ISSN 0494-7304 0207-4435

TARTU RIIKLIKU ÜLIKOOLI
TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

774

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1987

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1987

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 774 ВЫПУСК ОСНОВАНЫ В 1893.г.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1987

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1987

Kogumik "Kvantitatiivlingvistika ja tekstide automaatanalüüs" ilmub alates 1985.a. (jätkates sarja "Töid keelestatistika alalt", mis ilmus a. 1976 - 1984). Käesolevas kolmandas väljaandes on avaldatud Tartu Riikliku Ülikooli rakenduslingvistika uurimisgrupi liikmete ja välisautorite artiklid.

Сборник "Квантитативная лингвистика и автоматический анализ текстов" публикуется начиная с 1985 г. (сборник является продолжением серии "Труды по лингвостатистике" 1976 - 1984 гг.) Настоящий 3-й выпуск (1987 г.) содержит статьи сотрудников Группы прикладной лингвистики Тартуского государственного университета и исследователей из других городов.

The collections "Quantitative Linguistics and Automatic Text Analysis" have been published since 1985. The present issue No. 3 (1987) contains investigations by members of the Research Group of Applied Linguistics at Tartu State University and by guest authors.

О РАНГОВЫХ РАСПРЕДЕЛЕНИЯХ В КВАНТИТАТИВНОЙ ТИПОЛОГИИ ТЕКСТА

П.М. Алексеев

Квантитативная типология текста (КТТ) может иметь своим объектом текст не только индивидуальный, но и текст как обобщенный продукт речевых актов, реализующих язык и ту его часть, которая используется в какой-либо сфере его функционирования в тех или иных условиях, т.е. подъязык или даже идиолект. Поэтому уместно говорить о типологии именно текста, а не текстов; тогда объектом КТТ становятся и отдельный текст, и группа текстов, объединяемых темой, стилем, жанром, временем и т.д.

Анализ распределений лингвистических единиц, образующих текст, предлагается считать важнейшим средством КТТ, причем ему отводится роль, выходящая за пределы технической обработки результатов наблюдения. Распределения понимаются как квантитативная модель лингвистического объекта, его лексической, морфологической и других подсистем, и поэтому понятие распределения входит в концептуально-методологический аппарат КТТ.

Если далее признавать, что лингвистические явления образуют в своей совокупности систему, что они объединены в ней отношениями взаимосвязи и иерархии, то придется согласиться с тем, что определение количественной меры таких отношений - вопрос не только методики. Более того, группировка элементов, разбиение их на классы по любому из признаков, характеризующих их структурную или функциональную значимость, как раз и составляет содержание этого термина в самом общем смысле. Распределение поэтому - понятие не только гносеологическое, но и онтологическое, оно принадлежит и технике, и методике, и теории.

Ранее было обрисовано в целом место анализа распределений в КТТ и были обсуждены некоторые вопросы их классификации (Алексеев, 1983, 1985). Схема, которая строилась бы на базе трех стандартных статистических признаков изучаемого явления - количественного, качественного и рангового признаков, предлагалась не столько как альтернатива двум интересным подходам, впервые сформулированным в достаточно четком виде (Тулдава, 1982; Мартыненко, 1982), сколько в качестве дополнения, которое следовало бы учесть при разработке в бу-

дущем целостной теории лингвистических распределений.

Ранговые лингвистические распределения займут в любой классификации особое место и потребуют особого внимания. Это можно объяснить, во-первых, тем, что математическая традиция не принимает их, по-видимому, всерьез, к статистические средства их описания остаются недостаточно разработанными. Во-вторых, эти распределения характеризуют отношения внутри единой системной совокупности явлений, их связь друг с другом, и представляют собой одну из наиболее обобщающих моделей системного лингвистического объекта. Конечно, замена качественных или количественных значений признака рангами приводит к огрублению модели, к утрате более детальной информации, а это вызывает критическое отношение к такой подстановке. Ранговые распределения нередко представляются в логарифмическом масштабе, и это упрощение, делающее их анализ наглядным и доступным для лингвиста, усиливает критические замечания сторонников утонченной аналитической техники.

Однако факт остается фактом: ранговые распределения, в том числе и в логарифмической записи, привлекают именно своей простотой и возможностью представить в обозримом виде количественную упорядоченность системного объекта, состоящего из элементов с неодинаковым структурным или функциональным весом в системе. Недостаточно видеть в ранговом распределении лишь результат произвольного приписывания рангов частотам слов (или других элементов) текстовой выборки. Ранги здесь – это обозначение элементов системы, зафиксированных в выборке, или их количественных признаков. Целью настоящей статьи и является показать, что ранговые распределения, которые в общей классификации заняли бы не главное место, заслуживают гораздо большего внимания и отражают больше способов группировки или разбиения элементов системы, чем это обычно представляется в литературе.

Распределение в общем случае – это результат такого расположения элементов совокупности, когда каждому элементу с тем или иным признаком приписывается, сколько раз этот элемент (значение признака) зафиксирован в данной совокупности, т.е. приписывается его частота, численность. Пары значений, вариант, и их частот образуют вариационный ряд, или ряд распределения, называемый также просто распределением. Правило, связывающее варианты и их частоты в ряде распределения, называется законом распределения. Этого упрощенного объяснения вполне достаточно, чтобы, не используя вероят-

ностную терминологию, избежать вторжения в область вероятностей, до которой сегодняшней лингвостатистике при ее скромных масштабах наблюдения, доступных одному или даже целому коллективу исследователей, еще далековато.

О ранговым распределением имеем дело тогда, когда значения (варианты) количественного или качественного признака в ряде, построенном по убыванию или нарастанию частот, заменяются их порядковыми номерами - рангами вариант. Привычнее располагать ранжируемые варианты по убыванию частот, и тогда большим частотам соответствуют меньшие ранги вариант. Однако, расположив частоты по убыванию и приступая к ранжированию, мы нумеруем все-таки не частоты, а варианты, заменяемые рангами. Например, в классическом случае с частотным словарем (ЧС), ранжируются не частоты, а слова, имеющие ту или иную частоту: получают ранги сами слова, т.е. в данном случае варианты, конкретные выражения общего признака "принадлежность к словам". Об этом иногда забывают, оформляя ЧС и при словах указывая ранги как будто этих слов, но на самом деле ранги частот[+]. О необходимости различать ранг единицы ЧС и ранг ее частоты, а также о забавном смешении понятий (англ. "rank") и распространенности (англ. "range") говорилось в другом месте (Алексеев, 1975, с. 22-23), но тем не менее неключать возможность путаницы не стоило бы, если ух речь заходит о распределениях именно ранговых[++].

В табл. I, достаточно компактной для иллюстрации этих и последующих соображений, приведены полные данные о распределении терминологических словосочетаний в выборке из английских текстов по электронике на английском языке общей длиной 200 тыс. словоупотреблений (Частотный англо-русский...,1971, с. 285). Здесь представлено то, что обычно называют "лексическим спектром", "частотным спектром", "статистической структурой текста", "структурой ЧС" и т.п. Действительно, в таблице указаны все частоты обнаруженных в тексте терминосочетаний (т/с), количества т/с в группах с одинаковой часто-

---

[+] См., напр., серию ЧС-минимумов Лейпцигского университета им. К. Маркса, в частности (Fachwortschatz Physik, 1970).

[++] Составители одного из ЧС, нумеруя частоты слов и усомнившись в правомерности термина ранг, заключили его в кавычки (Частотный словарь... 1977, с. 895-915) и записали буквально: "[illegible] ("ранг")".

Таблица I

Частотный спектр терминологических словосочетаний в английском подъязыке электроники [+]

i	F	m	i	F	m
I	2	3	I	2	3
I	79	I	36-40	22	5
2	59	I	4I-43	2I	3
3	57	I	44-45	20	2
4-5	52	2	46-52	I9	7
6	49	I	53-59	I8	7
7	48	I	60-68	I7	9
8	46	I	69-78	I6	I0
9	43	I	79-92	I5	I4
I0	42	I	93-I04	I4	I2
II	4I	I	I05-III	I3	7
I2	40	I	II2-I28	I2	I7
I3	39	I	I29-I48	II	30
I4	38	I	I49-I68	I0	20
I5-I6	33	2	I69-207	9	39
I7-I8	30	2	208-256	8	49
I9-20	29	2	257-324	7	68
2I-22	28	2	325-4I6	6	92
23	27	I	4I7-557	5	I4I
24	26	I	558-795	4	238
25-28	25	4	796-I270	3	475
29-32	24	4	I27I-2443	2	II73
33-35	23	3	2444-8984	I	654I

[+] В этой таблице i - ранг т/с, F - частота т/с, m - количество т/с с частотой F.

той, порядковые номера - ранги всех 9 тыс. т/с. Хотя такую таблицу иногда называют таблицей рангового распределения, ранговое распределение представлено в ней колонками I-2. Колонки 2-3 - это уже другое распределение, не ранговое, но распределение по количественному признаку. И если ранги - это как бы элементарная ступень отвлеченности от самих лингвистических явлений, то числа в колонках 2-3 - это абстракция более "высокого" уровня.

Числа в колонке I - это варианты, т.е. в данном случае порядковые значения качественного признака, конкретного терминосочетания; в колонке 2 - частоты этих вариант. Для ряда, составленного из колонок 2-3, числа в колонке 2 - это уже варианты, а в колонке 3 - частоты этих вариант, т.е. если в колонке 2 помещены текстовые частоты т/с, то в колонке 3 - это частоты ("словарные") этих текстовых частот. Здесь повторение слова "частота" не должно приводить к путанице: каждый раз надо помнить, какое оно имеет содержание.

Таким образом, колонки 2-3 представляют не ранговое распределение, а распределение по количественному признаку, варианты которого (цифры в колонке 2) расположены по убыванию своих величин. Такой ряд можно разместить и по нарастанию величин, как это делается при рассмотрении количественных вариационных рядов, например радов распределения частот словоупотреблений текста или словоформ словаря этого текста по длине в буквах. В табл. 2 приведены данные для количественного вариационного ряда.

Таблица 2

Словарный спектр терминологических словосочетаний в английском подъязыке электроники*

i	m	n	m x n	i	m	n	m x n
I	2	3	4	I	2	3	4
I	6541	I	6541	I3	I4	I	I4
2	II73	I	II73	I4	I2	I	I2
3	475	I	475	I5	I0	I	I0
4	238	I	238	I6	9	I	9
5	I4I	I	I4I	I7-I9	7	3	2I
6	92	I	92	20	5	I	5
7	68	I	68	2I-22	4	2	8
8	49	I	49	23-24	3	2	6
9	39	I	39	25-30	2	6	I2
I0-II	20	2	40	3I-44	I	I4	I4
I2	I7	I	I7	Итого:			8984

* В этой таблице m - количество т/с с конкретной частотой, i - ранг этого количества, начиная с наибольшего, n - количество частот, приходящееся на m. Итоговая строка колонки 4 дает сумму разных т/с (ср. правое число в последней строке колонки I табл. I).

Здесь числа в колонке m размещены в порядке, обратном порядку соответствующей колонки табл. I. В результате получен ряд рангового распределения, в принципе отличный от "обычного" такого ряда, примером которого являются колонки I-2 в табл. I. Здесь, в отличие от первого случая, нумеруются уже не т/с, а частоты, приходящиеся на то или иное значение m. Объем выборки, измеряемый терминоупотреблениями (составными), по табл. I определить можно, если суммировать все произведения F на m. По табл. 2 этого сделать нельзя, можно лишь узнать объем словаря, для чего введена колонка произведений m на n.

Теперь необходимо пояснить, для чего понадобилась такая форма рангового распределения, как она представлена в табл. 2. Дело в том, что лингвостатистиков-"частотников" нередко упрекают за пренебрежение нижней зоной ЧС, отражаемой в "хвосте" рангового распределения, и за внимание только к верхней и средней зонам ЧС. К этому, правда, вынуждает и логарифмический масштаб, который уделяет все меньше места рангам по мере их удаления от нулевой оси, а частотам по мере их приближения к ней. Однако, говорят критики, подавляющая часть словаря сосредоточена именно в нижней, редкоупотребительной зоне ЧС. К тому же именно среди редких единиц ЧС содержатся семантически важные, несущие новую информацию единицы. Используя логарифмический масштаб и по сути сдавливая самую богатую информацией зону ЧС, мы сильно обедняем наши представления о моделируемой лингвистической системе. Все это совершенно справедливо, и поэтому здесь предлагается ранжировать частоты "хвоста" распределения, вынеся его на обозрение крупным планом и используя все то же свойство логарифмического масштаба.

На рис. I показаны графики этих двух ранговых распределений - в первом случае - это ранги и частоты т/с в тексте, во втором - ранги частот, начиная с наименьшей, приходящиеся на каждое отличное от других количество т/с с той или иной текстовой частотой. Здесь же показаны сглаживающие прямые как результат аппроксимации эмпирических распределений с помощью уравнения линейной регрессии (т.наз. "классический" случай закона Ципфа).

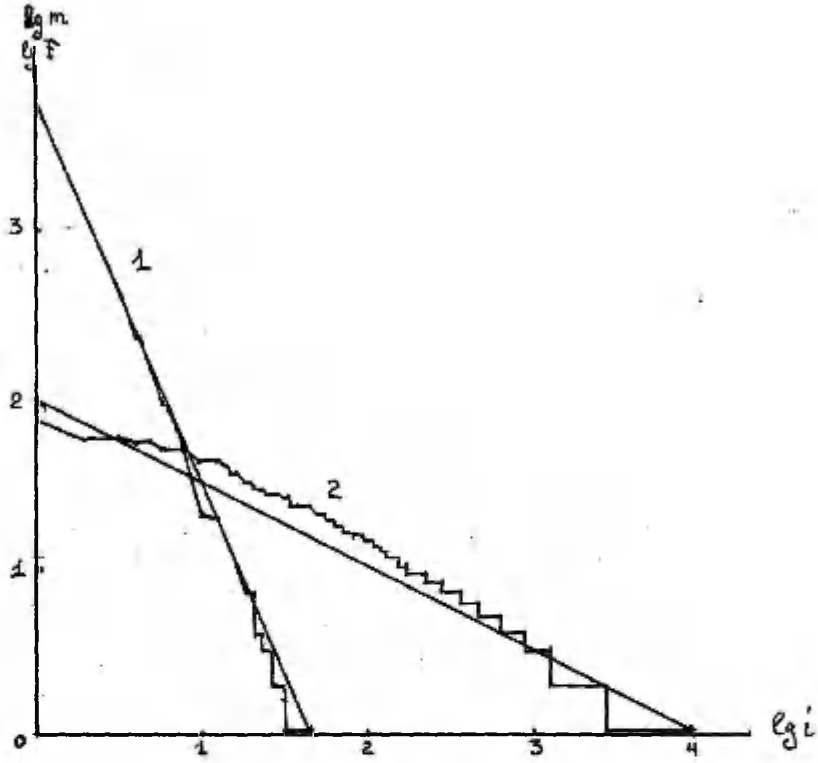

Рис. I. Ранговые распределения терминологических словосочетаний в английском подъязыке электроники. I - распределение i-m, 2 - распределение i-F.

Ранговое распределение вида i-m можно применять для сопоставления текстовых выборок, т.е. уже в целях КТТ. На рис. 2 показаны такие распределения в форме сглаживающих ципфовских графиков для однословных терминов в английских научных подъязыках[+].

Еще одну возможность описывать статистическую структуру текста и его словаря с помощью ранговых распределений иллюстрируют эмпирические и сглаживающие графики рис. 3, постро-

[+] Использованы данные из четырех ЧС (Лексико-терминологические материалы..., 1980, 1983; Учебные терминологические материалы..., 1982; Частотный англо-русский..., 1971). Выборка для каждого ЧС равна 200 тыс. словоупотреблений.

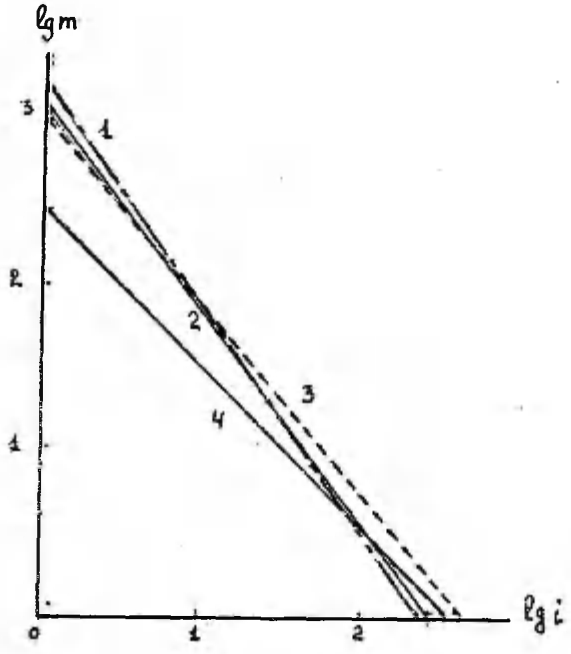

Рис. 2. Ранговые распределения вида i–m однословных терминов в английских подъязыках биологии (I), электроники (2), психологии (3) и математики (4).

ные по данным ЧС словосочетаний в английском газетном тексте[+]. Здесь видны резкие различия между распределениями текстовых частот словосочетаний и текстовых и словарных частот ("активности") структурных типов словосочетаний; тип сочетания определяется его составом в терминах частей речи: существительное + существительное, глагол + послелог и т.д. Сходство между распределениями текстовых и словарных частот типов словосочетаний как будто должно бы противоречить представлениям о принципиальных различиях между статистикой в тексте и статистикой в словаре. Однако оно объясняется очень просто: дело в объеме выборки. Хотя он и равен 100 тыс. словоупотреблений, для статистики на уровне словосочетаний он невелик. Сходство неизбежно исчезнет при росте выборки, особенно многократном. Количество новых сочетаний будет расти, увеличивая вес того или

[+] Соответствующие таблицы с полными цифровыми данными приведены в (Алексеев, 1974).

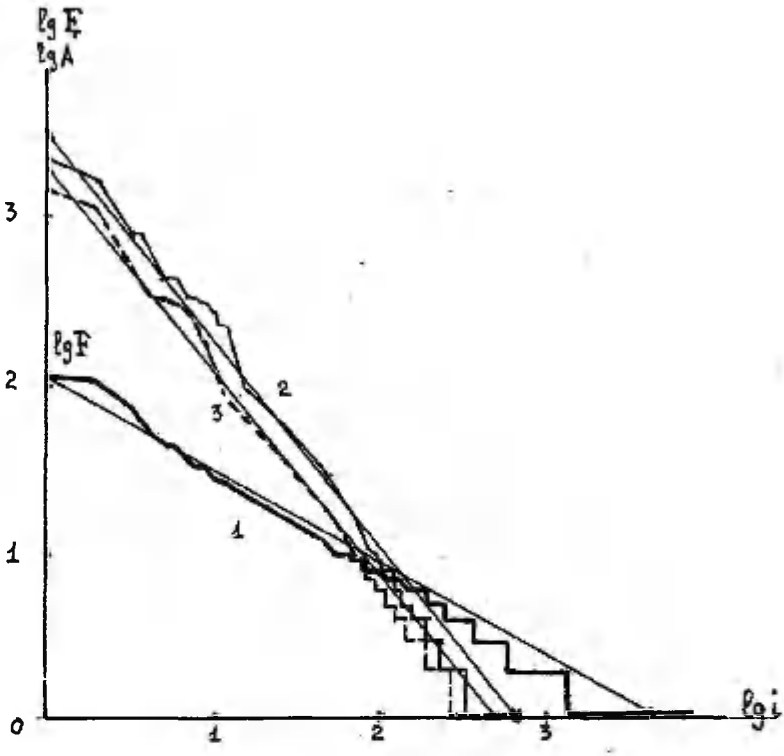

Рис. 3. Эмпирические и сглаживающие распределения текстовых частот словосочетаний (I), текстовых частот типов словосочетаний (2), словарных частот типов словосочетаний (3) в английских газетных текстах. Выборка равна 100 тыс. словоупотреблений. Здесь i - ранг, F - частота словосочетания, F_T - текстовая частота типа, A - словарная частота, "активность" типа.

иного типа на оси словаря. Вес этих же сочетаний в тексте будет оставаться ничтожным на фоне роста текстовых частот более употребительных сочетаний и типов.

При наличии соответствующих эмпирических данных можно было бы, очевидно, выносить дополнительные заключения об оптимальном ципфовском объеме" выборки - уже не только для уровня словоформ, но и для уровней словосочетаний, классов слов и т.д. и, возможно, также об "оптимальном ципфовском

объеме" словаря текстовой выборки[+].

За пределами настоящей статьи по необходимости остались другие случаи ранжирования единиц частотных инвентарей, их частот и других количественных характеристик. Здесь хотелось показать, что прежде чем завышать или занижать возможности ранговых распределений в лингвостатистике, предстоит рассмотреть их на более обширном и разнообразном материале, чем это пока делается, притом для разных условий наблюдения, для разных единиц, их признаков и классов, для разных способов образования классов. В квантитативной типологии текста становится все более очевидной необходимость учитывать, что оба аспекта лингвистической статистики – текстовой и словарный, могут проявляться не всегда привычным образом.

Л И Т Е Р А Т У Р А

Алексеев П.М. Методика квантитативной типологии текста. Учебное пособие. – Л.: ЛГПИ им. А.И. Герцена, 1983, 75 с.

Алексеев П.М. О нелинейных формулировках закона Ципфа. – Вопросы кибернетики. Вып. 41. Статистика речи и автоматический анализ текста. – М.–Л.: Научн. совет по комплексной проблеме "Кибернетика" АН СССР, 1978, с. 53–65.

Алексеев П.М. Об уровнях лингвистического анализа и о знаковости текста. – Инженерная лингвистика и романо-германское языкознание. – Л.: ЛГПИ им. А.И. Герцена, 1985, с. 5–19.

Алексеев П.М. Статистика словосочетаний в английском газетном тексте. – Вопросы статистической стилистики. –Киев: Наукова думка, 1974, с. 188–196.

Алексеев П.М. Статистическая лексикография. Учебное пособие. – Л.: ЛГПИ им. А.И. Герцена, 1975, 120 с.

Лексико-терминологические материалы для чтения текстов по биологии на английском языке. Частотный минимум. Сост. Л.Г. Берзиньш. – Л.: ЛГПИ им. А.И. Герцена, 1983, 71 с.

Лексико-терминологические материалы для чтения текстов по психологии на английском языке. Частотный минимум. Сост. Г.В. Басовская и А.В. Вербицкий. – Л.: ЛГПИ им. А. И. Герцена, 1980, 82 с.

[+] Некоторые соображения о причинах, влияющих на изменение формы ципфовского распределения, приводилось в (Алексеев, 1978).

Мартыненко Г.Я. Типология лингвостатистических распределений. – Учен. записки Тартуского университета. Вып. 628. – Тарту, 1982, с. 103–120.

Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (3): методика исследования. – Учен. записки Тартуского университета. Вып. 619. – Тарту, 1982, с. 123–143.

Учебные терминологические материалы для чтения текстов по математике на английском языке. Частотный минимум. Сост. Л.М. Сутягина. – Л.: ЛГПИ им. А.И. Герцена, 1982, 86 с.

Частотный англо-русский словарь-минимум по электронике. Сост. П.М. Алексеев. – М.: Воениздат, 1971, 302 с.

Частотный словарь русского языка. – М.: Русский язык, 1977, 935 с.

Fachwortschatz Physik. Häufigkeitswörterbuch russisch, englisch, französisch. – Leipzig: VEB Verlag Enzyklopädie, 1970, 107 S.

ON RANK DISTRIBUTION ANALYSIS IN QUANTITATIVE TYPOLOGY OF TEXT

Pavel Alekseev

S u m m a r y

An important role in the quantitative typology of text is played by the statistical distribution analysis of linguistic text units. Distributions are taken as quantitative models of complex linguistic systems and subsystems that are languages, sublanguages, text classes, individual texts, idiolects.

Even though rank distributions have been paid not too much attention to in mathematical statistics where they are considered to be an oversimplified presentation of data series, they deserve more of it, since they reflect certain universal and specific features of complex linguistic objects. It is proposed to enlarge the scope of rank distribution analysis covering word-classes, word-combinations and their classes, and other terms of possible statistical series.

An "inverted" Zipf series may be formed by assigning the first rank to the lowest frequency 1, the second one to frequency 2 etc, and filling the second row of the series with quantities of units that show equal frequency in a given count. In this way a distribution "tail" that is usually neglected in Zipf's law studies can be exposed to observation in terms of logarithmic scale. Linear approximation fits closely to experimental data.

УПОТРЕБИТЕЛЬНОСТЬ И МНОГОЗНАЧНОСТЬ СЛОВА

М.В. Арапов

I. В данной статье будет кратко рассмотрена связь между употребительностью слова и мерой его полисемии – числом значений данного слова. Решение поставленной задачи требует, чтобы были преодолены (или по крайней мере "обойдены") две существенные трудности.

Первая из них – хорошо известна. Она состоит в расплывчатости самого понятия "значение". Эта расплывчатость, во всяком случае на первый взгляд, вообще исключает возможность измерения. Действительно, даже если вести отсчет времени со средины 60-х годов, когда проблема значения стала центральной для лингвистики, то можно указать столь много разноречивых и даже просто несовместимых друг с другом суждений о природе значения, что было бы затруднительно привести даже самый краткий их обзор. Однако существует возможность обойти трудности, связанные с неопределенностью таких категорий как "отдельное значение", "полисемия", "омонимия" и проч. Эта возможность появляется, если отвлечься от теоретических споров и сосредоточить внимание на лексикографической практике.

Описание значений конкретных слов, выполненное лингвистами, стоящими на далеких друг от друга теоретических позициях, обнаруживает далеко идущее сходство, которое становится еще заметнее, если встать на достаточно абстрактную точку зрения. Будем рассматривать описание значения слова как организованный некоторым образом текст – т о л к о в а н и е. В словаре – одно- или двуязычном – толкование является частью словарной статьи, которая обычно включает грамматические, синтаксические и стилистические пометы, указания о происхождении слова, а также толкования других словарных единиц – устойчивых словосочетаний, содержащих данное слово. Но нас будет интересовать исключительно толкование, причем преимущественно его синтаксическая организация.

Лексикографы, стоящие на традиционной точке зрения, считают, что адекватное толкование вполне может быть записано средствами естественного языка (в толковом словаре – того же, к какому относятся толкуемые слова), если эти средства дополнить набором вспомогательных элементов, исторически сложившихся в лексикографической практике. Одни из этих эле-

ментов служат для разделения текста толкования на части и установления между этими частями определенных отношений, другие – для указания связей между толкованиями различных слов в одном словаре.

Части толкования могут быть равноправными, либо подчиненными друг другу. Например, в четырехтомном "Словаре русского языка" (далее сокращенно МАС – Малый академический словарь) равноправные части толкования разделяются арабскими цифрами. Внутри выделенных таким образом частей могут содержаться подчиненные фрагменты: например, толкование слова вагон в МАС состоит из одной части, но в ней – два подчиненные фрагмента. Считается, что у слова вагон одно значение: 'транспортное средство, специально оборудованное для перевозки пассажиров и грузов по рельсовым путям', но два "оттенка", выделенные знаками //: "количество груза, вмещающееся в один вагон" и "очень много, множество" (МАС, т. I, с. 133).

Толкования различных слов или равноправных частей этих толкований могут быть связаны ссылками различных типов. Составители словарей предполагают, что если толкование слова А содержит ссылку на слово Б, то по типу ссылки и толкованию Б читатель способен самостоятельно восстановить толкование А. Например, толкование слова вызолачиваться в МАС состоит из двух равноправных частей и каждая представляет собой ссылку: "I. Несов. к вызолотить. 2. Страдат. к вызолачивать. Поскольку слову вызолотить дается толкование 'покрыть позолотой', то часть I. у слова вызолачиваться должна была бы иметь вид 'покрывать позолотой', а часть 2 – 'подвергаться покрытию позолотой' или 'переносить покрытие позолотой'. Таким образом, при использовании ссылок 'синтаксическая структура" толкования сохраняется – по крайней мере на уровне самостоятельных частей, – хотя содержание частей более или менее закономерным образом меняется.

Современные критики (Апресян, 1974) традиционного подхода стремятся решить двуединую задачу: с одной стороны, так систематизировать систему ссылок; чтобы максимальное число связей между толкованиями было обозначено в словаре эксплицитно, а с другой стороны, настолько формализовать запись отдельной самостоятельной части, чтобы процесс "вычисления" одних значений по другим приблизился к алгоритмическому.

Признавая колоссальную важность данной задачи, мы хотели бы подчеркнуть существенный для нас момент: критики традиционной лексикографии сохраняют членение слова на отдельные части, и сам процесс членения для них логически предшествует формальной записи отдельных частей и отношений между ними. При этом членение по-прежнему опирается на интуицию лексикографа.

Естественно ожидать, что при переводе толкования на конкретный формальный язык членение толкования изменится по сравнению с членением в традиционном толковом словаре. Но фактически такая "перетасовка" значений происходит и при составлении каждого нового толкового словаря традиционного типа. Еще более уместным было бы, наверное, сравнение с процессом, который происходит при составлении двуязычного словаря: при переходе от одного словаря к другому членение меняется в зависимости от того, какие средства выражения предлагает язык, на который осуществляется перевод.

Наша гипотеза состоит в том, что количественной характеристикой о з н а ч а е м о г о является среднее число самостоятельных частей в толковании данного слова. Эта гипотеза в неявном виде присутствовала уже в самых ранних работах по квантитативной лингвистике, но их авторы, подметив, что с уменьшением употребительности слова убывает и число самостоятельных частей в его толковании, просто отождествляли это число с числом значений. Такая терминология удобна, но пользоваться ей нужно с осторожностью. Нельзя забыть, что толкование дает лишь приближенное представление о значении слова. Число частей в нем - только косвенная оценка сложности значения как число ломтей, на которые можно разрезать пирог, - косвенная характеристика его размеров.

В основе выбранного подхода лежит предположение (молчаливо принимаемое и нашими предшественниками), что при переходе от одного словаря к другому "масса" значения данного слова не изменяется. Если словари сопоставимы по типу и объему, то эта масса лишь иным образом перераспределяется по "ломтям", но их число остается грубой, но устойчивой характеристикой значения. При переходе к другому типу словаря (например, меньшего объема) размеры "ломтя" могут увеличиться, а число их сократиться. Поскольку для русского язы-

ка спектр толковых словарей очень ограничен, усреднять число частей в толковании одного слова не имеет смысла. Более целесообразен другой путь: использовать дополнительное предположение, что близкие по употребительности слова имеют и близкое число равноправных частей в толковании, и вести усреднение по словам с близкими показателями употребительности.

Но и тогда, однако, сопоставимость данных, относящихся к различным частям словаря, зависит от того, сохраняется ли с переходом от употребительных слов к редким неизменным само содержание понятия "самостоятельная часть толкования". Гарантировать этого нельзя. Некоторым косвенным показателем могут быть размеры выделяемых частей. Для ста наиболее употребительных в русском языке слов (по данным словаря Штейнфельдт (1963)) средняя длина самостоятельной части - 18 строк, для слов с рангом около 1000 - эта длина снижается до 10-11 и стабилизируется на уровне 7-8 строк для слов с рангом 3 000 - 4 000 (в словаре Засориной (1977)). То есть о линейной корреляции между числом частей в толковании и объемом толкования можно говорить только, если исключить из рассмотрения наиболее частые слова. Коэффициент корреляции достаточно высок (больше + 0,8 для взятых из МАС толкований слов с рангом большим 4 000 в словаре Засориной).

Как раз на стабилизацию, причем на стабилизацию не только зависимости между объемом словарной статьи и числом частей в ней, но и на стабилизацию зависимости между употребительностью и числом значений, мы и надеемся, пытаясь найти адекватную модель связи употребительности в полисемии слова.

Одновременно мы отдаем себе отчет, что экстраполяция на наиболее частные слова (в основном - служебные) связи между числом значений и показателями употребительности слова, найденная для слов с умеренной и малой частотой может привести к грубым ошибкам.

Но выбрав путь преодоления первой трудности, когда изучение значения слова заменяется изучением структуры его толкования в конкретном словаре, мы наталкиваемся на вторую трудность. Измеряя частоту употребления слова, его длину или возраст, мы знаем, что измеряем (возможно, с какой-то ошибкой) свойства самой лексической единицы. Но когда в ка-

честве измерительного инструмента используется толковый словарь, мы не можем быть уверены, что полученный результат характеризует только слово. Этот результат заведомо характеризует и сам словарь - метаязык, использованный для описания смысла.

Единственная возможность провести границу между тем, что определяется самим языком, а что - выбранным произвольно измерительным инструментом, состоит в сопоставлении результатов, полученных при использовании разных словарей. В данной работе мы делаем первый шаг в преодолении этой трудности, сопоставляя данные, полученные с помощью МАС и наиболее популярного толкового словаря С.И. Ожегова (Ожегов, 1986).

2. Полный и содержательный обзор работ, посвященных непосредственно употребительности слова и числа его значений, дан в работе С.И. Гиндина (Гиндин, 1982). Наши предшественники использовали обычно в качестве меры употребительности абсолютную частоту слова F и искали зависимость между F и средним числом $\bar{m}$ значений, $m = 1,2,3...$, характерным для слова с данной или близкими частотами. Наиболее общую из рассматриваемых гипотез выдвинул Ю.А. Тулдава, высказав предположение, что для широкого круга языков искомая зависимость должна описываться степенной функцией $\bar{m} = \alpha F^{\gamma}$. До этого Дж. Ципф (1943), а за ним П. Гиро (1954) считали, что $\gamma = 1/2$ в любом случае, позднее П.Ф. Андрукович и Э.И. Королев (1977) на исследованном им материале нашли, что $\gamma = 1/3$.

Таким образом, сложилась некоторая традиция изучения связи употребительности и многозначности слова. Нам представляется целесообразным отступить от этой традиции в трех следующих пунктах.

1) В качестве меры употребительности использовать не абсолютную частоту слова, а его ранг r в частотном словаре, r не зависит от объема выборки и позволяет легко сопоставлять данные, полученные на разном материале.

2) Исследовать не только динамику средних значений случайных величин M_r, но и характер их распределения в зависимости от ранга; ведь очевидно, что информация о среднем числе значений у редкого слова имеет совсем иную ценность, чем информация о среднем числе значений частого слова: в последнем случае вероятны большие уклонения от среднего значения.

3) Использовать для одного языка различные частотные и тол-

ковые слова, чтобы составить представление о зависимости полученных данных от исходного материала.

3. Если считать, что структура частот в текстах хотя бы в первом приближении описывается известным законом Ципфа

$$F \sim r^{-1} \qquad (I)$$

то предполагаемая степенная ("аллометрическая") зависимость между частотой и числом значений может быть переписана в виде:

$$\overline{m}_r = \alpha^{-1} r^{-\gamma} \qquad (2)$$

Однако, попытка экстраполировать зависимость (2) приводит к абсурду: при любом коэффициенте пропорциональности, начиная с какого-то r_0, оказывается, что среднее число значений у слов с данным и более высоким статусом должно быть меньше I. Остается предположить, что (2) аппроксимирует интересующую нас зависимость лишь на ограниченном интервале изменения аргумента, не превосходящем некоторого r_0. Можно было бы пытаться заменить (2) каким-либо выражением, которое с ростом r стремится к I, но нигде не достигает этого значения. Однако в данном случае более естественным представляется самое тривиальное решение: считать $\overline{m}_r$ заданной "кусочно": в интервале от I до r_0 она будет принимать значения от некоторого C до I, а для $r > r_0$ - тождественно равна I.

Смысл такого решения - введение границы (r_0), начиная с которой "разрешающая способность" словаря недостаточна для различения и противопоставления оттенков смысла. Словарь МАС, например, который располагает большей разрешающей способностью по сравнению со словарем Ожегова, последовательно приписывает словам типа ананас, апельсин, виноград, гранат и т.п. как минимум два значения, различая плодоносящее растение и его плод; тогда как словарь Ожегова столь же последовательно объединяет оба значения в одной самостоятельной части толкования. Но и у МАС разрешающая способность не безгранична: он противопоставляет значение растения и полезного продукта, получаемого из данного растения, у слов конопля, пшеница, рожь, хлопок и многих аналогичных им, но отступает от этого для слова ячмень. Предполагается, что при фиксированной "разрешающей способности" словаря у редкого слова, представленного в картотеке составителя малым количеством примеров, различается меньше значений, чем у частого. Вопрос

о том, существует ли для данного слова какой-либо верхний предел членения его значения представляется нам естественным, но преждевременным.

Что касается поведения $\bar{m}_r$ для $r \leq r_0$, то представляется, что эта функция убывает существенно медленнее, чем предполагалось ранее. Мы принимаем следующее предположение:

$$\bar{m}_r = \begin{cases} C - a \ln r & \text{, если } r \leq r_0 \\ 1 & \text{, если } r > r_0 \end{cases} \quad (3)$$

4. В предшествующем разделе мы рассмотрели, как с рангом изменяется среднее значение $\bar{m}$ случайной величины M – числа самостоятельных частей толкования. можно предположить, что M принимает значения $m = 1,2,3...$ с вероятностями

$$f_m = \binom{m+s-1}{m} p^s (1-p)^m \quad (4)$$

т.е. имеет отрицательно-биноминальное распределение с параметрами s и p, которые в общем случае зависят от ранга: $s = s(r)$, $p = p(r)$. Детальный вид этой зависимости нам неизвестен. Но некоторую информацию об изменении s и p мы можем извлечь из (3), если примем во внимание, что среднее значение $\bar{m}$ случайной величины M с распределением (4) выражается через s и p следующим образом:

$$\bar{m} = \frac{s(1-p)}{p} + 1 \quad (5)$$

Из (5) видно, что при $r > r_0$ распределение M становится вырожденным: при $p = 1$ оно сосредоточено в одной точке и имеет нулевую дисперсию. Если предположить, что с ростом ранга s убывает, стремясь к 1, и становится близким к этой величине при $r_1 < r_0$, то существует диапазон рангов от r_1 до r_0, в котором распределение (4) практически совпадает с геометрическим

$$f_m = p(1-p)^m$$

и зависит только от одного параметра p. Таким образом, форма распределения меняется от умеренно асимметричного при $s > 1$ и $p < 1$, к сильно скошенному при $s = 1$ и $p < 1$ и вырожденному при $p = 1$. Своего рода "индикатором" формы

распределения может служить доля слов с одним значением f_1 на данном интервале рангов: при $s > 1$ $f_1^{-1} > \bar{m}$, а при $s = 1$, $f_1^{-1} \approx \bar{m}$. При геометрическом распределении случайной величины M должно иметь место простое соотношение между ее средним значением $\bar{m}$ и дисперсией: последняя составляет $\bar{m}(\bar{m}-1)$.

5. Чтобы составить суждение о степени адекватности предложенной модели, мы взяли два толковых словаря русского языка - МАС и Ожегов (1986) - и два частотных - Штейнфельдт и Засориной, приписывая слову с определенным рангом из словаря Штейнфельдт число самостоятельных частей его толкования в словаре МАС, и аналогично - слову из словаря Засориной - число его значений в словаре Ожегова. Укажем наиболее существенные из принятых при этом соглашений:

а) Слова, отсутствующие в толковом словаре не учитывались при дальнейшей обработке. Исключение было сделано для многочисленных наречий, регулярно образованных от прилагательных (по-разному, пространно и т.п.), эти формы крайне непоследовательно приводятся в словаре Ожегова, мы приписывали им определенное число значений, учитывая толкование соответствующего прилагательного.

б) Если не было известно, какая из омонимичных форм представлена в частотном словаре [+], то указывалось число значений той формы, которая имеет их максимальное количество (МАС дает для склад^1 (оружия) - два значения, для склад^2 (ума) - четыре, для склад^3 (читать по складам) - одно значение, следовательно, мы приписывали форме склад четыре значения).

в) Если слово является ссылочным, то ему приписывалось столько же значений, сколько слову, на которое была дана ссылка, если в словаре не было на этот счет дополнительных указаний. Глаголу ухать приписано столько же значений сколько имеет совершенный вид этого глагола ухнуть (в МАС - 4, у Ожегова - 7); толкование слова учительница содержит ссылку на учитель, но только в одном из двух значений последнего.

Словарь Штейнфельдт (2 500 слов) и часть словаря Засориной (9 000, наиболее частых слов) были разбиты на группы, объединяющие слова с последовательными рангами. Все слова с

[+] В словарях Штейнфельдт и Засориной отдельно осуществляется подсчет для некоторых омонимов, но очень ограниченного числа.

одинаковой частотой употребления были отнесены к одной группе, размеры группы были не менее 100 слов. Для каждой группы было вычислено среднее число значений у составляющих ее слов и дисперсия числа значений в группе. Диаграмма изменения среднего числа $\bar{m}_r$ значений в обоих словарях приведена на рис. 1, числовые данные (для укрупненных групп) приводятся в таблице 1.

6. Далее мы оценили методом наименьших квадратов параметры C и α в уравнении (3), исключив при этом данные о первых, самых употребительных 100 словах, так как дисперсия на этом интервале рангов существенно превышала дисперсию на остальном участке изменения $\bar{m}_r$. Для словарей Ожегова/Засорина уравнение (3) приняло вид:

$$\bar{m}_r = 6,882 - 0,587 \ln r, \qquad (3')$$

а для словарей Штейнфельдт/МАС

$$\bar{m}_r = 7,128 - 0,565 \ln r \qquad (3'')$$

Пунктиром на рис. 1 показано изменение величины f_r^{-1}, которая для словаря Ожегова при $r \approx 8\,000$ уже очень близко подходит к прямой (3'). На рис. 2 приводится распределение числа значений для слов, которые имеют в словаре Засориной частоту 11 (ранг от 7 900 до 8 400). Пунктиром дано распределение случайной величины, подчиняющейся геометрическому закону, с тем же средним значением (хотя на глаз сходство очень большое, вероятность чисто случайного отклонения наблюдаемого распределения от теоретической модели мала: $\chi^2 = 11,2$ при четырех степенях свободы). Для сравнения приводится распределение по числу значений слов с рангом 601–700.

