О состоятельности порядковых статистик частотных словарей

Г.Я. Мартыненко

Санкт-Петербургский государственный университет

g.martynenko@spbu.ru

Аннотация

В статье рассмотрены некоторые порядковые статистики частотных словарей, представленных в виде ранговых распределений. Эти статистики анализируются с точки зрения их состоятельности, то есть проверяется, будут ли эмпирические характеристики по мере увеличения объема выборки устремляться к предельным теоретическим величинам. Тест на состоятельность порядковых статистик был осуществлен на основе корпуса «Один речевой день». Частотный словарь этого корпуса был представлен в виде динамической структуры, состоящей из 10 порций по 10 тыс. словоупотреблений каждая. Порции последовательно присоединялись друг к другу в случайном порядке. На каждом шаге строилось ранговое распределение. В результате проведенного исследования было показано, что анализируемые статистики действительно являются состоятельными. В статье также приводятся доли соответствующих порядковых статистик по семи разным частотным словарям и показаны "порядковые" профили для каждого из этих словарей. На материале разных, но однородных по составу корпусов установлено, что эти статистики обладают высоким стилеразличающим потенциалом. Это свойство порядковых статистик может использоваться для сравнения частотных словарей разной тематики, жанра и объема.

Ключевые слова: частотные словари, повседневная разговорная речь, статистическая состоятельность, статистические шкалы, выборка, динамические ряды, ранговое распределение, равномерное распределение.

1. Введение

Одна из основных целей статистического исследования в лингвистике — формирование полезного набора переменных (признаков) и их статистик, обеспечивающих лаконичное описание совокупности данных наблюдения или

эксперимента. Все статистики рассматриваются в качестве статистических оценок параметров генеральной совокупности, т. е. являются статистическими оценками последних [1].

В любом исследовании к статистическим оценкам должны предъявляться требования, обеспечивающие их надежность и практический смысл. Одним из таких требований является правило состоятельности статистических оценок, заключающееся в том, что эмпирические характеристики по мере увеличения объема выборки устремляются к предельным теоретическим величинам.

В лингвистике существует класс статистических распределений, в которых каноническое правило состоятельности ведет себя по-разному:

- оно или не соблюдается вообще, т. е. статистики не достигают предельных величин при любом, сколь угодно большом объеме выборки;
- оно на практике выполняется, но объемы выборки при этом могут быть очень большими;
- существует небольшая коллекция статистик, которые в полной мере отвечают требованию состоятельности, достигая предельных величин в выборках весьма ограниченного объема.

Последняя группа — это порядковые статистики, которые и рассматриваются в настоящей работе.

2. Частотные словари в системе измерительных шкал и статистических распределений

Важное место в квантитативной лингвистике занимает теория и практика создания частотных словарей, в которых реализуются все виды шкал, используемых в статистике: номинальная, ординальная (порядковая) и количественная. В номинальной шкале варьирует имя лексической единицы, в ординальной — порядок следования в ранжированном ряду, в количественной — значение количественного признака. При этом, если в стандартном частотном словаре представлены все три вида шкал, то в некоторых словарях информация, касающаяся одной из шкал, может отсутствовать. Например, может отсутствовать именная компонента. Тогда мы имеем два варианта: частотно-ранговую сетку или ранговое распределение. Может отсутствовать информация о рангах. В такой ситуации получаем классическое номинальное распределение, в котором именам поставлены в соответствие конкретные частоты. Если же исключить частоты, то получим рангово-номинальную последовательность. Такой вариант упорядочивания сплошь и рядом встречается в реальной жизни.

Способ упорядочивания данных, принятый в частотном словаре, не уникален. В ряде видов научной и практической деятельности способ упорядочивания тот же. Так, в таблицах чемпионатов по футболу (и другим игровым видам спорта) указывается ранг команды, ее имя и число набранных очков. Тот же метод используется в упорядочивании биологических видов по численности, миллиардеров — по величине активов в списке Forbes, ученых — по цитируемости и т. п.

Основную роль в лингвистике и других гуманитарных дисциплинах играет ранговое распределение — может быть, потому, что мир, в котором мы живем — это рейтинговый мир, в котором биологические виды выстраиваются по их численности в определенном сообществе организмов: ученые упорядочиваются по цитируемости, города — по численности населения, футболисты — по числу забитых голов и т. д. Не будет большим преувеличением сказать, что рейтинги правят миром. Среди способов рейтингования частотные словари занимают не последнее место.

