например, встретив в обследованном тексте словосочетание, в котором позицию Y занимают слова air, daisy или night (выступающие почти исключительно в роли определений), можно с очень высокой степенью вероятности утверждать, что компонент Y определяет Z и не является ядром сочетаний XY. Напротив, если в позиции Y находятся слова force (или aquisition), более вероятна структура, в которой компонент Y выполняет функцию ядра словосочетания XY.

3. В соответствии с одним из подходов устойчивость словосочетания определяется как вероятность, с которой один из компонентов предсказывает совместное появление остальных компонентов данного словосочетания.[7] Представляется интересным рассмотреть некоторые тенденции, проявляющиеся в процессе вычисления вероятностей, предсказания именного словосочетания по одному из компонентов.

Возможность предсказания одной языковой единицы посредством другой рассматривается в теории информации; она тесно связана с самим понятием «информация». Считается, что информация может доставляться лишь «неожиданным», т. е. неизвестным нам

[6] Л. Г. Кравец. Анализ структуры словосочетаний в английских научно-технических текстах. НТИ, № 10, 1963.

[7] И. А. Мельчук. О терминах «устойчивость» и «идеоматичность». ВЯ, № 4, 1960.

заранее сигналом (если бы мы были предупреждены о содержании сигнала, то не получили бы с ним никакой информации). При этом ценность информации в основном как раз и определяется «степенью неожиданности» сигнала. Такого рода соображения связывают понятие «информация» с понятием вероятности того или иного исхода опыта, которую можно описать как частоту появления именно этого исхода в длинной серии однотипных испытаний.

Этот вопрос применительно к словосочетаниям рассматривался Л. Р. Зиндером,[8] который ввел понятие лингвистической вероятности (т. е. вероятностных ограничений, характеризующих язык), увязав его с понятием количества избыточной информации, облегчающей догадку и тем самым понимание и восприятие речи.

Естественно поставить вопрос, не существует ли каких-либо закономерностей в размещении информации между компонентами словосочетаний в специальных текстах.[9]

Выявление компонентов, по которым словосочетания предсказываются с наибольшей вероятностью, имеет вполне определенный лингвистический смысл. Знание таких фактов способствует уяснению сущности некоторых структурно-семантических преобразований в пределах словосочетаний, например усечение сочетания, перенос на оставшийся компонент (компоненты) значения сочетания в целом и т. п.

В обследованном тексте отмечены случаи появления таких, например, единиц, как the defensive 'оборонительный бой', 'оборона', the offensive 'наступательный бой', 'наступление', the wounded 'раненые', the killed 'убитые', the first (от first lieutenant) 'первый лейтенант' и т. д.

По-видимому, упомянутые преобразования происходят постепенно: сосредоточение избыточной информации в одном из компонентов повышает вероятность предсказания другого компонента; члены языкового коллектива, стремясь к сокращению избыточности сообщения, нередко опускают компонент, несущий незначительную информацию; со временем подобное отклонение (при прочих благоприятных условиях) превращается в норму.

Поскольку мы имеем дело с двухкомпонентными сочетаниями, желательно установить, какой из двух компонентов рассматриваемых специальных словосочетаний (определение или ядро?) содержит избыточную информацию относительно чаще.

Согласно наблюдениям некоторых лингвистов, определения, как правило, предсказываются более легко, чем ядра соответ-

[8] Л. Р. Зиндер. О лингвистической вероятности. ВЯ, № 2, 1958.

[9] Данную задачу можно в какой-то степени сравнить с задачей, решавшейся в отношении букв, из которых строятся слова. Ср.: Р. Г. Пиотровский. Размещение информации в слове. Матер. межвузовск. конфер. по применению структурных и статистических методов исследования словарного состава языка, М., 1961.

ствующих словосочетаний (т. е. вероятность предсказания сочетания по ядру обычно выше).[10]

Анализ собранного нами материала заставляет усомниться в достоверности подобного заключения (по крайней мере по отношению к словосочетаниям из данной совокупности специальных текстов).

Как уже указывалось, в обследованной выборке зафиксировано 8100 различных двухкомпонентных комбинаций, в создании которых приняло участие 1981 определение (A) и 1499 ядер (Nc). Разница в реально зафиксированном числе определений и ядер позволяет предположить, что существуют в общем более благоприятные возможности предсказания сочетаний по определениям, а не по ядрам.

Известно, что вероятность предсказания какого-то исхода непосредственно зависит от общего числа возможных исходов (чем меньше число различных исходов, тем больше вероятность предсказания одного из них).

Поскольку каждое из 1499 ядер теоретически может сочетаться с 1981 определением, а каждое из 1981 определения — лишь с 1489 ядрами, средняя вероятность предсказания ядра по определению должна быть выше.

Данное предположение можно выразить и в иной форме: как 1981 определение, так и 1499 ядер вступают в одинаковое число (8100) различных сочетаний; отсюда на одно определение приходится $\frac{8100}{1931} = 4$ сочетания, а на одно ядро — $\frac{8100}{1499} = 5.4$ сочетания. Другими словами, если по одному ядру в среднем предсказывается 5.4 сочетания, то по одному определению — лишь 4. Естественно, что в последнем случае предсказывать легче.

То же самое наблюдается и в отношении выделенных выше частотных словосочетаний: в формировании 576 сочетаний, особенно часто встречающихся в обследованных текстах, принимают участие 262 определения и 212 ядер, т. е. и в данном случае средняя вероятность предсказания сочетания по одному из 262 определений должна быть выше, чем по одному из 212 ядер.

Для подтверждения правильности этого наблюдения было проведено сопоставление вероятностей предсказания конкретных

[10] «Вообще говоря, существительные предсказать труднее, чем прилагательные. Это справедливо по отношению к обычным существительным и обычным прилагательным. Существительное barn (амбар) уже содержит в себе определенные сведения о возможном круге прилагательных; данное существительное будет обычно описываться с точки зрения его размеров, цвета, формы, владельца и т. д. Что же касается прилагательного red (красный), то число возможных ядер оказывается практически безграничным, в связи с чем вероятность предсказания какого-либо конкретного ядра очень низка». (D. L. Bolinger. Stress and Information. American Speech, v. 33, № 1, 1958).

словосочетаний по каждому из двух компонентов. В качестве объекта анализа избраны упомянутые 576 наиболее частотных словосочетаний.

Сопоставление вероятностей предсказания всех сочетаний как по ядру, так и по определению показало, что в 356 случаях из 576 (62% случаев) сочетание с наибольшей вероятностью предсказывается по определению. В 5 случаях (0.9%) вероятности предсказания сочетания по определению и по ядру равны; в 215 случаях (37%) сочетания с наибольшей вероятностью предсказываются по ядру.

Наибольший интерес, естественно, представляют случаи, когда вероятность предсказания сочетания по одному из компонентов значительно превышает вероятность его предсказания по другому.

а) Прежде всего выделим сочетания, вероятность предсказания которых в пределах данной выборки равна 1 (предсказываемые по одному из компонентов со 100%-й вероятностью [11]). Например: по компоненту A: exploiting forces (27) 27/1571 [12] 'войска развития успеха'; world war (23) 23/258 'мировая война'; build-up area (15) 15/576 'район сосредоточения'; spoiling attack (11) 11/294 'упреждающий удар'; objective area (46) 46/576 'район цели' и т. д.; по компоненту Nc: target acquisition (10) 27/10 'обнаружение (захват) цели'; mobility differential (6) 10/6 'превосходство в подвижности'; troop safety (11) 38/11 'безопасность войск' и т. д.

б) Определенная часть сочетаний предсказывается по одному из компонентов с вероятностью, близкой к 1. Например: по компоненту A: covering force (84) 86/1501 'силы прикрытия'; cold war (32) 34/258 'холодная война'; frontal attack (7) 8/294 'фронтальный удар'; administrative support (180) 200/577 'административно-хозяйственное обеспечение'; task force (71) 76/1501 'оперативно-тактическая группа' и т. д.; по компоненту Nc: main body (30) 101/32 'главные силы'; armored cavalry (17) 112/19 'легкие бронетанковые войска'; high explosive (16) 82/20 'бризантное взрывчатое вещество' и т. д.

в) Многочисленную группу представляют сочетания, вероятность предсказания которых по определению значительно превышает вероятность их предсказания по ядру. Например: main attack 'главный удар'; air force 'военно-воздушные силы'; mobile

[11] Ограниченные размеры выборки позволяют говорить с достаточной достоверностью лишь об общей тенденции в распределении информации между компонентами анализируемых сочетаний. Что же касается выводов по конкретным сочетаниям, то они чаще всего имеют силу лишь в пределах обследованной выборки.

[12] Первая цифра в скобках означает число случаев появления данного сочетания в обследованном тексте. Числитель (вторая цифра) дроби означает общее число случаев появления в текстах данного определения, а знаменатель — число случаев появления ядра.

defense 'подвижная оборона'; airborne operation 'воздушно-десантная операция'; subordinate unit 'подчиненное подразделение' и т. д.

г) Сочетания, предсказываемые с наибольшей вероятностью по ядру, составляют не столь многочисленную группу; это та часть сочетаний, на которых в силу определенных причин не сказалась общая тенденция более высокой вероятности предсказания по компоненту А. Например: field manual 'полевой устав'; general outpost 'общее охранение'; nuclear parity 'равенство в ядерном вооружении' и т. д.

д) Наконец, имеется группа сочетаний, примерно с одинаковой вероятностью предсказываемых по любому из компонентов. Например: combat units 'боевые части'; close combat 'ближний бой'; defensive position 'оборонительная позиция' и т. д.

Сопоставление полученных данных позволяет заключить, что отмеченная тенденция в распределении информации между компонентами проявляется не только в абсолютном преобладании случаев более высокой вероятности предсказания сочетаний по компоненту А, но также и в сравнительно более частых случаях приближения вероятности предсказания сочетаний к 1 (100%), когда эти сочетания предсказываются именно по определяющему слову, а не по ядру.

Анализ конкретных примеров показывает, что сочетание с наибольшей вероятностью предсказывается по компоненту А, как правило, тогда, когда в роли компонента Nc выступают наиболее активные термины типа attack, defense, force, operation, support, war и т. п. Эти слова в силу их частой встречаемости (т. е. в силу высокой вероятности их появления в данном тексте) несут сравнительно незначительное количество информации.

Действительно, слова defence, operation и т. п., встречаясь в военном тексте, сами по себе говорят довольно мало из-за сильно разросшейся дифференциации понятий, обозначаемых словами «оборона», «операция» и т. п. Чтобы понять, что же в действительности имеется в виду, читающий или слушающий нуждается в одном из дифференциальных признаков, обычно сопровождающих основное понятие. Естественно поэтому, что именно слово-определение, выражающее этот признак, и несет основной «заряд» информации, выделяемой языком данному сочетанию.

Примерно то же можно сказать и об определениях типа enemy, military и т. п. В данном случае в силу частой встречаемости и широкой сочетаемости подобных определений дифференциальные функции переходят к ядрам (например, military strategy, а не military tactics), которые содержат основную часть информации.

Однако подобное распределение информации между компонентами сочетания сохраняется только в пределах текстов данной тематики: за пределами военных текстов определение military, например, снова приобретает заметные дифференцирующие свойства, и в сочетании military correspondent важным может оказаться то, что данный корреспондент «military», а не «political».

Из сказанного делается вывод об обратно пропорциональной зависимости между способностью слова-термина предсказывать сочетание и его распространенностью в текстах данной тематики: чем активнее используется данный термин, чем разнообразнее его сочетаемость, тем меньше его информационная нагрузка и тем ниже его способность предсказывать словосочетание.

Выявление закономерностей в предсказываемости словосочетаний может иметь различные приложения в теории и практике машинного перевода. Например, располагая данными о вероятности предсказывания сочетаний по различным компонентам, можно приписывать индексы вхождения слова в сочетание, содержащееся в автоматическом словаре оборотов, не первому (крайнему слева) компоненту, как это обычно делается, а именно тому слову, которое с наиболее высокой вероятностью предсказывает данное сочетание. Благодаря этому необходимость обращения к словарю оборотов в процессе перевода существенно снизится.

Л. В. Малаховский

НЕКОТОРЫЕ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ АНГЛИЙСКИХ ТЕКСТОВ ПО ЭЛЕКТРОНИКЕ

В области изучения количественных характеристик английской письменной речи сделано еще очень мало, если не считать работ, связанных с составлением частотных словарей. К тому же многие исследования в этой области обладают существенными недостатками, затрудняющими их использование: исследуются разнородные тексты; характер текста указывается слишком неопределенно («стандартный», «нормальный») или не указывается совсем; отсутствуют сведения об объеме изучаемого материала, а также о методике вычислений; не приводится никаких статистических показателей, позволяющих судить о дисперсии полученных величин и их достоверности.[1]

Настоящее исследование проведено на текстах, относящихся к одной и той же отрасли — радиоэлектронике.[2] Это обеспечивает бо́льшую однородность исследуемого материала, повышает достоверность получаемых результатов и, кроме того, позволяет получить данные, представляющие непосредственный интерес на современном этапе развития математической лингвистики, поскольку поиски в области машинного перевода и автоматического рефери-

[1] Л. Бриллюэн. Наука и теория информации. М., 1960.
C. L. Barber. In: Contributions to English syntax and philology. Gothenburg studies in English, № 14. Göteborg, 1962; Ch. D. Bourne and *D. F. Ford. Information and Control, 1961, 4, 1, pp. 48—67; G. Dewey.* Relative frequency of English speech sounds. Harvard univ., 1923; N. R. French, C. W. Carter, Jr., W. Koenig, Jr. «Bell System Technical Journal», 1930, 9, 2; R. T. Griffith. «Journal of Franklin Institute», 1949, 248, 5, pp. 399—436; G. Herdan. Language as choice and chance. Groningen, 1956; A. Nasvytis. Die Gesetzmässigkeiten kombinatorischer Technik. Berlin, 1954; F. Pratt. Secret and urgent. The story of codes and ciphers. 2nd ed. N. Y., 1942.

[2] В сборе материалов принимали участие А. И. Варшавская, Е. Е. Корди, И. Б. Фитиалова, И. М. Мальцева, Л. А. Медведева.

Таблица 1

Распределение вероятностей графем

Графемы	$\bar{x} \pm s$	p_1	p_2	Графемы	$\bar{x} \pm s$	p_1	p_2
Пробел	196±3	0,1625	—	*m*	24±1	0,020	0,024
e	128±2	0,106	0,127	*p*	23±1	0,019	0,023
t	103±2	0,085	0,102	*g*	20±2	0,016	0,019
i	89±3	0,074	0,089	*b*	15±1	0,012	0,014
a	74±2	0,062	0,074	*w*	13±1	0,0104	0,012
o	72±2	0,059	0,071	*y*	12±1	0,0101	0,012
n	72±3	*0,059*	*0,071*	*v*	*11±1*	*0,0089*	*0,011*
s	66±2	0,054	0,065	*q*	4±0,6	0,0031	0,004
r	65±2	0,054	0,065	Дефис	3±0,5	0,0023	—
h	47±2	0,038	0,045	*k*	2±0,4	0,0017	0,002
c	40±1	0,033	0,039	*x*	2±0,3	0,0017	0,002
l	40±2	0,032	0,038	*z*	2±0,5	0,0014	0,002
d	34±1	0,028	0,033	*j*	1±0,2	0,0007	0,001
u	29±1	0,024	0,029	Апостроф	0,1±0,01	0,0001	—
f	26±1	0,022	0,026	Итого		1,0000	1,000

Примечание. В табл. 1—4: $\bar{x}$ — частота рассматриваемого языкового элемента (в среднем на одну выборку); s — стандартная ошибка среднего значения частоты; ДС — длина слова (в букве); ДФ — длина фразы (в словах).

Таблица 2

Распределение позиционных вероятностей графем в словах длиной от 4 букв и более

Графемы	Позиционные вероятности (в тысячных)									
	место от начала слова					место от конца слова				
	1	2	3	4	5	5	4	3	2	1
a	62	90	130	49	65	160	58	95	80	2
b	43	10	6	13	5	11	16	14	4	0
c	126	6	32	50	30	63	73	14	66	7
d	36	8	14	34	8	18	20	17	24	97
e	57	166	76	155	218	118	52	113	202	190
f	44	4	16	8	8	13	21	1	0	1
g	21	0	25	18	8	26	12	8	33	33
h	23	116	4	32	22	16	44	62	1	39
i	55	114	74	64	122	109	91	176	49	0
j	6	0	4	0	0	0	1	0	0	0
k	2	0	3	16	2	5	8	2	1	6
l	30	14	74	72	64	41	39	30	66	51
m	45	30	38	25	24	20	32	9	9	27
n	19	44	76	86	59	38	36	83	97	96
o	34	152	50	24	50	24	48	80	119	8
p	57	17	51	42	14	23	16	18	2	3
q	4	20	2	4	0	1	0	1	0	0
r	45	86	108	59	59	79	63	54	74	95
s	87	22	58	80	63	58	54	18	38	150

Таблица 2 (*продолжение*)

Графемы	Позиционные вероятности (в тысячных)									
	место от начала слова					место от конца слова				
	1	2	3	4	5	5	4	3	2	1
t	100	16	75	87	104	80	202	110	61	115
u	15	54	22	45	38	41	30	58	47	0
v	37	8	26	16	14	22	17	16	17	0
w	48	2	5	14	3	26	48	4	6	5
x	0	14	14	0	0	4	0	9	1	4
y	1	6	17	8	20	8	12	5	2	71
z	3	0	0	0	0	0	4	2	0	0
Дефис	0	0	0	0	0	0	0	3	0	0
Апостроф	0	0	0	0	0	0	0	0	0	0

рования ведутся сейчас главным образом на базе узко ограниченного лексического материала.

Определялись следующие количественные характеристики английских текстов по радиоэлектронике: 1) распределение вероятностей букв во всем тексте; 2) распределение позиционных вероятностей букв в словах [3] длиной от 4 букв и более; 3) распределение вероятностей длин слов; 4) средняя длина слова; 5) распределение вероятностей длин фраз; 6) средняя длина фразы.

Вероятности графем (табл. 1) определялись на материале 29 выборок из различных текстов по 1000 букв (включая дефис и апостроф) в каждой выборке. Общее количество букв, считая и пробел, составило более 35 000. В графе p_1 даны вероятности всех графем; в графе p_2 для удобства сравнения с материалами других авторов приводятся вероятности, пересчитанные на 1000 букв (без пробела, дефиса и апострофа).

Сравнение с данными о распределении вероятностей букв в английских текстах общего характера [4] показывает, что для текстов по радиоэлектронике специфична более высокая вероятность появления букв *t*, *c* и *i* и менее высокая — для букв *h*, *w*, *y* и *d*, а также для пробела. Это объясняется, с одной стороны, высоким удельным весом слов латинского происхождения, в частности слов с аффиксами, содержащими буквы *c*, *t*, *i* (com-, in-, -ic, -ance, -at, -ion и т. п.), и высокочастотных специальных терминов (capacity, circuit, communication, component, current и др.); с другой стороны, — меньшей частотностью служебных слов, наречий и местоимений, содержащих буквы *w*, *h* и *y* (what, when, where, was, were, we, he, his, you, your, my и т. п.), а также меньшей употребительностью времен Past Indefinite, Present и Past Perfect, использующих глагольные формы с суффиксом -ed. Тот факт, что пробел имеет сравнительно малую вероятность в научно-технических текстах, говорит о бо́льшей длине слова в этих текстах см. ниже.

