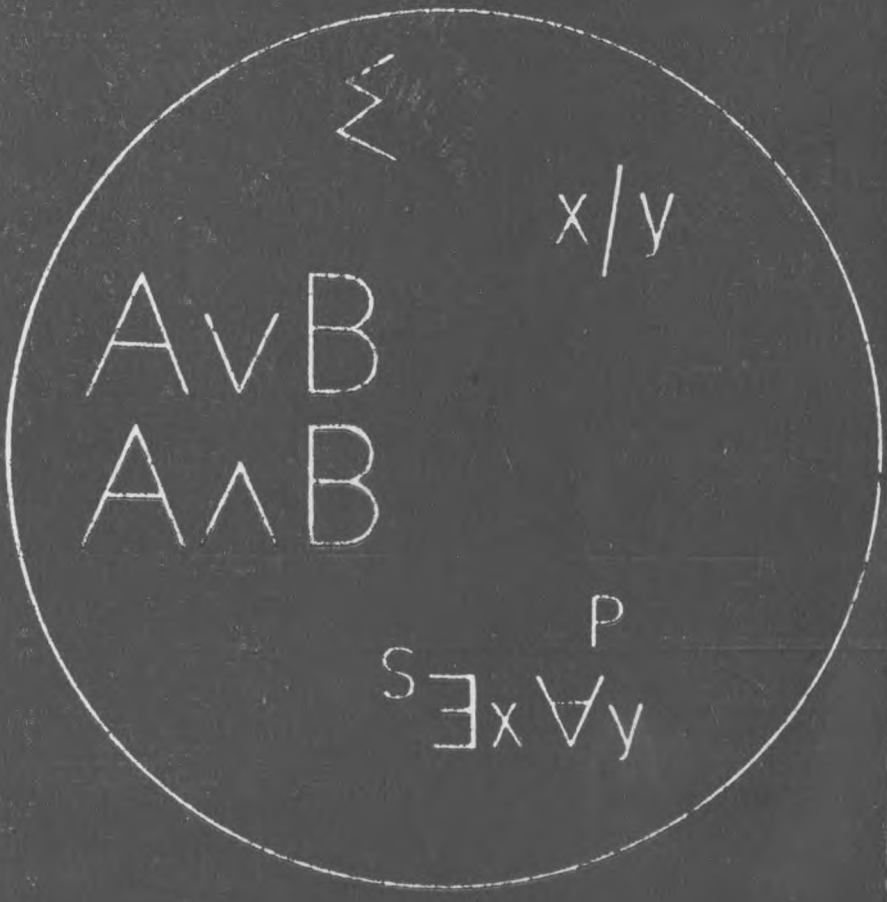
Структурная и прикладная лингвистика

3



ЛЕНИНГРАДСКИЙ ОРДЕНА ЛЕНИНА
И ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени А. А. ЖДАНОВА

СТРУКТУРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

Межвузовский сборник

Выпуск 3

Под редакцией А. С. Герда

ЛЕНИНГРАД
ИЗДАТЕЛЬСТВО ЛЕНИНГРАДСКОГО УНИВЕРСИТЕТА
1987

Печатается по постановлению
Редакционно-издательского совета
Ленинградского университета

Статьи сборника «Структурная и прикладная лингвистика» охватывают широкий круг проблем современной теоретической и прикладной лингвистики. В них рассматриваются актуальные вопросы семантики и синтаксиса предложения, применения математических методов в языкознании, проблемы лингвистической статистики. Сборник содержит статьи, посвященные разработке лингвистического обеспечения автоматических систем обработки текста.

2-й выпуск сборника «Структурная и прикладная лингвистика» вышел в 1983 году.

Сборник предназначен для филологов, специалистов по прикладной и математической лингвистике.

Редакционная коллегия: докт. филол. наук *В. В. Богданов*; докт. филол. наук. *Л. В. Бондарко;* докт. филол. наук *А. С. Герд* (отв. редактор); докт. филол. наук *Б. Ю. Городецкий.*

Рецензент докт. педагогических наук *А. В. Соколов.*

ИБ № 2522

Структурная и прикладная лингвистика

Межвузовский сборник

Выпуск 3

Редактор *Н. А. Никитина* Художественный редактор *А. Г. Голубев*
Технический редактор *А. В. Борщева*
Корректоры *В. А. Латыгина, Т. Г. Павлова*

Сдано в набор 04.06.86. Подписано в печать 22.12.86. М-26670. Формат 60×90$^{1}/_{16}$. Бумага тип. № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 12,0+0,25 вкладка. Усл. кр.-отт. 12,25. Уч.-изд. л. 13,7. Тираж 1013 экз. Заказ № 379. Цена 2 р. 10 к.
Издательство ЛГУ им. А. А. Жданова. 199034, Ленинград, Университетская наб., 7/9.

Типография Изд-ва ЛГУ им. А. А. Жданова. 199034, Ленинград, Университетская наб., 7/9.

С $\frac{1503000000-028}{076(02)-87}$ 46—87

© Издательство Ленинградского университета, 1987 г.

II

Г. Я. Мартыненко, Т. К. Чарская

К ВОПРОСУ О ГАУССОВОСТИ И НЕГАУССОВОСТИ ЛИНГВОСТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

О негауссовых представлениях в математической статистике и ее приложениях. Одна из основных целей, которые ставит перед собой лингвист, приступая к статистической обработке исходных данных, заключается в том, чтобы добиться удобной лаконичности описания свойств исследуемой совокупности, т. е. представления множества обрабатываемых данных в виде сравнительно небольшого числа обобщающих характеристик, построенных на основании исходных данных. Эти характеристики являются функциями от исходных результатов наблюдения или эксперимента и называются статистиками. К ним относятся все выборочные характеристики генеральной совокупности — различные виды счетных и распределительных средних, показатели вариации, коэффициенты асимметрии, эксцесса, ковариации, корреляции и т. п. Все статистики рассматриваются в качестве приближенных значений неизвестных параметров генеральной совокупности, т. е. являются статистическими оценками последних.

