НАУЧНЫЕ ДОКЛАДЫ

Г.Я. Мартыненко, Г.А. Мартинович

МНОГОПАРАМЕТРИЧЕСКИЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ АССОЦИАТИВНОГО ЭКСПЕРИМЕНТА

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА
2003

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Н А У Ч Н Ы Е Д О К Л А Д Ы

Г.Я. Мартыненко, Г.А. Мартинович

МНОГОПАРАМЕТРИЧЕСКИЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ АССОЦИАТИВНОГО ЭКСПЕРИМЕНТА

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА
2003

ББК 81
М29

Научный редактор серии *А.С. Герд*

Рецензенты: канд. филол. наук *А. К. Агибалов* (С.-Петерб. гос. акад. театр. искусства)
докт. филол. наук *М. А. Марусенко* (С.-Петерб. гос. ун-т)

Мартыненко Г.Я., Мартинович Г.А.
М29 Многопараметрический статистический анализ результатов ассоциативного эксперимента. — Изд-во С.-Петербургского ун-та, 2003. — 28 с.
ISBN 5-288-03107-X

Доклад посвящен актуальной и дискуссионной проблеме отбора статистических параметров для описания результатов ассоциативного эксперимента, представленных в виде частотного словаря слов-реакций на слово-стимул. Сложность данной проблемы заключается в том, что большинство таких параметров сильно зависят от объема выборки. В работе предлагается оригинальная экстраполяционная методика исследования зависимости априорно заданных параметров от объема выборки, т.е. их проверки на состоятельность в статистическом смысле. Авторы приходят к заключению, что наибольшей состоятельностью обладают энтропия, медиана по рангу, золотое сечение по рангу и др., которые при сравнительно небольшом объеме выборки практически достигают своих асимптотических значений.

Для широкого круга специалистов в области психолингвистики, экспериментальной и квантитативной лингвистики, стилометрии, статистической лексикографии.

ББК 81

ISBN 5-288-03107-X

© Г.Я. Мартыненко, Г.А. Мартинович, 2003

© Издательство С.-Петербургского университета, 2003

ВВЕДЕНИЕ

Одна из основных целей, которые ставит перед собой лингвист, приступая к обработке результатов статистического наблюдения или статистического эксперимента, заключается в том, чтобы добиться удобной лаконичности описания свойств исследуемой совокупности, т. е. представления множества обрабатываемых данных в виде небольшого числа обобщающих характеристик, построенных на основании этих данных. Эти характеристики являются функциями от результатов наблюдения или эксперимента и называются статистиками. 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. Эти требования формулируются обычно в виде свойств состоятельности, несмещенности и эффективности, вытекающих из закона действия больших чисел, определяемых обобщенной теоремой Чебышева, центральной предельной теоремой Ляпунова и рядом других теорем. В связи с этим 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 не пригоден для исследования совокупностей, которые с точки зрения классической (формальной) логики могут рассматриваться как собирательные понятия. Такие понятия, в отличие от разделительных понятий, характеризуют множества объектов, не обязательно однородных, но представляющих собой целостное образование (Мартыненко, 1988; Čebanov, Martynenko, 1998).

К совокупностям такого типа могут быть отнесены и результаты ассоциативного эксперимента, представляющие собой совокупность разнородных слов-реакций на некоторое слово-стимул. Такие совокупности организованы по принципу ассоциативного поля, целостность которого формируется с одной стороны благодаря организующей роли слова-стимула, а с другой – вследствие преобладания стереотипного начала в коллективном языковом сознании. Целью данной работы является исследование множества параметров, пригодных для изучения результатов ассоциативного эксперимента, и формирование такого комплекса параметров, который в лаконичной форме отражал бы системные свойства подобных полей, например, такие свойства, как: *разнообразие – ограничение разнообразия*, *концентрация – рассеяние*, *устойчивость – неустойчивость*, *однородность – неоднородность* и др.

1. МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ОСНОВАНИЯ ИССЛЕДОВАНИЯ

1.1. Частотный словарь как способ представления результатов ассоциативного эксперимента

Наиболее естественным способом формального результатов ассоциативного эксперимента является частотный словарь слов-реакций на слово-стимул. Как и любой частотный словарь, словарь такого типа представляет собой лексикографическое произведение, в котором каждая словарная статья содержит имя лексической единицы (словоформы или лексемы) в сопровождении различного рода статистических данных: частоты этой единицы, количества единиц с данной частотой, ранга лексической единицы и т. п. На основании информации, содержащейся в частотном словаре, могут быть построены статистические распределения определенного типа в зависимости от того, какая информация используется в качестве зависимой или независимой переменной. Основными среди них являются: полиномиальное распределение, ранговое и спектровое распределения. В полиномиальном распределении в качестве независимой переменной выступает варьирующее имя лексической единицы, а в качестве зависимой переменной – ее частота; в ранговом распределении в качестве независимой переменной выступает ранг лексической единицы, а в качестве зависимой – ее частота (имя в таком распределении «исчезает»); в спектровом распределении роль независимой переменной выполняет частота лексической единицы, а роль зависимой – число единиц, обладающих данной частотой (Нешитой, 1968; Тулдава, 1986; Мартыненко, 1989).

Детальный перечень способов систематизации лексико-статистичнских данных представлен в табл. 1. В таблице показано, что при переходе от кумулятивного распределения к ранговому значения случайной переменной и их статистические веса меняются местами: варианты становятся значениями зависимой переменной (т. е. функции), а статистические веса – значениями независимой переменной (т. е. аргумента). Обращает на себя внимание и то обстоятельство, что при переходе к ранговым распределениям накопляемые численности объектов «превращаются» в последовательность чисел натурального ряда, т. е. в ранговую последовательность.

1.2. О негауссовых представлениях в современной математической статистике

С теоретико-вероятностной точки зрения все реальные и мыслимые варианты статистических распределений распадаются на две группы: семейство гауссовых и семейство негауссовых распределений (Яблонский, 1975; Мартыненко, Чарская, 1987). К гауссовым относятся теоретические распределения, описываемые тем или иным вариантом экспоненциальной функции. В роли эталона этой группы выступает нормальное распределение, к которому сходятся все варианты гауссовых распределений (биномиальное, распределение Пуассона, показательное распределение и др.). В роли эталона второй группы выступает уравнение неравносторонней гиперболы. Это уравнение считается идеальным образцом (или асимптотикой), к которому устремляются все варианты распределений гиперболического типа (распределение Ципфа, Парето, Лотки, Мандельброта и др.)

(Шрейдер, Шаров, 1982). Важнейшей теоретической особенностью негауссовых распределений, отличающей их от распределений гауссового типа, является бесконечность моментов любого порядка, в частности дисперсии. Это означает, что в контексте негауссовых распределений традиционная теория моментов теряет смысл.

