В.В. Кромер

Новосибирский государственный педагогический университет
630126, Новосибирск-126, Вилюйская 28
e-mail: applied@nspu.nsu.ru

БЕСПАРАМЕТРИЧЕСКАЯ МОДЕЛЬ РАНГОВЫХ ПОЛИСЕМИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Известно, что семантическая нагрузка распределена по разным знаковым единицам неравномерно. Семантическую нагрузку слова принято определять по степени его полисемичности, а за меру полисемичности принимается количество его значений, зафиксированных в одноязычных толковых (авторских, текстовых) словарях.

Ранжирование лексики толкового словаря по убыванию степени полисемичности дает ранговое полисемическое распределение. Такие зависимости можно отображать графически в системе координат "ранг – степень полисемичности", а также описывать разной степени сложности эмпирическими формулами. Эмпирические формулы (особенно многопараметрические) позволяют достаточно точно описать наличное распределение, вместе с тем параметры формул не всегда имеют четкий лингвистический смысл. Анализ динамики изменения параметров эмпирических формул для гаммы толковых словарей (т.е. для ряда словарей последовательно возрастающего объема, отражающих лексику выбранного языка) наводит на мысль, что между параметрами существует функциональная зависимость, а их количество может быть сокращено без ущерба для точности описания. В предельном случае возможно "жесткое" беспараметрическое описание зависимости на основе одной или нескольких непосредственных характеристик самого распределения.

Примером беспараметрической модели является, например, модель Ю.К. Орлова, основанная на формуле Мандельброта для зависимости "ранг – частота" [1]. Целью данной работы является создание модели полисемического распределения с наименьшим числом подгоночных параметров и максимальной лингвистической "прозрачностью". В основе модели лежат некоторые теоретические положения и предположения, следствия из которых проверяются на экспериментальном материале.

Полисемические распределения невозможно рассматривать в отрыве от ранговых распределений частот слов. Общим (по определению) для 2 рассматриваемых типов распределений является то, что с ростом ранга

слова убывает как степень полисемичности слова, так и его частота в достаточно представительной выборке. Однако в общем случае ранги отдельного слова в полисемическом и частотном ранговых распределениях не совпадают, и связано это не только со ступенчатым характером хвостовых частей распределений (т.е. с неопределенностью ранга в пределах ступеньки), но и со статистическим характером самой зависимости "употребительность – полисемия". Статистический характер связи свидетельствует о том, что как полисемия, так и частота определяются некоторым общим набором факторов различной степени влияния, каждый из которых, возможно, действует строго детерминированно в отношении частоты или полисемии, но суммарная реакция не аддитивна относительно отдельных воздействий.

Из сказанного следует, что слову с определенным рангом в частотном списке свойственно некоторое вероятностное полисемическое распределение, т.е. оно может иметь с некоторой вероятностью 1, 2, 3 ... и т.д. значений. При анализе же зависимости "употребительность – полисемия" мы вынуждены сглаживать экспериментальные количества значений, при этом степень сглаживания данных должна определяться компромиссом между необходимостью выявления искомой зависимости, элиминирования флюктуаций и отсутствием значимых систематических смещений оценки.

Приводим положенные в основу при разработке модели теоретические предпосылки:

1. Каждый язык (подъязык, идиолект) базируется на определенном корпусе текстов (выборке) и вследствие этого ограничен.

2. Отдельный толковый словарь отражает состав конкретного языка. Корпус текстов, положенный в основу данного языка, предлагается назвать конститутивным для данного толкового языка.

3. Объективно каждое употребление слова в конститутивной выборке (корпусе текстов) дает новое его значение, т.е. количество значений слова равно общему количеству зафиксированных его употреблений, однако субъективно носитель языка не различает единичные значения слова, связанные с отдельным словоупотреблением.

4. Значения слов группируются носителем языка в группы значений, а количество групп (традиционно называемых количеством словарных значений) определяется в соответствии с известным основным

психофизическим законом Вебера-Фехнера.