К статистическим оценкам предъявляются определенные математические требования, обеспечивающие их надежность и практический смысл, которые формулируются обычно с помощью свойств состоятельности, несмещенности и эффективности, вытекающих из условий действия закона больших чисел, определяемых обобщенной теоремой Чебышева и центральной предельной теоремой Ляпунова.

Возникает вопрос, в какой мере отвечают перечисленным свойствам статистические оценки, вычисляемые на основе эмпирических распределений лингвистических единиц. Этот вопрос представляется крайне важным в теоретическом и практическом отношении, поскольку в последнее время все чаще высказывается мнение, что аппарат классической (гауссовой) статистики непригоден для исследования социальных явлений. Не-

гауссовость распределений социальных явлений усматривается в том, что они имеют более длинные «хвосты», чем распределения, встречающиеся в естественных науках, и в силу этого моменты этих распределений, с одной стороны, аномально велики, а с другой — аномалия тем значительней, чем больше объем выборки. При этом отмечается, что статистический аппарат, пригодный для описания социальных явлений, связан с распределением Ципфа в отличие от статистики природных явлений, основанной на распределении Гаусса.[1] Математической основой негауссовой статистики является предельная теорема Б. В. Гнеденко — В. Деблина, определяющая условия, при которых предельное распределение статистик существует, но отличается от распределения Гаусса. Оказывается, что для этого необходимо и достаточно, чтобы распределение статистик при больших значениях переменных имело форму распределения Ципфа с $\alpha<2$ или отличалось от него на множитель, являющийся медленно меняющейся функцией.[2]

Зависимость статистик негауссовых распределений от объема выборки ставит под сомнение использование метода моментов при оценивании параметров генеральной совокупности. Однако среди исследователей нет единодушия как в отношении выбора параметров, подлежащих оцениванию, так и в отношении методов оценивания. Вместо моментов предлагается оценивать параметры, слабо зависящие от объема выборки (показатель ципфового распределения γ, квантили этого распределения, его энтропию,[3] среднюю геометрическую[4] и др.). Однако бесспорного доказательства независимости этих оценок от объема выборки до сих пор нет. Более того, и по сей день достаточно строго не исследован характер зависимости эмпирических моментов негауссовых совокупностей от объема выборочного наблюдения.

Что касается категорической антиномии «природа гауссова — социальные явления негауссовы», то в ней просматривается несколько спорных моментов. Во-первых, возрождается устаревшее противопоставление наук о природе наукам о человеке, характерное для неокантианской школы и с позиций статистики убедительно опровергнутое А. А. Чупровым еще в начале XX столетия.[5] Во-вторых, среди распределений природных явлений встречаются «длиннохвостые» распределения, которые в принципе могут быть отнесены к негауссовым: распре-

[1] Яблонский А. И. Стохастические модели научной деятельности.— В кн.: Системные исследования. Ежегодник, 1975. М., 1976, с. 5—42.

[2] Гнеденко Б. В. К теории предельных теорем для сумм независимых случайных величин. — Изв. АН СССР. Сер. мат. 1939, с. 181—232, 643—657.

[3] Хайтун С. Д. Наукометрия: Состояние и перспективы. М., 1983.

[4] Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. М., 1980.

[5] Чупров А. А. Очерки по теории статистики. СПб., 1909.

деление химических элементов в земной коре по содержанию,[6] распределение малых космических тел по массе,[7] космических частиц по энергии[8] и т. п. В-третьих, социальные явления очень часто описываются гауссовыми распределениями. Например, в лингвистике накоплен довольно значительный багаж гауссовых аппроксимаций. В-четвертых, вопрос об адекватности закона Ципфа экспериментальным данным является в высшей степени дискуссионным. Хотя ни одному исследователю не удалось достигнуть достаточно корректной ципфовской аппроксимации, в исследовательской среде сформировалась парадигма безусловной веры в закон Ципфа, во многом созвучная с гауссовской парадигмой, сложившейся в начале XX в.[9]

Список трудных вопросов, возникающих при обсуждении гауссовости или негауссовости реальных распределений, конечно, намного длиннее приведенного нами. Значительная их часть должна решаться математиками. Что касается экспериментаторов, то они должны не только заботиться об эмпирической базе теоретических исследований, но и проявлять интерес к самим этим исследованиям. При этом важно иметь в виду, что многие спорные проблемы могут быть решены (хотя бы частично) средствами элементарной статистической техники с опорой на качественную природу исследуемых совокупностей и варьирующих признаков.

В данной работе решаются в основном две задачи:

1) уточнение качественной природы лингвостатических распределений, относимых к разряду негауссовых, и выяснение причин «длиннохвостости» и аномально большой вариации этих распределений;

2) исследование характера зависимости статистик распределений обсуждаемого типа от объема выборки и разыскания статистик, наименее чувствительных к объему выборки.

---

[6] БСЭ. 3-е изд. 1971, т. 6, с. 330—337.

[7] Идлис Г. М. Космические силовые поля и некоторые вопросы структуры и эволюции галактической материи. — Изв. АФИ АНКазССР, 1957, т. 4, с. 3—159.

[8] Fermi E. On the origin of the cosmic radiation. — Phys. Rev., 1949, vol. 75, N 8, p. 1169—1174.