Таблица 1

Формы упорядочивания данных статистического эксперимента

Формы упорядочивания	Статистические данные	
	варианты	статистические веса
Полиномиальное распределение	Имя ЛЕ*	Частота ЛЕ
Спектровое распределение	Частота ЛЕ	Число ЛЕ с данной частотой
Спектрово-кумулятивное распределение	Частота ЛЕ	Число ЛЕ, частота которых не превышает данную
Спектрово-декумулятивное распределение [1)]	Частота ЛЕ	Число ЛЕ, частота которых равна или превышает данную
Возрастающее ранговое распределение	Число ЛЕ, частота которых не превышает данную (ранг «текущей» ЛЕ)	Частота ЛЕ
Убывающее ранговое распределение	Число ЛЕ, частота которых равна или превышает данную (ранг «текущей» ЛЕ)	Частота ЛЕ
Кумулятивное возрастающее ранговое распределение	Ранг ЛЕ	Накопленная частота ЛЕ
Кумулятивное убывающее ранговое распределение	Ранг ЛЕ	Накопленная частота ЛЕ

*ЛЕ – лексическая единица

В последние годы негауссовы распределения интенсивно исследуются в разных дисциплинах на обширном эмпирическом материале. Установлено, что эти распределения обладают следующими свойствами: 1) они характеризуют структуру совокупностей, представляющих собой собирательные понятия; 2) в таких распределениях исследуемые объекты (единицы совокупностей) упорядочиваются по признаку, указывающему на их статус (функциональный вес); 3) статистики таких распределений неуклонно возрастают по мере возрастания объема выборки (Мартыненко, 1988), т.е. являются несостоятельными в статистическом смысле. Последнее обстоятельство ставит под сомнение возможность использования традиционных методов описательной статистики, в частности использование различного рода средних, показателей колеблемости признака, характеристик формы

[1)] Термин «декумулятивный» используется в эконометрии при построении распределений, в которых значение случайной величины образуют размер дохода, а статистическими весами являются численности лиц с доходом, превышающим данный (Ланге О., 1964).

эмпирического распределения и др. Однако при этом не вполне ясным остается вопрос, все ли статистики являются несостоятельными или только некоторые из них. Не менее важным вопросом является и «степень» состоятельности отдельных характеристик, т. е. скорость их сходимости к предельным величинам. Сказанное в полной мере относится и к ассоциативным полям, при описании которых могут быть использованы самые различные статистики, состоятельность которых до настоящего времени не подтверждена экспериментально.

2. МЕТОДИКА ИССЛЕДОВАНИЯ

2.1. Материал исследования

Исследование выполнено на материале совокупности ассоциатов слова-стимула *береза*. Эта совокупность получена в результате ассоциативного эксперимента, описанного в работе (Мартинович, 1993). В роли информантов этого эксперимента выступали студенты Санкт-Петербургского государственного университета, Педагогического университета им. А.А. Герцена и Гуманитарного университета профсоюзов. Всего к участию в эксперименте было привлечено 517 студентов, от которых было получено суммарно 2500 реакций, при этом реакции, имевшие вид относительно развернутых высказываний, были подвергнуты лексикографической обработке путем сведения их только к номинативным единицам.

Чтобы избежать влияния возмущающих факторов, порождаемых индивидуальными и коллективными различиями между информантами, все слова-реакции были подвергнуты рандомизации с помощью таблицы случайных чисел.

В итоге был получен частотный словарь слов-реакций, фрагмент которого приведен в табл. **2**. Словарь построен согласно стандартам, принятым в группе «Статистика речи» Санкт-Петербургского педагогического университета им. А.А. Герцена (Алексеев, 1977).

2.2. Типы шкал и соответствующие им статистики

В статистике упорядочивание значений варьирующего признака осуществляется в одной из трех измерительных шкал: количественной, порядковой (ординальной) и номинальной. Для различных шкал исторически сложились методы обработки данных, пригодные только для работы именно в данной шкале. Самая разработанная система техник сформировалась для количественных переменных. Здесь на основании теории моментов была разработана развитая система средних, показателей вариации, характеристик формы распределения и т.п. Здесь же продуктивно используется система порядковых статистик (мода, медиана, квантили и др.).

Система показателей для номинальной и порядковой шкал существенно беднее. Так, наиболее популярной характеристикой для номинальной шкалы является мода, а для ординальной - медиана.

Ранговые и спектровые распределения, приведенные в табл. 1, обладают всеми внешними свойствами количественной шкалы, но при этом характеризуются патологически большой вариацией признака как на порядковой шкале (ранговое распределение), так и на количественной шкале (спектровое распределение).

На основании литературных данных и собственных изысканий нами был сформирован достаточно полный перечень параметров, с помощью которых может быть

описана лексико-статистическая структура коммуникативно-тематического поля. В табл. 3 эти параметры разнесены по группам в соответствии с типом шкалы.

Таблица 2.

Фрагмент частотного словаря слов-ассоциатов слова-стимула *береза*

Ранг	**Слово**	**Частота**	**Ранг**	**Слово**	**Частота**	**Ранг**	**Слово**	**Частота**
1	Белый	129	39-41	Светлый	13	76-85	Почка	6
2	Стройный	98	42	Тишина	12		Природа	
3	Кудрявый	66	43-47	Гриб	11		Счастье	
4	Зеленый	62		Дрова			Тепло	
5	Дерево	53		Качаться			Шелестеть	
6	Русский	46		Красавица		86-100	Ансамбль	5
7	Роща	45		Одинокий			Гнездо	
8	Лес	41	48-51	Гибкий	10		Засохший	
9	Тонкий	39		Грустный			Карликовый	
10	Сок	38		Кривой			Корень	
11-13	Весна	30		Ствол			Костер	
	Девушка		52-55	Ветка	9		Крона	
	Нежный			Дом			Облако	
14-15	Белоствольный	29		Осина			Покой	
	Красивый			Хоровод			Поляна	
16-17	Береста	28	56-65	Ветер	8		Простор	
	Веник			Голубой			Пушистый	
18-19	Высокий	7		Карельский			Синий	
	Поле			Нежность			Чистота	
20	Родина	25		Песня			Шуметь	
21-23	Лето	23		Печальный		101-120	Анс. "Березка"	4
	Россия			Река			Грусть	
	Стоять			Родной			Ель	
24-25	Небо	21		Свежесть			Задумчивый	
	Плакучий			Черный			Запах	
26-27	Лист	20	66-75	Ветерок	7		Изящный	
	Подберезовик			Земля			Иней	
28-29	Молодой	19		Клен			Корявый	
	Трава			Листва			Ласковый	
30-33	Баня	17		Любовь			Листочек	
	Дуб			Молодость			Луг	
	Есенин			Птица			Маленький	
	Солнце			Снег			Невеста	
34	Сережки	16		Сосна			Облетевший	
35-36	Деревня	15		Старый			Пень	
	Красота		76-85	Березов. каша	6		Плакать	
37-38	Кора	14		Весенний			Русь	
	Расти			Озеро			Склониться	
39-41	Ветвистый	13		Окно			Стихи	
	Зелень			Полено			Стройность	

Таблица 3

Список статистических параметров коммуникативно-тематического поля

Номинальная шкала	**Количественная шкала**	**Порядковая (ранговая) шкала**
Мода (Mo) Объем словаря (n) Максимальная частота (F_{max}) Энтропия (E) Максимальная энтропия (E_{max}) Мера упорядоченности ($O=E/E_{max}$)	Средняя арифметическая (F_{cp}) Средняя геометрическая (F_g) Стандартное отклонение (σ) Среднее линейное отклонение (D_f) Коэффициент вариации по стандартному отклонению (V_σ) Коэффициент вариации по среднему линейному отклонению (V_{Df}) Коэффициент разнообразия (K)	Ранговое среднее (R_{cp}) Стандартное отклонение (σ_r) Коэффициент вариации (V_r) Медиана (Me_r) Золотое сечение (G_r) Среднее линейное отклонение (D_r) Коэффициент вариации по D_r (V_{Dr}) Индекс концентрации (γ)

Некоторые из статистик, приведенных в таблице 3, нуждаются в комментариях.

