

О частотно-ранговых распределениях интервалов оригинальных построений литературных произведений

Е.В. Морозенко

Омский государственный технический университет, пр. Мира, д. 11, г. Омск, 644050, Россия

morozenkoev@rambler.ru

Аннотация. Частотно-ранговые распределения широко используются при количественном описании объектов в информатике, лингвистике, биологии и многих других областях. Множество работ посвящено различным формам распределений типа Ципфа – Мандельброта (Брэдфорда, Лотка, Парето и др.). При этом среди объектов, характеризующихся ранговыми распределениями, особо выделяют литературные и музыкальные тексты, являющиеся продуктами непосредственной интеллектуальной деятельности человека.

В семидесятых годах XX века известный ученый-кибернетик Ю.К. Орлов открыл феномен целостности произведений. Проведенное им сопоставление полученных теоретических соотношений с эмпирическим материалом показало, что закон Ципфа-Мандельброта относительно более точно выполняется на полных текстах крупных литературных произведений и гораздо хуже для их частей или конгломератов. Таким образом, совпадение теоретического («циффовского») и эмпирического ранговых распределений можно рассматривать как признак высокой организованности, «правильности», «системности» исследуемого объекта, взятого как единое целое.

В данной работе приведены результаты обработки литературных произведений, позволяющие выдвинуть гипотезу о высокой степени их организации не только с точки зрения количественного состава, но и с точки зрения взаимного расположения элементов, определяющих этот состав. Данное предположение было сделано на основе относительно неплохого выполнения закона Ципфа-Мандельброта для ранговых распределений частот интервалов, определяющих строй исследуемых литературных текстов [1].

Ключевые слова: частотно-ранговые распределения, закон Ципфа-Мандельброта, строй знаковой последовательности, интервал, распределение интервалов

Закон Ципфа-Мандельброта

Напомним основные положения закона Ципфа-Мандельброта.

Объектом нашего рассмотрения является текст, понимаемый как список вхождений словоформ. Каждой словоформе соответствует некоторое слово. Совокупность (или список) всех слов, соответствующих словоформам, образующих текст, мы будем называть словарем данного текста. Для каждого слова из словаря мы можем указать целое число n , равное количеству имеющихся в тексте словоформ, которым соответствует данное слово. Упорядочим теперь слова в словаре по убыванию величин n . Номер слова в таком списке назовём рангом g . Итак, ранги, приписываемые словам в словаре, принимают значения от 1 до M , где M – объём словаря, т.е. общее количество слов в словаре [3].

Важно отметить, что можно по-разному определять, что является формой рассматриваемого слова. В данной работе, при автоматической сегментации текста, использовались слова, нормированные путем отбрасывания окончания. Критерием качества сегментации, являлась степень совпадения мощностей составов, полученных автоматически и вручную, для текстов приведенных в работах Орлова Ю.К.[2].

Итак, закон Ципфа-Мандельброта, для описанной выше модели текста имеет вид:

$$p_i = \frac{K}{(B + i)^\gamma}, \quad (1)$$

$$M = KZ - B, \quad (2)$$

где $K, B, \gamma - \text{const}$:

$$K = \frac{1}{\ln F_1}; \quad B = \frac{K}{p_1} - 1, \quad \gamma \approx 1.$$

p_i, p_1 – частоты, соответственно, i -ого, и 1-ого по рангу слов, F_1 – число вхождений самого частого слова, M – теоретический объем словаря текста длиной Z .

Чтобы согласовать зону низких частот с формулой Мандельброта, допускается, что число разных слов n_F , каждое из которых встретилось на выборке объемом Z единиц ровно F раз каждое, равно:

$$n_F = \frac{M}{F(F+1)}. \quad (3)$$

Таким образом, слов, встретившихся ровно один раз каждое ($F = 1$), должно быть $n_1 = M/2$, т. е. половина всего словарного запаса M выборки [2].