Поскольку особый интерес представляет характер распределения позиционных вероятностей букв в знаменательных словах, а также в словах, обладающих флексиями (а такие слова в английском языке имеют, как правило, длину более 3 букв), во всех словах текста длиной от 4 букв и более были определены вероятности появления букв в пяти позициях от начала и от конца слова (табл. 2). Вероятности появления букв в начальной и конечной позициях определялись на материале 7 выборок по 1000 слов в каждой; для остальных позиций использовалось по 2 выборки.

Вероятности длин слов получены на 6 выборках общим объемом свыше 10 000 слов (табл. 3). При подсчете длины слова сложное слово с дефисным написанием считалось за одно слово, а дефис и апостроф рассматривались как обычные буквы.

Таблица 3

Распределение вероятностей длин слов

ДС	$\bar{x} \pm s$	p	ДС	$\bar{x} \pm s$	p	ДС	$\bar{x} \pm s$	p
1	80±5	0.046	8	97±7	0.056	15	3±1	0.0016
2	330±18	0.190	9	116±9	0.067	16	2±1	0.0009
3	322±16	0.186	10	63±3	0.036	17	0.2±0.01	0.0001
4	197±15	0.114	11	39±4	0.023	18	0.3±0.03	0.0002
5	160±15	0.092	12	22±3	0.013	19	2±1	0.0012
6	108±5	0.062	13	20±3	0.012	20	0.3±0.03	0.0002
7	161±13	0.093	14	9±3	0.0055	22	0.5±0.5	0.0003

Как видно из таблицы, для текстов по радиоэлектронике (как, возможно, для научно-технических текстов вообще) характерна более высокая по сравнению с художественными текстами вероятность употребления длинных слов и более низкая — слов средней длины (от 3 до 6 букв).

Относительно высокая вероятность 7- и 9-буквенных слов является не случайной: разность между частотой каждого из этих двух классов слов и частотой любого из соседних классов превышает свою стандартную ошибку (s) более чем втрое. Такое распределение вероятностей длин слов является, по-видимому, специфичным для подъязыка радиоэлектроники, где имеется

[3] Под «словом» здесь и далее понимается отрезок текста, заключенный между двумя пробелами.

[4] Л. Бриллюэн, ук. соч., стр. 24; G. Herdan, ук. соч., стр. 74, 76, 134; F. Pratt, ук. соч., стр. 252—258.

много высокочастотных специальных терминов (carrier, cathode, circuit, current, emitter; capacitor, collector, conductor и др.).

Средняя длина слова вычислялась отдельно по каждому из двух массивов текста, использовавшихся при определении вероятностей графем и длин слов. Полученные величины — 5.15 ± 0.10 и 5.05 ± 0.03 буквы хорошо согласуются между собой: разность между ними превышает свою ошибку менее чем вдвое (достоверность различия средних $p < 0.95$). Это позволяет объединить оба массива и вычислить среднюю длину слова по всему

Таблица 4

Распределение вероятностей длин фраз

ДФ	p	ДФ	p	ДФ	p	ДФ	p
1	0.003	20	0.034	39	0.010	58	0.0014
2	0.014	21	0.038	40	0.009	59	0.0010
3	0.008	22	0.036	41	0.008	60	0.0008
4	0.009	23	0.032	42	0.007	61	0.0002
5	0.014	24	0.031	43	0.008	62	0.0006
6	0.013	25	0.030	44	0.003	63	0.0006
7	0.017	26	0.026	45	0.004	64	0.0006
8	0.019	27	0.029	46	0.005	66	0.0006
9	0.020	28	0.033	47	0.003	67	0.0002
10	0.026	29	0.026	48	0.005	68	0.0006
11	0.028	30	0.023	49	0.003	69	0.0002
12	0.030	31	0.017	50	0.004	70	0.0002
13	0.033	32	0.016	51	0.003	71	0.0006
14	0.035	33	0.020	52	0.003	72	0.0002
15	0.035	34	0.015	53	0.002	74	0.0002
16	0.035	35	0.016	54	0.002	76	0.0002
17	0.037	36	0.015	55	0.002	78	0.0006
18	0.036	37	0.012	56	0.0008	83	0.0002
19	0.035	38	0.012	57	0.0008	88	0.0002
						91	0.0002

объему текста. Она составляет 5.08 ± 0.08 буквы, или округленно 5.1 буквы. Эта величина существенно отличается от величин, полученных для художественных текстов — 4.55 буквы [5] и для текстов общего характера — 4.5 и 4.1 буквы,[6] а также для телефонных разговоров — 3.5 буквы.[7]

Вероятности длин фраз (табл. 4) и средняя длина фразы определялись на 5 выборках по 1000 фраз в каждой. Общий объем текста около 105 000 слов, откуда средняя длина фразы — 21 слово.

[5] Л. В. Малаховский и Т. А. Селютина. Сб. «Статистико-комбинаторное моделирование языков», изд. «Наука», М.—Л., 1965, стр. 318—326.

[6] F. Pratt, ук. соч.; Л. Бриллюэн, ук. соч.

[7] N. R. French, C. W. Carter, Jr., W. Koenig, Jr., ук. соч.

Наибольшей вероятностью обладают также фразы длиной в 21 слово. Такое совпадение говорит о том, что по характеру распределения длин фраз рассматриваемые тексты отличаются большой однородностью, свидетельствующей об их статистической однородности (ср. с данными по художественным текстам, полученными на материале русского языка [8] и на материале болгарского [9]).

Результаты, представленные в настоящей работе, могут оказаться полезными при решении ряда задач теоретического и прикладного языкознания (моделирование языка, проблема выделения подъязыков, машинный перевод, автоматическое реферирование, компрессия речи и др.).

[8] Г. А. Лесскис. ВЯ, 1963, № 3, стр. 92—112.

[9] Г. Я. Мартыненко. Сб. «Статистико-комбинаторное моделирование языков. Изд. «Наука», М.—Л., 1965, стр. 327—339.

Л. А. Новак

СТАТИСТИКА БУКВ И БУКВОСОЧЕТАНИЙ В РУМЫНСКОМ ПИСЬМЕННОМ ТЕКСТЕ

Ниже приводятся результаты статистического исследования употребления букв и буквосочетаний в румынских письменных текстах.

Исследовалось 100 текстов по 400 печатных знаков (буквы и пробелы между словами), взятых из четырех функциональных стилей современного румынского языка в следующей пропорции.

1) разговорный стиль (диалогическая речь в современной художественной литературе) — 30 текстов;
2) беллетристический стиль — 30 текстов;
3) научно-публицистический стиль — 30 текстов;
4) поэзия — 10 текстов.

Некоторые результаты этого исследования обобщены в табл. 1—4.

Таблица 1

Распределение частот появления отдельных букв на 100 букв по сравнению с данными Николау *

Буквы	Данные Николау	Данные Новак	Буквы	Данные Николау	Данные Новак
a	8.30	7.82	*m*	3,08	2.94
ă	2.55	3.46	*n*	4 46	5.15
b	0.73	0.83	*o*	1.97	3 43
c	4.34	4.30	*p*	2.08	2.55
d	3.22	2.81	*r*	4.86	5.41
e	9.80	9.37	*s*	1.93	3.38
f	0.65	1.04	*ş*	1.33	1.33
g	0.43	0.87	*t*	5.10	5.03

Таблица 2

Вероятность перехода от одной буквы к другой (× 1000)

	a	*ă*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*î*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*r*	*s*	*ş*	*t*	*ţ*	*u*	*v*	*z*	*x*	Δ	Σ
a	—	—	0.75	4.63	1.10	0.15	0.63	0.80	0.08	4.70	—	0.30	—	5.50	3.18	2.79	0.02	2.20	10.70	3.90	1.33	8.26	1.65	3.00	0.87	0.80	—	20.86	78.20
ă	—	—	0.05	1.25	0.60	0.02	0.02	0.25	0.02	0.70	—	0.10	—	0.40	1.10	0.55	—	0.38	3 40	1.05	0.38	1.63	0.43	0.60	0 20	1.00	—	20.52	34.65
b	1.07	1.00	—	—	0.05	0.92	—	—	—	1.32	0.10	—	—	0.42	—	0.03	1.05	—	0.55	0.22	—	—	0.05	1.00	—	0.02	—	0.49	8.29
c	5.60	5.20	—	0.02	—	7.95	—	—	1.90	3.70	2.00	—	0.02	0.50	0.15	0.02	3.33	—	1.38	—	0.02	1.15	0.20	7.05	—	—	—	2.27	42.46
d	2 90	0.80	—	—	—	13.13	—	0.02	—	3.65	0.10	—	—	—	0.05	0.05	1.90	—	1.08	—	—	—	—	2.63	0.05	—	—	1.75	28.10
e	9.00	—	0.45	2.83	0.93	0.42	0.35	1.13	0.02	2.53	0.02	0.13	—	5.50	2.08	4.35	0.55	1.45	6.33	4.08	1.60	2.20	0.73	1.00	1.05	1.70	0.40	42.90	93.73
f	1.60	1.33	—	—	—	1.13	—	—	—	2.25	0.38	—	—	0.75	—	—	1.45	—	0.53	—	—	0.05	—	0.75	—	—	—	0.20	10.42
g	0.88	0.93	—	—	—	1.40	—	—	0.43	1.20	0.38	—	—	0.25	—	0.02	0.55	—	1.30	—	—	—	—	0.98	0.02	—	—	0.39	8.73
h	0.20	0.05	—	—	—	0.60	—	—	—	1.81	0.03	—	—	—	—	0.02	0.20	—	—	—	—	—	—	0.13	—	—	—	0.05	3.09
i	3.50	—	0.38	4.53	1.18	3.90	0.40	0.45	0.08	4 58	—	0.25	—	3.35	3.34	10.53	1.00	0.83	1.35	2.20	0.50	4.80	0.88	1.43	1.00	0.50	0.02	33.99	84.97
î	—	—	—	—	—	0.02	0.02	—	0.02	0.80	—	0.02	—	0.33	0.90	13.34	—	—	0.90	0.02	0.40	1.08	0.18	0.08	0.02	—	—	0.07	18.20
j	0.18	0.05	—	—	0.08	0.12	—	0.02	—	0.12	0.02	—	—	0.12	0.05	—	0.25	—	—	—	—	—	—	0.35	—	—	—	—	1.36
k	—	—	—	—	—	—	—	—	—	0.02	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.04	0.06
l	4.38	1.50	0.20	0.25	0.30	8.23	—	0.10	—	2.75	0.12	—	—	—	0.15	0.17	3.50	0.75	—	0.05	—	1.95	0.30	4.40	0.05	—	—	6.57	35.72
m	5.08	2.40	0.77	—	—	4.83	0.10	—	—	4.63	1.40	—	—	—	—	0.83	1 00	1.55	—	—	—	0.02	0.12	2.62	0 02	—	—	4.00	29.37
n	2.62	1.65	—	2.35	4.02	5.93	0.18	1.38	—	4.80	0.12	0.9	—	0.10	—	0.02	2.25	—	—	1.65	0.05	7.00	1.52	5.53	0.28	0.05	—	9.89	51.48
o	3.97	—	0.80	2.07	0.75	0.05	0.35	0.75	—	1.60	—	—	0.02	1.80	1.80	2.00	—	1.15	7.30	1.80	0.20	1.52	0.30	0.50	0.60	0.25	0.07	4.72	34.37
p	2.13	2.12	—	0.08	—	5.10	—	—	—	1.45	0.80	—	—	1.15	—	—	2.40	—	4.90	0.25	—	1.20	0.08	2.95	—	—	—	0.96	25.57
r	5.05	4.90	0.70	0.65	0.70	13.13	0.20	0.93	0.02	9.83	0.75	—	—	0.10	1.25	0.70	2.35	0.08	—	0.33	0.33	1.50	0.30	3.65	0.15	0.15	0.02	6.34	54.11
s	1.83	5.50	—	2.30	—	6.00	0.52	—	—	1.60	0.55	—	—	0.09	0.25	0.15	1.34	2.06	—	—	—	6.00	0.02	2.20	—	—	—	3.44	33.85
ş	0.30	0.08	—	0.10	—	0.40	—	—	—	8.40	—	—	—	—	0.10	0.13	0.13	—	—	—	—	2.90	—	0.30	—	—	—	0.43	13.27
t	4.90	4.23	—	0.05	0.08	11.92	0.13	—	—	4.50	1.12	0.02	—	0.05	0.05	0.05	4.00	—	5.80	—	—	—	—	3.90	0.02	—	—	9.47	50.29
ţ	0.73	0.93	—	—	—	1.02	—	—	—	5.50	0.02	—	—	—	—	—	—	—	—	—	—	—	—	0.25	—	—	—	0.15	8.60
u	0.70	0.33	0.70	1.13	0.38	0.05	0.20	0 22	0.07	3.60	0.02	0.05	—	7.83	3.80	6.45	0.02	0.93	4.90	1.20	0.56	3.15	0.52	—	0.30	0.40	—	12.74	50.25
v	2.15	1.00	—	—	—	2.50	—	—	—	2.07	0.48	—	—	—	—	0.02	1.03	—	0.72	—	—	—	—	0.10	—	—	—	0.29	10.37
z	0.43	0.53	0.25	—	0.18	0.65	—	0.08	—	1.80	0.02	—	—	0.08	0.02	0.20	0.15	—	0.02	—	—	—	—	0.55	0.28	—	—	0.61	5.85
x	—	0.02	—	0.02	—	0.02	—	—	—	0.20	—	—	—	—	—	—	—	0.19	—	—	—	0.06	—	—	—	—	—	—	0.51
Δ	19.00	0.10	3.24	20.20	17.75	4.15	7.32	2.60	0 45	4.86	9.77	0.40	0.02	7.40	11.10	9.06	5.90	14.00	2.94	17.10	7.90	5.82	1.32	4.30	5.46	0.98	—	—	183.14
Σ	78.20	34.65	8.29	42.46	28.10	93.73	10.42	8 73	3.09	84.97	18.20	1.36	0.06	35.72	29.37	51.48	34.37	25.57	54.11	33.85	13.27	50.29	8.60	50.25	10.37	5.85	0.51	183.14	999.01

Таблица 1 (*продолжение*)

Буквы	Данные Николау	Данные Новак	Буквы	Данные Николау	Данные Новак
h	0.53	0.31	*t*	0.51	0.86
i	10.30	8.50	*u*	5.80	5.02
î	1.27	1.82	*v*	0.98	1.04
j	0.11	0.14	*z*	0.61	0.58
k	—	0.006	*x*	0.08	0.05
l	3.74	3.57	Пробел . .	20.80	18.31

* Данные Николау основаны на исследовании примерно 3000 букв из газеты «Информация Букурештюлуй»; представлен только публицистический стиль прессы. См.: E. N i c o l a u. Cibernetica şi Lingvistica. Studii şi cercetari Lingvistice, 1958, 4, p. 481.

Таблица 3

Распределение первых 100 трехбуквенных сочетаний по убывающей абсолютной частоте (*F*)

Сочетание	Частота	Сочетание	Частота	Сочетание	Частота	Сочетание	Частота
△*de*	407	△*pe* . . .	143	*ră*△ . . .	96	*a*△*d* . . .	78
de△	363	*ai*△ . . .	142	*i*△*m* . .	95	*ar*△ . . .	78
△*în*	336	△*ce* . . .	141	*au*△ . . .	94	△*al* . . .	77
şi△	303	△*ma* . . .	138	*ţi*△ . . .	94	△*da* . . .	77
△*şi*	269	*i*△*a* . . .	138	*e*△*m* . . .	93	*est*	77
te△	262	△*nu* . . .	133	△*di* . . .	92	*pri*	77
le△	239	△*se* . . .	129	*la*△ . . .	92	*ă*△*c*. . . .	76
e△*c*	217	*i*△*s* . . .	128	*rea*	92	*car*	76
re△	202	△*a*△ . . .	127	△*ac*. . . .	91	*ă*△*s*. . . .	74
e△*a*. . . .	196	△*pr* . . .	126	△*la* . . .	91	*i*△*n*. . . .	74
ea△. . . .	196	*ce*△ . . .	122	*ri*△ . . .	91	*un*△ . . .	74
ul△. . . .	186	*cu*△ . . .	122	△*co* . . .	90	*ace* . . .	73
are	177	*tă*△ . . .	121	*ci*△ . . .	90	*ă*△*d* . . .	73
△*cu*	175	*ele*	118	*in*△. . . .	89	*din* . . .	73
△*să*. . . .	173	*or*△. . . .	117	*ntr*	89	*e*△*n* . . .	72
i△*d*. . . .	168	*nu*△ . . .	116	*ui*△. . . .	89	*ine*	72

Таблица 3 (*продолжение*)

Сочетание	Часто-та	Сочетание	Часто-та	Сочетание	Часто-та	Сочетание	Часто-та
e△s	161	*△o△* . . .	115	*△mi* . . .	88	*lui*	72
ii△	159	*△un* . . .	115	*at△* . . .	84	*ie△* . . .	70
e△d	158	*înt*	114	*uri*	84	*△ar* . . .	69
e△p	157	*△că*	109	*a△s* . . .	82	*ale*	69
△ca	153	*pe△* . . .	109	*e△î* . . .	82	*ă△n* . . .	69
în△	152	*i△p*	104	*a△c*	81	*ile*	69
i△c	150	*mai* . .	101	*înd* . .	81	*ste*	69
că△	148	*ate*	100	*△lu*	81	*e△o* . . .	68
se△	147	*ne△* . . .	98	*lor*	80	*△ne* . . .	68

Таблица 4

Распределение первых 100 сочетаний по позиции букв (б) и пробелов (△)

△бб	*△de;*	*△în;*	*△și;*	*△cu;*	*△să;*	*△ca;*	*△pe;*	*△ce;*	24
	△ma;	*△nu;*	*△se;*	*△pr;*	*△un;*	*△că;*	*△di;*	*△ac;*	
	△la;	*△co;*	*△mi;*	*△lu;*	*△al;*	*△da;*	*△ar;*	*△ne.*	
ббб△	*de△;*	*și△;*	*te△;*	*le△;*	*re△;*	*ea△;*	*ul△;*	*ii△;*	31
	în△;	*că△;*	*se△;*	*ai△;*	*ce△;*	*cu△;*	*tă△;*	*or△;*	
	nu△;	*pe△;*	*ne△;*	*ră△;*	*au△;*	*ti△;*	*la△;*	*ri△;*	
	ci△;	*in△;*	*ul△;*	*at△;*	*ar△;*	*un△;*	*ie△.*		
б△б	*e△c;*	*e△a;*	*i△d;*	*e△s;*	*e△d;*	*e△p;*	*i△c;*	*i△a;*	23
	i△s;	*i△p;*	*i△m;*	*e△m;*	*a△s;*	*e△î;*	*a△c;*	*a△d;*	
	ă△c;	*ă△s;*	*i△n;*	*ă△d;*	*e△n;*	*ă△n;*	*e△o.*		
ббб	*are;*	*ele;*	*înt;*	*mai;*	*ate;*	*rea;*	*nir;*	*uri;*	20
	înd;	*lor;*	*est;*	*pri;*	*car;*	*ace;*	*din;*	*ine;*	
	lui;	*ale;*	*ile;*	*ste.*					
△△△	■*△a△;*	*△o△.*							2

Путем обработки только что указанных статистических данных получены следующие средние значения энтропии, приходящиеся на букву в румынском письменном тексте:

$$H_1 = 4.13,\ H_2 = 3.36,\ H_3 = 2.69.$$ [1]

[1] Методика расчета значений безусловной и условной энтропии см. в работах: К. Шеннон. Математическая теория связи. Сб. «Работы по теории информации и кибернетике», М., 1963, стр. 261—263; А. А. Петровская, Р. Г. Петровский, К. А. Разживин. Энтропия русского языка. ВЯ, 1962, 6, стр. 115, 116.