Экстраполируя (3') и (3''), мы можем оценить значение r_0. Поскольку для словаря Штейнфельдт "база экстраполяции" существенно уже, чем для словаря Засориной, мы решили проверить состоятельность прогноза, опирающегося на (3''), дополнительно обработав с помощью МАС, т.е. по той же методике, что и первые 2 500 слов, еще примерно 1000 слов с частотой 10–11 (7928 – 9036) из словаря Засориной и случайную выборку слов с частотой 1 (около 400 единиц) из того же словаря (средний ранг около 30 000). Найденные параметры всех трех дополнительных выборок (средние значения нанесены на диаграм-

му) мало отличаются от прогнозируемых (частота II: $\bar{m}$ =2,II, дисперсия I,85, частота I0: $\bar{m}$ = 2,07, дисперсия I,94, частота I - $\bar{m}$ = I,42, дисперсия - 0,4I).

На основе (3′) и (3″) мы получаем для МАС $\tau_0 \approx$ 50 000, для Ожегова $\tau_0 \approx$ 22 000.

7. Значение τ_0 могло бы играть роль количественной характеристики "разрешающей силы" словаря - признака, явно относящегося к изменительному инструменту, а не к объекту измерения. Однако τ_0 - очень грубая характеристика, так как она не несет никакой информации о том, как распределение случайной величины M стремится к своей предельной форме, а лишь о границе, на которой оно его достигает. Для МАС, разрешающая способность которого заведомо выше, чем у словаря Ожегова, больше не только τ_0, но и параметр s стремится к единице медленнее, чем для Ожегова (он достоверно равен I только для одноразовых слов в словаре Засориной).

Однако исследование динамики s (и p) представляет собой отдельную задачу, поскольку оценки этих параметров методом моментов имеют большую дисперсию, а оценки максимального правдоподобия требуют проведения громоздких вычислений. Собранный нами материал дает основание предположить (хотя это предположение нуждается, очевидно, в дальнейшей проверке), что инвариантной характеристикой самой лексики, а не определенного частотного или толкового словаря, является скорость убывания многозначной лексики, характеризующаяся угловым коэффициентом в выражении (3).

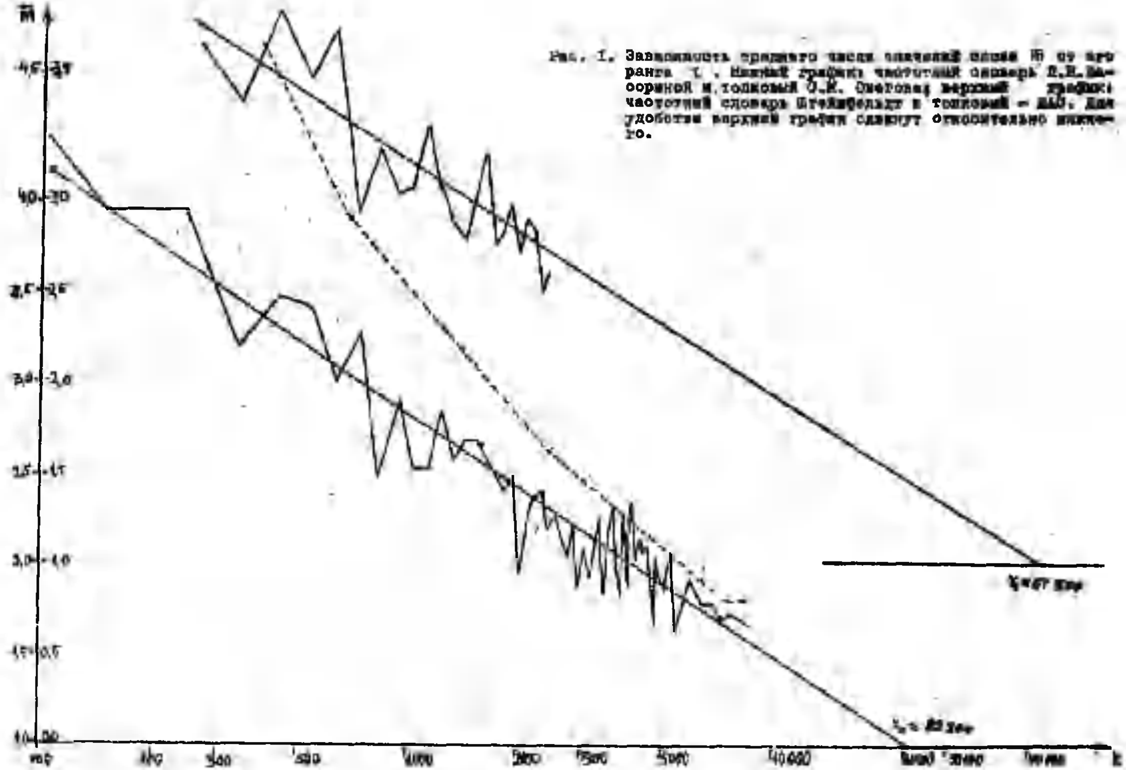

Рис. 1. Зависимость среднего числа значений слова $\bar{m}$ от его ранга i. Нижний график: частотный словарь Л.Н. Засориной и толковый С.И. Ожегова; верхний график: частотный словарь Штейнфельдт и толковый – БАС. Для удобства верхний график сдвинут относительно нижнего.

Таблица I.

Употребительность и среднее число значений у слова

Ранговый интервал	Словари			
	Штейнфельдт/МАС		Засорина/Ожегов	
	Среднее	Дисперсия	Среднее	Дисперсия
I - I00	6,38	38,23	4,73	I5,23
I0I - 200	4,27	II,I9	3,93	6,I0
20I - 300	3,98	7,79	3,94	6,47
30I - 400	3,54	6,70	3,2I	7,48
40I - 500	4,03	9,89	3,48	7,37
50I - 600	3,66	7,I0	3,40	5,36
60I - 700	3,95	9,43	3,00	4,38
70I - 800	2,94	4,28	3,27	7,22
80I - 900	3,29	6,03	2,49	2,90
90I - I000	3,03	4,67	2,88	5,55
I00I - II00	3,08	4,I6	2,53	2,4I
II0I - I200	3,4I	5,89	2,53	2,62
I20I - I300	3,06	4,82	2,84	4,I6
I30I - I400	2,84	3,55	2,58	3,I2
I40I - I500	2,77	3,96	2,67	4,5I
I50I - I600	3,04	6,47	2,67	4,I6
I60I - I700	3,26	6,63	2,58	3,06
I70I - I800	2,75	3,82	2,50	4,40
I80I - I900	2,82	3,34	2,4I	2,26
I90I - 2000	2,99	5,09	2,47	2,90
200I - 2500	2,70	3,20	2,23	2,07
250I - 3000	-	-	2,I0	I,56
300I - 4000	-	-	2,05	I,9I
400I - 5000	-	-	2,0I	I,83
500I - 6000	-	-	I,82	I,20
600I - 7000	-	-	I,82	I,36
700I - 8000	-	-	I,70	I,08
800I - 9036	-	-	I,68	0,9I

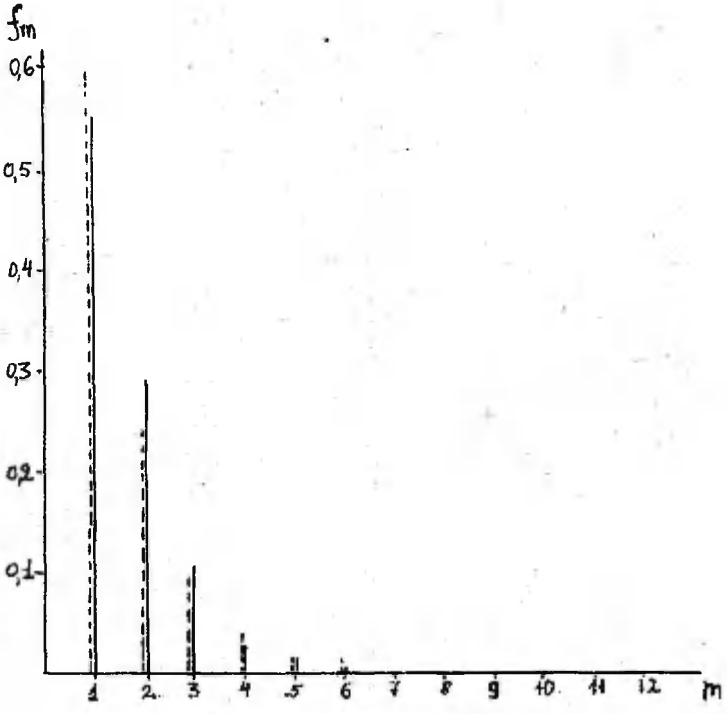

Рис. 2. Распределение слов по числу значений (m) по данным словарей Засорина/Ожегов для рангового интервала 7900–8400 (частота 11), пунктиром дано распределение случайной величины с тем же средним ($\bar{m}$ =1,68), подчиняющейся геометрическому закону. Ломаной указано распределение по числу значений слов из рангового интервала 601 – 700.

Л И Т Е Р А Т У Р А

Апресян Ю.Д. Лексическая семантика. Синонимические средства языка. – М.: Наука, 1974.

Андрукович П.Ф., Королев Э.И. О статистических и лексикографических свойствах слов. – Научно-техническая информация. Сер. 2, 1977, № 4, с. 1–9.

Гиндин С.И. Частота слова и его значимость в системе языка. – В кн.: Лингвостатистика и вычислительная лингвистика. Труды по лингвостатистике. Вып. 8. Тарту, 1982, с. 22–53.

Засорина Л.Н. Частотный словарь русского языка. – М.: Русский язык, 1977.

Ожегов С.И. Словарь русского языка. – М.: Русский язык, 1986.

МАС – Словарь русского языка в четырех томах. – М.: Русский язык, 1981 – 1984.

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. – Уч. зап. Тартуск. ун-та, вып. 502, Тарту, 1979, с. 107–141.

Штейнфельдт Э.А. Частотный словарь современного русского литературного языка. – Таллин, 1963.

Guiraud P. Les caractères statistiques du vocabulaire. – P.: Press universitaires, 1954.

Zipf G.K. The meaning-frequency relationship of words. – The journal of general psychology, 1945, v. 33, p. 2, p. 251–256.

USUALNESS AND POLYSEMY OF WORDS

Mikhail V. Arapov

S u m m a r y

The paper studies the stochastic relationship between the number of separate meanings of a word and its rank in the frequency dictionary (the rank is treated as an index of the word's usualness). The linear dependence of the mean number of meanings upon the logarithm of the rank is suggested. The slope of the graph is supposedly independent of the text sample and the dictionary as a source of definitions, and the place of intersection between the declining graph and the abscissa depends on the properties of that source (its 'discerning power'). The negative-binominal distribution of the number of meanings postulated for the most frequently used words subsequently changes into the geometric and the degenerated distribution for rare words.

ИНЖЕНЕРНАЯ ЛИНГВИСТИКА И СИСТЕМЫ "ПОНИМАНИЯ" ТЕКСТА

М.С. Бдехман

Последние годы развития языкознания в значительной степени характеризуются становлением инженерно-лингвистической методологии исследования и описания языка. Это связано как со стремлением языковедов строить модели, воспроизводящие реальные языковые факты, так и с объективной необходимостью разработки и внедрения в промышленность эффективных систем автоматической обработки текстовой информации (далее в тексте - ИС, информационные системы). Этим определяется "двуединый" характер инженерной лингвистики: в ее компетенцию входят, с одной стороны, вопросы общеметодологического характера - в первую очередь, построение методологии инженерно-лингвистического эксперимента для проверки лингвистических гипотез (ср. Пиотровский Р.Г., 1985), а с другой стороны - разработка конкретных процедур автоматической обработки информации. Настоящая работа принадлежит обоим направлениям. Для того, чтобы сформулировать ее цель, необходимо сделать следующее вводное замечание.

В настоящее время наметились два принципиальных подхода к построению лингвистического обеспечения ИС: (а) создание сравнительно простых систем, использующих минимально необходимые сведения о языке; (б) создание сложных систем, использующих максимально возможные сведения. Нам кажется беспредметным спор о монополии одного из этих подходов; единственно конструктивным представляется выявление классов лингвистических задач, для решения которых тот или иной подход является оптимальным. Не менее важно, с нашей точки зрения, выявить круг задач, решение которых требует человеческого интеллекта и принципиально плохо "поддается" машинным алгоритмам.

Целью данной работы является оценка задач инженерной лингвистики как в связи с построением реальных промышленных ИС, так и в связи с задачами теоретико-лингвистического характера. Для ответа на первый круг вопросов в работе делается попытка ранжировать степени "понимания" текста машиной, т.е. выявить уровни этого "понимания", сравнив его с человеческим. Затем на материале конкретных ИС, обеспечивающих определенный уровень "понимания" текста, выявляются некоторые актуальные задачи инженерной лингвистики и намечаются пути

развития инженерно-лингвистической методологии. Для ответа на второй круг вопросов в работе анализируется проблема соотношения лингвистической гипотезы и инженерно-лингвистического эксперимента.

Уровни "понимания" текста информационной системой

Прежде чем приступить к ранжированию степеней "понимания" текста машиной, зададим критерий этого ранжирования. По всей вероятности, при определении такого критерия следует исходить из сути информационной системы. Дело в том, что главная (если не единственная) реальная цель создания ИС любого типа - это удовлетворение информационных потребностей человека- специалиста, и, следовательно, оценивать уровень "понимания" текста информационной системой нужно именно с точки зрения пользователя системы.

Назовем уровнем "понимания" (УП) текста информационной системой такую переработку этого текста, которая обеспечивает определенную степень удобства для пользователя, стремящегося удовлетворить всю информационную потребность. Точнее, будем считать, что УП тем выше, чем больший процент имеющейся в тексте информации пользователь может получить от информационной системы. Выбор такого критерия кажется нам вполне логичным, т.к. только пользователь системы в состоянии дать объективную оценку УП, причем ясно, что, чем выше УП, тем больше в обработке информации система берет на себя и тем меньше оставляет пользователю-человеку.

Очевидно, практически максимальным был бы такой уровень "понимания" текста системой, который соответствовал бы уровню понимания его человеком-специалистом в данной области знаний. При этом специалист-пользователь мог бы получить от "специалиста"-ИС практически всю информацию, заложенную в текст автором, независимо от степени эксплицитности ее выражения. Нам представляется, однако, что между уровнем понимания текста человеком-специалистом и ИС любой мыслимой мощности имеется существенное различие, проявляющееся в том, что практически любой текст содержит, в частности, такую информацию, которая в принципе не может быть выявлена никакой информационной системой. Причиной этого "непонимания" является то, что человек понимает и план выражения, и план содержания текста,тогда как ИС "понимает" (с той или иной степенью глубины) только план выражения, и никакое лингвистическое обеспечение, каким бы мощным оно ни было, не может помочь системе понять до конца

план содержания, т.е. смысл текста. В связи с этим мы расцениваем как метафорические выражения типа "распознавание смысла текста информационной системой", поскольку смысл текста во всех случаях остается неуловимым для ЭВМ, а это, в свою очередь, означает невозможность истинного понимания текста машиной.

В самом деле, для того, чтобы машина могла понять смысл, мы должны предварительно объяснить ей, что это такое, и описать смысл языковых единиц и механизмов. Однако любое задание смысла при этом оказывается описательным, тогда как для понимания машиной смысла необходимо содержательное его задание. Со времен Ф. де Соссюра лингвисты, вслед за основоположником структурализма, пытаются описать смысл языковых единиц через их место в системе языка и/или структуре текста, полагая, что смысл единицы - это сумма ее окружений, некоторая комбинация (иерархия) элементарных семантических признаков и т.п. При этом смысл окружающих и составляющих единиц не задается, так что описание не носит содержательного характера. Можно, например, сказать, что семантический множитель "каузировать" входит в значение слов "убрать", "уничтожить", "заставить", но как объяснить системе, что такое "каузировать"?

Иными словами, чем более глубокое ("глубинное") описание семантики мы задаем, тем очевиднее становится порочный круг, к которому сводится этот процесс: для описания смысла единицы А необходимо аксиоматически задать смысл Б, а при необходимости описать смысл Б - аксиоматически задается смысл А. Так, например, значение слова "пóры" можно объяснить при помощи слов "дырочки", "отверстия" и т.п., но для объяснения смысла этих последних, в свою очередь, необходимо привлечение смысла слова "пóры". При таком описании смысла вряд ли можно ожидать понимания машиной смысла выражений "беспористый материал", "усадка детали" и т.п.

Разумеется, еще сложнее обстоит дело с описанием смысла синтаксических и, тем более, гиперсинтаксических отношений. В самом деле, как описать смысл причинно-следственных отношений между событиями? Возможно, так: "Если наступило событие А, то с вероятностью, равной единице, наступит и событие Б"? Однако такое утверждение по сути аналогично утверждению о том, что понедельник является причиной вторника.

В известной монографии Р. Шенка (1979) предлагается считать, что ИС понимает текст, если может (а) перифразиро-

вать его и (б) "вычислить" все его пресуппозиции. С нашей точки зрения, эти критерии "не работают" на реальных текстах. Дело в том, что для вычисления пресуппозиций одного знания языка оказывается явно недостаточно, ведь человек при вычислении пресуппозиций использует такие плохо формализуемые понятия, как, например, здравый смысл. Например, для высказывания "От Иванова ушла любимая жена" мы не строим пресуппозиции "Жена Иванова не была разбита параличом, т.е. могла ходить", "Иванов не держал жену взаперти" и т.п. С другой стороны, мы легко ответим на вопрос "Хорошо ли теперь Иванову?"

Итак, мы приходим к важному предположению: чрезвычайно сложным, если вообще возможным является создание ИС, для функционирования которых машине необходимо проявить реальное понимание текста, т.е. проникновение в его план содержания. ИС такого типа образуют гипотетический класс систем, который мы назовем классом систем высшего уровня, а понимание ими текста назовем высшим уровнем понимания. Приведем примеры таких гипотетических систем:

- ИС, самообучающиеся путем чтения научных текстов и выявляющие их новизну;
- ИС, проверяющие логику изложения в научных текстах;
- ИС, отвечающие на вопросы по научному тексту, требующие вычисления пресуппозиций.

Невозможность построения реальных (а не модельных, игрушечных) систем высшего класса объясняется, с нашей точки зрения, в первую очередь, тем, что такое понимание текста не допускает естественного для современных систем отрыва означающего от означаемого. Иными словами, никакое глубинно-синтаксическое и глубинно-семантическое описание текста не может компенсировать машине отсутствие у нее "доступа" к реальному явлению, лежащему в основе этого текста. Пользуясь метафорой Льюиса Кэрролла, можно сказать, что означающее без означаемого - это улыбка без кота. При отсутствии же этого "кота" любой удачный ответ системы на вопросы человека будет по сути...случайным,т.е.не обусловленным действительным пониманием описываемого текстом фрагмента действительности. Возьмем, к примеру, предложение из "Алисы в Зазеркалье": Twaa brillig. На вопрос How was it? система отвечает: Brillig, демонстрируя такое же "понимание" текста, как известная ИС "Элиза" Дж. Вейценбаума (1970). Ясно, что, чем сложнее (в языковом и смысловом планах) будут во-

просы, контролирующие понимание системой смысла текста, тем более явным будет полное непонимание ею такового. Вдумаемся, например, можно ли требовать от ИС анализа правильности логики авторов в нижеследующем рассуждении, если не объяснить ей содержательно значение причинно-следственного отношения между высказываниями:

"Исследования /I/, /2/ показали, что между особенностями синтаксической структуры элементов текста и информацией, заключенной в них, имеется некоторая связь. Тогда для выявления существенных элементов информации можно использовать синтаксическую структуру предложений." (Колтун А.Я., Пшеничная Л.Э. Использование терминов заглавия для автоматического реферирования текста научного документа. – Автоматическая обработка текста. Препринт 80-24 АН УССР, 1980, № 24, с.29).

Главной причиной такого "отторжения" (термин Р.Г. Пиотровского, 1979) языка машиной является, по нашему мнению, антропоморфизм (Степанов Ю.С., 1975) языка, его полная ориентированность на человека и, как следствие непонятность автомату, лишенному "человеческого фактора". В самом деле, любая грамматическая или функционально-грамматическая категория, любое синтаксическое отношение, любое семантическое обобщение (любое слово – уже абстракция – В.И. Ленин) "подогнаны" под человека, под его мировосприятие, удобны для человека, и только для него. Например, мы различаем выделенный и невыделенный из класса объекты, но не различаем "I97-й и не I97-й с конца". Мы используем понятия субъекта, предиката и т.п., потому что нам так привычно и удобно. Именно привычкой и удобством, вызванными "структурой" нашего мира и особенностями жизни человека в нем, объясняются эти и бесчисленные другие особенности человеческого языка. Человек постоянно сталкивается с причинно-следственными отношениями в окружающей жизни, поэтому они находят отражение в языке; если бы я не знал, что значит "поэтому", я бы не понял высказывания "Я мыслю, следовательно, я существую". Если бы я не знал что значит "очень", я бы не понял, что значит "Я очень люблю Баха". Для понимания того, что значит "скамейка", нужно владеть человеческой потребностью и способностью к обобщению. В самом деле, почему в один класс объектов объединяются именно скамейки, а не скамейки и лошади, ведь и у тех, и у других есть ноги, на них можно сидеть и т.п.? (Кстати, на воровском жаргоне "скамейкой" называется именно лошадь). В каждом языковом знаке, в каждой синтаксической структуре, высказывании,

5

тексте ярко проявляются также сугубо человеческие, "неуловимые" понятия, как "полезность", "удобство", "здравый смысл" и т.п. Говоря "У него голова, как мяч", мы имеем в виду не наличие ирама (похожего на шнуровку мяча), а скорее только форму головы. Говоря "Он пошел к врачу", мы скорее имеем в виду, что он болен, а не, скажем, отправился свести счеты с приятелем своей жены. Примеры эти можно было бы продолжить до бесконечности.

Как видим, способ организации языковых единиц и отношений в систему диктуется "сутью" человека и окружающего мира. Однако глубина этих понятий недоступна до конца даже самому человеку, не говоря уже об ЭВИ, для которой человека просто не существует и, следовательно, не существует потребности понимать то, что понимает человек. Отсутствие же такой потребности приводит к принципиальной невозможности понимать текст.

Таким образом, мы утверждаем невозможность построения систем высшего класса, ограничивая возможности ЭВМ "сверху". В то же время, наличие действующих ИС свидетельствует о том, что машина каким-то образом "понимает" текст. С другой стороны, и человек зачастую обрабатывает текст, фактически не понимая его. В частности, можно неплохо перевести текст с одного языка на другой, не имея ни малейшего понятия о сути описываемого в этом тексте явления, т.е. о плане содержания данного текста (многолетний опыт работы в качестве переводчика научно-технической литературы позволяет автору утверждать это).

Итак, машина "понимает" текст, не понимая его. Каковы же уровни этого "понимания"?

"Понимание" текста машиной

В этом разделе мы используем опыт разработки промышленных действующих систем для анализа вопросов, связанных с машинным "пониманием" текста. Промышленной ИС мы называем систему, работающую с текстами, не подготовленными специально для данной ИС, и имеющую конкретных пользователей. Мы не обращаемся к примеру различных экспериментальных "игрушечных" систем, т.к. их статус не позволяет, как нам кажется, делать окончательные выводы о характере "понимания" ими текста.

ИС - это такая система, в которой текст некоторым образом обрабатывается с целью удовлетворения информационных потребностей пользователя. В зависимости от этих потребностей

ЭВМ тем или иным образом "понимает" текст, не достигая, как мы видели, максимального уровня понимания. При этом нельзя просто сказать, что в одном случае система "понимает" текст лучше, чем в другом, если мы имеем дело с системами, удовлетворяющими разным потребностям пользователя (например, поисковой и переводческой), т.к. уровень понимания зависит от его близости к максимальному при решении данной задачи. Например, нельзя говорить, что система машинного перевода, использующая модель СМЫСЛ - ТЕКСТ, "понимает" текст лучше, чем система автоматического индексирования с "мелочной" грамматикой: она "понимает" его иначе. Важно, однако, что для достижения одного и того же уровня "понимания" в системах, решающих разные задачи, могут потребоваться принципиально разные средства описания языка[+].

Итак, ИС, анализируя текст, "понимает" его в том смысле, который вкладывается в слово "понимание" в системах данного типа, и преобразует в некоторую выходную запись. При этом возможны такие ситуации:

а) в выходной записи нет ничего, что не присутствовало бы в явном виде в оригинале или могло быть заранее поставлено в прямое соответствие его элементам, - нулевой уровень "понимания" ($УП_0$);

б) выходная запись содержит информацию, в явном виде не содержащуюся в тексте, т.е. некоторую имплицитно присутствующую в исходном тексте информацию, - первый уровень "понимания" ($УП_I$).

Сходство $УП_0$ и $УП_I$ заключается в том, что системы, обладающие способностью "понимать" текст на одном из этих уровней, не в состоянии, тем не менее, извлекать из него такую имплицитную информацию, выявление которой в тексте требует сугубо "человеческого" знания мира. Сходство же $УП_I$ и высшего уровня понимания заключается в том, что в обоих случаях понимание текста выражается в извлечении из текста некоторой имплицитной информации.

[+] В связи с этим можно высказать сомнение в справедливости предлагаемого Р. Шенком (1979) критерия "понимания" машиной текста как к способности к перифразированию и "вычислению" пресуппозиций. Этот критерий, как нам кажется, полностью определяется типом разработанной под руководством Р. Шенка экспериментальной системы "Марджи", тогда как, скажем, для системы индексирования данный критерий оказывается избыточным.

Как на первом, так и на нулевом уровне "понимание" текста может быть

а) "морфологическим" (система "понимает" морфологические характеристики лексических единиц);

б) "синтаксическим" (система "понимает" синтаксические отношения лексических единиц в предложении);

в) "семантическим" (система "понимает" семантические характеристики лексических единиц);

г) "гиперсинтаксическим" (система "понимает" гиперсинтаксические отношения между предложениями текста).

Разумеется, одна и та же ИС может обладать способностью к комбинации указанных разновидностей "понимания".

Эти разновидности "понимания" текста, точнее способы его "понимания", характеризуют только аппарат анализа текста, но никак не уровень "понимания". В частности, использование в одной ИС "семантико-синтаксического" понимания, а в другой, выполняющей сходный функции, "морфологического понимания", не дает нам права утверждать, что УП первой системы выше, чем второй.

Таким образом в дальнейшем мы сможем классифицировать любую ИС в зависимости от уровня и способа "понимания" ею текста.

Теперь, когда определены основные понятия, необходимые нам для дальнейшего изложения, перейдем к освещению некоторых теоретических и практических проблем, стоящих перед инженерно-лингвистической методологией с точки зрения выполнения ее важного социального заказа - построения систем, "понимающих" текст.

Задачи инженерной лингвистики

При построении промышленной ИС, ориентированной на решение некоторой конкретной задачи или круга задач, необходимо, разумеется, выбрать оптимальные уровень и способ "понимания" текста системой для решения данного круга задач. Выбирая тот или иной уровень, тот или иной способ "понимания", лингвист закладывает в основу разрабатываемой системы некоторую совокупность лингвистических гипотез. Таким образом, использование инженерно-лингвистической методологии при построении ИС способствует одновременно решению двух взаимосвязанных задач: выбору оптимального лингвистического обеспечения системы и проверке лингвистических гипотез, лежащих в основе данного лингвистического обеспечения. При этом линг-

вист, разрабатывающий лингвистическое обеспечение ИС, должен осознать, на каких гипотезах базируется разработка, и эксплицитно сформулировать эти гипотезы. Последние должны быть проверены посредством инженерно-лингвистического эксперимента для определения степени соответствия каждой гипотезы действительности. При недостаточно высокой (для решения поставленных практических задач) степени соответствия необходимо выбрать другой способ и/или уровень "понимания", что требует формулирования новых гипотез и экспериментальной их проверки. Подчеркнем очень важное, на наш взгляд, положение. Мы считаем, что такое существенное для информатики понятие, как качество работы ИС (качество реферирования, индексирования, перевода и т.п.), не носит абсолютного характера и сильно зависит от конечной цели создания системы. Так, качество автоматического реферирования в значительной степени определяется тем, кто именно и для чего будет использовать получаемый от ИС реферат. Иными словами, в понятие "качество" входит несколько составляющих, каждая из которых соответствует некоторой лингвистической гипотезе (группе гипотез), проверяемой в ходе эксперимента, ориентированного на проверку данной конкретной гипотезы. Таким образом, "центр тяжести" лингвистических исследований переносится на эксперимент. Именно необходимость экспериментальной проверки требует эксплицитной формулировки гипотезы, а иногда и определяет характер модели, основывающейся на этой гипотезе. Поэтому необходимо разработать методологию лингвистического эксперимента (ср. Пиотровский Р.Г., 1985), которая способствовала бы решению информационных и теоретико-лингвистических задач.

Перечислим требования, предъявляемые нами лингвистическому эксперименту.

1. Эксперимент должен проводиться на реальных, не подготовленных специально текстах.

2. В процессе эксперимента в явном виде должна проверяться каждая из эксплицитно и однозначно сформулированных гипотез.

3. Эксперимент должен периодически повторяться для контроля адаптации системы к возможным принципиальным изменениям структуры поступающих на ее вход текстов. Это требование распространяется на системы, стабильно функционирующие в промышленном режиме.

Суть лингвистического эксперимента заключается в следующем. Информационная система отрабатывает неподготовленные

заранее тексты, используя тот уровень и способ "понимания", в основе которых лежит эксплицитно сформулированная гипотеза о соответствующем лингвистическом объекте. Результаты работы ИС предъявляются либо лингвисту для прямой проверки гипотезы, либо конечному пользователю - для косвенной ее проверки. При прямой проверке эксперимент проводится в терминах гипотезы, а при косвенной - в терминах оценки пользователем качества работы ИС в целом. Иными словами, при прямой проверке эксперимент ведется в терминах лингвиста, а при косвенной - в терминах пользователя. Косвенная проверка отличается от прямой тем, что лингвист осуществляет ее на основании оценки пользователем, ничего не знающим о проверяемой гипотезе, результатов работы ИС, и именно на основании этой оценки лингвист оценивает свою гипотезу.

Проиллюстрируем каждый из этих методов эксперимента конкретными примерами.

Прямая проверка гипотез

Нами была сформулирована и подвергнута прямой проверке гипотеза о текстообразующем механизме категории определенности в английском научном тексте. Гипотеза была сформулирована следующим образом (см. Блехман М.С., 1985).

I) Механизм соотнесения грамматически определенного объекта с антецедентом, выступающим в качестве "адреса" этого объекта в классе подобных, лежит в основе маркированной денотативной связи предложений английского научного текста - А-связи предложений.

2) В зависимости от характера соотнесения объекта с его "адресом" в классе подобных, различаются следующие разновидности А-связи:

- эксплицитная прямая полная;
- эксплицитная прямая частичная;
- эксплицитная непрямая полная;
- эксплицитная непрямая частичная;
- тезаурусная прямая полная;
- тезаурусная прямая частичная;
- тезаурусная непрямая полная;
- тезаурусная непрямая частичная;
- списочная прямая полная;
- списочная непрямая полная;
- списочная непрямая частичная;
- имплицитная прямая;
- имплицитная непрямая.

3) А-связь предложений участвует в формировании синтаксической сверхфразовой структуры научного текста.[+]

С точки зрения инженерной лингвистики, эта гипотеза может иметь силу только при условии экспериментальной проверки ее состоятельности. Для осуществления такой проверки был построен аналог исследованного объекта - гипотетическая модель А-связи (Блехман М.С., 1985), после чего эта модель была представлена в виде алгоритма вявления А-связей в английских научных текстах. Алгоритм был ориентирован на функционирование в реальных ИС, осуществляющих автоматическое квазиреферирование, машинный перевод и автоматическое индексирование английских текстов. Все эти системы обладают нулевым уровнем "понимания" текста и используют в основном морфологический способ "понимания". Алгоритм базируется на формальном аппарате описания эксплицитной, тезаурусной и списочной А-связей и не позволяет выявлять в текстах имплицитную А-связь.

Эксперимент заключался в непосредственном анализе правильных и ошибочных решений алгоритма в каждой из указанных выше ИС и выявлении причин ошибок. При этом оказалось, что основная часть ошибок (т.е. выявление ложных "А-связей" либо невыявление истинных) вызвана морфологическим способом "понимания" текста данными системами, а не ошибками собственно модели А-связей, вызванными несовершенством сформулированной гипотезы.

Инженерная реализация модели подтвердила следующие положения сформулированной гипотезы.

I) Механизм соотнесения грамматически определенного объекта с антецедентом, выступающим в качестве "адреса" этого объекта в классе подобных, лежит в основе маркированной денотативной связи предложений английского научного текста. Данное предположение подтверждается тем, что реальные информационные системы, использующие аналог моделируемого лингвистического объекта, со сравнительно высокой надежностью выявляют в произвольно взятых научных текстах объективно существующие в них А-связи предложений, причем степень надежности, вероятно, может быть повышена при использовании синтаксического и семантического способов "понимания" текста данными ИС.

[+] Подробное изложение гипотезы с примерами А-связей см. Блехман М.С., 1985.

2) В зависимости от характера соотнесения объекта с его антецедентом - "адресом" в классе подобных объектов - различаются указанные выше разновидности эксплицитной, тезаурусной и списочной А-связи предложений. Данное предположение подтверждается тем, что используемый формальный аппарат, задающий условия существования в тексте каждой из разновидностей А-связи, кроме имплицитной, позволяет машине распознавать эти разновидности в реальных текстах при квазиреферировании, переводе и индексировании этих текстов. В частности, редактирование текстов в системе машинного перевода (Блехман М.С., 1985) основано на выявлении в них эксплицитной и списочной прямой полной А-связей и может быть распространено на тезаурусную прямую полную А-связь. Однако действие модели не распространяется на имплицитную А-связь, инженерному моделированию которой должно предшествовать углубленное теоретическое исследование ее механизмов.

3) А-связь двух предложений участвует в формировании синтаксической сверхфразовой структуры английского научного текста. Данное предположение подтверждается тем, что ИС, осуществляющая квазиреферирование английских текстов, выявляет в процессе распознавания сверхфразовой синтаксической структуры обрабатываемых текстов А-связи, удовлетворяющие определению синтаксической сверхфразовой (гиперсинтаксической) связи (см. Добрускина Э.М., Берсон В.Е., 1986), а именно:

- А-связь основана на насыщении синсемантичного предложения, содержащего грамматически определенный объект, предложением, содержащим "адрес" этого объекта в классе подобных;

- А-связью, выявляемой информационной системой, связываются предложения, находящиеся в логико-смысловой сверхфразовой связи того или иного типа.

Косвенная проверка гипотез

В предыдущем разделе мы проиллюстрировали принцип прямой проверки лингвистической гипотезы. Оказывается, однако, что применение этого способа к некотрым лингвистическим объектам является неэффективным, не позволяет получить надежную оценку сформулированной гипотезы. Это происходит в тех случаях, когда моделируемый лингвистический объект плохо поддается непосредственному наблюдению. Важно

подчеркнуть, что недоступность моделируемого объекта прямому наблюдению неизбежно приводит к субъективности в описании этого объекта. Так, например, гипотеза о возможности представления значения слова в виде совокупности "элементарных смыслов" нуждается именно в косвенной проверке, которая позволила бы объективно оценить правильность такого представления.

Проиллюстрируем теперь метод косвенной проверки на примере гипотезы о сверхфразовой синтаксической структуре английских текстов газетных информационных сообщений.

Газетное информационное сообщение - это текст длиной от 3 до 20 предложений, типичный для современной английской газеты. Гипотеза о его гиперсинтаксической структуре была сформулирована нами в следующем виде.

1) Текст газетного сообщения образует гиперсинтаксическую структуру (Добрускина Э.М., Берсон В.В., 1986).

2) Элементами этой структуры являются:

а) предложения (нижний уровень),

б) маркированные (имеющие специальное выражение) сверхфразовые единицы (МСЕ) - высший уровень.

3) В состав МСЕ входят: одно автосемантичное, т.е. не имеющее маркеров зависимости от контекста, предложение, а также, возможно, некоторое количество синсемантичных предложений (ср. Зарубина Н.Д., 1977).

4) Первое, автосемантичное предложение МСЕ является топиковым для всей единицы, т.е. содержит квинтэссенцию информации, заключенной в данной МСЕ, являясь своего рода "аннотацией" МСЕ.

5) Первое предложение в цепочке автосемантичных предложений является топиковым, т.е. несет ту же нагрузку, что и первое предложение МСЕ.

6) Количество топиковых предложений текста составляет не более 25% всех предложений этого текста.

7) Из топиковых предложений может быть составлен такой новый текст, который образует гиперсинтаксическую структуру.

Эта гипотеза легла в основу системы автоматического квазиреферирования текстов английских газетных сообщений. Система была реализована в качестве подсистемы многофункциональной ИС, разработанной в ЛГПИ им. А.И. Герцена (см. Чикаковский В.А., Беляева Л.Н., 1983). Текст для реферирования поступает после предварительной обработки, заключающейся в форматизации и разбивке на предложения. Некоторые

сообщения представляют собой формально самостоятельные фрагменты более крупных сообщений.

Для проверки указанной гипотезы была построена формальная модель сверхфразовой структуры текста, использующая аппарат коннекторов и квазиконнекторов (см. Добрускина Э.М., Берзон В.Е., 1986; Блехман М.О., 1984). На основании данного аппарата была разработана система квазиреферирования, извлекающая из исходного текста топиковые предложения и формирующая квазирефераты двух типов: (а) с указанием смысловых классов удаленных предложений (Добрускина Э.М., Берзон В.Е., 1986) - $УП_I$; (б) без указания этих классов - $УП_0$ (ср. Берзон В.Е. и др., 1984). Предложение считается топиковым, если не содержит коннекторов и квазиконнекторов и является первым в МСЕ либо в цепочке автосемантичных предложений. Система использует морфологический и гиперсинтаксический способы "понимания" текста.

Проверка гипотезы осуществлялась на массиве 17 произвольно отобранных газетных сообщений. Были введены следующие качественные характеристики квазирефератов:

а) полнота (с точки зрения передачи основного содержания документа);

б) точность (отсутствие в квазиреферате предложений, избыточных с точки зрения передачи основного содержания документа);

в) связность (в обычном смысле).

Были также введены следующие количественные оценки для каждой из перечисленных характеристик квазирефератов: "отлично" (5 баллов), "хорошо" (4), "удовлетворительно" (3), "плохо" (2), "очень плохо" (1).

Квазирефераты оценивались специалистом, знающим английский язык, но не знакомым с содержанием исходного текста. Оценки выставлялись исключительно с точки зрения будущего пользователя системы, в предположении, что квазиреферат в идеале должен иметь статус самостоятельного документа, т.е. давать пользователю четкое представление о теме исходного документа, информировать об основном его содержании, но не содержать при этом "лишней" (избыточной) информации.

Обрабатываемые документы были разделены нами на два класса: (а) поддающиеся естественному реферированию и (б) не поддающиеся таковому (например, перечень спортивных результатов). Оценки качества квазирефератов текстов обоих классов приведены в табл. 1 и 2.

Таблица I

Качество квазирефератов текстов, поддающихся естественному реферированию

№ текста	Коэфф. сжатия	Оценка полноты	Оценка точности	Оценка связности	Причины ошибок
I	3	3	4	3	Неправильная разбивка на предложения
3	3	4	5	5	
4	4	5	4	4	Непр. разбивка
5	4	5	5	3	Искусственная разбивка на формально самостоятельные фрагменты
6	5	5	5	4	- " -
7	3	5	5	5	
9	4,5	5	4	3	ошибка системы
I0	4	3	4	2	ошибки системы
II	4	5	5	4	
I2	4	5	4	3	искусств. разбивка
I3	4	5	4	4	- " -
I5	4	5	4	5	ошибка системы
I6	4	5	4	4	
I7	3	4	4	4	непр. разбивка на предложения

Таблица 2

Качество квазиреферирования текстов, не поддающихся естественному реферированию

№ текста	Коэфф. сжатия	Полнота	Точность	Связность	Тип документа
2	20	не поддаются оценке			биржевая сводка
8	3,5	- " -	" -		сводка спортивных результатов
I4	3,5	- " -	" -		сводка курсов валют

Объем полученных квазирефератов - от I до 3 предложений; в двух случаях объем составил 4 предложения: это были документы, не поддающиеся естественному реферированию.

Итак, эксперимент дал следующие результаты.

Во-первых, было установлено, что сформулированная гипо-

теза не относится к небольшой части не поддающихся реферкрованию текстов газетных сообщений.

Во-вторых, на материале большой части текстов были проверены все 7 пунктов гипотезы о гиперсинтаксической структуре текстов газетных информационных сообщений.

I) Текст газетного информационного сообщения образует гиперсинтаксическую структуру. Это предположение подтверждается тем, что система выделила в текстах реально существующие в них межфразовые связи, удовлетворяющие все условия синтаксической сверхфразовой связи (Добрускина Э.М., Берзон В.Е., 1986); именно эти отношения "спаивают" текст газетного сообщения в единое целое.

