УЧЕНЫЕ ЗАПИСКИ ПЕТРОЗАВОДСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Ноябрь, № 7. Т. 1 **Филология** **2013**

УДК 81’32, 801.82, 811.161.1’32, 811.161.1’373.46, 811.16

ЛЮБОВЬ ЮРЬЕВНА КОВРИГИНА
аспирант кафедры математической лингвистики филологического факультета, Санкт-Петербургский государственный университет (Санкт-Петербург, Российская Федерация)
lkovriguina@gmail.com

КОНЦЕПТУАЛЬНЫЕ РАЗЛИЧИЯ ПОДХОДОВ К ОПИСАНИЮ СТАТИСТИЧЕСКОЙ СТРУКТУРЫ ТЕКСТОВ (на примере «Сказания о Мамаевом побоище»)

Сопоставляются три метода моделирования статистической структуры текста как поликомпонентного объекта по одному из параметров на материале редакций «Сказания о Мамаевом побоище». Методы нацелены на выявление тематических и нетематических лексических единиц в эмпирическом распределении лексики текста. Производится апробация этих методов на одном материале и выявляется приоритетный параметр – точка *h* Хирша – Попеску, которая позволяет «отсеять» большинство нетематических единиц для конкретного текста.

Ключевые слова: вариативный текст, негауссовость, *H*-распределение, статистическая структура текста, ядерные и периферические элементы, неоднородность генеральной совокупности, пойнтер-точка *R* Б. И. Кудрина, точка *h* Дж. Хирша – И.-И. Попеску

В качестве предмета лексикостатистического исследования текст предстает как определенный набор лексических единиц (уникальных или повторяющихся с той или иной частотой). Если перенумеровать элементы словаря $V = \{x_1, x_2, ..., x_N\}$ так, чтобы частота F слова x_i была невозрастающей функцией его номера $F(x_1) > F(x_2) > ... > F(x_N)$, то ранговым распределением называется функция $\Phi(n) = F(x_N)$, которая ставит в соответствие номеру или рангу $n(x)$ слова $x \in V$ частоту $F(x)$ этого слова. В обсуждаемом далее контексте рассматриваются такие ранговые распределения, график которых приближается к гиперболе, в силу чего такие распределения обозначаются как *H*-распределения[1].

К настоящему времени накопилось довольно большое число разных подходов к описанию ранговых распределений и не прояснено, как эти подходы соотносить между собой, поскольку каждый из них разрабатывался внутри предметно-специфичной научной парадигмы и под конкретный материал.

Поэтому представляется практически важным исследование и сопоставление разных моделей и получаемых с их помощью результатов на одном материале. Предметом статьи выбрана задача соотнесения трех характеристик моделей, принадлежащих разным авторам (Б. И. Кудрину [4]; Г. Я. Мартыненко [6], [7], [8]; И. Попеску, Г. Альтманну и Я. Машутеку [15]).

Перечисленные ниже параметры моделей и характеристики эмпирических данных принимаются всеми исследователями: 1. Резкая неравночисленность разных классов распределения. 2. Большое количество одноэлементных классов. 3. Наличие небольшого количества высокочастотных классов. 4. Большой разрыв численности классов с рангами 1 и 2, 2 и 3, 3 и 4 и немного далее с уменьшением разницы по мере движения по этому ряду (в случае моделирования с помощью ранговых распределений). 5. Отсутствие простого описания соотношения численности классов распределения. 6. Более или менее симметричный относительно биссектрисы первого квадранта график рангового распределения. 7. Более или менее хорошая аппроксимация гиперболой графика рангового распределения. 8. Систематическое отклонение этих распределений от гиперболы в зоне средних частот.

При таких сходных допущениях среди исследователей *H*-распределений отсутствует единое мнение об их гауссовости/негауссовости и выполнения для них центральной предельной теоремы. Спорным моментом является и однородность/неоднородность генеральной совокупности, описываемой *H*-распределениями.

Большинство исследователей исходит из того, что они имеют дело с единой генеральной совокупностью (к их числу относится и Б. И. Кудрин). Как указывает С. В. Чебанов, «всячески обыгрывается то обстоятельство, что это такая хитрая генеральная совокупность, что она содержит принципиально непохожие друг на друга компоненты» [11; 75].

Иная точка зрения заключается в том, что *H*-распределение описывает смесь как минимум двух генеральных совокупностей, каждая из которых внутри однородна. Этой точки зрения придерживается Г. Я. Мартыненко, который считает более перспективным направление, учитывающее неоднородность статусных распределений [7; 67]. Убежденным сторонником неоднородности распределений, до этого аппроксимируемых

(в гуманитарных науках) преимущественно «гиперболой» Ципфа, был и Густав Хердан [14; 77–86]. Опишем подробнее каждый из подходов.