Таблица значений $-p \log_2 p$ при $0.000001 \leqslant p \leqslant 0.001$

p	$-\log_2 p$	$-p \log_2 p$	p	$-\log_2 p$	$-p \log_2 p$
0.000001	19.93273	0.000020	0.000051	14.25918	0.000727
0.000002	18.93273	0.000038	0.000052	14.23117	0.000740
0.000003	18.34738	0.000055	0.000053	14.20369	0.000753
0.000004	17.93215	0.000072	0.000054	14.17672	0.000766
0.000005	17.61010	0.000088	0.000055	14.15025	0.000778
0.000006	17.34699	0.000104	0.000056	14.12425	0.000791
0.000007	17.12454	0.000120	0.000057	14.09871	0.000804
0.000008	16.93186	0.000135	0.000058	14.07362	0.000816
0.000009	16.76190	0.000151	0.000059	14.04896	0.000829
0.000010	16.60987	0.000166	0.000060	14.02471	0.000841
0.000011	16.47234	0.000181	0.000061	14.00086	0.000854
0.000012	16.34680	0.000196	0.000062	13.97740	0.000867
0.000013	16.23130	0.000211	0.000063	13.95432	0.000880
0.000014	16.12438	0.000226	0.000064	13.93160	0.000892
0.000015	16.02483	0.000240	0.000065	13.90923	0.000904
0.000016	15.93171	0.000255	0.000066	13.88720	0.000917
0.000017	15.84424	0.000269	0.000067	13.86551	0.000929
0.000018	15.76177	0.000284	0.000068	13.84414	0.000941
0.000019	15.68376	0.000298	0.000069	13.82307	0.000954
0.000020	15.60975	0.000312	0.000070	13.80231	0.000966
0.000021	15.53936	0.000326	0.000071	13.78185	0.000978
0.000022	15.47224	0.000340	0.000072	13.76167	0.000991
0.000023	15.40810	0.000354	0.000073	13.74177	0.001003
0.000024	15.34670	0.000368	0.000074	13.72214	0.001015
0.000025	15.28780	0.000382	0.000075	13.70278	0.001028
0.000026	15.23121	0.000396	0.000076	13.68367	0.001040
0.000027	15.17676	0.000410	0.000077	13.66481	0.001062
0.000028	15.12429	0.000423	0.000078	13.64619	0.001064
0.000029	15.07366	0.000437	0.000079	13.62781	0.001077
0.000030	15.02475	0.000450	0.000080	13.60966	0.001089
0.000031	14.97744	0.000464	0.000081	13.59174	0.001101
0.000032	14.93164	0.000478	0.000082	13.57404	0.001113
0.000033	14.88724	0.000491	0.000083	13.55655	0.001125
0.000034	14.84417	0.000505	0.000084	13.53927	0.001137
0.000035	14.80235	0.000518	0.000085	13.52220	0.001149
0.000036	14.76170	0.000531	0.000086	13.50533	0.001161
0.000037	14.72217	0.000545	0.000087	13.48865	0.001173
0.000038	14.68370	0.000558	0.000088	13.47216	0.001186
0.000039	14.64622	0.000571	0.000089	13.45586	0.001198
0.000040	14.60969	0.000584	0.000090	13.43974	0.001210
0.000041	14.57407	0.000598	0.000091	13.42379	0.001222
0.000042	14.53930	0.000611	0.000092	13.40803	0.001234
0.000043	14.60535	0.000624	0.000093	13.39243	0.001245
0.000044	14.47219	0.000637	0.000094	13.37700	0.001257
0.000045	14.43976	0.000650	0.000095	13.36173	0.001269
0.000046	14.40805	0.000663	0.000096	13.34663	0.001281
0.000047	14.37702	0.000676	0.000097	13.33167	0.001293
0.000048	14.34665	0.000689	0.000098	13.31688	0.001305
0.000049	14.31690	0.000702	0.000099	13.30223	0.001317
0.000050	14.28775	0.000714	0.000100	13.28773	0.001329

Примечание. Таблица составлена А. В. Зубовым на ЭВМ «Минск-2».

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	p	$-\log_2 p$	$-p\log_2 p$
0.000101	13.27338	0.001341	0.000154	12.66479	0.001950
0.000102	13.25916	0.001352	0.000155	12.65545	0.001962
0.000103	13.24509	0.001364	0.000156	12.64618	0.001973
0.000104	13.23115	0.001376	0.000157	12.63696	0.001984
0.000105	13.21734	0.001388	0.000158	12.62780	0.001996
0.000106	13.20367	0.001400	0.000159	12.61870	0.002006
0.000107	13.19012	0.001411	0.000160	12.60965	0.002018
0.000108	13.17670	0.001423	0.000161	12.60066	0.002029
0.000109	13.16340	0.001435	0.000162	12.59173	0.002040
0.000110	13.15023	0.001447	0.000163	12.58285	0.002051
0.000111	13.13717	0.001458	0.000164	12.57403	0.002062
0.000112	13.12423	0.001470	0.000165	12.56526	0.002073
0.000113	13.11141	0.001482	0.000166	12.55654	0.002084
0.000114	13.09869	0.001493	0.000167	12.54787	0.002095
0.000115	13.08609	0.001505	0.000168	12.53926	0.002107
0.000116	13.07360	0.001517	0.000169	12.53070	0.002118
0.000117	13.06122	0.001528	0.000170	12.52219	0.002129
0.000118	13.04894	0.001540	0.000171	12.51372	0.002140
0.000119	13.03677	0.001551	0.000172	12.50531	0.002151
0.000120	13.02469	0.001563	0.000173	12.49695	0.002162
0.000121	13.01272	0.001575	0.000174	12.48863	0.002173
0.000122	13.00085	0.001586	0.000175	12.48037	0.002184
0.000123	12.98907	0.001598	0.000176	12.47215	0.002195
0.000124	12.97739	0.001609	0.000177	12.46397	0.002206
0.000125	12.96580	0.001621	0.000178	12.45584	0.002217
0.000126	12.95430	0.001632	0.000179	12.44776	0.002228
0.000127	12.94290	0.001644	0.000180	12.43972	0.002239
0.000128	12.93158	0.001655	0.000181	12.43173	0.002250
0.000129	12.92035	0.001668	0.000182	12.42378	0.002261
0.000130	12.90921	0.001678	0.000183	12.41588	0.002272
0.000131	12.89816	0.001690	0.000184	12.40801	0.002283
0.000132	12.88719	0.001701	0.000185	12.40019	0.002294
0.000133	12.87630	0.001713	0.000186	12.39242	0.002304
0.000134	12.86549	0.001724	0.000187	12.38468	0.002316
0.000135	12.85477	0.001735	0.000188	12.37699	0.002327
0.000136	12.84412	0.001747	0.000189	12.36933	0.002337
0.000137	12.83355	0.001758	0.000190	12.36172	0.002349
0.000138	12.82306	0.001770	0.000191	12.35415	0.002360
0.000139	12.81264	0.001781	0.000192	12.34661	0.002371
0.000140	12.80230	0.001792	0.000193	12.33912	0.002381
0.000141	12.79203	0.001804	0.000194	12.33166	0.002392
0.000142	12.78183	0.001815	0.000195	12.32425	0.002403
0.000143	12.77171	0.001826	0.000196	12.31687	0.002414
0.000144	12.76165	0.001838	0.000197	12.30952	0.002425
0.000145	12.75167	0.001849	0.000198	12.30222	0.002436
0.000146	12.74176	0.001860	0.000199	12.29495	0.002447
0.000147	12.73191	0.001872	0.000200	12.28772	0.002458
0.000148	12.72213	0.001883	0.000201	12.28052	0.002468
0.000149	12.71241	0.001894	0.000202	12.27336	0.002479
0.000150	12.70276	0.001905	0.000203	12.26624	0.002490
0.000151	12.69317	0.001917	0.000204	12.25915	0.002501
0.000152	12.68365	0.001928	0.000205	12.25209	0.002512
0.000153	12.67419	0.001939	0.000206	12.24507	0.002522

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	p	$-\log_2 p$	$-p\log_2 p$
0.000207	12.23809	0.002533	0.000260	11.90920	0.003096
0.000208	12.23114	0.002544	0.000261	11.90367	0.003107
0.000209	12.22422	0.002555	0.000262	11.89815	0.003117
0.000210	12.21733	0.002566	0.000263	11.89265	0.003128
0.000211	12.21048	0.002576	0.000264	11.88718	0.003138
0.000212	12.20365	0.002587	0.000265	11.88172	0.003149
0.000213	12.19686	0.002598	0.000266	11.87629	0.003159
0.000214	12.19011	0.002607	0.000267	11.87088	0.003169
0.000215	12.18338	0.002619	0.000268	11.86548	0.003180
0.000216	12.17669	0.002630	0.000269	11.86011	0.003190
0.000217	12.17002	0.002641	0.000270	11.85476	0.003201
0.000218	12.16339	0.002652	0.000271	11.84942	0.003211
0.000219	12.15679	0.002662	0.000272	11.84411	0.003222
0.000220	12.15021	0.002673	0.000273	11.83881	0.003232
0.000221	12.14367	0.002683	0.000274	11.83354	0.003242
0.000222	12.13716	0.002694	0.000275	11.82828	0.003253
0.000223	12.13067	0.002705	0.000276	11.82305	0.003263
0.000224	12.12422	0.002716	0.000277	11.81783	0.003274
0.000225	12.11779	0.002726	0.000278	11.81263	0.003284
0.000226	12.11139	0.002737	0.000279	11.80745	0.003294
0.000227	12.10503	0.002748	0.000280	11.80229	0.003305
0.000228	12.09868	0.002758	0.000281	11.79715	0.003315
0.000229	12.09237	0.002769	0.000282	11.79202	0.003325
0.000230	12.08608	0.002780	0.000283	11.78691	0.003336
0.000231	12.07982	0.002790	0.000284	11.78182	0.003346
0.000232	12.07359	0.002801	0.000285	11.77675	0.003356
0.000233	12.06739	0.002812	0.000286	11.77170	0.003367
0.000234	12.06121	0.002822	0.000287	11.76666	0.003377
0.000235	12.05506	0.002833	0.000288	11.76165	0.003387
0.000236	12.04893	0.002844	0.000289	11.75665	0.003398
0.000237	12.04283	0.002854	0.000290	11.75166	0.003408
0.000238	12.03676	0.002865	0.000291	11.74670	0.003418
0.000239	12.03071	0.002875	0.000292	11.74175	0.003429
0.000240	12.02468	0.002886	0.000293	11.73681	0.003439
0.000241	12.01868	0.002896	0.000294	11.73190	0.003449
0.000242	12.01271	0.002907	0.000295	11.72700	0.003459
0.000243	12.00676	0.002918	0.000296	11.72212	0.003470
0.000244	12.00084	0.002928	0.000297	11.71725	0.003480
0.000245	11.99493	0.002939	0.000298	11.71240	0.003490
0.000246	11.98906	0.002949	0.000299	11.70757	0.003501
0.000247	11.98321	0.002960	0.000300	11.70275	0.003511
0.000248	11.97738	0.002970	0.000301	11.69795	0.003521
0.000249	11.97157	0.002981	0.000302	11.69317	0.003531
0.000250	11.96579	0.002991	0.000303	11.68840	0.003542
0.000251	11.96003	0.003002	0.000304	11.68364	0.003552
0.000252	11.95429	0.003012	0.000305	11.67891	0.003562
0.000253	11.94858	0.003023	0.000306	11.67418	0.003572
0.000254	11.94289	0.003033	0.000307	11.66948	0.003583
0.000255	11.93722	0.003044	0.000308	11.66478	0.003593
0.000256	11.93157	0.003054	0.000309	11.66011	0.003603
0.000257	11.92595	0.003065	0.000310	11.65545	0.003613
0.000258	11.92035	0.003075	0.000311	11.65080	0.003623
0.000259	11.91476	0.003086	0.000312	11.64617	0.003634

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	p	$-\log_2 p$	$-p \log_2 p$
0.000313	11.64155	0.003644	0.000366	11.41587	0.004178
0.000314	11.63695	0.003654	0.000367	11.41193	0.004188
0.000315	11.63236	0.003664	0.000368	11.40801	0.004198
0.000316	11.62779	0.003674	0.000369	11.40409	0.004208
0.000317	11.62323	0.003685	0.000370	11.40019	0.004218
0.000318	11.61869	0.003695	0.000371	11.39629	0.004228
0.000319	11.61416	0.003705	0.000372	11.39241	0.004238
0.000320	11.60964	0.003715	0.000373	11.38854	0.004248
0.000321	11.60514	0.003725	0.000374	11.38468	0.004258
0.000322	11.60065	0.003735	0.000375	11.38082	0.004268
0.000323	11.59618	0.003746	0.000376	11.37698	0.004278
0.000324	11.59172	0.003756	0.000377	11.37315	0.004288
0.000325	11.58727	0.003766	0.000378	11.36933	0.004298
0.000326	11.58284	0.003776	0.000379	11.36552	0.004308
0.000327	11.57842	0.003786	0.000380	11.36171	0.004317
0.000328	11.57402	0.003796	0.000381	11.35792	0.004327
0.000329	11.56963	0.003806	0.000382	11.35414	0.004337
0.000330	11.56525	0.003817	0.000383	11.35037	0.004347
0.000331	11.56088	0.003827	0.000384	11.34661	0.004357
0.000332	11.55653	0.003837	0.000385	11.34285	0.004367
0.000333	11.55219	0.003847	0.000386	11.33911	0.004377
0.000334	11.54787	0.003857	0.000387	11.33538	0.004387
0.000335	11.54355	0.003867	0.000388	11.33166	0.004397
0.000336	11.53925	0.003877	0.000389	11.32794	0.004407
0.000337	11.53497	0.003887	0.000390	11.32424	0.004416
0.000338	11.53069	0.003897	0.000391	11.32054	0.004426
0.000339	11.52643	0.003907	0.000392	11.31686	0.004436
0.000340	11.52218	0.003918	0.000393	11.31318	0.004446
0.000341	11.51794	0.003928	0.000394	11.30952	0.004456
0.000342	11.51372	0.003938	0.000395	11.30586	0.004466
0.000343	11.50951	0.003948	0.000396	11.30221	0.004476
0.000344	11.50531	0.003958	0.000397	11.29857	0.004486
0.000345	11.50112	0.003968	0.000398	11.29494	0.004495
0.000346	11.49694	0.003978	0.000399	11.29132	0.004505
0.000347	11.49278	0.003988	0.000400	11.28771	0.004515
0.000348	11.48863	0.003998	0.000401	11.28411	0.004525
0.000349	11.48449	0.004008	0.000402	11.28052	0.004535
0.000350	11.48036	0.004018	0.000403	11.27693	0.004545
0.000351	11.47624	0.004028	0.000404	11.27336	0.004554
0.000352	11.47214	0.004038	0.000405	11.26979	0.004564
0.000353	11.46805	0.004048	0.000406	11.26623	0.004574
0.000354	11.46396	0.004058	0.000407	11.26268	0.004584
0.000355	11.45989	0.004068	0.000408	11.25914	0.004594
0.000356	11.45584	0.004078	0.000409	11.25561	0.004604
0.000357	11.45179	0.004088	0.000410	11.25209	0.004613
0.000358	11.44775	0.004098	0.000411	11.24857	0.004623
0.000359	11.44373	0.004108	0.000412	11.24507	0.004633
0.000360	11.43972	0.004118	0.000413	11.24157	0.004643
0.000361	11.43571	0.004128	0.000414	11.23808	0.004653
0.000362	11.43172	0.004138	0.000415	11.23460	0.004662
0.000363	11.42774	0.004148	0.000416	11.23113	0.004672
0.000364	11.42378	0.004158	0.000417	11.22767	0.004682
0.000365	11.41982	0.004168	0.000418	11.22421	0.004692

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	p	$-\log_2 p$	$-p \log_2 p$
0.000419	11.22076	0.004701	0.000472	11.04893	0.005215
0.000420	11.21732	0.004711	0.000473	11.04587	0.005225
0.000421	11.21389	0.004721	0.000474	11.04282	0.005234
0.000422	11.21047	0.004731	0.000475	11.03978	0.005244
0.000423	11.20705	0.004741	0.000476	11.03675	0.005253
0.000424	11.20365	0.004750	0.000477	11.03372	0.005263
0.000425	11.20025	0.004760	0.000478	11.03070	0.005273
0.000426	11.19686	0.004770	0.000479	11.02769	0.005282
0.000427	11.19348	0.004780	0.000480	11.02469	0.005291
0.000428	11.19010	0.004789	0.000481	11.02168	0.005301
0.000429	11.18673	0.004799	0.000482	11.01868	0.005311
0.000430	11.18338	0.004809	0.000483	11.01569	0.005321
0.000431	11.18002	0.004819	0.000484	11.01270	0.005330
0.000432	11.17668	0.004828	0.000485	11.00973	0.005340
0.000433	11.17335	0.004838	0.000486	11.00676	0.005349
0.000434	11.17002	0.004848	0.000487	11.00379	0.005359
0.000435	11.16670	0.004857	0.000488	11.00083	0.005368
0.000436	11.16338	0.004867	0.000489	10.99788	0.005378
0.000437	11.16008	0.004877	0.000490	10.99493	0.005387
0.000438	11.15678	0.004887	0.000491	10.99199	0.005397
0.000439	11.15349	0.004896	0.000492	10.98905	0.005407
0.000440	11.15021	0.004906	0.000493	10.98612	0.005416
0.000441	11.14693	0.004916	0.000494	10.98320	0.005426
0.000442	11.14367	0.004925	0.000495	10.98028	0.005435
0.000443	11.14041	0.004935	0.000496	10.97737	0.005445
0.000444	11.13715	0.004945	0.000497	10.97447	0.005454
0.000445	11.13391	0.004955	0.000498	10.97157	0.005464
0.000446	11.13067	0.004964	0.000499	10.96867	0.005473
0.000447	11.12744	0.004974	0.000500	10.96578	0.005483
0.000448	11.12421	0.004984	0.000501	10.96290	0.005492
0.000449	11.12100	0.004993	0.000502	10.96002	0.005502
0.000450	11.11779	0.005003	0.000503	10.95715	0.005511
0.000451	11.11458	0.005013	0.000504	10.95429	0.005521
0.000452	11.11139	0.005022	0.000505	10.95143	0.005530
0.000453	11.10820	0.005032	0.000506	10.94857	0.005540
0.000454	11.10502	0.005042	0.000507	10.94573	0.005549
0.000455	11.10185	0.005051	0.000508	10.94288	0.005559
0.000456	11.09868	0.005061	0.000509	10.94005	0.005568
0.000457	11.09552	0.005071	0.000510	10.93721	0.005578
0.000458	11.09236	0.005080	0.000511	10.93439	0.005587
0.000459	11.08922	0.005090	0.000512	10.93157	0.005597
0.000460	11.08608	0.005100	0.000513	10.92875	0.005606
0.000461	11.08295	0.005109	0.000514	10.92594	0.005616
0.000462	11.07982	0.005119	0.000515	10.92314	0.005625
0.000463	11.07670	0.005128	0.000516	10.92034	0.005635
0.000464	11.07359	0.005138	0.000517	10.91755	0.005644
0.000465	11.07048	0.005148	0.000518	10.91476	0.005654
0.000466	11.06738	0.005157	0.000519	10.91198	0.005663
0.000467	11.06429	0.005167	0.000520	10.90920	0.005673
0.000468	11.06120	0.005177	0.000521	10.90643	0.005682
0.000469	11.05812	0.005186	0.000522	10.90366	0.005692
0.000470	11.05505	0.005196	0.000523	10.90090	0.005701
0.000471	11.05198	0.005205	0.000524	10.89814	0.005711