2,3) Предложения этой гиперсинтаксической структуры объединяются в МСЕ. Данное предположение подтверждается чередованием в тексте сообщения автосемантичных и синсемантичных предложений.

4,5) Предположение о том, что автосемантичные предложения являются действительно топиковыми, оценивается следующим образом. Тот факт, что 12 из 14 квазирефератов (табл. I) имеют отличную или хорошую полноту, свидетельствует о том, что в 85% текстов предложения, которые мы предположительно назвали топиковыми, действительно содержат основную информацию, заключенную в исходных текстах. Кроме того, полученные квазирефераты содержат мало избыточной информации (ее наличие вызвано в основном ошибками, не связанными с качеством нашей модели). Таким образом, включенные в квазиреферат предложения содержат, как правило, основную информацию исходного текста, т.е. соответствуют определению топикового предложения.

6) Количество топиковых предложений, как правило, составляет не более 25% всех предложений этого текста (см. табл. I): коэффициент сжатия менее 4 получен только для очень коротких текстов.

7) Предположение о том, что из топиковых предложений может быть составлен новый текст, имеющий собственную гиперсинтаксическую структуру, частично опровергается результатами эксперимента: 5 квазирефератов из 14 (каждый третий) получили низкую оценку по параметру "связность", т.е. эти квазирефераты выглядят скорее как искусственное объединение предложений, относящихся к одной и той же теме, чем как связный текст. С другой стороны, основной причиной этого были внешние для нашей модели факторы, поэтому следует считать

полученный результат предварительным и нуждающимся в дополнительной проверке.

Уровень "понимания" и инженерно-лингвистический эксперимент

Безусловно, проблема выбора оптимального уровня "понимания", равно как и его способа, является принципиально важной при разработке ИС. В равной степени понятным представляется как стремление их разработчиков добиться максимальных результатов при использовании минимальных средств, так и желание оснастить систему максимально возможным лингвистическим обеспечением. В свете сказанного выше становится очевидным, что только лингвистический эксперимент позволяет ответить на вопрос о необходимости и достаточности того или иного уровня, того или иного способа "понимания" для достижения требуемого результата. Приведем примеры.

1) В литературе высказываются принципиально разные подходы к автоматизации реферирования: авторы предлагают строить системы с нулевым, первым и даже вторым УП (подробнее см. Добрускина В.И., Берзон В.Е., 1986). А за понятием уровня "понимания" стоят совершенно различные подходы к технологии реферирования, требующие принципиально различного лингвистического и программного обеспечения.

2) При построении информационно-поисковых систем, ориентированных на "понимание" текстов входных информационных потоков, можно ориентироваться как на нулевой, так и на первый уровень "понимания": в первом случае имеется в виду автоматическое индексирование документов путем выделения из них ключевых слов (точнее - автоматическое квазииндексирование), а во втором - усиление процедур индексирования путем использования тематического классифицирования, т.е. предшествующего индексированию определения тематики документа.

При выборе типа ИС роль инженерно-лингвистического эксперимента оказывается решающей, если мы не хотим пользоваться таким ненадежным критерием, как "общие соображения". Конечно, оценивая качество работы ИС можно в принципе ограничиться тривиальным экспериментом, не предполагающим оценки лингвистических гипотез, лежащих в основе лингвистического обеспечения данной системы. Однако при этом мы рискуем не заметить глубинную (собственно языковую) причину недостатков в работе системы и повторить те же ошибки в других системах.

В то же время, значение инженерно-лингвистического экс-

перимента этим не ограничивается. Не менее важным оно является для теоретической лингвистики. В этом случае эксперимент служит не только средством проверки уже сформулированной лингвистом гипотезы, а стимулирует выработку такой гипотезы и ее эксплицитную, непротиворечивую формулировку. Следует отметить, что проблематика лингвистического обеспечения информационных систем исключительно богата потенциальными гипотезами, и лингвисты, зачастую неосознанно, опираются на них при разработке лингвистического обеспечения. Поэтому создание ИС должно, по нашему убеждению, иметь сильную обратную связь с теорией языка. И наоборот, для каждой гипотезы, каждого нового понятия, вводимого лингвистом, необходимо искать систему, в рамках которой можно было бы проверить справедливость данной гипотезы и реальность данного понятия. Если же такую систему найти не удается, то данные гипотеза и понятие не могут быть признаны достоянием лингвистической теории.

Все сказанное позволяет нам говорить о важности дальнейшего развития теории и практики лингвистического эксперимента с обязательным учетом (а) потребностей в разработке промышленных информационных систем и (б) требований теоретической лингвистики.

Л И Т Е Р А Т У Р А

Берзон В.Е., Блехман М.С., Захаров А.А., Певзнер Б.Р. Реализация на ЭВМ системы, анализирующей синтаксические сверхфразовые связи. – НТИ, сер. 2, 1984, № 9, с. 25–31.

Блехман М.С. Эксплицитность и имплицитность межфразовых отношений в научном тексте. – НТИ, сер. 2, 1984, с. 25–31.

Блехман М.С. Инженерно-лингвистическое моделирование категории определенности при автоматической обработке связного текста (на материале английского языка). – Автореф. дисс. ... канд. филол. наук. – Л., 1985, – 21 с.

Вейценбаум Дж. Понимание связного текста вычислительной машиной. – В сб.: Распознавание образов. Исследование живых и автоматических распознающих систем. – М., 1970.

Добрускина З.М. Берзон В.Е. Синтаксические сверхфразовые связи и их инженерно-лингвистическое моделирование. – Кишинев: Штиинца, 1986.

Зарубина Н.Д. Методика обучения связной речи. – М.: Русский язык, 1977. – 48 с.

Пиотровский Р.Г. Инженерная лингвистика и теория языка. – Л.: Наука, 1979. – 112 с.

Пиотровский Р.Г. Лингвистические уроки машинного перевода. – Вопросы языкознания, 1985, № 4, с. 18–27.

Степанов Ю.С. Методы и принципы современной лингвистики. – М.: Наука, 1975. – 311 с.

Чижаковский В.А. Беляева Л.Н. Тезаурус в системах автоматической переработки текста. – Кишинев: Штиинца, 1983. – 163 с.

Шенк Р. Обработка концептуальной информации. Пер. с англ. – М.: Энергия, 1979. – 360 с.

ENGINEERING LINGUISTICS AND TEXT "UNDERSTANDING" SYSTEMS

Mikhail S. Elekhman

S u m m a r y

The tasks of engineering linguistics are analyzed in connection with the development of commercial text processing information systems, as well as with some problems of theoretical linguistics. Levels of computer's "understanding" of natural language texts are classified and compared with human understanding. Some information systems are described and their work analyzed in order to explicate some vital tasks of engineering linguistic methodology. Correlation between linguistic hypothesis and experiment is dealt with. Types of experiments are introduced and illustrated.

РИТМИКА АССОЦИАТИВНОГО ПОТОКА: К ПРОБЛЕМЕ КОЛИЧЕСТВЕННОГО АНАЛИЗА

М.Г.Борода, В.Э.Пашковский

Современный этап развития квантитативных исследований текста характеризуется тремя тенденциями, проявляющимися в работах последних лет все отчетливей и яснее и, по сути - определяющих это развитие в его главных чертах. Первая из этих тенденций связана с существенным р а с ш и р е н и е м о б ъ е к т а и с с л е д о в а н и я, включением в область интересов квантитативного анализа текста текстов на различного рода искусственных языках, текстов принципиально нелингвистической природы (например, музыкальных), и т.п. Вторая тенденция связана с попытками к о м п л е к с н о г о п о д х о д а к исследованию квантитативных характеристик текста как особого объекта сложной природы - в частности, с попытками интеграции квантитативной лингвистики текста с психологическими и психофизиологическими исследованиями процессов памяти, ассоциативного мышления, и т.д. Наконец, третья тенденция характеризуется подчеркнутым вниманием к и с х о д н ы м м е х а н и з м а м генерирования текста, механизмам, порождающим наблюдаемые в нем количественные закономерности. И хотя истоки каждой из названных тенденций могут быть, в принципе, прослежены в квантитативной лингвистике и ранее (достаточно вспомнить хотя бы работы Дж.Ципфа), все же определяющая их роль стала намечаться лишь в исследованиях недавнего - а точнее, последнего времени). Естественно, что в этих условиях приобретает особый интерес и значимость вопрос об общих закономерностях последовательного порождения текста, о характеристиках, влияющих на выбор из памяти каждого следующего его элемента. Именно поэтому так закономерен интерес, проявляемый исследователями квантитативной организации текста к результатам анализа а с с о ц и а т и в н ы х п о т о к о в, данных а с с о ц и а т и в н о г о э к с п е р и м е н т а. С одной стороны, обнаружившаяся в исследовании ассоциативных потоков их внутренняя цельность -

см., например классическую работу А.Н.Леонтьева (Леонтьев, 1983) - естественно наводит на мысль о трактовке человеком его речевой продукции как некоего т е к с т а. С другой стороны, закономерности, наблюдаемые в ассоциативном потоке (АП), когда отсутствуют многие ограничения и управляющие связи "высших уровней" организации обычного текста, отражают, повидимому, весьма общие принципы его генерирования - принципы, вне которых он вообще не может быть воспринят.

В настоящей статье кратко описаны результаты исследования некоторых квантитативных характеристик ассоциативных потоков, полученных в эксперименте по свободному ассоциированию от здоровых и душевнобольных испытуемых в возрасте от 18 до 30 лет. Исследовались, в основном, характеристики ассоциативного потока (АП), связанные с чередованием в нем слов различной длины и с различным местом ударения в слове а также - слов, принадлежащих различным семантическим группам (см. ниже). Показано, что характеристики эти - являющиеся, по существу, р и т м и ч е с к и м и характеристиками линейной организации АП - существенны для формирования АП как у здоровых испытуемых, так и у душевнобольных. Показано также, что имеется значимое, в целом, различие в "силе влияния" на АП этих характеристик у испытуемых первой и второй групп, а также - у больных разных нозологических групп. Однако основной целью статьи является, как это кажется авторам, не столько описание полученных результатов или их обсуждение, сколько демонстрация самой возможности и методики "чисто ритмического" анализа квазитекстов, подобных АП.

Методика. Для получения АП от испытуемого была использована методика континуальных свободных ассоциаций (испытуемый называет, в ответ на тестовое слово, серию любых приходящих в голову слов; как правило, длина каждой из таких "вызванных" ассоциативных цепей не превышает 50-60 слов), модифицированная одним из авторов настоящей статьи (Пашковский, 1984) следующим образом. Каждому испытуемому предлагалось называть любые приходящие в голову слова последовательно, б е з использования тестового слова - всего 500 слов в однократном эксперименте. На этот процесс называния

экспериментатором не накладывалось никаких ограничений (в частности, разрешались повторы слов), кроме ограничения на длину потока. Темп речи экспериментатором не задавался и по ходу эксперимента не корректировался, однако давалась общая инструкция называть слова в удобном для испытуемого быстром темпе. Названные слова последовательно записывались экспериментатором на предварительно пронумерованные поля тетради - это существенно упрощало как останов эксперимента при достижении объемом АП 500-словней границы, так и последующую обработку результатов.

Испытуемые. В экспериментах участвовало 33 испытуемых в возрасте от 18 до 30 лет, со средним образованием: 15 психически здоровых людей, 10 больных бредовой шизофренией и 8 - с диагнозом "олигофрения в степени легкой дебильности". Все больные исследовались вне острой психотической симптоматики и активного лечения. При отборе больных в группу соблюдался принцип однородности клинической картины внутри группы.

Результаты. Уже первые полученные в эксперименте ассоциативные потоки выявили общие принципы организации АП, характерные для испытуемых всех трех групп. Главным из них оказался принцип "мышления тематическими комплексами", когда АП распадался на большее или меньшее количество тематически однородных участков с плавными или, наоборот, резкими, переходами между ними. В эффекте этом, отмечавшемся и раньше рядом других исследователей, но выявившемся в данном случае на качественно большем по объему АП, сказывается организующее влияние на АП экстралингвистической ситуации, организующей слова в ассоциативном ряду. Здесь, на больших АП, со всей яркостью проявился принцип "текстовости высказывания", трактовки его как "потенциального текста"). С другой стороны, выявились и отчетливые различия в организации АП у испытуемых каждой из трех исследованных групп. В частности, для здоровых испытуемых было характерным политематическое строение АП, с плавным переходом между "темами", для испытуемых с диагнозом "олигофрения (дебильность)" - заметно меньшее количество тем, с преобладанием наглядно-образного типа организации АП,

для больных бредовой шизофренией - внутренняя неустойчивость АП (в частности, то плавные то резкие переходы между темами) и существенные межиндивидуальные различия. Однако, в полученных ассоциативных потоках привлекли внимание закономерности не только лексического уровня, но и другого, в каком-то смысле, более глубинного.

Именно, на целом ряде ассоциативных потоков можно было наблюдать, что нередко за длинным словом следует длинное же слово, за коротким - короткое. Более того, на отдельных участках АП за словом данной длины (в числе слогов) следовало слово такой же длины - независимо от наличия или отсутствия между словами семантической связи. Слова одинаковой длины на отдельных участках АП сцеплялись в своеобразные ц е п о ч к и. Похожие явления можно было наблюдать и в отношении м е с т а у д а р е н и я в слове: за словом с ударением, например, на I-м слоге следовало нередко слово (вовсе не обязательно равносложное первому) с ударением на I-м же слоге (напр., мать-скатерть-стол-ложка-вилка, и т.п.). Естественно, представляло интерес проверить, действительно ли такие цепочи равносложных или одинаковоударных слов существенны в формировании АП, одинаково ли они распространены в АП испытуемых трех групп или здесь существуют межгрупповые различия, проявляются ли здесь какие-либо общие эффекты, наблюдавшиеся при рассмотрении АП на лексическом (семантическом) уровне.

С этой целью исследовались р а с п р е д е л е н и я д л и н цепочек равносложных и (отдельно) одинаковоударных слов в каждом АП (как очевидно, любая последовательность из слов может быть разбита на такие цепочки, просмотром ее слева направо, без пропусков и однозначно). Анализ выявил следующее.

Во-первых, цепочки равносложных и одинаковоударных слов занимают в каждом из исследованных АП заметное место: относительная частота таких цепочек из 2-х, 3-х, и т.д. слов в АП - примерно 0.35-0.4.

Во-вторых, сравнение распределений длин цепочек одного и другого типа у здоровых испытуемых, олигофренов (дебилов) и больных бредовой шизофренией по хи-квадрат-критерию выяви-

ло статистически значимое ($P < 0.01$ и даже $P < 0.001$) различие между испытуемыми трех групп, в особенности сильное для больных шизофренией и олигофренов (дебилов)($P \ll 0.001$). Как показал анализ, "сила влияния", распространенность цепочек первого и второго типа в АП олигофренов выше, чем в АП здоровых, а для здоровых - несколько выше, чем для больных шизофренией. Сравнение распределений длин цепочек выявило и различие цепочек первого и второго типа: группировка равносложных слов с примерно одинаковой "силой" затрагивает АП у олигофренов и здоровых (значение χ^2 - 12.66 при 5 степенях свободы незначимо при $P=0.05$), в то время как "сила" влияния одинаковоударных группировок у них существенно различна - распределения длин соответствующих цепочек различаются при $P < 0.001$, и в АП олигофренов чаще, чем у здоровых, встречаются цепочки из 7,8 и более слов.

Наконец, в-третьих, анализ АП на уровне цепочек равносложных и одинаковоударных слов выявил существенно большие межиндивидуальные (внутригрупповые) различия в группе больных шизофренией и группе здоровых, чем в группе олигофренов. (В последней внутригрупповые различия распределений незначимы при $P=0.05$). Эта заметно большая внутренняя однородность АП олигофренов (дебилов) - точнее, внутригрупповая однородность -, так же как и отмеченная выше несколько более ясная, чем у испытуемых других групп, тенденция к образованию цепочек равносложных и одинаковоударных слов естественно связывается с большей инертностью испытуемых олигофренов, проявляющейся на л е к с и ч е с к о м уровне их АП. Повидимому здесь (так же, как и в случае цепочек в АП больных шизофренией) можно говорить о проявлении некоей целостной характеристики мышления, актуальной на различных этапах и уровнях текстопорождающего процесса.

Таким образом, исследование чисто ритмических характеристик ассоциативного потока - а чередование слов разной длины и с различным местом ударения несомненно задает р и т м АП - позволило выявить как некоторые, повидимому, неизвестные, индивидуальные (мехнозологические) различия в процессе порождения связного квазитекста, так и общие принципы, регулирующие этот процесс. В этих условиях естественен вопрос о

существовании "ритмической организации" на уровне лексики. И некоторые закономерности в этом плане действительно есть.

Именно, выявилось, что если разбить словарь АП на два класса: слова, относящиеся к "живому" - человек, животное, части тела, и т.п. - и "неживому" (критерий, по недостатку места, опускаем), и исследовать, насколько случайны чередования слов этих классов в конкретных АП (например, опираясь на критерий серий Вальда-Вольфовица), то окажется, что чередование это статистически значимо ($P < 0.01$) отлично от случайного - как на всем АП, так и, обычно, на его фрагментах (названное чередование исследовалось на последовательных фрагментах каждого АП: 1-100 слов, 100-200, и т.д.). При этом в АП олигофренов ясно наблюдалась не характерная для АП больных шизофренией или здоровых тенденция образовывать достаточно длинные цепочки из слов одного или другого семантического класса (типа:"поезд-машина-дорога-мост-столб...", "лошадь-корова-кабан-...-конь-овца-школа-библиотека-роща-огород-") АП здоровых характеризует в этом плане заметно более короткие цепочки и, нередко, тенденция к контрастированию соседних слов в определенном выше смысле. АП больных шизофренией отличается и по "ритму" чередования "живого-неживого" той же вариабильностью, что и в отношении цепочек равносложных и равноударных слов.

Отметим в заключение, что полученные (предварительные) результаты показывают определенные перспективы квантитативного исследования ритмических закономерностей организации лексики в тексте для изучения механизмов и принципов его генерирования.

Л И Т Е Р А Т У Р А

Леонтьев А.Н. Опыт структурного анализа цепных ассоциативных рядов. - В кн.: А.Н.Леонтьев. Избранные психологические произведения. - М.: Педагогика, 1983, с. 50-71.

RHYTHMIC OF THE ASSOCIATIVE STREAM: A QUANTITATIVE APPROACH

M.G.Boroda, V.E.Paškovskij

The paper deals with the analysis of "rhythmical" regularities of the Associative Stream on its lexical level. It is shown that the revealed regularities are connected with the basic principles of the process of text generation.

ЭТАЛОННЫЕ ТИПЫ МОРФОЛОГИЧЕСКИХ ПАРАДИГМ ДРЕВНЕСЛАВЯНСКИХ ТЕКСТОВ

А.С. Герд

В ряде предшествующих публикаций нами были затронуты некоторые вопросы общей типологии древнеславянских текстов на квантитативной основе. В частности были выявлены ареальные, стилистические и собственно жанровые группы славянских текстов XI-XIУ и ХУ-ХУI веков (Герд А.С. и др., 1974-1986).

Среди различных статистических параметров описания типологии текстов одним из наиболее ярких является средняя частота употребления флексии. Знание средних дает возможность проводить сравнение частоты флексий отдельных конкретных текстов со средними арифметическими их употребления и тем самым отнести впоследствии этот текст к той или иной совокупности текстов или типу языка. Анализ средних частот флексий позволяет по-новому поставить вопрос о выделении типовых древнеславянских текстов разных эпох.

В таком случае типовая морфологическая парадигма может строиться не только на собственно качественной основе - по принципу наличия флексии в тексте, но и на количественной основе, а именно с учетом средних частот употребительности флексии в текстах того или иного типа. Весьма существенно при этом, что сами условия выведения средней арифметической не позволяют исключать из исходных данных нулевые, отрицательные факты, то есть флексии, отсутствующие в том или ином тексте. Факт отсутствия флексии в памятнике имеет большое значение при сравнении данных на общеславянском фоне.

Таким образом, будучи построенной для той или иной совокупности текстов такая морфологическая парадигма, во-первых, материально представляет полную морфологическую систему падежей и флексий, во-вторых, показывает среднюю частоту употребления флексии, в третьих, выступает как один из возможных эталонов при сравнении и отнесении каждого отдельного славянского текста к той или иной совокупности текстов.

Ниже формирование исходных данных в таблицы проведено, исходя из следующих соображений: 1. Основным определяющим параметром характеристики древнеславянских текстов XI-ХУI веков является их деление по жанру на два общих типа: а) тип

деловых и летописно-хроникальных текстов, б) тип конфессионально-повествовательных и повествовательных текстов. 2. Для эпохи славянского средневековья конфессиональные южновосточнославянские тексты представляют один единый тип языка.

В то же время цель статьи - не типология отдельных текстов, а определение типовых парадигм, которые могли бы послужить основой для типологии древнеславянских текстов. Учтена необходимость равномерной репрезентации текстов в хронологическом и в ареальном отношениях (см. список источников).

Принципы формирования таблицы заданы в структуре списка источников, с. 67. Приводимые ниже в таблицах типовые парадигмы представляют достаточно полные системы сильных типов славянского именного склонения, характерных для тех или иных текстов в XI-XII, XIII-XIV и XV-XVI веках.

Поскольку статья построена на основе книг (Герд А.С. и др., 1974; 1977), то написание падежей, флексий и типов текстов приводится согласно (Герд А.С. и др., 1974; 1977).

Частота 1 в таблице обозначает, что средняя частота флексии не превышает единицу.

Знак "-" показывает отсутствие флексии в выборке. Естественно, что средние частоты по флексиям, приведенные в таблицах, отражают и активность типа основ, падежа, числа, взаимовлияние основ и связь с лексикой. Именно эти вопросы в лингвостатистическом аспекте достаточно подробно освещены в (Герд А.С. и др., 1974; 1977).

Приведенные данные вновь подтверждают, во-первых, общую взаимопротивопоставленность древнеславянских конфессионально-повествовательных и деловых текстов, а во-вторых, выделение и сильное обособление западнославянских текстов XV-XVI веков. При этом западнославянские тексты выделяются не только по средним частотам их флексий, но, нередко и материально, по наличию своих особых флексий.

В конечном счете, именно в рамках общего жанра деловых текстов каждый раз особо выделяются парадигмы текстов русских, западнорусских (староукраинских, старобелорусских) и текстов из Сербии, Боснии, Хорватии, Приморья, Дубровника.

Во всех падежах наблюдается резкий скачок в усилении продуктивности флексий от периода XI-XIV веков к эпохе XV-XVI веков, что свидетельствует о том, что язык XV-XVI веков

Таблица I

Тип текста	Конфессионально-повествовательные				Деловые		
Ареал	Юго-вост.-слав.			Зап.-сл.	Рус.	Зап.-рус.	Ю.
век	XI-XII	XIII-XIУ	XУ-XУI	XУ-XУI	XУ-XУI	XУ-XУI	XУ-XУI
Перечень падежей, основ и флексий							
I	2	3	4	5	6	7	8
Им.ед.м.р.^{x}o							
ъ/ъ/ø	230	265,7I	259,I4	304,57	4I8,I4	267,7I	5I3,85
o/e/Ѣ	1	1	6,85	I,I	I5,85	I7,57	I4,28
н/и	6,28	4,I4	-	-	-	-	-
Род.ед.м.р.^{x}o							
а/я/Ѧ	I08,I4	I20	I58,I5	99,I4	263,I4	292,42	399,7I
у/ю/Ѫ	I,85	3	2,85	46,57	32,I4	6I	I,28
н/и	1	1	3,I4	-	I0,57	-	-
е/Ѣ	1	2,57	4,28	26,28	7	-	6,57
Дат.ед.м.р.^{x}o							
у/ю/Ѫ	70,7I	75,28	I26	40,I4	I96,28	III,85	2I2,42
ови/еви	9,85	II,I4	I3,42	I9,57	I	4,28	-
е/Ѣ	-	-	-	-	2,42	I	I,85
ы/и	1	-	1	5,I4	-	-	2,I4
Вин.ед.м.р.^{x}o							
ъ/ъ/ø	98,85	90,28	I24,14	84,28	I75,57	I50,I4	I49,28
а/я/Ѧ	67	72,85	I03	55,I4	III,28	34,7I	83,85
о/е/Ѣ	0,42	-	I,42	I2,28	2,28	-	2,57
н/и	0,42	-	-	-	-	-	-

Таблица 2

I	2	3	4	5	6	7	8
Тв.^{x}ед.м.р.^{x}о							
ъмъ/ьмь	37,28	43,42	8I,57	57,I4	86,85	76,42	96,57
нмъ/нмь	1	-	-	-	-	-	-
Местм.ед.м.р.^{x}о							
ѣ/е	25	42,7I	52	32,I4	88	46	7,42
м/м	5	6,I4	II	3	3,7I	5,I4	I3,7I
Зв.ед.м.р.^{x}о							
о/е/ѣ	39,28	I7,57	2,7I	I2,7I	23,7I	-	9,28
у/ю	2,57	I,I4	-	-	-	-	-
Им.ед.ср.р.^{x}о							
о/е/ѣ	49,57	6I,28	5I	32,42	25,42	20,57	54,85
ъ/ь/ѳ	-	-	-	-	-	-	-
Род.ед.ср.р.^{x}о							
а/я/Ѧ	48	67,42	I34,28	50,85	79,85	I06,I4	I08,42
у/ю/Ѫ	4,85	-	-	-	-	I,7I	-
е/ѣ	I3,85	-	-	20,I4	-	-	3,42
м/м	-	-	4,7I	-	-	-	
Дат.ед.ср.р.^{x}о							
у/ю/Ѫ	I3,85	I9,85	48,7I	II,I4	20	I5	30,7I
м/м	1	-	4,42	I4,I4	-	-	I,57

Таблица 3

I	2	3	4	5	6	7	8
Вин.ед.ср.р. ^{x}о							
о/е/ѣ	88,42	86,28	I62,57	74,I7	78,28	55,85	I30,7I
а/я/ѧ	1	-	-	-	3,I4	-	-
и	-	-	-	I5,I4	-	-	-
Тв.ед.ср.р. ^{x}о							
ъмь/ьмь	29,85	25,I4	72	I7,85	I7	32,42	23,28
имъ/имъ	1	-	-	I3,85	-	-	1
ымъ	1	-	-	-	-	-	-
Мести.ед.ср.р. ^{x}о							
ѣ/е	8,85	8,7I	I2	I5,I4	30,I4	I3,42	3,85
и/и	25,42	27,7I	27,85	20,42	3,85	9	5,85
у/ю	-	1	1	I2,28	-	8	I7
Им.ед.ж.р. ^{x}а							
а/я/ѧ	46,7I	53,I4	63,7I	67,7I	I39,85	46	68,7I
и/и	1	-	-	-	-	-	
ѣ/е/о	4,2	1	-	5,42	-	-	1

Таблица 4

1	2	3	4	5	6	7	8
Род.ед.ж.р.^{x}a							
м/и	37,57	45,28	59,7I	63,I4	I73,I4	I4I,57	I7,I4
ѣ/е/Ѧ	I2	I2,7I	II,7I	I8,7I	8,28	I	I03,I4
а/я/Ѧ	5,42	2,7I	6,85	-	-	-	-
Дат.ед.ж.р.^{x}a							
ѣ/е	I0,28	I5,85	20,85	I4,5I	76,28	20,7I	I0,85
м/и	8,57	7,85	II,28	4,I4	3,42	6	29,28
Вин.ед.ж.р.^{x}a							
у/ю/Ѫ	II8,7I	I05,7I	I4I	65,42	I96,7I	70	II5,57
ѣ/е	-	-	-	-	-	-	3,42
а/я/Ѧ	3,28	I,42	-	2,7I	24,7I	I,85	-
у/i	-	-	-	2	-	-	-
Тв.ед.ж.р.^{x}a							
оѪ/ою	I7,85	22,I4	36	I7	59	38,7	I
и	-	-	-	8,57	-	-	-
омь	-	-	-	-	-	-	I8,85
Местн.ед.ж.р.^{x}a							
ѣ/е	I5,7I	I5,7I	2I,28	I8,I4	64,42	30,7	II,85
и/ы	I6,57	II	20,I4	I3,85	II,I4	I0,7I	27,7I
у/ю	-	-	-	-	-	-	5,85
Зв.ед.ж.р.^{x}a							
о/е	4,57	8,I4	2,7I				
ѣ/Ѧ	6	I,I4	-	-	-	-	-
Им.ед.ж.р.^{x}i							
ъ/ь/ø	I5,7I	24,28	28,7I	39,57	22,42	28,57	II,7I
и/ы	I,I4	-	-	-	-	-	-
Род.ед.ж.р.^{x}i							
и/ы	22,85	29,42	47,42	4I,42	42,28	84,42	I4,42
ѣ/е/Ѧ	I,57	-	8	2,28	-	-	-

Таблица 5

I	2	3	4	5	6	7	8
Дат.ед.м.р.ж,у i							
и/ы	4,57	6,85	I6,42	II,28	4,7I	I7,57	I6,42
Вин.ед.ж.р.ж							
ъ/ъ/Ø	40,42	35,85	75,42	42,7I	36	28	53,I4
и/ы	I	-	-	-	-	-	-
Тв.ед.м.р.ж,у i							
ию/ью	I7,28	I4,85	35,85	29,57	9,28	28,85	2I,7I
омь	-	-	-	-	-	-	2,7I
инь	-	-	-	-	-	-	I
у/ю	-	-	-	-	-	-	I,42
Мести.ед.м.р.х i							
и/ы	9,I4	II,7I	20,I4	22,28	I4,57	7,28	6,85
ѣ/е	I	-	-	-	-	I	-
у	-	-	-	-	-	-	I
Им.мн.м.р.ж о							
и/ы	69,7I	59,I4	38,57	32	90,I4	46,57	28,7I
ѣ/е	3,28	3,42	I0,85	I4,57	9,7I	I3	24,57
а/я/ѧ	I,85	-	I	I,9I	2,28	2,28	2,7I
ове/еве	I	I	-	30,42	-	I0,I4	3,85
ие	I	2,I4	-	-	-	-	-
Род.мн.м.р.х о							
ъ/ъ/Ø	24,85	I8,85	33,I4	8,I4	I0,7I	7,42	62,I4
овъ/евъ	4,57	I,7I	7,7I	-	56,I4	60,7I	I4,7
ей/ей	I	I,57	-	5,7I	6,7I	20,57	-
ии	I	-	-	-	-	-	-

Таблица 6

I	2	3	4	5	6	7	8
Дат.мн.м.р.$^{*}_{o}$							
омъ/емъ	35,I4	30,28	33,57	27,7I	45,85	26,85	47,57
овомь	-	-	-	-	-	-	1
амъ/ямъ	-	-	-	2,42	9,I4	I,I4	3,85
Вин.мн.м.р.$^{x}_{o}$							
н/м	43,I4	42,42	60	36,I4	25,7I	38	23,7I
ъ/ь/Ø	1	-	I,7I	1	3	4,28	1
е/ѣ	I,I4	-	2,85	4,28	-	-	20,57
а/я/Ѧ	2,57	I,42	-	-	-	-	4,I4
овъ/евъ	-	1	1	26,7I	9,7I	I2,57	1
ове	-	-	-	-	-	-	3,85
еи/ей	-	-	2,42	-	2,57	I,I4	-
Тв.мн.м.р.$^{x}_{o}$							
м/н	I0,I4	I0,42	I6,7I	20,7I	29,28	36,7	2I,7I
ами/ями	-	-	-	-	-	I5,85	-
ьии/ми	1	-	1	-	-	4	3,28
Местн.мн.м.р.$^{x}_{o}$							
ѣхъ/ехъ	5,42	7,28	I3,7I	6,7I	I3,57	4	7,85
ихъ/мхъ	0,7I	-	-	3,57	-	-	6,57
ахъ/ямъ	1	-	1	-	5,28	5,I4	-
охь	-	-	-	-	-	-	2
Им.мн.ср.р.$^{x}_{o}$							
а/я/Ѧ	4,85	5,42	8	8,28	9,42	5,28	4
н/м	1	1	-	I,85	-	-	-
е/ѣ	1	-	-	3,7I	-	-	-

Таблица 7

I	2	3	4	5	6	7	8
Род.мн.ср.р. x_o							
ъ/ь/ѣ	5,85	8,85	II,85	I4,42	I5	II,85	II,7I
ии/ий	2,57	-	1	1	-	-	
е/ѣ	1	-	-	-	-	-	
Дат.мн.ср.р. x_o							
омъ/емъ	2,7I	3,42	5,57	1	-	I,42	1
амъ/емъ	-	-	-	-	-	1	1
Вин.мн.ср.р. x_o							
а/я/ѧ	I8,I4	I8,85	28,28	5,42	I5,42	I0,42	4,42
и/и	1	-	4,7I	I,42	-	-	
е/ѣ	1	-	-	-	-	-	
Тв.мн.ср.р. x_o							
и/и	2,57	3,42	8,28	I,42	II,28	I0,28	2,I4
ими	1	-	-	-	-	-	
Местн.мн.ср.р. x_o							
ѣхъ/ехъ	I,7I	3	8	4,28	I3,57	1	2,I4
ихъ	5,I4	7,I4	-	1	5,28	-	I,57
ахъ	-	-	-	-	-	1	1
Им.мн.х.р. x_a							
и/и	8	6,7I	7	I4,57	23,57	6,I4	2,7I
ѣ/е	1	1	-	2	-	-	3
а/я/ѧ	I,85	I,28	-	-	-	-	-

Таблица 8

I	2	3	4	5	6	7	8
Род.мн.ж.р.^{x}а							
ъ/ь/ø	I4	II,42	22,28	I5,7I	48,28	44,42	26,I4
ich	-	-	-	2	-	-	-
i	-	-	-	I,85	-	-	-
Дат.мн.ж.р.^{x}а							
амъ/ямъ	8	4	I2,85	2,7I	23,I4	5	2,57
Вин.мн.ж.р.^{x}а							
м/и	I7,7	I5,7I	2I,28	26,42	38,7I	27,I4	5,85
а/я/Ѧ	4,42	5,28	-	-	-	-	-
ѣ/е	I,28	I,57	5	3,57	-	-	23,57
ъ/ъ/ø	-	-	-	-	-	3,57	
Тв.мн.ж.р.^{x}а							
ами/ями	I0,85	7,42	2I,57	8,42	20,42	2I	I0,42
еми	-	-	-	-	-	-	4,I4
Местн.мн.ж.р.^{x}а							
ахъ/яхъ	7,28	6,I4	5,7I	I2,7I	I8	7,28	7,57
ѣхь	-	-	-	-	-	-	I
Им.мн.ж.о.^{x}							
и/м	3,42	2,85	I9,42	5,28	6,7I	4,I4	2,85
е	-	-	-	-	-	-	I
Род.мн.ж.р.^{xy}							
еи	I	-	5,42	-	-	6	I
ии	6,28	I,28	-	6,57	5,85	-	I

Таблица 9

1	2	3	4	5	6	7	8
Дат.ми.ж.р.^{x_1}							
ъмь/ѣ м	2,57	I,85	I	I	2,7I	I	I
амъ	-	-	3,28	-	I,28	I	I
Вин.ми.ж.р.^{x_1}							
и/ы	I0	I0,28	7,42	I2,85	5,85	4,7I	I,42
ѣ/е	I	-	-	-	-	-	-
Тв.ми.ж.р.^{x_1}							
ьми	3,85	I,85	3,57	I,42	5,I4	4,7I	I,57
Мести.ми.ж.р.^{x_1}							
ьхъ/ъхъ	I,I4	3	4,42	3,I4	3,57	4,7	-
ихь	-	-	-	-	-	-	I
яхь	-	-	-	-	-	5	-

представляет в целом качественно новый этап в жстории славянских языков.

Хотя при комплектовании исходных данных памятники XI–XII вв. и XIII–XIУ вв. и были сознательно разделены на две разные сонокупности, однако материалы типовых парадигм показывают, что в целом южновосточнославянские конфессионально-повествовательные тексты XI–XII вв. и XIII–XIУ вв. обнаруживают почти полное единство не только по наличию или отсутствию флексий, но и в средних частотах по всем флексиям.

Таким образом, эти типы текстов должны быть объединены в один тип. Приведенные статистические данные на новом материале подтверждают традиционные выводы о единстве языка древнеславянских текстов XI–XIУ веков. Парадигмы конфессионально-повествовательных южновосточнославянских текстов XУ–XУI веков близки к парадигмам текстов XI–XIУ веков, но отличаются, как правило, более высокими средними для каждой флексии. Таким образом, по данным именного склонения выделяются следующие общие типы морфологических парадигм именного склонения, характерные для древнеславянских текстов XI–XУI веков.

I. Тип парадигмы южновосточнославянских конфессионально-повествовательных текстов XI–XIУ веков.

2. Тип парадигмы южновосточнославянских конфессионально-повествовательных текстов XУ-XУI веков.
3. Тип парадигмы западнославянских конфессионально-повествовательных текстов XУ-XУI веков.
4. Тип парадигмы русских деловых текстов XУ-XУI веков.
5. Тип парадигмы западнорусских (староукраинских, старобелорусских) деловых текстов XУ-XУI веков.
6. Тип парадигмы сербо-хорвато-боснийских-приморских деловых текстов XУ-XУI веков.

Каждая из выделенных выше типовых парадигм может интерпретироваться как морфологический эталон, представляющий собой абстрактный инвариантный текст, по отношению к которому отдельные древнеславянские тексты выступают как его варианты.

Тогда и любой другой текст, который материально характеризуется наличием соответствующих флексий, такой же парадигмой и близкими средними, должен быть отнесен к одному из типов, древнеславянских текстов, выделенных выше.

Л И Т Е Р А Т У Р А

Герд А.С., Капорулина Л.В., Колесов В.В. и др. Именное склонение в славянских языках XI-XIУ вв. Л., 1974.

Герд А.С. и др. Именное склонение в славянских языках XУ-XУI вв. Л., 1977.

Герд А.С. Ареальная типология славянских текстов XУ-XУI веков. - Советское славяноведение, 1982, № 5, с. 74-82.

Герд А.С. Ответы на вопросы по славянскому языкознанию к IX международному съезду славистов (Киев, сентябрь, 1983 г.) - Вопросы языкознания, 1983, № 4, с. 41-44.

Герд А.С. Древнеславянский язык и его типы по лингвостатистическим данным. - Ученые зап. Тартуск. гос. ун-та, 1985, вып. 711, с. 9-21.

Герд А.С. О синхроническом описании языка древнеславянских центров письменности. - Вестник ЛГУ, 1986, № 1, сер. 2, с. 62-67.

Список источников

Ниже в списке источников каждая отдельная ареально-хронологическая совокупность включает выборку из 7 источников (групп источников) по I2.000 словоупотреблений в каждом. Таким образом, каждая такая совокупность представлена выборкой в 84.000 словоупотреблений.

Конфессионально-повествовательные тексты.

Юго-восточно-славянские тексты.

XI-XII вв. (колонка 2).

I. Мариинское ев. - Мариинское четвероевангелие. Спб., I883, с. 73-I3I.

2. Чудовская пс. - Чудовская псалтырь XI в. Спб., I9I0, с. 42-85.

3. Новг. минеи - Служебные минеи в церковнославянском переводе по русским рукописям I095-97 гг. Спб., I886, с. 03-023 (сентябрь), I-I9 (октябрь), 267-286 (ноябрь).

4. Погодинская пс. - Погодинская псалтырь. - В кн.: Словѣньска псалтырь. Psalterium bononiense. Vindobone, I907, с. I-200.

5. Мирославово ев. - Мирославово евангелие. Београд, I897, с. I-I8, 9I-I08, I80-I97, 280-297.

6. Галицкое ев. - Четвероевангелие Галицкое II44 г., сличенное с древнеславянскими рукописями евангелия XI-XVII вв. М., I882, с. 3I-329.

7. Житие Ф. Печерского - Житие Феодосия Печерского. - В кн.: Сб. XII в. Московского Успенского собора. Вып. I. М., I899, с. 40-72.

XIII-XIУ вв. (колонка 3)

I. Добрейшево ев. - Добрейшево четвероевангелие. - В кн.: Български старини. Кн. I. София, I906, с. 95-I03, I33-I40, I55-I62, 2I6-226.

2. Вуканово ев. - Вуканово еванѣље. Београд, I967, п. 65-95.

3. Житие Нифонта - Житие Нифонта. - В кн.: Матеріяли з історії византійсько-слов'яньскої літератури та мови. Одесса, I928, с. 252-293.

4. Софийская пс. - Софийская псалтырь. - В кн.: Словѣньска псалтырь. Psalterium bononiense. Vindobone, I907, с. I-200.

5. Шишатовацкий ап. - Шишатовацкий апостол. - В кн.: Apostolus e codice Monasterii Šišatovac (paleoslovenice), edidit F. Miklosioh. Vindobone, I853, с. I-3I, 77-88, I49-I58.

6. Чудов. новый завет - Евангелие от Марка ... Сергиев Посад, I894, с. 89-404; Послания св. апостола Павла... Вып. I. Сергиев Посад, I892, с. 57-I46.

7. Евф. Тырновский - Сочинения Евфимия Тырновского. Leben des Johannes Rylski. - In: Werke des patriarchen von bulgarien Euthymius (1375-1393). - Wien, I90I, с. 5-26; Leben Hilarions, bishofs von Moglen. - Ibid., с. 27-46.