[9] В этой связи нам кажется уместным привести высказывания известных математиков о законе Гаусса и законе Ципфа. В 1912 г. А. Пуанкаре писал: «Каждый уверен в справедливости нормального закона: экспериментаторы потому, что они думают, что это математическая теорема; математики — потому, что они думают, что это экспериментальный факт» (цит. по: Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М., 1983, с. 70). В 1961 г. Г. Хердан почти дословно высказал ту же мысль, но уже применительно к закону Ципфа: «Математики верят в него потому, что лингвисты постановили считать его лингвистическим законом, а лингвисты, со своей стороны, верят в него потому, что математики постановили считать его математическим законом» (Herdan G. The calculus of linguistic observation. The Hague, 1962).

Статусные распределения лингвистических единиц и их неоднородность. В последнее время в лингвистической статистике серьезное внимание уделяется классификации лингвостатистических распределений в зависимости от типа единиц, образующих совокупности, и качественной природы признака, по которому упорядочиваются элементы совокупности.[10] Остановимся на некоторых типах распределений, релевантных с точки зрения решаемых нами задач.

1. С точки зрения противопоставления «элемент — совокупность элементов» выделятся два типа распределений: о д н о п р е д м е т н ы е и м н о г о п р е д м е т н ы е. Распределения первого типа описывают активность одного и того же элемента (например, слова) в совокупности ситуаций (например, в множестве текстов или отрезков текста); распределения второго типа описывают активность разных элементов (например, множество слов в одном и том же тексте) относительно одной и той же ситуации (например, конкретного текста).

2. При построении лингвостатических распределений могут быть использованы две контрастные категории варьирующих признаков: а) признаки, отражающие внутреннюю структуру единиц совокупности; б) признаки, указывающие на статус (функциональный вес) единиц, образующих совокупность. В соответствии с этими категориями признаков можно говорить о двух содержательных типах распределений: с т р у к т у р н ы х и с т а т у с н ы х. Текстовые словоформы могут быть упорядочены, например, по их размеру (числу букв, морфем, фонем, слогов), т. е. по признаку, отражающему внутреннюю структуру словоформы. В этом случае совокупность словоформ образует структурное распределение. Но та же совокупность словоформ может быть упорядочена по функциональному признаку, например по их встречаемости в тексте. В этом случае образуется статусное распределение.

3. Вопрос формирования однородной в качественном отношении совокупности в большинстве случаев решается на стадии планирования эксперимента или наблюдения путем качественного анализа совокупности с последующей ее стратификацией или рандомизацией на качественно однородные группы. В результате строятся несколько распределений, состоящих из однородных в качественном отношении единиц. Однако в ряде случаев качественная специфика лингвистических объектов такова, что принадлежность конкретного объекта к данной подсовокуп-

---

[10] А л е к с е е в П. М. Квантитативная типология текста: Автореф. докт. дис. Л., 1977; М а р т ы н е н к о Г. Я. Типология лингвостатических распределений. — В кн.: Лингвостатистика и вычислительная лингвистика. Тарту, 1982, с. 103—120 (Учен. зап. Тартуского ун-та, вып. 628); Т у л д а в а Ю. А. О теоретико-методологических основах квантитативно-системного анализа лексики (3): методика исследования. — В кн.: Linguistica. Тарту, 1982, с. 123—143.

ности прямо не распознается, т. е. неоднородность априорно неустранима. Это дает основание говорить о двух типах распределений: распределениях с априорно устранимой и распределениях с априорно неустранимой неоднородностью.

Распределения, которые обычно считаются негауссовыми, по нашей терминологии следует отнести к многопредметным статусным распределениям с априорно неустранимой неоднородностью (именно такие распределения строятся для описания структуры частотных словарей различного типа). Первые два признака (многопредметность и статусность) очевидны, третий (априорно неустранимая неоднородность) нуждается в обосновании. Следует подчеркнуть, что отнесение распределений обсуждаемого типа к разряду распределений с априорно неустранимой неоднородностью является принципиальным в следующем аспекте: если такие совокупности действительно неоднородны, то распределение их единиц не может быть в принципе описано какой-либо единой математической моделью, независимо от того, принадлежит она к гауссову или же к негауссову семейству теоретических законов распределения.

Итак, необходимо проверить справедливость гипотезы о неоднородности статусных распределений. В качестве экспериментального материала был использован частотный словарь предикатных слов, составленный одним из авторов настоящей работы на корпусе текстов отечественных патентных документов.[11] Ранговое распределение этих слов, построенное на выборке в 50 тыс. словоупотреблений, приведено в табл. 1.

В теории статистики практически нет четко разработанных критериев, позволяющих решить, однородна совокупность в качественном отношении или нет. Обычно исследователи руководствуются двумя симптомами неоднородности: 1) величиной коэффициента вариации: чем больше коэффициент вариации, тем больше шансов на то, что совокупность неоднородна; 2) внешним видом графика эмпирического распределения: если кривая распределения многовершинна, то почти наверняка исходная совокупность состоит из качественно неоднородных подсовокупностей.

Рассмотрим, каким образом проявляются эти симптомы в распределении, приведенном в табл. 1.

Коэффициент вариации этого распределения равен 528%. В теории статистики господствует убеждение, что коэффициент вариации, равный 50% и более, говорит о том, что совокупность неоднородна. С этой точки зрения наша совокупность неоднородна в высшей степени, она аномально неоднородна.

---

[11] Чарская Т. К. Статистическая структура текста патентного документа: Канд. дис. Л., 1983, ЛГУ.

