УДК 811.111'371

ОСОБЕННОСТИ РАСПРЕДЕЛЕНИЯ ФУНКЦИОНАЛЬНОЙ НАГРУЗКИ МЕЖДУ ЗНАЧЕНИЯМИ СЛОВ

И. А. Терентьева, Г. Д. Селезнев

Воронежский государственный университет

Поступила в редакцию 19.04.2012 г.

Аннотация: в статье исследуется распределение функциональной нагрузки многозначных слов, производится аппроксимация экспериментального материала различными законами распределений, а также статистическая обработка.

Anotation. The article explores the distribution of functional load of polysemantic words. The approximation of the experimental data based on different laws of distributions is carried out. The static handling of coefficients is carried out.

Ключевые слова: полисемия, функциональная нагрузка многозначных слов, экспоненциальное распределение, энтропия.

Keywords: polysemy, functional load of polysemantic words, exponential function, entropy.

ВВЕДЕНИЕ

Лексическая многозначность успешно изучалась различными способами и методами [1, 2, 3]. В данной работе рассматривается распределение функциональной нагрузки (ФН) между значениями многозначных слов. Под функциональной нагрузкой на значение многозначного слова понимается доля употреблений слова в данном значении от общего числа употреблений слова в выборке. Предметом исследования является распределение функциональной нагрузки между значениями многозначных слов и выявление закономерностей, управляющих функционированием значений таких слов.

1. ИССЛЕДОВАНИЕ ЛЕКСИЧЕСКОЙ МНОГОЗНАЧНОСТИ

Первым этапом работы было создание электронной базы частотно-семантического словаря "The semantic count of English Words" [4, 5], для этого из данного словаря И. Лорджа и Э. Торндайка были выбраны все многозначные слова. Значения слов в этом словаре распределены согласно Oxford English Dictionary [6], 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.

Поскольку в словаре допущены опечатки и неточности, а также округление данных до нуля, то к рассмотрению были приняты только статистически достоверные данные словаря Торндайка–Лорджа. Были выбраны, во-первых, только те слова, в которых сумма ФН на значения лежит в интервале $1000 \pm 10‰$; во-вторых, слова, в которых ФН ни на одно из значений не равна 0; в-третьих, только те слова, в которых ФН последовательно убывает, т.е. количество рангов частоты равно количеству значений.

Таким образом, у нас осталось 1418 слов, а общее количество групп сократилось до 9, но мы не рассматриваем однозначные слова, так как они имеют всего лишь одно значение. Результаты показаны в таблице 1.

2. АППРОКСИМАЦИЯ ДАННЫХ ЭКСПЕРИМЕНТА

Чтобы понять каким именно закономерностям подчиняется полученные нами данные, был применен метод аппроксимации. Он состоит в выборе подходящей математической функции, в наименьшей степени отличающейся от табличных данных. Выбор аппроксимирующей функции определяется природой исследуемого явления и во многом зависит от опыта и инту-

Таблица 1

Распределение функциональной нагрузки многозначных слов

Число значений слова n	Количество слов N_n	Сумма ‰	Номер значения k								
			1	2	3	4	5	6	7	8	9
			M_{nk} – функциональная нагрузка в ‰								
1	8047	1000	1000								
2	1418	1000	815	185							
3	1035	1000	676	238	86						
4	523	1000	608	242	109	41					
5	225	1001	550	240	125	63	23				
6	114	1000	504	235	130	74	42	15			
7	64	999	492	221	130	76	47	23	10		
8	29	999	412	245	147	86	52	35	15	7	
9	15	1000	372	215	154	98	65	42	28	17	9
Сумма	11470										

Здесь n – число значений слова; N_n – количество слов в словаре, имеющих n – значений; k – номер значения, $k = 1, 2, \ldots, n$; M_{nk} – функциональная нагрузка, т.е. частота встречаемости n-значного слова в k-ом значении, выраженная в промилле

иции исследователя. Если известен вид аппроксимирующей функции, то задача сводится к отысканию коэффициентов, входящих в функцию.

Для экспериментальных данных, представленных в Таблице 1, наилучшее качество аппроксимации распределения слов по числу значений N_n достигается экспоненциальной функцией

$$\tilde{N}_n = A\exp\left(-\alpha n\right), \qquad (1)$$

где i – номер значения слова, а A, α – подбираемые коэффициенты.

Подбор параметров в формулах производился с помощью процедуры EXCEL Сервис / «Поиск решения». С помощью этой процедуры производится автоматический поиск таких значений параметров, при которых величина «достоверности аппроксимации» R^2 принимает максимально возможное значение.

Качество аппроксимации экспериментальных данных выбранной функцией оценивается величиной достоверности аппроксимации R^2, вычисляемой по формуле, принятой в Microsoft Excel:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - f_i\right)^2}{\sum_{i=1}^{n} y_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}, \qquad (2)$$

где y_i – экспериментальные данные, а f_i – теоретические значения модели, n – количество значений слова.