Начнем с номинальной шкалы.

Известно, что **мода** (Mo) – это вариант с наибольшей частотой. Поскольку шкала является номинальной, то мода также номинальна. Например, если в частотном словаре указаны только полиграфические облики слов и их частоты, то модой в такой ситуации будет имя слова. Так, в большинстве канонических частотных словарей самым частым словом является союз «и», который вследствие этого и является модой.

Объем словаря (n) - число разноименных единиц (лексем, словоформ, терминов и т.п.) в частотном словаре.

Энтропия (E) является мерой упорядоченности (организованности) систем самой разнообразной природы. Она вычисляется по формуле:

$$E = -\sum_{i=1}^{n} p_i \log p_i ,$$

где p_i – относительные частоты единиц (например, слов), образующих конкретную совокупность.

При фиксированном объеме словаря энтропия максимальна в том случае, когда частоты всех слов одинаковы, т.е. распределены равномерно. **Максимальная энтропия** (E_{max}) вычисляется по формуле:

$$E_{max} = \log n ,$$

где n – число разноименных элементов (т.е. объем словаря).

Мера упорядоченности ($O=E/E_{max}$), представляющая собюой частное от деления энтропии на максимальную энтропию – статистический индекс, указывающий на степень неравномерности распределения частот. Этот показатель всегда меньше единицы. Его близость к единице свидетельствует о слабой упорядоченности данной совокупности.

Перейдем к количественной (частотной) шкале.

Средняя арифметическая (F_{cp}) вычисляется по формуле:

$$F_{cp} = \frac{\sum_{i=1}^{n} f_i \cdot n_i}{n_i},$$

где n – объем словаря.

Для **средней геометрической** (F_g) формула имеет вид:

$$F_g = \sqrt[n]{\prod_{i=1}^{n} f_i \cdot n_i},$$

где n – объем словаря. Логарифмируя это выражение, получаем

$$\lg F_g = \frac{1}{n} \sum_{i=1}^{n} (\lg f_i) \cdot n_i .$$

Рабочая формула имеет вид:

$$F_g = \text{antilg}\left[\frac{1}{n} \sum_{i=1}^{n} (\lg f_i) \cdot n_i\right].$$

Средняя геометрическая, также как и энтропия, иногда интерпретируется как показатель неравномерности распределения (Трофимов, 1985). Этот вид средней в теоретико-статистических исследованиях фигурирует как наиболее оптимальный для закона Ципфа с точки зрения принципа максимального правдоподобия (Хастингс, Пикок, 1980).

Стандартное отклонение (σ) представляет собой корень квадратный из дисперсии, вычисляемой по формуле:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (f_i - F_{cp})^2 \times n_i}{n}$$

Среднее линейное отклонение (D_f) вычисляется по формуле:

$$D_f = \frac{\sum_{i=1}^{n} |f_i - F_{cp}| \cdot n_i}{n}$$

Этот показатель активно использовался на начальном этапе становления статистики как науки. Впоследствии он был вытеснен дисперсией и стандартным отклонением. В самое последнее время этот показатель начинает возвращать утраченные позиции (Айвазян, Енюков, Мешалкин, 1983).

Коэффициент вариации по стандартному отклонению (V_σ) вычисляется следующим образом:

$$V_\sigma = \frac{\sigma}{F_{cp}} .$$

Подобным же образом вычисляется и **коэффициент вариации по среднему линейному отклонению** (D_r):

$$V_{D_f} = \frac{D_f}{F_{cp}}.$$

Коэффициент разнообразия (K) представляет собой долю слов с однократным употреблением по отношению ко всему словарю:

$$K = \frac{t}{n},$$

где t – число однократных слов.

Последняя группа статистик относятся к порядковой шкале.

Ранговое среднее (R_{cp}) вычисляется по формуле:

$$R_{cp} = \frac{\sum_{i=1}^{n} r_i \cdot f_i}{\sum_{i=1}^{n} f_i},$$

где r_i – ранг лексической единицы в ранговом распределении. Эта статистика была впервые предложена одним из авторов данной работы (Мартыненко, 1988) и получила обоснование с позиций предельных теорем теории вероятностей (Мартыненко, Фомин, 1989).

R_{cp} изменяется в диапазоне:

$$1 \le R_{cp} \le \frac{n+1}{2},$$

т.е. достигает максимума при равномерном распределении.

Ранговые показатели, которые рассматриваются ниже, впервые были использованы при исследовании распределения ключевых слов в поисковых образах документов (Мартыненко, 2000).

Стандартное отклонение (σ_r) вычисляется по формуле:

$$\sigma_r = \frac{\sqrt{\sum_{i=1}^{n} (r_i - R_{cp})^2 \cdot f_i}}{\sum_{i=1}^{n} f_i}.$$

Коэффициент вариации (V_r) соответственно равен:

$$V_\sigma = \frac{\sigma_r}{R_{cp}}.$$

Медиана (Me_r) - величина ранга, соответствующая накопленной половине объема выборки (числа словоупотреблений) в кумулятивном убывающем ранговом распределении (см. табл. 1).

Золотое сечение (G_r) – значение ранга, соответствующее 38,2% от накопленного объема выборки в кумулятивном убывающем ранговом распределении (см. табл. 1).

Известны случаи применения золотого сечения при исследовании ритмики поэтического текста, например, сонета и Онегинской строфы (Гринбаум, 2000). При исследовании лексического уровня эта мера ранее не использовалась.

Среднее линейное отклонение (D_r) может быть получено по формуле:

$$D_r = \frac{\sum_{i=1}^{n} \left| r_i - R_{cp} \right| \cdot f_i}{\sum_{i=1}^{n} f_i}.$$

Коэффициент вариации по среднему линейному отклонению D_r (V_{Dr}):

$$V_{D_r} = \frac{D_r}{R_{cp}}.$$

Индекс концентрации (индикатор Ципфа) (γ) – показатель степени в ранговой форме закона Ципфа

$$f(x) = \frac{C}{r^{\gamma}}.$$

Лингво-математический смысл этого показателя был раскрыт в работе (Мартыненко, 1989), в которой приводится способ вычисления степени концентрации для ранговой формы ряда теоретических распределений. Для закона Ципфа этот индекс вычисляется следующим образом:

$$\gamma = \frac{n+1}{R_{cp}} - 1.$$

При $R_{cp} = 1$ γ максимален и равен n,
при $R_{cp} = (k+1)/2$ γ минимален и равен 1, т.е.

$$1 \le \gamma \le n.$$

Чем больше перепад между «головой» и «хвостом» рангового распределения, тем больше уровень концентрации в этом распределении.

Из представленного списка параметров в прикладных исследованиях чаще всего используются следующие: объем словаря (напр., Нешитой, 1975), максимальная частота (напр., Хайтун, 1983), средняя арифметическая (напр., Бондалетов, 1976), энтропия (напр., Овчинникова, Штерн, 1989) и индикатор Ципфа (напр., Хайтун, 1983). Из табл. 2 видно, что наш список параметров значительно шире. В частности он включает большое число ранговых параметров, которые используются только в последние годы (Мартыненко, 1988, 1989 и 2000; Гребенников, 1998); для некоторых из них было сформулировано математическое обоснование их теоретической состоятельности (Мартыненко, Фомин, 1989).