XУ-XУI вв. (колонка 4)

I. Сербия - Повѣсть Иакова архiепископа Іерусалимскаго на рождьство прѣславнне владичице нане Богородице и приснодѣвы Маріе. - In: Novaković S. Apokrifno protojevanđelje Jakovljevo. - Sterine Kn.X.Zagreb, 1878, s. 62-71; Afroditijana Persijanea priča o rodjeniju Hristovu. - Ibid., s. 74-80; Teodosija mniha Hilandarca djelo o Petru Koriškom. Priobčio dopisni član St. Novaković. - In: Starine. Kn. XVI. Zagreb, 1884, s. 12-27; Život sprskoga patriarha Jefrema. Priobćio dopisni član St. Novaković. - Ibid., s. 36-40 (всего II 890 словоупотреблений).

2. Москва (Епифаний Премудрый) - Житие святаго Стефана епископа Пермскаго. Спб., I897, с. I-39 (I2 000 словоупотреблений).

3. Москва (Пахомий Логофет) - Пахомий Серб. Житие пр. Кирилла Бѣлозерскаго. - В кн.: Яблонский В. Пахомий Серб и его агиографические писания. Спб., I908, Приложение, с. I - XLIII (I2 000 словоупотреблений).

4. Болгврия (Владислав Граматик) - Rukopis Vladislava Gramatika piaan godine 1469. - In: Starine. Kn. I. Zagreb, 1869, s. 60-85; Život Sv. Konstantine řečeneho Kyrilla. - Praha, 1868, s. 1-25; Památky dřevního písemnictví juhoslovanův. Praha, 1873, s. 1-25 (всего I2 000 словоупотреблений).

5. Болгария (Константин Костенечский) - Константин философ и ѣ егов Живот Стефана Лазаревића деспота српскога. - В кн.: Гласник српског ученог друштва. Кн. XXII. Београд, I875, с. 244-328 (I2 000 словоупотреблений).

6. Москва (Четьи-Минеи Макария) - Великия Минеи четіи.Ноябрь, - Тетрадь III. Изд. археографической комиссии. М., I9I4, с. 270I-272I, 3092-3I04, 289I-29I2 (I2 000 словоупотреблений).

7. Болгария (Матей Граматик) - Граматик Матей. Житие св. Николая новаго Софийскаго. - В кн.: Сирку П. Очеркк из исторки литературных сношений болгар и сербов в XIУ-XУII веках. Спб., I90I, с. 27-78 (I2 000 словоупотреблений).

Западно-славянские тексты XУ-XУI вв. (колонка 5).

I. Чехия (Троянская история) - Historie Trojanská. - In: Výbor z literatury české, díl druhý, od počátku XV až do konce XVI století. Praha, I868, s. 75-I23 (I2 000 словоупотреблений).

2. Чехия (Ян Гус) - Hus J. Knížky o svatokupectví. Praha, I954 (I2 000 словоупотреблений).

3. Польша - Biblja szaroszpataoka. - In: Vrtel-Wierczyński S. Wybor teketów staropolskih. Warszawa, 1963, s. 70-84; Kazania gnieźnieńskie. - Ibid., s. 33-43; Modlitwy godzienne. - Ibid., s. 49-51; Modlitwy, czyli godzinki Wacława. - Ibid., s. 373-376; Psalter z puławski. - Ibid., s. 52-59 (всего I2 000 словоупотреблений).

4. Польша - Fortuny i cnoty Hosność, 1524. - In: Vrtel-Wierczyński S. Wybor tekstów staropolskich. Warszawe, 1963, s. 270-273; Hietorja o poncianie, 1540. - Ibid., s. 288-292; Rozmowy Salomona z Marcholtem, 1521. - Ibid., s. 264-269; Powieść o papieżu Urbanie, 1514. - Ibid., s. 219-222; Hiatorja traech króli. - Ibid., s. 111 - 114;

M. Bielski. Zywoty filozofów, 1535. - Ibid., s. 276-279; Rozmyślsnie przemyskie o żywocie pana Jezusa. - Ibid., s. 87-99; Sprawa Chędoga o Męce pana chrystusowej, 1544. - Ibid., s. 99-108; Historje rzymskie, 1543. - Ibid., s. 298-300 (I2 000 словоупотреблений).

5. Польша (Петр Скарга) - Skarga P. O jednósci kościola Bożego pod jednym pasterzem, 1577. - РИБ. т. УП. Спб., I882, с. 223-266 (I2 000 словоупотреблений).

6. Чехия - Ian Táborský z Klokotské hory. Z vypsání orleje pražského, 1570. - In: Výbor z literatury české, díl druhý od počátku XV sž do konce XVI století. Praha, 1868, s. 1523-1528; Viterin Kornelius ze Všehrad: 1) Z knihy o napravení padlého; 2) Z knih o právích země české. - Ibid., s. 1039-1058; Mikuláš Konač z Hodíšťkova. Z Kroniky české Aenea Sylvia, 1510. - Ibid., s. 1189-1205; Martin Kuthen ze Šprínsberka. Kronika velmi pěkná o urozeném statečném rytíři Isnovi Žižkovi, 1564. - Ibid., s. 1511-1522 (всего I2 000 словоупотреблений).

7. Польша - Zołtarz vróbla, 1539. - In: Vrtel-Wierzyński S. Wybór tekstów staropolskich. Warszawa, 1963, s.252-259; Psalterz krakowski, 1532. - Ibid., s. 245 - 251; Ecclesiastes, 1522. - Ibid., s. 225-229; Baj dusemny, 1514. - Ibid., s. 223-225; Modlátewsnik siostry Kouztancji, 1527. - Ibid., s. 145-148; Kasania Jana z Szamotuł Paterka. - Ibid., s. 141-145 (I2 000 словоупотреблений).

Деловые тексты

Русские тексты XУ - XVI вв. (колонка 6).

I. Северная Двина - Двинские грамоты. - В кн.: Грамоты великого Новгорода и Пскова. М.-Л., I949 (II 800 словоупотреблений).

2. Москва - Московские грамоты. - В кн.: Акты социально-экономической истории Северо-восточной Руси. Т. I. М., I952 (II 800 словоупотреблений).

3. Псков - Псковские грамоты. - В кн.: Грамоты великого Новгорода и Пскова. М.-Л., I949, № 335-340, 346, 348; Псковская судная грамота. Спб., I9I4 (всего 8 I20 словоупотреблений).

4. Москва - Московские грамоты. - В кн.: Акты социально-экономической истории Северо-восточной Руси. Т.П. М., I950, № 422, 423, 423а, 426-433; т. III. М., I964, № 9-I2, 2I, 22, 24, 25, 27, 45-48, 6I, 65, 66, 67 (I2 000 словоупотреблений).

5. Москва (Судебник) - Судебник I589 года. - В кн.: Судебники XУ-XVI веков. М.-Л., I952, с. 366-4I4 (I2 000 словоупотреблений).

6. Тверь - Тверские акты. Вып. I. Изд. Тверской ученой архивной комиссии. Тверь, I896, с. 5I-II6, I53-I55, I59, I60, I67-I79 (I2 000 словоупотреблений).

7. Рязань - Рязанские грамоты. - В кн.: Акты социально-экономической истории Северо-восточной Руси. Т. II. М., I964, с. 36I-363, 370-372, 383-385, 388-409; Древние грамоты и акты рязанского края, собранные А.Н. Пискаревым. Спб., I854, с. 23-25, 28-30, 32-44 (всего I0 527 словоупотреблений).

Западно-русские тексты (староукраинские, старобелорусские)
XУ-XУI вв. (колонка 7)

I. Юго-западная Русь - Юго-западные русские грамоты. - В кн.: Акты, относящиеся к истории Южной и Западной России, собранные и изданные археографическою комиссиею. Т. I. Спб., I863 (I2 000 словоупотреблений).

2. Западная Русь - Западнорусские грамоты. - В кн.: Акты западной и юго-западной России. Т.I. Спб., I856, № I3-I9, 22-24, 38, 58, 60, 6I, 64, 67, 70, 90, 9I, I0I, I27, I28, I44, I45 (II 800 словоупотреблений).

3. Брест - Брестские акты. - В кн.: Акты, издаваемые комиссией для разбора древних актов в Вильне. Т. УI. Вильна, I872, № I, 5, 6, I0, I3, I59 (I2 000 словоупотреблений).

4. Вильно - Виленские акты. - В кн.: Акты, издаваемые комиссией для разбора древних актов в Вильне. Т. УIII. Вильна, I875, с. I8-I9, 405-4II, 424-426, 460, 46I; Акты, относящиеся к истории Западной России, изданные археографической комиссией. Т. III (I544-I587 гг.). Спб., I848, с. I5-I8, 2I-22, 95-97, I07-LI5 (I2 000 словоупотреблений).

5. Юго-западная Русь - Акты Правобережной Украины. - В кн.: Акты, относящиеся к истории Южной и Западной России, собранные и изданные археографическою комиссиею. Т. I. Спб., I863, № 45, 48, 50, 57, 59, 68, 98, I03, II0,II7, I26, I33, I35, I39, I43, I54, I63, 222, 238 (I2 000 словоупотреблений).

6. Вильно - Акты, издаваемые Виленскою археографическою комиссией. Вильна, I875, т. УIII, с. 9- 4; Там же, т. II, с. 5-43.

Тексты с территории Югославии (сербия, Хорватия, Босния, Приморье, Дубровник) XУ-XУI вв. (колонка 8).

I. Сербия - Monumenta Serbica spectantia historiam Serbiae Bosnae, Ragusii, edidit Fr. Miklosich. Viennae, I858; Старе српске повеље и писма. Кн. I, први део. Београд, Ср. Карловци, I929; кн. I. други део. Београд, Ср. Карловци, I934 (всего I2 000 словоупотреблений).

2. Хорватия - Hrvatski Spomenici (Acta Croatica). Sv. I. - In: Monumenta historico-juridica slavorum meridionalium V.VI. Zagreb, 1898; Red i Zakon od primljen'ja na dil dobroga činjen'ja sestar naših reda svetoga otca našega Dominika. - In: Starine. Kn. I. Zagreb, 1869, s. 225 - 226; Rukoviet jugoslavenskih listina. - In: Starine. Kn. X. Zagreb, 1878, s. 4-5; Statut Kastavski. - In: Monumenta historico-juridica slavorum meridionalium. Pars 1, v. IV. Statuta lingua Croatica conscripta. Zagreb, 1890, s. 181-193; Statut otoka Krka. - In: Arkiv za pověstnicu jugoslavensku. Kn. II. Uredio Ivan Kukuljević Sakcinski. Zagreb, 1852, s. 293-296, 311-312 (List od konfini med Bakrani, Grobnicani i Tersacani, god 1455) (всего I0 409 словоупотреблений).

3. Босния - Monumenta Serbica spectantia historiam Serbiae Bosnae Ragusii edidit Fr. Miklosich. Viennae, 1858 (I2 000 словоупотреблений).

4. Дубровник - Monumenta Serbica spectantia historiam Serbiae, Bosnae, Ragusii edidit Fr. Miklosich. Viennae, 1858; Старе српске повеље и писма. Кн. I, први део. Београд, Ср. Карловци, 1929; кн. I, други део. Београд, Ср. Карловци, 1934 (всего 12 000 словоупотреблений).

5. Хорватия - Urbar grižanski od g. 1544. - In: Hrvatski Urbari. Sv I. Monumenta historico-juridica slavorum meridionalium. V.V. Zagreb, 1894, s. 85-93; Urbari grada Dubovca od g. 1579 i 1581. - Ibid., s. 97-116; Rukoviet listina sa urbarskimi ustanovami. - Ibid., s. 384- 386; Statut Vrbanski a donekle i svega krčkoga otoka. - In: Hrvatski pisani Zakoni. Monumenta historico-juridica slavorum meridionalium. Pars I, V, IV. Zagreb, 1890, s. 171-172; Statut Kastavski. - Ibid., s. 194-198; Statut Veprinački. - Ibid., s. 211-216; Naredbe biskupie Modruskeod god 1589. - In: Arkiv za pověstnicu jugoslavensku. Kn. II. Uredio IV an Kukuljević Sakcinski. Zagreb, 1852, s. 86-89; Rukoviet jugoslavènskih listina.Tursko-Mletačke listine. - In: Starine. Kn. X. Zagreb, 1878, s. 7-16; Mletačke listine, 1574. - Ibid., s. 37;Hrvatski spomenici (Acta Croatica). Sv. I. Monumenta historico-juridica slavorum meridionalium. V. VI. Zagreb, 1898 (всего 12 000 словоупотреблений).

6. Хорватия (Полический статут) - Статут Полички. - In: Hrvatski pisani Zaconi. Monumenta historico-juridica slavorum meridionalium. Pars I, V, IV. Zagreb, 1890, s. 27-180 (12 000 словоупотреблений).

7. Monumenta serbica spectantia historium Serbiae, Bosnae, Ragusii, Vienne, 1858, c. 469-513.

STANDARD TYPES MORPHOLOGICAL PARADIGMS IN OLD SLAVONIC TEXTS

Alexandr S. Heard

S u m m a r y

In the article an attempt is made to determine the standard morphological paradigms for types of Old Slavonic texts of different time periods, areas and purposes (functions). The texts belong to the 11th - 16th centuries.

The arithmetical mean of inflexion frequencies is used as a statistical parameter.

СТИЛЕДИФФЕРЕНЦИРУЮЩИЕ ПОТЕНЦИИ ОТГЛАГОЛЬНЫХ СУЩЕСТВИТЕЛЬНЫХ БЕЗАФФИКСНОГО ТИПА В НЕМЕЦКОМ ЯЗЫКЕ

Т.С. Глушак, И.И. Большаков

В трактовке современных лингвистов средства всех языковых уровней, прежде всего лексические, не остаются "безучастными" к функционально-стилистической дифференциации языка, будучи по-разному используемы и, значит, специализируемы в текстах, представляющих функциональные стили. Суть их специализации следует понимать не как формирование для каждого стиля наборов лишь ему присущих единиц и категорий, а как установление различных вероятностей использования одних и тех же элементов языка, поскольку именно вероятность "меняется от стиля к стилю, а при условии существенных расхождений - дифференцирует стили" (Головин Б.П., 1968, с 15). Тем самым правильно признается в лингвистике, что "стили различаются между собой не столько наличием специфичных элементов, сколько специфичным их распределением" (Арнольд И.В., 1981, с. 76), являющимся следствием неодинакового характера развертывания стилеобразующей потенции у различных языковых единиц, форм, структур (типов, категорий и т.д.).

Множество выполненных до сих пор работ, подтвердив неодинаковую активность реализации названной потенции различными единицами языка, доказало одновременно и бесспорность тезиса, согласно которому отражением качественных расхождений в системах стилей являются количественные различия в текстах. Изучение качественной специфики каждой функционально-стилевой подсистемы может и должно поэтому базироваться на показателях количественного распределения тех или иных явлений в соответствующих текстах, т.е. исходить из принципа, что качественный и количественный аспекты функционирования языка безусловно взаимосвязаны.

Этот принцип кладется здесь в основу проведения анализа, который должен раскрыть стиледифференцирующие потенции одного из типов отглагольных имен - безаффиксных девербативов немецкого языка. Получение и сопоставление в ходе анализа статистических данных призвано показать, в каком (или каких) стиле анализируемые единицы лексики составляют достаточно часто употребляющийся тип слов, а в каком (или ка-

ких) стиле их представленность не относится к признакам, характерным или значимым. Фиксация же относительно устойчивых расхождений между стилями по признаку использования безаффиксных девербативов позволит установить новый (на фоне многих других, установленных в прежних исследованиях) параметр функционально-стилистической дифференциации современного немецкого языка.

Для исследования привлекаются тексты четырех функциональных стилей – официально-делового, научно-технического, газетно-публицистического и литературно-художественного. Но поскольку стили сами по себе не являются гомогенными образованиями и в них продолжается действие процесса дифференциации, то логично учитывать при изучении функциональных стилей и их жанрово-субстилевое расслоение. Анализ избранного типа имен будет опираться на тексты четырех жанровых разновидностей в каждом из четырех функциональных стилей. Учет жанрово-субстилевого расслоения априорно создает то преимущество, что к исследованию привлекается неоднородный текстовый материал в пределах одного и того же функционального стиля, а это, естественно, повышает достоверность результатов.

Для получения цифровых данных о функционально-реализационных особенностях имен в разных жанрах и функциональных стилях, а также для выявления специфики их количественного участия в формировании соответствующих текстов, используется одна из разновидностей вероятностно-статистического метода. Это не противоречит стремлению осветить главным образом качественную сторону роли отглагольных дериватов, поскольку с самого начала признается взаимозависимость качественной и количественной сторон в функционировании языка, а также тот факт, что через цифровые данные и их лингвотеоретическую интерпретацию возможно проникнуть внутрь этой взаимозависимости.

Посредством расчетной процедуры предполагается получить следующую статистическую информацию о функциональных свойствах безаффиксных девербативов:

а) средний показатель реализации имен в каждом стиле;

б) отклонения от среднестилевого показателя в текстах отдельных жанровых субстилей;

в) "интервал независимого распределения" имен в общей текстовой выборке из каждого функционального стиля;

г) стиледифференцирующие потенции анализируемого типа именных слов.

Как многократно описано и доказано, опорную роль в статистическом изучении языковых явлений играют средние частоты, поскольку именно в них находят свое отражение вероятностные характеристики. Получив средние частоты употребления безаффиксных девербативов в жанровых субстилях и функциональных стилях, можно будет со значительной степенью достоверности судить об интервалах, содержащих вероятности функционирования изучаемого языкового явления внутри каждой функционально-стилевой системы. Будет продемонстрирована и разница между числовой представленностью имен в текстах отдельных жанровых субстилей и их среднеарифметическим числом в функциональном стиле. Данные позволят затем определить границы распределения имен в рамках каждой функционально-стилевой системы, уже независимо от жанровой дифференциации текстового материала.

Статистическая обработка общей выборки из каждого функционального стиля для определения независимого распределения безаффиксных девербативов проводится с оценкой квадратического отклонения по формуле

$$\sigma = 2 \cdot \sqrt{\frac{\Sigma(\bar{x}i - xi)}{n \cdot (n-1)}},$$

где xi – квадратическое отклонение результата всего ряда измерений,

$\bar{x}i$ – арифметическая средняя, x – отдельное измерение,

n – число выборок,

Σ – знак суммы,

2 – коэффициент надежности[+].

Получаемые по данной формуле численные значения интервала вариации измеряемой величины соответствуют коэффициенту доверительной вероятности, равному 0,95, и представляют собой существенный количественный показатель функциональной реализации наблюдаемых единиц. Статистические данные в работе извлекаются из текстовых выборок по 25 000 словоупотреблений на каждый жанровый субстиль, в жанре романа общий объем выборки составит 100 000 словоупотреблений. Таким образом, официально-деловой, научно-технический и газетно-публицистический стили предстанут в выборках по 100 000 словоупотреблений, а литературно-художественный стиль – в выборке

[+] О методике вычисления "интервала независимой вариации" см. Дмитриева Л.Ф., 1977.

из 175 000 словоупотреблений.

Результатом статистической обработки материала являются частоты распределения имен в текстах обследуемых жанровых разновидностей четырех стилей, приводимые в таблице:

Жанрово-субстилевые разновидности	Абсолютная частота БД	Среднеарифметическая стилевая частота БД
Военное законодательство	871	
Конституционное законодательство	858	
Уголовное законодательство	903	878
Пенсионное законодательство	880	
Всего по официально-деловому стилю	3512	
Точное приборостроение	740	
Обработка металла	728	
Машиностроение	762	736
Энергетика	713	
Всего по научно-техническому стилю	2943	
Передовые статьи	604	
Политические комментарии	598	
Выступления государственных деятелей	566	596
Информационные сообщения	617	
Всего по газетно-публицистическому стилю	2385	
Роман	266	
в т.ч. у Йобста	257	
Ремарка	263	
Штриттматтера	266	
Гайдучека	277	294
Новелла	303	
Рассказ	291	
Художественный очерк	317	
Всего по литературно-художественному стилю	1177	

Данные таблицы позволяют произвести еще некоторые расчеты, существенные с точки зрения лингвистической стилистики, - вывести границы независимого распределения имен в общей выборке из текстов каждого стиля. Вначале устанавливает-

ся разница между числовой представленностью анализируемых имен в каждой жанровой выборке и их среднестилевым показателем, что находит свое выражение в следующих цифрах (на примере жанрово-субстилевых разновидностей официально-делового стиля): первая выборка -20, вторая -7, третья +25, четвертая +2. Полученные данные возводятся в квадрат, квадраты суммируются и получается число 1078, которое делится на 12 - число степеней свободы ($I = n \cdot (n - 1)$, число выборок 4, $n - 1 = 3$, $4 \times 3 = 12$), и в результате получается 89,8. Извлекается корень из $\sqrt{89,8} = 9,5$, что является средним квадратическим отклонением от средней арифметической частоты реализации безаффиксных девербативов в стиле. Так как для определения границ колебания средних величин принято использовать удвоенное среднее квадратическое отклонение, то умножается $9,5 \times 2 = 19$ и получается отклонение от среднего числа 878 случаев реализации безаффиксных девербативов в официально-деловом стиле, в обе стороны равное 19, т.е. границы интервала независимой вариации обследуемых имен в текстах стиля находятся в пределах от 859 до 897. Дальнейший расчет показывает, что отклонение от среднестилевого показателя использования анализируемых имен в текстах научно-технического стиля равняется в обе стороны 21, т.е. границы интервала их независимой реализации в стиле находятся в пределах от 715 до 757. Общее отклонение от среднестилевого показателя употребительности имен в газетно-публицистическом стиле равняется в обе стороны числу 22, а это значит, что границы интервала их независимой вариации в стиле находятся в пределах от 574 до 618. В художественных текстах безаффиксные отглагольные производные не отличаются особой активностью. По сравниваемым жанрам разброс абсолютных частот свидетельствует о существовании некой средней величины, или усредненной вероятностной закономерности количественного использования данных имен в стиле, равной 270-318 единицам на 25 000 словоупотреблений. Положение в таблице цифровых показателей по индивидуально-авторским стилям позволяет считать, что именно жанровая разновидность романа характеризуется весьма равномерным распределением имен. Хотя выборка здесь была увеличенной (до 100 000 словоупотреблений), это не меняет сколь-нибудь существенно среднежанровой величины.

Уже отмечалось, что общий объем материала, который был включен в количественное обследование, исчисляется в 475 000

словоупотреблений, а доля безаффиксных девербативов в этом материале составляет 10 814 единиц. Трудно ответить на вопрос, много это или мало, такой вопрос и не ставится. Важнее сопоставить количественные характеристики для выявления своеобразия функциональных стилей. В практике конкретных исследований уже давно принято различать два типа вероятностей в реализации языковых средств. В том случае, если количественные показатели употребления некоего средства или некоторых языковых единиц существенно расходятся (в двух или нескольких функциональных стилях), они объявляются дифференциальными стилевыми вероятностями (частотами). При несущественном расхождении цифровых показателей их считают нейтральными стилевыми вероятностями (частотами) (Головин Б.Н., 1968, с. 11-19). С точки зрения такого различения интересно проследить, воплощают ли безаффиксные девербативы в особенностях своего употребления дифференциальный, стилеразличительный признак, или же их употребление не содержит в себе стиледифференцирующих возможностей. Для выполнения расчетов в исследование вводится понятие Z-критерия проверки статистических данных, который и должен послужить выяснению того, могут ли безаффиксные девербативы с показателями их употребительности быть дифференциальным признаком функционально-стилистического членения современного немецкого языка. Указанный критерий выражается следующей формулой:

$$Z = \frac{|f_1 - f_2|}{\sqrt{\frac{f_1(1-f_1)}{N_1} + \frac{f_2(1-f_2)}{N_2}}}$$

(Пиотровский Р.Г., 1977, с. 319)

где f_1 и f_2 – средние частоты реализации девербативов в сопоставляемых стилях;

N_1 и N_2 – объемы сравниваемых выборок стилей.

При условии, если Z окажется больше 1,96, разница между частотами безаффиксных девербативов должна считаться существенной, т.е. их употребительность может служить критерием разграничения стилей; наоборот же, при Z, меньшем 1,96, признается несущественным различие между функциональными стилями по употребительности в них обследуемых отглагольных имен.

Расчетная операция по приведенной выше формуле обеспечивает получение данных, главный вывод из которых сводится к следующему: Z-критерий во всех случаях оказывается больше заданного числа. Показатели Z-критерия при сопоставлении

соседствующих (по общей количественной реализации имен) функциональных стилей соответственно равны: между литературно-художественным и газетно-публицистическим стилями – 13,6; газетно-публицистическим и научно-техническим – 4,3; научно-техническим и официально-деловым – 2,8. Полученные статистические данные позволяют оценивать наличие анализируемых имен в текстах как признак (лексический) спецификации функциональных стилей.

Результаты статистического анализа придают бесспорную актуальность известному в функциональной стилистике положению, которое гласит: "структура речи всегда своя у каждого функционального стиля" (Кожина М.Н., 1968, с. 116). "Своя структура" означает в данном случае, во-первых, существование значительных расхождений между частотами актуализации языковых явлений (здесь – безаффиксных девербативов) в текстах, соотносимых с разными стилями; во-вторых, более приближенное или более отдаленное положение по отношению друг к другу тех или иных функциональных стилей. Согласно показателям таблицы, однозначно оценить как взаимно сближенные можно официально-деловой и научно-технический стили. Высокая насыщенность (значительно выше, чем в других стилях) безаффиксными девербативами выпукло оттеняет характерную черту лексического состава этих стилей, отвечающую и их экстралингвистической специфике. Предопределенная извне тенденция к конденсированному воплощению информации, а тем самым к экономии языкового выражения, неизбежно влечет за собой наполнение текстов существительными отглагольного происхождения, поскольку именно они способны, в силу имплицированной в них валентности, к синтаксически расширенному группообразованию, то есть как раз к созданию компактных (экономных) структур выражения. Стиль художественной литературы помещается по отношению к названным стилям на противоположном полюсе, обладая наименьшей частотностью рассматриваемых имен, что также вытекает из экстралингвистически заданной специфики этого стиля: в нем не действует обязательное требование компрессии, предпочтительности компактных языковых форм выражения, прежде всего структур номинализации, и т.д. Выводом из проведенного исследования может служить следующее: установленная вариативность реализации имен по стилям свидетельствует о четкой функциональной "реакции" данного типа отглагольных дериватов на качественное изменение экстралингвистической основы, а значит может служить существенным дополнением к тому

набору стиледифференцирующих признаков, которым уже располагает функциональная стилистика современного немецкого языка.

ЛИТЕРАТУРА

Арнольд И.В. Стилистика современного английского языка. – Л.: Просвещение, 1981.

Головин Б.Н. О стилях языка и их изучении. – Русский язык в школе, 1968, № 4.

Дмитриева Л.Ф. Функциональное проявление черт временного и видового планов английского глагола в различных подъязыках. КД. Минск, 1977.

Кожина М.Н. К основаниям функциональной стилистики. – Пермь: Изд. Пермского ун-та, 1968.

Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высш. школа, 1977.

ZUR FRAGE ÜBER DIE STILDIFFERENZIERENDEN POTENZEN DER SUFFIXLOSEN DEVERBATIVA IM DEUTSCHEN

T.S. Glußak, I.I. Bolschakow

R e s ü m e e

In der modernen Funktionalstilistik behauptete sich die These, daß die wesentlichen (qualitativen) Unterschiede der Funktionalstile in ihren quantitativen Unterschieden sichtbar werden. Aus diesem Grund kann sich die Erforschung der qualitativen Spezifik jedes Funktionalstils auf die Angaben über die quantitative Verteilung der oder jener Spracherscheinungen in entsprechenden Texten stützen. Die suffixlosen Deverbativa, die durch ihre Fähigkeit zur Gruppensetzung und damit zur Bildung von kompakten Einheiten der Ausdrucksebene gekennzeichnet sind, entsprechen dem extralinguistischen Wesen der Funktionalstile der Sachprosa und sind gerade in solchen Texten besonders häufig anzutreffen. Ihre vergleichende Betrachtung nach vier Funktionalstilen ermöglicht die Schlußfolgerung, daß dieser Typ von Substantiven als ein charakteristisches Merkmal der funktionalstilistischen Differenzierung der deutschen Sprache angesehen werden kann.

СТАЦИОНАРНАЯ МОДЕЛЬ ПОРОЖДЕНИЯ СВЯЗНОГО ТЕКСТА

Ю.К. Крылов

Одной из наиболее актуальных задач квантитативного языкознания является необходимость создания теории, которая в той или иной степени позволила бы понять и объяснить закономерности статистической организации лексики в связном тексте естественного языка.

В своем обзоре Ю.А. Тулдава (1985) приводит множество математических формул, которые к настоящему времени предложены для аналитического выражения зависимости между частотой и рангом слова. Большинство из этих формул, представляющих разновидности формулировок закона Ципфа, получено на чисто эмпирической основе и направлено на уточнение описания экспериментально наблюдаемых зависимостей. Однако вне системного рассмотрения механизмов порождения текста как целостного объекта предлагаемые зависимости способны ответить лишь на вопрос о том, каковы статистические закономерности организации лексики, но оставляют в стороне наиболее существенный вопрос о причинах, ответственных за структуру этой организации. В частности, даже в работах Ю.К. Орлова (1976; 1978), в которых проведен детальный анализ следствий, вытекающих из выполнимости закона Ципфа-Мандельброта на определенной длине текста (равной "объему Ципфа"), сам факт существования "объема Ципфа" у любого связного текста рассматривается как некоторый постулат - эмпирическая данность, не получающая теоретического обоснования.

В предлагаемой работе делается попытка расширить понимание статистической организации лексики в связном тексте, исходя из вариационных принципов. Возможность такого подхода к обоснованию закона Ципфа была убедительно показана М.В. Араповым и Ю.А. Шрейдером (1977; 1978), применимость указанного метода к описанию других лингвостатистических распределений продемонстрирована в работах (Крылов Ю.К., Якубовская М.Д.,1977; Арапов М.В., Крылов Ю.К.,1980; Крылов Ю.К., 1982).

Эмпирические распределения и принцип наибольшего правдоподобия

В качестве основополагающего принципа, лежащего в основе наших рассуждений, возьмем принцип, который в статистике получил название метода наибольшего правдоподобия Р.А.Фишера

(см., например, Б.Л. Ван дер Варден, 1960), а в физике уже давно известен как принцип максимума статистической энтропии наблюдаемой системы. Согласно этому принципу наблюдаемые распределения частотных характеристик элементов системы могут отличаться от наиболее вероятных распределений лишь за счет сравнительно небольших случайных флуктуаций. Другими словами, в наблюдении нам даны лишь те состояния, которые обладают наибольшей вероятностью осуществления.

Рассмотрим множество всех потенциально допустимых текстов с фиксированной длиной в N словоупотреблений. Будем характеризовать частотную структуру любого из них матрицей

$$M = \begin{pmatrix} m_1 & m_2 \dots & m_k \dots & 1 \\ 1 & 2 & k & F_1 \end{pmatrix}, \qquad (I)$$

которую назовем матрицей первичной спецификации, или просто спецификацией, текста. В этой матрице k - й элемент первой строки имеет смысл числа слов данного текста, вошедшего в него ровно k раз. Так, в представленном тексте было m_1 одноразовых слов, m_2 - двухразовых, и т.д., причем последний столбец написан с учетом того, что для подавляющего большинства текстов кратность слова первого ранга (его абсолютная частота обозначена F_1) равна I. Таким образом в дальнейшем термины спецификация, частотная структура и лексический спектр текста в нашем изложении будут употребляться практически как синонимы.

В принятых обозначениях словарь текста дается выражением:

$$\mathcal{L} = \sum_k m_k , \qquad (2)$$

а его длина равна

$$N = \sum_k k m_k . \qquad (3)$$

Обозначим также через

$$\gamma_s = \sum_k k^s m_k \qquad s = 0, 1, 2 \dots \qquad (4)$$

начальные моменты частотной структуры (I) s - го порядка.

Очевидно, что частотная структура (I) однозначно определяет все начальные моменты $\gamma_0 = \mathcal{L}$, $\gamma_1 = N$ и γ_s ($s \geq 2$). В соответствии с теоремой Крамера о единственности решения системы линейных алгебраических уравнений с определителем не равным нулю (в данном случае в качестве определителя системы выступает так называемый определитель Вандермонда (см., например, Бугров Я.С., Никольский С.М.,

1980, с. 14), который отличен от нуля) верно и обратное: задание μ первых начальных моментов (μ равно числу ненулевых элементов второй строки матрицы (I)) достаточно для однозначного определения частотной структуры текста, так как система (4) полностью детерминирует значения элементов этой строки.

В реальной ситуации, однако, в качестве параметров, характеризующих данный текст, выступают не все начальные моменты, а лишь его словарь $\mathcal{L}$ и длина N. Вследствие этого допустимыми оказываются различные частотные структуры, от которых лишь требуется, чтобы их проекции одовлетворяли уравнениям (2) и (3).

В результате порождение текстов с различными спецификациями $\mathcal{M}$ обладает неодинаковыми вероятностями $p(\mathcal{M})$, так как чаще будут наблюдаться те частотные спектры, которым соответствует большее число потенциально допустимых способов их осуществления.

Поясним сказанное упрощенным числовым примером. Отвлекаясь от лингвистического содержания вопроса, допустим сначала, что число способов порождения текстов с заданной спецификацией просто равно числу принципиально допустимых различных представлений упорядоченных представлений частотных структур. Так для текста длиной в 7 словоупотреблений со словарем в 4 лексических единицы существует только 20 различных способов такого предетавления:

4+I+I+I	3+2+I+I	I+3+2+I	2+2+2+I
I+4+I+I	3+I+2+I	I+3+I+2	2+2+I+2
I+I+4+I	3+I+I+2	I+2+3+I	2+I+2+2
I+I+I+4	2+3+I+I	I+2+I+3	I+2+2+2
	2+I+3+I	I+I+3+2	
	2+I+I+3	I+I+2+3.	

При этом структура $\mathcal{M}_1=\begin{pmatrix}3&1\\1&4\end{pmatrix}$ реализуется четырмя различными способами, структура $\mathcal{M}_2=\begin{pmatrix}2&1&1\\1&2&3\end{pmatrix}$ двенадцатью, а $\mathcal{M}_3=\begin{pmatrix}1&3\\2&1\end{pmatrix}$ снова четырмя способами. Соответственно и вероятности случайных реализаций этих спецификаций будут равны $p(\mathcal{M}_1)=0{,}2=p(\mathcal{M}_3)$, а $p(\mathcal{M}_2)=0{,}6$, т.е. структура $\mathcal{M}_2$ обладает преимуществом и в среднем должна реализоваться в три раза чаще, чем каждая из $\mathcal{M}_1$ и $\mathcal{M}_3$.

При переходе к конкретным текстам числа N и $\mathcal{L}$ сильно увеличиваются. Соответственно резко возрастает и число допустимых лексических спектров, причем это увеличение

происходит фантастически быстро. Если снять ограничение на объем словаря и рассмотреть множество текстов, у которых фиксирована только длина N, то число различных допустимых частотных структур (удовлетворяющих лишь уравнению (3)) будет ничем иным, как числом различных разбиений $R(N)$ числа N. Поиск точной зависимости $R(N)$ был предметом исследования многих крупнейших математиков и, наконец, эта задача (см., например, Г. Энрюс, 1982) была решена Харди и Рамануджаном. Установлено, что $R(10) = 42$, $R(50) = 204226$, $R(100) = 190\,569\,292$, $R(200) = 3\,972\,999\,029\,388$. Для не слишком малых N справедливо асимптотическое представление

$$R(N) = \frac{1}{4\sqrt{3}N} \exp\left(\pi\sqrt{\frac{2N}{3}}\right),$$

которое показывает, сколько различных лексических спектров могут иметь тексты заданной длины.

Почему же практически реализуются далеко не все из этих возможностей? Дело в том, что с ростом N и $\mathcal{L}$ резко увеличивается и различие в числе комбинаций, способствующих осуществлению каждой конкретной спецификации. Это различие становится столь резко выраженным, что лишь для немногих из логически допустимых частотных структур вероятность их реализации значимо отличается от нуля. В конкретных текстах оказываются реализованы лишь некоторые из наиболее вероятных распределений. Соответственно, наиболее вероятный лексический спектр (которому соответствует наибольшее число способов его осуществления) может рассматриваться как теоретическое распределение, отклонение от которого в реальных спектрах обусловлено лишь незначительными случайными флуктуациями.

Порождение текста как квазистационарный процесс

Вернемся к лингвистически содержательной стороне вопроса. Чтобы получить теоретический спектр для текста с заданными значениями параметров N и $\mathcal{L}$, мы должны научиться подсчитывать число различных способов реализации любой из его допустимых спецификаций. В соответствии с принятой терминологией назовем это число $G(M)$ статистическим весом соответствующей спецификации. Строго говоря, для того чтобы правильно вычислить статистический вес $G(M)$ мы должны знать по крайней мере главные из ограничений, которые накладываются на последовательность появления различных лексических единиц в связном тексте законами его порождения.

К сожалению, наши знания в этом направлении пока крайне ограничены. По-видимому, задача может быть решена лишь последовательными итерациями в построении все более и более лингвистически содержательных моделей. Начав с простейших предположений, мы можем выявить те противоречия, к которым эти предположения приводят и путем устранения этих противоречий перейти к моделям, более адекватным действительности. В этой связи подчеркнем, что предлагаемая ниже модель ни в коей мере не претендует на окончательное решение задачи и является лишь одним из необходимых первых шагов в этом направлении.

Отметим, что для вычисления $G(M)$ вряд ли применимы комбинаторные рассуждения, базирующиеся на предположении, что связный текст построен из лексики языка чисто случайным образом, допускающим любые сочетания лексических единиц. Такие последовательности с необходимостью будут включать в себя комбинации, когда одно и то же слово появляется несколько раз подряд, что заведомо, если и происходит, то представляет аномальную ситуацию в реальном связном тексте. Учитывая сказанное, постараемся сформулировать правила вычисления $G(M)$ в понятиях, абстрагированных от порядка расположения в тексте конкретных лексических единиц и учитывающих это распределение лишь косвенным образом.

Начнем с того, что любой текст однозначно порождает последовательность из N чисел, каждое из которых равно кратности слова, стоящего в соответствующей позиции (ячейке) текста. В качестве примера рассмотрим отрывок из "главы никакой" "Приключений Алисы в стране чудес".

> "Надо знать, кто такие антиподы и что такое параллели и меридианы, надо знать, когда что случилось и что такое ткань повествования; надо знать, из чего не делается горчица и как правильно играть в крикет."

После того, как проделаны необходимые подсчеты, легко убедиться, что этой фразе на уровне словоформ соответствует спектр $M=\begin{pmatrix}19 & 1 & 3 & 1\\ 1 & 2 & 3 & 4\end{pmatrix}$ с параметрами $N=34$, $L=24$ и последовательность A:

33III432I4I33I3I432II33IIIII4IIIII (A)

В этой последовательности I9 единичек стоят на местах одноразовых словоформ, две двойки замещают единственную двухразовую словоформу "такое", девять троек расположены в

ячейках трехкратных словоформ "надо", "знать", "что", наконец четыре четверки соответствуют употреблениям союза "и". Любой другой текст с аналогичным спектром породит либо ту же самую последовательность A, либо другую, например такую:

$$I13I4III42I3III33I322I43IIII3I34IIII3 \qquad (A_I)$$

с тем же количеством единичек, двоек, троек и т.д.

Для вычисления статистического веса $G(M)$ нам "надо знать", сколько различных последовательностей A_i может соответствовать данной частотной структуре M и каким числом разных способов $S(A_i)$ может быть реализована каждая из таких последовательностей. При этом

$$G(M) = \sum_{i=1}^{J(M)} S(A_i), \qquad (5)$$

где суммирование производится по всем разным последовательностям, а $J(M)$ – их общее количество.

Расчеты существенно упростятся, если предположить, что любая из допустимых последовательностей A_i осуществляется с одинаковой вероятностью и одним и тем же числом разных способов. Тогда $S(A_i) = S(A) = Const$ для любого i и сумма (5) превращается в произведение:

$$G(M) = J(M) \cdot S(A) \qquad (6)$$

Можем ли мы принять сформулированную выше гипотезу, которую назовем гипотезой квазистационарности порождения связного текста? Прежде всего отметим, что эта гипотеза имеет определенное эмпирическое обоснование. В частности, как показано в работе (Крылов Ю.К., 1985), из нее как необходимое следствие вытекают формулы В.М. Калинина (1964), которые, как известно, довольно хорошо работают и на связных текстах.

Подчеркнем, что гипотеза квазистационарности не означает, что появление новой конкретной лексической единицы случайно по отношению к смыслу и содержанию предыдущего контекста, а лишь предполагает, что для любого K все вхождения K – кратных слов в совокупности покрывают текст квазислучайно. Другими словами, мы предполагаем, что нет существенных причин, которые систематически приводили бы к значительному скоплению, скажем, редкочастотной или высокочастотной лексики в фиксированном месте всех текстов, например, в их начале, середине или конце. Таким образом, предлагаемая гипотеза не отрицает как существования известного эффекта "сгущения" отдельных лексических единиц, когда одно появле-

ние какого-нибудь слова повышает вероятность его повторного появления в ближайшем контексте, так и, например, возможности неоднородного распределения по тексту различных форм повествования, в частности авторской речи и диалога. Важно лишь, что если некоторые отклонения от случайности в покрытии текста отдельными словами существуют, чтобы для всех K эти отклонения были пренебрежимо малы по сравнению со случайными флуктуациями в перераспределении совокупности положения всех K -разовых элементов при переходе от одного текста к другому.