Закон Вебера-Фехнера, справедливый для простейших видов ощущений (зрения, слуха, осязания), может быть распространен и на более сложные виды психической деятельности. В нашем случае множеством внешних сигналов является множество всех предъявлений конкретного слова в процессе разворачивания конститутивной выборки. Диапазон потенциально воспринимаемых сигналов (область адекватного отражения) естественным образом ограничивается снизу предъявлением одноразовых слов и сверху – абсолютной частотой самого частого слова выборки. Шаг стимульного ряда также определяется естественным образом в одно предъявление слова.

Основной психофизический закон Вебера-Фехнера устанавливает соответствие между внешним сигналом-стимулом S и ответной реакцией (ощущением) R [2, с. 139]:

$$R = a \ln S + b \ , \qquad (1)$$

где a и b – некоторые постоянные, определяемые чувствительностью и порогом восприятия приемника сигналов.

Для интересующей нас модальности (последовательное предъявление одного и того же слова в процессе разворачивания выборки) естественны следующие соотношения в области малых сигналов: $R(0) = 0$; $R(1) = 1$. Чтобы увязать эти соотношения с выражением (1), справедливым в области средних сигналов по определению, достаточно положить $a = 1$; $b = C = 0{,}5772...$ (постоянная Эйлера), и зависимость реакции в функции стимула (слова с частотой F) выражается через пси-функцию [3, с. 175]:

$$R = \sum_{k=1}^{F} \frac{1}{k} = \psi(F+1) + C \, . \qquad (2)$$

Выражение (2) асимптотически стремится к $(\ln F + C)$ при больших F, и оно также справедливо и при $F = 0$, поскольку $\psi(1) = -C$.

Сделаем предположение, что конститутивной выборке свойственно ципфовское распределение частот слов:

$$F = \frac{K}{i^{\gamma}} , \qquad (3)$$

где i – ранг слова, K и γ – ципфовские параметры распределения. В соответствии с принятой моделью математическое ожидание количества словарных значений составляет:

$$m_i = \psi(F+1) + C\,. \tag{4}$$

Ранее М.В. Араповым [4, с. 140] предложена следующая кусочно-линейно заданная зависимость (обозначения наши):

$$m = \{D - a\ln i : i \le i_0\ ; 1 : i > i_0\}, \tag{5}$$

где i_0 – граница, начиная с которой, "разрешающая способность" словаря, по мысли автора, недостаточна для различения и противопоставления различных значений одного слова.

Мы, в отличие от автора формулы (5), полагаем, что подобной границы не существует, поскольку даже в группе самых редких слов выявляются слова с 2, 3 и т.д. значениями. Потребуем равенства частоты самого редкого слова конститутивной выборки (и соответственно количества его значений в толковом словаре, поскольку $\psi(2) + C = 1$), единице.

Обозначим за L количество слов в рассматриваемом языке (что соответствует количеству слов в соответствующем толковом словаре) и введем условие нормировки (равенство суммы математических ожиданий количеств значений по всем рангам в соответствии с выражением (4) и общего количества значений всех слов толкового словаря). Оба выдвинутых требования можно записать в виде системы:

$$\begin{cases} \dfrac{K}{L^{\gamma}} = 1 \\ \displaystyle\sum_{i=1}^{L}\left[\psi\left(\frac{K}{i^{\gamma}} + 1\right) + C\right] = M \end{cases}, \tag{6}$$

где $M = \sum_{i=1}^{L} m(i)$ – суммарное количество значений всех слов словаря. Решая систему (6), находим значения K и γ. Поскольку $m_i = \psi(F+1) + C \approx \ln F + C = \ln\frac{K}{i^{\gamma}} + C = \ln K + C - \gamma\ln i$, формулу (5)

можно представить как упрощенный вариант формулы (4), с заменой пси-функции от (F+1) на логарифм от F. В области больших F (малых рангов) обе зависимости практически сливаются. В области высоких рангов значения $\psi(F+1)$ существенно больше значения $\ln F$, т.е. зависимость m_i по нашей модели лежит выше аналогичной зависимости по М.В. Арапову. На рис. 1 приведены сглаженные фактические и 2 теоретические подобные зависимости для толкового словаря русского языка [5]. Ранги слов определялись по частотному словарю русского языка [6].