3. Параметризация частотных словарей с помощью порядковых статистик

Статистические распределения структурируются с помощью самых разнообразных статистик. Речь идет преимущественно о количественных характеристиках. Это различного рода средние, меры колеблемости, меры формы распределения (асимметрии, эксцесса) и некоторые другие.

В ряде работ предприняты усилия для изучения состоятельности этих характеристик [2] [3] [4] [5] [6] [7]. В этих исследованиях выборка разбивалась на несколько частей, которые затем последовательно присоединялись друг к другу. При этом на каждом шаге, т.е. на нарастающем объеме выборки, вычислялись соответствующие статистики. В итоге возникал динамический ряд, визуальное впечатление от которого позволяло сделать первичные выводы о специфике тренда числовой последовательности. Затем данные динамического ряда выравнивались методами, принятыми в математической статистике: методом скользящей средней, аналитическим выравниванием с помощью метода наименьших квадратов, экспоненциального сглаживания и т. п. Это позволяло делать заключение о состоятельности конкретной статистической характеристики, о ее сходимости к определенным теоретическим величинам.

Так в работах [4] [5] [7] установлено, что ряд статистических характеристик по мере увеличения объема выборки не обнаруживают очевидной сходимости даже при многомиллионных объемах (например, объем словаря, средняя частота), другие (например, число однократных слов или ранговое среднее) продемонстрировали медленную, а третьи (речь идет о порядковых статистиках — квартилях и медиане) — очень быструю сходимость.

Напомним, что порядковые статистики — это характеристики совокупности, зависящие от порядка следования элементов в ранжированном ряду. Таковыми являются наибольшее и наименьшее значения, медиана и квантили — значения переменной, занимающие в ранжированной совокупности определенное место: десятое, двадцать пятое, пятидесятое и т. д. На практике чаще всего используются квартили, разбивающие исследуемую совокупность на четыре части, и говорят о первом, втором (совпадающем с медианой) и третьем квартиле.

Эффект быстрой сходимости порядковых статистик был обнаружен впервые на материале частотных словарей слов-ассоциатов, построенных в ходе ассоциативного эксперимента [5]. Приведем из этой работы данные о динамике медианы при возрастании объема выборки (табл. 1). Таблица приводится с небольшими сокращениями.

Таблица 1. Динамика медианы в частотном словаре слов-ассоциатов

Выборка	100	200	300	400	500	600	700	800	900	1000	1500	1800	2100	2500
Медиана	21	28	32	33	34	35	37	37	39	37	39	39	39	39
Объем словаря	71	121	166	200	235	267	296	321	346	362	473	535	591	643

Из табл. 1 видно, что при возрастании объема наблюдения (в данном случае — числа ассоциатов при слове-стимуле «береза») число единиц словаря неуклонно возрастает, но медиана сравнительно быстро достигает верхнего предела.

Словарь конкретных ассоциатов невелик, в его составе нет строевых элементов и мало слов с широким значением, поэтому его организация довольно быстро стабилизируется. При этом нет полной уверенности в том, проявится ли эффект состоятельности в традиционных частотных словарях, имеющих более сложную и громоздкую организацию, включающую сотни тысяч и даже миллионы словоупотреблений. Однако пространство исследования таких словарей ограничено лишь теми из них, где составители разбивают выборочную совокупность на части, последовательно присоединяемых друг к другу. Такой подход позволяет следить за перестройкой словаря по мере возрастания объема выборки, одновременно осуществляя проверку на состоятельность разных параметров, относящихся к любым шкалам, в том числе порядковой.

4. О состоятельности порядковых статистик

Тест на состоятельность порядковых статистик осуществлен нами на основе корпуса «Один речевой день» [8] [9]. Частотный словарь этого корпуса представлен в виде динамической структуры, состоящей из 10 порций по 10 тыс. словоупотреблений каждая. Порции последовательно присоединяются друг к другу в случайном порядке. На каждом шаге строится ранговое распределение.

В этом можно убедиться, обратившись к результатам эксперимента, в котором на данном материале исследованы порядковые статистики, занимающие в ранжированном ряду определенное место, т.е. обладающие конкретным рангом: первым, десятым, сотым и тысячным [10]. Частоты порядковых статистик приведены в табл. 2, в скобках указаны доли соответствующих частот.