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	p	$-\log_2 p$	$-p\log_2 p$
0,000525	10.89539	0.005720	0.000578	10.75664	0.006217
0.000526	10.89265	0.005730	0.000579	10.75415	0.006227
0.000527	10.88991	0.005739	0.000580	10.75166	0.006236
0.000528	10.88717	0.005748	0.000581	10.74917	0.006245
0.000529	10.88444	0.005758	0.000582	10.74669	0.006255
0.000530	10.88172	0.005767	0.000583	10.74422	0.006264
0.000531	10.87900	0.005777	0.000584	10.74174	0.006273
0.000532	10.87629	0.005786	0.000585	10.73927	0.006282
0.000533	10.87358	0.005796	0.000586	10.73681	0.006292
0.000534	10.87087	0.005805	0.000587	10.73435	0.006301
0.000535	10.86817	0.005814	0.000588	10.73189	0.006310
0.000536	10.86548	0.005824	0.000589	10.72944	0.006320
0.000537	10.86279	0.005833	0.000590	10.72700	0.006329
0.000538	10.86011	0.005843	0.000591	10.72455	0.006338
0.000539	10.85743	0.005852	0.000592	10.72211	0.006347
0.000540	10.85475	0.005862	0.000593	10.71968	0.006357
0.000541	10.85208	0.005871	0.000594	10.71725	0.006366
0.000542	10.84942	0.005880	0.000595	10.71482	0.006375
0.000543	10.84676	0.005890	0.000596	10.71240	0.006385
0.000544	10.84410	0.005899	0.000597	10.70998	0.006394
0.000545	10.84146	0.005909	0.000598	10.70757	0.006403
0.000546	10.83881	0.005918	0.000599	10.70515	0.006412
0.000547	10.83617	0.005927	0.000600	10.70275	0.006422
0.000548	10.83354	0.005937	0.000601	10.70035	0.006431
0.000549	10.83091	0.005946	0.000602	10.69795	0.006440
0.000550	10.82828	0.005956	0.000603	10.69555	0.006449
0.000551	10.82566	0.005965	0.000604	10.69316	0.006459
0.000552	10.82304	0.005974	0.000605	10.69078	0.006468
0.000553	10.82043	0.005984	0.000606	10.68839	0.006477
0.000554	10.81783	0.005993	0.000607	10.68601	0.006486
0.000555	10.81522	0.006002	0.000608	10.68364	0.006496
0.000556	10.81263	0.006012	0.000609	10.68127	0.006505
0.000557	10.81003	0.006021	0.000610	10.67890	0.006514
0.000558	10.80745	0.006031	0.000611	10.67654	0.006523
0.000559	10.80486	0.006040	0.000612	10.67418	0.006533
0.000560	10.80228	0.006049	0.000613	10.67182	0.006542
0.000561	10.79971	0.006059	0.000614	10.66947	0.006551
0.000562	10.79714	0.006068	0.000615	10.66712	0.006560
0.000563	10.79458	0.006077	0.000616	10.66478	0.006569
0.000564	10.79202	0.006087	0.000617	10.66244	0.006579
0.000565	10.78946	0.006096	0.000618	10.66010	0.006588
0.000566	10.78691	0.006105	0.000619	10.65777	0.006597
0.000567	10.78436	0.006115	0.000620	10.65544	0.006606
0.000568	10.78182	0.006124	0.000621	10.65312	0.006616
0.000569	10.77928	0.006133	0.000622	10.65080	0.006625
0.000570	10.77675	0.006143	0.000623	10.64848	0.006634
0.000571	10.77422	0.006152	0.000624	10.64616	0.006643
0.000572	10.77170	0.006161	0.000625	10.64385	0.006652
0.000573	10.76918	0.006171	0.000626	10.64155	0.006662
0.000574	10.76666	0.006180	0.000627	10.63925	0.006671
0.000575	10.76415	0.006189	0.000628	10.63695	0.006680
0.000576	10.76164	0.006199	0.000629	10.63465	0.006689
0.000577	10.75914	0.006208	0.000630	10.63236	0.006698

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	p	$-\log_2 p$	$-p\log_2 p$
0,000631	10.63007	0.006708	0.000684	10.51371	0.007191
0,000632	10.62779	0.006717	0.000685	10.51161	0.007200
0,000633	10.62551	0.006726	0.000686	10.50950	0.007210
0,000634	10.62323	0.006735	0.000687	10.50740	0.007219
0,000635	10.62095	0.006744	0.000688	10.50530	0.007228
0,000636	10.61868	0.006753	0.000689	10.50321	0.007237
0,000637	10.61642	0.006763	0.000690	10.50111	0.007246
0,000638	10.61415	0.006772	0.000691	10.49902	0.007255
0,000639	10.61189	0.006781	0.000692	10.49694	0.007264
0,000640	10.60964	0.006790	0.000693	10.49486	0.007273
0,000641	10.60739	0.006799	0.000694	10.49277	0.007282
0,000642	10.60514	0.006808	0.000695	10.49070	0.007291
0,000643	10.60289	0.006818	0.000696	10.48862	0.007300
0,000644	10.60065	0.006827	0.000697	10.48655	0.007309
0,000645	10.59841	0.006836	0.000698	10.48448	0.007318
0,000646	10.59618	0.006845	0.000699	10.48242	0.007327
0,000647	10.59394	0.006854	0.000700	10.48036	0.007336
0,000648	10.59172	0.006863	0.000701	10.47830	0.007345
0,000649	10.58949	0.006873	0.000702	10.47624	0.007354
0,000650	10.58727	0.006882	0.000703	10.47419	0.007363
0,000651	10.58505	0.006891	0.000704	10.47213	0.007372
0,000652	10.58284	0.006900	0.000705	10.47009	0.007381
0,000653	10.58063	0.006909	0.000706	10.46804	0.007390
0,000654	10.57842	0.006918	0.000707	10.46600	0.007399
0,000655	10.57622	0.006927	0.000708	10.46396	0.007408
0,000656	10.57401	0.006937	0.000709	10.46192	0.007417
0,000657	10.57182	0.006946	0.000710	10.45989	0.007427
0,000658	10.56962	0.006955	0.000711	10.45786	0.007436
0,000659	10.56743	0.006964	0.000712	10.45583	0.007445
0,000660	10.56524	0.006973	0.000713	10.45381	0.007454
0,000661	10.56306	0.006982	0.000714	10.45179	0.007463
0,000662	10.56088	0.006991	0.000715	10.44977	0.007472
0,000663	10.55870	0.007000	0.000716	10.44775	0.007481
0,000664	10.55653	0.007010	0.000717	10.44574	0.007490
0,000665	10.55436	0.007019	0.000718	10.44373	0.007499
0,000666	10.55219	0.007028	0.000719	10.44172	0.007508
0,000667	10.55002	0.007037	0.000720	10.43971	0.007517
0,000668	10.54786	0.007046	0.000721	10.43771	0.007526
0,000669	10.54570	0.007055	0.000722	10.43571	0.007535
0,000670	10.54355	0.007064	0.000723	10.43371	0.007544
0,000671	10.54140	0.007073	0.000724	10.43172	0.007553
0,000672	10.53925	0.007082	0.000725	10.42973	0.007562
0,000673	10.53710	0.007091	0.000726	10.42774	0.007571
0,000674	10.53496	0.007101	0.000727	10.42575	0.007580
0,000675	10.53282	0.007110	0.000728	10.42377	0.007588
0,000676	10.53069	0.007119	0.000729	10.42179	0.007597
0,000677	10.52855	0.007128	0.000730	10.41981	0.007606
0,000678	10.52643	0.007137	0.000731	10.41784	0.007615
0,000679	10.52430	0.007146	0.000732	10.41587	0.007624
0,000680	10.52218	0.007155	0.000733	10.41390	0.007633
0,000681	10.52006	0.007164	0.000734	10.41193	0.007642
0,000682	10.51794	0.007173	0.000735	10.40997	0.007651
0,000683	10.51582	0.007182	0.000736	10.40800	0.007660

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	p	$-\log_2 p$	$-p \log_2 p$
0,000737	10.40605	0,007669	0.000790	10.30586	0.008142
0,000738	10.40409	0,007678	0.000791	10.30403	0.008150
0,000739	10.40214	0,007687	0.000792	10.30221	0.008159
0,000740	10.40018	0,007696	0.000793	10.30039	0.008168
0,000741	10.39824	0,007705	0.000794	10.29857	0.008177
0,000742	10.39629	0,007714	0.000795	10.29676	0.008186
0.000743	10.39435	0.007723	0.000796	10.29494	0.008195
0,000744	10.39241	0,007732	0.000797	10.29313	0.008204
0,000745	10.39047	0,007741	0.000798	10.29132	0.008212
0,000746	10.38853	0,007750	0.000799	10.28951	0.008221
0,000747	10.38660	0,007759	0.000800	10.28771	0.008230
0,000748	10.38467	0,007768	0.000801	10.28591	0.008239
0,000749	10.38274	0,007777	0.000802	10.28411	0.008248
0,000750	10.38082	0,007786	0.000803	10.28231	0.008257
0,000751	10.37890	0,00779	0.000804	10.28051	0.008266
0,000752	10.37698	0,007803	0.000805	10.27872	0.008274
0,000753	10.37506	0,007812	0.000806	10.27693	0.008283
0,000754	10.37315	0,007821	0.000807	10.27514	0.008292
0,000755	10.37123	0,007830	0.000808	10.27335	0.008301
0,000756	10.36932	0,007839	0.000809	10.27157	0.008310
0,000757	10.36742	0,007848	0.000810	10.26979	0.008319
0,000758	10.36551	0,007857	0.000811	10.26801	0.008327
0,000759	10.36361	0,007866	0.000812	10.26623	0.008336
0,000760	10.36171	0,007875	0.000813	10.26445	0.008345
0,000761	10.35981	0,007884	0.000814	10.26268	0.008354
0,000762	10.35792	0,007893	0.000815	10.26091	0.008363
0,000763	10.35603	0,007902	0.000816	10.25914	0.008371
0,000764	10.35414	0,007911	0.000817	10.25737	0.008380
0,000765	10.35225	0,007919	0.000818	10.25561	0.008389
0,000766	10.35037	0,007928	0.000819	10.25385	0.008398
0,000767	10.34848	0,007937	0.000820	10.25209	0.008407
0,000768	10.34660	0,007946	0.000821	10.25033	0.008416
0,000769	10.34473	0,007955	0.000822	10.24857	0.008424
0,000770	10.34285	0,007964	0.000823	10.24682	0.008433
0,000771	10.34098	0,007973	0.000824	10.24507	0.008442
0,000772	10.33911	0,007982	0.000825	10.24332	0.008451
0,000773	10.33724	0,007991	0.000826	10.24157	0.008460
0,000774	10.33538	0,008000	0.000827	10.23982	0.008468
0,000775	10.33351	0,008008	0.000828	10.23808	0.008477
0,000776	10.33165	0,008017	0.000829	10.23634	0.008486
0,000777	10.32980	0,008026	0.000830	10.23460	0.008495
0,000778	10.32794	0,008035	0.000831	10.23286	0.008503
0,000779	10.32609	0,008044	0.000832	10.23113	0.008512
0,000780	10.32424	0,008053	0.000833	10.22939	0.008521
0,000781	10.32239	0,008062	0.000834	10.22766	0.008530
0,000782	10.32054	0,008071	0.000835	10.22593	0.008539
0,000783	10.31870	0,008080	0.000836	10.22421	0.008547
0,000784	10.31686	0,008088	0.000837	10.22248	0.008556
0,000785	10.31502	0,008097	0.000838	10.22076	0.008565
0,000786	10.31318	0,008106	0.000839	10.21904	0.008574
0,000787	10.31135	0,008115	0.000840	10.21732	0.008583
0,000788	10.30951	0,008124	0.000841	10.21560	0.008591
0,000789	10.30768	0,008133	0.000842	10.21389	0.008600

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	p	$-\log_2 p$	$-p \log_2 p$
0,000843	10.21218	0,008609	0,000896	10.12421	0,009071
0,000844	10.21047	0,008618	0,000897	10.12260	0,009080
0,000845	10.20876	0,008626	0,000898	10.12099	0,009089
0,000846	10.20705	0,008635	0,000899	10.11939	0,009097
0,000847	10.20535	0,008644	0,000900	10.11778	0,009106
0,000848	10.20365	0,008653	0,000901	10.11618	0,009115
0,000849	10.20195	0,008661	0,000902	10.11458	0,009123
0,000850	10.20025	0,008670	0,000903	10.11298	0,009132
0,000851	10.19855	0,008679	0,000904	10.11139	0,009141
0,000852	10.19686	0,008688	0,000905	10.10979	0,009149
0,000853	10.19516	0,008696	0,000906	10.10820	0,009158
0,000854	10.19347	0,008705	0,000907	10.10661	0,009167
0,000855	10.19179	0,008714	0,000908	10.10502	0,009175
0,000856	10.19010	0,008723	0,000909	10.10343	0,009184
0,000857	10.18841	0,008731	0,000910	10.10184	0,009193
0,000858	10.18673	0,008740	0,000911	10.10026	0,009201
0,000859	10.18505	0,008749	0,000912	10.09868	0,009210
0,000860	10.18337	0,008758	0,000913	10.09709	0,009219
0,000861	10.18170	0,008766	0,000914	10.09552	0,009227
0,000862	10.18002	0,008775	0,000915	10.09394	0,009236
0,000863	10.17835	0,008784	0,000916	10.09236	0,009245
0,000864	10.17668	0,008793	0,000917	10.09079	0,009253
0,000865	10.17501	0,008801	0,000918	10.08922	0,009262
0,000866	10.17334	0,008810	0,000919	10.08764	0,009271
0,000867	10.17168	0,008819	0,000920	10.08608	0,009279
0,000868	10.17001	0,008828	0,000921	10.08451	0,009288
0,000869	10.16835	0,008836	0,000922	10.08294	0,009296
0,000870	10.16669	0,008845	0,000923	10.08138	0,009305
0,000871	10.16504	0,008854	0,000924	10.07982	0,009314
0,000872	10.16338	0,008862	0,000925	10.07826	0,009322
0,000873	10.16173	0,008871	0,000926	10.07670	0,009331
0,000874	10.16008	0,008880	0,000927	10.07514	0,009340
0,000875	10.15843	0,008889	0,000928	10.07358	0,009348
0,000876	10.15678	0,008898	0,000929	10.07203	0,009357
0,000877	10.15513	0,008906	0,000930	10.07048	0,009366
0,000878	10.15349	0,008915	0,000931	10.06893	0,009374
0,000879	10.15185	0,008923	0,000932	10.06738	0,009383
0,000880	10.15185	0,008932	0,000933	10.06583	0,009391
0,000881	10.14857	0,008941	0,000934	10.06429	0,009400
0,000882	10.14693	0,008950	0,000935	10.06274	0,009409
0,000883	10.14530	0,008958	0,000936	10.06120	0,009417
0,000884	10.14366	0,008967	0,000937	10.05966	0,009426
0,000885	10.14203	0,008976	0,000938	10.05812	0,009435
0,000886	10.14040	0,008984	0,000939	10.05658	0,009443
0,000887	10.13878	0,008993	0,000940	10.05505	0,009452
0,000888	10.13715	0,009002	0,000941	10.05351	0,009460
0,000889	10.13553	0,009010	0,000942	10.05198	0,009469
0,000890	10.13390	0,009019	0,000943	10.05045	0,009478
0,000891	10.13228	0,009028	0,000944	10.04892	0,009486
0,000892	10.13067	0,009037	0,000945	10.04740	0,009495
0,000893	10.12905	0,009045	0,000946	10.04587	0,009503
0,000894	10.12743	0,009054	0,000947	10.04435	0,009512
0,000895	10.12582	0,009063	0,000948	10.04282	0,009521

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	p	$-\log_2 p$	$-p\log_2 p$
0.000949	10.04130	0.009529	0.000975	10.00231	0.009752
0.000950	10.03978	0.009538	0.000976	10.00083	0.009761
0.000951	10.03826	0.009546	0.000977	9.999355	0.009769
0.000952	10.03675	0.009555	0.000978	9.997879	0.009778
0.000953	10.03523	0.009564	0.000979	9.996405	0.009786
0.000954	10.03372	0.009572	0.000980	9.994932	0.009795
0.000955	10.03221	0.009581	0.000981	9.993461	0.009804
0.000956	10.03070	0.009589	0.000982	9.991991	0.009812
0.000957	10.02919	0.009598	0.000983	9.990522	0.009821
0.000958	10.02768	0.009607	0.000984	9.989056	0.009829
0.000959	10.02618	0.009615	0.000985	9.987590	0.009838
0.000960	10.02468	0.009624	0.000986	9.986126	0.009846
0.000961	10.02317	0.009632	0.000987	9.984664	0.009855
0.000962	10.02167	0.009641	0.000988	9.983203	0.009863
0.000963	10.02017	0.009649	0.000989	9.981743	0.009872
0.000964	10.01868	0.009658	0.000990	9.980285	0.009880
0.000965	10.01718	0.009667	0.000991	9.978829	0.009889
0.000966	10.01569	0.009675	0.000992	9.977374	0.009898
0.000967	10.01419	0.009684	0.000993	9.975920	0.009906
0.000968	10.01270	0.009692	0.000994	9.974468	0.009915
0.000969	10.01121	0.009701	0.000995	9.973017	0.009923
0.000970	10.00972	0.009709	0.000996	9.971568	0.009932
0.000971	10.00824	0.009718	0.000997	9.970120	0.009940
0.000972	10.00675	0.009727	0.000998	9.968674	0.009949
0.000973	10.00527	0.009735	0.000999	9.967229	0.009957
0.000974	10.00379	0.009744	0.001000	9.965786	0.009966

Таблица значений $-p\log_2 p$, $-q\log_2 q$ $(q=1-p)$ и H при $0.0011 \leqslant p \leqslant 0.5000$

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.0011	9.828281	0.010811	0.012397	0.001586	0.001588	0.9989
0.0012	9.702750	0.011643	0.013373	0.001730	0.001732	0.9988
0.0013	9.587273	0.012463	0.014338	0.001874	0.001877	0.9987
0.0014	9.480357	0.013273	0.015291	0.002018	0.002021	0.9986
0.0015	9.380822	0.014071	0.016234	0.002162	0.002166	0.9985
0.0016	9.287712	0.014860	0.017167	0.002306	0.002310	0.9984
0.0017	9.200249	0.015640	0.018091	0.002450	0.002455	0.9983
0.0018	9.117787	0.016412	0.019007	0.002595	0.002599	0.9982
0.0019	9.039785	0.017176	0.019914	0.002739	0.002744	0.9981
0.0020	8.965784	0.017932	0.020814	0.002882	0.002888	0.9980
0.0021	8.895395	0.018680	0.021707	0.003026	0.003033	0.9979
0.0022	8.828281	0.019422	0.022593	0.003170	0.003177	0.9978
0.0023	8.764150	0.020158	0.023472	0.003314	0.003322	0.9977
0.0024	8.702750	0.020887	0.024345	0.003458	0.003467	0.9976
0.0025	8.643856	0.021610	0.025212	0.003602	0.003611	0.9975
0.0026	8.587273	0.022327	0.026073	0.003746	0.003756	0.9974
0.0027	8.532825	0.023039	0.026929	0.003890	0.003901	0.9973