2.3. Методика аппроксимации зависимости параметров ассоциативного поля от объема выборки

Каждый из параметров, указанных в табл. 2, был подвергнут тесту на состоятельность по определенной методике, основанной на использовании метода наименьших квадратов с

некоторыми существенными модификациями, обусловленными спецификой исследуемого материала.

Исходные данные для этого теста представлены в таблице 4, где указаны значения статистических параметров (см. табл. 3) в зависимости от объема выборки. Исследование данной зависимости осуществлялось на макровыборке объемом 2500 словоупотреблений, разделенной на 25 микровыборок равного объема, последовательно присоединяемых друг к другу.

Графики, отражающие типичную динамику исследуемых показателей, представлены на рис. 1 – 3, из которых видно, что все показатели по мере увеличения объема выборки возрастают с разной, но всегда с затухающей скоростью, при этом, однако, не совсем ясно, стремятся ли они к фиксированному верхнему пределу или нет.

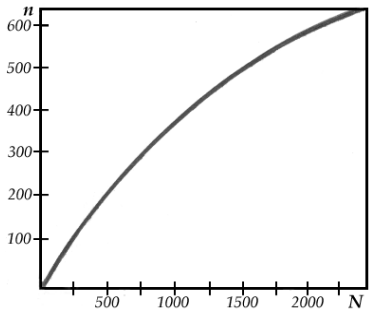

Рис. 1. График зависимости объема словаря (n) от объема выборки (N).

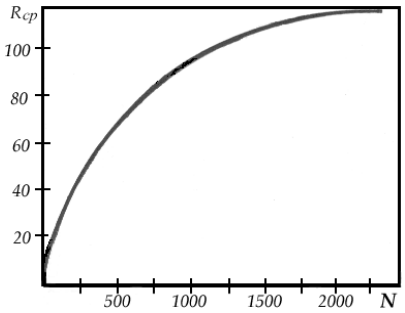

Рис. 2. График зависимости рангового среднего (R_{cp}) от объема выборки (N).

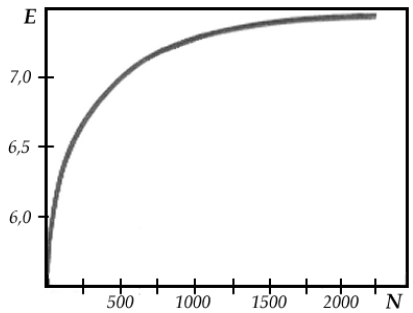

Рис. 3. График зависимости энтропии (E) от объема выборки (N).

Для аппроксимации эмпирических зависимостей между объемом выборки и величиной конкретного параметра был использован широкий спектр теоретических функций асимптотического и неасимптотического роста. Некоторые из них используются в экстраполяционном прогнозировании (Гражданников, 1974; Haustein, 1970) и в квантитативной лингвистике (Тулдава, 1980).

Для неасимптотической зависимости использовались четыре простейшие функции:

$y = ax + b$	линейная функция
$y = ax^b$	степенная функция
$y = ae^{bx}$	экспоненциальная функция
$y = a\,(\ln x)^b$	логарифмическая функция

Чтобы определить численные значения коэффициентов a и b линейной функции, должна быть решена система двух нормальных уравнений с двумя неизвестными:

$$a\Sigma x + nb = \Sigma y$$
$$a\Sigma x^2 + \Sigma bx = \Sigma yx\ ,$$

где n – количество значений независимой переменной. Параметры уравнения можно определить по следующим формулам:

$$a = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - \Sigma x\Sigma y}\ ;$$
$$b = \frac{\Sigma y\Sigma x^2 - \Sigma xy\Sigma x}{n\Sigma x^2 - \Sigma x\Sigma y}\ .$$

Что касается трех оставшихся уравнений, то здесь может быть осуществлен переход к линейной функции путем их логарифмирования. В результате этой операции уравнения приобретают вид:

$\ln y = \ln a + b \ln x$	степенная функция
$\ln y = \ln a + bx$	экспоненциальная функция
$\ln y = \ln a + b \ln \ln x$	логарифмическая функция

Список использованных нами функций асимптотического роста значительно шире. Эти функции распадаются на три группы.

Первая группа объединяет разностно-дробные функции:

$y = k - \frac{a}{x^b}$	степенная функция
$y = k - ke^{-\boldsymbol{axb}}$	экспоненциально-степенная функция (функция Вейбулла)
$y = k - \frac{a}{(\ln x)^b}$	логарифмически-степенная функция

Вторая группа состоит из вариантов дробно-экспоненциальной функции:

$y = \frac{k}{e^{\frac{\boldsymbol{a}}{\boldsymbol{x^b}}}}$	экспоненциально-степенная функция
$y = \frac{k}{e^{\frac{\boldsymbol{a}}{\boldsymbol{e^{bx}}}}}$	двойная экспоненциальная функция
$y = \frac{k}{e^{\frac{\boldsymbol{a}}{\boldsymbol{(\ln x)^b}}}}$	экспоненциально-логарифмическая функция

При $b = 1$ функция Вейбулла превращается в показательную.

И, наконец, третью группу образуют логистические функции:

$y = \frac{k}{1 + \frac{a}{x^b}}$	степенная логиста (замедленная логистическая функция)
$y = \frac{k}{1 + \frac{a}{e^{bx}}}$	экспоненциальная логиста
$y = \frac{k}{1 + \frac{a}{(\ln x)^b}}$	логарифмическая логиста

Каждая из 9 функций асимптотического роста путем однократного или повторного логарифмирования также может быть преобразована в линейную зависимость. Линейные варианты этих функций имеют вид:

Первая группа:

$\ln(k-y) = \ln a - b \ln x$	степенная функция
$\ln \ln \frac{k}{k-y} = \ln a - b \ln x$	экспоненциально-степенная функция (функция Вейбулла)
$\ln(k-y) = \ln a - b \ln \ln x$	логарифмически-степенная функция

Вторая группа:

$\ln \ln \frac{k}{y} = \ln a + b \ln x$	экспоненциально-степенная функция
$\ln \ln \frac{k}{y} = \ln a - bx$	двойная экспоненциальная функция
$\ln \ln \frac{k}{y} = \ln a - b \ln \ln x$	экспоненциально-логарифмическая функция

Третья группа:

$\ln \frac{k}{y-1} = \ln a - b \ln x$	степенная логиста
$\ln \frac{k}{y-1} = \ln a - bx$	экспоненциальная логиста
$\ln \frac{k}{y-1} = \ln a - b \ln \ln x$	логарифмическая логиста

Как говорилось выше, система нормальных уравнений для линейной зависимости очень проста, и ее решение относительно неизвестных параметров не вызывает сомнений. Однако в нашем случае один из параметров (асимптота) входит в состав зависимой переменной, что исключает использование метода наименьших квадратов в чистом виде. Нами был найден следующий выход из сложившейся ситуации:

1) асимптоте k мы задавали конкретные значения через фиксированный шаг, начиная с максимального значения конкретного параметра, достигнутого в эмпирическом ряду;

2) далее на каждом шаге при фиксированном значении асимптоты методом наименьших квадратов вычисляли значения других коэффициентов (a и b) с помощью системы уравнений, упомянутой выше;

3) на каждом шаге вычислялась сумма отклонений по модулю теоретической кривой от эмпирической;

4) затем строился график зависимости величины отклонений от последовательности пошагово задаваемых значений асимптоты; график такой зависимости имеет вид U-образной кривой с ярко выраженным минимумом (см. рис. 4);

5) среди множества величин отклонений выбиралось минимальное, которое и принималось за теоретическое значение асимптоты.