ЦЕНОЛОГИЧЕСКИЙ ПОДХОД Б. И. КУДРИНА

Б. И. Кудриным начиная с середины 1970-х годов разрабатывается претендующий на всеобщность ценологический подход [5][2].

По Б. И. Кудрину, аналитическим выражением для *H*-распределения является функция вида $\Omega(x) = W_1/x^{1+\alpha}$, где x – численность класса ($x = 1, 2, \ldots n$, где n – численность самого высокочастотного класса); W_1 – количество классов с численностью 1 (*hapax legomena*); $\alpha > 0$ – параметр (см. подробное описание, напр., в [3; 388]).

Функция $\Omega(x)$ не вполне хорошо описывает эмпирическое распределение, поэтому Кудрин вводит понятие особой точки, точки перегиба – пойнтер-точки *R,* которая фиксируется на эмпирическом распределении. Гипербола делится точкой *R* на две ветви: слева $x = 1, 2, \ldots, R$ – неоднородные группы, где каждая образована множеством классов; справа $x = R + 1, R + 2, \ldots, F_{max}$ – однородные группы (F_{max} – эмпирическая частота слова, встретившегося в тексте с максимальной частотой) [4] (см. столбец «*n*» табл. 1, где пойнтер-точка выделена полужирным шрифтом).

Б. И. Кудриным были получены интересные результаты при изучении распределения персонажей в романе М. А. Булгакова «Мастер и Маргарита»: вокруг пойнтер-точки сгруппировались персонажи, определяющие отличие романа Булгакова от «Фауста» Гёте: Левий Матвей, Босой, Варенуха, Римский, Стёпа Лиходеев [4].

ЗОНЫ КОНЦЕНТРАЦИИ И РАССЕЯНИЯ ПО Г. Я. МАРТЫНЕНКО

Г. Я. Мартыненко подвергает радикальному сомнению однородность генеральной совокупности и возможность аппроксимации ее распределения с помощью функции Ципфа – Парето [6; 157], справедливо указывая, что, «если по данным наблюдения построить убывающее ранговое распределение, то его характерные особенности могут ускользнуть из поля зрения исследователя ввиду чрезмерной растянутости графика вдоль оси рангов. Как правило, такой график отождествляется с крайне асимметричной *J*-образной кривой. Высокий пик таких распределений и очень растянутый хвост затушевывают некоторые ненормальности в поведении кривой, которые чаще всего списываются на счет ошибок наблюдения» [6; 140]. Под «ненормальностями» подразумевается характерный бугорок на кривой распределения в области средних частот, который Г. Я. Мартыненко квалифицирует как результат наложения одного распределения на другое (графики распределения элементов смешанной совокупности характеризуются более сложным рельефом), причем делает это уже в 1978 году [7; 67–69].

Чтобы установить неоднородность совокупности, Г. Я. Мартыненко предлагает следующие критерии: 1) величину коэффициента вариации: чем больше этот коэффициент, тем больше шансов, что совокупность неоднородна; 2) внешний вид графика эмпирического распределения: если кривая распределения многовершинна, то есть веские основания полагать, что исходная совокупность состоит из нескольких качественно однородных фрагментов [6; 137].

Для разделения совокупности на зоны можно воспользоваться методикой вычисления прироста скользящего коэффициента вариации (СКВ). Процедура расчетов с примерами подробно описана в той же работе [6; 150–153]. По данным автора статьи, на текстах достаточного объема на графике прироста СКВ явно прослеживается следующая тенденция: сначала функция монотонно убывает, потом убывание чередуется с возрастанием и, наконец, монотонно возрастает (при построении графика по оси абсцисс откладывается $ln(n)$, по оси ординат – $ln(dV)$, значения коэффициента вариации и его прироста см. в столбцах *V* и *dV* табл. 1).

Г. Я. Мартыненко выделяет в *H*-распределениях ядерную и периферическую зоны и переходную зону. В ядерные элементы попадает служебная лексика и другие единицы со стертой семантикой, на периферии – низкочастотная лексика. Правая граница монотонности на графике прироста скользящего коэффициента вариации (значения подчеркнуты в столбцах табл. 1) отграничивает ядерные элементы.

ПОДХОД И.-И. ПОПЕСКУ, Г. АЛЬТМАННА И Я. МАШУТЕКА. ТОЧКА *h*

Попеску, Машутек и Альтманн принимают и неоднородность, и негауссовость лингвостатистических распределений.

