УДК 811.111’371

# ФУНКЦИОНАЛЬНЫЕ ЗАКОНОМЕРНОСТИ АНГЛИЙСКОЙ ПОЛИСЕМИИ

**И. А. Терентьева**

***Воронежский государственный университет***

Поступила в редакцию 18 марта 2010 г.

**Аннотация:** *в статье исследуется распределение функциональной нагрузки многозначных слов, производится аппроксимация экспериментального материала различными законами распределений, а также подбор коэффициентов для формул и их статистическая обработка.*
**Ключевые слова:** *полисемия, функциональная нагрузка многозначных слов, экспоненциальное распределение, закон Ципфа, закон Ципфа-Мандельброта.*

**Abstract:** *the article explores the distribution of functional load of polysemic words. The approximation of the experimental data based on different laws of distributions is carried out. Coefficients for formulae are selected and the statistic analysis of coefficients is carried out.*
**Key words:** *polysemy, functional load of polysemic words, exponential function, Zipf’s law, Zipf-Mandelbrot law.*

Данная работа посвящена исследованию распределения функциональной нагрузки между значениями многозначных слов. Объект исследования – частотно-семантический словарь [1]. Предметом исследования являются закономерности, управляющие функционированием значений многозначных слов.

Первым этапом работы было создание электронной базы частотно-семантического словаря «The semantic count of English Words», для этого из частотно-семантического словаря И. Лорджа и Э. Торндайка были выбраны все многозначные слова.

Значения слов в словаре распределены согласно Oxford English Dictionary, а относительная частота употребительности значения слова в словаре И. Лорджа и Э. Торндайка указана в промилле.

Следующий этап работы – разделение многозначных слов на группы по количеству значений и анализ этих групп. Значения слов в группах распределены согласно убыванию их частотности, в каждой группе должно быть не менее четырех слов, поскольку меньшие вариативные ряды не обладают достаточной статистической информативностью. Таким образом, в группах оказалось 11,914 слов и общее количество групп – 53.

Для того чтобы понять, каким именно закономерностям подчиняется распределение значений многозначных слов, была проведена аппроксимация этих данных следующими типами распределений:

– экспоненциальным ($N_i = K_{exp}(-\alpha i)$, где $i$ – количество значений слова, а $K$, $\alpha$ – подбираемые коэффициенты);

– гиперболическим Ципфа ($N_i = Ai^{\alpha}$, где $i$ – количество значений слова, а $\alpha$ – степень с изменяющимся значением);

– распределением Ципфа-Мандельброта ($N_i = A(B+i)^{-\beta}$, где $i$ – количество значений слова, $A$, $B$ – подбираемые коэффициенты, $\beta$ – степень с изменяющимся значением) [2, с. 172];

– распределением, полученным из «Словаря языка Пушкина» [3, с. 293].

Распределение имеет следующий вид:

$$Фн\ \underset{п.\ знач.гл}{1зн.} = \frac{1000}{(n-1)} * k\ ,$$

где $n$ – количество значений глагола, $K = 0{,}7$. Формула функциональной нагрузки для второго значения имеет следующий вид:

$$Фн\ \underset{п.\ знач.гл}{2зн.} = \frac{1000 - Фн1зн.}{n-1}\ .$$

Таким образом, мы получили теоретическую пропорцию 700/300. Чтобы получить пропорцию для остальных значений, необходимо каждый раз делить последнее число пропорции на два и полученное частное отнять от первого числа пропорции.

Степень близости аппроксимации экспериментальных данных выбранной функцией оценивается коэффициентом детерминации $R^2$ в Microsoft Excel.

Формула для вычисления коэффициента детерминации:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2},$$

где $y_i$ — экспериментальные данные, а $f_i$ — теоретические значения модели. Таким образом, если есть несколько подходящих вариантов типов аппроксимирующих функций, можно выбрать функцию с большим коэффициентом детерминации (стремящимся к 1).

**Формула, полученная из «Словаря языка Пушкина».** Наибольшая близость аппроксимации функциональной нагрузки у слов с двумя и тремя значениями, она составляет 99,9 % и 98,3 % соответственно. Среднее значение составляет 73,6 %. На графике 1 изображено изменение точности функциональной нагрузки для каждой группы многозначных слов.

**Экспоненциальное распределение.** Аппроксимация экспериментальных данных осуществляется путем построения их графика для каждого значения многозначного слова с последующим подбором подходящей аппроксимирующей функции, в данном случае экспоненциальной.

Наибольшая точность распределения функциональной нагрузки (99,6 %) наблюдается у двузначных слов. Минимальная близость аппроксимации у слов с сорока девятью значениями (63,4 %). Средняя точность составляет (94,8 %). График 2 показывает точность распределения значений.