Таким образом, если есть несколько подходящих вариантов типов аппроксимирующих функций, можно выбрать функцию с большей достоверностью аппроксимации, как можно более близкой к единице.

Наибольшая точность распределения функциональной нагрузки 100% наблюдается у двузначных слов, но это объясняется тем, что у данных слов всего две точки (два значения), поэтому для них легко подобрать подходящее распределение. Пример аппроксимации данных представлен на рисунке 1.

Наименьшая точность наблюдается у слов с шестью и восемью значениями. Средняя достоверность аппроксимации составляет 99,5%.

Рис. 1. Качество аппроксимации функциональной нагрузки девятизначных слов экспоненциальным распределением

Результаты аппроксимации, а также коэффициенты A, α представлены в таблице 2.

Таблица 2

Качество аппроксимации ФН экспоненциальным распределением

Кол-во значений	A	α	R2
1		∞	
2	3590	1,483	1
3	1887	1,031	0,9999
4	1477	0,889	0,9986
5	1194	0,769	0,994
6	972	0,666	0,9895
7	861	0,615	0,9927
8	792	0,563	0,9897
9	612	0,446	0,996

Если рассмотреть изменение коэффициент α(см. рис. 2), то очевидно, что он меняется в соответствии с логарифмическим распределением, которое описывается следующей формулой: $N_i = -0,465\ln(i) + 1,4243$,где i – номер значения слова. Достоверность аппроксимации составляет 98,4%. Это говорит о том, что коэффициентом α управляют определенные закономерности, а значит, мы можем предсказывать его значения для каждой группы многозначных слов.

ЗАКЛЮЧЕНИЕ

В работах [7, 8 и 9] были предложены варианты объяснения природы экспоненциальности распределения (1). В [8] была предложена эвристическая модель на основе модели однородной цепи Маркова. В [10] было высказано предположение о том, что «распределение количества слов по числу их значений в словарях романских языков является экспоненциальным распределением, аналогичным распределению Больцмана в статистической термодинамике». По результатам работы [9] можно сделать вывод о том, что данное предположение справедливо и для славянских языков, а результат настоящей работы позволяет предположить, что это справедливо вообще для всех, по крайней мере, европейских языков.

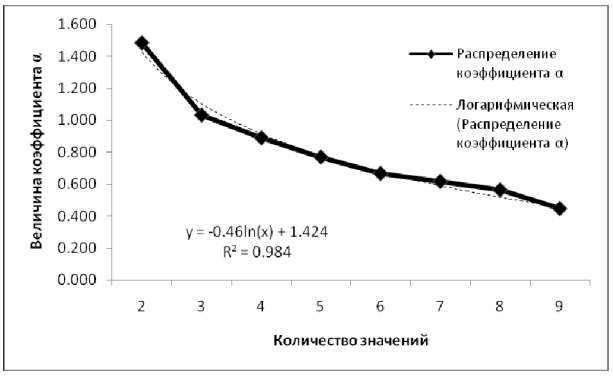

Рис. 2. Качество аппроксимации коэффициента α

В работах [7, 10] была предложена гипотеза о существовании в статистической лингвистике законов, аналогичных законам статистической физики: «Закона сохранения общего количества значений всех слов некоторой замкнутой языковой системы» – в данном случае словаря и «Закона возрастания энтропии распределения значений по группам однозначных, двузначных, трехзначных и т.д. слов». Экспоненциальное распределение может быть прямым следствием этих двух законов. Эта физическая аналогия позволила ввести в статистическую лингвистику новые параметры: величину обратную показателю экспоненты в формуле (1) $T = 1 / \alpha$, который был назван «семантической температурой», и величину H – «семантическую энтропию распределения слов по числу значений».

В таблице 3 представлены показатели «температуры» и энтропии, которые характеризуют многозначные слова частотно-семантического словаря.

Таблица 3

«Температура» и энтропия многозначных слов

Кол-во значений	T = 1/Альфа	Энтропия распред.
1	0	0
2	0,6744	0,478891
3	0,970026	0,817334
4	1,125239	1,018433
5	1,301067	1,19199
6	1,501953	1,276695
7	1,624959	1,420609
8	1,775884	1,572139
9	2,242152	1,736663

Мы видим, что изменение «температуры» и энтропии в зависимости от количества значений слова подчиняется определенным закономерностям (см. рис. 3, 4).