Таблица 1. **Ранговое распределение предикатных слов в выборке объемом 50 тыс. словоупотреблений**

Ранг	Частота	Число повторений	Ранг	Частота	Число повторений
1	6530	1	58	66	1
2	5931	1	59	62	1
3	4675	1	60	59	1
4	3797	1	61—62	58	2
5	3168	1	63	57	1
6	2738	1	64—65	56	2
7	2310	1	66	55	1
8	2291	1	67	53	1
9	1850	1	68—69	49	2
10	1603	1	70	48	1
11	1289	1	71	46	1
12	1187	1	72—75	45	4
13	1126	1	76	44	1
14	690	1	77—78	42	1
15	498	1	79	41	1
16	389	1	80	40	1
17	362	1	81	39	1
18	346	1	82—83	36	2
19	327	1	84—85	35	2
20	325	1	86	34	1
21	314	1	87	33	1
22	277	1	88	32	1
23	220	1	89—91	31	3
24	217	1	92	30	1
25	201	1	93	29	1
26	198	1	94	28	1
27	185	1	95—100	27	6
28	174	1	101	26	1
29	156	1	102—103	25	2
30—31	155	2	104—107	24	4
32	149	1	108	23	1
33	143	1	109	22	1
34	141	1	110	21	1
35	136	1	111—114	20	4
36	127	1	115—118	19	4
37	121	1	119—122	18	4
38	118	1	123—127	17	5
39—40	117	2	128	16	1
41	111	1	129—134	15	6
42	110	1	135—138	14	4
43—44	107	2	139—140	13	2
45	94	1	141—146	12	6
46—47	92	1	147—152	11	6
48	88	1	153—166	10	14
49	83	1	167—175	9	9
50	81	1	176—187	8	12
51	80	1	188—198	7	11
52	76	1	199—216	6	18
53	74	1	217—232	5	16
54	72	1	233—253	4	21
55	71	1	254—283	3	30
56	68	1	284—359	2	76
57	67	1	360—465	1	106

Что касается геометрического критерия, то здесь он тоже срабатывает довольно однозначно. Кривая рангового распределения предикатных слов, показанная на рис. 1, имеет «двояковыпуклый» контур, четко выраженный при разных объемах выборки, начиная от минимальной (200 словоупотреблений) и кончая максимальной (50 тыс. словоупотреблений).

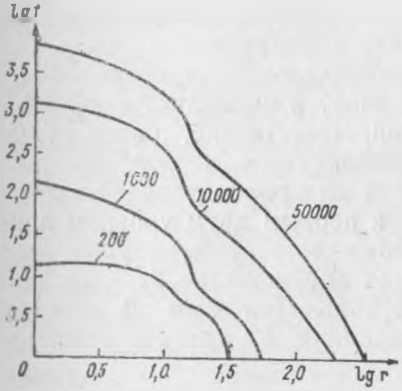


Рис. 1. Ранговые распределения предикатных слов при разных объемах выборки.

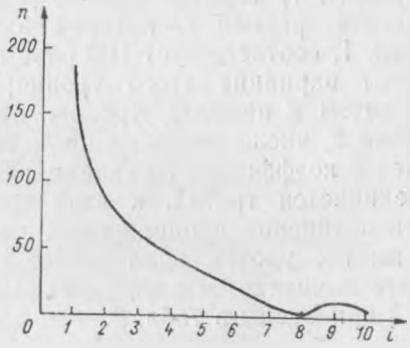


Рис. 2. Дифференциальное распределение предикатных слов.

Естественно предположить, что каждая выпуклость кривой распределения описывает один из двух фрагментов словаря: подмножество высокочастотных и подмножество низкочастотных слов.

Чтобы лишний раз убедиться в неоднородности распределения предикатных слов, мы построили также кривую частотного (дифференциального) распределения. При этом были решены технические трудности, связанные с тем, что распределение характеризуется чрезвычайно широким вариационным размахом, с крайне неравномерным распределением численности единиц на его различных участках. Статистики в таких случаях пользуются неравными интервалами, принцип организации которых в значительной степени произволен. Для примерной ориентации в выборе числа интервалов (s) можно воспользоваться формулой Стерджесса

$$s \approx \log_2 n + 1,$$

где n — объем выборки.

Применительно к распределению, приведенному в табл. 1, s оказалось равным 9,86 (или 10 с точностью до единицы). Далее все варианты были прологарифмированы и вся область изменения аргумента, измеренная в логарифмической шкале, была разделена на 10 равных интервалов, которым соответствуют неравные интервалы в «декартовой» шкале.

График частотного распределения показан на рис. 2, из которого видно, что кривая распределения имеет достаточно откровенную бимодальную форму.

Для получения полной уверенности в неоднородности совокупности предикатных слов мы дополнительно воспользовались процедурой, основная идея которой заимствована у В. В. Сиськова.[12] Этот прием можно условно назвать методом скользящего коэффициента вариации. Смысл его заключается в следующем: 1) берется первый уровень вариационного ряда (т. е. частота, равная 1, которой в распределении, приведенном в табл. 1, соответствует 106 разных слов) и вычисляется коэффициент вариации этого уровня; он, естественно, равен нулю; 2) затем к первому уровню присоединяется второй (частота равна 2, число разных слов — 76) и для обоих уровней вычисляется коэффициент вариации; 3) к первым двум уровням присоединяется третий, к этой тройке — четвертый, и процедура присоединения продолжается до тех пор, пока не будут исчерпаны все уровни ряда (всего их 106). При этом на каждом шаге вычисляется коэффициент вариации. Результаты вычислений приведены в табл. 2.