**Формула Ципфа.** Экспериментальные данные каждой группы многозначных слов из частотно-семантического словаря были аппроксимированы с помощью закона Ципфа.

Максимальная точность распределения функциональной нагрузки у слов с тремя и четырьмя значе-

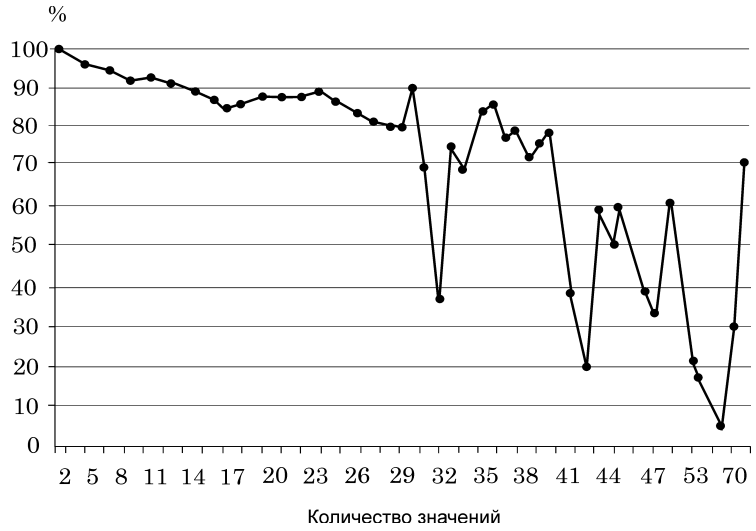


*Рис. 1.* Качество аппроксимации функциональной нагрузки формулой из словаря Пушкина

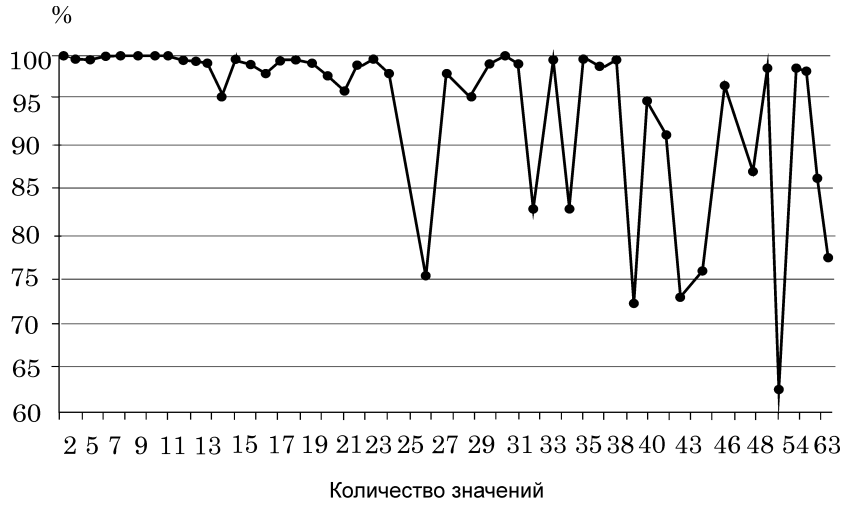


*Рис. 2.* Качество аппроксимации функциональной нагрузки экспоненциальным распределением

ниями (99,9 %). Наименьшая – у слов с шестьюдесятью одним значением (82,3 %). Средняя точность составляет 98,3 %. График 3 показывает, с какой точностью описываются экспериментальные данные законом Ципфа.

**Формула Ципфа-Мандельброта.** Наибольшая близость аппроксимации функциональной нагрузки у двузначных слов (100 %). Минимальная точность – у пятидесятишестизначных слов (97,1 %). Средняя точность – 99,5 %. На графике 4 изображено изменение точности функциональной нагрузки для каждой группы многозначных слов.

Как мы видим, наиболее стабильно описывает имеющееся у нас множество закон Ципфа-Мандельброта. Ему уступает на 1,2 % формула Ципфа, на третьем месте находится экспоненциальное распределение, которое имеет на 5,1 % и на 3,5 % меньшую точность, чем распределение Ципфа-Мандельброта и Ципфа соответственно. На последнем месте находится формула, полученная из «Словаря языка Пушкина», что может объясняться использованием в этой формуле констант в отличие от трех других, где используются коэффициенты с изменяющейся величиной. Поэтому было решено подобрать постоянные коэффициенты для этих законов.