На рис.1 в билогарифмическом масштабе представлены теоретические и фактические частотные кривые для полного произведения «Дворянское гнездо» (Тургенев А.С.) и для его части ($\sim 1/3$). Для удобства расположения кривых по оси ординат отложены значения не относительных, а абсолютных частот. Теоретические кривые получены умножением p_i на объем текста. Легко заметить, что для полного текста получилось хорошее соответствие, в то время как для отрывка наблюдается существенное отклонение теоретических и эмпирических данных.

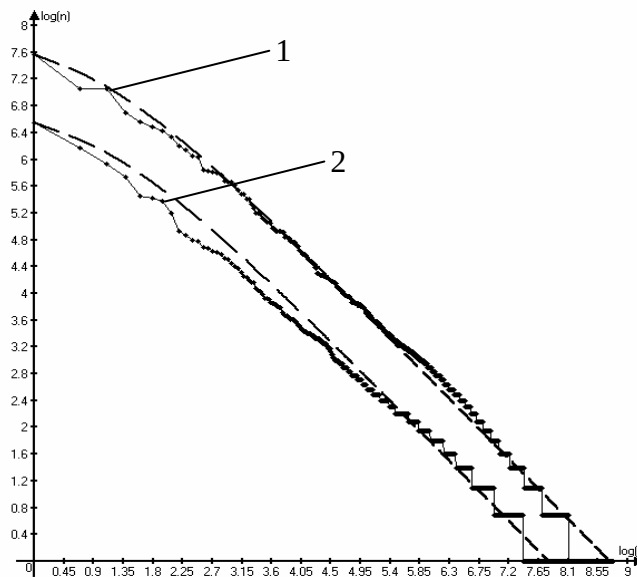


Рис. 1. Фактические (—•—) и теоретические (---) частотно-ранговые распределения слов для произведения Тургенева А.С. «Дворянское гнездо»

1 – полный текст, 2 – отрывок текста;

$\log(n), \log(r)$ – логарифмы абсолютной частоты и ранга слова.

В таблицах 1 и 2 представлены характеристики нескольких обследованных произведений совместно с теоретическими величинами, вычисленными по формулам, приведенным выше. Здесь наблюдается такая же картина – для отдельных произведений относительное соответствие лучше, чем для отрывков.

Таблица 1. Числовые характеристики слов для полных произведений

Произведение	F1	Z	M (ф.)	M (т.)	$\delta(M)$ %	n1 (ф.)	n1 (т.)	$\delta(n1)$ %
Тургенев И.С. "Дворянское гнездо"	0.0412	46795	6480	6185	4.77	3155	3092	2.02
Тургенев И.С. "Рудин"	0.0369	35498	5019	4943	1.54	2463	2472	-0.34
Шолохов М.А. "Поднятая целина" (книга 1)	0.0322	94776	12830	11810	8.64	6213	5905	5.22
Шолохов М.А. "Поднятая целина" (книга 2)	0.0373	112561	12282	13489	-8.95	5618	6744	-16.70
Пушкин А.С. "Капитанская дочка"	0.0398	29312	4789	4148	15.45	2355	2074	13.55
Куприн А.И. "Поединок"	0.0460	66753	9007	8313	8.35	4422	4156	6.39
Гоголь Н.В. "Тарас Бульба"	0.0516	36934	6167	4890	26.11	3212	2445	31.37

Таблица 2. Числовые характеристики слов для отрывков произведений

Произведение	F ₁	Z	M (ф.)	M (т.)	δ(M) %	n1 (ф.)	n1 (т.)	δ(n1) %
Тургенев И.С. "Дворянское гнездо"	0.0434	16000	3763	2444	53.97	2102	1222	72.01
Тургенев И.С. "Рудин"	0.0406	12000	2761	1937	42.54	1535	968	58.49
Шолохов М.А. "Поднятая целина" (книга 1)	0.0307	36000	7386	5134	43.86	4130	2567	60.89
Шолохов М.А. "Поднятая целина" (книга 2)	0.0369	40000	7187	5479	31.17	3893	2740	42.11
Пушкин А.С. "Капитанская дочка"	0.0388	10000	2567	1675	53.25	1491	838	78.03
Куприн А.И. "Поединок"	0.0457	22000	4733	3181	48.79	2646	1590	66.36
Гоголь Н.В. "Тарас Бульба"	0.0501	13000	3343	2005	66.73	1967	1002	96.21