Из табл. 2 видно, что по мере возрастания объема выборки, порядковые статистики (относительные величины в скобках) остаются практически постоянными, не ускоряясь и не замедляясь, проявляя удивительную устойчивость. Причем эта устойчивость с возрастанием объема выборки тоже парадоксально возрастает.

Полученный эффект состоятельности может быть распространен на любые частотные словари умеренного и большого объема. При этом за основу берется максимальный объем словаря, достигнутый в конкретном исследовании. Порядковые статистики вычисляются для этого объема и экстраполируются на

любые объемы любого частотного словаря. В табл. 3 приводятся доли соответствующих порядковых статистик по семи частотным словарям. В скобках указаны ранги каждого из семи частотных словарей, упорядоченных по убыванию соответствующих статистик. На рис. 1–4, в соответствии с этими рангами, показаны "порядковые" профили каждого из семи словарей. Из рисунков видно, что частотные словари образуют родственные группы.

Таблица 2. Зависимость частоты и доли порядковых статистик от объема выборки

Объем выборки	**Ранг порядковой статистики**			
	1	**10**	**100**	**1000**
10 тыс.	382 (0,0382)	156 (0,0156)	15 (0,0015)	1 (0,0001)
20 тыс.	815 (0,0408)	314 (0,0157)	29 (0,0015)	2 (0,0001)
30 тыс.	1154 (0,0385)	481 (0,0160)	45 (0,0015)	3 (0,0001)
40 тыс.	1504 (0,0376)	600 (0,0150	60 (0,0015)	4 (0.0001)
50 тыс.	1915 (0,0383)	784 (0,0157)	73 (0,0015)	5 (0,0001)
60 тыс.	2374 (0,0396)	980 (0,0163)	89 (0,0015)	6 (0,0001)
70 тыс.	2783 (0,0398)	1112 (0,0159)	105 (0,0015)	7 (0,0001)
80 тыс.	3286 (0,0411)	1238 (0,0150)	119 (0,0015)	8 (0,0001)
90 тыс.	3564 (0,0396)	1356 (0,0151)	129 (0,0014)	9 (0,0001)
100 тыс.	3924 (0,0394)	1434 (0,0150)	143 (0,0015)	10 (0,0001)
Среднее значение	(0,0393)	(0,0155)	(0,0015)	(0,0001)

Таблица 3. Систематизация частотных словарей в пространстве порядковых статистик

Частотные словари	**Ранг порядковой статистики**			
	1	**10**	**100**	**1000**
Штабные документы [11]	0,295 (1)	0,0142 (2)	0,00221 (1)	0,000058 (7)
Частотный словарь Л. Андреева [12]	0,0724 (2)	0,0101 (5)	0,00108 (6)	0,000121 (2)
Частотный словарь А.П.Чехова [13]	0,0516 (3)	0,0121 (4)	0,00118 (5)	0,000115 (4)
Частотный словарь А.И.Куприна [14]	0,0472 (4)	0,0068 (7)	0,00099 (7)	0,000116 (3)
Бытовые письма [15]	0,0410 (5)	0,0133 (3)	0,00144 (4)	0,000108 (5)
Устная речь [10]	0,0394 (6)	0,0143 (1)	0,00149 (3)	0,000098 (6)
Электроника [16]	0,0330 (7)	0,0075 (6)	0,00222 (2)	0,000145 (1)