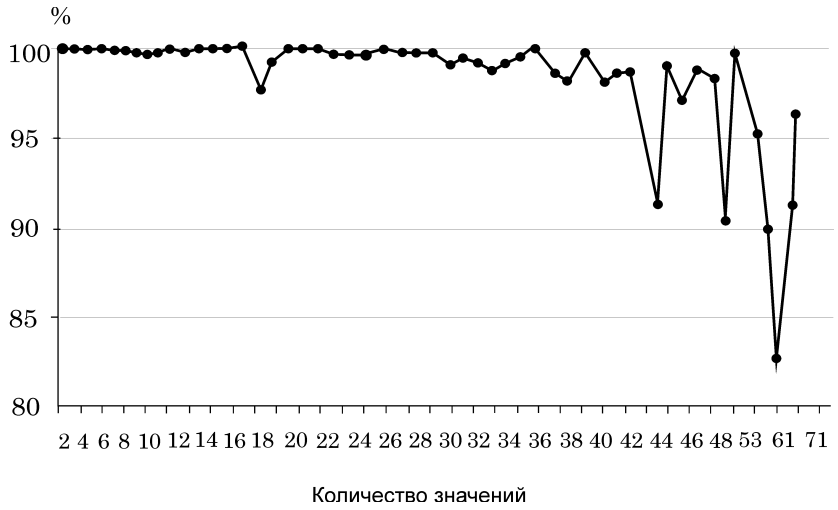


*Рис. 3.* Качество аппроксимации функциональной нагрузки значений законом Ципфа

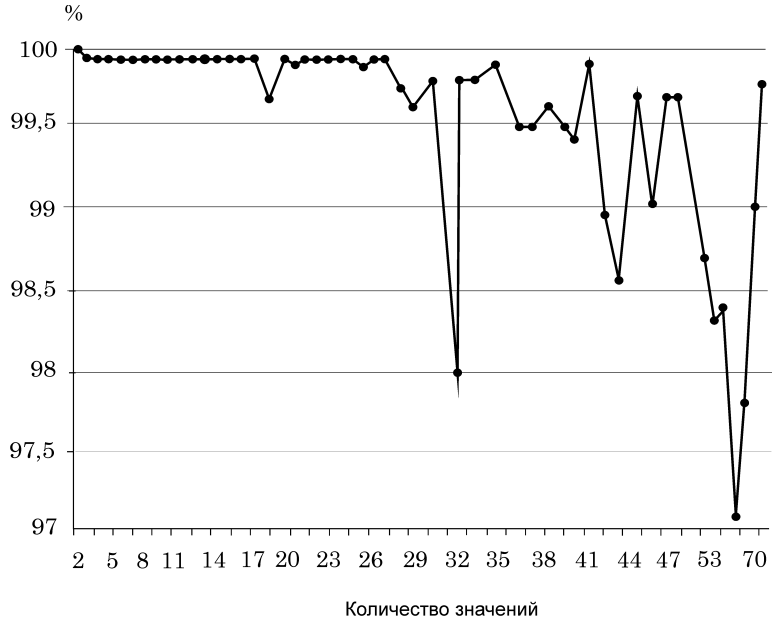


*Рис. 4.* Качество аппроксимации функциональной нагрузки значений законом Ципфа-Мандельброта

Во всех группах многозначных слов для каждого распределения имеются свои коэффициенты. Так как групп у нас 53, то получается, что, например, коэффициент $K$ в экспоненциональном распределении имеет 53 различных значения. Для того чтобы получить единый коэффициент, было решено использовать его среднеарифметическое значение.

Таким образом, мы получили усредненные коэффициенты для экспоненты $K$ = 847,24 и $\alpha$ = -0,61. Затем мы подставили полученные значения в каждую группу многозначных слов, и после проделанной операции экспоненциальное распределение описывает имеющиеся у нас экспериментальные данные с точностью 89,13 %.

Аналогичная работа была проделана для формулы Ципфа и закона Ципфа-Мандельброта. В таблицах представлены коэффициенты для этих распределений.