Приняв гипотезу квазистационарности, легко вычислить множитель $S(M)$ в формуле (6). Для этого достаточно сосчитать, сколько различных последовательностей можно составить из N элементов спецификации M, с учетом того, что среди этих элементов есть одинаковые (m_1 - единичек, $2m_2$ - двоек и т.д.). Если бы все элементы А были различными, то число разных последовательностей просто равнялось бы $N!$ - числу перестановок из N различных элементов. Действительно, тогда на первое место мы могли бы поместить любой из N элементов. Таких разных способов было бы N. На второе - любой из $(N-1)$ оставшихся. Всего получим $N(N-1)$ разных возможных способов расположения первых двух элементов. Они независимо комбинируют с $(N-2)$ возможностями выбора третьего члена последовательности и т.д. Всего было бы

$$N(N-1)(N-2)\ldots 3\cdot 2\cdot 1 = N!$$

разных последовательностей. Однако любая перестановка местами двух одинаковых элементов, скажем двух единичек, не изменит исходной последовательности А. Всего единички можно переставить опять $m_1!$ способами, каждый из которнх не приведет к появлению новой последовательности. Все эти перестановки независимо комбинируют с $(2m_2)!$ неразличными перестановками двоек и т.д. с $(km_k)!$ тождественными перестановками K - кратных элементов. В результате

$$m_1!(2m_2)!(3m_3)!\ldots(km_k)!\ldots = \prod_K (km_k)!$$

перестановок оставят исходную последовательность без изменения. Аналогичные рассуждения справедливы и для любой другой последовательности A_1. Всего таких разных последовательностей $J(M)$. Перемножая $J(M)$ и $\prod_K (km_k)!$ получим все допустимые перестановки, т.е. $N!$ откуда

$$J(M) = \frac{N!}{\prod_{k} (k m_k)!} \qquad (7)$$

Функции слова в контексте. Лингвистическая интерпретация рассматриваемой модели

При вычислении $J(M)$ мы воспользовались, вообще говоря, известной формулой для числа различных перестановок в последовательности, содержащей повторяющиеся элементы. При этом по сравнению с предложенной ранее моделью М.В. Арапова и Ю.А. Шрейдера (1977, 1978), в которой эквивалентными считались вхождения в контекст лишь одного и того же слова, мы отождествили все вхождения k - кратных элементов. Можно ли считать такое отождествление правомерным с лингвистической точки зрения? Действительно, как понимать, что в приведенном выше примере лексемы "надо", "знать", "что" неразличимы? Ведь это различные лексемы!

Конечно, разные. Но их различие - словарное различие. Оно обусловлено классификацией по отношению к их функциям в языке в целом, к функциям, которые эти лексемы выполняют на множестве самых различных контекстов. Эта классификация внешняя по отношению к нашему тексту, ибо она осуществлена вне его и независимо от него и существовала бы даже если Л. Кэррол и не написал "Алисы в стране чудес". Функции же всех k - кратных слов, которые эти слова выполняют по отношению к формированию данного текста, статистически эквивалентны.

Основное назначение одноразовых слов - служить той средой, тем массивом, в которой "погружаются" все другие слова. Вряд ли можно утверждать, что достаточно большой текст существенно трансформируется, если какое-либо из одноразовых слов мы заменим подходящим синонимом. Скорее всего мы просто не заметили бы такой замены в тексте.

Двухразовые слова, с одной стороны, могут выполнять функции одноразовых слов (при независимых вхождениях), но обладают и дополнительными собственными функциями, например, могут способствовать локальному "скреплению" контекста. С их помощью, в частности, может уже осуществляться объединение предложений в сверхфразовые единства и т.п.

С увеличением k число функций, которые способны выполнять в тексте слова данной кратности, монотонно возрастает, так как очевидно, что эти функции характеризуются отношением включения: любое вхождение в контекст k - разового слова может исполнять как любую функцию, присущую более низ-

кочастотным словам, так и нести специфическую функцию.

Таким образом, диалектика порождения текста заключается в том, что в тексте разные слова сначала как бы обезличиваются. Они, сохраняя общеязыкоое значение, в то же время становятся как бы эквивалентными по отношению к тексту. Однако, вступив в новую систему отношений, они начинают различаться по своему статусу в рамках этой системы, по тем функциям, которые обусловлены их присутствием в этой системе. Это различие выступает сначала еще не как истиное различие, а лишь ea уровне антитезиса, с одной стороны, как отрицание общеязыковых различий, с другой - как отрицание полной эквивалентности по отношению к тексту, т.е. как статистическое различие функций слов одной кратности по отношению к функциям слов других кратностей.

Синтез, истинное различие слов в тексте, связан с изменением их семантики, с приобретением ими контекстуальных и текстовых значений, обусловленных всей семантикой текста. Индивидуальность слов воспроизводится в новом качестве, делая их абсолютно различимыми, - различимыми как в системе общеязыковых отношений, так и в системе организации смыслов данного текста. Соответственно, уже каждое вхождение даже одного и того же слова в текст оказывается строго индивидуальным. Изменяется статус различных вхождений слов в текст, необходимость присутствия слова в конце может индуцировать необходимость его предшествующих появлений, линейный порядок воспроизведения текста оказывается несущественным, так как не соответствует глубинной организации семантики в системе ее причинно-следственных отношений.

На уровне организации смыслов текста все слова выступают в первую очередь как индивидуальные, т.е. как однократные ($m_i = N$). Соответственно, семантическая значимость отдельных слов уже теряет непосредственную связь с их частотностью и, следовательно, не может быть определена чисто формально, особенно в художественных произведениях. Сказанное выше показывает, что семантическая и, особенно, художественная ценность текста вне целостного анализа его семантики вряд ли может определяться статистическими методами. Целостность (по терминологии М.В. Арапова (1982)) статистической организации текста обусловлена системными отношениями более низкого уровня, однако индивидуальность каждого текста должна быть учтена и в нашем рассмотрении, так как именно она оказывается ответственной за то, что

$G(M)$ - статистический вес каждой частотной структуры еще нельзя однозначно выразить через $S(A)$ - число различных числовых последовательностей, соответствующих этой структуре (см. формулу (7)).

Организация функций слова в контексте

Чтобы окончательно определить статистический вес $G(M)$, нам необходимо еще вычислить $S(A)$ - число различных способов реализации каждой конкретной числовой последовательности. Пусть T общее число ее возможных текстовых реализаций, как различных, так и эквивалентных. Обозначим количество последних $Q(A)$. Тогда, аналогично предыдущему, $S(A)\,Q(A) = T$ и, следовательно:

$$S(A) = \frac{T}{Q(A)} \qquad (8)$$

Для определения T и $Q(A)$ нам снова необходимо сделать дополнительные предположения о статистической организации связного текста. Очевидно, что T определяется всей системой языка в целом и не зависит от того, какая конкретная лексика наполняет каждый конкретный текст. Это позволяет допустить, что T функционально связано лишь с макропараметрами рассматриваемого множества текстов, их длиной N и словарем $\mathcal{L}$, т.е. $T = T(N, \mathcal{L})$. Назовем эту рабочую гипотезу - гипотезой индифферентности (в широком смысле) речетворческого процесса по отношению к его конкретному содержанию. Если принять гипотезу индифферентности, то зависимость числа различных текстов $S(A)$ от конкретного вида их лексического спектра M найдет свое опосредованное выражение в $Q(A)$ - числе эквивалентных текстов, каждый из которых порождает последовательность А с заданными значениями $m_1, m_2, \ldots, m_k, \ldots$

Чтобы ввести понятие о структурно эквивалентных текстах, заметим, что каждый текст наряду с последовательностью А может быть представлен и другой числовой последовательностью, которую обозначим В. В последовательности В расположим все единички на местах первых вхождений каждого слова в текст, двойками отметим все вторые вхождения и т.д. Для примера, приведенного ранее, последовательность В запишется в виде:

11111111121221213321133111114111111 (В)

Очевидно, что в последовательности В число единичек равно объему словаря $\mathcal{L}$ (в нашей фразе $\mathcal{L} = 24$), число

двоек совпадает с количеством слов употребленных в тексте два и более раз, число троек равно объему словаря слов, присутствующих в тексте не менее трех раз и т.д.

Будем называть два текста структурно эквивалентными, или одинаково лексически организованными, если для них полностью совпадают обе последовательности, как А так и В. У таких текстов для любого K одинаково не только расположение всех K - разовых элементов, но и для каждого слова совпадают "ячейки", занятые его первым, вторим т.д. вхождениями. Однако, как отмечалось выше, линейный порядок расположения слов в тексте, находящий свое отражение в последовательности В, в плане рассмотрения семантики всего текста, утрачивает свое значение. На уровне абсолютной различимости лексических единиц последующие вхождения способны индуцировать предидущие, функции, выполняемые отдельными вхождениями, могут не совпадать с кратностью слова. Функционально значимым оказывается лишь характер связей между различными вхождениями в текст одного и того же слова.

Учитывая вышесказанное, при подсчете $Q(A)$ допустим, что любое вхождение слова в контекст может быть либо изолированным (самонаправленным), либо обладать связью с любым другим вхождением этого слова. При этом естественно считать, что характер связи каждого из вхождений не зависит от того, как ориентированы связи других вхождений. Тогда для каждого K - кратного слова последовательности А существует K эквивалентных способов порождения последовательности В. Действительно, первое вхождение может быть напрвлено разными способами. Они независимо комбинируют с K возможностями ориентации направления связи второго вхождения. Всего для первого и второго вхождений K^2 способов их ориентации. Третье вхождение также обладает K допустимыми направлениями его связей, что приводит к $K^2 \cdot K = K^3$ разных комбинаций и т.д. Так как всего в тексте m_K K- разовых слов, то им соответствует

$$(K^K)^{m_K} = K^{(Km_K)}$$

комбинаций, а всего при фиксированной А последовательность В может быть порождена $Q(A) = \prod_K K^{(Km_K)}$ эквивалентнными способами, откуда с учетом (8)

$$S(A) = \frac{T(N, \Sigma)}{\prod_K K^{(Km_K)}}$$

и пользуя (7) получаем окончательное выражение для $G(M)$:

$$G(M) = \frac{N!\,T(N,\mathcal{L})}{\prod_k (km_k)!} \qquad (9).$$

Лексический спектр оптимальной частотной структуры

Полученная выше формула (9) позволяет вычислить теоретическое распределение численностей k - разовых слов в связном тексте - лексический спектр наиболее вероятной из всех допустимых структур с фиксированными значениями макропараметров: словаря

$$\mathcal{L} = \sum_k m_k \qquad (2)$$

и длины текста

$$N = \sum_k k m_k \qquad (3)$$

Для этого достаточно найти те значения m_k, которые обеспечивают максимум статистического веса (9) при дополнительных условиях (2) и (3).

Так как логарифм любой функции достигает экстремальной величины при тех же значениях аргументов, при которых экстремальна и сама функция, при отыскании максимума вместо $G(M)$ воспользуемся его логарифмом:

$$\ln G(M) = \ln N! + \ln T(N,\mathcal{L}) - \sum_k k m_k \ln k - \sum_k \ln (km_k)! \qquad (10)$$

Задача на условный экстремум выражения (10) может быть решена методом Лагранжа (см., например, В.И. Смирнов,1948). С этой целью приравняем нулю частные производные по всем m_k от функции

$$g = \ln G(M) + \alpha\left(\sum_k m_k - \mathcal{L}\right) + \beta\left(\sum_k k m_k - N\right) \qquad (11)$$

где α и β - неопределенные множители Лагранжа, значения которых впоследствии должны быть найдены из уравнений (2) и (3). Выполняя дифференцирование и воспользовавшись известным представлением факториала с помощью Г - функции Эйлера: (см., например, Янке Е., Эмде Ф., 1949 или Кратцер А., Франц В., 1963)

$$(km_k)! = \Gamma(km_k + 1) \qquad (12)$$

получим $k \ln k + k\,\psi(km_k) = \alpha + \beta k$

(13)

где

$$\psi(x)=\frac{d\ln\Gamma(x+1)}{dx}=-\lambda_0+\sum_{\nu=1}^{\infty}\left(\frac{1}{\nu}-\frac{1}{x+\nu}\right) \quad (14)$$

– логарифмическая производная Г – функции,

λ_0 = 0,5772156649..)

– постоянная Эйлера–Маскерони.

При $km_k > 1$, что имеет место в нашем случае, для $\psi(x)$ справедливо асимптотическое разложение (Янке Е., Эмде Ф., 1949)

$$\psi(x)=\ln x+\frac{1}{2x}-\frac{1}{12x^2}+\frac{1}{120x^4}-\cdots \quad (15)$$

Ограничиваясь первым членом ряда (15) с достаточной степенью точности можно считать $\psi(x)=\ln x$, откуда

$$k\ln k+k\ln(km_k)=k\ln(k^2m_k)=\alpha+\beta k, \quad (16)$$

что приводит к окончательному выражению для теоретического лексического спектра

$$k^2m_k=e^{\frac{\alpha}{k}+\beta}\Leftrightarrow m_k=\frac{C}{k^2}e^{\frac{\alpha}{k}}, \quad (C=e^{\beta}) \quad (17)$$

Качественный анализ теоретического распределения

Переходя к обсуждению полученного теоретического распределения (17) прежде всего заметим, что при $\alpha = 0$ оно равносильно

$$k^2m_k=C=Const, \quad (18)$$

что совпадает с одной из формулировок распределения слов по частотам, предложенной Дж. Ципфом (см. Арапов М.В., 1982) Однако входящие в формулу (17) параметры C и α должны удовлетворять соотношениям (2) и (3) и, следовательно, не являются лингвистическими константами, а зависят от параметров текста $C = C(N,\mathcal{L})$ и $\alpha = \alpha(N,\mathcal{L})$. В связи с этим условие $\alpha(N,\mathcal{L}) = 0$ не может выполняться для текстов любой длины N.

Действительно, α легко выразить в функции количества однократных

$$m_1=Ce^{\alpha} \quad (19)$$

и двухкратных

$$m_2 = \frac{C}{4} e^{\frac{\alpha}{2}} \qquad (20)$$

слов в тексте. Если разделить (I9) на (20) и решить полученное уравнение относительно α, то легко убедиться, что

$$\alpha = 2 \ln \frac{m_1(N, \mathcal{L})}{4 m_2(N, \mathcal{L})} \qquad (21)$$

Для коротких текстов всегда $\frac{m_1}{4m_2} > 1$ и $\alpha > 0$. С увеличением длины рассматриваемых текстов (или их фрагментов) при любом способе подсчета лексических единиц (лексемы, словоформы, ЛСВ, гиперлексемы и т.д.) отношение $\frac{m_1}{m_2}$ (с точностью до флуктуаций) может лишь монотонно уменьшаться. Следовательно, для <u>любого текста</u> и для <u>любых единиц</u> существует (по крайней мере потенциально) такая длина $N = Z$, для которой $\frac{m_1(Z)}{4m_2(Z)} = 1$, что равносильно $\alpha(Z) = 0$.

Таким образом из рассматриваемой теории <u>как необходимое следствие</u> вытекает гипотеза Ю.К. Орлова (1976, 1978) о том, что для любого текста существует длина $N = Z$ (объем Ципфа), на которой закон Ципфа-Мандельброта выполняется в каноническом виде. Более того, теоретически объясняется и установленный М.Г. Бородой и А.А. Поликарповым (1984) факт уменьшения Z при переходе от более мелких к более крупным единицам подсчета лексики. Ведь чем крупнее единица, тем меньше словарь при фиксированной длине текста, тем быстрее убывает отношение $\frac{m_1}{m_2}$ в функции N и тем самым при меньших длинах достигается условие $\alpha = 0$.

Необходимо особо подчеркнуть однако, что в отличие от Ю.К. Орлова (1976, 1978) и М.Г. Бороды и А.А. Поликарпова (1984), которые склонны считать выполнимость закона Ципфа на полной длине текста проявлением "сознательного (или бессознательного) стремления автора к организации повторов в тексте", с точки зрения представлений развиваемых в данной работе условие $\alpha = 0$ не является характеристикой целостности связного текста. Положенная в основу вывода формулы (17) гипотеза квазистационарности, наоборот, предполагает, что статистическая организация лексики любого фрагмента связного текста не обладает какими-либо существенными особенностями по сравнению с полными текстами аналогичной длины. Соответственно и значение $\alpha = 0$ не является какой-нибудь особой "качественной" точкой на множестве всех допустимых значений α.

Определение параметров теоретического распределения

Перейдем к вопросу вычисления параметров распределения (17). Казалось бы, наиболее простой способ их оценки состоит в использовании метода наименьших квадратов в соответствии с формулой (16). Однако полученные таким способом оценки не будут удовлетворять естественным условиям:

$$\sum_{k=1}^{k_0} \frac{C}{k^{\alpha}} e^{-\frac{k}{k}} = \tilde{\mathcal{L}} \qquad (2a)$$

и

$$\sum_{k=1}^{k_0} \frac{C}{k} e^{-\frac{k}{k}} = \tilde{N} \qquad (3a)$$

которые мы положили в основу наших рассуждений. Применив метод наименьших квадратов для определения α и β в конкретном тексте и используя затем полученные оценки при вычислении сумм (2а) и (3а), мы, строго говоря, получим значения словаря текста $\mathcal{L}$ и его длины N, не совпадающие с их реальными значениями $\tilde{\mathcal{L}}$ и $\tilde{N}$. В связи с вышесказанным для вычисления C и α будем непосредственно использовать уравнения (2а) и (3а).

Подчеркнем, что при аппроксимации реального распределения любой непрерывной функцией возникает неоднозначность в возможных способах отображения непрерывного распределения на теоретический дискретный спектр. Эта неопределенность имеет принципиальный характер и прежде всего связана с тем, что при больших k теоретические значения m_k становятся меньше единицы, что в свою очередь приводит к тому, что для многих k наблюдаемые значения $\tilde{m}_k = 0$. Следствием указанной неоднозначности является возможность различного определения верхнего предела суммирования k_0 в суммах (2а) и 3а), что делает его еще одним параметром теории. Сказанное выше относится не только к данной, но и к любой другой модели. В частности в неявной форме k_0 присутствует и в модели Ю.К. Орлова (1976, 1978), что и обеспечивает необходимое число степеней свободы этой модели, чтобы удовлетворить условиям $\mathcal{L} = \tilde{\mathcal{L}}$, $N = \tilde{N}$ и $F_1 = \tilde{F}_1$, где $\tilde{F}_1$ – наблюдаемая частота самого частого слова.

Будем считать, что в высокочастотной области (при $m < 1$) слову ранга z в суммах (2а) и (3а) соответствуют слагаемые с индексом k, пробегающим значения $k = k_{z-1}+1$, $k_{z-1}+2$ ..... k_z, причем

$$\sum_{k=k_{z+1}}^{k=k_{z-1}} m_k = 1 \qquad \text{и} \qquad \sum_{k=k_{z+1}}^{k=k_{z-1}} k m_k = F_z \tag{22}$$

где F_z – количество вхождений этого слова в текст. Тогда накопленная частота слов первых z рангов S_z оказывается равной

$$S_z = \sum_{z=1}^{z} F_z = \sum_{k=k_{z+1}}^{k_0} k m_k = N(k_0, C, \alpha) - N(k_z, C, \alpha) \tag{23}$$

а сам ранг определяется выражением:

$$z = \sum_{k=k_{z+1}}^{k_0} = \mathcal{L}(k_0, C, \alpha) - \mathcal{L}(k_z, C, \alpha) \tag{24}$$

Если потребовать, чтобы для некоторого z значение $S_z = \tilde{S}_z$ – совпадало с наблюдаемой суммой частот z первых рангов (в простейшем варианте $z = I$ и $\tilde{S}_1 = \tilde{F}_1$), то система уравнений (2а), (3а), (23), (24) оказывается замкнутой относительно неизвестных C, α, k_0, k_z и однозначно определяет их значения. Для ее решения разложим $e^{\frac{\alpha}{k}}$ в ряд по степеням $\frac{\alpha}{k}$ и для сокращения записи объединим оба уравнения (2а) и (3а) в одно:

$$\tilde{\alpha}_j = \sum_{k=1}^{k_0} \frac{C}{k^j} e^{\frac{\alpha}{k}} = \sum_{k=1}^{k_0} \frac{C}{k^j} \sum_{s=0}^{\infty} \frac{\alpha^s}{s! k^s} \tag{25}$$

Здесь $j = 2$ для (2а) и $j = I$ в случае (3а). Меняя порядок суммирования и используя известную формулу (Крамер А., Франц В., 1963, с. 34)

$$\sum_{k=1}^{k_0} \frac{1}{k^{s+2}} = \frac{(-1)^{s+1}}{(s+1)!} \left[\psi^{(s+1)}(k_0) - \psi^{(s+1)}(0) \right] \tag{26}$$

$\tilde{\alpha}_j$ можно записать в виде ряда: $\tilde{\alpha}_j = \sum_{s=0}^{\infty} \frac{\alpha^s}{s!} \sum_{k=1}^{k_0} \frac{1}{k^{s+j}} =$

$$= \sum_{s=0}^{\infty} \frac{(-1)^{s+j+1} \alpha^s}{s!(s+j-1)!} \left[\psi^{(s+j-1)}(k_0) - \psi^{(s+j-1)}(0) \right] \tag{27}$$

Чтобы вычислить $\psi(k_0)$ и ее производные

$$\psi^{(s+j-1)}(k_0) = \frac{d^{s+j-1} \psi(x)}{dx^{s+j-1}} \Big|_{x=k_0}$$

в точке $x = k_0 >> 1$, достаточно воспользоваться асимптотическим разложением (15). Значения же $\psi(0) = \lambda_0 = 5{,}772156610^{-1}$ и $\psi^{(s+j-1)}(0)$ известны (см., например, Прудников А.П., Брычков Ю.А., Маричев О.И., 1981, с.651)

$$\psi^{(s+j-1)}(0) = (-1)^{s+j}(s+j-1)!(1+b_{s+j})$$

где $b_1 = \lambda_0 - 1 = -0{,}4227843$, $b_2 = 0{,}64493407$, $b_3 = 0{,}2020569$, $b_4 = 0{,}08232323$, $b_5 = 0{,}03692776$ и т.д.

Если в выражениях для $\psi(k_0)$ и ее производных пренебречь членами порядка $\frac{1}{k_0^2}$ по сравнению с $\frac{1}{k_0}$ и $\frac{1}{k_0}$ по сравнению с $\ln k_0$, что допустимо, т.к. $k_0 \sim F_1 >> 1$ то из (27) следует, что:

$$\mathcal{L}(k_0) = C\left(e^{\alpha} + \sum_{s=0}^{\infty} \frac{\alpha^s}{s!} b_{s+2} - \frac{1}{k_0}\right) \quad (28)$$

$$N(k_0) = C\left(e^{\alpha} + \sum_{s=0}^{\infty} \frac{\alpha^s}{s!} b_{s+1} + \ln k_0\right) \quad (29)$$

Формулы (28) и (29) справедливы не только при $k = k_0$, но и для любого $k >> 1$. Практически ими можно пользоваться как в высокочастотной, так и в среднечастотной зоне спектра, ограничиваясь в суммах членами порядка α^3. В результате с достаточной степенью точности для большинства реальных текстов приближенно можно считать

$$\mathcal{L}(k_0) = C\left(e^{\alpha} + b_2 + \alpha b_3 + \frac{\alpha^2}{2} b_4 + \frac{\alpha^3}{6} b_5 - \frac{1}{k_0}\right) \quad (30)$$

$$\mathcal{L}(k_2) = C\left(e^{\alpha} + b_2 + \alpha b_3 + \frac{\alpha^2}{2} b_4 + \frac{\alpha^3}{6} b_5 - \frac{1}{k_2}\right) \quad (31)$$

$$N(k_0) = C\left(e^{\alpha} + \ln k_0 + b_1 + \alpha b_2 + \frac{\alpha^2}{2} b_3 + \frac{\alpha^3}{6} b_4\right) \quad (32)$$

$$N(k_2) = C\left(e^{\alpha} + \ln k_2 + b_1 + \alpha b_2 + \frac{\alpha^2}{2} b_3 + \frac{\alpha^3}{6} b_4\right) \quad (33)$$

Вычитая из (30) (31) и из (32) (33), получим:

$$z = \mathcal{L}(k_0) - \mathcal{L}(k_2) = C\left(\frac{1}{k_2} - \frac{1}{k_0}\right) \Rightarrow \frac{1}{k_2} = \frac{zk_0 + C}{Ck_0}$$

$$S_2 = N(k_0) - N(k_2) = C \ln \frac{k_0}{k_2}$$

откуда

$$S_2 = C \ln\left(1 + \frac{zk_0}{C}\right) = C \ln\left(1 + \frac{z}{B}\right) \quad (34)$$

где обозначено $B = \frac{C}{\kappa_0}$. Потребовав $\mathcal{L}(\kappa_0) = \tilde{\mathcal{L}}$, $N(\kappa_0) = \tilde{N}$, и $S_2 = \tilde{S}_2$ систему (30), (31), (34) можно решить численно методом простых итераций. Сходимость вычислительного процесса обеспечивает последовательность итераций по схеме:

$$\kappa_{0i} = \frac{C_i}{\tilde{\mathcal{L}}}\left(\exp\frac{\tilde{S}_2}{C_i} - 1\right)$$

$$\alpha_{i+1} = \ln\left(\frac{\tilde{\mathcal{L}}}{C_i} - b_2 - \alpha_i b_3 - \frac{\alpha_i^2}{2} b_4 - \frac{\alpha_i^3}{6} b_5 + \frac{1}{\kappa_{0i}}\right)$$

$$C_{i+1} = \frac{\tilde{N}}{\ln\kappa_{0i} + b_1 + e^{\alpha_{i+1}} + \alpha_{i+1} b_2 + \frac{\alpha_{i+1}^2}{2} b_3 + \frac{\alpha_{i+1}^3}{6} b_4},$$

причем в качестве нулевого приближения достаточно положить

$$C_0 = \tilde{\mathcal{L}}, \quad \alpha_0 = 0.$$

Программа обеспечивающая вычисление параметров на микрокалькуляторах типа Б3-34 и МК-56 приводится ниже:

ИП6	ИПС	+	Fe^x	1	−	ИПС	x	ИП7	÷
П1	ИП8	ИПС	÷	ИП2	−	ИПА	ИП3	x	−
ИПА	Fx^2	2	÷	ПВ	ИП4	x	−	ИПВ	ИПА
x	3	÷	ИП5	x	−	ИП1	$F\frac{1}{x}$	+	Fln
ПА	Fe^x	ИП0	+	ИПА	ИП2	x	+	ИП1	Fln
+	ИПА	Fx^2	2	÷	ПВ	ИП3	x	+	ИПВ
ИПА	x	3	÷	ИП4	x	+	2	$F\frac{1}{x}$	ИПА
−	ИП1	÷	+	$F\frac{1}{x}$	ИП9	x	ИПС	XY	ПС
−	ИПС	÷	Fx^2	$F\sqrt{}$	ИПД	−	$Fx<0$	00	СП

Перед началом вычислений необходимо записать в ячейки памяти П2–П5 значения констант b_2 = 0,64493407, b_3 = 0,2020569, b_4 = 0,08232323, b_5 = 0,03692776. В регистр П0 заносится $b_1 = \lambda_0 - 1 = -0{,}4227843$. Величины $\tilde{N}$ и $\tilde{S}_2$ записываются соответственно в регистры П9 и П6, а $\tilde{\mathcal{L}}$ присваивается сразу двум регистрам П8 и ПС. Кроме того в ячейку ПА заносится 0, а ранг $\tilde{\mathcal{L}}$, соответствующий сумме $\tilde{S}_2$ в П7. (При условии $\tilde{S}_2 = f_1$, в П7 заносится просто 1) Наконец, необходимая точность вычислений $\delta = \frac{|\alpha C|}{C} \sim \frac{1}{\tilde{\mathcal{L}}} \sim 10^{-5}$ хранится в регистре ПД. Пуск программы осуществляется командами ВО СП. Время вычислений зависит от точности δ и быстродействия конкретного микрокалькулятора, но обычно не превышает нескольких минут. В результате вычислений значения параметров находятся: α в ячейке А, С в ячейке С, κ_0 в яч. I.

После того, как параметры определены, теоретические частоты при малых z естественно вычислять по формуле (17). Для расчета частотной структуры в голове распределения удобнее всего исходить из уравнения (34). Учитывая, что $F_z = S_z - S_{z-1}$, сразу получим:

$$F_z = C \ln\left(1 + \frac{1}{B+z-1}\right). \qquad (55)$$

При $\frac{1}{z+B-1} \ll 1$, что всегда имеет место в среднечастотной зоне приближенно

$$F_z \approx \frac{C}{z+B-1},$$

т.е. формулу Мандельброта с $\gamma = 1$. Таким образом, рассматриваемая теория предсказывает наличие линейного участка на графике зависимости $\ln F_z = \ln F_z(\ln z)$ в среднечастотной зоне спектра для любого текста. Однако величина этого участка оказывается зависящей от того, где справедливы вышесформулированные допущения и может варьироваться от текста к тексту.

Подробный анализ соответствия полученного теоретического распределения эмпирическим данным, выяснение достоинств и недостатков предлагаемой модели по сравнению с другими статистическими моделями представляет самостоятельное исследование и выходит за рамки возможностей данной публикации. Предварительное сопоставление теории с экспериментом показало, что модель удовлетворительно работает на связных текстах практически любого объема (от самых коротких типа стихотворения в прозе И.О. Тургенева "Как хороши, как свежи были розы" вплоть до текстов, превышающих по длине 100 000 словоупотреблений, например, "Поднятая целина" М.А. Шолохова) как для словоформ, так и для спектров лексем. Систематическое отклонение теоретических спектров от экспериментальных пока было обнаружено лишь для сводных частотных словарей объемом в 10^6 и более словоупотреблений, для описания спектров которых данная модель и не предназначена, так как в ней рассматриваются лишь связные тексты, а не их конгломераты.

Л И Т Е Р А Т У Р А

Арапов М.В., Шрейдер Ю.А. Классификации и ранговые распределения. – Научно-техническая информация. Серия 2, М., 1977, № 11-12, с. 15-21.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. – К кн.: Семиотика и информатика. Вып. 10, М7, 1978, с. 74-95.

Арапов М.В. Текст и язык – целостность и организованность. – Учен. зап. Тарт. ун-та. Вып. 628. Квантитативная лингвистика и автоматический анализ текстов, 1982, с. 55-66.

Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единицы различных уровней организации текста. – Учен. зап. Тарт. ун-та. Вып. 689. Квантитативная лингвистика и автоматический анализ текстов, 1984, с. 35-61.

Бугров Я.С., Никольский С.И. Элементы линейной алгебры и аналитической геометрии. М.: Наука, Г.И.Ф.-М.Л., 1980.

Ван дер Варден Б.Л. Математическая статистика. – М.: И.Л., 1960, с. 184.

Калинин В.М. Некоторые статистические законы математической лингвистики. – Проблемы кибернетики. Вып. 11. М., 1964, с. 244.

Кратцер А., Франц В. Трансцендентные функции. – М.: ИИЛ, 1963.

Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова. – Научно-техническая информация. Серия 2. Вып. 3. М., 1977, с. 1-7.

Крылов Ю.К. Об одной парадигме лингвостатистических распределений. – Учен. зап. Тарт. ун-та. Вып. 628. Труды по лингвостатистике, 1982, с. 80-102.

Крылов Ю.К. К вопросу о динамике нарастания объема словаря случайной выборки и связного текста. – Учен. зап. Тарт. ун-та. Вып. 711. Квантитативная лингвистика и автоматический анализ текстов, 1985, с. 55-66.

Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. – В кн.: Вычислительная лингвистика. М.: Наука, 1976, с. 179-202.

Орлов Ю.К. Модель частотной структуры лексики. – В кн.: Исследования в области вычислительной лингвистики. М.: Из-во МГУ, 1978, с. 59-118.

Прудников А.П., Брычков Ю.А., Маричев О.И. Интегралы и ряды. Элементарные функции. – М.: Наука, 1984.

Смирнов В.И. Курс высшей математики. Том I. – М.; Л.: ГИТТЛ, 1948.

Туллава Ю.А. Частотная структура текста и закон Ципфа. – Учен. зап. Тарт. ун-та. Вып. 711. Квантитативная лингвистика и автоматический анализ текстов, 1985, с. 93–116.

Эндрюс Г. Теория разбиений. – М.: Наука, 1982, с. 80–99.

Янке Е., Эмде Ф. Таблицы функций с формулами и кривыми. М.; Л., 1949, с. 106–120.

Arapov M.V., Krylov Yu.K. – Mathematical Models of classification in Application to Some Problems of Statistical Linguistics, – In: Computational Linguistics and Related Topics. Symposium. Tallinn, 1980, pp. 14–15.

Krylov Yu.K., Yakubovskaya M.D. – On Some Possibilities of Applying the Quantum-mechanical Formalism when Constructing Stochastic Models of Text Generation and Recognition. – Symposium on Grammars of Analysis and Synthesis and Their Representation in Computational Structures. – Summaries, Tallinn, 1983, pp. 49–51.

A STATIONARY MODEL OF COHERENT TEXT GENERATION

Yu.K. Krylov

S u m m e r y

Unlike the traditional approach, with the word as the element of the probabilities epace describing text generation, the article regards the theoretical frequency structure as the most probable point in the epace of all the potentially possible texte with given values of macroparameters - text length and vocabulary size. The probability of occurrence of each of the possible frequency structures is determined axiomatically on the basis of a priori natural combinatorial considerations. It is shown that the new model acoounts for the main empirical regularities of rank distributions in coherent texts. A set of methods has been elaborated for establishing the parameters of the theoretical distributions and approximate expressions are given that allow a simple comparison of the theory with the experiment.

О МЕРЕ БЛИЗОСТИ ТЕОРЕТИЧЕСКОГО И ЭМПИРИЧЕСКОГО РАСПРЕДЕЛЕНИЙ

/на материале распределений длин лексических единиц в тексте/

Н.С. Манасян

Проблема изучения статистических характеристик длины лексических единиц имеет немаловажное значение в изучении самых различных аспектов языка - в фонологии, функциональной стилистике, прикладной лингвистике и др. Эта проблема привлекала и привлекает многих исследователей. Одни авторы ограничивались либо просто нахождением частот встречаемости слов с той или иной длиной, либо попытками как-то объяснить или прогнозировать распределение частот в генеральной совокупности по тенденции их изменения на уровне первичной статистической структуры /Allén ; Никонов/. Другие исследователи подходят к проблеме длины слова как к проблеме аналитического описания распределений длин слов. Как правило в таких случаях за основу берется словоупотребление в виде словоформы /ср. Herdan ; Мартыненко; Тулдава/.

В настоящей работе ставятся следующие задачи:

I/ Проверка логнормального распределения на пригодность по описанию распределения длины словоупотребления в тексте.

2/ χ^2 критерий как мера близости распределений длин словоупотреблений.

3/ Нахождение альтернативы критерию χ^2[1].

Нами была предпринята попытка рассмотреть распределения длин словоупотреблений на материале данных ряда частотных словарей, представленных в виде IO эмпирических распределений /см. табл. I/.

---

[1]Автор пользуется случаем, чтобы выразить самую глубокую благодарность канд. ф.-м. наук И.М.Сливняку за помощь в решении задачи, за консультации и ценные советы на протяжении работы над статьей.

Таблица I.

№ пп	язык/подъязык	тип учетных единиц	объем выборки
I.	шведский	словоформа	I 000 669
2.	шведский	лексема	I03 4I6
3.	английский	словоформа	I 0I4 232
4.	английский	лексема	50 406
5.	английский/электроника	словоформа	20I 865
6.	английский/электроника	лексема	I0 486
7.	украинский	лексема	62 587
8.	украинский	словоформа	364 982
9.	английский/механика	словоформа	I 79I
I0.	английский/биохимия	словоформа	II 373

Примечание. Для шведского языка данные взяты из /Allén/, для английского из / Kučera, Francis/. По английскому подъязыку электроники П.М.Алексеев любезно предоставил автору полученные им ряды распределений длин лексических единиц по /Алексеев/. Эмпирические распределения по украинским текстам были получены П.М.Алексеевым совместно с автором по /Частотний словник.../. Данные по английским подъязыкам механики и биохимии были получены автором статьи.

Судя по данным, распределения длин лексических единиц, как лексем, так и словоформ, по частоте есть одновершинные распределения, скошенные вправо /см., например, рис. I и 2/, представляющие типичные такие графики, где на оси абсцисс lg длины, а на оси ординат - lg ее частоты/. Кривые напоминают усеченное логнормальное распределение.

В литературе по аналитическому описанию длины слова в

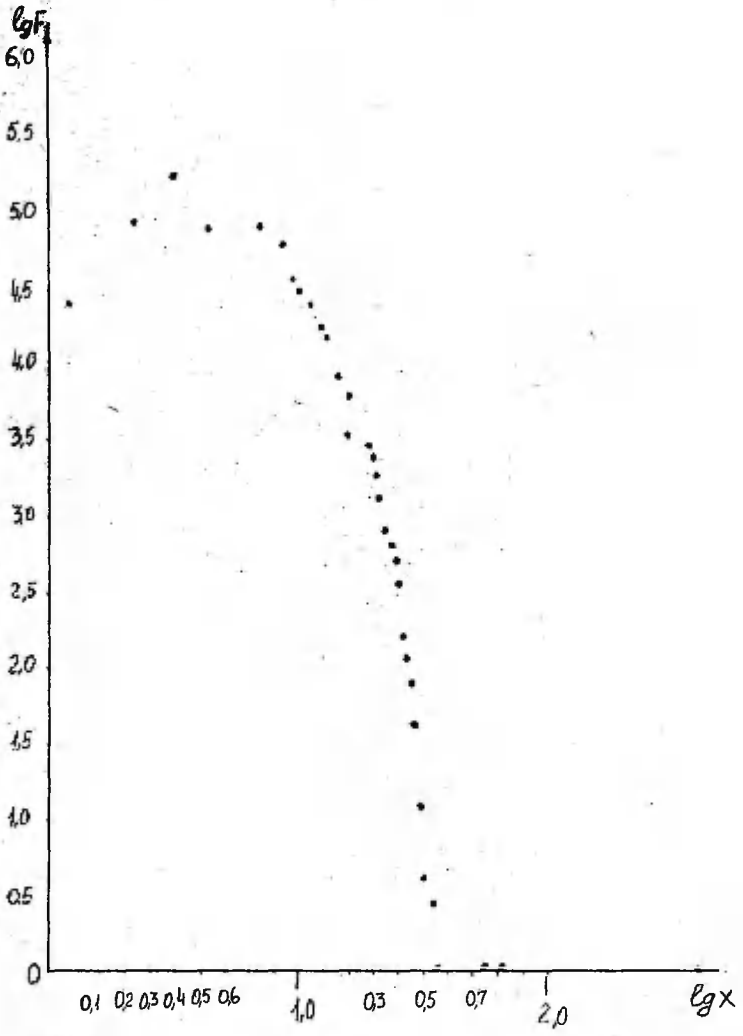

Рис. I. Распределение длин словоформ в шведском тексте.

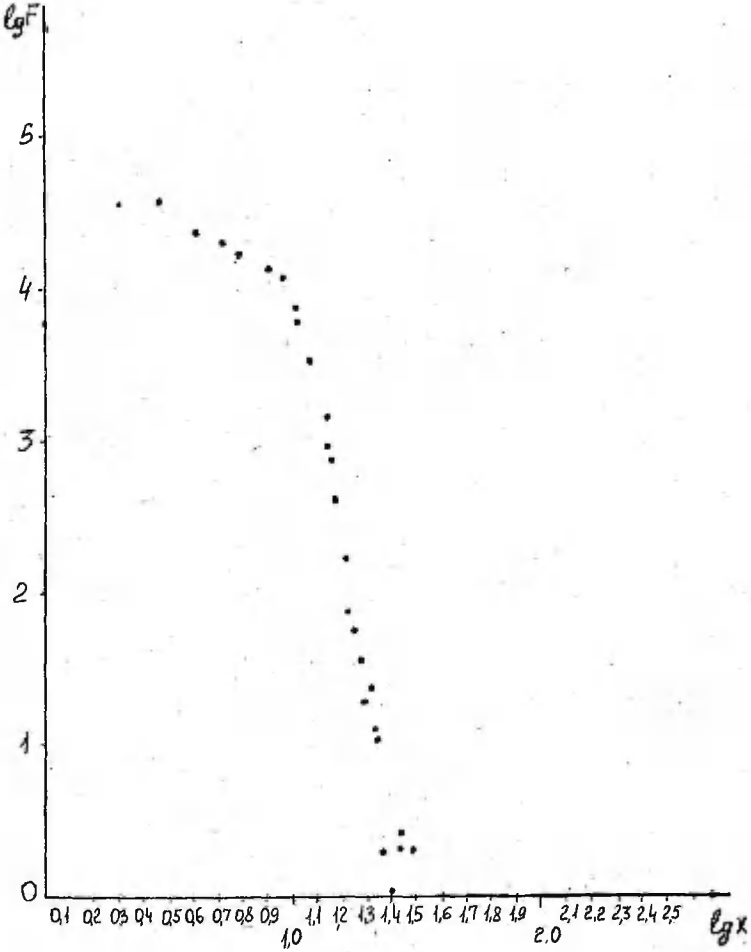

Рис. 2. Распределение длин словоформ в английских текстах по электронике.

графемах можно встретить утверждение о том, что распределение длины лексической единицы имеет логнормальный характер /Herdan; Пиотровский, Бектаев, Пиотровская; Тулдава/. Причем данные одних авторов /Herdan , с.134/ показывают, что логнормальное распределение хорошо описывает длину лексической единицы как на словарной, так и на текстовой оси. Данные других /см. Тулдава, с.20/, обнаруживают хорошее согласие с теоретическим логнормальным законом только на словарной оси.