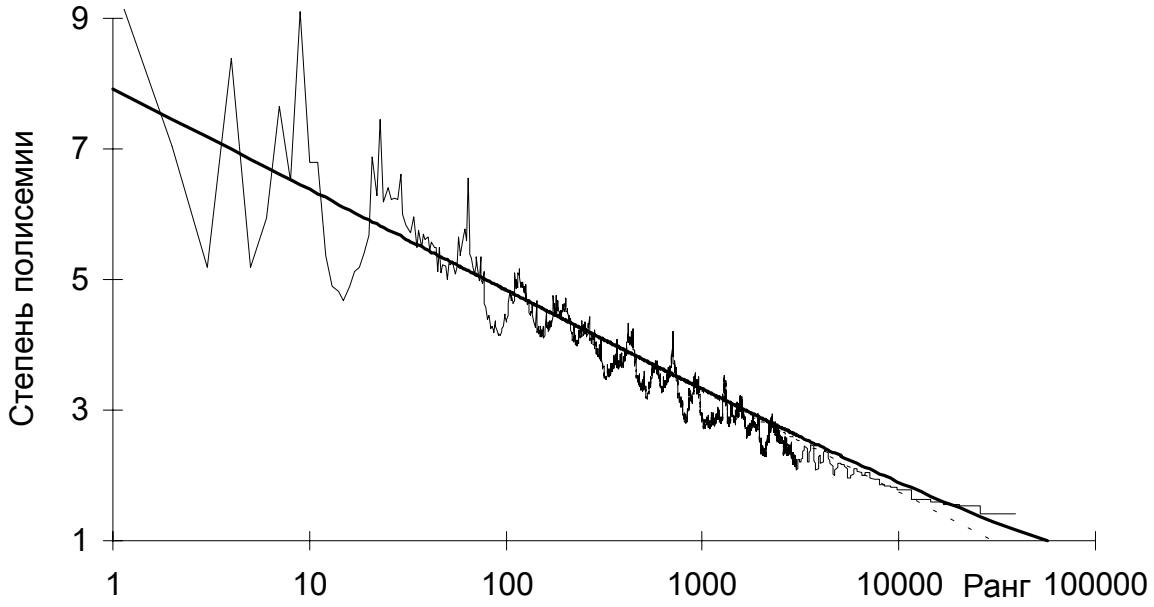

Рис. 1.

Тонкой линией на рис. 1 отображены сглаженные эмпирические зависимости $m(i)$, толстой линией – зависимость m_i по нашей модели с определенными из решения системы (6) параметрами K и γ. Пунктирной линией отражена зависимость $m(i)$ для тех же K и γ при замене пси-функции на логарифм. Масштаб по оси рангов выбран логарифмическим, по оси значений линейным, поскольку степень полисемичности слова в рассматриваемой модели сама по себе является логарифмической величиной, как и всякая величина, являющаяся следствием проявления закона Вебера-Фехнера.

Рассмотрим экспериментальную зависимость вероятности слова с рангом i иметь 1, 2, 3 ... и т.д. значений. Такая зависимость может быть прослежена на равночастотных массивах слов. В качестве примера на рис. 2 непрерывной линией приведено ранговое полисемическое распределение для массива слов с частотой 6 по частотному словарю [6].