Таблица 2. **Расчетные данные «скользящего» коэффициента вариации**

Номер шага	$v_{ск}$	Номер шага	$v_{ск}$
1	0,000	55	1,452
5	0,589	60	1,508
10	0,823	65	1,584
15	0,927	70	1,696
20	1,015	75	1,772
25	1,077	80	1,871
30	1,143	85	2,013
35	1,200	90	2,257
40	1,262	95	3,263
45	1,337	100	4,339
50	1,404	105	5,156

Соответствующий график показан на рис. 3, из которого видно, что скользящий коэффициент вариации ведет себя двойственно: сначала он возрастает по восходящей кривой, затем резко замедляет свой рост, после чего столь же резко начинает возрастать. Если бы распределение было качественно однородным, то возрастание коэффициента вариации происходило бы по простейшему закону без всяких (и тем более резких) изме-

[12] Сиськов В. В. Корреляционный анализ в экономических исследованиях. М., 1975.

нений кривой распределения. По-видимому, из приведенных нами аргументов в пользу неоднородности статусных распределений этот является наиболее сильным. В данном случае аномально высокая вариация признака обусловлена тем, что в итоговом распределении совмещены два распределения с резко отличными математическими ожиданиями. Проведем элементарные расчеты. Из рис. 1—3 видно, что резкое изменение геометрии кривых наблюдается в промежутке между 10-м и 20-м рангами. В качестве первого приближения можно допустить, что нижней границей высокочастотной зоны является 15-й ранг. Для оценки уровня межгрупповой и внутригрупповой вариации воспользуемся правилом сложения дисперсий:

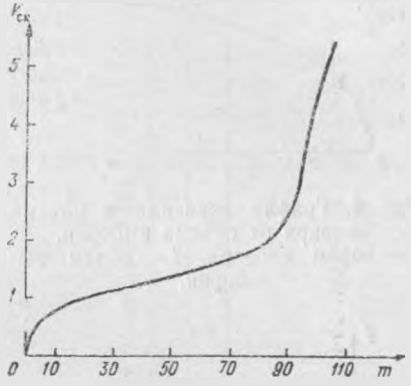


Рис. 3. График зависимости «скользящего» коэффициента вариации от числа последовательно присоединяемых вариант дифференциального распределения.

$$\sigma^2 = \delta^2 + \bar{\sigma}^2,$$

σ^2 — общая дисперсия, $\bar{\sigma}^2$ — средняя групповых дисперсий, а δ^2 — межгрупповая дисперсия. Расчеты показали, что в общей дисперсии основную массу составляет межгрупповая дисперсия. При этом эмпирическое корреляционное отношение, вычисляемое по формуле

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}},$$

оказалось равным 0,82. Это означает, что фактор неоднородности играет ведущую роль в формировании аномально высокой вариации признака. Что касается внутригрупповой вариации, то она сравнительно невелика: для высокочастотной зоны коэффициент вариации равен 0,69, а для низкочастотной — 1,15. Из этого следует, что каждая из двух подсовокупностей в отдельности может быть в принципе аппроксимирована какой-либо функцией, каждая из которых принадлежит к гауссовому семейству.[13] Это, в свою очередь, означает, что принадлежность распределений обсуждаемого типа к семейству негауссовых распределений не является бесспорной.

---

[13] Мартыненко Г. Я. Некоторые закономерности концентрации и рассеяния элементов в лингвистических и других сложных системах. — В кн.: Структурная и прикладная лингвистика, вып. 1. Л., 1978, с. 63—79.

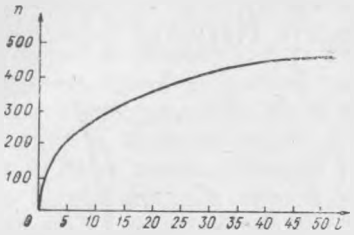


Рис. 4. График зависимости объема словаря от объема выборки.
n — объем словаря, L — объем выборки.

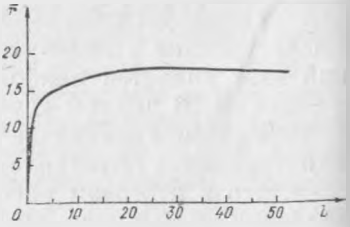


Рис. 7. Зависимость ранговой средней от объема выборки.
$\bar{r}$ — ранговая средняя, L — объем выборки.

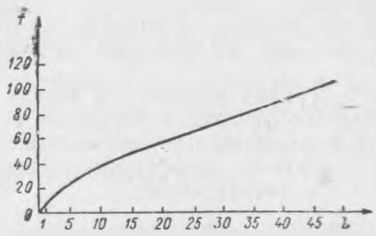


Рис. 5. График зависимости средней частоты от объема выборки.
f — средняя частота, L — объем выборки.

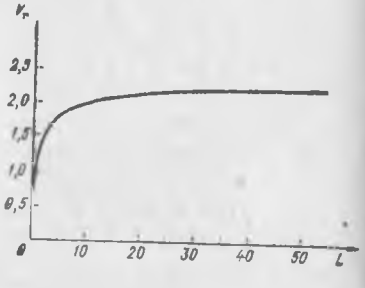


Рис. 8. Зависимость рангового коэффициента вариации от объема выборки.
V_r — ранговый коэффициент вариации, L — объем выборки.

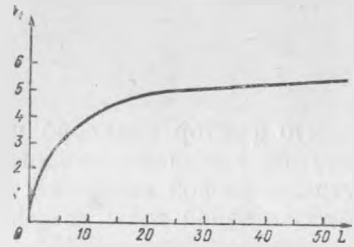


Рис. 6. График зависимости частотного коэффициента вариации от объема выборки.
V_f — частотный коэффициент вариации, L — объем выборки.