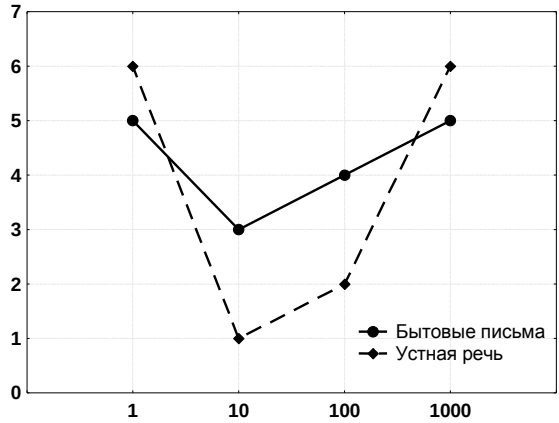

Рис. 1. Рисунок профиля словарей устной речи (по корпусу ОРД) и бытовых писем

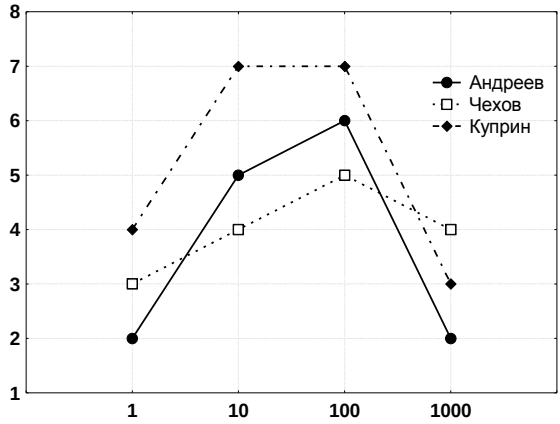

Рис. 2. Профили авторских словарей (Л.Н.Андреев, А.П.Чехов, А.И.Куприн)

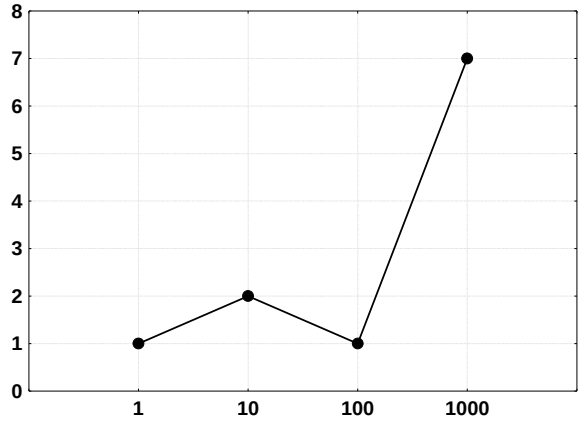

Рис. 3. Профиль словаря штабных документов (ограниченный подъязык)

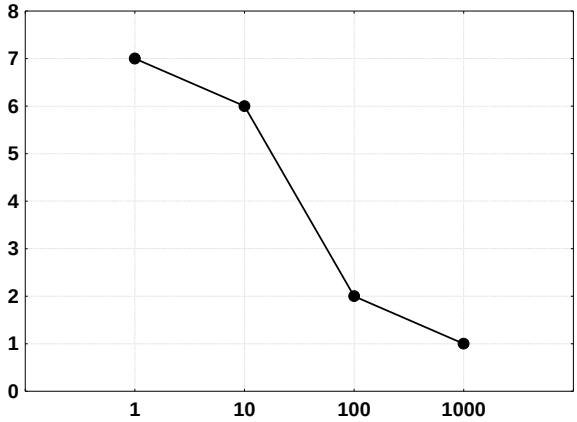

Рис. 4. Профиль словаря по электронике (специализированный язык)

5. Заключение

В статье рассмотрены некоторые порядковые статистики частотных словарей, представленных в виде ранговых распределений. На материале корпуса расшифровок русской устной спонтанной речи со статистической достоверностью установлено, что статистики, занимающие в ранжированном ряду определенное место, практически не зависят от объема выборки. С другой стороны, на материале разных, но однородных по составу корпусов, эти порядковые статистики отличаются, что демонстрирует их высокий стилеразличающий потенциал.

Полученные результаты позволяют предположить, что для сравнения частотных словарей разного объема с помощью порядковых статистик нет нужды, чтобы эти объемы были достаточно велики. Более того, начиная с некоего небольшого объема выборки[1], размеры сравниваемых по структуре словарей практически не имеют значения. Видится целесообразной апробация состоятельности данных статистик на других корпусных данных, чтобы оценить перспективы использования порядковых статистик для сравнения структуры частотных словарей разного объема и разных языковых жанров.

Литература

[1] Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Методы моделирования и первичная обработка данных. Справочное издание. М.: Финансы и статистика, 1983. 471 с.

[2] Нешитой В.В. Математические модели роста словаря и информационных потоков // Уч. записки Тартуск. ун-та. Вып. 872. Квантитативная лингвистика и автоматический анализ текстов. Тарту, 1989. С. 83–102.