Так, тексты значительной длины считаются неоднородными, поскольку написаны не за один раз [15; 8]. Интересно также мнение авторов об устойчивости статистических характеристик лексических единиц в корпусе текстов: они отвергают принцип «чем больше выборка, тем устойчивее статистические характеристики», заменяя его на принцип «чем больше выборка, тем более неоднородным становится текст» [15; 8]. Такой «ход» означает фактическое признание негауссового характера лингвистических распределений. В пользу негауссовости свидетельствует и указание этих авторов на слабую пригодность традиционных критериев (χ-квадрат) для оценки качества аппроксимации распределений единиц в тексте [15; 15].

На гиперболической кривой рангового распределения Альтманн и Попеску выделяют особую точку *h*, для которой выполняется условие

$r = f(r)$ (где r – ранг, $f(r)$ – частота слова с рангом r) [15; 24]. По их мнению, эту точку следует интерпретировать как границу между словами со стертой семантикой (synsemantics) и словами, значимыми для этого текста (autosemantics), но для них эта точка не является маркером неоднородности совокупности, как для Г. Я. Мартыненко. Попеску с соавторами считают, что те полнозначные слова, которые оказываются среди отграниченной точкой h служебной лексики, отражают тематику текста.

Точка перегиба h (h-point), предложенная Дж. Хиршем в 2005 году, определяется как точка, в которой кривая рангового распределения пересекается с прямой $y = x$. Этот параметр мгновенно увязывается с симметрийными идеями Ю. А. Шрейдера [2] и пойнтер-точкой R Б. И. Кудрина. Б. И. Кудрин, что интересно, тоже характеризует этот введенный им параметр как «точку перегиба». Попеску и Альтманн задают точку h так:

$$h = \begin{cases} r, & \text{если } \exists\ r = f(r); \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{если } \overline{\exists}\ r = f(r). \end{cases}$$

В большинстве случаев в эмпирическом ряду есть точка, в которой $r = f(r)$. Если такой точки нет, то берут соседние точки $f(i)$ и $f(j)$, такие, что $f(i) > ri$ и $f(j) < rj$. Чаще всего $r_{i+1} = r_j$.

СОПОСТАВЛЕНИЕ МЕТОДИК

Несмотря на различие позиций Б. И. Кудрина и Г. Я. Мартыненко, обоих исследователей объединяет глубинный принцип сегментации кривой распределения на функциональные зоны, но картина мира, в которой описываются эти представления о сегментации, разная: Кудрин маркирует особые точки (пойнтер-точка, начало и конец распределения), а Мартыненко защищает позицию о смешанном характере итоговой совокупности, хотя обоим в итоге нужна содержательная интерпретация количественных данных. Странным кажется то обстоятельство, что в научном сообществе вовлеченных в «ципфиаду» практически не обсуждаются количественные соотношения между маркированными точками: можно ли по началу распределения предсказать его хвост, общее количество классов и т. п., в общем случае по части реконструировать целое.

Представляется небезынтересным сравнить результаты, получаемые с помощью применения всех трех параметров, на одном материале, а именно на частотных словарях лексем, построенных для 8 списков «Сказания о Мамаевом побоище»[3]. Для исследования статистической структуры вариативного текста были получены в электронном виде 8 списков памятника, представляющие 5 основных редакций и нетиповые и компилятивные списки. В пределах статьи приводятся статистические данные только для четырех списков: 1. Список Основной редакции РНБ O.IV.22, XVI век. 2. Список Летописной редакции СПбОИИ № 251, XVI век. 3. Список Киприановской редакции БАН 32.14.8, XVI век. 4. Список Распространенной редакции РНБ Q.IV.354, XIX век[4].

Для каждого списка были построены частотные словари и произведена лемматизация. Таблица 1 содержит спектровые[5] распределения лексем для частотных словарей 4 списков «Сказания» и данные о приросте скользящего коэффициента вариации и его значениях в характеристических точках распределения. Ячейка, содержащая пойнтер-точку R Б. И. Кудрина, выделена полужирным шрифтом; ячейка, содержащая точку h Хирша – Попеску, – серым фоном. Подчеркиванием выделена строка, содержащая значение точки, с которой начинается монотонное возрастание прироста коэффициента вариации (точки B). Даже визуальный анализ таблицы показывает близость или даже совпадение пойнтер-точки R и точки B. Полученные результаты подталкивают к вполне недвусмысленному выводу о равноценности сопоставляемых методов в практической деятельности. Точка h Хирша – Попеску ожидаемо располагается значительно ближе к центру распределения.

Обратимся теперь к анализу лексем, отграничиваемых с помощью рассматриваемых параметров (табл. 2). Лексика, выделенная с помощью пойнтер-точки R и правой границы монотонности графика СКВ, – это союзы, местоимения, клише или коллокаты (титул «великий князь»), а также 2 высокочастотных глагола («быти» и «глаголати») и имя и отчество великого князя. Роль Дмитрия в походе против татар, его поступки и речи постоянно подчеркиваются в тексте.