Здесь и далее во всех таблицах:

F₁ – частота вхождения самого частого слова (интервала);

Z – длина текста, определяемая числом всех слов;

M(ф.) – количество разных слов (интервалов) в данном произведении (мощность фактического словаря);

M(т.) – мощность расчетного словаря, вычисленная по формуле (2) закона Ципфа-Мандельброта;

n1(ф.) – фактическое количество одноразовых слов (интервалов) в данном произведении;

n1(т.) – теоретическое количество одноразовых слов (интервалов), вычисленная по формуле (3);

δ(M) – отклонение мощности фактического словаря слов (интервалов) от расчетного;

δ(n1) – отклонение фактического количества одноразовых слов (интервалов) от расчетного.

Строй знаковой последовательности и ранговые распределения интервалов

До недавнего времени при исследовании текстов почти не уделялось внимания закономерностям конкретного расположения знаков или слов, составляющих отдельную символическую последовательность.

Рассмотрим знаковую последовательность конечной длины. **Строй цепи** определяется как особый кортеж, компонентами которого являются натуральные числа, представляющие знаки данной последовательности следующим образом:

- при просмотре цепи слева направо каждый первый встречный отличный от других знак обозначается очередным натуральным числом,
- повторные вхождения этого же знака заменяются тем же числом.

Таким образом, строй представляет взаимное расположение различных и одинаковых элементов цепи независимо от их обозначения, веса и содержимого [1].

Число пустых позиций (плюс один) между любыми (i и $i+1$) двумя ближайшими вхождениями одинаковых знаков j названо интервалом «времени» – Δ_{ij} . *Первый или единственный интервал (если знак j встретился только раз) определяется от первого вхождения предыдущего знака ($j-1$) до первого вхождения текущего знака j .* Интервал для первого вхождения первого знака j определяется от начала текста и равен $\Delta_{11} = 1$. Принципиальность подобной привязки «по алфавиту» прокомментирована ниже. В итоге каждая однородная цепь представлена вектором интервалов, а полная знаковая цепь – матрицей интервалов.

Очевидно, что мощность полученного множества интервалов равна длине текста Z , так как каждому слову соответствует «его» интервал. Подсчитав частоты вхождения одинаковых интервалов, мы можем построить их ранговое распределение.

На рис.2 представлены теоретические и эмпирические частотно-ранговые кривые интервалов, описывающих оригинальный строй двух литературных произведений – «Поднятая целина» (Шолохов М.А., книга 1) и «Капитанская дочка» (Пушкин А.С.). Как видно из графиков, фактическое ранговое распределение интервалов для оригинального построения текстов хорошо совпадает с теоретическим, демонстрируя отклонение только в зоне низких частот.

При просмотре ранговых распределений интервалов для нескольких десятков литературных произведений наблюдалась аналогичная картина. Закономерное отклонение распределений в зоне низких частот (фактическое количество интервалов всегда больше) приводит и к отклонению фактического объема «словаря интервалов» от расчетного в среднем на 30 % (таблица 3).

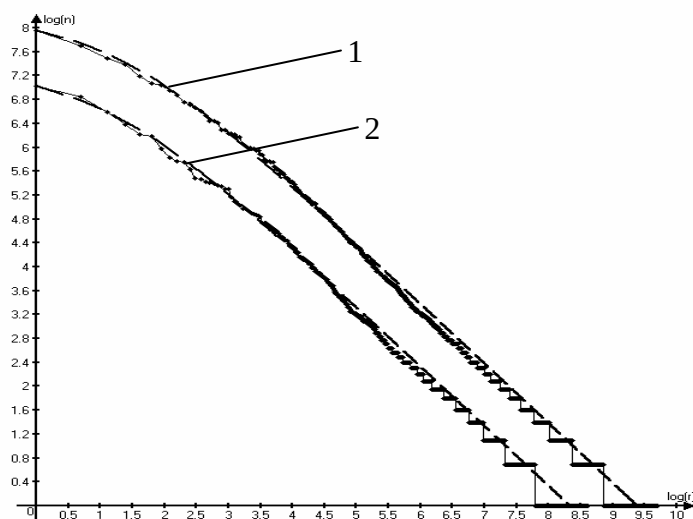


Рис. 2. Фактические (—♦—) и теоретические (---) частотно-ранговые распределения интервалов для произведений:
1 – Шолохов М.А. «Поднятая целина» (книга 1), 2 – Пушкин А.С. «Капитанская дочка»;
 $\log(n)$, $\log(r)$ – логарифмы абсолютной частоты и ранга интервала.