Таблица 1

*Значения подбираемых коэффициентов для закона Ципфа-Мандельброта*

| Кол-во значений | A | β | B | Кол-во значений | A | β | B | Кол-во значений | A | β | B |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 799,96 | -1,36 | 0,1 | 20 | 800,02 | -1,54 | 0,74 | 38 | 800 | -1,56 | 0,88 |
| 3 | 799,97 | -1,5 | 0,19 | 21 | 800,03 | -1,6 | 0,63 | 39 | 799,96 | -1,57 | 0,82 |
| 4 | 800 | -1,6 | 0,22 | 22 | 799,86 | -1,63 | 0,57 | 40 | 800,03 | -1,59 | 0,8 |
| 5 | 800,16 | -1,6 | 0,29 | 23 | 800,14 | -1,6 | 0,56 | 41 | 801,29 | -1,57 | 0,82 |
| 6 | 800,02 | -1,6 | 0,34 | 24 | 800,07 | -1,58 | 0,64 | 42 | 799,99 | -1,46 | 1,8 |
| 7 | 800,05 | -1,63 | 0,36 | 25 | 800,08 | -1,6 | 0,66 | 43 | 800,08 | -1,35 | 2,01 |
| 8 | 800,57 | -1,65 | 0,35 | 26 | 799,44 | -1,63 | 0,63 | 44 | 800,24 | -1,52 | 1,07 |
| 9 | 800,13 | -1,57 | 0,49 | 27 | 800 | -1,56 | 0,76 | 45 | 800,24 | -1,47 | 1,28 |
| 10 | 800,5 | -1,55 | 0,51 | 28 | 800,04 | -1,55 | 0,81 | 46 | 800,22 | -1,51 | 1,07 |
| 11 | 800,17 | -1,6 | 0,5 | 29 | 800,21 | -1,5 | 0,8 | 47 | 800,4 | -1,47 | 1,23 |
| 12 | 800,07 | -1,57 | 0,54 | 30 | 800 | -1,65 | 0,62 | 48 | 800,42 | -1,44 | 1,51 |
| 13 | 800,03 | -1,53 | 0,64 | 31 | 800,01 | -1,5 | 0,96 | 49 | 800,4 | -1,56 | 0,82 |
| 14 | 800,06 | -1,63 | 0,5 | 32 | 799,99 | -1,44 | 1,18 | 53 | 800,82 | -1,49 | 1,24 |
| 15 | 799,99 | -1,61 | 0,55 | 33 | 800,18 | -1,53 | 0,9 | 56 | 800,39 | -1,5 | 1,47 |
| 16 | 800 | -1,57 | 0,65 | 34 | 799,99 | -1,53 | 0,94 | 61 | 801,84 | -1,45 | 1,42 |
| 17 | 800,09 | -1,54 | 0,71 | 35 | 800,02 | -1,58 | 0,71 | 70 | 800,65 | -1,4 | 1,9 |
| 18 | 800,12 | -1,56 | 0,64 | 36 | 799,99 | -1,65 | 0,54 | 71 | 800,12 | -1,46 | 1,65 |
| 19 | 800,07 | -1,53 | 0,71 | 37 | 800,06 | -1,54 | 0,91 | Среднеариф. | 800,17 | -1,54 | 0,82 |

Таблица 2

*Значения подбираемых коэффициентов для закона Ципфа*

| Кол-во значений | *A* | *α* | Кол-во значений | *A* | *α* | Кол-во значений | *A* | *α* |
|---|---|---|---|---|---|---|---|---|
| 2 | 706,65 | -1,27 | 20 | 340,82 | -1,2 | 38 | 298,5 | -1,21 |
| 3 | 618,26 | -1,36 | 21 | 368,13 | -1,25 | 39 | 314 | -1,17 |
| 4 | 577,83 | -1,44 | 22 | 384,39 | -1,27 | 40 | 312,22 | -1,24 |
| 5 | 530,17 | -1,36 | 23 | 385 | -1,3 | 41 | 314,55 | -1,22 |
| 6 | 502 | -1,4 | 24 | 364,37 | -1,25 | 42 | 256,39 | -1,08 |
| 7 | 486,85 | -1,36 | 25 | 355,66 | -1,25 | 43 | 238,97 | -0,96 |
| 8 | 485,52 | -1,38 | 26 | 362,56 | -1,24 | 44 | 265,29 | -1,08 |
| 9 | 426,26 | -1,17 | 27 | 330,89 | -1,15 | 45 | 268,29 | -1,07 |
| 10 | 418,88 | -1,25 | 28 | 319,62 | -1,16 | 46 | 268,17 | -1,1 |
| 11 | 422,8 | -1,27 | 29 | 329,24 | -1,14 | 47 | 280,41 | -1,11 |
| 12 | 402,81 | -1,24 | 30 | 361,24 | -1,33 | 48 | 297,49 | -1,12 |
| 13 | 376,14 | -1,19 | 31 | 291,15 | -1,08 | 49 | 314,64 | -1,19 |
| 14 | 415,56 | -1,34 | 32 | 261 | -1,01 | 53 | 260,86 | -1,08 |