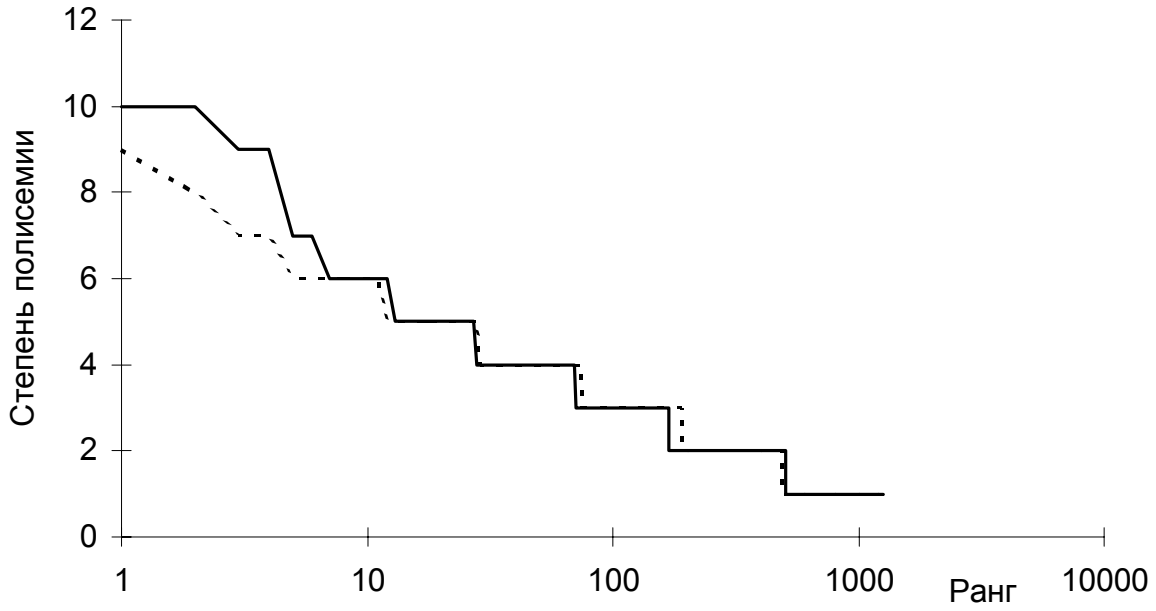

Рис. 2.

Всего в данном массиве 1279 слов, и они ранжированы по степени их полисемии согласно толковому словарю [5]. Бросается в глаза равная длина ступенек (в логарифмическом масштабе) на рангах выше 10 при высоте ступенек равной единице. Аналогичный вид имеют распределения и для других частот. Общим является линейный при выбранных осях (если отвлечься от ступенчатости) вид графика. Распределения подобного типа следует относить к ципфовским, несмотря на линейный масштаб по оси ординат, поскольку по оси ординат откладывается логарифмическая (в данной модели) величина – количество значений.

Возможно, что ципфовский характер полисемической ранговой зависимости для равночастотных слов носит системный характер. Так, в [7] показано, что распределение частот слов для слов одинаковой длины в ряде языков также является ципфовским.

Обозначим за n_k количество слов с k значениями из общего количества N_F слов с частотой F. Обозначим за p_1 вероятность слова принимать 1 значение (т.е. долю однозначных слов в массиве):

$$p_1 = \frac{n_1}{N_F}. \qquad (7)$$

Приняв гипотезу о ципфовском характере полисемического рангового

распределения в пределах равночастотной группы слов, можно определить вероятность слова принимать k значений:

$$p_k = p_1(1-p_1)^{k-1}. \qquad (8)$$

Математическое ожидание количества значений у слова с частотой F составляет

$$m_F = \sum_{k=1}^{\infty} k p_k = \sum_{k=1}^{\infty} k p_1 (1-p_1)^{k-1} = \frac{1}{p_1}\ , \qquad (9)$$

и тогда $p_1 = \frac{1}{m_F}$, а

$$p_k = p_1(1-p_1)^{k-1} = \frac{1}{m_F}\left(1-\frac{1}{m_F}\right)^{k-1} = \frac{(m_F-1)^{k-1}}{m_F^k}\ . \qquad (10)$$

Рассчитанная по выражению (10) теоретическая зависимость отражена на рис. 2 пунктирной линией. Во всем диапазоне рангов от 1 до L количество k-значных слов составит

$$N_k = \sum_{i=1}^{L} p_k = \sum_{i=1}^{L} \frac{(m_F-1)^{k-1}}{m_F^k}\ , \qquad (11)$$

где m_F определяется по формуле (4), а F – по формуле (3) с параметрами K и γ, определенными из системы (6). Для проверки предлагаемой модели нами использованы приведенные в [8] данные по 5 толковым и 1 авторскому словарям. В таблицу 1 сведены исходные данные и вычисленные по системе (6) значения K и γ. Количество слов в словаре L определялось как сумма количеств слов во всех группах слов с одинаковой степенью полисемии.