Как видно на рисунках, «температура» и энтропия изменяются согласно логарифмическому распределению, которое описывается следующей формулой:

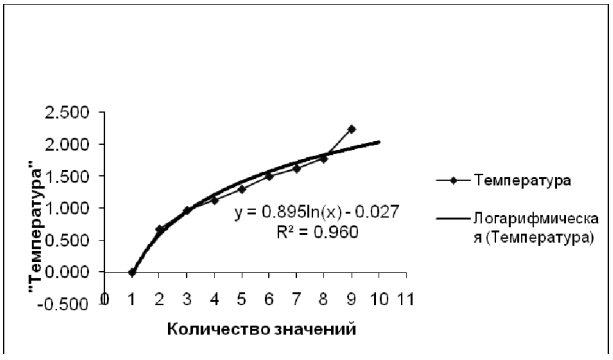

Рис. 3. "Температура" распределения слов по частоте встречаемости в зависимости от числа значений слова

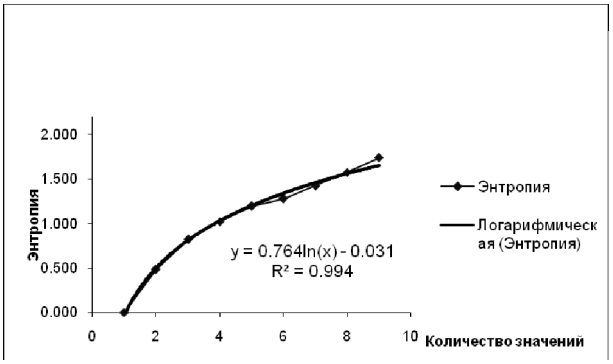

Рис. 4. Энтропия распределения слов по частоте встречаемости в зависимости от числа значений слова

Для «температуры»

$$N_i = 0,8955 \ln(i) - 0,0276,$$

где i – номер значения слова. Достоверность аппроксимации равна 96%.

Для энтропии

$$N_i = 0,7649 \ln(i) - 0,0311,$$

где i – номер значения слова. Достоверность аппроксимации составляет 99,4%.

Средний показатель «температуры» для английского языка, по данным словаря-источника, составляет 1,4; энтропии – 1,19.

Из приведенных выше результатов следует, что исследуемая в работе функциональная нагрузка на значения многозначных слов подчиняется определенным математическим законам, а именно, экспоненциальному распределению. Достигаемая точность аппроксимации не менее 99,5% позволяет считать, что найденные теоретические распределения адекватны практическим.

Не исключено, что полученные показатели «температуры» и энтропии связаны с определенными свойствами данного языка.

СПИСОК ЛИТЕРАТУРЫ

1. *Апресян Ю.Д.* Исследования по семантике и лексикографии. Т. 1: Парадигматика / Ю. Д. Апресян – М., 2009 – 586 с.

2. *Зализняк А. А.* Многозначность в языке и способы ее представления / Анна А. Зализняк. – М., 2006. – 672 с.

3. *Тулдава Ю.А.* О некоторых квантитативно-системных характеристиках полисемии / Ю. А. Тулдава // Уч. зап. Тартуского у-та. – Тарту: Изд-во Тартуского ун-та. – 1979. – Вып. 502. – С. 107–141.

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

5. *Lorge I.* The semantic count of 570 Commonest English Words / I. Lorge. – New York, 1949.

6. Oxford English Dictionary on CD-ROM. Version 3.1. (2009). Oxford: Oxford University Press.

7. *Селезнев Г.Д.* Природа экспоненциального распределения слов по числу значений / Г. Д. Селезнев // Вестник Воронежского государственного университета. Сер. Лингвистика и межкультурная коммуникация. № 2. 2007. – Воронеж. – С. 42–46.

8. *Селезнев Г.Д.* Природа экспоненциального распределения слов по числу значений / Г. Д. Селезнев // Проблемы компьютерной лингвистики. Сб. научн. трудов. – Вып. 2. – Воронеж, 2005. С. 169–173.

9. *Селезнев Г.Д.* Как распределяются слова по числу значений / Г. Д. Селезнев // Проблемы компьютерной лингвистики. Сб. научн. трудов. – Вып. 3. – Воронеж, 2008. – С. 220–225.

10. *Селезнев Г.Д.* Природа распределения длин слов в словарях романских языков / Г. Д. Селезнев // Проблемы компьютерной лингвистики. Сб. научн. трудов. – Вып. 5. – Воронеж, 2011. – С. 329–334.

11. *Ландау Л.Д.* Статистическая физика. / Л. Д. Ландау, М. Е. Лившиц – М., Наука, 1964. – 568 с.

Терентьева Ирина Анатольевна – преподаватель кафедры теоретической и прикладной лингвистики, Воронежский государственный университет. Тел.: 2204149. E-mail: irina1985_2004@mail.ru

Terentyeva I. A. – Lecturer of the Theoretical and Applied Linguistics Department, Voronezh State University. Tel.: 2204149. E-mail: irina1985_2004@mail.ru

Селезнев Геннадий Данилович – доц. кафедры теоретической и прикладной лингвистики, Воронежский государственный университет. Тел.: 2204149. E-mail: selforg@newmail.ru

Seleznev G. D. – Associate Professor, the dept. of the Theoretical and Applied Lingvistics, Voronezh State University. Tel.: 2204149. E-mail: selforg@newmail.ru