Зависимость статистик статусных распределений от объема выборки. Порождение речи протекает на фоне двух тенденций: стремления к расширению состава разноименных единиц (увеличение разнообразия) и стремления к его сокращению (ограничение разнообразия).[14] Этим противоборствующим тенденци-

[14] Тулдава Ю. А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. — В кн.: Лингвостатистика и квантитативные закономерности текста: Труды по лингвостатистике, вып. 6. Тарту, 1980, с. 113—144.

ям соответствуют два типа количественной вариативности лингвистических единиц: вариативность состава (номенклатурная колеблемость знакотипов) и рекуррентная вариативность (колеблемость числа знакоупотреблений каждого знакотипа). Распределения, связывающие накопляемое количество знакотипов с накопляемым числом знакоупотреблений, принято называть распределениями знакотип — знакоупотребление.[15] В статистической лексикографии этот тип распределения используется для изучения характера зависимости объема словаря и индекса разнообразия (средней частоты) от объема текста или выборки.

В нашем исследовании система показателей значительно шире. В дополнение к упомянутым показателям — объему словаря (n) и средней частоте ($\bar{f}$) — мы использовали также коэффициент вариации по частоте (v_f), а также два ранговых показателя: средний ранг ($\bar{r}$) и коэффициент вариации по рангу (v_r).

Зависимость пяти показателей от объема выборки исследовалась на макровыборке в 50 тыс. словоупотреблений, состоящей из 10 микровыборок равного объема, последовательно присоединяемых друг к другу. Числовые данные, отражающие динамику всех показателей, приведены в табл. 3, а соответствующие графики на рис. 4—8.

Из рис. 4—8 видно, что все показатели по мере увеличения объема выборки возрастают с затухающей скоростью, при этом, однако, не совсем ясно, стремятся ли они к фиксированному верхнему пределу или нет.

Поскольку поведение всех показа-

Таблица 3. Зависимость статистик частотного словаря от объема выборки

Статистики	Объем выборки, тыс. словоупотреблений										
	1	5	10	15	20	25	30	35	40	45	50
n	105	221	266	321	361	398	417	433	447	456	465
$\bar{f}$	9,52	22,62	37,59	46,73	55,40	62,81	71,94	80,83	89,48	98,68	107,53
v_f	2,34	3,61	3,99	4,36	4,66	4,86	4,96	5,06	5,13	5,20	5,28
$\bar{r}$	13,13	16,15	15,84	16,97	17,43	18,12	18,22	18,06	18,41	18,12	17,94
v_r	1,41	1,88	1,98	2,07	2,14	2,01	2,20	2,21	2,24	2,25	2,21

[15] Carroll J. B. On sampling from lognormal model of word-frequency distribution.— In: Kučera H., Francis W. W. Computational analysis of present-day American English. Providence, 1967.

телей в динамическом ряду зависит не только от объема выборки, но и от объема словаря, необходимо в первую очередь выяснить характер зависимости объема словаря от объема выборки.

Естественно предположить, что объем словаря конечен. Проверку этой гипотезы можно осуществить путем построения теоретической функции, аппроксимирующей данные наблюдения. При этом одновременно можно решить задачу прогнозирования максимального объема словаря и установить, какой объем выборочного наблюдения нужно произвести, чтобы список разных слов был близок к тому, который характерен для генеральной совокупности.

В качестве аппроксимирующих были выбраны три функции асимптотического роста:

$$\text{степенная} — y = K - Kx^{-d},$$
$$\text{экспоненциальная} — y = K - Ke^{-cx},$$
$$\text{функция Вейбулла (комбинация степенной и экспоненциальной)} — y = K - Ke^{-(cx)d}.$$

Данные наблюдения (см. табл. 3, строка 1) выравнивались перечисленными функциями с помощью метода наименьших квадратов. Применение этого метода значительно облегчается тем, что каждая из этих функций путем однократного или повторного логарифмирования может быть преобразована в линейную. Линейные варианты перечисленных функций имеют вид:

$$\ln(K - y) = \ln K - d \ln x \text{ (степенная функция)},$$
$$\ln(K - y) = \ln K - cx \text{ (экспоненциальная функция)},$$
$$\ln \ln(K - y) = \ln \ln K - d \ln cx \text{ (функция Вейбулла)}.$$

Известно, что система нормальных уравнений для линейной зависимости очень проста и ее решение относительно неизвестных параметров не вызывает особых затруднений. Однако в нашем случае один из параметров (асимптота) входит в состав зависимой переменной, а это исключает использование метода наименьших квадратов в его чистом виде. Нами был найден следующий выход из сложившейся ситуации. Асимптоте мы задавали конкретные значения через шаг, равный пяти единицам, начиная с максимального объема словаря, достигнутого в эмпирическом ряду. Далее, на каждом шаге при фиксированном значении асимптоты с помощью метода наименьших квадратов вычислялись значения других параметров и на каждом же шаге с помощью критерия χ^2 проверялась гипотеза об отсутствии существенных различий между эмпирическим и теоретическим распределением при пятипроцентном уровне значимости. Если функция ни на одном шаге не обнаруживала согласия с эмпирическими данными (а именно это произошло со степенной и экспоненциальной функциями), то мы прекращали работу и переходили к другой функции и искали затем область

значений, удовлетворяющую требованиям критерия χ^2. Функция Вейбулла в определенном интервале в этом смысле обнаружила удовлетворительное согласие с эмпирическим распределением. Зависимость величины критерия χ^2 от величины максимального объема словаря показана на рис. 9. Здесь же показан интервал допустимых значений максимального объема словаря (от 225 до 825 слов), а также оптимальный для эмпирического ряда максимальный объем словаря (575 слов).