Таблица 3. Числовые характеристики интервалов для полных произведений

Произведение	F1	Z	M (ф.)	M (т.)	$\delta(M)$ %	n1 (ф.)	n1 (т.)	$\delta(n1)$ %
Тургенев И.С. "Дворянское гнездо"	0.0331	46795	8213	6368	28.97	4546	3184	42.78
Тургенев И.С. "Рудин"	0.0346	35498	6213	4988	24.56	3441	2494	37.97
Шолохов М.А. "Поднятая целина" (книга 1)	0.0303	94776	16377	11902	37.60	9323	5951	56.66
Шолохов М.А. "Поднятая целина" (книга 2)	0.0213	112561	17869	14464	23.54	10262	7232	41.90
Пушкин А.С. "Капитанская дочка"	0.0385	29312	5467	4168	31.17	3047	2084	46.21
Куприн А.И. "Поединок"	0.0337	66753	11509	8646	33.11	6449	4323	49.18
Гоголь Н.В. "Тарас Бульба"	0.0394	36934	6599	5068	30.21	3819	2534	50.71

Отметим, что привязка однородных цепей (слов) любого другого типа (до конца текста – привязка к «финишу», от начала текста – привязка к «старту», от последнего вхождения слова до первого – замыкание в кольцо) или вовсе ее отсутствие, приводило не только к гораздо большему (чем наблюдаемое) отклонению в зоне низких частот (до ~100%), но и вовсе к очевидному несовпадению теоретической и эмпирической частотных кривых (рис.3).

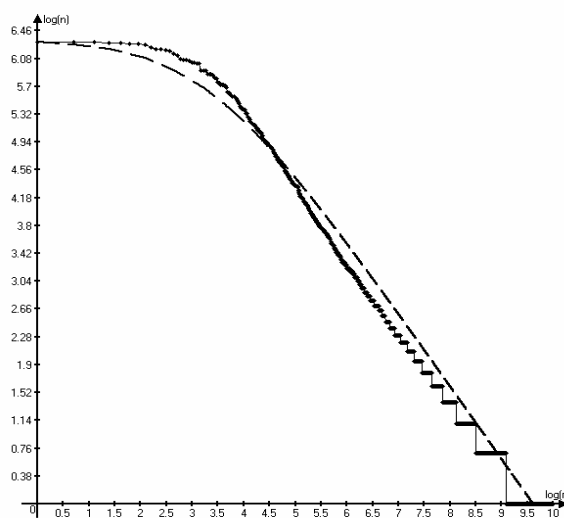


Рис. 3. Фактическое (—♦—) и теоретическое (---) частотно-ранговое распределение интервалов для произведения Шолохова М.А. «Поднятая целина» (книга 1), в случае привязки к «финишу»
 $\log(n)$, $\log(r)$ – логарифмы абсолютной частоты и ранга интервала.

Так же интерес представляют ранговые распределения интервалов для оригинальных построений частей произведений. На графике (рис.4) и в таблице 4 приведены данные для рассмотренных выше отрывков текстов.

Как видно, отклонение по мощности «словарей интервалов», в сравнении с оригинальным строем полного текста, увеличилось в среднем на 10% (для сравнения для этих же отрывков расхождение по словам увеличилось на ~ 50%), но кривые продолжают демонстрировать неплохое совпадение. На основе полученных результатов можно сделать предположение об относительной «стабильности» строя на протяжении всего литературного произведения.