Анализ лексем, отсекаемых с помощью точки h, подтверждает выводы И.-И. Попеску с соавторами: предложенный ими параметр позволяет отделить большинство нетематических единиц для конкретного текста (что невозможно сделать с помощью параметров Г. Я. Мартыненко и Б. И. Кудрина) и имеет практическое значение. Выше точки h располагается идеологически маркированная лексика («Господь», «Бог», «святыи», «брат», «русский», «полк») и лексика, характеризующая противостоящие силы («русский» против «поганыи»).

Ни один из этих параметров не является стилеразличающим (в данном случае, списки Киприановской – БАН 32.14.8 и Летописной – СПбОИИ № 251 редакций, бытующих в составе летописных сводов, никак не маркированы по составу или количеству выделенной лексики).

Таблица 1

Совмещение параметров методик

Основная, O.IV.22				Летописная, СПбОИИ № 251				Киприановская, 32.14.8				Распространенная, Q.IV.354			
n	**f(n)**	**V**	**dV**	**n**	**f(n)**	**V**	**dV**	**n**	**f(n)**	**V**	**dV**	**n**	**f(n)**	**V**	**dV**
1	896	0,000		1	885	0,000		1	778	0,000		1	940	0,000	
2	318	0,349	0,349	2	308	0,348	0,348	2	289	0,350	0,350	2	279	0,342	0,342
3	172	0,478	0,129	3	140	0,469	0,121	3	142	0,472	0,122	3	149	0,478	0,136
4	92	0,560	0,083	4	75	0,554	0,085	4	88	0,562	0,090	4	115	0,586	0,108
5	59	0,627	0,067	5	53	0,629	0,075	5	59	0,633	0,071	5	65	0,652	0,067
6	43	0,687	0,060	6	44	0,700	0,071	6	31	0,680	0,047	6	40	0,703	0,051
7	23	0,727	0,040	7	25	0,747	0,047	7	31	0,737	0,057	7	25	0,744	0,041
8	25	0,779	0,052	8	29	0,808	0,061	8	33	0,801	0,064	8	26	0,795	0,051
9	24	0,832	0,054	9	18	0,850	0,041	9	13	0,829	0,028	9	9	0,816	0,021
10	15	0,869	0,037	10	12	0,882	0,032	10	19	0,876	0,047	10	20	0,871	0,055
11	16	0,912	0,043	11	8	0,907	0,025	11	12	0,908	0,032	11	15	0,914	0,043
12	9	0,939	0,027	12	8	0,936	0,029	12	15	0,951	0,043	12	17	0,964	0,050
13	13	0,981	0,042	13	9	0,973	0,037	13	7	0,973	0,022	13	17	1,013	0,049
14	9	1,012	0,031	14	10	1,016	0,043	14	9	1,005	0,031	14	11	1,046	0,032
15	8	1,041	0,029	15	7	1,047	0,031	15	7	1,031	0,026	15	10	1,077	0,031
16	4	1,058	0,016	16	6	1,076	0,029	16	7	1,060	0,028	16	13	1,117	0,041
17	6	1,085	0,027	17	5	1,101	0,026	17	9	1,097	0,038	17	5	1,134	0,016
18	6	1,113	0,029	18	5	1,129	0,027	18	7	1,127	0,030	18	3	1,145	0,011
19	5	1,138	0,025	19	5	1,158	0,029	19	5	1,149	0,022	19	5	1,165	0,021
20	2	1,149	0,011	20	4	1,182	0,024	20	3	1,163	0,014	20	3	1,179	0,014
21	8	1,194	0,045	21	2	1,195	0,013	21	5	1,189	0,026	21	3	1,194	0,015
22	6	1,228	0,033	22	4	1,223	0,028	22	5	1,215	0,026	22	2	1,205	0,011
23	8	1,270	0,043	23	7	1,271	0,048	23	1	1,221	0,006	23	3	1,223	0,018
24	4	1,291	0,021	24	5	1,304	0,033	24	2	1,234	0,013	24	3	1,242	0,019
25	1	1,297	0,006	25	1	1,311	0,007	25	4	1,260	0,026	25	4	1,268	0,026
26	2	1,310	0,013	26	7	1,360	0,049	26	2	1,274	0,014	26	2	1,282	0,014
27	3	1,329	0,020	27	2	1,374	0,014	27	6	1,315	0,041	27	2	1,297	0,015
28	1	1,336	0,007	28	1	1,381	0,007	28	1	1,322	0,007	28	3	1,319	0,023
29	1	1,344	0,008	29	1	1,389	0,008	29	5	1,358	0,036	29	2	1,335	0,016