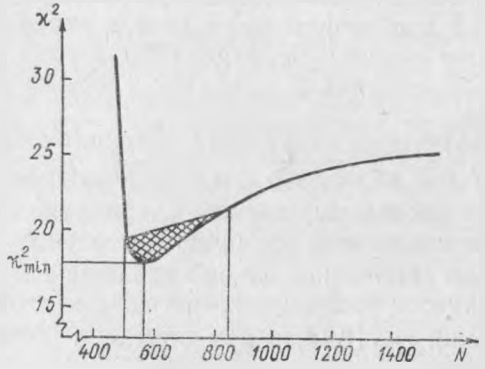


Рис. 9. Зависимость величины критерия x^2 от величины верхнего предела объема словаря (K).

Таким образом, рабочая гипотеза о конечности объема словаря не противоречит опытным данным.

Оптимальная формула, описывающая зависимость объема словаря от объема выборки, имеет вид:

$$N = 575 - 575e^{-0{,}18L^{0{,}51}}$$

По этой формуле можно приблизительно рассчитать максимальный объем выборки, при котором число разных слов будет ничтожно мало отличаться от предельного объема словаря. Сформулируем, например, такой вопрос: при каком объеме выборки накопленный объем словаря составляет 99% от полного словаря, то есть меньше последнего на шесть единиц. Подставим в эмпирическую формулу, приведенную выше, вместо N накопленное число разных слов (569) и решим ее относительно L. Получаем: $L = 398$ тыс. словоупотреблений.

Перейдем теперь к анализу остальных обобщающих показателей.

Из рис. 5 видно, что частотная средняя f неуклонно возрастает с тенденцией к стабилизации темпа возрастания около постоянного уровня. Это представляется естественным, так как по мере увеличения объема выборки постепенно затухающий темп роста словаря оказывает все меньшее воздействие на рост f.

Когда же будет достигнут полный объем словаря, величина частотной средней будет полностью определяться объемом выборки и далее будет расти неограниченно и линейно по мере роста объема последней. Однако ситуация меняется коренным образом при переходе к шкале частостей: средняя частость находится в строгой обратно пропорциональной зависимости от объема словаря и конечна в той же мере, в какой конечен объем словаря. Что касается частотного коэффициента вариации (v_f), то он с возрастанием объема выборки устремляется к верхнему пределу, который в конечном итоге определяется объемом словаря. Причем этот показатель безразличен к тому, в какой шкале измеряется варьирующий признак: в шкале частот или в шкале частостей.

Ранговые статистики ведут себя принципиально иначе. Из табл. 3 и рис. 7 и 8 видно, что $\bar{r}$ и v_r значительно быстрее, чем $\bar{f}$ и v_f достигают предельных значений, обнаруживая почти полное безразличие и к объему выборки, и к объему словаря. Уже при сравнительно небольшом объеме выборки r и v_r стабилизируются у постоянного уровня: так, при объеме выборки в 5 тыс. словоупотреблений $\bar{r}=16{,}15$, а при объеме в 50 тыс. словоупотреблений $\bar{r}=17{,}94$.

Самобытность поведения ранговых статистик определяется тем, что они реагируют не столько на распределение частот в совокупности, сколько на концентрацию частот в верхней части частотного словаря, указывая на критическую точку, разделяющую словарь на две качественно однородные зоны.

Из сказанного следует, что при оценивании параметров генеральных совокупностей, единицы которых упорядочены по функциональному признаку, следует пользоваться ранговыми статистиками, при расчете которых в роли варьирующего признака выступают ранги, а в качестве статистических весов — частоты, соответствующие конкретным рангам.

Заканчивая статью, можно сделать следующие выводы.

1. Распределения, относимые традиционно к разряду негауссовых, с содержательной точки зрения принадлежат к классу многопредметных статусных распределений с априорно неустранимой неоднородностью. Это означает, что такие распределения в принципе не могут быть аппроксимированы единой математической моделью, независимо от того, гауссова она или не гауссова. Если каждая из подсовокупностей (зона высокочастотных и зона низкочастотных слов) описывается гауссовым законом, то вопрос о негауссовости таких распределений оказывается беспредметным.

2. При оценивании параметров статусных распределений следует пользоваться ранговыми статистиками, обнаруживающими слабую чувствительность к объему выборки и к числу разноименных элементов совокупности. При расчете этих ста-

тистик в роли варьирующего признака выступают ранги, а в роли статистических весов — частоты, соответствующие конкретным рангам.

CONTENTS

I

Bogdanov V. V., Bondarko L. V., Butorov V. D., Gerd A. S. Applied linguistics and language theory (Leningr. Univ.) 3
Bogdanov V. V. On three aspects of sentence organization as a sign object (Leningr. Univ.) 16
Kovalik Yu. N., Iljin Yu. V. On coextensiveness, equivalence of meaning and synonymy (Magnitogorsk polytechnical institute) . . . 25
Dikareva S. S. Semantics of anaphore (Simferopol Univ.) . . 29
Fitialov S. Ya. A model of a situation of denotative noun identification (Leningr. Univ.) 38
Grinbaum O. N. Some variants of denotative noun identification in Russian texts (Gossnab) 44
Zubkova T. I. Grammar as an hierarchically organized structure (Leningr. Univ.) 48
Gorodetsky B. Yu. Term and its linguistic properties (Mosk. Univ.) 54