Т а б л и ц а 1

Параметр	СЯП	СО	МАС	ССРЛЯ	Hornby	Shorter
L	21201	57003	82159	120481	44372	79801
M	28820	78457	123486	204490	60331	160551
K	126	257	1724	30760	182	111800

γ	0,486	0,507	0,659	0,883	0,486	1,234
L^*				157417	28129	93006
M^*				241426	44088	173756
K^*				4073	1844	220300
γ^*				0,695	0,734	1,075

СЯП – “Словарь языка Пушкина” (1956–1961); СО – “Словарь русского языка” С.И. Ожегова (1972 – 9-е изд.); МАС – “Словарь русского языка” в 4-х тт. под ред. А.П. Евгеньевой (1957–1961); ССРЛЯ – “Словарь современного русского литературного языка” в 17-ти тт. (1948–1965); Hornby – A.S. Hornby, “Oxford Advanced Learner’s Dictionary of Current English” (1982); Shorter – “Shorter Oxford English Dictionary” (1962).

Вычисленные по формуле (11) теоретические значения N_k сравниваются с приведенными в [8, с. 146] экспериментальными N_e. Распределение N_k принимаем (в первом приближении) нормальным, с дисперсией, равной N_k , и стандартным отклонением $\sqrt{N_k}$. Вычисленные по формуле

$$s_\sigma = \frac{N_e - N_k}{\sqrt{N_k}} \qquad (12)$$

значения отклонения N_e от N_k в сигма-единицах в зависимости от k приведены в таблице 2. Значения s_σ в таблице 2 приведены для значений k, соответствующих $N_k > 5$.

Таблица 2

k	СЯП	СО	МАС	ССРЛЯ	Hornby	Shorter	ССРЛЯ*	Hornby*	Shorter*
1	- 0,37	0,33	-0,02	-17,32	8,57	-11,70	0,00	0,00	0,00
2	1,67	-0,14	-0,77	34,03	-21,73	13,14	0,90	-0,37	-4,17
3	-0,38	-0,64	0,62	10,63	-1,80	14,29	0,65	3,17	6,65
4	-0,73	-1,12	1,48	0,02	-0,94	5,45	-1,11	-2,30	2,61
5	0,11	-0,75	0,74	-4,46	2,58	0,61	-1,97	-1,21	0,12
6	-2,16	1,04	-1,15	-4,99	2,10	-1,18	-0,98	-2,33	-0,47
7	0,49	0,28	-0,63	-6,01	5,54	-1,69	-1,54	0,38	-0,33
8	-1,04	-2,59	1,30	-6,24	3,42	-3,47	-1,78	-1,15	-1,82
9	-1,11	2,49	-0,99	-5,22	6,47	-4,61	-0,86	1,28	-2,87
10	-0,75	1,82	1,30	-4,23	5,05	-2,93	-0,03	0,48	-1,06
11		0,20	-0,68	-2,39	2,37	-5,48	1,81	-1,07	-3,82
12		1,60	-0,87	-2,03		-1,96	1,90	0,46	-0,09
13		-1,89	-0,72	-4,32		-3,51	-1,37	1,62	-1,87
14			-0,80	-2,17		-0,55	0,99	1,14	1,29

15			-0,27	-0,75		-1,65	2,51	0,90	-0,03
16			1,51	-1,40		-1,43	1,36		0,12
17			-0,91	-3,24		-1,08	-1,44		0,42
18				-1,19		-2,12	1,09		-0,87
19				0,29		-1,06	2,94		0,24
20				1,27		-0,48			0,82
21				0,44		-1,72			-0,71
22				-0,84		-0,78			0,31
23						-0,69			0,33
24						-0,75			0,19
25						-1,32			-0,56
26						-0,92			

В соответствии с общепринятой практикой случайными считаются отклонения, не превышающие ±1,96 сигма-единиц. Из данных таблицы 2 видно, что предлагаемой модели соответствуют словари СЯП, СО и МАС. Что же касается словарей ССРЛЯ, Hornby и Shorter, для них отклонения экспериментальных значений от теоретических слишком велики, чтобы считать их случайными.