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.0028	8.480357	0.023745	0.027779	0.004034	0.004045	0.9972
0.0029	8.429731	0.024446	0.028624	0.004178	0.004190	0.9971
0.0030	8.380822	0.025142	0.029464	0.004322	0.004335	0.9970
0.0031	8.333516	0.025834	0.030299	0.004465	0.004479	0.9969
0.0032	8.287712	0.026521	0.031130	0.004609	0.004624	0.9968
0.0033	8.243318	0.027203	0.031956	0.004753	0.004769	0.9967
0.0034	8.200249	0.027881	0.032778	0.004897	0.004914	0.9966
0.0035	8.158429	0.028555	0.033595	0.005041	0.005058	0.9965
0.0036	8.117787	0.029224	0.034408	0.005184	0.005203	0.9964
0.0037	8.078259	0.029890	0.035218	0.005328	0.005348	0.9963
0.0038	8.039785	0.030551	0.036023	0.005472	0.005493	0.9962
0.0039	8.002310	0.031209	0.036825	0.005616	0.005638	0.9961
0.0040	7.965784	0.031863	0.037622	0.005759	0.005782	0.9960
0.0041	7.930160	0.032514	0.038417	0.005903	0.005927	0.9959
0.0042	7.895395	0.033161	0.039207	0.006047	0.006072	0.9958
0.0043	7.861448	0.033804	0.039994	0.006190	0.006217	0.9957
0.0044	7.828281	0.034444	0.040778	0.006334	0.006362	0.9956
0.0045	7.795859	0.035081	0.041559	0.006477	0.006507	0.9955
0.0046	7.764150	0.035715	0.042336	0.006621	0.006652	0.9954
0.0047	7.733123	0.036346	0.043110	0.006765	0.006797	0.9953
0.0048	7.702750	0.036973	0.043881	0.006908	0.006942	0.9952
0.0049	7.673002	0.037598	0.044650	0.007052	0.007087	0.9951
0.0050	7.643856	0.038219	0.045415	0.007195	0.007232	0.9950
0.0051	7.615287	0.038838	0.046177	0.007339	0.007377	0.9949
0.0052	7.587273	0.039454	0.046936	0.007482	0.007522	0.9948
0.0053	7.559792	0.040067	0.047693	0.007626	0.007667	0.9947
0.0054	7.532825	0.040677	0.048447	0.007769	0.007812	0.9946
0.0055	7.506353	0.041285	0.049198	0.007913	0.007957	0.9945
0.0056	7.480357	0.041890	0.049946	0.008056	0.008102	0.9944
0.0057	7.454822	0.042492	0.050692	0.008200	0.008247	0.9943
0.0058	7.429731	0.043092	0.051436	0.008343	0.008392	0.9942
0.0059	7.405069	0.043690	0.052177	0.008487	0.008537	0.9941
0.0060	7.380822	0.044285	0.052915	0.008630	0.008682	0.9940
0.0061	7.356975	0.044878	0.053651	0.008774	0.008827	0.9939
0.0062	7.333516	0.045468	0.054385	0.008917	0.008973	0.9938
0.0063	7.310432	0.046056	0.055116	0.009060	0.009118	0.9937
0.0064	7.287712	0.046641	0.055845	0.009204	0.009263	0.9936
0.0065	7.265345	0.047225	0.056572	0.009347	0.009408	0.9935
0.0066	7.243318	0.047806	0.057296	0.009490	0.009553	0.9934
0.0067	7.221623	0.048385	0.058018	0.009634	0.009699	0.9933
0.0068	7.200249	0.048962	0.058739	0.009777	0.009844	0.9932
0.0069	7.179188	0.049536	0.059457	0.009920	0.009989	0.9931
0.0070	7.158429	0.050109	0.060172	0.010063	0.010134	0.9930
0.0071	7.137965	0.050680	0.060886	0.010207	0.010280	0.9929
0.0072	7.117787	0.051248	0.061598	0.010350	0.010425	0.9928
0.0073	7.097888	0.051815	0.062308	0.010493	0.010570	0.9927
0.0074	7.078259	0.052379	0.063015	0.010636	0.010716	0.9926
0.0075	7.058894	0.052942	0.063721	0.010780	0.010861	0.9925
0.0076	7.039785	0.053502	0.064425	0.010923	0.011006	0.9924
0.0077	7.020926	0.054061	0.065127	0.011066	0.011152	0.9923
0.0078	7.002310	0.054618	0.065827	0.011209	0.011297	0.9922
0.0079	6.983932	0.055173	0.066525	0.011352	0.011443	0.9921

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0,0080	6,965784	0,055726	0,067222	0,011495	0,011588	0,9920
0,0081	6,947862	0,056278	0,067916	0,011638	0,011733	0,9919
0,0082	6,930160	0,056827	0,068609	0,011781	0,011879	0,9918
0,0083	6,912673	0,057375	0,069300	0,011925	0,012024	0,9917
0,0084	6,895395	0,057921	0,069989	0,012068	0,012170	0,9916
0,0085	6,878321	0,058466	0,070676	0,012211	0,012315	0,9915
0,0086	6,861448	0,059008	0,071362	0,012354	0,012461	0,9914
0,0087	6,844769	0,059549	0,072046	0,012497	0,012606	0,9913
0,0088	6,828281	0,060089	0,072729	0,012640	0,012752	0,9912
0,0089	6,811979	0,060627	0,073409	0,012783	0,012897	0,9911
0,0090	6,795859	0,061163	0,074088	0,012926	0,013043	0,9910
0,0091	6,779918	0,061697	0,074766	0,013069	0,013189	0,9909
0,0092	6,764150	0,062230	0,075442	0,013212	0,013334	0,9908
0,0093	6,748554	0,062762	0,076116	0,013354	0,013480	0,9907
0,0094	6,733123	0,063291	0,076789	0,013497	0,013625	0,9906
0,0095	6,717857	0,063820	0,077460	0,013640	0,013771	0,9905
0,0096	6,702750	0,064346	0,078130	0,013783	0,013917	0,9904
0,0097	6,687800	0,064872	0,078798	0,013926	0,014062	0,9903
0,0098	6,673002	0,065395	0,079464	0,014069	0,014208	0,9902
0,0099	6,658356	0,065918	0,080129	0,014212	0,014354	0,9901
0,0100	6,643856	0,066439	0,080793	0,014355	0,014500	0,9900
0,0102	6,615287	0,067476	0,082116	0,014640	0,014791	0,9898
0,0104	6,587273	0,068508	0,083433	0,014926	0,015083	0,9896
0,0106	6,559792	0,069534	0,084745	0,015211	0,015374	0,9894
0,0108	6,532825	0,070555	0,086051	0,015497	0,015666	0,9892
0,0110	6,506353	0,071570	0,087352	0,015782	0,015958	0,9890
0,0112	6,480357	0,072580	0,088647	0,016067	0,016249	09888
0,0114	6,454822	0,073585	0,089938	0,016353	0,016541	0,9886
0,0116	6,429731	0,074585	0,091223	0,016638	0,016833	0,9884
0,0118	6,405069	0,075580	0,092503	0,016923	0,017125	0,9882
0,0120	6,380822	0,076570	0,093778	0,017208	0,017417	0,9880
0,0122	6,356975	0,077555	0,095048	0,017493	0,017709	0,9878
0,0124	6,333516	0,078536	0,096314	0,017778	0,018001	0,9876
0,0126	6,310432	0,079511	0,097574	0,018063	0,018293	0,9874
0,0128	6,287712	0,080483	0,098831	0,018348	0,018586	0,9872
0,0130	6,265345	0,081449	0,100082	0,018633	0,018878	0,9870
0,0132	6,243318	0,082412	0,101329	0,018917	0,019170	0,9868
0,0134	6,221623	0,083370	0,102572	0,019202	0,019463	0,9866
0,0136	6,200249	0,084323	0,103810	0,019487	0,019755	0,9864
0,0138	6,179188	0,085273	0,105044	0,019771	0,020048	0,9862
0,0140	6,158429	0,086218	0,106274	0,020056	0,020340	0,9860
0,0142	6,137965	0,087159	0,107499	0,020340	0,020633	0,9858
0,0144	6,117787	0,088096	0,108721	0,020625	0,020926	0,9856
0,0146	6,097888	0,089029	0,109938	0,020909	0,021219	0,9854
0,0148	6,078259	0,089958	0,111151	0,021193	0,021511	0,9852
0,0150	6,058894	0,090883	0,112361	0,021477	0,021804	0,9850
0,0152	6,039785	0,091805	0,113566	0,021761	0,022097	0,9848
0,0154	6,020926	0,092722	0,114768	0,022046	0,022390	0,9846
0,0156	6,002310	0,093636	0,115966	0,022330	0,022683	0,9844
0,0158	5,983932	0,094546	0,117160	0,022614	0,022977	0,9842
0,0160	5,965784	0,095453	0,118350	0,022897	0,023270	0,9840
0,0162	5,947862	0,096355	0,119537	0,023181	0,023563	0,9838

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0,0164	5,930160	0,097255	0,120720	0,023465	0,023856	0,9836
0,0166	5,912673	0,098150	0,121899	0,023749	0,024150	0,9834
0,0168	5,895395	0,099043	0,123075	0,024033	0,024443	0,9832
0,0170	5,878321	0,099931	0,124248	0,024316	0,024737	0,9830
0,0172	5,861448	0,100817	0,125417	0,024600	0,025030	0,9828
0,0174	5,844769	0,101690	0,126582	0,024883	0,025324	0,9826
0,0176	5,828281	0,102578	0,127744	0,025167	0,025618	0,9824
0,0178	5,811979	0,103453	0,128903	0,025450	0,025911	0,9822
0,0180	5,795859	0,104325	0,130059	0,025733	0,026205	0,9820
0,0182	5,779918	0,105195	0,131211	0,026017	0,026499	0,9818
0,0184	5,764150	0,106060	0,132360	0,026300	0,026793	0,9816
0,0186	5,748554	0,106923	0,133506	0,026583	0,027087	0,9814
0,0188	5,733123	0,107783	0,134649	0,026866	0,027381	0,9812
0,0190	5,717857	0,108639	0,135788	0,027149	0,027675	0,9810
0,0192	5,702750	0,109493	0,136925	0,027432	0,027969	0,9808
0,0194	5,687800	0,110343	0,138058	0,027715	0,028263	0,9806
0,0196	5,673002	0,011191	0,139189	0,027998	0,028558	0,9804
0,0198	5,658356	0,112035	0,140316	0,028281	0,028852	0,9802
0,0200	5,643856	0,112877	0,141441	0,028563	0,029146	0,9800
0,0202	5,629501	0,113716	0,142562	0,028846	0,029441	0,9798
0,0204	5,615287	0,114552	0,143681	0,029129	0,029735	0,9796
0,0206	5,601212	0,115385	0,144796	0,029411	0,030030	0,9794
0,0208	5,587273	0,116215	0,145909	0,029694	0,030325	0,9792
0,0210	5,573467	0,117043	0,147019	0,029976	0,030619	0,9790
0,0212	5,559792	0,117868	0,148126	0,030259	0,030914	0,9788
0,0214	5,546245	0,118690	0,149231	0,030541	0,031209	0,9786
0,0216	5,532825	0,119509	0,150332	0,030823	0,031504	0,9784
0,0218	5,519258	0,120326	0,151431	0,031105	0,031799	0,9782
0,0220	5,506353	0,121140	0,152527	0,031388	0,032094	0,9780
0,0222	5,493297	0,121951	0,153621	0,031670	0,032389	0,9778
0,0224	5,480357	0,122760	0,154712	0,031952	0,032684	0,9776
0,0226	5,467533	0,123566	0,155800	0,032234	0,032979	0,9774
0,0228	5,454822	0,124370	0,156886	0,032516	0,033274	0,9772
0,0230	5,442222	0,125171	0,157969	0,032797	0,033570	0,9770
0,0232	5,429731	0,125970	0,159049	0,033079	0,033865	0,9768
0,0234	5,417348	0,126766	0,160127	0,033361	0,034160	0,9766
0,0236	5,405069	0,127560	0,161202	0,033643	0,034456	0,9764
0,0238	5,392895	0,128351	0,162275	0,033924	0,034751	0,9762
0,0240	5,380822	0,129140	0,163346	0,034206	0,035047	0,9760
0,0242	5,368849	0,129926	0,164413	0,034487	0,035343	0,9758
0,0244	5,356975	0,130710	0,165479	0,034769	0,035638	0,9756
0,0246	5,345198	0,131492	0,166542	0,035050	0,035934	0,9754
0,0248	5,333516	0,132271	0,167603	0,035331	0,036230	0,9752
0,0250	5,321928	0,133048	0,168661	0,035613	0,036526	0,9750
0,0252	5,310432	0,133823	0,169717	0,035894	0,036822	0,9748
0,0254	5,299028	0,134595	0,170770	0,036175	0,037118	0,9746
0,0256	5,287712	0,135365	0,171822	0,036456	0,037414	0,9744
0,0258	5,276485	0,136133	0,172870	0,036737	0,037710	0,9742
0,0260	5,265345	0,136899	0,173917	0,037018	0,038006	0,9740
0,0262	5,254289	0,137662	0,174961	0,037299	0,038303	0,9738
0,0264	5,243318	0,138424	0,176004	0,037580	0,038599	0,9736
0,0266	5,232430	0,139183	0,177043	0,037861	0,038895	0,9734

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.0268	5.221623	0.139939	0.178081	0.038141	0.039192	0.9732
0.0270	5.210897	0.140694	0.179116	0.038422	0.039488	0.9730
0.0272	5.200249	0.141447	0.180149	0.038703	0.039785	0.9728
0.0274	5.189680	0.142197	0.181180	0.038983	0.040081	0.9726
0.0276	5.179188	0.142946	0.182209	0.039264	0.040378	0.9724
0.0278	5.168771	0.143692	0.183236	0.039544	0.040675	0.9722
0.0280	5.158429	0.144436	0.184261	0.039825	0.040972	0.9720
0.0282	5.148161	0.145178	0.185283	0.040105	0.041269	0.9718
0.0284	5.137965	0.145918	0.186303	0.040385	0.041566	0.9716
0.0286	5.127841	0.146656	0.187322	0.040665	0.041863	0.9714
0.0288	4.117787	0.147392	0.188338	0.040945	0.042160	0.9712
0.0290	5.107803	0.148126	0.189352	0.041226	0.042457	0.9710
0.0292	5.097888	0.148858	0.190364	0.041506	0.042754	0.9708
0.0294	5.088040	0.149588	0.191374	0.041786	0.043051	0.9706
0.0296	5.078259	0.150316	0.192382	0.042065	0.043349	0.9704
0.0298	5.068544	0.151043	0.193388	0.042345	0.043646	0.9702
0.0300	5.058894	0.151767	0.194392	0.042625	0.043943	0.9700
0.0302	5.049308	0.152489	0.195394	0.042905	0.044241	0.9698
0.0304	5.039785	0.153209	0.196394	0.043184	0.044538	0.9696
0.0306	5.030325	0.153928	0.197392	0.043464	0.044836	0.9694
0.0308	5.020926	0.154645	0.198388	0.043744	0.045134	0.9692
0.0310	5.011588	0.155359	0.199382	0.044023	0.045431	0.9690
0.0312	5.002310	0.156072	0.200375	0.044302	0.045729	0.9688
0.0314	4.993092	0.156783	0.201365	0.044582	0.046027	0.9686
0.0316	4.983932	0.157492	0.202353	0.044861	0.046325	0.9684
0.0318	4.974829	0.158200	0.203340	0.045140	0.046623	0.9682
0.0320	4.965784	0.158905	0.204325	0.045420	0.046921	0.9680
0.0322	4.956795	0.159609	0.205307	0.045699	0.047219	0.9678
0.0324	4.947862	0.160311	0.206288	0.045978	0.047517	0.9676
0.0326	4.938984	0.161011	0.207268	0.046257	0.047816	0.9674
0.0328	4.930160	0.161709	0.208245	0.046536	0.048114	0.9672
0.0330	4.921390	0.162406	0.209220	0.046815	0.048412	0.9670
0.0332	4.912673	0.163101	0.210194	0.047093	0.048711	0.9668
0.0334	4.904008	0.163794	0.211166	0.047372	0.049009	0.9666
0.0336	4.895395	0.164485	0.212136	0.047651	0.049308	0.9664
0.0338	4.886833	0.165175	0.213104	0.047930	0.049606	0.9662
0.0340	4.878321	0.165863	0.214071	0.048208	0.049905	0.9660
0.0342	4.869860	0.166549	0.215036	0.048487	0.050204	0.9658
0.0344	4.861448	0.167234	0.215999	0.048765	0.050502	0.9656
0.0346	4.853084	0.167917	0.216960	0.049044	0.050801	0.9654
0.0348	4.844769	0.168598	0.217920	0.049322	0.051100	0.9652
0.0350	4.836501	0.169278	0.218878	0.049600	0.051399	0.9650
0.0352	4.828281	0.169955	0.219834	0.049878	0.051698	0.9648
0.0354	4.820107	0.170632	0.220788	0.050157	0.051997	0.9646
0.0356	4.811979	0.171306	0.221741	0.050435	0.052296	0.9644
0.0358	4.803897	0.171979	0.222692	0.050713	0.052596	0.9642
0.0360	4.795859	0.172651	0.223642	0.050991	0.052895	0.9640
0.0362	4.787866	0.173321	0.224589	0.051269	0.053194	0.9638
0.0364	4.779918	0.173989	0.225536	0.051547	0.053494	0.9636
0.0366	4.772013	0.174656	0.226480	0.051824	0.053793	0.9634
0.0368	4.764150	0.175321	0.227423	0.052102	0.054093	0.9632
0.0370	4.756331	0.175984	0.228364	0.052380	0.054392	0.9630

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.0372	4.748554	0.176646	0.229304	0.052657	0.054692	0.9628
0.0374	4.740818	0.177307	0.230242	0.052935	0.054992	0.9626
0.0376	4.733123	0.177965	0.231178	0.053212	0.055291	0.9624
0.0378	4.725470	0.178623	0.232113	0.053490	0.055591	0.9622
0.0380	4.717857	0.179279	0.233046	0.053767	0.055891	0.9620
0.0382	4.710284	0.179933	0.233977	0.054045	0.056191	0.9618
0.0384	4.702750	0.180586	0.234908	0.054322	0.056491	0.9616
0.0386	4.695255	0.181237	0.235836	0.054599	0.056791	0.9614
0.0388	4.687800	0.181887	0.236763	0.054876	0.057091	0.9612
0.0390	4.680382	0.182535	0.237688	0.055153	0.057392	0.9610
0.0392	4.673002	0.183182	0.238612	0.055430	0.057692	0.9608
0.0394	4.665661	0.183827	0.239534	0.055707	0.057992	0.9606
0.0396	4.658356	0.184471	0.240455	0.055984	0.058293	0.9604
0.0398	4.651088	0.185113	0.241374	0.056261	0.058593	0.9602
0.0400	4.643856	0.185754	0.242292	0.056538	0.058894	0.9600
0.0402	4.636661	0.186394	0.243208	0.056815	0.059194	0.9598
0.0404	4.629501	0.187032	0.244123	0.057091	0.059495	0.9596
0.0406	4.622376	0.187668	0.245036	0.057368	0.059796	0.9594
0.0408	4.615287	0.188304	0.245948	0.057644	0.060096	0.9592
0.0410	4.608232	0.188938	0.246858	0.057921	0.060397	0.9590
0.0412	4.601212	0.189570	0.247767	0.058197	0.060698	0.9588
0.0414	4.594225	0.190201	0.248675	0.058474	0.060999	0.9586
0.0416	4.587273	0.190831	0.249581	0.058750	0.061300	0.9584
0.0418	4.580353	0.191459	0.250485	0.059026	0.061601	0.9582
0.0420	4.573467	0.192086	0.251388	0.059303	0.061902	0.9580
0.0422	4.566613	0.192711	0.252290	0.059579	0.062204	0.9578
0.0424	4.559792	0.193335	0.253190	0.059855	0.062505	0.9576
0.0426	4.553003	0.193958	0.254089	0.060131	0.062806	0.9574
0.0428	4.546245	0.194579	0.254986	0.060407	0.063108	0.9572
0.0430	4.539519	0.195199	0.255882	0.060683	0.063409	0.9570
0.0432	4.532825	0.195818	0.256776	0.060958	0.063711	0.9568
0.0434	4.526161	0.196435	0.257670	0.061234	0.064012	0.9566
0.0436	4.519528	0.197051	0.258561	0.061510	0.064314	0.9564
0.0438	4.512295	0.197666	0.259452	0.061786	0.064616	0.9562
0.0440	4.506353	0.198280	0.260341	0.062061	0.064917	0.9560
0.0442	4.499810	0.198892	0.261228	0.062337	0.065219	0.9558
0.0444	4.493297	0.199502	0.262114	0.062612	0.065521	0.9556
0.0446	4.486812	0.200112	0.262999	0.062887	0.065823	0.9554
0.0448	4.480357	0.200720	0.263883	0.063163	0.066125	0.9552
0.0450	4.473931	0.201327	0.264765	0.063438	0.066427	0.9550
0.0452	4.467533	0.201933	0.265646	0.063713	0.066730	0.9548
0.0454	4.461164	0.202537	0.266525	0.063989	0.067032	0.9546
0.0456	4.454822	0.203140	0.267404	0.064264	0.067334	0.9544
0.0458	4.448509	0.203742	0.268280	0.064539	0.067636	0.9542
0.0460	4.442222	0.204342	0.269156	0.064814	0.067939	0.9540
0.0462	4.435963	0.204942	0.270030	0.065089	0.068241	0.9538
0.0464	4.429731	0.205540	0.270903	0.065363	0.068544	0.9536
0.0466	4.423526	0.206136	0.271775	0.065638	0.068846	0.9534
0.0468	4.417348	0.206732	0.272645	0.065913	0.069149	0.9532
0.0470	4.411195	0.207326	0.273514	0.066188	0.069452	0.9530
0.0472	4.405069	0.207919	0.274382	0.066462	0.069755	0.9528
0.0474	4.398969	0.208511	0.275248	0.066737	0.070058	0.9526