[1] Можно предположить, что уже 40000-ный объем словаря в рамках однородной совокупности будет достаточен для оценки его структуры и ее сравнения с другими.

[3] Тулдава Ю.А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста // Учен. Зап. Тартуского ун-та. Вып. 549. Лингвостатистика и квантитативные закономерности текста. Тарту, 1986. С. 139–162.
[4] Мартыненко Г.Я. Основы стилеметрии. Л.: Изд-во Ленингр.ун-та, 1988. – 165 с.
[5] Мартыненко Г.Я., Мартинович Г.А. Многопараметрический статистический анализ результатов ассоциативного эксперимента. СПб: Изд-во С.-Петерб. ун-та, 2003. 26 с.
[6] Martynenko G. Statistical Consistency of Keywords Dictionary Parameters // Content-Based Multimedia Information Access: Conference Proceeding. Vol. 2. Paris, 2000. P. 1541–1547.
[7] Гребенников А.О. О состоятельности статистик частотного словаря художественной прозы // Структурная и прикладная лингвистика. Вып. 5, СПб: Изд-во С.-Петерб. ун-та, 1998. С. 110–123.
[8] Богданова-Бегларян Н.В., Шерстинова Т.Ю., Баева Е.М., Блинова О.В., Мартыненко Г.Я., Ермолова О.Б., Рыко А.И. и др. Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография. СПб.: Издательство "Лайка", 2016. 244 с.
[9] Богданова-Бегларян Н.В., Шерстинова Т.Ю., Блинова О.В., Мартыненко Г.Я. Корпус "Один речевой день" в исследованиях социолингвистической вариативности русской разговорной речи // Анализ разговорной русской речи (АР^3-2017): Труды седьмого междисциплинарного семинара. СПб.: Политехника-принт, 2017. С. 14–20.
[10] Косарева Е.О., Мартыненко Г.Я. Отношение текст-словарь в повседневной устной речи // Структурная и прикладная лингвистика. Вып. 11. СПб: Изд-во С.-Петерб. ун-та, 2015. С. 220–228.
[11] Колгушкин А.Н. Лингвистика в ваенном деле (разработка и использование частотных словарей военной лексики). М.: Военное издательство Министерства Обороны, 1970.
[12] Частотный словарь Л.Н.Андреева. Под ред. Г.Я.Мартыненко. Автор-составитель А.О.Гребенников. СПб.: Изд-во С.-Петерб. Ун-та, 2003. 398 с.
[13] Частотный словарь А.П.Чехова. Под ред. Г.Я.Мартыненко. Автор-составитель А.О.Гребенников. Спб.: Изд-во С.-Петерб. Ун-та, 1999. 172 с.
[14] Частотный словарь рассказов А.И.Куприна. Под ред. Г.Я.Мартыненко. Авторы-составители: А.О.Гребенников, Н.А.Данилова. Издательство СПбГУ, 2012.
[15] Алексеева А.М. Лексика частных писем. Учебные материалы по русской некодифицированной речи. Составители: П.М.Алексеев, А.С.Григорьева. Л.: ЛГПИ им. А.А.Герцена, 1981.
[16] Калинина Е.А. Изучение лексико-статистических закономерностей на основе вероятностной модели // Статистика речи. Л.: Наука, 1968.

On the consistency of order statistics of word frequency lists

G. Ya. Martynenko

St. Petersburg State University

The article considers order statistics of frequency dictionaries presented in the form of rank distributions. These statistics are analyzed from the point of view of their consistency (i.e., whether their empirical data converges to some limiting theoretical values as the sample size increases). The consistency of ordinal statistics were tested on the data of ORD speech corpus. Frequency word list for this corpus was compiled in the form of a dynamic structure consisting of 10 fragments of 10 thousand words each. These fragments are sequentially joined each other in a random order. At each step, a rank distribution was constructed. As a result of the conducted research, it was proved that these order statistics are indeed consistent. The shares of the corresponding order statistics for seven different frequency word lists, as well as the "order" profiles for each of these dictionaries are given. On the material of different, but homogeneous corpora, it was found that these statistics have high potential style differentiating capacity. This property of order statistics can be used for comparing frequency word lists of different themes, genres and sizes.

Keywords: frequency dictionaries, everyday spoken language, statistical consistency, statistical scales, sampling, dynamic series, rank distribution, uniform distribution.