Вместе с тем, зона однозначных слов представляется очень неопределенной для учета и для представления в полисемических распределениях [8, 147]. Как отмечается далее в этой же работе, произвольное изъятие части состава словаря сказывается на его системно-количественных характеристиках, и зона однозначных слов выпадает из общей тенденции соотношения объемов групп слов разной полисемии. По оценке из [8, с. 148], в словаре ССРЛЯ лексикографами упущено около 40-50 тыс. однозначных слов. Возможно также и переполнение той же самой периферийной области однозначных слов за счет привлечения слов, относящихся к лексической сфере за пределами словарного типа.

Выскажем предположение, что в словарях ССРЛЯ, Hornby и Shorter в зоне однозначных слов наблюдается дефицит (или профицит) слов (и соответственно значений). Перепишем систему (6), добавив в нее еще одно неизвестное L^* и еще одно уравнение:

$$\begin{cases} \dfrac{K^*}{(L^*)^{\gamma^*}} = 1 \\ \sum\limits_{i=1}^{L^*}\left[\psi\left(\dfrac{K^*}{i^{\gamma^*}}+1\right)+C\right] = M + L^* - L \\ \sum\limits_{i=1}^{L^*}\dfrac{1}{\psi\left(\dfrac{K^*}{i^{\gamma^*}}+1\right)+C} = N_1 + L^* - L \end{cases}, \qquad (13)$$

где N_1 – экспериментальное количество однозначных слов в словаре. Последнее уравнение системы (13) требует равенства теоретического значения количества однозначных слов эмпирическому. Таким образом, беспараметрическая модель переводится в однопараметрический режим с подбираемым параметром L^*, где L^* – количество слов, а K^* и γ^* – параметры модифицированного словаря (словаря с измененным количеством слов в зоне однозначных слов для сохранения тенденции соотношения объемов групп разной полисемии). Найденные из системы (13) значения L^*, K^* и γ^* для словарей ССРЛЯ, Hornby и Shorter приведены в таблице 1. За $M^* = M + L^* - L$ обозначено суммарное количество значений слов для модифицированного словаря.

Пересчитанные по формулам (11) и (12) значения отклонения N_e от N_k в сигма-единицах для 3 модифицированных словарей приведены в таблице 2 в колонках ССРЛЯ*, Hornby* и Shorter*. Как видно, для модифицированных словарей соответствие лучше и отклонения экспериментальных значений от теоретических находятся в пределах случайных отклонений.

Таким образом, предлагаемая беспараметрическая модель полисемических ранговых распределений удовлетворительно описывает полисемические ранговые распределения уравновешенных (по соотношению центральных и периферийных зон) толковых словарей. Идейной платформой модели является высказанное в [8, с. 140] положение о том, что общим источником вариативности полисемии, длины и частоты знаков является размер знакового набора, и он же определяет распределение функциональной нагрузки по всем знакам набора при заданном социальной практикой наборе смыслов, обслуживаемом данным языком.

Для неуравновешенных по соотношению центральных и периферийных

зон толковых словарей перевод модели в однопараметрический режим позволяет определить дефицит или профицит однозначных слов в словаре, а тем самым и общее количество слов в словаре при условии уравновешивания зоны однозначных слов.

Литература

1. Орлов Ю.К. Статистическое моделирование речевых потоков // Вопросы кибернетики: Статистика речи и автоматический анализ текста. – М.; Л., 1978. – Вып. 41. – С. 66–99.
2. Забродин Ю.М., Лебедев А.Н. Психофизиология и психофизика. – М.: Наука, 1977.
3. Кромер В.В. Некоторые особенности горизонтального распределения слов русского языка по жанровым группам текстов // Развитие личности в системе непрерывного образования: Тез. докл. II Междунар. конф. – Новосибирск, 1997. – Ч. 3. – С. 174–176.
4. Арапов М.В. Квантитативная лингвистика. – М.: Наука, 1988.
5. Ожегов С.И. Словарь русского языка. – М.: Рус. яз., 1985.
6. Частотный словарь русского языка / Под ред. Л.Н. Засориной – М.: Рус. яз., 1977.
7. Leopold E. Frequency Spectra within Word-Length Classes // J. of Quantitative Linguistics. – 1998. – Vol. 5. – № 3. – P. 224–231.
8. Поликарпов А.А. Полисемия: системно-квантитативные аспекты // Учен. зап. Тартус. ун-та. – Тарту, 1987. – Вып. 774. – С. 135–154.