30	1	1,352	0,008	30	3	1,414	0,025	30	2	1,372	0,014	30	1	1,344	0,008
31	2	1,370	0,017	31	1	1,423	0,009	31	1	1,379	0,008	31	3	1,370	0,026
32	4	1,405	0,035	32	1	1,433	0,009	32	3	1,403	0,024	32	2	1,388	0,018
33	2	1,422	0,017	35	5	1,488	0,055	33	2	1,419	0,016	33	2	1,406	0,018
34	3	1,449	0,027	37	1	1,500	0,012	34	1	1,428	0,008	36	1	1,417	0,011
35	3	1,475	0,026	38	3	1,535	0,035	35	4	1,461	0,034	37	2	1,441	0,023
38	1	1,486	0,011	40	1	1,548	0,013	37	1	1,471	0,009	38	2	1,464	0,024
39	2	1,508	0,022	42	1	1,562	0,014	38	1	1,480	0,010	39	4	1,510	0,045
43	1	1,521	0,014	43	1	1,577	0,015	39	5	1,528	0,047	40	1	1,521	0,011
44	1	1,536	0,014	44	1	1,591	0,015	41	2	1,547	0,020	42	2	1,546	0,025
47	2	1,568	0,032	45	1	1,607	0,015	42	2	1,567	0,020	43	1	1,558	0,012
48	1	1,584	0,016	48	1	1,624	0,018	43	3	1,595	0,029	46	1	1,573	0,015
49	1	1,601	0,016	52	1	1,646	0,021	45	1	1,606	0,010	47	2	1,602	0,029
53	2	1,639	0,038	54	1	1,668	0,022	46	1	1,616	0,011	49	1	1,617	0,015
54	3	1,692	0,053	56	1	1,691	0,023	47	1	1,627	0,011	50	1	1,633	0,016
56	1	1,710	0,018	57	2	1,736	0,045	49	1	1,639	0,012	54	2	1,670	0,037
57	1	1,728	0,018	58	1	1,758	0,022	50	2	1,663	0,024	56	2	1,707	0,037
60	2	1,767	0,039	64	1	1,785	0,027	56	1	1,679	0,016	57	2	1,742	0,035
62	1	1,787	0,020	65	1	1,812	0,027	57	2	1,711	0,031	58	1	1,759	0,017
66	1	1,809	0,023	67	1	1,839	0,027	58	1	1,726	0,015	60	1	1,777	0,018
69	1	1,834	0,024	74	2	1,905	0,065	59	1	1,741	0,015	62	1	1,796	0,019
73	1	1,860	0,027	80	1	1,941	0,036	60	1	1,757	0,015	63	1	1,815	0,019
81	1	1,894	0,034	85	2	2,018	0,077	61	1	1,772	0,015	64	2	1,852	0,037
82	1	1,926	0,032	87	2	2,088	0,070	63	2	1,804	0,032	69	1	1,873	0,021
83	1	1,958	0,031	97	2	2,170	0,082	**67**	**1**	1,822	0,018	76	1	1,900	0,027
84	1	1,988	0,030	112	2	2,272	0,102	68	1	1,840	0,018	86	1	1,935	0,035
87	1	2,019	0,031	120	2	2,374	0,102	70	1	1,858	0,018	87	1	1,969	0,034
89	1	2,050	0,031	**123**	**1**	2,422	0,047	86	1	1,890	0,032	93	1	2,007	0,038
90	1	2,080	0,030	158	1	2,507	0,085	89	1	1,922	0,033	94	1	2,043	0,036
97	1	2,115	0,035	165	1	2,591	0,084	93	1	1,956	0,034	100	1	2,083	0,040
104	2	2,190	0,075	173	1	2,675	0,084	96	1	1,991	0,035	101	1	2,121	0,038
114	**1**	2,233	0,043	184	1	2,762	0,087	105	1	2,032	0,041	107	1	2,163	0,041
141	1	2,303	0,070	196	1	2,852	0,090	106	1	2,071	0,039	110	1	2,204	0,041
144	1	2,370	0,067	246	1	2,994	0,142	175	1	2,198	0,127	113	1	2,244	0,041
148	1	2,434	0,064	281	1	3,162	0,167	182	1	2,318	0,120	115	1	2,284	0,040
153	1	2,498	0,063	690	1	4,123	0,962	187	1	2,429	0,111	132	1	2,337	0,053
174	1	2,578	0,080					188	1	2,527	0,097	139	1	2,393	0,056
184	1	2,661	0,083					203	1	2,630	0,103	143	1	2,448	0,055
206	1	2,759	0,098					210	1	2,728	0,098	148	1	2,503	0,055
210	1	2,849	0,090					215	1	2,818	0,090	154	1	2,558	0,055
269	1	2,998	0,149					227	1	2,910	0,091	187	2	2,718	0,159
306	1	3,171	0,173					284	1	3,052	0,143	**196**	**1**	2,793	0,075
760	1	4,173	1,002					301	1	3,191	0,139	230	1	2,896	0,103
								1203	1	5,162	1,971	288	1	3,053	0,158
												340	1	3,252	0,199
												796	1	4,246	0,994