Известно, что интегральная кривая нормального /логнормального/ распределения будет спрямлена, если на ординате берется нормированный масштаб. Имеется метод проверки логарифмической нормальности, который предусматривает графическое расположение точек эмпирического распределения на логарифмической гауссовой бумаге, на которой одна из осей координат распределена по интегральной функции нормального распределения, а другая имеет логарифмическую шкалу. Если эмпирический ряд логнормален, то графическое изображение эмпирической функции будет представлять собой прямую /подробно об этом см. /Aitchieon, Brown/.

Построив подобные графики на основе приведенных данных, мы получили группы точек, расположенных приблизительно по прямой линии. Типичные графики представлены на рис. 3 и 4. Следует заметить, что большинство графиков имеют точки перелома. Следствием этого факта является аппроксимация эмпирического распределения по частям; результаты этой аппроксимации отражены в табл.2.

Была составлена программа по проверке гипотезы о логнормальном распределении для исследуемых лингвостатистичеоких совокупностей. Параметры логнормального распределения для анализируемых 10-ти рядов распределения были определены аналитически, методом наименьших квадратов. Проверка соответствия эмпирических распределений теоретическому дала отрицательные результаты *.

Исходя из общих соображений нельзя было ожидать, что данные эмпирические распределения являются в точности логнормальными. Речь, как нам кажется, может идти только о степени сходства с логнормальным распределением.

---

* Это еще раз подтверждает наблюдения некоторых лингвостатистиков /ср.Сегал/ о том, что очень часто Г.Херданом выдвигались постулаты, которые не были проверены надостаточно представительном материале.

Причем здесь следует учесть то, что если величина выборки не очень велика, то возможности различения теоретического и эмпирического распределения ограниченны.* Для того, чтобы уверенно отвергнуть указанную гипотезу, расхождение между этими распределениями должно быть достаточно сильно. Поэтому часто для малых выборок критерий согласия χ^2 не отвергает гипотезу о логнормальности /хотя это и не означает, что логнормальный закон распределения доказан/. Если же неограниченно увеличивать объем словаря, то становится возможным улавливать все более тонкие различия между теоретическими и эмпирическими распределениями /с формальной точки зрения величина критерия χ^2 при заданных частотах прямо пропорциональна объему словаря, и следует ожидать, что гипотеза о логнормальности будет отвергнута/. Это и наблюдалось для всех достаточно больших словарей /ср. результаты вычислений в табл.2/.

С самого начала вопрос должен был ставится иначе – не о совпадении эмпирического распределения с логнормальным, а только о степени сходства.

Естественно с этой точки зрения сравнивать различные распределения между собой по степени их близости к логнормальному распределению, подсчитанной с помощью какой-либо меры "расстояния" между распределениями.

Что касается меры расхождения χ^2 между двумя распределениями, то она обладает двумя недостатками.

I/ Она меняется от 0 до ∞, а хотя бы из соображений удобства хотелось бы, чтобы она менялась от нуля до I.

2/ Она зависит от объема выборки. Если мы возьмем два экземпляра одного и того же словаря, т.е. умножим объем выборки на 2, то χ^2 также увеличится в два раза. Для меры расхождения это плохо. Свойства у словаря тежеже, а статистическая картина оказывается некорректно представленной, так как словарь тот же, просто он дублирован.

Путем простого преобразования этих двух неудобств можно легко избежать. Это преоразование имеет вид, аналогичный формуле коэффициента сопряженности /можно без знака квадратного корня/ /ср.Закс/; обозначим его через χ.

---

*Ср. /Арапов, Ефимова, Шрейдер, с.I3/, где приблизительно такая же мысль развивается относительно закона Ципфа.

Таблица 2

Значения критерия χ^2 и меры $\mathcal{Z}$ для I0-ти исследуемых рядов /нумерация рядов распределения соответсвует нумерации в табл. I/

№№ пп	границы интервалов	χ^2	$\mathcal{Z}$
I.	7 - 20	$4.64.10^4$	0.22
	I - 7	$I.66.10^4$	0.I3
2.	7 - 24	$I.I2.10^3$	0.I0
	I - 8	$I.6I.10^3$	0.I2
3.	8 - I5	85 550	0.09
	I - 8	I9 800	0.I4
4.	8 - 20	I I90	0.I5
	I - 9	64.3	0.036
5.	7 - I5	3 850	0.I4
	I - 7	6 9I0	0.I8
6.	3 - 20	35.8	0.58
	I - 3	3I5	0.I7
7.	I - I2	8.35.I0	0.086
8.	5 - I5	$I.02.10^4$	0.24
	I - 5	$2.20.10^4$	0.I65
9.	7 - I3	3.48.I0	0.I4
	I - 7	2.28.I0	0.II
I0.	6 - I8	$2.09.10^2$	0.I35
	I - 6	$2.04.10^4$	0.I33

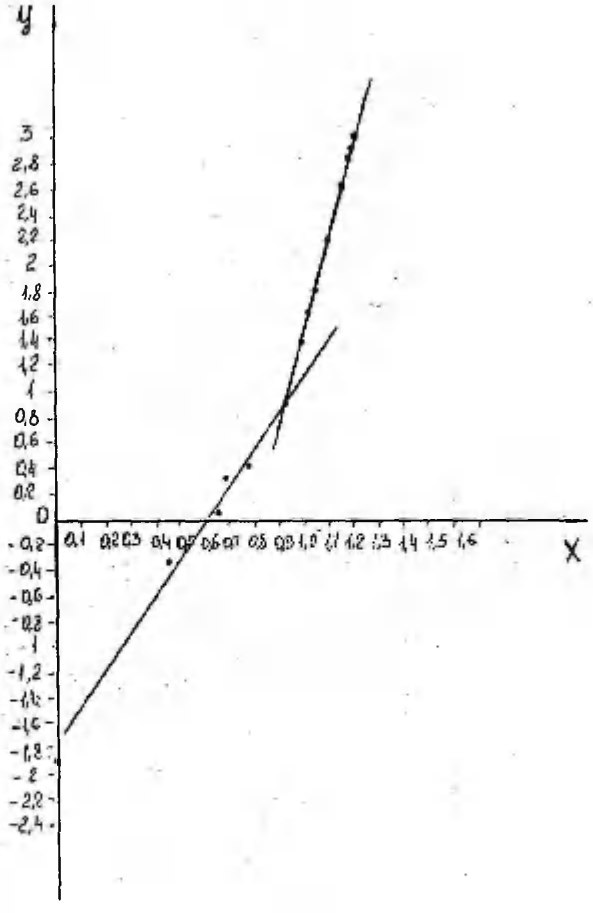

Рис.3. Спрямленный график распределения длин словоупотреблений в шведском тексте.

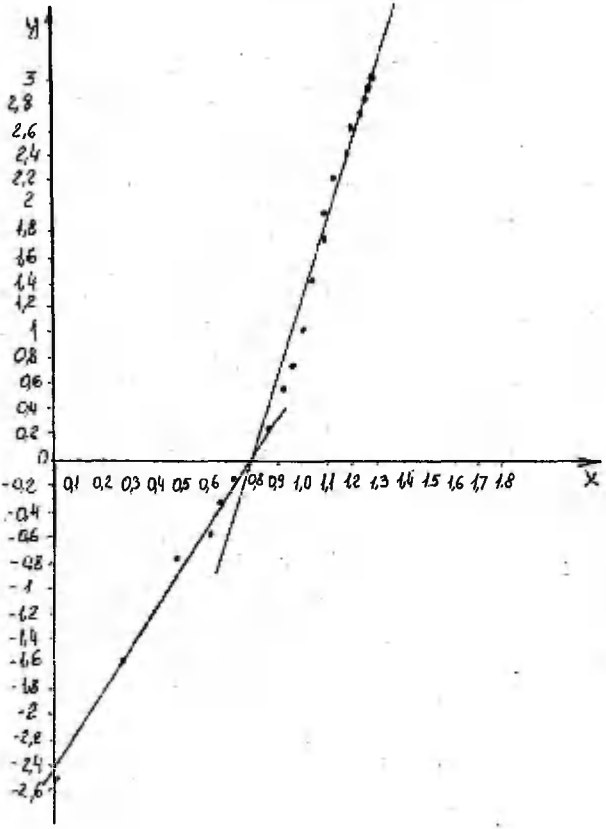

Рис.4. Спрямленный график распределения длин словоупотреблений в английских текстах по электронике.

$$\mathcal{Z} = \sqrt{\frac{\chi^2}{\chi^2 + n}} . \qquad / I /$$

Как следует из формулы /I/, величина $\mathcal{Z}$ меняется в пределах от 0 до I и не зависит от объема словаря*.

Учитывая все вышесказанное, нами были проанализированы на логнормальность IO рядов распределения длин лексических единиц. Затем для каждого ряда были получены значения величины $\mathcal{Z}$ /ср. табл. 2/.

Думается, что при наличии достаточного количества сопоставимых частотных распределений по различным языкам и подъязыкам можно будет классифицировать частотные словари при помощи меры близости $\mathcal{Z}$. Это может явиться предметом последующих исследований.

---

*Следует отличать меру $\mathcal{Z}$ от коэффициента сопряженности Пирсона. Эти две величины имеют следующие содержательные различия: коффициент сопряженности предназначен для оценки связи между двумя признаками, а в нашем случае сравниваются два распределения, эмпирическое и теоретическое.

ЛИТЕРАТУРА

Алексеев П.М. Частотный англо-русский словарь-минимум по электронике. - М.: Воениздат, 1971.

Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А. О смысле ранговых распределений. НТИ, сер. 2, № 1, 1975.

Закс Л. Статистическое оценивание. М.: Статистика, 1976.

Мартыненко Г.Я. Некоторые статистические наблюдения на материале болгарского языка. - Статистико-комбинаторное моделирование языков. М.-Л.: Наука, 1965.

Никонов В.А. Длина слова. ВЯ, 1978, № 6.

Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. М.: Высшая школа, 1977.

Тулдава Ю.А. Проблемы и методы квантитативного исследования лексики. АДД. Тарту, 1984.

Частотний словник сучасної украиньскої художньої прози. Киев: Наукова думка, 1981.

Aitchison J., Brown J.A.C. The lognormal distribution. Cambridge, 1957.

Allén S. Nusvensk frekvensorbok baserad på tidningstext. Stockholm, 1970-1971.

Herdan G. Type-token mathematics. London, 1964, c.61-62.

Kučera H., Fransis W.N. Computational analysis of present-day American English. Providence, 1967.

ON THE PROXIMITY MEASURE OF THEORETICAL AND EMPIRICAL DISTRIBUTIONS
(the lexical length distribution)

Nariney Manasyan

S u m m a r y

It is shown that lexical units length distribution may be approximated by the lognormal law. It is also shown that χ^2- criterion is not sufficient since its value ranges from 0 to ∞ and is dependent on the volume of the sample. As an alternative to χ^2 another index (Z) is suggested which is free from such disadvantages. It is maintained that this measure can be applied in frequency dictionary classifications.

ЧАСТОТНЫЙ СЛОВАРЬ СЛОВОСОЧЕТАНИЙ В АНГЛОЯЗЫЧНЫХ ГАЗЕТНЫХ ТЕКСТАХ

И.А. Мацукова

Материалом для формирования выборки и составления частотного словаря газетных словосочетаний послужили статьи на внутри- и внешнеполитические темы из 100 номеров газеты **"Morning Star"** за 1983-1985 г.г. Общая длина обследованных текстов - 200 тыс.словоупотреблений /35470 случаев употребления словосочетаний/. Из входящих в выборку текстов вручную выписывались целостные, регулярно воспроизводимые словосочетания с различной степенью устойчивости. Статьи обрабатывались от начала до конца. Для каждой единицы словаря указаны ее ранг /порядковый номер по убыванию частоты/ и абсолютная частота. Ниже приводятся 685 словосочетаний с частотой не менее 7.

Использованы следующие сокращения:

adj-прилагательное;
d -название дня недели;
f -цифровая форма числительного;
g n -географическое название;
m -название месяца;
n -существительное;
num -буквенная форма числительного;
pron-местоимение;
s,es-окончания множественного числа существительного;
t -название отрезка времени /неделя, месяц, год/;
v -глагол;
y -цифровая форма числительных, обозначающих порядок летоисчисления;

заключение в скобки суффикса **'s** означает возможность употребления впереди стоящего существительного как в общем, так и в притяжательном падеже;

многоточие означает наличие переменного компонента внутри словосочетания; многоточие в скобках указывает на возможность присутствия или отсутствия переменного компонента у словосочетания.

В целях снятия грамматической и лексико-грамматической омонимии все глаголы в начале словосочетаний употреблены с частицей **"to"**; отсутствуют глаголы в форме **Past Indefinite**; послелоги в отличие от предлогов отмечены знаком *.

1	there be	321
2	per cent	271
3	trade union	226
4	on (d)	176
5	general secretary	151
6	Labour Party	124
7-8	between ... and, to have to (v)	107
9	to call for	100
10	to be to (v)	98
11	Communist Party	95
12	on (m) (f)	94
13-14	one of, South Africa	90
15	more than	81
16	last night	80
17-18	Prime Minister, support for	78
19-20	to accuse ... of, at the weekend	73
21-22	in (m), National Union of	70
23-24	Labour MP, trade unionist	62
25	nuclear weapons	61
26	Cruise missiles	58
27	attack on	54
28	Coal Board	53
29-30	at least, to take part (in)	51
31	because of	50
32-34	last week, last year, to take place	49
35-38	at (f) a m(p m), to call on (upon) ... to (v), picket line, to set up*	48
39	in support of	47
40	local authority	45
41-44	Greenham Common, Labour government, power station, South African (n)	44
45	both ... and	43
46	to point out* (that)	42
47-49	Labour leader, shop steward, Soviet Union	41
50-55	as well as, (to be) able to (v), Labour movement, Northern Ireland, nuclear disarmament, on strike	40
56-57	general election(s), involved in	39
58-59	according to, general council	38
60-61	Morning Star, nuclear war	37
62-65	a week, to describe ... as, to lead to, some of	36

66-67	at (g n), from ... to	35
68	South Wales	34
69-73	to be going to (v), call for, civil service, many of, (m) (f)	33
74-77	as ... as, for ... years, Foreign Minister, peace movement	32
78-84	(to be) due to (v), to be expected to (v), cut(s) in, People's March for Jobs, this year, trades council, transport union	31
85-86	press conference, to take action	30
87	on (d) night	29
88-93	a number of, civil servant, in order to (v), opposition to, this week,year(s) ago	28
94-99	British government, industrial action, to make it clear that, peace camp, strike action, Transport and General Workers' Union	27
100-104	action against, at the same time, (to be) likely to (v), debate(s) on, to go on*	26
105-111	aimed at, all over (n), to deal with, High Court, to see ... as, to send ... to, to take over*	25
112-117	to carry out* , County Council,Defence Secretary, demand(s) for, last month, meeting with	24
118-126	apartheid regime, (to be) opposed to, Home Office, in protest at, plan(s) for, talks with, West Midlands, White House, young people	23
127-134	a year, to come from, to fail to (v), health authority, increase in, last (d), mass meeting, public service(s)	22
135-145	annual conference, at (the) ... level, Greater London Council, human rights, more ... than, next week, nuclear missile(s), post office, to return to work, so far, such as	21
146-155	(to be) responsible for, executive committee, to fight for, manual workers, national executive, on behalf of, out of work, to result in, Tory Party, West Germany	20
156-164	to agree to, health service, in favour of, local government, nuclear arms, to step up*, to take step(s), the rest (of), Tory MP	19

165-174 arms race, to ask for, to be part of, Downing Street, end to, less than, loss of jobs, pay offer, peace campaigner, to spend ... on 18

175-189 along with, at the end of, civil defence, to comment on, executive council, fight for, to lose (...) job(s), Ministry of Defence, National Coal Board, need for, to protest against, striking miner, the number of, throughout the country, town hall 17

190-206 a series of, to arrive in, as part of, ban on, Fleet Street, to go back*, to go on strike, government('s) policy, miners' strike, responsibility for, return to work, Scottish area, Secretary of State, so that, social service, to talk about, union leader 16

207-227 as a result of, at a (the) meeting, based on(upon), to be confident, British people, campaign for, Central America, to come to, engineering union, job loss(es), to look at, next (d), over the weekend, overtime ban, peace woman, reduction in, to respond to, solidarity with, strike by, Tory government, Western Europe 15

228-260 a total of, at present, to be against, to be out, to campaign for, Christian CND, claim for, commitment to, Common Market, Employment Secretary, government('s) plan, to hold (...) meeting(s), in the ... (num) years, instead of, to join in, Labour candidate, most of, New Zealand, next month, no longer, to pay for, to protest at, spokesman for, such a (n), to take (...) decision, the people of (g n), this morning, to turn up*, United Nations, up to (f), Upper Heyford, US base, week(s) ago 14

261-293 to back up*, basis for, to be determined to (v), to be for, to be necessary, (to be) organised by, to concentrate on, day(s) of action, engineering worker(s), for ... weeks, general strike, to get rid of, to go up*, to hand over*, in line with, in response to, in the Commons, Labour councillor, last (m), national officer, new technology, public sector, referring to, to seek to (v), to spell out*, support from, the same (n), this weekend, transport workers, to urge ... to (v), to vote for, to work 13

with, to write to

294-327 a lot of, a (the) majority of, about (f) (n), 12
Austin Rover, to ban ... from, to be available, British Telecom, close to, Communist candidate, compulsory redundancies, death squad, Defence Minister, to give ... support to, to go ahead*, in the area, in particular, in the face of, letter to, miners' president, over the (adj) ... years, pay rise, pit closure, preparation(s) for, to press for, pressure on (upon), public employees, rather than, Reagan administration, together with, to transfer ... to, to turn ... into, to vote to (v), working class, working week

328-363 aim is to (v), to allow ... to (v), to appeal to, 11
to be among, (to be) faced with, to be joined by, (to be) led by, to bring in*, Cabinet minister, to charge ... with, democratic rights, to depend on, to do (...) work, to draw up*, Energy Secretary, fight against, Foreign Office, House of Commons, industrial relations, to insist on, lack of, near (g n), news conference, not only, nuclear-free zone, to refer to, to save (...) jobs, Scottish miners, Scottish TUC, to set out*, to speak at, to speak to, to stand for, talks on, task force, West German (n)

364-421 action(s) by, to add ... to, Anti-Apartheid Move- 10
ment, as many (much) as, at a (the) rally, at the (...) moment, to be on (n), (to be) ready to (v), change(s) in, city council, CND group, to come in*, to come out*, to come to power, to condemn ... as, Council of Churches, County Hall, discussion(s) on, district secretary, to do nothing, to do so, East London, economic policy, for ... hours, for ... months, to force ... to (v), Foreign Secretary, to give up*, to go into, to go through, House of Representatives, to impose ... on, in addition to, in the interest(s) of, inquiry into, involvement in, main gate(s), to make ... clear, next year, power workers, protest(s) against, to put forward*, resistance to, response to, Royal Ulster Constabulary, since (y), social security, support group, to take up*, talks between, teacher('s)

union, union official, United Democratic Front, up to, to use ... to (v), to wait for, to withdraw ... from, working people

422-478 answer(s) to, appeal(s) for, to arrive at, as a whole, 9
at home, at (...) risk, at the time, to be certain, (to be) designed to (v), to be important, (to be) sent to, (to be) supported by, to blame ... for, branch secretary, to coincide with, to come into, compared with, contact(s) with, coordinating committee, delegate conference, East Midlands, executive member, Falklands war, far from, to go ahead* with, Greenham women, to have (...) job(s), in (pron) case, in Parliament, in the (this) election, in the wake of, incomes policy, it is ... who, job centre, to lay off*, local council, London borough, low pay, National Association of, national organiser, nuclear base, on the eve of, to pass a (the) resolution, people's march, Pershing-(f) missiles, public opinion, to put up*, to rely on, right(s) to (v), South Yorkshire, to take away*, to tell a (the) meeting, to turn down*, visit to (g n), Warsaw Pact, West London, yesterday morning

479-558 anti-union law, to apply for, as soon as, at the 8
beginning of, at (the) ... gate(s), at work, to be forced to (v), (to be) hit by, to be prepared to (v), (to be) used as, (to be) used by, to become more (adj), to belong to, biggest union, Bishop of (g n), black township, to break the (...) strike, to bring ... into, to build up*, by (m), by (y), to call off*, Cammel Laird, to carry on*, chemical weapons, city centre, Civil and Public Services Association, to close down*, conference('s) decision, to create (...) jobs, to criticise ... for, to deprive ... of, deputy leader, election campaign, Foreign Ministry, GLC leader, (f)-hour week, hundreds of thousands, in (the) court, in the event of, in (the) future, in the region, inflation rate, to join ... in, large number(s) of, links with, London Transport, to look for, Lord Justice, to make (...) statement(s), mass unemployment, to meet in (g n), mining community, national

committee, nerve gas, on the dole, other countries, Palestine Liberation Organisation, parliamentary candidate, peace group(s), to pick up , political activity, pressure group, to reach (...) agreement, to regard ... as, regional council, report(s) on, to rise by (f), school meal(s), to set off , Socialist Party, special conference, to talk to, tens of thousands of, threat to, Trafalgar Square, to vote against, to walk out , washing-up time, withdrawal from
559-685 action group, African National Congress, to agree 7
to (v), agreement on, aircraft carrier, to amount to, annual meeting, apart from, to apply to, as far as... is concerned, at a time when, at ... meeting, at the start of, to be (...) affected by, (to be) critical of, to be due, to be (...) essential, to be followed by, to be opposed by, (to be) paid for, (to be) sacked for, (to be) threatened with, to be used to (v), black people, black section, black workers, to break up , British army, British Rail, campaign against, challenge to, to claim that, collective bargaining, council meeting, to cut ... by, defence ministry, to demand that, detention orders, discussions with, district council, to do (...) job(s), to draw attention to, during the (...) dispute, engineering section, Falkland Islands, for the ... time, General, Municipal, Boilermakers and Allied Trades Union, to give ... backing to, to go out , to have (...) effect on (upon), to hold (...) rally(ies), Home Secretary, (f)-hour strike, in accordance with, in addition, in (an) attempt to (v), in fact, in the country, in (pron) view, in (m) (y), to join the strike, kind of, Labour group, Labour('s) policy, later this (t), to launch a (the) campaign, to leave (g n), living standards, to make... (v), to make clear, to mark the ... anniversary, May Day, to meet with, Militant Tendency, military dictatorship, miners' leader, mining industry, national conference, national demonstration, negotiating committee, nuclear bomber base, on the ... day, peace initiative, perimeter fence, to play a (...) part in, political prisoner(s), proposal(s) for, to provide ... with,

public ownership, to pull out , to pull out of, rail union, reason for, to reply to, Royal Navy, to rule out , sanctions against, Scottish secretary, to shout at, to sit down , social workers, South London, Southern Africa, to speak for, to speak out , statement by, struggle for, to supply ... with, to tell a (the) conference, the same (...) as, Third World, this (d), train drivers, TUC guidelines, TUC leaders, under (...) pressure, unemployed centre, union('s) conference, unions involved, until (m) (f), US officials, to use ... as, wages councils, water workers, whether or not, to work out , world war

С частотой 6 зарегистрировано 176 словосочетаний, с частотой 5 - 231, с частотой 4 - 405, с частотой 3 - 798, с частотой 2 - 1916 словосочетаний.

A FREQUENCY LIST OF SET EXPRESSIONS OF ENGLISH NEWSPAPER TEXTS

I. Matsukova

S u m m a r y

On the basis of articles dealing with home policy and foreign affairs from 100 issues of the newspaper "Morning Star" for the period of 1983-1985, a frequency list of set expressions has been compiled. Complete articles were used, totalling 200 000 words (35 470 word combinations) on the token level. The degrees of stability of different word combinations were taken into account. The list of the set expressions with absolute frequencies of 7 and more (685 expressions) is given with the indication of the rank and the frequency.

ФОРМА ПРЕДСТАВЛЕНИЯ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

В.В.Нешитой

В статье обсуждается традиционная форма представления ранговых распределений слов в виде "прямой Ципфа" в логарифмических координатах, отмечаются ее недостатки и предлагается новая форма представления такого рода распределений. Она значительно расширяет возможности исследования статистических ранговых распределений. Рассматриваются причины, порождающие иллюзию справедливости закона Ципфа. Для описания статистических ранговых распределений однородных единиц предлагаются обобщенные распределения, одним из частных случаев которых является закон Ципфа.

Традиционная форма представления ранговых распределений. Наиболее широкое применение ранговые распределения имеют в лингвистике и информатике. Если все разные слова, которые употребились в некотором тексте (выборке), упорядочить по невозрастанию абсолютных или относительных частот и каждому слову приписать порядковый номер (ранг), то зависимость между относительной частотой слова p_z и его рангом z в первом приближении может быть описана законом Дж.К.Ципфа (Zipf G.K., 1940):

$$p_z = \frac{\kappa}{z^{\gamma}}, \qquad (1)$$

где по данным автора этого закона $\kappa \approx 0{,}1$, $\gamma \approx 1$.

В информатике этим законом описывается ранговое распределение журналов по числу опубликованных в них статей на определенную тему. Действительно, если p_z обозначает долю статей из общего их числа (по данной тематике), опубликованных в журнале с рангом z, то накопленная доля статей в z первых журналах на основании (I) будет равна (при γ =I)

$$F(z) = \sum_{i=1}^{z} p_i = \sum_{i=1}^{z} \frac{\kappa}{i} \approx \kappa(\ln z + C). \qquad (2)$$

Выражение (2) по форме совпадает с формулировкой закона рассеяния публикаций С.Брэдфорда в ранговой интегральной форме (Bradford, 1948) $X(z) = a + b\log z$, где $X(z)$ -накопленное число статей в z первых журналах; a, b - параметры.

Выражение (I) при логарифмировании преобразуется в прямую

$$\ln p_z = \ln k - \gamma \ln z , \qquad (3)$$

которая утвердилась как одна из основных форм представления ранговых распределений. Однако график зависимости $\ln p_z$ от $\ln z$, построенный по опытным данным, близок к прямой лишь в средней части. Наличие кривизны в областях низких и высоких рангов принуждает исследователей либо вводить поправки в модель Ципфа-Брэдфорда, либо искать новые, более подходящие модели.

Однако несмотря на многочисленные попытки усовершенствовать модель Ципфа-Брэдфорда, в настоящее время не существует единого уравнения (распределения), которое с достаточной точностью описывало бы все многообразие статистических ранговых распределений. Причин неудовлетворительных аппроксимаций подобных распределений много. Рассмотрим наиболее существенные из них.

Первая причина: неудачно выбрана форма представления статистических ранговых распределений (в виде прямой (3)), не отражающая в должной мере их характерных особенностей. "Уже верное отражение природы,- писал Ф.Энгельс, - дело трудное, продукт длительной истории опыта" [К.Маркс, Ф.Энгельс, с.639] .

Принятая форма представления ранговых распределений несет слишком мало информации о статистическом распределении. На таком графике колебания частот мало заметны, поскольку последние изображены в логарифмическом масштабе. Кроме того, такое преобразование кривой распределения не имеет вероятного смысла.

Новая форма представления ранговых распределений. В связи с вышесказанным нам представляется целесообразным перейти к другой форме представления ранговых распределений, а именно $rp_z = f(\ln z)$. По оси ординат будем откладывать произведение ранга на относительную частоту слова с данным рангом, а по оси абсцисс - натуральный логарифм ранга. При построении такого графика будем использовать координаты середин ступенек статистического дискретного распределения (см.рис.I), т.е. те точки дискретного распределения, через которые проходит выравни-

вающая непрерывная кривая распределения. Этими точками на рис.I являются:

z	0,5	I,5	2,5	...	6	I0	I8
m_z	I0	7	5	...	2,5	I,5	0,5

График зависимости $zp_z = f(\ln z)$ имеет принципиальные преимущества перед традиционной формой представления ранговых распределений. Во-первых, он представляет собой кривую распределения, что легко доказать следующим образом. Пусть $p(t)$ - невозрастающая плотность распределения, аппроксимирующая относительные частоты p_z. Тогда t будет соответствовать рангу z. Следовательно,

$$\int_{-\infty}^{\infty} tp(t)d\ln t = \int_{0}^{\infty} tp(t)\frac{dt}{t} = \int_{0}^{\infty} p(t)dt = 1,$$

т.е. площадь под кривой $tp(t) = f(\ln t)$ равна единице (условие, обязательное для кривых распределения). Во-вторых, на такой кривой видны колебания самих частот (по оси ординат), а не их логарифмов. В-третьих, статистические ранговые распределения однородных случайных величин имеют одновершинную кривую распределения. Это дает возможность устанавливать однородность или неоднородность ранговых распределений, выделять неоднородную часть, оценивать минимально необходимый объем выборки для установления типа выравнивающей кривой и нахождения оценок параметров и т.д. [Нешитой В.В., I984] .

Вторая причина неудовлетворительных аппроксимаций статистических ранговых распределений заключается в попытке описать одним уравнением распределения неоднородных случайных величин. Частотный словарь, как правило, представляет собой неоднородную совокупность элементов. Действительно, роль полнозначных и служебных слов в тексте различна, при этом служебные слова употребляются значительно чаще и поэтому находятся главным образом в начале частотного списка. Это обстоятельство позволяет весьма просто выделять неоднородную часть рангового распределения с помощью графика зависимости $zp_z = f(\ln z)$.

На рис.2 представлена кривая рангового распределения слов в романе Л.Н.Толстого "Война и мир" ["Частотный словарь романа Л.Н.Толстого "Война и мир", с.357-367] . Объем текста сос-

тавил x=409407 словоупотреблений, объем словаря y=I95I9 лексем. Кривая распределения имеет неправильную форму, поскольку изображает распределение неоднородных элементов, при этом границей неоднородной части является наинизшая точка впадины на кривой распределения с абсциссой $\ln z_0$=4,25 (z_0=70). Удалим из частотного словаря первые 70 слов, на которые приходится x_0=I80844 словоупотребления текста, а оставшимся словам припишем новые ранги $z'=z-z_0$. Тогда количество оставшихся в словаре разных слов будет $y'=y-z_0$=I9449, а их общая частота $x'=x-x_0$=228563 словоупотребления. Если теперь построить график зависимости $z'p_{z'}=f(\ln z')$, то получим весьма плавную одновершинную кривую, которая отличается закономерным характером возрастания и убывания (рис.3). Такая же одновершинная кривая получается в случае распределения любых однородных лингвистических единиц (терминов, дескрипторов и т.д.). Для описания такого рода статистических распределений существует объективная возможность нахождения выравнивающей кривой распределения, в то время как для описания неоднородных случайных величин (рис.2) такой возможности не существует.

Третья причина неудовлетворительных аппроксимаций статистических распределений заключается в том, что исследовались выборки недостаточного объема. Из рис.3 видно, что статистический закон распределения однородных единиц проявляет себя в полной мере лишь при том условии, если крайняя справа точка близка к горизонтальной оси. Только при этом условии можно подбирать выравнивающее непрерывное распределение.

Ордината крайней справа точки по построению равна

$$zp_z = \frac{y\bar{m}_z}{x} = \frac{y}{2x} ,$$

поскольку $z_{max}=y$, $m_{z=y}=1$, $m_{y+1}=0$, $\bar{m}_{z=y}=(1+0)/2=1/2$. Объем выборки можно считать достаточным, если отношение ординаты крайней справа точки к наибольшей ординате $(zp_z)_c$ кривой распределения не превышает наперед заданного числа δ

$$\frac{zp_z}{(zp_z)_c} = \frac{y}{2x(zp_z)_c} \leq \delta , \qquad (4)$$

где $0 < \delta < 0,3$. Чем меньше δ, тем точнее могут быть оцене-

ны параметры выравнивающего распределения.

Статистическую кривую распределения $zp_2 = f(\ln z)$ можно использовать для расчета необходимого объема выборки при построении достоверного словаря заданного объема. Пусть достоверная частота m_2 и наибольший ранг z, т.е. объем словаря, заданы. Тогда из равенства

$$zp_2 = z\,\frac{m_2}{x}$$

находим необходимый объем выборки x

$$x = \frac{z\,m_2}{zp_2}\,. \qquad (5)$$

Произведение zp_2, входящее в формулу (5), берется из графика зависимости $zp_2 = f(\ln z)$.

Из последней формулы и рис.2 видно, что между объемом словаря z (при постоянной достоверной частоте m_2) и необходимым объемом выборки x нет линейной зависимости: объем выборки растет значительно быстрее объема достоверного словаря, поскольку произведение zp_2 с ростом z уменьшается (см.правую ветвь кривой распределения на рис.2).

Последние две причины неудовлетворительных аппроксимаций ранговых распределений проливают свет на происхождение закона Ципфа. Иллюзия справедливости этого закона порождается двумя факторами: во-первых, неоднородностью лексического состава частотного словаря, которая влияет на форму начала кривой распределения, поскольку мы имеем композицию по крайней мере двух законов распределения слов (служебных и полнозначных); во-вторых, ограниченностью объема выборки, из-за чего последняя справа точка не успевает приблизиться к горизонтальной оси и может оказаться вблизи воображаемой прямой Ципфа $zp_2 = K$.

Обратимся к фактам. На рис.4 (кривая I) изображено статистическое распределение по данным "Частотного словаря немецкого подъязыка хирургии" [Яблонская Н.Н., 1978] ($x = 200000$, $y = 41041$). Здесь объем выборки весьма ограничен, но композиция двух законов распределения налицо, о чем говорилось выше. Однако ни о какой (даже воображаемой) прямой Ципфа ($zp_2 = K$) здесь не может быть речи.

На том же рис.4 изображена статистическая кривая 2 рангового распределения слов в литературном тексте Дж.Джойса "Улисс" [Zipf G.K ,1949] (x=260430, y=29899). На этом примере Ципф иллюстрировал свой закон прямолинейной зависимости между частотой и рангом (в логарифмических координатах), при этом угловой коэффициент прямой (3) по абсолютной величине равен единице. Однако в системе координат ($\ln z$; zp_z) четко видно, что статистическая кривая распределения на рис.4 не может быть аппроксимирована прямой $zp_z = K$ (т.е. законом Ципфа).

Итак, приведенных на рис.2-4 статистических распределений достаточно, чтобы сделать однозначный вывод: закон Ципфа не может быть использован для описания даже в первом приближении ранговых распределений слов, поскольку его не подтверждает ПРАКТИКА как КРИТЕРИЙ ИСТИНЫ.

Обобщенные распределения. Для описания распределений однородных случайных величин автором разработаны системы непрерывных распределений, заданные обобщенными плотностями. На эмпирическом материале было установлено, что ранговые распределения периодических изданий по числу помещенных в них статей на заданную тему могут быть описаны обобщенным распределением вида

$$p(t) = N t^{\gamma-1}\left(1-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}, \qquad (6)$$

где α, β, γ, u - параметры распределения; N - нормирующий множитель; t -случайная величина, которая соответствует рангу z статистического распределения; $p(t)$ - непрерывная плотность, аппроксимирующая относительные частоты p_z.

Этим же законом (6) хорошо описываются статистические распределения: слов по длине (в словаре), фраз по количеству словоупотреблений, словосочетаний по длине, терминов по длине и др., а также ранговое распределение научных сотрудников по продуктивности.

Статистические ранговые распределения однородных лингвистических единиц (лексем, словоформ, терминов, дескрипторов и др.)

хорошо описываются обобщенным логарифмическим распределением вида

$$p(Y) = \frac{N(\ln Y)^{\gamma-1}}{Y}\left(1-\alpha u \ln^{\beta} Y\right)^{\frac{1}{u}-1}, \qquad (7)$$

где $Y=y+1$; величина y соответствует рангу z статистического распределения.

Выравнивающее распределение для примера на рис.3 имеет параметры: u =0,30; β =2,25105; γ =3,60168; αu=1/219,656; нормирующий множитель N=1/283,430. С учетом оценок параметров плотность распределения (7) можно представить в виде

$$Yp(Y) = \frac{(\ln Y)^{2,60168}}{283,430}\left[1-\frac{(\ln Y)^{2,25105}}{219,655}\right]^{2,33333}$$

$$1 < Y < 58226.$$

Метод нахождения оценок параметров непрерывных распределений изложен в работе [Нешитой В.В.,1985].

Место закона Ципфа в системе непрерывных распределений. Отметим, что обобщенное распределение (7) при $\alpha>0, \beta, \gamma, u$ =1 переходит в закон Ципфа: $p(Y)=\alpha/Y\ (1<Y<e^{1/\alpha})$. Следовательно, закон Ципфа относится к семейству логарифмических распределений (7), является весьма частным его случаем и может быть использован для описания распределений однородных случайных величин (так же как и обобщенная плотность (7)). На практике же его пытаются применить для описания распределений неоднородных случайных величин (без всякого на то основания).

Далее, при $u\to 0$, β,γ=1, $\alpha>0$ из (7) имеем фактически формулу Б.Мандельброта

$$p(Y) = \frac{N}{Ye^{\alpha \ln Y}} = \frac{N}{Y^{1+\alpha}} = \frac{N}{(y+1)^{1+\alpha}}$$

а при $u\to 0$, β =2, γ =1, $\alpha>0$ - "четвертое приближение закона Ципфа" по П.М.Алексееву (логнормальный закон)

$$p(Y) = \frac{N}{Ye^{\alpha \ln^2 Y}} = \frac{N}{Y^{1+\alpha \ln Y}}$$

Если начало отсчета значений случайной величины Y поместить в центр распределения $Y_1 = M[\ln Y]$, то последняя формула примет вид, более близкий к модели П.М.Алексеева [Алексеев П.М., 1978]

$$p(Y) = \frac{N'}{Y^{1-2\alpha Y_1 + \alpha \ln Y}}.$$

При описании ранговых распределений однородных лингвистических единиц реализуется, как правило, четырехпараметрическая модель (7), а в некоторых, весьма редких случаях, логнормальный закон или четырехпараметрическая модель (6).

Заключение. Введение новой формы представления ранговых распределений позволило не только раскрыть причины, порождающие иллюзию справедливости закона Ципфа, но также решить некоторые важные практические задачи: дать графический метод выделения неоднородной части ранговых распределений z, метод оценки необходимого объема выборки для установления типа выравнивающего распределения и нахождения оценок параметров, а также метод расчета необходимого объема выборки x для построения частотного словаря заданного объема z с заданной наименьшей частотой слов m_z.

Показано, что законы Ципфа, Мандельброта и "четвертое приближение закона Ципфа" по П.М.Алексееву являются частными случаями обобщенного логарифмического распределения (7). Последняя модель хорошо описывает статистические ранговые распределения однородных лингвистических единиц.

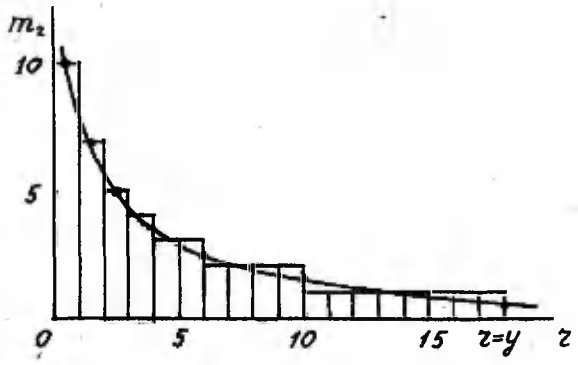

Рис.I. Выравнивание дискретного рангового распределения непрерывным.

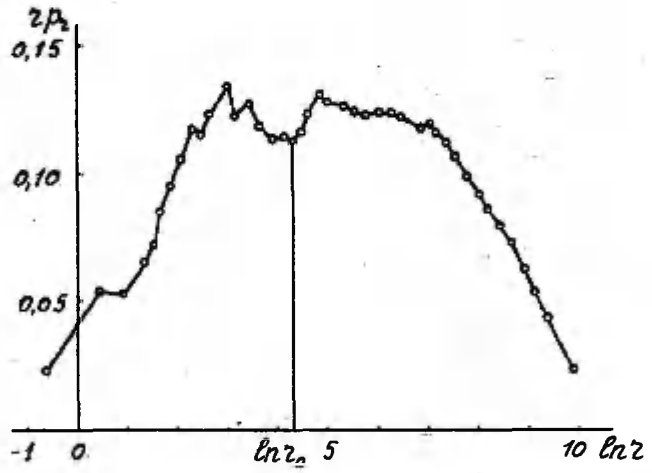

Рис.2. Ранговое распределение слов в романе Л.Н.Толстого "Война и мир"

17*

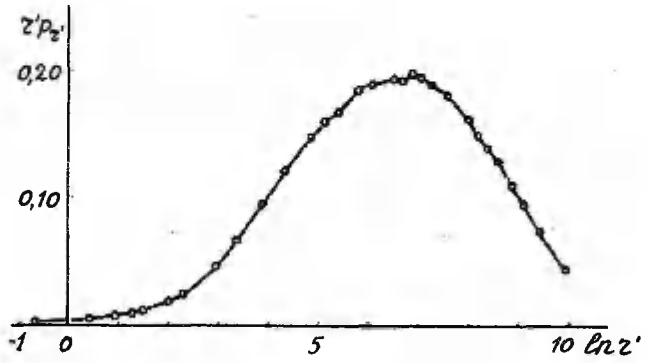

Рис.3. Ранговое распределение слов в романе Л.Н.Толстого "Война и мир" (без первых r_0=70 лексем)

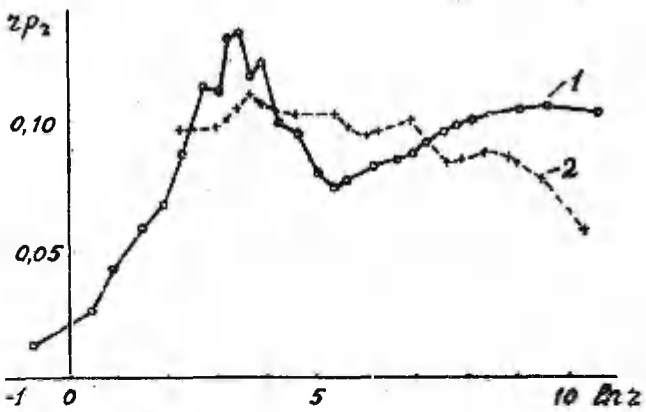

Рис.4. Ранговое распределение слов:

I - в немецком подъязыке хирургии;
2 - в тексте Дж.Джойса "Улисс".