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.0476	4.392895	0.209102	0.276113	0.067011	0.070360	0.9524
0.0478	4.386846	0.209691	0.276977	0.067286	0.070663	0.9522
0.0480	4.380822	0.210279	0.277840	0.067560	0.070967	0.9520
0.0482	4.374823	0.210866	0.278701	0.067834	0.071270	0.9518
0.0484	4.368849	0.211452	0.279561	0.068109	0.071573	0.9516
0.0486	4.362900	0.212037	0.280420	0.068383	0.071876	0.9514
0.0488	4.356975	0.212620	0.281277	0.068657	0.072179	0.9512
0.0490	4.351074	0.213203	0.282134	0.068931	0.072483	0.9510
0.0492	4.345198	0.213784	0.282989	0.069205	0.072786	0.9508
0.0494	4.339345	0.214364	0.283843	0.069479	0.073090	0.9506
0.0496	4.333516	0.214942	0.284695	0.069753	0.073393	0.9504
0.0498	4.327710	0.215520	0.285547	0.070027	0.073697	0.9502
0.0500	4.321928	0.216096	0.286397	0.070301	0.074001	0.9500
0.0502	4.316169	0.216672	0.287246	0.070574	0.074304	0.9498
0.0504	4.310432	0.217246	0.288094	0.070848	0.074608	0.9496
0.0506	4.304719	0.217819	0.288940	0.071121	0.074912	0.9494
0.0508	4.299028	0.218391	0.289786	0.071395	0.075216	0.9492
0.0510	4.293359	0.218961	0.290630	0.071668	0.075520	0.9490
0.0512	4.287712	0.219531	0.291473	0.071942	0.075824	0.9488
0.0514	4.282088	0.220099	0.292315	0.072215	0.076128	0.9486
0.0516	4.276485	0.220667	0.293155	0.072489	0.076432	0.9484
0.0518	4.270904	0.221233	0.293995	0.072762	0.076737	0.9482
0.0520	4.265345	0.221798	0.294833	0.073035	0.077041	0.9480
0.0522	4.259806	0.222362	0.295670	0.073308	0.077345	0.9478
0.0524	4.254289	0.222925	0.296506	0.073581	0.077650	0.9476
0.0526	4.248793	0.223487	0.297341	0.073854	0.077954	0.9474
0.0528	4.243318	0.224047	0.298174	0.074127	0.078259	0.9472
0.0530	4.237864	0.224607	0.299007	0.074400	0.078564	0.9470
0.0532	4.232430	0.225165	0.299838	0.074673	0.078868	0.9468
0.0534	4.227016	0.225723	0.300668	0.074945	0.079173	0.9466
0.0536	4.221623	0.226279	0.301497	0.075218	0.079478	0.9464
0.0538	4.216250	0.226834	0.302325	0.075491	0.079783	0.9462
0.0540	4.210897	0.227388	0.303152	0.075763	0.080088	0.9460
0.0542	4.205563	0.227942	0.303977	0.076036	0.080393	0.9458
0.0544	4.200249	0.228494	0.304802	0.076308	0.080698	0.9456
0.0546	4.194955	0.229045	0.305625	0.076580	0.081003	0.9454
0.0548	4.189680	0.229594	0.306447	0.076853	0.081308	0.9452
0.0550	4.184425	0.230143	0.307268	0.077125	0.081614	0.9450
0.0552	4.179188	0.230691	0.308088	0.077397	0.081919	0.9448
0.0554	4.173970	0.231238	0.308907	0.077669	0.082225	0.9446
0.0556	4.168771	0.231784	0.309725	0.077941	0.082530	0.9444
0.0558	4.163591	0.232328	0.310542	0.078213	0.082836	0.9442
0.0560	4.158429	0.232872	0.311357	0.078485	0.083141	0.9440
0.0562	4.153286	0.233415	0.312172	0.078757	0.083447	0.9438
0.0564	4.148161	0.233956	0.312985	0.079029	0.083753	0.9436
0.0566	4.143054	0.234497	0.313798	0.079301	0.084058	0.9434
0.0568	4.137965	0.235036	0.314609	0.079572	0.084364	0.9432
0.0570	4.132894	0.235575	0.315419	0.079844	0.084670	0.9430
0.0572	4.127841	0.236113	0.316228	0.080116	0.084976	0.9428
0.0574	4.122805	0.236649	0.317036	0.080387	0.085282	0.9426
0.0576	4.117787	0.237185	0.317843	0.080659	0.085589	0.9424
0.0578	4.112787	0.237719	0.318649	0.080930	0.085895	0.9422

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.0580	4.107803	0.238253	0.319454	0.081201	0.086201	0.9420
0.0582	4.102837	0.238785	0.320258	0.081473	0.086507	0.9418
0.0584	4.097888	0.239317	0.321060	0.081744	0.086814	0.9416
0.0586	4.092956	0.239847	0.321862	0.082015	0.087120	0.9414
0.0588	4.088040	0.240377	0.322663	0.082286	0.087427	0.9412
0.0590	4.083141	0.240905	0.323462	0.082557	0.087733	0.9410
0.0592	4.078259	0.241433	0.324261	0.082828	0.088040	0.9408
0.0594	4.073393	0.241960	0.325059	0.083099	0.088347	0.9406
0.0596	4.068544	0.242485	0.325855	0.083370	0.088654	0.9404
0.0598	4.063711	0.243010	0.326650	0.083641	0.088960	0.9402
0.0600	4.058894	0.243534	0.327445	0.083911	0.089267	0.9400
0.0605	4.046921	0.244839	0.329427	0.084588	0.090035	0.9395
0.0610	4.035047	0.246138	0.331402	0.085264	0.090803	0.9390
0.0615	4.023270	0.247431	0.333371	0.085940	0.091571	0.9385
0.0620	4.011588	0.248718	0.335334	0.086615	0.092340	0.9380
0.0625	4.000000	0.250000	0.337290	0.087290	0.093109	0.9375
0.0630	3.988504	0.251276	0.339240	0.087965	0.093879	0.9370
0.0635	3.977100	0.252546	0.341185	0.088639	0.094649	0.9365
0.0640	3.965784	0.253810	0.343123	0.089313	0.095420	0.9360
0.0645	3.954557	0.255069	0.345055	0.089986	0.096190	0.9355
0.0650	3.943416	0.256322	0.346981	0.090659	0.096962	0.9350
0.0655	3.932361	0.257570	0.348902	0.091332	0.097733	0.9345
0.0660	3.921390	0.258812	0.350816	0.092004	0.098506	0.9340
0.0665	3.910502	0.260048	0.352724	0.092676	0.099278	0.9335
0.0670	3.899695	0.261280	0.354627	0.093348	0.100051	0.9330
0.0675	3.888969	0.262505	0.356524	0.094019	0.100824	0.9325
0.0680	3.878321	0.263726	0.358415	0.094689	0.101598	0.9320
0.0685	3.867752	0.264941	0.360301	0.095360	0.102372	0.9315
0.0690	3.857260	0.266151	0.362181	0.096030	0.103147	0.9310
0.0695	3.846843	0.267356	0.364055	0.096699	0.103922	0.9305
0.0700	3.836501	0.268555	0.365924	0.097369	0.104697	0.9300
0.0705	3.826233	0.269749	0.367787	0.098037	0.105473	0.9295
0.0710	3.816037	0.270939	0.369644	0.098706	0.106249	0.9290
0.0715	3.805913	0.272123	0.371497	0.099374	0.107026	0.9285
0.0720	3.795859	0.273302	0.373343	0.100041	0.107803	0.9280
0.0725	3.785875	0.274476	0.375185	0.100709	0.108581	0.9275
0.0730	3.775960	0.275645	0.377021	0.101376	0.109359	0.9270
0.0735	3.766112	0.276809	0.378851	0.102042	0.110137	0.9265
0.0740	3.756331	0.277968	0.380677	0.102708	0.110916	0.9260
0.0745	3.746616	0.279123	0.382497	0.103374	0.111695	0.9255
0.0750	3.736966	0.280272	0.384312	0.104039	0.112475	0.9250
0.0755	3.727380	0.281417	0.386121	0.104704	0.113255	0.9245
0.0760	3.717857	0.282557	0.387926	0.105369	0.114035	0.9240
0.0765	3.708396	0.283692	0.389725	0.106033	0.114816	0.9235
0.0770	3.698998	0.284823	0.391519	0.106696	0.115597	0.9230
0.0775	3.689660	0.285949	0.393308	0.107360	0.116379	0.9225
0.0780	3.680382	0.287070	0.395093	0.108023	0.117161	0.9220
0.0785	3.671164	0.288186	0.396872	0.108685	0.117944	0.9215
0.0790	3.662004	0.289298	0.398646	0.109348	0.118727	0.9210
0.0795	3.652901	0.290406	0.400415	0.110009	0.119510	0.9205
0.0800	3.643856	0.291508	0.402179	0.110671	0.120294	0.9200
0.0805	3.634867	0.292607	0.403939	0.111332	0.121079	0.9195

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.0810	3.625934	0.293701	0.405693	0.111992	0.121863	0.9190
0.0815	3.617056	0.294790	0.407443	0.112653	0.122648	0.9185
0.0820	3.608232	0.295875	0.409187	0.113312	0.123434	0.9180
0.0825	3.599462	0.296956	0.410927	0.113972	0.124220	0.9175
0.0830	3.590745	0.298032	0.412663	0.114631	0.125006	0.9170
0.0835	3.582080	0.299104	0.414393	0.115289	0.125793	0.9165
0.0840	3.573467	0.300171	0.416119	0.115948	0.126580	0.9160
0.0845	3.564905	0.301234	0.417840	0.116606	0.127368	0.9155
0.0850	3.556393	0.302293	0.419556	0.117263	0.128156	0.9150
0.0855	3.547932	0.303348	0.421268	0.117920	0.128945	0.9145
0.0860	3.539520	0.304399	0.422975	0.118577	0.129734	0.9140
0.0865	3.531156	0.305445	0.424678	0.119233	0.130523	0.9135
0.0870	3.522841	0.306487	0.426376	0.119889	0.131313	0.9130
0.0875	3.514573	0.307525	0.428070	0.120544	0.132104	0.9125
0.0880	3.506353	0.308559	0.429759	0.121200	0.132894	0.9120
0.0885	3.498179	0.309589	0.431443	0.121854	0.133685	0.9115
0.0890	3.490051	0.310615	0.433123	0.122509	0.134477	0.9110
0.0895	3.481968	0.311636	0.434799	0.123162	0.135269	0.9105
0.0900	3.473931	0.312654	0.436470	0.123816	0.136062	0.9100
0.0905	3.465938	0.313667	0.438137	0.124469	0.136854	0.9095
0.0910	3.457990	0.314677	0.439799	0.125122	0.137648	0.9090
0.0915	3.450084	0.315683	0.441457	0.125774	0.138442	0.9085
0.0920	3.442222	0.316684	0.443111	0.126426	0.139236	0.9080
0.0925	3.434403	0.317682	0.444760	0.127078	0.140030	0.9075
0.0930	3.426625	0.318676	0.446405	0.127729	0.140826	0.9070
0.0935	3.418890	0.319666	0.448046	0.128379	0.141621	0.9065
0.0940	3.411195	0.320652	0.449682	0.129030	0.142417	0.9060
0.0945	3.403542	0.321635	0.451314	0.129680	0.143213	0.9055
0.0950	3.395929	0.322613	0.452943	0.130329	0.144010	0.9050
0.0955	3.388355	0.323588	0.454566	0.130978	0.144808	0.9045
0.0960	3.380822	0.324559	0.456186	0.131627	0.145605	0.9040
0.0965	3.373327	0.325526	0.457802	0.132276	0.146403	0.9035
0.0970	3.365871	0.326490	0.459413	0.132923	0.147202	0.9030
0.0975	3.358454	0.327449	0.461020	0.133571	0.148001	0.9025
0.0980	3.351074	0.328405	0.462623	0.134218	0.148801	0.9020
0.0985	3.343732	0.329358	0.464223	0.134865	0.149601	0.9015
0.0990	3.336428	0.330306	0.465818	0.135511	0.150401	0.9010
0.0995	3.329160	0.331251	0.467409	0.136157	0.151202	0.9005
0.1000	3.321928	0.332193	0.468996	0.136803	0.152003	0.9000
0.1005	3.314733	0.333131	0.470579	0.137448	0.152805	0.8995
0.1010	3.307573	0.334065	0.472158	0.138093	0.153607	0.8990
0.1015	3.300448	0.334996	0.473733	0.138737	0.154410	0.8985
0.1020	3.293359	0.335923	0.475304	0.139381	0.155213	0.8980
0.1025	3.286304	0.336846	0.476871	0.140024	0.156016	0.8975
0.1030	3.279284	0.337766	0.478434	0.140668	0.156820	0.8970
0.1035	3.272297	0.338683	0.479993	0.141310	0.157624	0.8965
0.1040	3.265345	0.339596	0.481549	0.141953	0.158429	0.8960
0.1045	3.258425	0.340505	0.483100	0.142595	0.159235	0.8955
0.1050	3.251539	0.341412	0.484648	0.143236	0.160040	0.8950
0.1055	3.244685	0.342314	0.486192	0.143877	0.160847	0.8945
0.1060	3.237864	0.343214	0.487732	0.144518	0.161653	0.8940
0.1065	3.231075	0.344109	0.489268	0.145158	0.162460	0.8935

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.1070	3.224317	0.345002	0.490800	0.145798	0.163268	0.8930
0.1075	3.217591	0.345891	0.492329	0.146438	0.164076	0.8925
0.1080	3.210897	0.346777	0.493854	0.147077	0.164884	0.8920
0.1085	3.204233	0.347659	0.495375	0.147716	0.165693	0.8915
0.1090	3.197600	0.348538	0.496892	0.148354	0.166503	0.8910
0.1095	3.190997	0.349414	0.498406	0.148992	0.167312	0.8905
0.1100	3.184425	0.350287	0.499916	0.149629	0.168123	0.8900
0.1105	3.177882	0.351156	0.501422	0.150266	0.168933	0.8895
0.1110	3.171368	0.352022	0.502925	0.150903	0.169745	0.8890
0.1115	3.164884	0.352885	0.504424	0.151539	0.170556	0.8885
0.1120	3.158429	0.353744	0.505919	0.152175	0.171368	0.8880
0.1125	3.152003	0.354600	0.507411	0.152811	0.172181	0.8875
0.1130	3.145605	0.355453	0.508899	0.153446	0.172994	0.8870
0.1135	3.139236	0.356303	0.510384	0.154080	0.173807	0.8865
0.1140	3.132894	0.357150	0.511864	0.154715	0.174621	0.8860
0.1145	3.126580	0.357993	0.513342	0.155348	0.175436	0.8855
0.1150	3.120294	0.358834	0.514816	0.155982	0.176251	0.8850
0.1155	3.114035	0.359671	0.516285	0.156615	0.177066	0.8845
0.1160	3.107803	0.360505	0.517753	0.157247	0.177882	0.8840
0.1165	3.101598	0.361336	0.519216	0.157880	0.178698	0.8835
0.1170	3.095420	0.362164	0.520676	0.158511	0.179515	0.8830
0.1175	3.089267	0.362989	0.522132	0.159143	0.180332	0.8825
0.1180	3.083141	0.363811	0.523584	0.159774	0.181149	0.8820
0.1185	3.077041	0.364629	0.525034	0.160404	0.181968	0.8815
0.1190	3.070967	0.365445	0.526480	0.161035	0.182786	0.8810
0.1195	3.064917	0.366258	0.527922	0.161664	0.183605	0.8805
0.1200	3.058894	0.367067	0.529361	0.162294	0.184425	0.8800
0.1205	3.052895	0.367874	0.530796	0.162923	0.185245	0.8795
0.1210	3.046921	0.368677	0.532228	0.163551	0.186065	0.8790
0.1215	3.040972	0.369478	0.533657	0.164179	0.186886	0.8785
0.1220	3.035047	0.370276	0.535083	0.164807	0.187707	0.8780
0.1225	3.029146	0.371070	0.536505	0.165434	0.188529	0.8775
0.1230	3.023270	0.371862	0.537293	0.166061	0.189351	0.8770
0.1235	3.017417	0.372651	0.539339	0.166688	0.190174	0.8765
0.1240	3.011588	0.373437	0.540750	0.167314	0.190997	0.8760
0.1245	3.005782	0.374220	0.542159	0.167939	0.191821	0.8755
0.1250	3.000000	0.375000	0.543564	0.168564	0.192645	0.8750
0.1255	2.994241	0.375777	0.544966	0.169189	0.193470	0.8745
0.1260	2.988504	0.376552	0.546365	0.169814	0.194295	0.8740
0.1265	2.982791	0.377323	0.547761	0.170438	0.195120	0.8735
0.1270	2.977100	0.378092	0.549153	0.171061	0.195946	0.8730
0.1275	2.971431	0.378857	0.550542	0.171684	0.196773	0.8725
0.1280	2.965784	0.379620	0.551928	0.172307	0.197600	0.8720
0.1285	2.960160	0.380381	0.553310	0.172929	0.198427	0.8715
0.1290	2.954557	0.381138	0.554689	0.173551	0.199255	0.8710
0.1295	2.948976	0.381892	0.556065	0.174173	0.200084	0.8705
0.1300	2.943416	0.382644	0.557438	0.174794	0.200913	0.8700
0.1305	2.937878	0.383393	0.558808	0.175415	0.201742	0.8695
0.1310	2.932361	0.384139	0.560174	0.176035	0.202572	0.8690
0.1315	2.926865	0.384883	0.561538	0.176655	0.203402	0.8685
0.1320	2.921390	0.385623	0.562898	0.177274	0.204233	0.8680
0.1325	2.915936	0.386361	0.564255	0.177893	0.205064	0.8675