Таблица 2

Лексемы, отсекаемые с помощью точки h (весь столбец), пойнтер-точки R (полужирный шрифт) и правой границы монотонности СКВ (серый фон)

r, ранг	Основная, РНБ О.IV.22	$f(r)$	Летописная, СПбОИИ 251	$f(r)$	Киприановская, БАН 32.14.8	$f(r)$	Распространенная, РНБ Q.IV.354	$f(r)$
1	**И**	760	**И**	690	**И**	1203	**И**	796
2	**ЖЕ**	306	**ЖЕ**	281	**КНЯЗЬ**	301	**ЖЕ**	340
3	**КНЯЗЬ**	269	**КНЯЗЬ**	246	**ОН**	284	**КНЯЗЬ**	288
4	**НА**	210	**НА**	196	**ВЕЛИКИИ**	227	**ВЕЛИКИИ**	230
5	**ВЕЛИКИИ**	206	**ОН**	184	**НА**	215	**НА**	196
6	**СВОИ**	184	**СВОИ**	173	**ЖЕ**	210	СВОИ	187
7	**БЫТИ**	174	**ВЕЛИКИИ**	165	**БЫТИ**	203	БЫТИ	187
8	**ОН**	153	**БЫТИ**	158	**ВЕСЬ**	188	ОН	154
9	**ТЪ**	148	**НЕ**	123	**С**	187	В	148
10	**В**	144	С	120	**В**	182	РЕЧИ	143
11	**С**	141	В	120	**СВОИ**	175	ОНИ	139
12	**НЕ**	114	ТЪ	112	**К**	106	С	132
13	ОНИ	104	ОНИ	112	**ОНИ**	105	К	115
14	МЫ	104	РЕЧИ	97	**СЬ**	96	ЯКО	113
15	РЕЧИ	97	А	97	**ТЪ**	93	НЕ	110
16	А	90	МЫ	87	**НЕ**	89	ОТ	107
17	ЯКО	89	БО	87	**ОТ**	86	ВЕСЬ	101
18	ОТ	87	ОТ	85	**МНОГИИ**	70	ТЪ	100
19	АЗ	84	НО	85	**ГЛАГОЛАТИ**	68	АЗ	94
20	ВЕСЬ	83	К	80	**ДМИТРИИ**	67	СЬ	93
21	БО	82	ЯКО	74	ЯКО	63	МЫ	87
22	ГОСПОДЬ	81	АЗ	74	ЦАРЬ	63	ДМИТРИИ	86
23	К	73	ВЕСЬ	67	ПО	61	А	76
24	ЦАРЬ	69	ДА	65	ДА	60	ЦАРЬ	69
25	РУССКИИ	66	ПО	64	АЗ	59	ПО	64
26	ПО	62	ГОСПОДЬ	58	ГОСПОДЬ	58	НАШ	64
27	НАШ	60	ЦАРЬ	57	УБО	57	БОГ	63
28	БОГ	60	СЬ	57	МЫ	57	ДА	62
29	ТВОИ	57	ДМИТРИИ	56	СИЛА	56	БО	60
30	БРАТ	56	БРАТ	54	ПРИИТИ	50	ВИДЕТИ	58
31	СЬ	54	ПОЛК	52	МАМАИ	50	ИВАНОВИЧ	57
32	ПОЛК	54	НАШ	48	БОГ	49	БРАТ	57
33	ДА	54	МНОГИИ	45	СВЯТЫИ	47	ТЫ	56
34	СВЯТЫИ	53	ТЫ	44	ЗА	46	ГОСПОДЬ	56
35	ДМИТРИИ	53	ТВОИ	43	РЕЧИ	45	ТВОИ	54
36	МОИ	49	БОГ	42	О	43	МНОГИИ	54
37	ТЫ	48	РУССКИИ	40	ИВАНОВИЧ	43	ПРИИТИ	50
38	ПОГАНЫИ	47	СВЯТЫИ	38	А	43	О	49
39	НЫНЕ	47			ТЫ	42	ПОЛК	47
40	АКИ	44			ВИДЕТИ	42	МОИ	47
41	МНОГИИ	43			ЗЕЛО	41	НО	46
42	НАЧАТИ	39					ДАТИ	43
43							СВЯТЫИ	42

ПРИМЕЧАНИЯ

[1] Термин «H-распределение» предложен Б. И. Кудриным, см., напр., [3], [4].