Окончание табл. 2

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 15 | 393,21 | -1,27 | 33 | 300,33 | -1,18 | 56 | 260,86 | -1,14 |
| 16 | 365,17 | -1,21 | 34 | 291,09 | -1,13 | 61 | 298,56 | -1,14 |
| 17 | 348,63 | -1,15 | 35 | 342,09 | -1,25 | 70 | 242,31 | -1,04 |
| 18 | 369,9 | -1,21 | 36 | 390,65 | -1,38 | 71 | 228,63 | -1,14 |
| 19 | 353,19 | -1,17 | 37 | 295,71 | -1,15 | Среднеариф. | 358,38 | -1,2 |

Итак, мы получили следующие константы: для закона Ципфа при $A = 358{,}38$ и $\alpha = -1{,}2$ близость аппроксимации данной функции практического материала составляет 91,9 %. Для формулы Ципфа-Мандельброта при $A = 800{,}17$, $\beta = -1{,}54$ и $B = 0{,}82$ точность описания экспериментального материала составляет 93,7 %.

Для этих коэффициентов были посчитаны основные статистические характеристики, такие, как дисперсия, стандартное отклонение и доверительный интервал.

Дисперсия – числовая характеристика случайной величины, характеризующая рассеяние ее возможных значений около математического ожидания [4, с. 94]. Она вычисляется по следующей формуле:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n},$$

где $\sigma^2$ – дисперсия, $x$ – выборочное среднее значение, а $n$ – размер выборки.

Для коэффициентов закона Ципфа $A\ \sigma^2 = 9522$ и для $\alpha\ \sigma^2 = 0{,}01$, для коэффициентов формулы Ципфа-Мандельброта для $A\ \sigma^2 = 0{,}13$, для $\beta\ \sigma^2 = 0{,}005$ и для $B\ \sigma^2 = 0{,}19$.

Среднеквадратическое (стандартное) отклонение $\sigma$ получается из дисперсии извлечением квадратного корня. Оно используется наряду с дисперсией для характеристики степени рассеивания случайной величины и оказывается в ряде случаев более удобным и естественным, в первую очередь из-за своей однородности в единицах измерения, что и признак [5, с. 105].

Для коэффициентов закона Ципфа $A\ \sigma = 97{,}58$ и для $\alpha\ \sigma = 0{,}1$, для коэффициентов формулы Ципфа-Мандельброта для $A\ \sigma = 0{,}35$, для $\beta\ \sigma = 0{,}07$ и для $B\ \sigma = 0{,}43$.

Во избежание ошибки в расчетах и для того, чтобы получить достоверное суждение об этих коэффициентах, был вычислен доверительный интервал, который рассчитывает предельное значение предельной ошибки выборки.

Таким образом, на основании проведенного исследования с уровнем надежности 95 % можно предположить, что для закона Ципфа $A$=358,38±26,5, $\alpha$=-1,2±0,3 и для формулы Ципфа-Мандельброта $A$=800,17±0,1, $\beta$=-1,54±0,02 и $B$=0,82±0,12.

Из приведенных выше результатов следует, что исследуемая в работе функциональная нагрузка между значениями многозначных слов подчиняется статистическим законам. Достигаемая точность аппроксимации не менее 94 % позволяет считать, что найденные теоретические распределения адекватны практическим. Таким образом, формула Ципфа-Мандельброта является наиболее подходящей для описания закономерностей функциональной нагрузки многозначных слов в частотно-семантическом словаре Лорджа и Торндайка.

ЛИТЕРАТУРА

*1. Lorge I.* A Semantic Count of English Words / I. Lorge, E. L. Thorndike. – Columbia, 1938. – 1177 p.

*2. Селезнев Г. Д.* Математическая модель динамики лексической системы / Г. Д. Селезнев // Проблемы лингвистической прогностики. – Воронеж, 2007. – Вып. 4. – С. 171–175.

*3. Терентьева И. А.* Распределение функциональной нагрузки между значениями многозначных слов / И. А. Терентьева, А. А. Кретов // Проблемы компьютерной лингвистики. – Воронеж, 2010. – Вып. 4. – С. 293–301.

4. *Макарова Н. В.* Статистика в Excel / Н. В. Макарова, В. Я. Трофимец. – М., 2006. – 368 с.

*5. Айвазян С. А.* Прикладная статистика и основы эконометрики / С. А. Айвазян. – М., 1998. – 1022 с.

*Воронежский государственный университет*

*Терентьева И. А., аспирантка кафедры теоретической и прикладной лингвистики*

*E-mail: irina1985_2004@mail.ru*

*Тел.: 8 (910)344-70-26*

*Voronezh State University*

*Terentyeva I. A., Post-graduate Student of the Department of Theoretical and Applied Linguistics*

*E-mail: irina1985_2004@mail.ru*

*Tel.: 8 (910) 344-70-26*