Л И Т Е Р А Т У Р А

К.Маркс, Ф.Энгельс. Соч. 2-е изд., т.20.

Алексеев П.М. О нелинейных формулировках закона Ципфа.- Вопросы кибернетики, вып.41, М. Л., 1978, с.53-65.

Нешитой В.В. Исследование ранговых распределений. - НТИ.Сер.2, 1985, №2, с.16-20.

Нешитой В.В. Критерий однородности лексического состава частотного словаря. - Веснік Беларускага дзяржаўнага ун-та імя У.І.Леніна. Сер.4. 1984, №1, с.54-56.

Частотный словарь романа Л.Н.Толстого "Война и мир". Тула, 1978, Тульский гос.пед.инст.им.Л.Н.Толстого.

Яблонская Н.Н. Частотный словарь немецкого подъязыка хирургии.- В сб.:Структурная и прикладная лингвистика. Вып.I, Л.: изд-во ЛГУ, 1978.

Bradford S.C. Documentation. - London, 1948.

Zipf G.K. Human Behavior and the principle of least effort. - Cambridge, 1949.

ABOUT THE FORM OF REPRESENTING RANK DISTRIBUTIONS

V.V. Nešitoj

S u m m a r y

The article points out the principal demerits of the tradition of representing rank distributions in logarithmic coordinates (the Zipf straight line). A new method is put forward, with the coordinates ($\ln r$, rp_r), where r is the rank of an event and p_r - its relative frequency. It is demonstrated that in these coordinates the distribution curves of homogeneous random quantities have one peak and a regular character of increase and decrease. Generalized distributions are presented for the description of such curves.

ПОЛИСЕМИЯ: СИСТЕМНО-КВАНТИТАТИВНЫЕ АСПЕКТЫ

А.А.Поликарпов

I. Полисемия знаковых единиц человеческого языка продолжает оставаться характеристикой, до сих пор недооцененной и теоретически, и практически (в практике преподавания языков, в разработках автоматизированных систем обработки текстов и т.п.). Эта недооценка связана, прежде всего, с непроясненностью в общем языкознании основных принципов языковой коммуникации, что, в свою очередь, связано с длительным господством в науке о языке статично-ограничительных представлений о её предмете. Только ситуативно-коммуникативный подход к языку заставляет обратить внимание на моменты вероятностного намекания и вытекающей отсюда полисемии знаков в языковом общении как принципиально определяющие его сущность (Поликарпов А.А., 1976; 1979).

Недостаточно к настоящему времени проработан и вопрос о реальной распространенности этого явления в разных языках в целом, а также в отдельных их подсистемах, стилях, жанрах, что и не позволяло разработчикам различных автоматизированных информационных систем, имеющих дело на входе с естественно-языковыми текстами, оценить реальную значимость этого явления и предпринять усилия, необходимые для машинного моделирования процессов разрешения многозначности[+]. Например, оценки распространенности в языках мира явления полисемии совсем недавно колебались в чрезвычайно широких пределах – от признания многозначности слов нетипичным для языка явлением (Милевский Т., 1963) до оценки 80% слов в языке в качестве многозначных (Будагов Р.А., 1958). Только исследования последних лет (Папп Ф., 1967; 1969; Вишнякова С.М., 1976; Поликарпов А.А., 1976; 1979; 1981; Крылов Ю.К., Якубовская М.Д. 1977; Андрукович П.Ф., Королев Э.И., 1977; Гулдава Ю.А., 1979; Борода М.Г., Поликарпов А.А., 1984; Денисов П.Н., 1984; Поликарпов А.А., Бушуева О.В., 1985) позволили собрать достаточно представительные и надежные экспериментальные данные о представленности лексической полисемии в словаре и те-

---

[+] Можно отметить лишь две наиболее серьезные попытки учесть эту проблему и разработать средства её решения в практике построения АИС – (Kelly E.F. and Stone P.J., 1975; Марчук Ю.Н., 1976).

кстах разных языков, о соотношении и суммарных характеристиках полисемической и омонимической многозначности слов, о соотношении полисемии и омонимии с частотными, контекстуальными и некоторыми другими характеристиками употребления и строения слов[+], о зависимости характеристик многозначности от типа языка, о динамике реализации полисемического потенциала слов в текстовых массивах в связи с ростом их объёма и т.д. В частности, теперь не голословно, а на основе объёмных эмпирических подсчётов можно утверждать, что если в словаре языка в целом доля многозначных слов, как правило, значительно меньше, чем доля однозначных[++], то употребления многозначных слов в тексте составляют подавляющую его часть - от 80% до 99% его объёма в зависимости от типа языка и тематики текста[+++]. То есть, для правильного вычленения смысла из любого текста необходимо практически к каждому очередному словоупотреблению прилагать процедуру разрешения многозначности, вероятностного определения того значения из числа присущих слову, которое реализовано в данном случае. Это соотношение наглядно иллюстрируется на материале русского языка. Так, если, например, средняя лексическая полисемия в большом академическом толковом словаре современного русского языка (Словарь, 1948-1965), по нашим данным, равна 1,7 значений на одно слово, то в русских литературно-художественных текстах каждое словоупотребление обладает в среднем 6,6 потенциальными значениями. Соответствующие данные по английскому языку для большого толкового словаря американского варианта английского языка (Webster's, 1961) - 2,3 значения на одно словарное слово, а для английских текстов, параллельных к обследованным русским текстам, - 9,5 значений на одно текстовое словоупотребление.

2. Еще совсем недавно многозначность лексических и других знаковых единиц считали, если уж не неким ущербным, то,

---

[+]Пионером исследования этого круга вопросов был Дж.К. Ципф (Zipf G.K., 1945; 1949).

[++]Исключение составляют такие языки, как китайский, при их анализе на уровне однослогов. См. статью Н.В.Обуховой (1986).

[+++]Противопоставленными здесь являются языки ярко синтетические (типа русского) и ярко аналитические (типа китайского).

во всяком случае, вынужденным свойством человеческого языка (Виноградов В.В., 1947, с.10). На самом деле, многозначность является специфическим выражением глубинной сущности языка, языкового общения. В ходе общения то, что намеревается сообщить отправитель , передаётся получателю и воспринимается им отнюдь не в буквальном и неизменном виде, как это может представляться обыденному сознанию. "Мы говорим только необходимыми намеками. Раз они вызывают в слушателе нужную нам мысль, цель достигается, и говорить иначе было бы безрассудной расточительностью" (Поливанов Е.Д., 1968, с.296).

Принцип общения с помощью намеков отражает действительную общесемиотическую суть общения. Знаки – это физические посредники между общающимися, несущие намекательный минимум информации в своей структуре, похожие своей формой на намекаемое содержание (иконические знаки) или возбуждающие его по условно-наработанной связи в памяти общающихся (условные, или произвольные знаки). Отработанный, социально значимый намекательный информационный минимум – значение знака (" общее", "собственно", "ядерное" значение во многих современных концепциях). Типовые интерпретации значения в тех или иных типовых контекстах – его узуальные смыслы (или "значения", как чаще принято говорить). Любой знак в одном из узуальных смыслов – семантический вариант.

Интерпретации знака могут быть самой разной глубины и индивидуальности. Язык – социально обработанный и относительно идентичный у всех членов данного коллектива системный набор знаков в их типовых семантических вариантах, через которые можно намекать на самые индивидуальные смыслы.

Надо полагать, что чем более бедным является содержание общего значения знака, тем более широкой оказывается возможная сфера его типовых интерпретаций, то есть тем более полисемичным оказывается в итоге данный знак. При каждом употреблении знака (например, слова), являющегося более многозначным, чем другой, можно утверждать, что при прочих равных условиях он будет с большим трудом правильно воспринят, чем менее многозначный знак. Т.е. более многозначный знак более семантически неопределенен, сложен, чем менее многозначный[+].

[+]Общее значение как некоторое пересечение признаков узуальных смыслов знака может быть выделено не всегда, многозначность же является универсальной характеристикой широты его отнесенности.

Однако эффективность общения на языках с высокой степенью полисемии их лексических единиц, то есть способность с достаточно высокой точностью указать-намекнуть с их помощью на какое-то смысловое содержание, отнюдь не ниже, чем на языках с лексикой менее многозначной, то есть в языках более синтетического устройства. Дело в том, что то или иное "полученное" содержание является результатом совокупного действия целой серии знаков-намеков. Каждый последующий намек уточняет, конкретизирует ту смысловую картину, которая к этому моменту сложилась. Принципиальная разница в том, что для достижения одной и той же цели носителями типологически отличных языков будет применяться разное число знаков данного уровня: чем более многозначными в среднем являются лексические единицы данного языка, тем относительно большее их число необходимо употреблять в сообщениях на нем.

Принципиально важно отметить то, что при сопоставлении параллельных текстов языков, заметно отличных друг от друга по общей величине многозначности их единиц, обнаруживается относительное удлинение текста, пропорциональное не линейной величине средней общеязыковой многозначности его текстовых единиц ($_m$), а логарифму этой многозначности. Логарифмическая мера количества значений у слова (или среднего количества значений для всех слов данного текста) и есть $H_{сем.}$ - мера его семантической неопределенности ($H_{сем.} = \log m$), отражающая при прочих равных условиях меру его сложности, трудности выбора для получателя нужного значения из числа всех значений, присущих данному слову (или среднюю меру семантической неопределенности каждого из слов в данном тексте как интегральную характеристику). Введение логарифмического масштаба для перехода от многозначности к семантической неопределенности связано с тем, что именно логарифмическая мера отражает существо ситуации выбора[+].

Введенная нами величина семантической неопределенности позволяет онтологически естественным образом членить состав

[+]Отметим, что более точное значение $H_{сем.}$ может быть получено с использованием данных о вероятностях (относительных частотах) употребления в текстах слов в их разных значениях. Однако данные этого рода менее доступны, чем информация об общем количестве значений у слова. Кроме того, удовлетворительная сопоставимость характеристики $H_{сем.}$ у разных слов достигается и при использовании приведенной нами формулы.

словаря того или иного текста или всего языка в целом на ряд слоев (взяв за основание логарифмов 2):

I) слова нулевой степени семантической неопределенности: с одним значением (log I=0);

2) слова первой степени семантической неопределенности: с двумя значениями (log 2=I);

3) слова второй степени: с 3–4 значениями (log 4=2);

4) слова третьей степени: с 5–8 значениями (log8=3) и т.п.

Введем также величину семантической специфичности S, являющейся обратной по отношению к величине семантической неопределенности ($S = I/H_{сем.}$). Возрастание многозначности слова ведет к повышению его семантической неопределенности, или к понижению его семантической специфичности.

Если учесть, что относительное уменьшение длины (W) словоупотреблений в текстах языка, развивающегося в сторону аналитизма, или относительное уменьшение длины слов при переходе от рассмотрения языка, относительно синтетического, к рассмотрению языка, относительно более аналитического, осуществляется в той же степени, в какой понижается их средняя семантическая специфичность, то величина семантической специфичности, приходящаяся в среднем на ту или иную единицу различения (фонему, дифференциальный признак), должна быть в разных языках (в том числе, и одном языке, но в разные периоды его типологического развития) величиной, относительно одинаковой, константной. Что и обнаруживается на английском, русском, французском, немецком материале. Эта универсальная величина D ($D=S/W$) определена опытным путем и равняется ок. 0,07 усл.ед. сем. специфичности на I фонему (Поликарпов, I98I).

Обнаружение подобной квантитативной универсалии имеет и прикладное значение, поскольку знание ее позволяет рассчитывать те или иные неизвестные значения ряда языковых параметров по другим известным параметрам. Например, по той или иной текстовой выборке очень легко определяется средняя длина слов (или других знаковых единиц – морфем, словосочетаний и т.п.), что позволяет, далее, вычислить на основе указанных соотношений и среднюю величину многозначности в данном тексте. Подобные расчеты могут быть проведены и по представительной выборке из того или иного сводного общеязыкового словаря.

Существование квантитативной универсалии удельной семантической специфичности тесно связано с теми оптимизацион-

ными отношениями, которые осуществляются в системе "полисемия – длина – частота знаков". Оптимизация заключается в согласовании этих характеристик, в достижении баланса между экономией и достаточной различительной силой сообщений, производимых на том или ином естественном языке. При любом сочетании значений каждого из этих параметров достигается постоянное, оптимальное семантико-различительное отношение (отражающееся в указанной квантитативной универсалии).

3. В чем же источник существенной (но взаимно согласованной) вариативности каждого из трех параметров – полисемии, длины, частоты? Он располагается вне данного треугольника. Существо отношений в нем может быть понято только в том случае, если мы осознаем, что и полисемия, и частота, и длина знаков являются хотя и связанными друг с другом, но зависящими в своих значениях от четвертого параметра – размера знакового набора. При заданном и относительно фиксированном социальной практикой наборе смыслов, которые нужно обслуживать данному языку, та функциональная нагрузка, которая падает в среднем на каждый знак и некоторым специфическим образом распределяется по всем знакам набора, зависит от длины (размера) этого набора, является функцией от числа разных единиц в нем. Чем более богатым является знаковый набор, тем меньше в среднем приходится узуальных смыслов на каждый из знаков и тем, видимо, меньший удельный вес в нем занимают знаки с большим количеством значений.

Аналогично этому – число знаковых единиц в наборе задает и необходимую длину знаков в нем – при удлинении знакового набора должно расти в нем число единиц большой длины.

Поскольку "наложение" знакового набора определенного размера на совокупное смысловое поле данного коллектива можно, видимо, рассматривать как случайный процесс, то и распределение значений по знакам является случайным. Неслучайным должно являться лишь то, в большей или меньшей степени, в общем, окажутся полисемически нагружены знаковые единицы в таких случаях в зависимости от размера знакового набора. Это должно отображаться и в значимых различиях общего характера таких случайных распределений.

Эти соображения и предопределяют интерес к рассмотрению полисемических распределений, т.е. распределений разной полисемии по группам слов разного объема в словарях языков разного типа и в словарях разного типового объема одного языка.

4. Полисемические распределения как особый и весьма

важный вид распределений привлекли к себе внимание совсем недавно. Сначала Ф.Папп (1967) на материале толкового словаря венгерского языка обнаружил существование закономерности в эмпирическом распределении слов по числу значений, аппроксимируемом выражением $y = V/2^x$ (I), где y – доля в словаре слов, имеющих x значений, V – объем словаря. Но уже в этой работе Ф.Папп отмечал, что русский материал описать таким образом нельзя.

В работе (1979) Ю.А.Тулдава отмечал, что предложенная Ф.Паппом формула является частным случаем более общего экспоненциального распределения типа $p = ae^{-bm^c}$ (2), где p – вероятность слов с данным количеством значений, m – количество значений, а и в – константы, величины которых зависят от рассматриваемого языка. На основе экспериментальных данных (по русскому, английскому и венгерскому языку) Ю.А.Тулдава установил, что вероятности слов (p) с хорошим приближением зависят экспоненциально от квадратного корня от количества значений слов. Функция в этом случае принимает вид:

$$p = ae^{-b\sqrt{m}} \qquad (3).$$

Интерпретируя $\sqrt{m}$ как единицу измерения семантического объема слова, Ю.А.Тулдава выделяет следующие естественные полисемические подклассы слов:

I) Нулевая степень полисемичности – слова с одним значением;

2) Первая степень полисемичности – слова с 2–4 значениями;

3) Вторая степень полисемичности – слова с 5–9 значениями;

4) Третья степень полисемичности – слова с I0–I6 значениями и т.п.

Как видно, этот способ полисемической квантификации словарного состава частично перекрещивается с предложенной нами квантитативной классификацией по степени семантической неопределенности слов. Однако и по существу, и по структуре эти характеристики (особенно при больших *m*) сильно расходятся.

Отметим, далее, что представление квантитативной структуры полисемии в виде модифицированного экспоненциального распределения является одной из возможных интерпретаций эмпирических фактов, полученных из словарей. Другая возможность связана с представлением этих фактов в виде распределения

гиперболического типа: $n_p = (\frac{A}{p+C} - c)^{\alpha}$ (4),

где n_p – объем группы слов с данным количеством значений; p – количество значений; A, C, α – константы распределения. Величина A зависит от типа (объема) рассматриваемого словаря (точнее – от типового общего количества значений слов в нем); α зависит от типа языка и демонстрирует темп нарастания количества слов с падением количества значений у них; C зависит от степени выпуклости распределения в билогарифмической системе координат, степени его однородности : чем более однородным (выпуклым) является распределение, тем больше величина этой поправки.

Аналогичная эмпирическая формула может быть предложена и для расчета зависимости полисемии от ранговых характеристик слов, упорядоченных в нисходящей по величине их полисемии последовательности:

$$p_i = \frac{B}{i^{\gamma} + K} - K \qquad (5),$$

где p_i – полисемия i -го слова в ранжированной по этой величине последовательности слов, i – наибольший порядковый номер из группы слов с одинаковой полисемией, γ, B, K – параметры.

5. Предложенные формулы характеризуют распределения гиперболического класса и весьма наглядно описывают связь хода, например, рангово-полисемического распределения как с типологическими характеристиками языка, который рассматривается, так и с типом словарей внутри одного языка. Связанный с типом языка параметр γ аналогично параметру γ в ципфовском частотном распределении в основном отвечает за угол наклона графической формы распределения в билогарифмической системе координат. Чем более язык аналитичен, тем больше места в полисемических распределениях, полученных из его словарей, занимают слова высокой полисемии, тем более значительной должна быть величина параметра γ. Например, по ряду толковых словарей относительно синтетического русского языка эта величина устойчиво равна 0,29 – 0,30, а по ряду толковых словарей аналитического английского языка – 0,33 – 0,35.

Параметр A отображает объемные характеристики словаря, его тип в классификации типов толковых словарей : при

+Предлагаемые формулы сконструированы и экспериментально проверены совместно с А.В.Маловым.

переходе от описания кратких словарей к средним и большим мы последовательно наблюдаем возрастание этой величины.

Особое место в этой формуле занимает параметр К. Именно он в наибольшей степени определяет кривизну распределения в принятой билогарифмической системе координат, действуя в определенной мере аналогично поправке В в формуле закона Ципфа -Мандельброта. Однако особенностью нашей формулы является то, что она включает поправки одинаковой величины для обеих ветвей гиперболы. Симметричность геометрии полисемического распределения - его характерное свойство, наблюдаемое на всех привлеченных к рассмотрению словарях и отличающее его от частотных распределений в тексте. В отличие от поправки Мандельброта в формуле закона Ципфа-Мандельброта наша поправка К действует не так локально, а на всем протяжении распределения (с более постепенным затуханием её действия при движении от одной ветви гиперболы к другой, чем в сопоставляемой формуле). Это свойство данного распределения реализуется за счет того, что под показатель γ (кот. существенно меньше единицы, т.е. он обладает уменьшающим действием) подводится не все выражение ($1+K$), как это имеет место в формуле закона Ципфа-Мандельброта, а лишь одно i .Это увеличивает удельный вес поправки K , делает ее действие более заметным на более протяженном ранговом интервале.

6. Сравнение сконструированной формулы (5) с ципфо-мандельбротовской не случайно. С самого начала поисков наиболее точного вида рангово-полисемической зависимости мы отправлялись от логарифмического масштаба рангов и полисемии, как от наиболее естественных для организации такой информационной системы, как язык. Логарифм полисемии выше уже был проинтерпретирован как онтологически значимая мера семантической неопределенности слова, мера его семантической сложности. Логарифм же рангов отражает порядковые отношения в словарном составе, его концентрическое расслоение на пропорционально уменьшающиеся по объему группы при переходе ко все более высоким величинам семантической неопределенности.

Сходство масштабов представления аргумента и функции в рангово-частотном и рангово-полисемическом распределениях и позволяет обнаружить некоторые существенные различия в их параметрах, различия в зависимостях, представленных в этих распределениях. Вместе с тем , это позволяет осознать то, что

вместе они являются представителями обширного семейства гиперболических распределений. Например, возможен вариант формул с комбинированием в них обоих типов поправок - мандельбротовской и нашей, а также с неодинаковыми величинами двух поправок С ($C_1 \neq C_2$) и т.п.

Собственно, последний упомянутый вариант и выводится Ю.К.Крыловым теоретически для рангово-частотных структур, исходя из общих вероятностно-комбинаторных соображений в работе (1982): $$y = \frac{A}{x + C_1} - C_2 \qquad (6),$$ где (в его обозначениях) у - вероятность слова, х - порядковый номер, А, C_1, C_2 - константы.

Исходя из тех же общих соображений, но несколько меняя условия вывода, можно теоретически получить и формулу, идентичную нашей. Это свидетельствует о высокой степени общности подхода Ю.К.Крылова, об изоморфизме закономерностей организации полисемических и частотных характеристик слов, а также, возможно, о вхождении задачи полисемической классификации слов в языке в общий класс физических задач о стохастическом распределении частиц в ячейках " фазового пространства", или "пространства классификации" (Крылов Ю.К., 1982, с.80).

Об общности подхода Ю.К.Крылова, видимо, свидетельствует и тот факт, что для описания квантитативно-системных закономерностей организации собственно полисемии, исходя из тех же общих вероятностно-комбинаторных соображений, Ю.К. Крылов выводит и использует совсем другой аппарат, не схожий с нашим: уравнение, описывающее полисемическое распределение, у него является дискретным аналогом нормального закона. Существенно заметить, что этим распределением описываются не сами численности классов, а их разности (Крылов Ю.К., 1982, с.92). В связи с этим, видимо, можно говорить не только о выводимости разных форм описания количественных закономерностей полисемии из общего теоретического источника, не только о возможности их преобразования друг в друга, но и об их взаимодополнительности, описании с их помощью разных аспектов организации единого объекта[+].

7. Из приводимых ниже наших данных по трем основным толковым словарям современного русского литературного языка разного типа: большому академическому - 120 тыс.слов (Словарь,

---

[+] См. также (Крылов Ю.К., Якубовская М.Д., 1977).

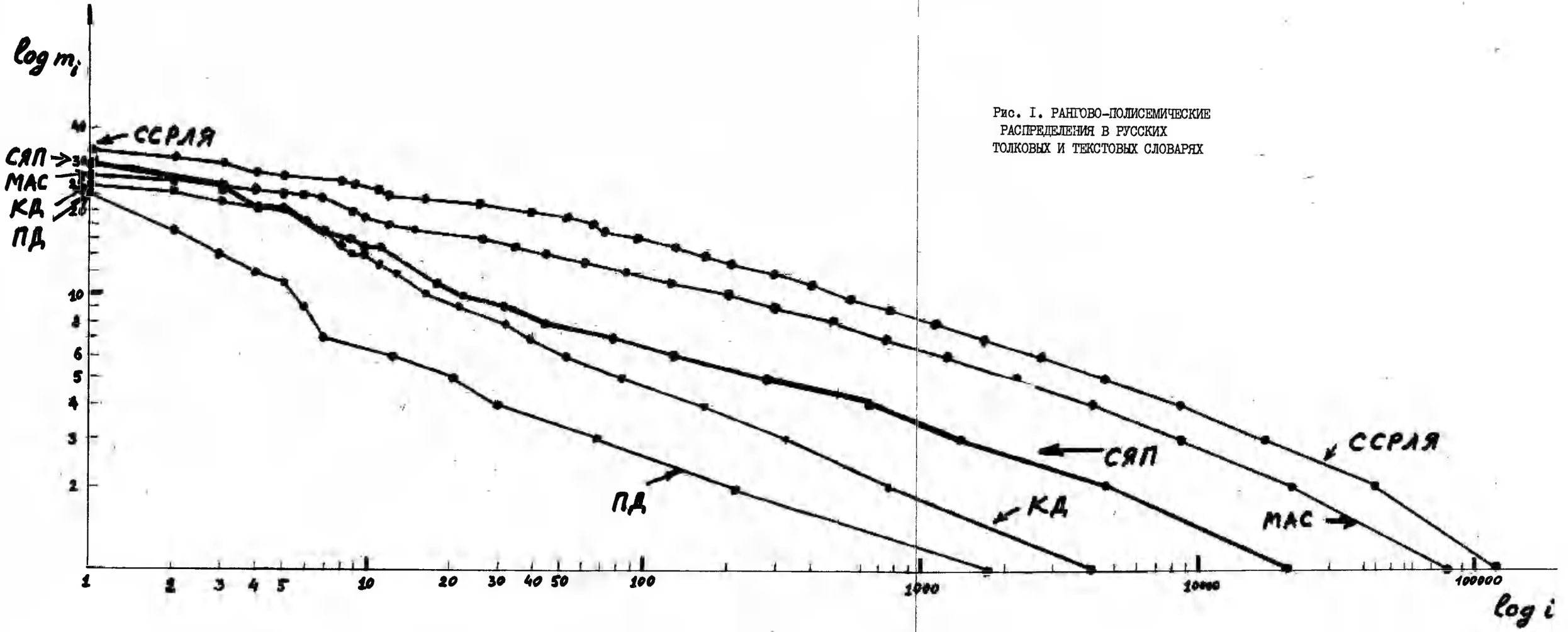

Рис. I. РАНГОВО-ПОЛИСЕМИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ В РУССКИХ ТОЛКОВЫХ И ТЕКСТОВЫХ СЛОВАРЯХ

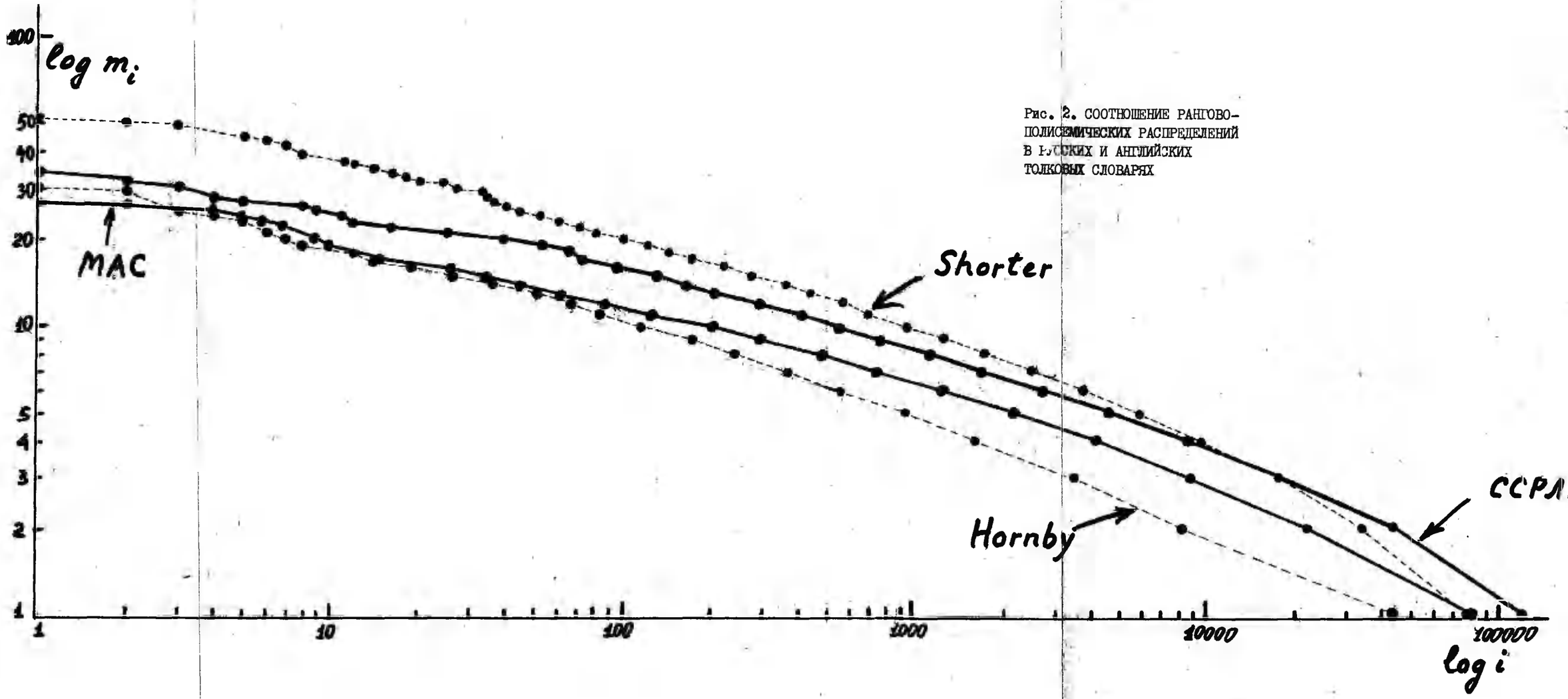

Рис. 2. СООТНОШЕНИЕ РАНГОВО-ПОЛИСЕМИЧЕСКИХ РАСПРЕДЕЛЕНИЙ В РУССКИХ И АНГЛИЙСКИХ ТОЛКОВЫХ СЛОВАРЯХ

I948-I965),среднему, так называемому МАС'у - 82 тыс. слов (Словарь,I957-I96I) и краткому -"Словарю русского языка" С.И. Ожегова - 57тыс. слов (Ожегов С.И.,I972);по словарю языка писателя (Словарь языка Пушкина) - около 20тыс. слов (Словарь, I956-I96I));по двум толковым словарям английского языка:среднему "Shorter Oxford English Dictionary" -около 80тыс. слов (Shorter ,I962),краткому, с элементами учебного (Hornby , I962) - 50тыс. слов - смотри таблицы I и 2 и рис.I и 2 - можно сделать следующие выводы.

I) Рангово-полисемические распределения независимо от типа толкового словаря данного языка сохраняют устойчивое типологическое сходство, заключающееся в почти одинаковом угле их наклона ($\gamma_{рус.}$= 0,29-0,30; $\gamma_{англ.}$=0,33-0,35),в повторяющейся пропорции слов разной полисемии в словаре, независимо от его объема (типа).Это означает, что отбор слов и значений в словари меньшего объема происходит не случайным образом, а на основе достаточно четких критериев, системно соотносящих их состав с общим составом лексической системы языка. Эти критерии наличествуют в реальной языковой действительности, и лексикографы их не придумывают, а лишь учитывают. Они сводятся к тому,что краткий толковый словарь типа "Словаря русского языка" С.И.Ожегова отражает в своем составе (словнике и объеме толкований) преимущественно пересечение всех активных словников и значений культурных носителей языка данной эпохи,т.е. ядро актуального словоупотребления. Средний толковый словарь (типа МАС'а или словаря под ред.Д.Н.Ушакова (I935-I940)) отражает преимущественно объединение всех актуальных активных индивидуальных словарей. Наконец , большой словарь (типа I7-томного) является объединением всех активных и пассивных словарей носителей данного языка. т.е. содержит в себе и неупотребляемую сейчас лексику, но содержащуюся в культурных текстах, доставшихся нам от прошлых эпох и понятную образованным носителям данного языка.

Глубина исторического взгляда" словаря зависит,как правило,от того,до каких исторических пределов простирается так называемый "современный литературный язык" данного народа,т.е. от какого рубежа тексты,написанные ранее,продолжают оставаться понятными,продолжают циркулировать и в современном обществе. Т.е. большой словарь - это максимальное объединение всего того,что продолжает сохранять коммуникативную ценность в современности и в прежнем языковом материале.

Таблица I

Соотношение количества слов разной полисемии по данным толковых, авторских и текстовых словарей русского и английского языков

К-во слов с данной полис. / Полисемия:	ССРЯН	МАС	СО	СЯП	ПД	КД	ДС	Shorter	Hornby
I	2	3	4	5	6	7	8	9	I0
5I								I	
50								I	
49								I	
45								2	
44								I	
42								I	
39								I	
37								3	
36								I	
35								2	
34								2	
33	I							2	
32								2	
3I	I							4	
30	I			I				3	I
29		I		I				6	I
28	I							I	
27	I							2	
26	3		2					4	
25	I	3		2		I		4	I
24	2	I						7	I
23	I	I			I			9	I
22	4					I		II	
2I	9					I		I0	
20	I4	2	I	I		I		I9	I
I9	I4	I						2I	I
I8	I2	2	3			I		2I	4
I7	7	3		2	I	I		35	2
I6	22	II	2					43	5
I5	35	9	3	2		I		55	7
I4	38	II	6		I	2		84	I0
I3	37	I7	I			2		79	I5
I2	83	25	I3	3	I	2		I3I	I6
II	II8	40	I4	5	I		3	I30	I6
I0	I5I	79	30	5		3	I	237	34
9	222	97	5I	8	I	5	2	308	57
8	342	I9I	42	I6		9	2	489	66
7	587	282	II6	4I	I	7	2	792	I29
6	I057	49I	23I	55	5	I5	4	I240	I80
5	I928	996	426	I54	9	28	4	2078	366
4	3989	2060	960	336	9	86	8	3825	707
3	9316	4680	2546	9II	39	I40	30	7771	I853
2	26I02	I3236	8390	3I64	I39	473	III	I6397	4688
I	76382	59920	44I66	I6494	I683	3913	I2I3	45965	362I0
В с е г о:	I2048I	82017	56998	20196	I89I	4687	I380	79823	44374

Таблица 2

Параметры рангово-полисемических распределений по толковым, авторским и текстовым словарям русского и английского языков

Словарь / Параметры	ССРЯН	МАС	СО	СЯП	ПД	КД	ДС	Shorter	Hornby
эксп.	I2048I	82017	56996	20196	I89I	4694	I380	79823	44374
теор.	I64000	82500	57000	20280	I000	4960	I400	80000	28200
i	0,295	0,30	0,30	0,30	0,325	0,35	0,357	0,35	0,33
i_{max}	94	7I	54	28	II,2	I9,6	I3,2	I60	62
C	I,6	I,28	0,79	0,4	0	0	0	2,II	I,I4
C/log i_{max}	0,093	0,079	0,050	0,028	0	0	0	0,I3	0,078

ССРЯН – "Словарь современного русского литературного языка" в I7-ти т.т. (I948–I965); МАС – "Словарь русского языка" под ред. А.П. Евгеньевой (I957–I96I); СО – "Словарь русского языка" С.И. Ожегова в 4-х т.т. (I972 – 9-е издание); СЯП – "Словарь языка Пушкина" (I956–I96I); ПД – "Пиковая дама" А.С. Пушкина; КД – "Капитанская дочка" А.С. Пушкина; ДС – "Дама с собачкой" А.П. Чехова; Shorter – "Shorter Oxford English Dictionary" (1962); Hornby – A.S. Hornby. "Oxford Advanced Learner's Dictionary of Current English" (1982).

Таким образом, реальное системное членение словарного состава отражается в иерархии словарей, в их системной соотнесенности[+].

2) Имея устойчивые структурные характеристики, оцениваемые ранее приведенными уравнениями (4) и (5), эмпирические полисемические распределения имеют и зоны структурной неопределенности, ненадежности относящихся к ним данных. Таковых зон – две: в самом начале и в самом конце распределений, в зоне самых многозначных и в зоне однозначных слов. В зоне самых многозначных слов это связано со сложностью семантической структуры служебных слов, с их широкозначностью (в противовес многозначности), с непредметной, как правило, денотативной отнесенностью, с тонкими разграничениями между значениями, с их широкой, неспецифической сочетаемостью. Это и предопределяет субъективные трудности (отражающиеся, неизбежно, и в труде лексикографов) по выделению значений этих слов. Довольно часто лексикографическая мысль здесь идет по пути подмены выделения значений слов выделением их типовых употреблений в тех или иных контекстуальных условиях. Количество значений в сравнении с реальностью, вследствие этого, в определенной степени здесь завышается. Несколько деформируется и экспериментальная кривая полисемического распределения. Зона многозначных слов в полисемическом распределении поэтому и должна рассматриваться особо.

Еще более неопределенной для учета и для представления в полисемических распределениях предстает зона однозначных слов. Это участок открытости словаря, его роста, возможности постоянной пополняемоти. Какие из слов этой зоны входят в состав литературного языка, а какие не входят, какие включать, а какие не включать в состав даже самого большого словаря – это болезненные вопросы для лексикографов и для теоретиков языка. Очевидно, что это периферия языка, что однозначная лексика – это, чаще всего, либо специальная, употребительная лишь в узких областях жизнедеятельности общества

[+]Наше понимание словарной иерархии является развитием идей С.И.Ожегова (1953). Реальность указанного соотношения состава разных словарей подтверждается психолингвистическими экспериментами по тестированию индивидуального словарного знания (Поликарпов А.А., 1984; Поликарпова А.О., Поликарпов А.А., 1986).

терминология, либо, хотя и бытовые, но совсем свежие индивидуальные новообразования. Где же граница между этой периферией и всем остальным словарным составом, каковы внешние границы терминологического состава языка, как отделить слова-однодневки, индивидуально-окказиональные образования от новообразований, вошедших в язык? Все это, как видно, вопросы, связанные с представлениями о соотношении социального и индивидуального, типового и окказионального в языке, с проблемами определения границ литературного языка. Сложившаяся в русской лексикографии практика, опирающаяся на суженные представления о составе и границах литературного языка, характеризуется недооценкой терминологической лексики как составного компонента литературного языка+, что, например, отразилось в факте слабого включения этих единиц в состав 17-томного академического словаря. По определению самих составителей (см. указанное предисловие) лексика подобного плана допускалась только в том случае, если она употребляется не в одной, а в нескольких специальных сферах общения, является междисциплинарной. Разумеется, это ограничение находится в противоречии с целевой предназначенностью словаря такого типа – полным охватом всей литературной лексики. Сказывается ли это произвольное изъятие части состава словаря на его системно-количественных характеристиках, например, на том, не выпадает ли зона однозначных слов из общей тенденции во всем словаре соотношения объемов групп слов разной полисемии? Как оказывается, да, и весьма заметно.

Произвольное изъятие из системы части элементов сразу обнаруживается в виде зияющего провала: рангово-полисемическое распределение в зоне перехода от двузначных слов к однозначным надламывается, демонстрируя здесь дефицит самых семантически специфичных слов. Подсчеты показывают, что лексикографы, возможно, упустили 40–50 тыс. однозначных слов.

Лексическая система существует не в воображении лингвистов, а объективно. При попытках ее урезанного представления она совершенно явным образом "протестует" в виде деформации некоторых своих квантитативно-системных характеристик (в данном случае – квантитативно-полисемических).

Соответствие наших предположений о возможности подобной деформации квантитативно-полисемической структуры словаря и реальной представленности этой деформации в нем само

+ См. предисловие в 1-м томе к (Словарь, 1948–1965).

по себе является вполне убедительным для признания эффективности теоретических представлений, которые лежали в основе подобного прогноза. Требуется, однако, неоднократная демонстрация подобного эксперимента, чтобы идея, лежащая в его основе, была окончательно доказана, в том числе, и со стороны корректности "условий проведения эксперимента".

Самое прямое, самое радикальное подтверждение подобных представлений появилось совсем недавно. В словарном академическом секторе Института русского языка АН СССР под руководством Р.П.Рогожниковой был создан сводный словник всех толковых, терминологических и энциклопедических словарей русского языка. В настоящее время он готовится к печати и включает в себя около 163тыс. лексических единиц, т.е. на 43тыс. больше, чем в большом академическом словаре. Это, в основном, та специальная лексика, которая не попала в большой словарь. Являясь преимущественно однозначной, она, в основном, и покрывает тот дефицит однозначных слов, который был обнаружен в семнадцатитомном ССРЛЯ.

Эти факты, видимо, свидетельствуют о том, что границы, пусть и размытые, вероятностные у литературно-нормативного словаря существуют. Класс однозначных слов, действительно, самый неопределенный по своему составу, но при использовании достаточно четких критериев отбора из него единиц (например, (1) "нормативность", т.е. социальная признанность, (2) "определенная историческая глубина", (3)"полнота", т.е. критериев отбора лексики в большой литературно-нормативный словарь), его границы тоже в определенной степени проясняются.

Эти факты позволяют нам поспорить с утверждением о некорректности постановки вопроса об объёме словаря литературного языка. Определенный язык (русский, английский и т.д.), в определенное время (сейчас, последние 20, 100, 200, 900 лет и т.д.), определенного социального статуса (с вульгарными и индивидуально-окказиональными слоями, только нормативно-литературный и т.д.) обладает словарем определенного, вероятностно оцениваемого объема. Наблюдаемые в тех или иных случаях отклонения в практике составления словарей от тех принципов, на которых основывается тот или иной тип словаря, ведут либо к неполному отражению лексической сферы, подпадающей под словарный тип в ее периферийной, чаще всего однозначной области (это, чаще, наблюдается для больших и средних словарей), либо к переполнению в кратких и средних словарях

той же самой периферийной области малозначных слов, за счет привлечения таких слов, которые относятся к более широкой сфере, чем та, которая должна по типовой целеустановке охватываться такими словарями.

Русские средние и краткие словари оказываются относительно уравновешены по соотношению центральных и периферийных полисемических зон. Английские средние и краткие толковые словари обладают избытком однозначной лексики (см. рис.I и 2).

Подобное положение дел заставляет при оценке характеристик полисемических распределений ориентироваться на их основную, среднюю часть, не принимая во внимание зону самых полисемических и однозначных слов.