[2] См. персональный сайт Б. И. Кудрина. Режим доступа: www.kudrinbi.ru

[3] Обращение к такому неожиданному источнику при существующих многомиллионных корпусах текстов вызвано тем, что некоторые из моделей описания структуры поликомпонентных объектов (текстов), которые нет возможности обсудить в рамках данной статьи, предназначены для описания диахронического изменения структуры объекта (структурно-топологическая динамика В. В. Фуфаева [9], [10] и модель изменения словаря во времени М. В. Арапова – М. М. Херц [1]). Эти модели могут быть успешно применены только к текстам, для которых зафиксированы их изменения во времени (вариативным текстам). В силу культурных особенностей такие письменные тексты довольно большого объема обнаруживаются только в средневековой литературе.

[4] Указана дата списка, а не время возникновения редакции.

[5] Т. е. распределения типа: n – численность класса, $f(n)$ – количество классов с такой численностью.

СПИСОК ЛИТЕРАТУРЫ

1. Арапов М. В., Херц М. М. Изменение словаря во времени (опыт теории) // Информационные вопросы семиотики, лингвистики и автоматического перевода. Вып. 3. М.: ВИНИТИ, 1972. С. 3–85.
2. Арапов М. В., Шрейдер Ю. А. Закон Ципфа и принцип диссимметрии системы // Семиотика и информатика. Вып. 10. М.: ВИНИТИ, 1978. С. 74–95.
3. Кудрин Б. И. Математика ценозов: видовое, ранговидовое, ранговое по параметру гиперболические H-распределения и законы Лотки, Ципфа, Парето, Мандельброта // Философские основания технетики. М., 2002. С. 357–412.
4. Кудрин Б. И. Мои семь отличий от Ципфа // Общая и прикладная ценология. М., 2007. № 4. С. 25–33.
5. Кудрин Б. И. Распределение электрических машин по повторяемости как некоторая закономерность // Электрификация металлургических предприятий Сибири. Вып. 2. Томск, 1974. С. 31–40.
6. Мартыненко Г. Я. Введение в теорию числовой гармонии текста. СПб., 2009. 252 с.
7. Мартыненко Г. Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других

сложных системах // Структурная и прикладная лингвистика. Вып. 1. Л., 1978. С. 63–80.
8. Мартыненко Г. Я. Основы стилеметрии. Л., 1988. 174 с.
9. Фуфаев В. В. Общеценологический метод структурно-топологического анализа самоорганизующихся систем // Общая и прикладная ценология. М., 2007. № 3. С. 23–31.
10. Фуфаев В. В. Основы теории динамики структуры техноценозов. Математическое описание ценозов и закономерности технетики // Ценологические исследования. Вып. 1. Абакан, 1996. С. 156–193.
11. Чебанов С. В. Концепции ранговых распределений: консенсусный анализ // Ценологические исследования. Вып. 46. М., 2012. С. 72–86.
12. Чебанов С. В. Оптимальность и экстремальность в культуре, ципфиада и закон Лотмана // Ценологические исследования. Вып. 28. М., 2005. С. 411–428.
13. Чебанов С. В. Распределения с неопределенными центральными моментами, размерная структура природных тел и натуральнозначные функции натуральных аргументов // Философские основания технетики. Ценологические исследования. Вып. 19. М., 2002. С. 436–444.
14. Herdan G. Quantitative linguistics. London, 1964. 284 p.
15. Popescu I.-I., Mačutek J., Altmann G. Aspects of word frequencies. RAM-Verlag, 2009. IV + 198 p. Available at: http://www.iipopescu.com/Aspects_of_Word_Frequencies.pdf

---

Kovrigina L. Yu., Saint Petersburg State University (Saint Petersburg, Russian Federation)

METHODOLOGICAL DIFFERENCES IN MODELING TEXTS' STATISTICAL STRUCTURE (on the example of "The Tale of The Rout of Mamai")

Three methods of modeling statistical structure of the text are analyzed. The obtained comparative results were derived by the employment of different statistical models to the same material ("The Tale of The Rout of Mamai"). All compared models are designed to separate autosemantic words from synsemantic words of the plot. The results received during models' testing are provided . The *h*-point introduced by Hirsch – Popescu is shown to be the most suitable parameter helping to separate content words from structure-class words. The *h*-point marks the biggest part of non-thematic words for a certain text.

