*УДК 81'32, 811.13, 519.765*

# ПРИРОДА ЭКСПОНЕНЦИАЛЬНОГО РАСПРЕДЕЛЕНИЯ СЛОВ ПО ЧИСЛУ ЗНАЧЕНИЙ

**Г. Д. Селезнев**

*Воронежский государственный университет*

Предлагается математическая модель динамики распределения слов по числу значений. Модель основана на математическом аппарате цепей Маркова и используется для анализа словарей романских языков.

В работах В. Т. Титова [1, 2] представлены распределения количества слов в словарях романских языков по числу их значений [1, табл. 5]. Частотный анализ текстов и словарей показывает, что чаще встречаются слова с малым числом значений — одним, двумя; гораздо реже встречаются слова с тремя, четырьмя и более значениями (до 14 в латинском языке). Чем больше значений имеет слово, тем реже оно встречается. В настоящей работе предпринята попытка математического анализа и теоретического осмысления этих экспериментальных данных.

## 1. ЭКСПОНЕНЦИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ СЛОВ ПО ЧИСЛУ ЗНАЧЕНИЙ

Для корректности сравнения распределения для разных языков нормировались, т.е. вычислялась вероятность количества слов с данным числом значений, и результат умножался на 10000; тем самым предполагалось, что такое количество слов содержит некий усредненный словарь. Средствами табличного процессора EXCEL проводилась аппроксимация этих распределений. Наилучший результат достигается при аппроксимации данных экспоненциальным распределением вида

$$N_i = Vp_i = K\exp(-\alpha i), \qquad (1)$$

где $N_i$ — количество слов с числом значений равным $i$, $i = 1, 2, \ldots, n$; $V$ — объем словаря соответствующего языка (принимался равным 10000); $p_i$ — вероятность распределения слов по числу значений; $\alpha$, $K$ — подбираемые при аппроксимации показатель экспоненты и масштабный коэффициент.

Для увеличения достоверности аппроксимации для некоторых языков не принимались в расчет слова с самым большим количеством значений; не более 6 слов для румынского словаря (0.08 % от объема словаря).

Результаты аппроксимации приведены в табл. 1 и рис. 1 и 2.

Таблица 1

| Язык | $K$ | $\alpha$ | Достоверность аппроксимации (%) |
|---|---|---|---|
| Латинский | 12265 | 0.73 | 99.2 |
| Итальянский | 48841 | 1,88 | 98.7 |
| Испанский | 33774 | 1,54 | 99.5 |
| Португальский | 50771 | 1,73 | 99.8 |
| Французский | 28645 | 1,47 | 99.8 |
| Румынский | 37144 | 1,69 | 96.3 |

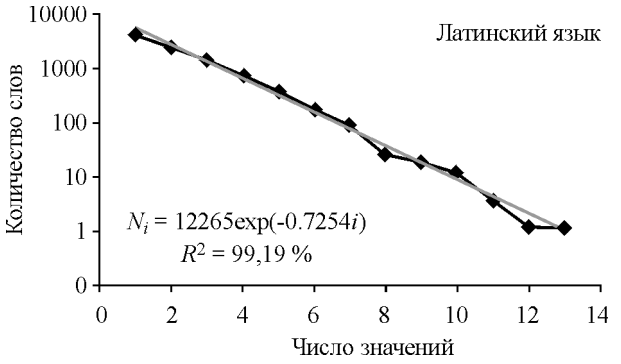



*Рис. 1.*

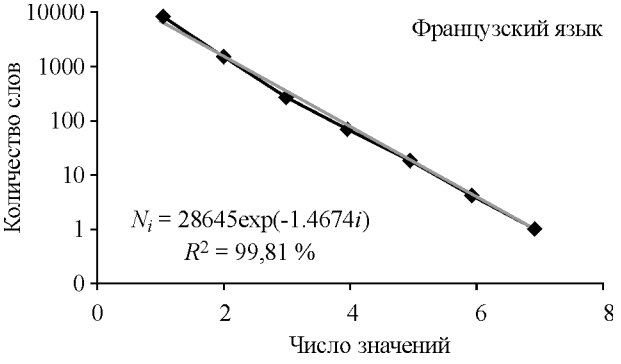



*Рис. 2.*

По результатам аппроксимации может быть сформулирована гипотеза о том, что, по крайней мере, для романских языков, ***распределение количества слов по числу их значений является экспоненциальным распределением***.