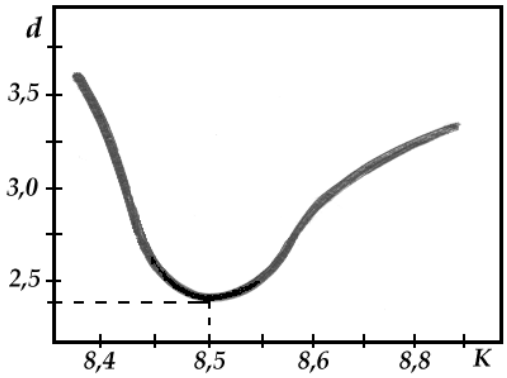

Рис. 4. График зависимости суммы отклонений от величины асимптоты *K*, задаваемый при аппроксимации зависимости энтропии от объема выборки с помощью степенной логисты. *K* = 8,50 – оптимальная асимптота, соответствующая минимуму отклонений.

3. ПАРАМЕТРИЗАЦИЯ РЕЗУЛЬТАТОВ АССОЦИАТИВНОГО ЭКСПЕРИМЕНТА

3.1. Результаты аппроксимации экспериментальных данных

Все эмпирические зависимости, приведенные в табл. 4, были подвергнуты обработке с помощью указанной последовательности операций с использованием всех 12 теоретических функций асимптотического и неасимптотического роста. В результате было получено 264 обработки (22 параметра × 12 теоретических функций). Все 12 теоретических функций с точки зрения их согласованности могут быть разбиты на три группы: 1) функции с хорошим согласованием – функция Вейбулла и степенная логиста, 2) функции с относительно удовлетворительным согласованием (все они являются функциями асимптотического роста) – степенная, экспоненциально-степенная, экспоненциальная логиста, 3) функции с плохим согласованием, сюда относятся в частности все функции неасимптотического роста.

В данном исследовании мы ограничимся интерпретацией поведения функций первой группы, т. е. функции Вейбулла и степенной логисты. Обе функции относятся к разряду функций асимптотического роста. Из этого следует, что все параметры, аппроксимируемые с помощью обеих функций, имеют теоретический верхний предел, несмотря на качественные различия между этими параметрами и скоростью их стремления к верхнему пределу.

На примере трех параметров (объем словаря, ранговое среднее и энтропия) рассмотрим характер зависимости каждого из них от объема выборки (*N*) под эмпирическим и теоретическим углом зрения (см. табл. 5). Прежде всего обращает на себя внимание тот факт, что эмпирические данные и результаты аппроксимации с помощью двух функций очень хорошо согласуются друг с другом. Более того, практически нет никаких различий между двумя вариантами аппроксимации: на рис. 5 – 7 видно, что аппроксимации с помощью функций Вейбулла и степенной логисты практически сливаются в одну линию (с целью достижения обозримости графики выполнены в билогарифмических координатах).

Этот результат представляется довольно неожиданным, поскольку функция Вейбулла и степенная логиста с математической точки зрения являются достаточно контрастными функциями. Первая из них основана на экспоненциальном росте, а вторая – на степенном. Обращает на себя внимание и тот факт, что реальные аппроксимации с помощью данных функций имеют по каждому параметру весьма различающиеся асимптоты (*K*).

По-видимому, высокая и примерно одинаковая степень согласованности с эмпирическими данными обеих функций объясняется тем, что несмотря на весьма существенные различия между асимптотическими уровнями, обе функции имеют примерно одинаковый «ход» при относительно малых объемах выборки, как это имеет место и в нашем эксперименте. С возрастанием объема выборки кривые этих функций начинают постепенно «разбегаться», устремляясь к своей индивидуальной асимптоте, причем эти асимптоты для обоих распределений, как это видно, например, из рис. 5, довольно существенно отличаются друг от друга.

Характер сходимости параметров ассоциативного поля к предельным величинам (асимптотам) наглядно просматривается на примере зависимости объема словаря от объема текста. Из табл. 5 (прогноз) видно, что функция Вейбулла и степенная логиста начинают «разбегаться» примерно при объеме выборки в 5 тыс. словоупотреблений, в 2 раза превышающем объем выборки, достигнутый в нашем эксперименте. По мере приближения к асимптотам, пространственный разрыв между кривыми постепенно достигает максимума: для объема словаря это происходит приблизительно в области 100 тыс. словоупотреблений.

Таблица 4

Зависимость параметров ассоциативного поля от объема выборки

Объем выборки	Объем словаря	Энтро-пия	Мера упоря-доч.	Средняя частота	Среднее гео-метр.	Стан-дарт. отклон.	Среднее лин. откл.	Коэф. вар.по σ	Коэф. вар.по D_f
N	n	E	O	F_{cp}	F_g	σ	D_f	$V\sigma$	V_{Df}
100	71	5,942	0,966	1,408	1,256	0,897	0,610	0,637	0,433
200	121	6,546	0,946	1,653	1,352	1,498	0,939	0,906	0,568
300	166	6,897	0,935	1,807	1,403	1,929	1,138	1,068	0,630
400	200	7,049	0,922	2,000	1,462	2,443	1,380	1,222	0,690
500	235	7,183	0,912	2,128	1,483	2,887	1,603	1,357	0,753
600	267	7,289	0,904	2,247	1,510	3,292	1,778	1,465	0,791
700	296	7,389	0,900	2,365	1,545	3,619	1,937	1,530	0,819
800	321	7,442	0,894	2,492	1,581	4,060	2,100	1,629	0,843
900	346	7,538	0,894	2,601	1,629	4,246	2,211	1,632	0,850
1000	362	7,515	0,884	2,762	1,656	4,832	2,453	1,749	0,888
1100	391	7,591	0,882	2,813	1,663	5,091	2,531	1,810	0,900
1200	414	7,639	0,879	2,899	1,679	5,368	2,643	1,852	0,912
1300	431	7,663	0,876	3,016	1,714	5,700	2,783	1,890	0,923
1400	454	7,682	0,870	3,084	1,706	6,070	2,930	1,968	0,950
1500	473	7,708	0,867	3,171	1,726	6,386	3,053	2,014	0,963
1600	495	7,742	0,865	3,232	1,734	6,661	3,158	2,061	0,977
1700	517	7,771	0,862	3,288	1,734	6,943	3,252	2,112	0,989
1800	535	7,799	0,860	3,364	1,758	7,228	3,334	2,149	0,991
1900	554	7,817	0,858	3,364	1,762	7,478	3,454	2,223	1,027
2000	572	7,831	0,855	3,497	1,766	7,741	3,566	2,214	1,020
2100	591	7,839	0,851	3,553	1,766	8,121	3,653	2,286	1,028
2200	610	7,855	0,849	3,607	1,770	8,400	3,745	2,329	1,038
2300	628	7,869	0,847	3,662	1,778	8,719	3,834	2,381	1,047
2400	643	7,883	0,845	3,733	1,791	9,012	3,924	2,414	1,051
2500	658	7,899	0,844	3,799	1,803	9,241	4,022	2,432	1,059