Таблица 4. Числовые характеристики интервалов для отрывков произведений

Произведение	F1	Z	M (ф.)	M (т.)	$\delta(M) \%$	n1 (ф.)	n1 (т.)	$\delta(n1) \%$
Тургенев И.С. "Дворянское гнездо"	0.0686	16000	3369	2285	47.44	1970	1142	72.43
Тургенев И.С. "Рудин"	0.0707	12000	2480	1779	39.40	1418	890	59.42
Шолохов М.А. "Поднятая целина" (книга 1)	0.0582	36000	7009	4706	48.94	4021	2353	70.89
Шолохов М.А. "Поднятая целина" (книга 2)	0.0460	40000	7213	5320	35.58	4136	2660	55.49
Пушкин А.С. "Капитанская дочка"	0.0781	10000	2015	1501	34.24	1142	750	52.17
Куприн А.И. "Поединок"	0.0674	22000	4367	3012	44.99	2481	1506	64.74
Гоголь Н.В. "Тарас Бульба"	0.0788	13000	2637	1875	40.64	1499	938	59.89

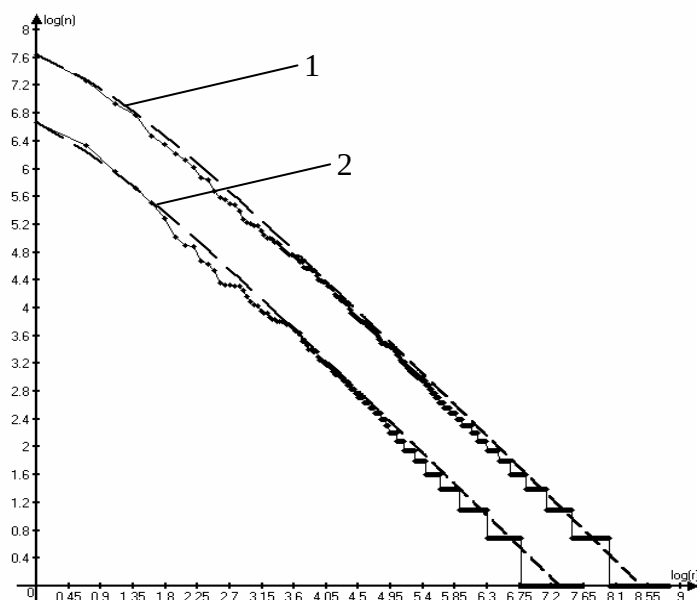


Рис. 4. Фактические (—◆—) и теоретические (— — —) частотно-ранговые распределения интервалов для отрывков произведений:

1 – Шолохов М.А. «Поднятая целина» (книга 1), 2 – Пушкин А.С. «Капитанская дочка»;
 $\log(n)$, $\log(r)$ – логарифмы абсолютной частоты и ранга интервала.

Рассмотрим ранговые распределения интервалов для модификаций строя литературных произведений. Над исследуемыми текстами проводилась «операция перемешивания», заключающаяся в многократной случайной перестановке слов. Таким образом, количественный состав произведения оставался таким же, в то время как оригинальный строй текста (взаимное расположение слов) менялся. Для полученных модификаций также определялись матрицы интервалов и строились частотно-ранговые распределения.

Как и в предыдущем случае, расхождение мощностей теоретических и фактических словарей увеличилось в среднем на 10% (таблица 5), но вот частотные кривые стали демонстрировать закономерное отклонение в зоне высоких частот (рис.5). Любые модификации строя направленные не на перемешивание, а напротив, на упорядочивание текста (так чтобы одинаковые слова шли друг за другом) приводили к подавляющему преобладанию интервалов в единицу, и как следствие, к абсолютному несоответствию теоретической и эмпирической частотных кривых.

Таким образом, результаты обработки нескольких десятков литературных произведений, свидетельствуют о наилучшем совпадении распределений для текстов с оригинальным строем.