Подобные гипотезы высказывались и ранее [3, 4]. Так, в работе Ю. А. Тулдавы [3] приводится выражение, аналогичное (1), но с показателем экспоненты, в котором стоит не величина $i$, как в формуле (1), а $\sqrt{i}$. Такая зависимость, согласно [3], дает лучший результат аппроксимации для английского, венгерского и русского (только глаголы) языков. В работе Ю. А. Шрейдера [4] приводится вывод экспоненциального распределения количества слов от некой числовой величины, интерпретируемой автором как «сложность» или «энергия» порождения слова. Вывод основан, по выражению автора [4], на «некоторых термодинамических аналогиях».

## 2. «ТЕРМОДИНАМИЧЕСКАЯ» МОДЕЛЬ ИСТОРИЧЕСКОЙ ДИНАМИКИ ЛЕКСИЧЕСКОЙ СИСТЕМЫ

Для обоснования и вывода выражения (1) также воспользуемся «термодинамической аналогией». Известно, что в физике, точнее в статистической термодинамике, также встречается экспоненциальное распределение. Это так называемое распределение Больцмана [5]. Согласно этому распределению, ансамбль систем, которые могут находиться в состояниях с различными энергиями, в состоянии термодинамического равновесия с окружающей средой подчиняется экспоненциальному распределению вероятностей

$$p_i = A\exp\left(-\frac{E_i}{kT}\right), \qquad (2)$$

где $i$ — номер состояния с энергией $E_i$, $i = 1,2,\ldots, n$; $T$ — температура системы (абсолютная); $k$ — константа Больцмана; $A$ — нормировочный коэффициент; $P_i$ — вероятность пребывания системы в состоянии с энергией $E_i$.

Данное распределение является следствием двух фундаментальных законов физики:

— закона сохранения энергии (первое начало термодинамики), согласно которому в замкнутой системе энергия со временем не изменяется

$$E = \sum_{i=1}^{n} E_i = const; \qquad (3)$$

— закона возрастания энтропии (второе начало термодинамики), согласно которому в замкнутой системе энтропия $H$ возрастает, а в состоянии термодинамического равновесия принимает максимальное значение

$$H = k\sum_{i=1}^{n} p_i \ln p_i = \max(E_i, T). \qquad (4)$$

Т.е. из (3) и (4) следует (2).

По аналогии с законами термодинамики можно предположить, что в языке также имеют место некий «закон сохранения», но теперь уже не энергии, а количества значений, и закон возрастания энтропии распределения этих значений.

1. Закон сохранения общего, суммарного количества значений всех слов некоторой замкнутой языковой системы; иными словами — закон сохранения объема семантической информации. Разумно предположить, что такой замкнутой системой может являться текст или словарь, или даже язык в целом на данном этапе его исторического развития.

2. Закон возрастания энтропии распределения значений, содержащихся в тексте или словаре по однозначным, двузначным, трехзначным,…, $i$ — значным словам. В устойчивой — замкнутой в семантическом отношении языковой системе — энтропия такого распределения, вычисляемая по формуле

$$H = \sum_{i=1}^{n} p_i \ln p_i \qquad (5)$$

принимает максимальное значение.

Если предположить, что в формуле (1) величина $i$ — это некая семантическая «энергия», семан-

Таблица 2

| Язык | Семантическая «температура» $1/\alpha$ | Семантическая энтропия $H$ | Отношение вероятностей переходов ($p/q$) |
|---|---|---|---|
| Итальянский | 0.86 | 0.70 | 0.31 |
| Испанский | 0.96 | 0.98 | 0.35 |
| Португальский | 0.81 | 0.95 | 0.29 |
| Французский | 0.75 | 0.94 | 0.26 |
| Румынский | 0.81 | 0.77 | 0.29 |

тическая «емкость» слова, то величина обратная показателю экспоненты $1/\alpha$ может быть условно названа «семантической температурой» текста, т.е. в формулах (1) и (2)

$$i \sim E_i,\ 1/\alpha \sim kT.$$

В той степени, в которой распределения, приведенные в работе В. Т. Титова [1], полученные на основе анализа словарей, отражают некие универсальные свойства романских языков, эта семантическая «температура» и семантическая энтропия характеризует состояние данного конкретного языка (табл. 2).