Продолжение табл. 4

Объем выборки	Ранговое среднее	Коэффи ц. разно-обр.	Стандарт. откл. по R	Коэффиц. вар. по R	Медиана по R	Золотое сечение	Средн. лин. откл. по R	Коэф. вар. по D	Индекс концент.
N	Rcp	K	σ_r	V_r	Me_r	G_r	D_r	$\mathbf{V_{Dr}}$	γ
100	28	0,746	22	0,786	21	14	19	0,679	1,157
200	41	0,719	37	0,902	28	16	32	0,780	1,976
300	53	0,705	50	0,943	32	19	43	0,811	2,151
400	59	0,690	60	1,017	33	19	50	0,847	2,407
500	65	0,698	69	1,062	34	19	58	0,892	2,631
600	71	0,689	78	1,099	35	19	65	0,915	2,775
700	76	0,679	85	1,118	37	20	70	0,921	2,908
800	79	0,660	91	1,152	37	21	74	0,937	3,076
900	84	0,642	97	1,155	39	22	78	0,929	3,131
1000	84	0,649	100	1,190	37	21	80	0,952	3,321
1100	89	0,639	107	1,202	38	21	86	0,966	3,404
1200	92	0,638	113	1,228	39	22	89	0,967	3,511
1300	94	0,626	116	1,234	39	22	91	0,968	3,596
1400	96	0,648	121	1,260	39	21	95	0,990	3,740
1500	99	0,634	126	1,273	39	22	98	0,989	3,788
1600	101	0,630	131	1,297	39	22	101	1,000	3,911
1700	104	0,629	136	1,308	40	22	105	1,010	3,981
1800	107	0,622	140	1,308	40	22	108	1,009	4,009
1900	108	0,625	144	1,333	39	22	110	1,019	4,139
2000	110	0,622	148	1,345	39	22	113	1,027	4,209
2100	112	0,626	152	1,357	39	22	116	1,036	4,286
2200	114	0,628	156	1,368	39	22	118	1,035	4,360
2300	116	0,623	160	1,379	39	22	121	1,043	4,422
2400	117	0,619	163	1,393	39	22	123	1,051	4,504
2500	119	0,616	166	1,395	39	22	125	1,050	4,535

В ассоциативном эксперименте этот уровень практически не достижим. Причем для степенной логисты предельный уровень достигается при более значительных объемах выборки. Поэтому вопрос о верхнем пределе объема словаря в теоретическом аспекте остается открытым, поскольку на объемах реального эксперимента обе функции, как было сказано выше, практически идентичны. Примерно та же картина наблюдается и для других параметров, лишь с той разницей, что скорость практического достижения асимптоты по каждому параметру различна для обеих аппроксимирующих функций.

Таблица 5

Результаты аппроксимации зависимости объема словаря, ранговой средней и энтропии от объема выборки с помощью:

функции Вейбулла $$y = k - ke^{-\boldsymbol{a}\boldsymbol{x}^{\boldsymbol{b}}}$$

степенной логисты $$y = \frac{k\,x^b}{a + x^b}$$

N	Объем словаря			Ранговое среднее			Энтропия		
	Э	Теория		Э	Теория		Э	Теория	
		Вейбулл $k = 2250$ $a = 0{,}034$ $b = 0{,}725$	Логиста $k = 4000$ $a = 52{,}83$ $b = 0{,}728$		Вейбулл $k = 175$ $a = 0{,}184$ $b = 0{,}564$	Логиста $k = 240$ $a = 7{,}244$ $b = 0{,}607$		Вейбулл $k = 8{,}15$ $a = 0{,}298$ $b = 1{,}319$	Логиста $k = 8{,}5$ $a = 0{,}427$ $b = 0{,}530$
1	71	74,383	74,312	28	29,429	29,112	5,942	5,972	5,941
2	121	121,604	121,624	41	41,699	41,693	6,546	6,541	6,553
3	166	161,626	161,713	53	50,702	50,860	6,897	6,843	6,861
4	200	197,412	197,542	59	57,970	58,209	7,049	7,041	7,055
5	235	230,238	230,391	65	64,117	64,387	7,183	7,183	7,193
6	267	260,810	260,971	71	69,463	69,732	7,289	7,292	7,298
7	296	289,572	289,729	76	74,200	74,449	7,389	7,378	7,381
8	321	316,830	316,974	79	78,457	78,674	7,442	7,448	7,448
9	346	342,803	342,929	84	82,321	82,499	7,538	7,507	7,505
10	362	367,659	367,763	84	85,858	85,995	7,151	7,557	7,554
11	391	391,528	391,609	89	89,117	89,213	7,591	7,600	7,596
12	414	414,515	414,571	92	92,137	92,192	7,639	7,638	7,633
13	431	436,706	436,738	94	94,948	94,966	7,663	7,671	7,666
14	454	458,171	458,182	96	97,576	97,558	7,682	7,700	7,695
15	473	478,972	478,965	99	100,041	99,992	7,708	7,727	7,722
16	495	499,159	499,137	101	102,360	102,284	7,742	7,750	7,746
17	517	518,778	518,746	104	104,548	104,449	7,771	7,772	7,768
18	535	537,866	537,830	107	106,616	106,500	7,799	7,791	7,788
19	554	556,460	556,425	108	108,577	108,448	7,817	7,809	7,807
20	572	574,588	574,561	110	110,439	110,301	7,831	7,825	7,824
21	591	592,278	592,266	112	112,210	112,068	7,839	7,840	7,840
22	610	609,555	609,564	114	113,898	113,756	7,855	7,854	7,855
23	628	626,441	626,479	116	115,508	115,371	7,869	7,867	7,869
24	643	642,954	643,030	117	117,047	116,920	7,883	7,879	7,883
25	658	659,115	659,235	119	118,519	118,406	7,899	7,890	7,895
П Р О Г Н О З									
50		981,26	985,47		142,115	143,345		8,032	8,073
100		1377,10	1405,22		160,217	166,351		8.105	8,201
500		2142,42	2544,67		174,617	205,710		8,148	8,371
1000		2235,22	2973,48		174,980	216,323		8,150	8,411

N – объем выборки (в сотнях словоупотреблений)
Э – экспериментальные данные

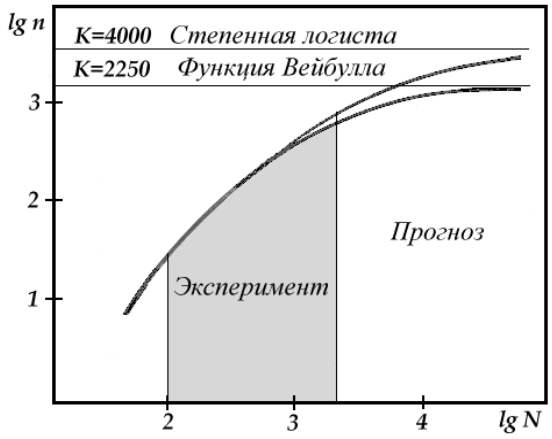

Рис. 5. Результаты аппроксимации зависимости объема словаря (n) от объема выборки (N) в билогарифмических координатах