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.1330	2.910502	0.387097	0.565609	0.178512	0.205896	0.8670
0.1335	2.905088	0.387829	0.566959	0.179130	0.206728	0.8665
0.1340	2.899695	0.388559	0.568307	0.179748	0.207561	0.8660
0.1345	2.894322	0.389286	0.569652	0.180365	0.208394	0.8655
0.1350	2.888969	0.390011	0.570993	0.180982	0.209228	0.8650
0.1355	2.883635	0.390733	0.572331	0.181599	0.210062	0.8645
0.1360	2.878321	0.391452	0.573667	0.182215	0.210897	0.8640
0.1365	2.873027	0.392168	0.574999	0.182830	0.211732	0.8635
0.1370	2.867752	0.392882	0.576328	0.183446	0.212568	0.8630
0.1375	2.862496	0.393593	0.577654	0.184061	0.213404	0.8625
0.1380	2.857260	0.394302	0.578977	0.184675	0.214240	0.8620
0.1385	2.852042	0.395008	0.580297	0.185289	0.215077	0.8615
0.1390	2.846843	0.395711	0.581614	0.185903	0.215915	0.8610
0.1395	2.841663	0.396412	0.582928	0.186516	0.216753	0.8605
0.1400	2.836501	0.397110	0.584239	0.187129	0.217591	0.8600
0.1405	2.831358	0.397806	0.585547	0.187741	0.218430	0.8595
0.1410	2.826233	0.398499	0.586852	0.188353	0.219270	0.8590
0.1415	2.821126	0.399189	0.588154	0.188964	0.220110	0.8585
0.1420	2.816037	0.399877	0.589453	0.189575	0.220950	0.8580
0.1425	2.810966	0.400563	0.590749	0.190186	0.221791	0.8575
0.1430	2.805913	0.401246	0.592042	0.190796	0.222633	0.8570
0.1435	2.800877	0.401926	0.593332	0.191406	0.223475	0.8565
0.1440	2.795859	0.402604	0.594619	0.192016	0.224317	0.8560
0.1445	2.790859	0.403279	0.595904	0.192625	0.225160	0.8555
0.1450	2.785875	0.403952	0.597185	0.193233	0.226004	0.8550
0.1455	2.780909	0.404622	0.598464	0.193841	0.226848	0.8545
0.1460	2.775960	0.405290	0.599739	0.194449	0.227692	0.8540
0.1465	2.771027	0.405956	0.601012	0.195056	0.228537	0.8535
0.1470	2.766112	0.406618	0.602282	0.195663	0.229382	0.8530
0.1475	2.761213	0.407279	0.603549	0.196270	0.230228	0.8525
0.1480	2.756331	0.407937	0.604813	0.196876	0.231075	0.8520
0.1485	2.751465	0.408593	0.606074	0.197481	0.231922	0.8515
0.1490	2.746616	0.409246	0.607332	0.198086	0.232769	0.8510
0.1495	2.741783	0.409896	0.608588	0.198691	0.233617	0.8505
0.1500	2.736966	0.410545	0.609840	0.199295	0.234465	0.8500
0.1505	2.732165	0.411191	0.611090	0.199899	0.235314	0.8495
0.1510	2.727380	0.411834	0.612337	0.200503	0.236164	0.8490
0.1515	2.722610	0.412475	0.613581	0.201106	0.237013	0.8485
0.1520	2.717857	0.413114	0.614823	0.201709	0.237864	0.8480
0.1525	2.713119	0.413751	0.616061	0.202311	0.238715	0.8475
0.1530	2.708396	0.414385	0.617297	0.202913	0.239566	0.8470
0.1535	2.703689	0.415016	0.618530	0.203514	0.240418	0.8465
0.1540	2.698998	0.415646	0.619760	0.204115	0.241270	0.8460
0.1545	2.694321	0.416273	0.620988	0.204715	0.242123	0.8455
0.1550	2.689660	0.416897	0.622213	0.205315	0.242977	0.8450
0.1555	2.685014	0.417520	0.623435	0.205915	0.243831	0.8445
0.1560	2.680382	0.418140	0.624654	0.206514	0.244685	0.8440
0.1565	2.675765	0.418757	0.625870	0.207113	0.245540	0.8435
0.1570	2.671164	0.419373	0.627084	0.207711	0.246395	0.8430
0.1575	2.666576	0.419986	0.628295	0.208309	0.247251	0.8425
0.1580	2.662004	0.420597	0.629503	0.208907	0.248108	0.8420
0.1585	2.657445	0.421205	0.630709	0.209504	0.248965	0.8415

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.1590	2.652901	0.421811	0.631912	0.210101	0.249822	0.8410
0.1595	2.648372	0.422415	0.633112	0.210697	0.250680	0.8405
0.1600	2.643856	0.423017	0.634310	0.211293	0.251539	0.8400
0.1605	2.639355	0.423616	0.635504	0.211888	0.252398	0.8395
0.1610	2.634867	0.424214	0.636696	0.212483	0.253257	0.8390
0.1615	2.630394	0.424809	0.637886	0.213077	0.254117	0.8385
0.1620	2.625934	0.425401	0.639073	0.213671	0.254978	0.8380
0.1625	2.621488	0.425992	0.640257	0.214265	0.255839	0.8375
0.1630	2.617056	0.426580	0.641438	0.214858	0.256700	0.8370
0.1635	2.612637	0.427166	0.642617	0.215451	0.257563	0.8365
0.1640	2.608232	0.427750	0.643794	0.216043	0.258425	0.8360
0.1645	2.603841	0.428332	0.644967	0.216635	0.259288	0.8355
0.1650	2.599462	0.428911	0.646138	0.217227	0.260152	0.8350
0.1655	2.595097	0.429489	0.647306	0.217818	0.261016	0.8345
0.1660	2.590745	0.430064	0.648472	0.218409	0.261881	0.8340
0.1665	2.586406	0.430637	0.649635	0.218999	0.262746	0.8335
0.1670	2.582080	0.431207	0.650796	0.219588	0.263612	0.8330
0.1675	2.577767	0.431776	0.651954	0.220178	0.264478	0.8325
0.1680	2.573467	0.432342	0.653109	0.220767	0.265345	0.8320
0.1685	2.569180	0.432907	0.654262	0.221355	0.266212	0.8315
0.1690	2.564905	0.433469	0.655412	0.221943	0.267080	0.8310
0.1695	2.560643	0.434029	0.656560	0.222531	0.267948	0.8305
0.1700	2.556393	0.434587	0.657705	0.223118	0.268817	0.8300
0.1705	2.552156	0.435143	0.658847	0.223705	0.269686	0.8295
0.1710	2.547932	0.435696	0.659987	0.224291	0.270556	0.8290
0.1715	2.543720	0.436248	0.661125	0.224877	0.271426	0.8285
0.1720	2.539520	0.436797	0.662260	0.225462	0.272297	0.8280
0.1725	2.535332	0.437345	0.663392	0.226047	0.273169	0.8275
0.1730	2.531156	0.437890	0.664522	0.226632	0.274041	0.8270
0.1735	2.526992	0.438433	0.665649	0.227216	0.274913	0.8265
0.1740	2.522841	0.438974	0.666774	0.227799	0.275786	0.8260
0.1745	2.518701	0.439513	0.667896	0.228383	0.276660	0.8255
0.1750	2.514573	0.440050	0.669016	0.228966	0.277534	0.8250
0.1755	2.510457	0.440585	0.670133	0.229548	0.278409	0.8245
0.1760	2.506353	0.441118	0.671248	0.230130	0.279284	0.8240
0.1765	2.502260	0.441649	0.672360	0.230711	0.280159	0.8235
0.1770	2.498179	0.442178	0.673470	0.231292	0.281036	0.8230
0.1775	2.494109	0.442704	0.674577	0.231873	0.281912	0.8225
0.1780	2.490051	0.443229	0.675682	0.232453	0.282790	0.8220
0.1785	2.486004	0.443752	0.676785	0.233033	0.283668	0.8215
0.1790	2.481968	0.444272	0.677885	0.233612	0.284546	0.8210
0.1795	2.477944	0.444791	0.678982	0.234191	0.285425	0.8205
0.1800	2.473931	0.445308	0.680077	0.234769	0.286304	0.8200
0.1805	2.469929	0.445822	0.681170	0.235347	0.287184	0.8195
0.1810	2.465938	0.446335	0.682260	0.235925	0.288065	0.8190
0.1815	2.461959	0.446845	0.683347	0.236502	0.288946	0.8185
0.1820	2.457990	0.447354	0.684433	0.237079	0.289827	0.8180
0.1825	2.454032	0.447861	0.685516	0.237655	0.290709	0.8175
0.1830	2.450084	0.448365	0.686596	0.238231	0.291592	0.8170
0.1835	2.446148	0.448868	0.687674	0.238806	0.292475	0.8165
0.1840	2.442222	0.449369	0.688750	0.239381	0.293359	0.8160
0.1845	2.438307	0.449868	0.689823	0.239955	0.294243	0.8155

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.1850	2.434403	0.450365	0.690894	0.240529	0.295128	0.8150
0.1855	2.430509	0.450859	0.691962	0.241103	0.296013	0.8145
0.1860	2.426625	0.451352	0.693028	0.241676	0.296899	0.8140
0.1865	2.422752	0.451843	0.694092	0.242249	0.297786	0.8135
0.1870	2.418890	0.452332	0.695153	0.242821	0.298673	0.8130
0.1875	2.415037	0.452820	0.696212	0.243393	0.299560	0.8125
0.1880	2.411195	0.453305	0.697269	0.243964	0.300448	0.8120
0.1885	2.407364	0.453788	0.698323	0.244535	0.301337	0.8115
0.1890	2.403542	0.454269	0.699375	0.245105	0.302226	0.8110
0.1895	2.399730	0.454749	0.700424	0.245675	0.303116	0.8105
0.1900	2.395929	0.455226	0.701471	0.246245	0.304006	0.8100
0.1905	2.392137	0.455702	0.702516	0.246814	0.304897	0.8095
0.1910	2.388355	0.456176	0.703559	0.247383	0.305788	0.8090
0.1915	2.384584	0.456648	0.704599	0.247951	0.306680	0.8085
0.1920	2.380822	0.457118	0.705637	0.248519	0.307573	0.8080
0.1925	2.377070	0.457586	0.706672	0.249086	0.308466	0.8075
0.1930	2.373327	0.458052	0.707705	0.249653	0.309359	0.8070
0.1935	2.369595	0.458517	0.708736	0.250219	0.310254	0.8065
0.1940	2.365871	0.458979	0.709765	0.250785	0.311148	0.8060
0.1945	2.362158	0.459440	0.710791	0.251351	0.312044	0.8055
0.1950	2.358454	0.459899	0.711815	0.251916	0.312939	0.8050
0.1955	2.354759	0.460355	0.712836	0.252481	0.313836	0.8045
0.1960	2.351074	0.460811	0.713856	0.253045	0.314733	0.8040
0.1965	2.347399	0.461264	0.714873	0.253609	0.315630	0.8035
0.1970	2.343732	0.461715	0.715887	0.254172	0.316528	0.8030
0.1975	2.340075	0.462165	0.716900	0.254735	0.317427	0.8025
0.1980	2.336428	0.462613	0.717910	0.255297	0.318326	0.8020
0.1985	2.332789	0.463059	0.718918	0.255859	0.319226	0.8015
0.1990	2.329160	0.463503	0.719924	0.256421	0.320126	0.8010
0.1995	2.325539	0.463945	0.720927	0.256982	0.321027	0.8005
0.2000	2.321928	0.464386	0.721928	0.257542	0.321928	0.8000
0.2010	2.314733	0.465261	0.723294	0.258662	0.323733	0.7990
0.2020	2.307573	0.466130	0.725910	0.259780	0.325539	0.7980
0.2030	2.300448	0.466991	0.727888	0.260897	0.327348	0.9770
0.2040	2.293359	0.467845	0.729856	0.262011	0.329160	0.7960
0.2050	2.286304	0.468692	0.731816	0.263124	0.330973	0.7950
0.2060	2.279284	0.469532	0.733767	0.264235	0.332789	0.7940
0.2070	2.272297	0.470366	0.735709	0.265344	0.334607	0.7930
0.2080	2.265345	0.471192	0.737642	0.266451	0.336428	0.7920
0.2090	2.258425	0.472011	0.739567	0.267556	0.338250	0.7910
0.2100	2.251539	0.472823	0.741483	0.268660	0.340075	0.7900
0.2110	2.244685	0.473629	0.743390	0.269761	0.341903	0.7890
0.2120	2.237864	0.474427	0.745288	0.270861	0.343732	0.7880
0.2130	2.231075	0.475219	0.747178	0.271959	0.345564	0.7870
0.2140	2.224317	0.476004	0.749059	0.273055	0.347399	0.7860
0.2150	2.217591	0.476782	0.750932	0.274150	0.349235	0.7850
0.2160	2.210897	0.477554	0.752796	0.275242	0.351074	0.7840
0.2170	2.204233	0.478319	0.754652	0.276333	0.352916	0.7830
0.2180	2.197600	0.479077	0.756499	0.277422	0.354759	0.7820
0.2190	2.190997	0.479828	0.758337	0.278509	0.356606	0.7810
0.2200	2.184425	0.480573	0.760167	0.279594	0.358454	0.7800
0.2210	2.177882	0.481312	0.761989	0.280677	0.360305	0.7790

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.2220	2.171368	0.482044	0.763803	0.281759	0.362158	0.7780
0.2230	2.164884	0.482769	0.765608	0.282838	0.364013	0.7770
0.2240	2.158429	0.483488	0.767404	0.283916	0.365871	0.7760
0.2250	2.152003	0.484201	0.769193	0.284992	0.367732	0.7750
0.2260	2.145605	0.484907	0.770973	0.286066	0.369595	0.7740
0.2270	2.139236	0.485607	0.772745	0.287138	0.371460	0.7730
0.2280	2.132894	0.486300	0.774509	0.288209	0.373327	0.7720
0.2290	2.126580	0.486987	0.776264	0.289277	0.375197	0.7710
0.2300	2.120294	0.487668	0.778011	0.290344	0.377070	0.7700
0.2310	2.114035	0.488342	0.779750	0.291408	0.378944	0.7690
0.2320	2.107803	0.489010	0.781481	0.292471	0.380822	0.7680
0.2330	2.101598	0.489672	0.783204	0.293532	0.382702	0.7670
0.2340	2.095420	0.490328	0.784919	0.294591	0.384584	0.7660
0.2350	2.089267	0.490978	0.786626	0.295648	0.386468	0.7650
0.2360	2.083141	0.491621	0.788325	0.296704	0.388355	0.7640
0.2370	2.077041	0.492259	0.790016	0.297757	0.390245	0.7630
0.2380	2.070967	0.492890	0.791698	0.298808	0.392137	0.7620
0.2390	2.064917	0.493515	0.793373	0.299858	0.394032	0.7610
0.2400	2.058894	0.494134	0.795040	0.300906	0.395929	0.7600
0.2410	2.052895	0.494748	0.796699	0.301952	0.397828	0.7590
0.2420	2.046921	0.495355	0.798350	0.302996	0.399730	0.7580
0.2430	2.040972	0.495956	0.799994	0.304038	0.401635	0.7570
0.2440	2.035047	0.496551	0.801629	0.305078	0.403542	0.7560
0.2450	2.029146	0.497141	0.803257	0.306116	0.405451	0.7550
0.2460	2.023270	0.497724	0.804876	0.307152	0.407364	0.7540
0.2470	2.017417	0.498302	0.806488	0.308186	0.409278	0.7530
0.2480	2.011588	0.498874	0.808093	0.309219	0.411195	0.7520
0.2490	2.005782	0.499440	0.809689	0.310249	0.413115	0.7510
0.2500	2.000000	0.500000	0.811278	0.311278	0.415037	0.7500
0.2510	1.994241	0.500554	0.812859	0.312305	0.416962	0.7490
0.2520	1.988504	0.501103	0.814433	0.313330	0.418890	0.7480
0.2530	1.982791	0.501646	0.815998	0.314352	0.420820	0.7470
0.2540	1.977100	0.502183	0.817557	0.315373	0.422752	0.7460
0.2550	1.971431	0.502715	0.819107	0.316392	0.424688	0.7450
0.2560	1.965784	0.503241	0.820650	0.317409	0.426625	0.7440
0.2570	1.960160	0.503761	0.822185	0.318424	0.428566	0.7430
0.2580	1.954557	0.504276	0.823713	0.319438	0.430509	0.7420
0.2590	1.948976	0.504785	0.825234	0.320449	0.432455	0.7410
0.2600	1.943416	0.505288	0.826746	0.321458	0.434403	0.7400
0.2610	1.937878	0.505786	0.828252	0.322465	0.436354	0.7390
0.2620	1.932361	0.506279	0.829749	0.323471	0.438307	0.7380
0.2630	1.926865	0.506766	0.831240	0.324474	0.440263	0.7370
0.2640	1.921390	0.507247	0.832723	0.325476	0.442222	0.7360
0.2650	1.915936	0.507723	0.834198	0.326475	0.444184	0.7350
0.2660	1.910502	0.508193	0.835666	0.327473	0.446148	0.7340
0.2670	1.905088	0.508659	0.837127	0.328468	0.448115	0.7330
0.2680	1.899695	0.509118	0.838580	0.329462	0.450084	0.7320
0.2690	1.894322	0.509573	0.840026	0.330453	0.452057	0.7310
0.2700	1.888969	0.510022	0.841465	0.331443	0.454032	0.7300
0.2710	1.883635	0.510465	0.842896	0.332431	0.456009	0.7290
0.2720	1.878321	0.510903	0.844320	0.333416	0.457990	0.7280
0.2730	1.873027	0.511336	0.845737	0.334400	0.459973	0.7270