Таблица 5. Числовые характеристики интервалов для произведений с модифицированным строем

Произведение	F1	Z	M (ф.)	M (т.)	$\delta(M) \%$	n1 (ф.)	n1 (т.)	$\delta(n1) \%$
Тургенев И.С. "Дворянское гнездо"	0.0336	46795	8612	6354	35.54	4675	3177	47.15
Тургенев И.С. "Рудин"	0.0355	35498	6570	4970	32.19	3584	2485	44.23
Шолохов М.А. "Поднятая целина" (книга 1)	0.0307	94776	17335	11880	45.92	9492	5940	59.80
Шолохов М.А. "Поднятая целина" (книга 2)	0.0234	112561	19026	14285	33.19	10548	7142	47.68
Пушкин А.С. "Капитанская дочка"	0.0426	29312	5893	4109	43.42	3168	2054	54.20
Куприн А.И. "Поединок"	0.0328	66753	12185	8676	40.44	6637	4338	53.00
Гоголь Н.В. "Тарас Бульба"	0.0439	36934	7209	4995	44.32	3950	2498	58.16

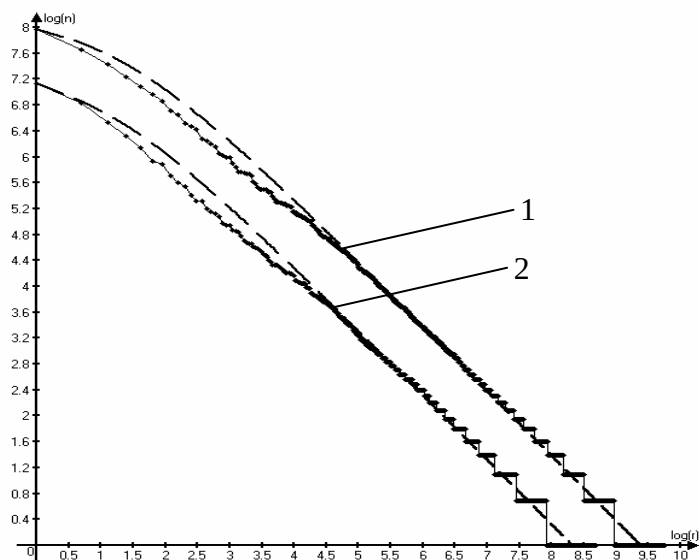


Рис. 5. Фактические (—♦—) и теоретические (—) частотно-ранговые распределения интервалов для произведений с модифицированным строем:

1 – Шолохов М.А. «Поднятая целина» (книга 1), 2 – Пушкин А.С. «Капитанская дочка»;
 $\log(n)$, $\log(r)$ – логарифмы абсолютной частоты и ранга интервала.

Для обоснования, а возможно и формализации, факта отклонения фактического и теоретического распределений в зоне низких частот требуются дальнейшие исследования. Но и на данном этапе можно говорить о том, что *частотно-ранговые распределения интервалов оригинальных построений литературных произведений относительно хорошо описываются законом Ципфа-Мандельброта*. Этот факт свидетельствует о *системности* природы такого объекта как строй литературного текста и подтверждает необходимость его активного исследования для информационных последовательностей не только лингвистической, но и любой другой природы.

В работах Орлова Ю.К., Шрейдера Ю.А. и других отмечается невозможность использования традиционного вероятностного аппарата для исследования объектов, описываемых подобными ранговыми распределениями. Поэтому, наряду с вопросами о природе подобных распределений и причинах их распространенности, актуальным является вопрос и о новых методах исследования подобных объектов. Одним из таких подходов для рассматриваемого в данной работе объекта является непосредственный анализ строя информационных последовательностей [1].

Литература

1. Гуменюк А.С., Богорад Д.М., Об исчислении взаимного расположения элементов в массивах данных, Труды РНТОРЭС им. Попова, серия: научная сессия, посвященная Дню радио. Выпуск: LIX- 2. 2004. – С.114-116.
2. Орлов Ю.К., Частотные структуры конечных сообщений в некоторых естественных информационных системах. Канд. диссертация, Тбилиси: Издательство тбилисского университета, 1974.
3. Шрейдер Ю.А. Ранговые распределения как системное свойство. // Математическое описание ценозов и закономерности технетики. Вып.1 и вып.2 "Ценологические исследования". Ред. и сост.Кудрин Б.И. – Абакан: Центр системных исследований, 1996. - с.33-42.