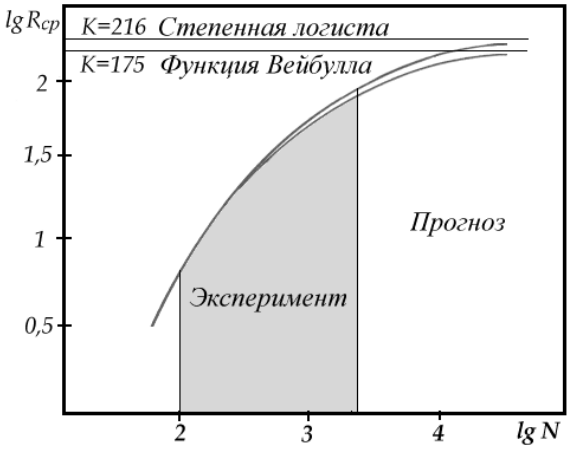

Рис. 6. Результаты аппроксимации зависимости ранговой средней (R_{cp}) от объема выборки (N) в билогарифмических координатах

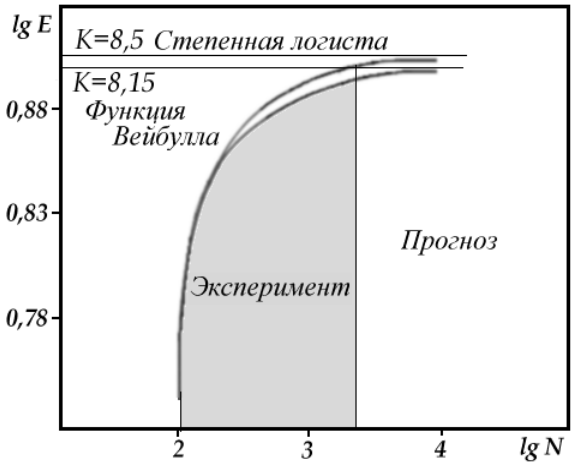

Рис. 7. Результаты аппроксимации зависимости энтропии (E) от объема выборки (N) в билогарифмических координатах.

3.2. Зависимость поведения аппроксимирующих функций от прогностической базы

В табл. 6 приведены значения функции Вейбулла и степенной логисты, спрогнозированные при нарастающих объемах выборки, т. е. при прогностической базе в 500, 1000, 1500, 2000 и 2500 словоупотреблений. Прогноз осуществлен для выборок очень большого объема: 10, 20 и 30 тыс. словоупотреблений. Соответствующие прогностические данные приведены в табл. 6, из которой видно, что все параметры достаточно сильно зависят от размера прогностической базы за исключением энтропии, медианы по рангу и коэффициента рассеивания. Последние очень быстро достигают своей асимптоты, причем между прогностическими данными для обеих функций нет существенных различий. Аналогичным образом ведут себя и некоторые другие функции, не показанные в табл. 6: золотое сечение по рангу, коэффициент разнообразия и мера упорядоченности. Это означает, что с одной стороны перечисленные параметры быстро сходятся к предельным величинам уже при сравнительно небольших объемах выборки, практически не выходящих за пределы нашего эксперимента, а с другой стороны – выбор аппроксимирующей функции в данной ситуации не имеет существенного значения, поскольку результаты аппроксимации по обеим функциям практически сливаются в одну линию (см. напр. рис. 6).

Другие параметры с точки зрения результатов аппроксимации с помощью функции Вейбулла и степенной логисты могут быть разбиты на 2 группы. К первой группе можно отнести параметры, которые, как мы предполагаем, при незначительном увеличении выборки более четко и однозначно проявят характер своей устремленности к предельным величинам. Такими параметрами можно считать, например, ранговое среднее, среднее линейное отклонение по рангу, коэффициент вариации по рангу и некоторые другие. Что же касается второй группы параметров, то у нас пока нет достаточных оснований, чтобы быть уверенными в их статистической состоятельности. Это относится в первую очередь к самому популярному среди исследователей параметру – объему словаря. В частности, как видно из табл. 6, асимптота степенной логисты для этого параметра почти в 2 раза превышает асимптоту для функции Вейбулла (имеем соответственно асимптоту *K,* равную 4000 и 2250). Столь же внушительные различия по этому параметру наблюдаются и в результатах прогноза. Учитывая особый интерес филологического сообщества к объему словаря как важной языковой характеристике, необходимо продолжить эксперименты, направленные на выявление не только теоретической, но и практической состоятельности этого параметра.

Таблица 6

ОБЪЕМ СЛОВАРЯ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	709,1	714,9	715	715	729	735	735	735	954	1009	1016	1017	1237	1475	1558	1620	1377	1780	1974	2250
Степенная логиста	902	1013	1061	1200	916	1031	1080	1225	1053	1234	1318	1600	1292	1654	1853	2800	1405		2186	4000

ЭНТРОПИЯ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	7,74	7,74	7,74	7,74	8,00	8,02	8,02	8,02	7,99	8,01	8,01	8,01	8,10	8,14	8,14	8,15	8,11	8,14	8,14	8,15
Степенная логиста	8,08	8,16	8,19	8,30	8,15	8,23	8,30	8,42	8,20	8,29	8,33	8,50	8,22	8,32	8,36	8,55	8,20	8,29	8,33	8,50

РАНГОВОЕ СРЕДНЕЕ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	92	92	92	92	114	114	114	114	129	130	130	130	138	140	140	140	160	170	173	175
Степенная логиста	143	153	157	170	158	174	181	210	143	153	157	170	158	173	181	210	166	186	196	240

ЧАСТОТНЫЙ КОЭФФИЦИЕНТ ВАРИАЦИИ ПО СТАНДАРТНОМУ ОТКЛОНЕНИЮ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	3,16	3,33	3,37	3,40	2,84	2,93	2,94	2,95	2,85	2,97	2,99	3,00	3,15	3,33	3,41	3,45	3,67	4,26	4,55	5,23
Степенная логиста	4,47	5,54	6,20	13,0	3,18	3,52	3,69	4,49	3,17	3,51	3,67	4,46	3,33	3,74	3,95	5,07	3,72	4,42	4,81	8,26

МЕДИАНА ПО РАНГУ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	35,0	35,0	35,0	35,0	39,5	39,5	39,5	39,5	39,8	39,8	39,8	39,8	40,5	40,5	40,5	40,5	40,1	40,1	40,1	40,1
Степенная логиста	37,3	37,4	37,4	37,4	41,5	41,9	42,0	42,0	41,5	41,8	41,8	42,0	41,1	41,3	41,4	41,5	40,3	40,4	40,4	40,5

КОЭФФИЦИЕНТ РАССЕИВАНИЯ

(прогноз при разных объемах выборки)

Выборка	500				1000				1500				2000				2500			
	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k	Прогноз (тыс. словоупотр.)			k
	10	20	30		10	20	30		10	20	30		10	20	30		10	20	30	
Функция Вейбулла	0,77	0,77	0,77	0,77	0,76	0,76	0,76	0,76	0,78	0,78	0,78	0,78	0,80	0,81	0,81	0,81	0,88	0,90	0,91	0,92
Степенная логиста	0,75	0,76	0,76	0,76	0,83	0,86	0,86	0,87	0,83	0,85	0,85	0,87	0,84	0,86	0,86	0,88	0,86	0,88	0,89	0,92