Разумно предположить, что если провести подобный анализ не языка в целом, а каких-либо текстов, даже с учетом контекста, то получатся такого же типа экспоненциальные распределения, и тогда можно получить оценку семантической температуры текстов одного типа или одного стиля. Можно предположить, что «мягкие» — в терминологии А. А. Налимова [6] — художественные, поэтические тексты, изобилующие многозначностью, богатством семантики, будут иметь высокую температуру (малую величину $\alpha$) и пологий вид кривых распределения. А жесткие тексты деловой, научной прозы, наоборот, будут «холодными», будут иметь большую величину показателя $\alpha$ в формуле (1). При переводах текстов с одного языка на другой требование адекватности перевода равносильно закону сохранения общего числа значений в тексте — оригинале и тексте — переводе.

## 3. МОДЕЛЬ СЛУЧАЙНОГО БЛУЖДАНИЯ

В работе В. В. Кромера [7] предложена модель исторической динамики лексической системы как процесса случайного блуждания. Случайное блуждание понимается как случайный процесс приобретения словом нового значения или утрата прежде существовавшего значения, т.е. «блуждание» слова по полисемическим классам («зонам» в терминологии В. В. Кромера) одно-, двух-, трех- … $n$-значных слов. Стационарным состоянием лексической системы считается состояние динамического равновесия, при котором количество слов — лексем в словаре и их распределение по полисемическим классам остается неизменным.

Целью настоящей работы является согласование моделей случайного блуждания В. В. Кромера [7] и предлагаемой «термодинамической» модели, вывод формулы (1) на основании представлений модели случайного блуждания.

В качестве модели исторической динамики лексической системы как процесса случайного блуждания [2] используем математический аппарат цепей Маркова. Рассмотрим цепь со следующим графом (рис. 3).

Здесь $c$, $d$, $r_{11}$, $p$, $q$ — вероятности переходов цепи Маркова, для которых выполняются следующие очевидные условия:

$$c + p + r_{11} = 1,\ q + p + r = 1,\ q > p.$$

В отличие от модели В. В. Кромера, на вероятности переходов не накладывается никаких других дополнительных условий; в частности, будем счи-

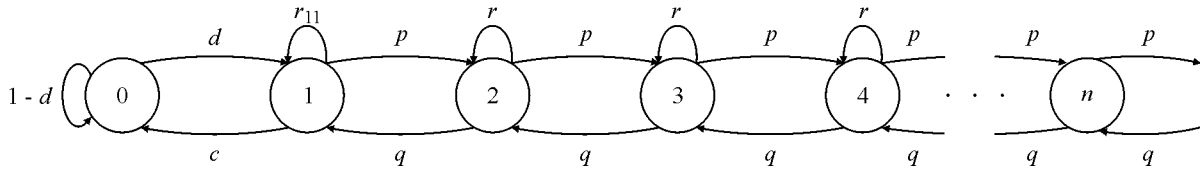
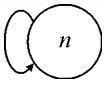



*Рис. 3.*

*Обозначения*

| | |
|---|---|
|  | Состояние Цепи Маркова, соответствующее тому, что слово имеет ровно $n$ значений, где $n = 0, 1, 2, \ldots$, т.е. принадлежит к классу $n$-значных слов. При $n = 0$ слово не имеет значений; полагается, что таких слов бесконечно много. |
|  | Увеличение на единицу значений слова — слово приобретает новое значение, переходит из класса $n$-значных в класс $(n+1)$-значных. |
|  | Уменьшение на единицу значений слова — слово утрачивает одно из значений, переходит из класса $n$-значных в класс $(n-1)$-значных. |
|  | Количество значений слова остается неизменным за условную единицу времени. |

тать их не зависящими от количества слов в полисемическом классе.