II

Martynenko G. Ya., Charskaya T. K. To the problem of haussness and non-haussness of linguostatistic distribution (Leningr. Univ.) . 63
Turygina L. A. System approach to the modelling of formal properties of language (Leningr. Naval communications school) . . . 77
Marusenko M. A. Selection of informative parameters in the problems of stylistic diagnostics (Leningr. Univ.) 84
Manasyan N. S. Evaluation of parameters of linguistic distributions of a certain class (Yerevan polytechnical institute) 94
Malakhovsky L. V. On representation in language of the concept of "roundness" of a number (on the material of English numerals) (Institute of linguistics of the Acad. of Sci. of the USSR) 97
Vasiljeva L. R. Frequency vocabulary of terminological word-groups in the sublanguage of organic semiconductors (Lvov polytechnical institute) 104
Averina S. A., Gerd A. S., Demidov D. G., Popov M. B. Statistic data on noun suffixes in Old Slavonic manuscripts (Leningr. Univ.) . 109
Kolesov V. V. Statistic distribution of verb forms in Old Russian manuscripts of the XI—XIV centuries (Leningr. Univ.) . . . 115
Ruskova M. P. Statistic characteristics of article forms in the Bulgarian of the XVIII century (Leningr. Univ.) 131
Koltunova Ye. A. Statistic analysis of pronoun declination in oficial papers of the XIV—XVII centuries in Nizhny Novgorod (Gorky Univ.) 138

III

Kiyak T. R. Motivation as an objective criterion for the unification of terminology (Chernovtsy Univ.) 144
Zakharov V. P. Linguistic background for polythematic documentary retrieval (Leningr. Centre of Sci.-Tech. Inf.) 149
Belyaeva L. N., Matorina L. V., Piotrovsky R. G., Yatsenko T. V. Principles of machine translation of nominal terminological wordgroups (Leningr. Ped. Institute, Leningr. Institute of Avia Instrumentmaking, Dagestan Univ.) 154
Otkupshchikova M. I. The problem of uniformity in the task of man-machine dialogue (Leningr. Univ.) 165
Tisenko E. V. Representation of natural language grammatical features in an information retrieval language (Leningr. Univ.) . . . 173
Berkov V. P. L. V. Shcherba and problems of bilingual lexicography (Leningr. Univ.) 180
Kozlov Yu V. Problems of compression of terminological glossaries (Leningr. Univ.) 187

СОДЕРЖАНИЕ

I

Богданов В. В., Бондарко Л. В., Буторов В. Д., Герд А. С. Прикладная лингвистика и теория языка (Ленингр. ун-т) 3
Богданов В. В. О трех аспектах организации предложения как знакового объекта (Ленингр. ун-т) 16
Ковалик Ю. Н., Ильин Ю. В. О коэкстенсивности, смысловом тождестве и синонимии (Магнитогорск. политехн. ин-т) 25
Дикарева С. С. Семантика анафоры (Симферопольский ун-т) . 29
Фитиалов С. Я. Модель ситуации денотативного отождествления существительных (Ленингр. ун-т) 33
Гринбаум О. Н. Способы денотативного отождествления существительных в русских текстах (ВЦ Госснаба СССР) 44
Зубкова Т. И. Грамматика как иерархически организованная структура (Ленингр. ун-т) 48
Городецкий Б. Ю. Термин и его лингвистические свойства (Моск. ун-т) 54

II

Мартыненко Г. Я., Чарская Т. К. К вопросу о гауссовости и негауссовости лингвостатистических распределений (Ленингр. ун-т, Саранский ун-т) 63
Турыгина Л. А. Системный подход к моделированию формальных свойств языка (Ленингр. высшее военное училище связи) . . . 77
Марусенко М. А. Отбор информативных параметров в задачах стилистической диагностики (Ленингр. ун-т) 84
Манасян Н. С. Об оценке параметров лингвистических распределений определенного класса (Ереванск. политехн. ин-т) . . . 94
Малаховский Л. В. Об отражении в языке понятия «круглости» числа (на материале английских числительных) (Ин-т языкознания АН СССР) 97
Васильева Л. Р. Частотный словарь терминологических словосочетаний английского подъязыка органических полупроводников (Львовский политехн. ин-т) 104
Аверина С. А., Герд А. С., Демидов Д. Г., Попов М. Б. Статистика суффиксов имен существительных в памятниках древнеславянской письменности (Ленингр. ун-т) 109
Колесов В. В. Статистическое распределение глагольных форм в древнерусских источниках XI—XIV вв. (Ленингр. ун-т) . . . 115
Рускова М. П. Статистическая характеристика членных форм в болгарском языке XVIII в. (Ленингр. ун-т) 131
Колтунова Е. А. Статистический анализ местоименного склонения в деловой нижегородской письменности XIV—XVII вв. (Горьковск. ун-т) 138

III

Кияк Т. Р. Мотивированность как объективный критерий унификации терминологии (Черновицк. ун-т) 144
Захаров В. П. Лингвистическое обеспечение политематического документального поиска (Ленингр. ЦНТИ) 149
Беляева Л. Н., Маторина Л. В., Пиотровский Р. Г., Ященко Т. В. Принципы машинного перевода именных терминологических словосочетаний (ЛГПИ им. А. И. Герцена, Ленингр. ин-т авиац. приборостроения, Дагестанск. ун-т) 154
Откупщикова М. И. Проблема однородности в задаче общения пользователя с ЭВМ (Ленингр. ун-т) 165
Тисенко Э. В. Представление знаний о грамматических признаках единиц естественного языка на информационно-поисковом языке (Ленингр. ун-т) 173
Берков В. П. Л. В. Щерба и вопросы двуязычной лексикографии (Ленингр. ун-т) 180
Козлов Ю. В. Вопросы компрессии терминологических словников (Ленингр. ун-т) 187