ЗАКЛЮЧЕНИЕ

В работе получены следующие основные результаты:

1. По литературным данным и на основании собственных изысканий был сформирован исходный перечень параметров результатов ассоциативного эксперимента.
2. Были построены 22 экспериментальных кумулятивно-динамических распределения, отражающих зависимость параметров коммуникативно-тематического поля от объема выборки.
3. На основании опыта, накопленного в экстраполяционном прогнозировании, в теории и практике анализа информационных потоков, а также в лексико-статистических исследованиях, был установлен набор теоретических функций асимптотического и неасимптотического роста, предназначенных для аппроксимации экспериментальных данных.
4. С помощью методики, основанной на использовании метода наименьших квадратов и адаптированной к лингвистическому материалу, осуществлена аппроксимация каждого кумулятивно-динамического ряда с помощью теоретических функций.
5. По каждому параметру исследована степень согласованности теоретических и эмпирических данных. Все теоретические функции были разбиты на 3 класса в соответствии с их обобщенной согласованностью с эмпирическими данными. Наилучшее согласование с эмпирическими данными обнаружили функция Вейбулла и степенная логиста.
6. Установлено, что все параметры обладают теоретической состоятельностью, однако практической состоятельностью с точки зрения проведенного ассоциативного эксперимента обладают лишь шесть из них (энтропия, медиана по рангу, коэффициент рассеивания, золотое сечение по рангу, коэффициент разнообразия и мера упорядоченности), поскольку они уже при сравнительно небольшом объеме выборки (например, в том, который достигнут в нашем эксперименте), практически достигают своих асимптотических значений.
7. Несмотря на принципиальные математические различия между функцией Вейбулла и степенной логистой, обе функции дают весьма близкие прогностические результаты для отмеченных выше практически состоятельных параметров. Это означает, что при работе с выборками, близкими по объему к нашей, при аппроксимации экспериментального материала пока нет достаточно веских оснований для того, чтобы отдать решительное предпочтение одной из этих функций. Обе функции в данной ситуации «работают» практически одинаково.
8. Обоснованность полученных в данной работе выводов в дальнейшем предполагается усилить путем расширения состава ассоциативных полей, увеличения числа параметров, наращивания объема выборки для каждого поля, а также посредством использования коммуникативно-тематических полей, построенных на основании дистрибутивно-статистического анализа текста и корпуса текстов.

ЛИТЕРАТУРА

Айвазян С.А., Енюков И.С., Мешалкин Е.Д. Прикладная статистика. Основы моделирования и первичной обработки данных. М., 1983.

Алексеев П.М. Статистическая лексикография (типология, составление и применение частотных словарей). – Л.: ЛГПИ им. А.И.Герцена, 1975.

Бондалетов В.Д. Русский именник, его состав, статистическая структура и особенности изменения (мужские и женские имена) // Ономастика и норма. М., 1976. С. 12–46.

Васильев Л.М. Теория семантических полей // Вопросы языкознания. 1971. N 5, с. 110.

Ворончак Е. Методы вычисления показателей лексического богатства текстов // Семиотика и искусствометрия. М., 1972. С. 232–250.

Гражданников Е.Д. Прогностические модели социально-демографических процессов. Новосибирск, 1974.

Гребенников А.О. О состоятельности статистик частотного словаря художественной прозы // Структурная и прикладная лингвистика. Вып. 5. СПб, 1998.

Гринбаум О.Н., Мартыненко Г.Я ."Онегинская строфа" и сонет: ритмические параллели // Труды межд. науч. конф. "А.С.Пушкин и мировая культура", М., 1999. – С.51–53.

Ланге О. Введение в эконометрику. М.: Прогресс, 1964.

Мартинович Г.А. Вербальные ассоциации и организация лексикона человека // Филологические науки. – 1989, № 3. С. 39–45.

Мартинович Г.А. К проблеме онтологии языка // Вестник Ленинградского университета. Сер. 2, 1989. Вып. 3. С. 54-60.

Мартинович Г.А. Текст и эксперимент. СПб, 1993.

Мартинович Г.А. Текст и эксперимент (исследование коммуникативно-тематического поля в русском языке). Автореф. дис. ... докт. филол. наук. - СПб, 1994.

Мартыненко Г.Я. О статистических характеристиках ранговых распределений // Квантитативная лингвистика и автоматический анализ текста. Вып. 872. Тарту, 1989. – С. 50–68.

Мартыненко Г.Я. Основы стилеметрии. Л., Изд-во ЛГУ, 1988.

Мартыненко Г.Я., Фомин С.В. Ранговые моменты // Научно-техническая информация. Серия 2, 1989, № 5. – С. 23–29.

Мартыненко Г.Я., Чарская Т.К. К вопросу о гауссовости и негауссовости лингвостатистических распределений // Структурная и прикладная лингвистика. Вып. 3, Л., Изд-во Ленингр. ун-та, 1987. – С. 63–76.

Нешитой В.В. Длина текста и объем словаря: Показатели лексического богатства текста // Методы изучения лексики. Минск, 1975. С. 110–118.

Нешитой В.В. О взаимосвязи ранговых распределений со спектровыми. – Научно-техническая информация. Сер. 2, 1968, 10. – С.19–24.

Овчинникова И.Г., Штерн А.С. Ассоциативная сила русского слова // Психолингвистические исследования фонетики и лексики. Калинин, 1989.

Трофимов В.П. Логическая структура статистических моделей. М., 1985.

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, 1980. С. 113–114.

Тулдава Ю.А. О квантитативных характеристиках богатства лексического состава художественных текстов // Linguistica. Тарту, 1977. IX. С. 159–175.

Тулдава Ю.А. О частотном спектре лексики текста. – Учен. зап. Тартуского ун-та, Вып. 745. Квантитативная лингвистика и автоматический анализ текста. Тарту, 1986. С. 139–162.

Хайтун С.Д. О Наукометрия. Состояние и перспективы. – М.: Наука, 1983.

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. Пер. с англ. – М.: Статистика, 1980.

Частотный словарь рассказов А.П.Чехова. Под ред. Г.Я.Мартыненко, Составитель А.О.Гребенников. СПб, Изд-во Санкт-Петербург. ун-та, 1998.

Чебанов С.В., Мартыненко Г.Я. Семиотика описательных текстов: типологический аспект. – СПб.: Изд-во С.-Петербург. ун-та, 1999. – 424 с.

Шрейдер Ю.А., Шаров А.А. Системы и модели. М., Радио и связь, 1982.

Яблонский А.И. Стохастические модели научной деятельности // Системные исследования. Ежегодник. М., 1975. С.5–42.

*Čebanov S., Martynenko G.*Text as Real Population in A.A.Čuprov's Sense. In: Journal of Quantitative Linguistics. Volume 5, Number 3. December 1998. – P. 163–166.

Haustein H.-D. Prognoseverfahren in den sozialistischen Wirtschaft. Berlin, 1970/

Martynenko G. On Measuring of Lexicosemantic Concentration in Text and Text Corpus // In: Workshop on Computational Lexicography and Multimedia Dictionaries. Patras University, Greece, 22–23 September 2000. – P. 11–14.

Martynenko G. Statistical Consistency of Keywords Dictionary Parameters // Content-Based Multimedia Information Access. Conference Proceedings. Volume 2, Paris, 2000. – P. 1541–1547.