Зная вероятности ***переходов*** из состояния в состояние цепи Маркова, можно получить выражения для финальных (при $t \to \infty$) вероятностей ***состояний*** этой цепи. В данной интерпретации это будут вероятности того, что слово, наугад выбранное из словаря, будет иметь ровно $n$ значений, т.е будет принадлежать к классу $n$-значных слов. Опуская промежуточные математические выкладки, приведем итоговые формулы для вероятностей:

$$p_0 = \frac{c}{d+c},$$
$$p_n = \frac{d(q-p)}{(d+c)p}\left(\frac{p}{q}\right)^n \text{ при } n > 0. \qquad (6)$$

Данное выражение согласуется с полученным выше результатом об экспоненциальном распределении слов по числу значений. Действительно, обозначим объем словаря $V = Kp_1$, где $p_1 = \frac{d(q-p)}{(d+c)p}$, и произведем в формуле (2) замену переменных

$$\ln\left(\frac{p}{q}\right) = -\alpha, \text{ откуда}$$
$$\left(\frac{p}{q}\right)^n = \exp(-\alpha n). \qquad (7)$$

В этих обозначениях выражение (2) переходит в (1).

Выражение (2) описывает устойчивое, стационарное состояние лексической системы, т.е. состояние динамического равновесия, при котором распределение вероятностей слов по полисемическим классам совпадает с финальным распределением состояний цепи Маркова и не меняется со временем. Отношение вероятностей ($p$/$q$) может быть получено на основе анализа соответствующих национальных словарей. Например, в табл. 2 приведены численные значения величины «семантической температуры» $1/\alpha = kT$ романских языков. Зная $\alpha$, из выражений (7) из можно получить значения ($p$/$q$) (табл. 2).

## ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Из (6) следует, что вероятности распределения слов по классам $p_n$ не зависят от самих значений вероятностей переходов слов из класса в класс $p$ и $q$, но определяются их отношением — величиной ($p$/$q$). Т.е. распределение слов по полисемическим классам не зависит от «скорости случайного блуждания» слов по классам. А величина отношения ($p$/$q$) не зависит от $n$, не изменяется от класса к классу.

Для получения теоретических численных значений количеств $n$-значных слов в словаре, т.е объем соответствующего полисемического класса $N_n$, следует умножить величину объема словаря $V$ на соответствующее значение вероятности $p_n$, т.е $N_n = Vp_n$.

СПИСОК ЛИТЕРАТУРЫ

1. *Титов В.Т.* Общая квантитативная лексикология романских языков. —Воронеж: Изд-во Воронежского государственного университета, 2002, — 240 с.

2. *Титов В.Т.* Лексико-семантическая парадигматика романских языков. Вестник ВГУ, сер. «Лингвистика и межкультурная коммуникация» Вып. 1. 2004. — С. 3.

3. *Тулдава Ю.А.* Проблемы и методы квантитативно-системного исследования лексики. Таллин: Валгус, 1987. — 204 с.

4. *Шрейдер Ю.А.* О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). Проблемы передачи информации. Т. 3, вып. 1. М., 1967. — С. 57—63.

5. *Ландау Л.Д., Лившиц М.Е.* Статистическая физика. М.: Наука, 1964. — 568 с.

6. *Налимов В.В.* Вероятностная модель языка. О соотношении естественных и искусственных языков. М.: Наука. 1979. — 304 с.

7. *Кромер В.В.* Историческая динамика лексической системы. Проблемы компьютерной лингвистики: Сб-к. научн. трудов / Под ред. А. А. Кретова. — Вып. 1. — Воронеж: РИЦ ЕФ ВГУ, 2004. — С. 49—53.

8. *Селезнев Г.Д.* Природа экспоненциального распределения слов по числу значений. Проблемы компьютерной лингвистики: Сб-к. научн. трудов / Под ред. А. А. Кретова. — Вып. 2. — Воронеж: РИЦ ЕФ ВГУ, 2005. — С. 169—173.

9. *Селезнев Г.Д.* Математическая модель динамики лексической системы. Проблемы лингвистической прогностики: Сб-к. научн. трудов / Под ред. А. А. Кретова. — Вып. 4. — Воронеж: Изд. ВГУ, 2007. — С. 171—175.