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.2740	1.867752	0.511764	0.847146	0.335382	0.461959	0.7260
0.2750	1.862496	0.512187	0.848548	0.336362	0.463947	0.7250
0.2760	1.857260	0.512604	0.849943	0.337339	0.465938	0.7240
0.2770	1.852042	0.513016	0.851331	0.338315	0.467932	0.7230
0.2780	1.846843	0.513422	0.852711	0.339289	0.469929	0.7220
0.2790	1.841663	0.513824	0.854085	0.340261	0.471929	0.7210
0.2800	1.836501	0.514220	0.855451	0.341230	0.473931	0.7200
0.2810	1.831358	0.514612	0.856810	0.342198	0.475936	0.7190
0.2820	1.826233	0.514998	0.858162	0.343164	0.477944	0.7180
0.2830	1.821126	0.515379	0.859506	0.344128	0.479955	0.7170
0.2840	1.816037	0.515755	0.860844	0.345089	0.481968	0.7160
0.2850	1.810966	0.516125	0.862175	0.346049	0.483985	0.7150
0.2860	1.805913	0.516491	0.863498	0.347007	0.486004	0.7140
0.2870	1.800877	0.516852	0.864814	0.347963	0.488026	0.7130
0.2880	1.795859	0.517207	0.866124	0.348916	0.490051	0.7120
0.2890	1.790859	0.517558	0.867426	0.349868	0.492079	0.7110
0.2900	1.785875	0.517904	0.868721	0.350817	0.494109	0.7100
0.2910	1.780909	0.518244	0.870009	0.351765	0.496142	0.7090
0.2920	1.775960	0.518580	0.871291	0.352711	0.498179	0.7080
0.2930	1.771027	0.518911	0.872565	0.353654	0.500218	0.7070
0.2940	1.766112	0.519237	0.873832	0.354595	0.502260	0.7060
0.2950	1.761213	0.519558	0.875093	0.355535	0.504305	0.7050
0.2960	1.756331	0.519874	0.876346	0.356472	0.506353	0.7040
0.2970	1.751465	0.520185	0.877593	0.357408	0.508403	0.7030
0.2980	1.746616	0.520491	0.878832	0.358341	0.510457	0.7020
0.2990	1.741783	0.520793	0.880065	0.359272	0.512514	0.7010
0.3000	1.736966	0.521090	0.881291	0.360201	0.514573	0.7000
0.3010	1.732165	0.521382	0.882510	0.361128	0.516636	0.6990
0.3020	1.727380	0.521669	0.883722	0.362053	0.518701	0.6980
0.3030	1.722610	0.521951	0.884927	0.362976	0.520769	0.6970
0.3040	1.717857	0.522228	0.886126	0.363897	0.522841	0.6960
0.3050	1.713119	0.522501	0.887317	0.364816	0.524915	0.6950
0.3060	1.708396	0.522769	0.888502	0.365733	0.526992	0.6940
0.3070	1.703689	0.523033	0.889680	0.366647	0.529073	0.6930
0.3080	1.698998	0.523291	0.890851	0.367560	0.531156	0.6920
0.3090	1.694321	0.523545	0.892016	0.368470	0.533242	0.6910
0.3100	1.689660	0.523795	0.893173	0.369379	0.535332	0.6900
0.3110	1.685014	0.524039	0.894324	0.370285	0.537424	0.6890
0.3120	1.680382	0.524279	0.895469	0.371189	0.539520	0.6880
0.3130	1.675765	0.524515	0.896606	0.372092	0.541618	0.6870
0.3140	1.671164	0.524745	0.897737	0.372992	0.543720	0.6860
0.3150	1.666576	0.524972	0.898861	0.373890	0.545824	0.6850
0.3160	1.662004	0.525193	0.899978	0.374785	0.547932	0.6840
0.3170	1.657445	0.525410	0.901089	0.375679	0.550043	0.6830
0.3180	1.652901	0.525623	0.902193	0.376571	0.552156	0.6820
0.3190	1.648372	0.525831	0.903291	0.377460	0.554273	0.6810
0.3200	1.643856	0.526034	0.904381	0.378347	0.556393	0.6800
0.3210	1.639355	0.526233	0.905466	0.379233	0.558517	0.6790
0.3220	1.634867	0.526427	0.906543	0.380116	0.560643	0.6780
0.3230	1.630394	0.526617	0.907614	0.380997	0.562772	0.6770
0.3240	1.625934	0.526803	0.908678	0.381876	0.564905	0.6760
0.3250	1.621488	0.526984	0.909736	0.382752	0.567041	0.6750

Продолжение

p	$-\log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.3260	1.617056	0.527160	0.910787	0.383627	0.569179	0.6740
0.3270	1.612637	0.527332	0.911832	0.384499	0.571322	0.6730
0.3280	1.608232	0.527500	0.912870	0.385370	0.573467	0.6720
0.3290	1.603841	0.527664	0.913901	0.386238	0.575615	0.6710
0.3300	1.599462	0.527822	0.914926	0.387104	0.577767	0.6700
0.3310	1.595097	0.527977	0.915945	0.387968	0.579922	0.6690
0.3320	1.590745	0.528127	0.916957	0.388829	0.582080	0.6680
0.3330	1.586406	0.528273	0.917962	0.389689	0.584241	0.6670
0.3340	1.582080	0.528415	0.918961	0.390546	0.586406	0.6660
0.3350	1.577767	0.528552	0.919953	0.391402	0.588574	0.6650
0.3360	1.573467	0.528685	0.920939	0.392255	0.590745	0.6640
0.3370	1.569180	0.528813	0.921919	0.393105	0.592919	0.6630
0.3380	1.564905	0.528938	0.922892	0.393954	0.595097	0.6620
0.3390	1.560643	0.529058	0.923859	0.394801	0.597278	0.6610
0.3400	1.556393	0.529174	0.924819	0.395645	0.599462	0.6600
0.3410	1.552156	0.529285	0.925772	0.396487	0.601650	0.6590
0.3420	1.547932	0.529393	0.926720	0.397327	0.603840	0.6580
0.3430	1.543720	0.529496	0.927661	0.398165	0.606035	0.6570
0.3440	1.539520	0.529595	0.928595	0.399000	0.608232	0.6560
0.3450	1.535332	0.529689	0.929523	0.399834	0.610043	0.6550
0.3460	1.531156	0.529780	0.930445	0.400665	0.612637	0.6540
0.3470	1.526992	0.529866	0.931360	0.401494	0.614845	0.6530
0.3480	1.522841	0.529949	0.932269	0.402321	0.617056	0.6520
0.3490	1.518701	0.530027	0.933172	0.403145	0.619271	0.6510
0.3500	1.514573	0.530101	0.934068	0.403967	0.621488	0.6500
0.3510	1.510457	0.530170	0.934958	0.404788	0.623710	0.6490
0.3520	1.506353	0.530236	0.935842	0.405605	0.625934	0.6480
0.3530	1.502260	0.530298	0.936719	0.406421	0.628162	0.6470
0.3540	1.498179	0.530355	0.937590	0.407234	0.630394	0.6460
0.3550	1.494109	0.530409	0.938454	0.408046	0.632629	0.6450
0.3560	1.490051	0.530458	0.939313	0.408855	0.634867	0.6440
0.3570	1.486004	0.530503	0.940165	0.409661	0.637109	0.6430
0.3580	1.481969	0.530545	0.941010	0.410466	0.639355	0.6420
0.3590	1.477944	0.530582	0.941850	0.411268	0.641604	0.6410
0.3600	1.473931	0.530615	0.942683	0.412068	0.643856	0.6400
0.3610	1.469929	0.530644	0.943510	0.412866	0.646112	0.6390
0.3620	1.465938	0.530670	0.944331	0.413661	0.648372	0.6380
0.3630	1.461959	0.530691	0.945145	0.414454	0.650635	0.6370
0.3640	1.457990	0.530708	0.945953	0.415245	0.652901	0.6360
0.3650	1.454032	0.530722	0.946755	0.416034	0.655171	0.6350
0.3660	1.450084	0.530731	0.947551	0.416820	0.657445	0.6340
0.3670	1.446148	0.530736	0.948341	0.417604	0.659723	0.6330
0.3680	1.442222	0.530738	0.949124	0.418386	0.662004	0.6320
0.3690	1.438307	0.530735	0.949901	0.419166	0.664288	0.6310
0.3700	1.434403	0.530729	0.950672	0.419943	0.666576	0.6300
0.3710	1.430509	0.530719	0.951437	0.420718	0.668868	0.6290
0.3720	1.426625	0.530705	0.952195	0.421491	0.671164	0.6280
0.3730	1.422752	0.530687	0.952948	0.422261	0.673463	0.6270
0.3740	1.418890	0.530665	0.953694	0.423029	0.675765	0.6260
0.3750	1.415037	0.530639	0.954434	0.423795	0.678072	0.6250
0.3760	1.411195	0.530609	0.955168	0.424558	0.680382	0.6240
0.3770	1.407364	0.530576	0.955896	0.425320	0.682696	0.6230

Продолжение

p	$-\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.3780	1.403542	0.530539	0.956617	0.426078	0.685014	0.6220
0.3790	1.399730	0.530498	0.957333	0.426835	0.687335	0.6210
0.3800	1.395929	0.530453	0.958042	0.427589	0.689660	0.6200
0.3810	1.392137	0.530404	0.958745	0.428341	0.691989	0.6190
0.3820	1.388355	0.530352	0.959442	0.429091	0.694321	0.6180
0.3830	1.384584	0.530296	0.960133	0.429838	0.696658	0.6170
0.3840	1.380822	0.530236	0.960818	0.430583	0.698998	0.6160
0.3850	1.377070	0.530172	0.961497	0.431325	0.701342	0.6150
0.3860	1.373327	0.530104	0.962170	0.432065	0.703689	0.6140
0.3870	1.369595	0.530033	0.962836	0.432803	0.706041	0.6130
0.3880	1.365871	0.529958	0.963497	0.433539	0.708396	0.6120
0.3890	1.362158	0.529879	0.964151	0.434272	0.710756	0.6110
0.3900	1.358454	0.529797	0.964800	0.435002	0.713119	0.6100
0.3910	1.354759	0.529711	0.965442	0.435731	0.715486	0.6090
0.3920	1.351074	0.529621	0.966078	0.436457	0.717857	0.6080
0.3930	1.347399	0.529528	0.966708	0.437181	0.720232	0.6070
0.3940	1.343732	0.529431	0.967332	0.437902	0.722610	0.6060
0.3950	1.340075	0.529330	0.967951	0.438621	0.724993	0.6050
0.3960	1.336428	0.529225	0.968563	0.439337	0.727380	0.6040
0.3970	1.332789	0.529117	0.969169	0.440051	0.729770	0.6030
0.3980	1.329160	0.529006	0.969769	0.440763	0.732165	0.6020
0.3990	1.325539	0.528890	0.970363	0.441472	0.734563	0.6010
0.4000	1.321928	0.528771	0.970951	0.442179	0.736966	0.6000
0.4010	1.318326	0.528649	0.971533	0.442884	0.739372	0.5990
0.4020	1.314733	0.528522	0.972108	0.443586	0.741783	0.5980
0.4030	1.311148	0.528393	0.972678	0.444286	0.744197	0.5970
0.4040	1.307573	0.528259	0.973242	0.444983	0.746616	0.5960
0.4050	1.304006	0.528122	0.973800	0.445678	0.749038	0.5950
0.4060	1.300448	0.527982	0.974352	0.446370	0.751465	0.5940
0.4070	1.296899	0.527838	0.974898	0.447060	0.753896	0.5930
0.4080	1.293359	0.527690	0.975438	0.447748	0.756331	0.5920
0.4090	1.289827	0.527539	0.975972	0.448433	0.758770	0.5910
0.4100	1.286304	0.527385	0.976500	0.449116	0.761213	0.5900
0.4110	1.282790	0.527227	0.977023	0.449796	0.763660	0.5890
0.4120	1.279284	0.527065	0.977539	0.450474	0.766112	0.5880
0.4130	1.275786	0.526900	0.978049	0.451149	0.768568	0.5870
0.4140	1.272297	0.526731	0.978553	0.451822	0.771027	0.5860
0.4150	1.268817	0.526559	0.979051	0.452493	0.773491	0.5850
0.4160	1.265345	0.526383	0.979544	0.453160	0.775960	0.5840
0.4170	1.261881	0.526204	0.980030	0.453826	0.778432	0.5830
0.4180	1.258425	0.526022	0.980511	0.454489	0.780909	0.5820
0.4190	1.254978	0.525836	0.980985	0.455150	0.783390	0.5810
0.4200	1.251539	0.525646	0.981454	0.455808	0.785875	0.5800
0.4210	1.248108	0.525453	0.981917	0.456463	0.788365	0.5790
0.4220	1.244685	0.525257	0.982373	0.457116	0.790859	0.5780
0.4230	1.241270	0.525057	0.982824	0.457767	0.793357	0.5770
0.4240	1.237864	0.524854	0.983269	0.458415	0.795859	0.5760
0.4250	1.234465	0.524648	0.983708	0.459061	0.798366	0.5750
0.4260	1.231075	0.524438	0.984141	0.459704	0.800877	0.5740
0.4270	1.227692	0.524224	0.984569	0.460344	0.803393	0.5730
0.4280	1.224317	0.524008	0.984990	0.460982	0.805913	0.5720
0.4290	1.220950	0.523788	0.985405	0.461618	0.808437	0.5710

Продолжение

p	$-p\log_2 p$	$-p\log_2 p$	H	$-q\log_2 q$	$-\log_2 q$	q
0.4300	1.217591	0.523564	0.985815	0.462251	0.810966	0.5700
0.4310	1.214240	0.523338	0.986219	0.462881	0.813499	0.5690
0.4320	1.210897	0.523107	0.986617	0.463509	0.816037	0.5680
0.4330	1.207561	0.522874	0.987008	0.464134	0.818579	0.5670
0.4340	1.204233	0.522637	0.987394	0.464757	0.821126	0.5660
0.4350	1.200913	0.522397	0.987775	0.465378	0.823677	0.5650
0.4360	1.197600	0.522154	0.988149	0.465995	0.826233	0.5640
0.4370	1.194295	0.521907	0.988517	0.466611	0.828793	0.5630
0.4380	1.190997	0.521657	0.988880	0.467223	0.831358	0.5620
0.4390	1.187707	0.521403	0.989237	0.467833	0.833927	0.5610
0.4400	1.184425	0.521147	0.989588	0.468441	0.836501	0.5600
0.4410	1.181149	0.520887	0.989933	0.469046	0.839080	0.5590
0.4420	1.177882	0.520624	0.990272	0.469648	0.841663	0.5580
0.4430	1.174621	0.520357	0.990605	0.470248	0.844251	0.5570
0.4440	1.171368	0.520088	0.990932	0.470845	0.846843	0.5560
0.4450	1.168123	0.519815	0.991254	0.471439	0.849440	0.5550
0.4460	1.164884	0.519538	0.991570	0.472031	0.852042	0.5540
0.4470	1.161653	0.519259	0.991880	0.472621	0.854649	0.5530
0.4480	1.158429	0.518976	0.992184	0.473207	0.857260	0.5520
0.4490	1.155213	0.518690	0.992482	0.473792	0.859876	0.5510
0.4500	1.152003	0.518401	0.992774	0.474373	0.862496	0.5500
0.4510	1.148801	0.518109	0.993061	0.474952	0.865122	0.5490
0.4520	1.145605	0.517814	0.993342	0.475528	0.867752	0.5480
0.4530	1.142417	0.517515	0.993617	0.476102	0.870387	0.5470
0.4540	1.139236	0.517213	0.993886	0.476673	0.873027	0.5460
0.4550	1.136062	0.516908	0.994149	0.477241	0.875672	0.5450
0.4560	1.132894	0.516600	0.994407	0.477807	0.878321	0.5440
0.4570	1.129734	0.516288	0.994658	0.478370	0.880976	0.5430
0.4580	1.126580	0.515974	0.994904	0.478930	0.883635	0.5420
0.4590	1.123434	0.515656	0.995144	0.479488	0.886299	0.5410
0.4600	1.120294	0.515335	0.995378	0.480043	0.888969	0.5400
0.4610	1.117161	0.515011	0.995607	0.480595	0.891643	0.5390
0.4620	1.114035	0.514684	0.995829	0.481145	0.894322	0.5380
0.4630	1.110916	0.514354	0.996046	0.481692	0.897006	0.5370
0.4640	1.107803	0.514021	0.996257	0.482237	0.899695	0.5360
0.4650	1.104697	0.513684	0.996462	0.482778	0.902389	0.5350
0.4660	1.101598	0.513345	0.996662	0.483317	0.905088	0.5340
0.4670	1.098506	0.513002	0.996856	0.483853	0.907793	0.5330
0.4680	1.095420	0.512656	0.997043	0.484387	0.910502	0.5320
0.4690	1.092340	0.512308	0.997225	0.484918	0.913216	0.5310
0.4700	1.089267	0.511956	0.997402	0.485446	0.915936	0.5300
0.4710	1.086201	0.511601	0.997572	0.485971	0.918660	0.5290
0.4720	1.083141	0.511243	0.997737	0.486494	0.921390	0.5280
0.4730	1.080088	0.510882	0.997896	0.487014	0.924125	0.5270
0.4740	1.077041	0.510517	0.998049	0.487531	0.926865	0.5260
0.4750	1.074001	0.510150	0.998196	0.488046	0.929611	0.5250
0.4760	1.070967	0.509780	0.998337	0.488557	0.932361	0.5240
0.4770	1.067939	0.509407	0.998473	0.489066	0.935117	0.5230
0.4780	1.064917	0.509031	0.998603	0.489572	0.937878	0.5220
0.4790	1.061902	0.508651	0.998727	0.490076	0.940645	0.5210
0.4800	1.058894	0.508269	0.998846	0.490577	0.943416	0.5200
0.4810	1.055891	0.507884	0.998958	0.491074	0.946194	0.5190

Продолжение

p	$-p \log_2 p$	$-p \log_2 p$	H	$-q \log_2 q$	$-\log_2 q$	q
0.4820	1.052895	0.507495	0.999065	0.491570	0.948976	0.5180
0.4830	1.049905	0.507104	0.999166	0.492062	0.951764	0.5170
0.4840	1.046921	0.506710	0.999261	0.492551	0.954557	0.5160
0.4850	1.043943	0.506313	0.999351	0.493038	0.957356	0.5150
0.4860	1.040972	0.505912	0.999434	0.493522	0.960160	0.5140
0.4870	1.038006	0.505509	0.999512	0.494003	0.962969	0.5130
0.4880	1.035047	0.505103	0.999584	0.494482	0.965784	0.5120
0.4890	1.032094	0.504694	0.999651	0.494957	0.968605	0.5110
0.4900	1.029146	0.504282	0.999711	0.495430	0.971431	0.5100
0.4910	1.026205	0.503867	0.999766	0.495900	0.974262	0.5090
0.4920	1.023270	0.503449	0.999815	0.496367	0.977100	0.5080
0.4930	1.020340	0.503028	0.999859	0.496831	0.979942	0.5070
0.4940	1.017417	0.502604	0.999896	0.497292	0.982791	0.5060
0.4950	1.014500	0.502177	0.999928	0.497751	0.985645	0.5050
0.4960	1.011588	0.501748	0.999954	0.498206	0.988504	0.5040
0.4970	1.008682	0.501315	0.999974	0.498659	0.991370	0.5030
0.4980	1.005782	0.500880	0.999988	0.499109	0.994241	0.5020
0.4990	1.002888	0.500441	0.999997	0.499556	0.997117	0.5010
0.5000	1.000000	0.500000	1.000000	0.500000	1.000000	0.5000

ОГЛАВЛЕНИЕ

СТАТИСТИКА РЕЧИ

Утверждено к печати
Научным советом по кибернетике
при Президиуме АН СССР

Художник *В. В. Грибакин*
Технический редактор *Г. А. Бессонова*
Корректоры *Л. М. Бова, Ш. А. Иванова*
и *Н. В. Лихарева*

Сдано в набор 30/XII 1966 г. Подписано к печати 16/IV 1968 г. РИСО АН СССР № 4-175В. Формат бумаги 60×90¹/₁₆. Бум. л. 8⁷/₁₆. Печ. л. 16¹/₄ + 2 вкл. (⁵/₈ печ. л.) = 16⁷/₈ усл. печ. л. Уч.-изд. л. 20.15. Изд. № 3047. Тип. зак. № 1372. М-11490. Тираж 3500. Бумага типографская № 1. *Цена 1 р. 37 к.*

Ленинградское отделение издательства «Наука»
Ленинград, В-164, Менделеевская лин., д. 1

1-я тип. издательства «Наука». Ленинград, В-34, 9 линия, д. 12

ИСПРАВЛЕНИЯ И ОПЕЧАТКИ

Страница	*Строка*	*Напечатано*	*Должно быть*
13	7 сверху, формула (10)	H''_n	$H''_{n'}$
17	6 сверху, формула (12)	$\frac{\gamma_s}{n}$	$\frac{\gamma_s}{N}$
17	2 снизу	Analising	Analysing
57	4 сверху, формула (1)	$S_{pS}\log_2 S$	$Sp_S \log_2 S$
57	13 сверху, формула (2)	$p_N \log_2 N$	$p_S \log_2 p_S$
59	Таблица 5, столб. 5	H_n	$\underline{H}_n$
216	18 снизу	aquisition	acquisition