Key words: text variants, text component structure, non-gaussian distributions, *H*-distribution, concentration and dispersion of elements in linguistic distributions, population heterogeneity, Kudrin's *R*-point, Hirsch – Popescu's *h*-point (*h*-index)

REFERENCES

1. Arapov M. V., Herts M. M. Temporal Variation of Vocabulary (Experience Practice) [Izmeneniye slovarya vo vremeni (opyt teorii)]. *Informatsionnye voprosy semiotiki, lingvistiki i avtomaticheskogo perevoda* [Informatic Problems of Semiotics, Linguistics and Automatic Translation]. Vol. 3. Moscow: VINITI Publ., 1972. P. 3–85.
2. Arapov M. V., Schreyder Yu. A. Zipf's Law and a Principle of the System's Dissymmetry [Zakon Tsipfa i printsip dissimmetrii sistemy]. *Semiotika i informatika* [Semiotics and Informatics]. Vol. 10. Moscow: VINITI Publ., 1972. P. 74–95.
3. Kudrin B. I. Mathematics of Coenoses: Hyperpolic *H*-distributions (Species Distribution, Rank-species Distribution, Rank-parameter Distribution) and Lotka's, Zipf's, Pareto's, Mandelbrot's Laws [Matematika tsenozov: vidovoye, rangovidovoye, rangovoye po parametru giperbolicheskiye *H*-raspredeleniya i zakony Lotki, Tsipfa, Pareto, Mandel'brota]. *Filosofskiye osnovaniya tekhnetiki* [Philosophical Foundations of Technetics]. Moscow, 2002. P. 357–412.
4. Kudrin B. I. There are Seven Points That I Differ From Zipf [Moi sem' otlichiy ot Tsipfa]. *Obshchaya i prikladnaya tsenologiya* [General and Applied coenology]. Moscow, 2007. № 4. P. 25–33.
5. Kudrin B. I. A Pattern of Distribution of Electric Machinery on Its Frequency of Occurrence [Raspredeleniye elektricheskikh mashin po povtoryaemosti kak nekotoraya zakonomernost']. *Elektrifikatsiya metallurgicheskikh predpriyatiy Sibiri* [Motorization of Iron and Steel Plants of Siberia]. Vol. 2. Tomsk, 1974. P. 31–40.
6. Martynenko G. Ya. *Vvedeniye v teoriyu chislovoy garmonii teksta* [Introduction to the Theory of Numeric Harmony of a Text]. Saint-Petersburg, 2009. 252 p.
7. Martynenko G. Ya. Some Regularities in Concentration and Dispersion of Elements in Linguistics and Other Complex Systems [Nekotorye zakonomernosti kontsentratsii i rasseyaniya elementov v lingvisticheskikh i drugikh slozhnykh sistemakh]. *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics]. Vol. 1. Leningrad, 1978. P. 63–80.
8. Martynenko G. Ya. *Osnovy stilemetrii* [Basics of Stylemetrics]. Leningrad, 1988. 174 p.
9. Fufaev V. V. Structure-topological Analysis: A Universal Coenological Method for Self-organizing Systems [Obshchetsenologicheskiy metod strukturno-topologicheskogo analiza samoorganizuyushchikhsya sistem]. *Obshchaya i prikladnaya tsenologiya* [General and Applied Coenology]. Moscow, 2007. № 3. P. 23–31.
10. Fufaev V. V. Theoretical Basics of Structural Dynamics of Technocoenoses [Osnovy teorii dinamiki struktury tekhnotsenozov]. *Tsenologicheskiye issledovaniya* [Coenological Research]. Vol. 1. Abakan, 1996. P. 156–193.
11. Chebanov S. V. Conceptions of Rank Distributions: Consensus Analysis [Kontseptsii rangovykh raspredeleniy: konsensusnyi analiz]. *Tsenologicheskiye issledovaniya* [Coenological Research]. Vol. 46. Moscow, 2012. P. 72–86.
12. Chebanov S. V. Optimality and Extremality Within the Culture, "Zipfiade" and Lotman's Law [Optimal'nost' i ekstremal'nost' v kul'ture, tsipfiada i zakon Lotmana]. *Tsenologicheskiye issledovaniya* [Coenological Research]. Vol. 28. Moscow, 2005. P. 411–428.
13. Chebanov S. V. Distributions with Undetermined Central Moments, Size Structure of Natural Bodies and Natural Value Functions of Natural Value Arguments [Raspredeleniya s neopredelennymi tsentral'nymi momentami, razmernaya struktura prirodnykh tel i natural'noznachnye funktsii natural'nykh argumentov]. *Filosofskiye osnovaniya tekhnetiki* [Philosophical Foundations of Technetics]. Vol. 19. Moscow, 2002. P. 436–444.
14. Herdan G. Quantitative Linguistics. London, 1964. 284 p.
15. Popescu I.-I., Mačutek J., Altmann G. Aspects of Word Frequencies. RAM-Verlag, 2009. IV + 198 p. Available at: www.iipopescu.com/Aspects_of_Word_Frequencies.pdf

Поступила в редакцию 10.06.2013