Следует, однако, дополнительно подчеркнуть, что все сказанное отнюдь не свидетельствует о принципиальной ненадежности данных о квантитативно-системных характеристиках слов, получаемых из толковых словарей. Позитивистский скепсис в отношении объективности словарей и системности лексического состава языка в целом в настоящее время должен быть решительно преодолен, как не соответствующий реальному положению вещей. Вместе с тем, информация, представленная в словарях, должна оцениваться с учетом возможности некоторых систематических смещений в ее организации (типа указанных выше).

3) Важным системным параметром полисемических распределений, как уже указывалось выше, является параметр Ж. Он характеризует степень однородности, степень концентрации единиц разного семантического объема в пределах распределения. Увеличение значений параметра Ж (выражающееся в графической форме распределения в увеличении степени выпуклости) при равенстве объемов сопоставляемых словарей свидетельствует об усилении в нем средней (в данном случае - среднеполисемичной) зоны.

В тех случаях, когда мы сопоставляем словари неодинакового объёма, поправка Ж должна быть взвешена в отношении логарифмической меры объёма словаря.

Полученные нами данные по толковым словарям русского и английского языков свидетельствуют о том, что степень выпуклости полисемических распределений систематически увеличивается при переходе от краткого словаря к среднему и от среднего - к большому. Одна из возможных интерпретаций этого явления заключается в том, что, видимо, полисемическое распре-

деление больших по объему, по широте охвата лексики словарей оказывается более усредненным, более компактно сконцентрированным в средней зоне, чем распределение меньших по объему словарей.

В целом, эта динамика напоминает динамику постепенного усиления искривления частотной структуры слов, если последовательно идти от отдельного текста ко все более и более объемным совокупностям текстов. Исходный пункт квантитативно-полисемической динамики тот же, что и в случае частотной динамики, – отдельный текст. Рангово-полисемическое распределение отдельного текста аппроксимируется прямой линией в билогарифмической системе координат (см. на графиках распределения, например, для отдельных пушкинских текстов), как и рангово-частотное распределение. Рангово-полисемическое и рангово-частотное распределения, получаемые на основе анализа массивов, состоящих из все большего и большего числа разных текстов, демонстрируют тенденцию ко все более значительному искривлению своей геометрии в указанной системе координат. Словарь языка текстов писателя – один из промежуточных пунктов в этой динамике. Закономерно то, что и степень криволинейности графика, измеряемая отношением K к логарифму объема словаря последовательно растет при движении от словаря текста к словарю всех текстов одной личности, а далее – к словарю ядерной части всей актуальной лексики данного народа, всей актуальной лексики и, наконец, – к словарю всего нормативного лексического состава данного языка, в т.ч. и с его пассивным слоем (слоем устаревшей, не употребляемой ныне живущими, но понятной им лексики).

Как видно, использование поправки K имеет отнюдь не формальный, а глубоко содержательный, типологический смысл, помогает понять важные структурные отличия между словарями разного типа.

8. Разумеется, приведенные соображения и данные отнюдь не исчерпывают всей проблематики системно-квантитативного исследования полисемии языковых единиц. Взят был уровень единиц, самых очевидных языковому сознанию, самых исследованных, самых описанных (в словаря) – уровень лексических единиц. Ожидают еще своего серьезного описания и квантитативного исследования типовые семантические функции (значе-

ния) морфем, словосочетаний.

Сопоставимости результатов исследования квантитативных характеристик полисемии слов, получаемых по разным языкам, до сих пор мешают различия в критериях выделения этих единиц. При отборе лексических единиц в словари языков разного типа. нерешенными во многих случаях остаются вопросы отграничения сложного слова от словосочетания и вопросы степени необходимого и однородного обобщения текстовых единиц при движении от словоформы к лексеме и, далее, к гиперлексеме+.

Не были затронуты и проблемы корреляции квантитативно - полисемических характеристик слов с рядом других их характеристик - частотой, длиной, морфемной сложностью, стилистическим статусом и др.

Все это - вопросы для дальнейших исследований.

Л И Т Е Р А Т У Р А

Андрукович П.Ф., Кондолев Э.И. О статистических и лексико-грамматических свойствах слов. - НТИ, сер.2, 1977, № 4, с.1-9.

Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единиц различных уровней организации текста. - Учен. зап. Тарт. ун-та, вып. 689. Труды по лингвостатистике. Тарту, 1984, с.35-60.

Булагов Р.А. Многозначность слова. - НДВШ. Филологические науки. 1958, № 1, с.5-18.

Виноградов В.В. Русский язык. М., 1947.

Вишнякова С.М. Опыт статистического исследования многозначности слов в английском языке. - В кн.: "Вычислительная лингвистика", М., "Наука", 1976, с.168-178.

Денисов Н.Н. Место и роль самых многозначных слов в лексической системе языка. - В кн.: "Слово в грамматике и словаре". М., "Наука", 1984, с.142-158.

Кондлов Ю.К. Об одной парадигме лингвостатистических распределений. - Учен.зап. Тарт. ун-та, вып. 628. Труды по лингвостатистике. Тарту, 1982, с.80-102.

Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова. - НТИ, сер.2, 1977, № 3, с.1-6.

Марчук Ю.Н. (сост.) Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. Ч.I-II, М., ВШ, 1976.

Милевский Т. Предпосылки типологического языкознания. - В кн.: "Исследования по структурной типологии" (отв. ред. Т.Н.Молошная). Изд-во АН СССР, М., 1963, с.3-31.

+О разграничении лексемы и гиперлексемы см. [Борода М. Г., Поликарпов А.А., 1984].

Обухова Н.В. О квантитативно-системных характеристиках полисемии в китайском языке на иероглифическом и фонетическом уровнях. – В кн.: "Квантитативная лингвистика и автоматический анализ текстов. 1986." Тарту, изд-во Тарт. гос. ун-та, 1986.

Ожегов С.И. О трёх типах толковых словарей современного русского языка. – ВЯ, 1952. №2.

Ожегов С.И. Словарь русского языка. Под ред. Н.Ю.Шведовой. 9-е изд. М., "Русский язык", 1972.

Папп Ф. О некоторых количественных характеристиках словарного состава языка. – "Slavica", т.VII, Дебрецен, 1967, с.51-58.

Папп Ф. О машинной обработке одноязычных словарей (на материале венгерского словаря). – НТИ, сер.2, 1969, №3, с.20-29.

Поликарпов А.А. Факторы и закономерности аналитизации языкового строя. Канд.дисс., М., МГУ, 1976.

Поликарпов А.А. Элементы теоретической социолингвистики. М: изд-во Моск. ун-та, 1979.

Поликарпов А.А. Квантитативная универсалия удельной семантической специфичности: Логика теоретического выведения, измерения и соположения с другими языковыми параметрами. – В кн.:"Количественные методы в гуманитарных науках". М: изд-во Моск. ун-та, 1981, с.166-171.

Поликарпов А.А. Оппозиции "Ядро-периферия" и "специфичность-усредненность" – основания двух типов словарной иерархии. – В кн.: "III Всесоюзная конференция по теоретическим вопросам языкознания. "Типы языковых общностей и методы их изучения". (Тезисы)". М., АН СССР, 1984, с.124-125.

Поликарпов А.А., Бушуева О.В. О закономерных системно-количественных соотношениях полисемических и контекстуальных признаков слов. – В кн.: "Международная конференция "Теория и практика научно-технического перевода" (Москва, 2-6 декабря 1985 г.). Тезисы докладов и сообщений". М., ВШ, 1985, с.49-51.

Поликарпова А.О., Поликарпов А.А. Опыт изучения уровня и характера знания русской лексики. – В кн.:"Квантитативные аспекты системной организации текста. Мат-лы межвуз. сем." Тбилиси, 1986.

Поливанов Е.Д. Статьи по общему языкознанию. М., 1968.

Словарь русского языка. Под ред. А.П.Евгеньевой. Тт.I-4, М., Изд-во АН СССР, 1957-1961 г.

Словарь современного русского литературного языка. Тт.I-17, М-Л.: изд-во АН СССР, 1948-1965.

Словарь языка Пушкина. Тт.I-4. М., изд-во АН СССР, 1956-1961.

Толковый словарь русского языка. Под ред. Д.Н.Ушакова. Тт.I-IУ. М., 1935-1940.

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. – "Linguistica", XI, (Учен.зап. Тарт. ун-та, вып.502), Тарту, Изд-во Тарт.гос.ун-та, 1979, с.107-141.

Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики (на материале эстонского языка) Докт.дисс. Тарту, Тарт.гос.ун-т, 1984.

Hornby A.S. Oxford Advanced Learner's Dictionary of Current English. M., "Русский язык ", Oxford, " Oxford University Press", 1982.
Kelly E.F., Stone Ph.J. Computer Recognition of English Word Senses. Amsterdam - Oxford, North-Holland Publishing Company. 1975.
Lorge I. Semantic Count of the 680 Commonest English Words. N.Y., 1949.
Roget's Thesaurus of English Words and Phrases. New edition completely revised and modernized by R.A.Dutch.L., Longmans, 1963.
The Shorter Oxford English Dictionary, vols. I - II, 3d ed . Oxford, 1962.
Webster's International Dictionary of English Language (3-d edition). Springfield (Mass.), Merriam-Webster, 1961.
Zipf G.K. The Meaning-Frequency Relationship of Words - "Journal of General Psychology", 1945, v.33,N2,P. 251-255.
Zipf G.K. Human Behavior and the Principle of Least Effort . Boston (Mass.), 1949.

QUANTITATIVE ASPECTS OF POLYSEMY

Anatoliy A. Polykarpov

S u m m a r y

The paper deals with the problem of some quantitative features of polysemy viewed as a language universal. New notions such as semantic uncertainty and semantic specificity have been introduced and some measures quantifying them suggested. An equation relating the number of meanings to the rank of the word (with the words ranged in the descending order of the number of meanings) $\left(p_i = \frac{B}{i^{\gamma} + k} - k\right)$ (where p_i is the number of meanings, i - the rank of the word, B , γ , k - parameters) has been constructed, analysed and checked up by using data from several Russian and English general, author's and textual dictionaries. The parameters of the equation have a definite typological and lexicographic significance. The parameter γ is an indicator of the type (the degree of analyticity) of the language whose dictionary is under consideration, the parameter B - of the dictionary type (complete, medium and short types), the parameter k - of the degree of dictionary homogeneity (increasing with the growth of the volume of the sense field covered by the dictionary).

К ВОПРОСУ О КЛАССИФИКАЦИИ И ИНТЕРПРЕТАЦИИ ЛИНГВИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Ю.А. Тулдава

Одним из важнейших методов исследования в квантитативной лингвистике является моделирование с помощью распределений. В статье рассматриваются некоторые общие вопросы, связанные с классификацией (типологией) лингвистических распределений и их качественной интерпретацией.

Основные типы лингвистических распределений. Само понятие распределения можно истолковать в широком смысле как упорядоченную совокупность результатов измерения; последнее понимается как отношение объекта (X) к количественно выраженному значению (У) какого-либо признака (Р), допускающего количественную оценку (частота, объем, длина и др.). Если X - лингвистический объект, то можно выделить три основных типа лингвистических распределений.+

I. О д н о о б ъ е к т н о е распределение (один объект, несколько признаков):

	P_1	P_2	...	P_n
X	y_1	y_2	...	y_n

По этой исходной схеме один объект соотносится с результатами измерения в разных условиях. Например, если X - конкретная лингвистическая единица (или класс этих единиц), P_i - частоты в разных текстах, y_i - конкретные значения частот, то удается проследить "поведение" одной определенной лингвистической единицы в серии испытаний. Если ранжировать данные по убыванию или возрастанию количественных значений y_i (и соответственно P_i), то можно:

а) составить вариационный ряд с указанием частот или вероятностей появления y_i; в этом случае образуется т.н. спектр, или спектральное распределение;

б) приписать ранги (порядковые номера) значениям y_i, в результате чего получается т.н. ранговое распределение.

По спектральному однообъектному распределению определя-

+ Подробнее см. Тулдава Ю.А., 1982; другие возможные подходы к классификации лингвистических распределений см. Алексеев П.М., 1978 и 1985; Мартыненко Г.Я., 1982.

ется вид распределения (в отношении лингвистических объектов это чаще всего распределение "гауссового семейства", т.е. биноминальное, нормальное, пуассоновское и др.). Ранговое распределение при данной схеме однообъектных распределений обычно не представляет интереса при изучении лингвистических объектов. Особнй случай возникает при ранжировании (упорядочении) $У_i$ по качественным соображениям, например, при рассмотрении динамических (диахронных) процессов. В таких случаях можно исследовать распределение как "тренд" – в виде функциональной зависимости значений $У_i$ от $P(t_i)$, где t_i – моменты времени.

2. М н о г о о б ъ е к т н о е распределение (несколько объектов, один признак):

	P
X_I	$У_I$
X_2	$У_2$
...	...
X_m	$У_m$

Согласно этой схеме разные объекты измеряются по одному общему признаку. Например, X_i – разные лингвистические единицы (или классы единиц), P – частота в одном тексте, $У_i$ – конкретные значения частот. По этой схеме составлен, например, обыкновенный частотный словарь слов или классов слов (частей речи, фонетических, морфологических, семантических и др. типов слов), а также любой частотный список разных других лингвистических единиц. При этом получается распределение частот разных единиц относительно друг друга в данной совокупности (в частности в одном индивидуальном тексте или в совокупности текстов, рассматриваемой в эксперименте как одно целое, например, в текстах одного жанра или подъязыка).

Здесь, так же как и при однообъектном распределении, различаются две разновидности (формы представления):

а) спектральное распределение (спектральная форма распределения), когда одинаковые результаты измерений объединяются в группы с указанием числа объектов с данным результатом измерения; например, когда исследуется зависимость между частотой слова в тексте и количеством слов с данной частотой (т.н. частотный, или лексический спектр);

б) ранговое распределение, при котором ранжированным значениям частот $У_i$ приписываются ранги (i) и исследуется зависимость между $У_i$ и i (например, ранговое распределение

частот слов).

Как спектральное, так и ранговое многообъектное распределения в лингвистике относятся, как правило, к распределениям "негауссового семейства". В них проявляется одно из характерных свойств коммуникативных систем: асимметричное распределение элементов по "значимости", вследствие чего основную функциональную нагрузку несет небольшое число доминирующих элементов ("ядро").

3. К о м п л е к с н о е (многомерное) распределение (несколько объектов, несколько признаков):

	P_I	P_2	...	P_n
X_I	y_{II}	y_{I2}	...	y_{In}
X_2	y_{2I}	y_{22}	...	y_{2n}
...				
X_m	y_{mI}	y_{m2}	...	y_{mn}

Данная схема комбинируется из двух первых схем. В простейшей форме число объектов (X) или число признаков (P) равняется двум, например, когда сравниваются частоты разных слов в двух разных текстах или частоты разных классов слов в двух аспектах: в словаре и тексте. На основе схемы комплексного распределения исследуются "многомерные" задачи: взаимосвязь и совместная вариация ряда объектов или признаков. Связь между распределениями количественных значений (по "горизонтали" или "вертикали") может быть выражена функциональной зависимостью (уравнением регрессии), а сила этой связи может быть измерена коэффициентом корреляции. Внутренние связи по всей совокупности данных можно установить методами факторного анализа, кластер-анализа и др.

Итак, все рассмотренные типы распределений могут служить квантитативному исследованию и моделированию лингвистических объектов. При этом возможны различные способы представления распределений. Как уже было отмечено, рассмотренные типы распределений могут иметь спектральную или ранговую форму, кроме того, они могут быть представлены либо в дифференциальной (некумулятивной), либо в интегральной (кумулятивной) форме, а сами частоты могут быть абсолютными или относительными (последние могут быть интерпретированы как вероятности). С точки зрения технического оформления всякое распределение может быть задано в виде таблицы (ряда или матрицы распределения), графика или может быть выражено математически в виде формулы (обычно функциональной зависимос-

ти). В силу некоторых теоретических и практических соображений дискретное распределение лингвистических единиц обычно описывается через непрерывную функцию распределения. Вполне возможно также описание некоторых видов распределений в терминах теории нечетких множеств.

Отличительной чертой данного подхода к классификации (типологии) лингвистических распределений является то, что здесь исходят из качественных представлений о распределении, как неком результате измерения, при котором соотносятся друг с другом три основных компонента: объект (X), признак (P) и значение признака (У). Хотя в строгом смысле собственно распределением можно считать лишь ряд У в спектральной или ранговой форме, все же при конкретном анализе нельзя забывать о "происхождении" этого ряда, т.е. о связи компонента У с компонентами X и P. Это необходимо не только для выработки предварительных гипотез (о форме распределения и т.д.), но и для содержательной интерпретации результатов квантитативного анализа, с учетом специфики и задач л и н г в и с т и ч е с к о г о исследования. Кроме того, как было показано выше, именно учет всех компонентов процедуры измерения (X, P и У) позволяет наиболее четко и естественно различать основные типы лингвистических распределений: однообъектных ("горизонтальных"), многообъектных ("вертикальных") и комплексных.

На практике допускается "гибкая" трактовка понятий объекта, признака и значения признака. Объектом может быть индивидуальная единица (например, конкретное слово) или класс единиц (часть речи и т.п.). Признак (точнее, наименование признака) может быть простым или сложным, например, "частота в тексте T_i". Значение признака может быть количественно-пропорциональным, интервальным или порядковым (в т.ч. "балльным"). Вследствие такой гибкости в определении компонентов измерения возможны различные варианты классификаций лингвистических распределений в зависимости от конкретных условий и задач исследования.

Распределение как модель вероятностной системы. При квантитативно-системном подходе (Тулдава Ю.А., 1980) изучаемые лингвистические объекты рассматриваются как вероятностные системы в том широком смысле, который придается им в современных системных исследованиях (см., например, Оачков Ю.В., 1971; Кравец А.С., 1976). Вероятностная система при

таком понимании характеризуется единством признаков устойчивости (детерминированности) и иррегулярности (флуктуации, вариативности), проявляющихся на базе единства философских категорий необходимости и случайности. Важнейшей структурной характеристикой вероятностной системы считается распределение, которое отражает не только упорядоченность в системе, но и взаимодействие между элементами и общность в их поведении, т.е. целостность системы. В более общем плане распределение указывает на устойчивость и регулярность в массе вероятностно-случайных событий.

Исследование лингвистических распределений начинается, как правило, с установления эмпирического, или частотного распределения, которое можно считать моделью вероятностной системы в первом приближении. Частотное распределение может дать достаточно хорошее представление об устойчивых чертах в структуре и функционировании исследуемой системы. Моделирование данных с помощью какого-либо теоретического распределения поднимает исследование на более высокую ступень обобщения. Известна важная роль теоретических распределений в отображении закономерностей материального мира, и можно предположить, что в отношении некоторых классов лингвистических объектов такое моделирование открывает широкие возможности не только для решения многих актуальных прикладных задач, но и для выявления более глубоких квантитативных закономерностей структуры и функционирования языка.

Ориентация на моделирование с помощью распределений при квантитативно-системном подходе объясняется спецификой системного исследования, которая состоит в том, что изучение объекта осуществляется именно в том аспекте, в каком он представляет систему при данном подходе. В этом смысле лингвистическое распределение является м о д е л ь ю - о п и с а н и е м тех языковых объектов, которые можно представить себе как вероятностные системы в вышеуказанном понимании. В самом процессе моделирования направление мысли идет в начале от объекта к модели (формирование модели), а затем от модели обратно к объекту (интерпретация модели, получение нового знания об объекте).

Интерпретация лингвистических распределений. По своей природе лингвистические распределения не являются моделями "чистой математики", а их надо рассматривать как интерпретируемые знаковые системы. Выявление свойств и отношений, а так-

же внутренних закономерностей, которые определяют характер лингвистического распределения в целом, должно сопровождаться качественным (содержательным) анализом результатов исследования.

Интерпретацию (объяснение, толкование, раскрытие смысла) лингвистических распределений как моделей изучаемых вероятностных систем можно представить себе как многоуровневый (многостадиальний) процесс качественно-количественного анализа и синтеза. Задача интерпретации состоит в том, чтобы раскрыть сущность явления, однако здесь могут быть разные пути к достижению поставленной цели. Различаются индуктивный и дедуктивный подходы, причем высшей формой интерпретации считается научное объяснение, которое должно показать, что "данний научный факт является проявлением определенного закона или что объясняемый закон вытекает из более общего закона (или же теории)" (Друянов Л.А., 1980, с. 53). В зависимости от условий и задач исследования и по характеру материала при исследовании лингвистических распределений можно условно выделить три типа интерпретаций: структурно-функциональную, прагматическую (стилистическую) и генетическую (причинную). На практике эти три типа интерпретаций, будучи взаимосвязанными и взаимопроникаемыми, могут сосуществовать в одном конкретном исследовании.

I. Структурно-функциональная интерпретация

Важным условием моделирования является проведение предмодельного анализа физической сущности изучаемого явления с целью формирования некоторой априорной информации для вывода общего вида искомой модели. Наиболее простым и удобным средством представления информации о распределении является графико-геометрический метод. Графики (рисунки) позволяют представить данные в наглядной форме при минимальной их обработке. По графическому изображению можно сделать обоснованные выводы об общей форме распределения и о характере взаимосвязи элементов изучаемой системы (симметричность – несимметричность, линейность – нелинейность, унимодальность – мультимодальность и т.д.). S-образная кривая говорит о том, что данное распределение может отражать динамический процесс, подчиняющийся логистическому закону роста, а гипербола на графике может указать на феномен "концентрации и рассеяния" элементов системы. Графики могут быть представлены в обыкновенной декартовой системе координат или в модифицированной

форме, например, в логарифмическом масштабе, на котором легко можно удостовериться в соответствии или несоответствии данных определенным законам распределения. Важно отметить ценность графиков при попытках добыть новую информацию по отклонениям от общей тенденции в отдельных частях распределения, по точкам перегиба и т.п. Так, например, анализ формы кривой распределения частот слов в билогарифмических координатах дал толчок для выявления особого "нелинейного" типа рангового распределения лексики в больших текстах и для формулировки тезиса о "четвертом приближении закона Ципфа" (Алексеев П.М., 1978).

В отличие от графиков формулы обладают тем важным преимуществом, что допускают проведение различных операций, которые сами по себе могут служить основанием для раскрытия новых, неожиданных сторон изучаемого явления. Формулу можно рассматривать как символическую (аналитическую) запись структуры данного явления, но формула может дать также представление о функционировании системы, о динамическом процессе, о развитии и т.п.

В качестве аналитических записей распределений в квантитативной лингвистике обычно употребляются различные функции или дифференциальные уравнения. Представляя распределение в виде функции (функциональной зависимости)+, часто начинают с определения вида функции по опытным данным. Такой индуктивно-эмпирический подход может иногда дать вполне приемлемые результаты. Например, если установлено, что данное эмпирическое распределение описывается степенной функцией типа $y = ax^b$ (где a и b – параметры), то можно сделать вывод об "аллометрическом" законе изменения y в зависимости от x, причем параметр b ("коэффициент относительной эластичности") позволяет определить средний процент изменения y в связи с изменением x на I %, а параметр a указывает на начальное значение y при $x = I$. Безусловно, качественная интерпретация формулы и ее параметров требует анализа явления или процесса по существу, по внутренней логике или по физической сущности с целью формирования адекватного представления об изучаемом явлении. Далее, представляя эту же

+ В отношении вероятностных систем функциональная зависимость трактуется как вероятностная функция распределения. Формально между ними нет разницы, пока мы не вкладываем определенного смысла в математические символы.

формулу в дифференциальной записи, например, в форме уравнения $\frac{dy/y}{dx/x} = b$, мы видим, что соотношение относительных приростов y и x постоянное (сохраняет устойчивость) и тем самым явление подчиняется закону "постоянного относительного роста" (Ланд К.Ч., 1977, с. 388) - одному из важнейших законов, характеризующих некоторые типы сложных самоорганизующихся систем.

Аналогично может интерпретироваться экспоненциальная функция как "лавинообразный рост", логарифмическая функция как "закон адаптационного торможения" (Налимов В.В., Мульченко З.М., 1969, с. 41) или "закон пропорционально убывающего относительного роста" (Ланд К.Ч., 1977, с. 388), экспоненциальная функция как "рост с начальным ускорением и последующим замедлением (и насыщением)", функция Вейбулла как "обобщенная модель прогрессивного роста (включающая экспоненциальный закон как частный случай)" (см. Добров Г.М., 1969, с. 158) и т.д.

На другом (дедуктивном) уровне научного анализа используются гипотетико-теоретические модели вероятностных систем, т.е. когда те или другие функции-модели выводятся гипотетически, на основе некоторых теоретических постулатов большей или меньшей степени общности. Интерпретация таких моделей включает в себе как сами исходные постулаты, так и возможные выводы, сделанные на основе анализа конкретного материала. В квантитативной лингвистике известны попытки вывода закона Ципфа, используя понятие вероятностного процесса (например, Simon H.A., 1955) или прибегая к аналогии с термодинамикой (Mandelbrot B., 1954). Некоторым авторам удалось вывести формулы квантитативной структуры текста, исходя из комбинаторно-вариационных принципов (Арапов М.В., Шрейдер Ю.А., 1978; см. также статью Ю.К. Крылова в настоящем сборнике). Модель "обобщенного закона Ципфа-Мандельброта" (Орлов Ю.К., 1976) выводится теоретически, но тезис об особой роли "объема Ципфа" обосновывается и интерпретируется, исходя из эмпирических соображений, причем делается вывод, что выполнение закона Ципфа (соответствие "объему Ципфа") свидетельствует о "высокой степени организованности целостного текста".

В квантитативной лингвистике практикуется также промежуточный "гипотетико-эмпирический" подход, при котором исходная модель получается на основе некоторых относительно самостоятельных теоретических схем (гипотез), а конкретная форма распределения определяется эмпирически итеративным пу-

тем, придерживаясь требований исходной теоретической модели (см., например, Тулдава Ю.А., 1980а).

Далее, имеются интересные попытки сконструировать и интерпретировать теоретические модели, исходя из анализа взаимных зависимостей между лингвистическими объектами (например, трактовка закона Менцерата по Г. Альтманну, см. Altmann G., 1980) или исходя из разветвленной системы содержательных предпосылок (Köhler R., 1986). Конечной целью в этих исследованиях объявляется построение адекватной лингвистической теории в рамках общей теории саморегулирующихся систем.

2. Прагматическая (стилистическая) интерпретация

Под этим названием подразумевается объяснение, которое основывается на связи формальных (квантитативных) показателей с некоторыми прагматико-стилистическими, в т.ч. оценочными характеристиками изучаемых явлений. В квантитативной лингвистике, в частности в ее подразделении – квантитативной лингвостилистике, или стилеметрии, оперируют такими содержательными (стилистическими) понятиями как "богатство" или "разнообразие" словаря, "целостность" и "художественная завершенность" текста, "стереотипность" и "экономия" высказывания и т.п. Все подобные понятия должны по замыслу исследователей охарактеризовать некоторые реальные свойства реальных объектов, однако эти свойства прямо не наблюдаемы и тем более прямо не измеряемы. Тогда обращаются к внешним, наблюдаемым сторонам изучаемых реальных объектов (словарей, текстов) и пытаются таким путем косвенно добраться до сути изучаемых явлений.

Теоретически такую ситуацию можно сравнивать с положением, известным из теории измерения, где рассматриваются переменные двух типов: I) латентные (скрытые) переменные, т.е. то, что исследователь выбирает, фиксирует для своего анализа; 2) индикаторы – то, что непосредственно измеряется (Хайтун С.Д., 1983, с. I6). При этом необходимым условием анализа является существование определенной связи между латентными переменными и индикаторами.

К формальным индикаторам можно отнести и лингвистические распределения. Известно, например, что "богатство" словаря определяется по особенностям формы рангового распределения слов (в частности по углу наклона кривой или по ципфовскому параметру γ) или по комплексному распределению, отражающему связь между ростом объема словаря и ростом объема

текста (по этой связи можно прогнозировать тенденцию роста, т.е. "потенциальное богатство" словаря). Асимметричное распределение частот слов истолковывается как осуществление принципа "предпочтения" или "значимости" определенной части словаря в данных условиях, что приводит к "концентрации и рассеяния" единиц. "Действенность" и "качественность" стиля определяются по распределению частей речи в тексте. И т.д.

При таком анализе необходимо помнить, что связь квантитативных индикаторов с латентными переменными (свойствами, характеристиками) носит вероятностный характер. Это значит, во-первых, что интерпретация данных может дать положительный эффект лишь на представительном материале, т.е. при достаточно больших выборках и при воспроизводимости результатов испытаний. Во-вторых, надо отдать себе отчет в том, что формальные индикаторы могут лишь косвенно отражать содержательно-латентные свойства изучаемых явлений. Возникает также вопрос, в какой мере выбранные индикаторы способны охватывать все существенные стороны явления и не может ли случиться искажение ("деформация") картины при неправильном выборе или недостаточной представительности индикаторов. Оправдано ли соотнесение данного индикатора (или данных индикаторов) к данной латентной переменной, в каждом конкретном случае проверяется практикой.

3. Генетическая (причинная) интерпретация

Генетической можно назвать такую интерпретацию, которая пытается объяснить изучаемое явление, указывая на его происхождение и тем самым прямо или косвенно устанавливая причину его существования. Здесь, так же как и при предыдущих типах интерпретации, надо иметь в виду, что мы имеем дело с вероятностным подходом к исследованию языковых явлений. При таком подходе надо исходить из вероятностной концепции причинности, согласно которой "причинность есть нечто, могущее присутствовать в большей или меньшей степени, а не только быть или не быть" (Винер Н., 1964, с. 309). Надо также признать принцип множественности причин, обусловливаемой множественностью связей данного явления с другими явлениями.

При исследовании лингвистических распределений можно, конечно, кое-что объяснить "внутренними" причинами, например, особенностями морфологической структуры данного языка.

Однако такие внутренние причины всегда связаны и переплетаются с внешними (внелингвистическим) причинами (контакты языков, общественные потребности и т.п.). В последнее время особый интерес представляют попытки психологического (психолингвистического), психофизиологического и филогенетического объяснения лингвистических распределений как моделей определенных сторон речевой деятельности.

Так, например, "гиперболическое" распределение частот слов (в виде степенной зависимости переменных), известное под названием закона Ципфа, пытаются связать с особенностями психики человека, его потребностью общения, с одной стороны, и стремлением свести к минимуму свои умственные и физические усилия, с другой стороны. Этот широко известный принцип "наименьшего усилия" (Zipf G.K., 1949) основывается, таким образом, на взаимодействии двух противоборствующих тенденций в психике (подсознании) человека. Взаимодействием противоположных тенденций при порождении речи (разнообразие - ограничение разнообразия) и ассоциативными свойствами человеческой памяти (при ограниченности объема кратковременной памяти) объясняются и некоторые другие известные лингвистичесие распределения, например, комплексное распределение, отражающее более медленный темп роста объема словаря по сравнению с темпом роста объема текста (Тулдава Ю.А., 1980а, с. 119 и след.).

Представителями психофизиологии некоторые типы лингвистических распределений как моделей речевой деятельности связываются со структурными особенностями мозга, в частности с пространственно-временной организацией периодических (циклических) процессов в мозге. Исходя из представления о кодировании образов слов "пакетами волн нейронной активности", А.Н. Лебедев выводит формулу, совпадающую с формулой Ципфа, для описания распределения частот слов в речи, и другую формулу, описывающую связь между ростом объема словаря и ростом объема текста (Лебедев А.Н., 1983 и 1986). Можно констатировать, что предположения о связи между особенностями квантитативной структуры текста и некоторыми закономерностями деятельности мозга, по-видимому, не лишены основания, и исследования в этой области приобретают актуальность.

В филогенетическом плане лингвистические распределения объясняются эволюционным развитием человеческого языка, который формировался на протяжении тысячелетий в результате приспособления к внешнему миру. Этот эволюционный процесс

развития языка можно "отдаленно уподобить органической эволюции на основе естественного отбора" (Панов Е.Н., 1980. с. 147). Известно высказывание А.Н. Бернштейна (1966) о том, что сущность речевой деятельности именно как деятельности заключается в оптимизации пути достижения поставленной цели. Такое системное свойство основных лингвистических распределений как иерархичность можно объяснить адаптацией к увеличению числа элементов (при порождении речи), учитывая, что иерархическая структура минимизирует число связей (см. Козачков Л.С., 1978, с. 15). Устойчивость некоторых лингвистических распределений (сохранение общей формы) может указать на стремление к равновесию, оптимальности и целесообразности самоорганизующейся сложной системы - языка.

Безусловно, многие из названных свойств имеют всеобщий характер, они обнаруживаются в живой и неживой природе. Это свидетельствует о "единстве мира", которое состоит, в частности, в том, что "более общие законы "ниже" лежащих уровней бытия сохраняют свою силу для всех "выше" лежащих уровней" (Брушлинский А.В., 1979, с. 47), причем эта универсальность "не только исключает, а, наоборот, предполагает наличие специфических закономерностей" (там же).

В итоге можно констатировать, что генетический метод объяснения, ставящий на первое место отыскание причин явлений, так же как и прагматический метод, основанный на анализе латентных переменных, занимают прочное место среди методов интерпретации лингвистических распределений (как моделей вероятностных лингвистических систем). Они расширяют возможности количественно-качественного анализа лингвистических явлений, обращают внимание на психологию и онтологию языкового творчества, могут дать импульс для новых открытий и для развития перспективных направлений исследования языка. Все же для полной и всесторонней научной интерпретации прагматический и генетический методы не всегда достаточны и должны быть дополнены структурно-функциональным объяснением.

Л И Т Е Р А Т У Р А

Алексеев П.М. О нелинейных формулировках закона Ципфа. - В кн.: Вопросы кибернетики. Вып. 41. М.; Л., 1978, с. 53-65.

Алексеев П.М. Лингвистические распределения (Элементы количественного анализа текста). - Л.: ЛГПИ, 1985. - 56 с.

Арапов М.В., Шрейдер Ю.А. Закон Ципфа и принцип диссимметрии системы. – В кн.: Семиотика и информатика. Вып. 10. М., 1978, с. 74–95.

Бернштейн Н.А. Очерки по физиологии движений и физиологии активности. – М.: Медицина, 1966.

Брушлинский А.В. Мышление и прогнозирование (логико-психологический анализ). – М.: Мысль, 1979. – 230 с.

Винер Н. Я – математик. – М.: Наука, 1964.

Добров Г.М. Прогнозирование науки и техника. – М.: Наука, 1969. – 208 с.

Друянов Л.А. Законы науки, их роль в познании. – М.: Знание, 1980. – 64 с.

Козачков Л.С. Информационные системы с иерархической ("ранговой") структурой. – Научно-техническая информация. Серия 2. М., 1978, № 8, с. 15–24.

Кравец А.С. Природа вероятности (философские аспекты). – М.: Мысль, 1976.

Ланд К.Ч. Сравнительная статика в социологии. – В кн.: Математика в социологии./Перев. с англ. – М.: Мир, 1977, с. 371–401.

Лебедев А.Н. Закономерности повторения слов в речи. – Психологический журнал, том 4. М., 1983, № 5, с. 11–22.

Лебедев А.Н. Нейрофизиологические пределы памяти человека и богатства его лексики. – Учен. зап. Тартуского ун-та, 1986, вып. 745, с. 95–108.

Мартыненко Г.Я. Типология лингвостатистических распределений. – Учен. зап. Тартуского ун-та, 1982, вып. 628, с.103–120.

Налимов В.В., Мульченко З.М. Наукометрия. Изучение развития науки как информационного процесса. – М.: Наука, 1969. – 192 с.

Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. – В кн.: Вычислительная лингвистика. – М.: Наука, 1976, с. 176–202.

Панов Е.Н. Знаки, символы, языки. – М.: Знание, 1980, – 192 с.

Сачков Ю.В. Введение в вероятностный мир. Вопросы методологии. М.: Наука, 1971. – 208 с.

Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (I). – Учен. зап. Тартуского ун-та, 1980, вып. 544, с. 143–158.

Тулдава Ю.А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. – Учен. зап. Тартуского ун-та, 1980а, вып. 549, с. 113-144.

Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (3): методика исследования. – Учен. зап. Тартуского ун-та, 1982, вып. 619, с. 123-143.

Хайтун С.Д. Наукометрия – состояние и перспективы. – М.: Наука, 1983. – 344 с.

Altmann G. Prolegomena to Menzerath's Law. – In: Glottometrika 2. Bochum: Brookmeyer, 1980, pp. 1-10.

Köhler R. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. – Bochum: Brockmeyer, 1986. – 201 S.

Mandelbrot B. Structure formelle des textes et communication: deux études. – Word, vol. 10, 1954, No. 1., pp. 1-27.

Simon H.A. On a Class of Skew Distribution Functions. – Biometrika, vol. 42. 1955, pp. 425-440.

Zipf G.K. Human Behavior and the Principle of Least Effort. – Cambridge (Mass.): Addison-Wesly Press, 1949.

ON CLASSIFICATION AND INTERPRETATION OF LINGUISTIC DISTRIBUTIONS

Juhan Tuldava

S u m m a r y

One of the main methods of investigation in quantitative linguistics is modelling with the help of distributions. The typology of linguistic distributions and their interpretation are discussed. Three main "schemes" of linguistic distributions are proposed including mono-objective, poly-objective and complex (multidimensional) distributions according to the possible combinations of the object (X), its characteristics (P), and the value of the characteristics (Y). The interpretation has been carried out on structural-functional, pragmatic (stylistic), and genetic (causal) levels.

ЧАСТОТНЫЕ СЛОВАРИ И ОПЫТ ИХ ИСПОЛЬЗОВАНИЯ.

Рецензия на книгу: В.С. Перебийніс, М.П. Муравицька, Н.П. Дарчук. Частотні словники та їх використання. – Київ: Наукова думка, 1985. – 204 с.

Рецензируемая монография обобщает опыт исследований коллектива отдела структурно-математической лингвистики Института языковедения им. А.А.Потебни АН УССР (г. Киев) по исследованию методов составления частотных словарей и их использованию. Работа состоит из следующих частей: "Методы составления частотных словарей" (автор В.И.Перебейнос), "Лингво-статистические свойства грамматических и семасиологических категорий" (автор М.П.Муравицкая), "Индивидуальное и общее в лексической системе авторского стиля" (автор Н.П.Дарчук).

Первая часть посвящена наиболее общим проблемам анализа того опыта составления частотных словарей и их использования , который накоплен в современной квантитативной лингвистике, а также выработке новых, всесторонне аргументированных рекомендаций для составителей словарей этого типа.

В первом разделе этой части проводится анализ словарей по применяемым единицам счета (словоформы, лексемы, словосочетания, идиомы и т.д.), по объему и характеру выборки, характеру представления материала в ЧС. Кроме того, анализируются статистические характеристики единиц, представленных в ЧС.

Во втором разделе описывается подготовительный этап составления ЧС, зависящий от целей его создания: определяется рациональный объем выборки, решается вопрос о лингвистической однородности выборки, рассматривается вопрос о характере текстов, которые войдут в выборку (полные тексты или их части, а если части, то какие по объему), создается схема подсчетов, определяется характер статистических показателей, которые будут приводиться в словаре.

В третьем разделе подробно излагаются этапы составления ЧС, а именно: организация выборки, составление алфавитно-частотного и рангово-частотного списков, подсчеты статистических показателей (средней частоты, меры ее колебания для измеряемых единиц и т.д.), описывается несколько способов проверки надежности и эффективности ЧС.

В четвертом разделе содержатся сведения о сфере применения ЧС: отбор лексических минимумов, определение существенности расхождения частот сопоставляемых единиц из ЧС, в т.ч. в разных тематических или жанровых частях выборки и т.д.

В целом, первая часть рецензируемой работы может быть

определена как весьма концентрированное изложение опыта работы с частотными словарями, накопленного к настоящему времени. Анализ типов ранее вышедших словарей близок к исчерпывающему. Предлагаемые решения представляют собой относительно завершенную на сегодняшний день систему. Особую ценность этой системе придает то, что она получила развернутую проверку в ходе составления "Частотного словаря современной украинской художественной прозы" (1981).

Следующие две части и посвящены многоаспектному анализу этого словаря и связанным с ним текстовым материалам.

Во второй части исследуется функционирование грамматических и семасиологических категорий. В частности, определяется частота в текстах каждого из аспектологических классов глаголов. Ставится также задача исследования закономерностей функционирования в текстах двух типов омонимов - лексико-грамматических и лексических. Рассматриваются стилистические характеристики взаимодействия синтаксиса и семантики в функционировании многозначного украинского слова, особенности отношения антонимичных слов между собой, а также выражения в тексте эстетической концепции автора.

Третья часть монографии посвящена такой важной теоретической проблеме, как установление общих и индивидуальных черт в лексической системе авторских стилей. Исследование ведется по материалам статистической картотеки к указанному частотному словарю современной украинской художественной прозы. Описываются закономерности статистической структуры текстов, привлеченных в выборку, закономерности прироста новых слов по ходу текста и обусловливающие их причины (соотношение прямой и авторской речи, увеличение количества описываемых ситуаций, особенности композиционной структуры текста и т.п.). Предлагается комплекс квантитативно-стилистических характеристик текстов (различные индексы лексического разнообразия, индексы исключительности и концентрации). Исследуется соотношение в тексте и словаре текста между разными частями речи, в т.ч. между группой знаменательных частей речи и группой служебных частей речи. Это определяется как важный стиледифференцирующий критерий при сравнении текстов разных авторов.

Важным представляется сопоставление словарного состава текстов разных украинских авторов. Выявляется общая часть словников исследуемых произведений и проводится сопоставительный анализ каждой берущейся пары произведений.

Сравнение текстов позволило выявить разные группы лекси-

ки, в частности, слова, разница в средней частоте которых несущественна во всех произведениях, и слова с существенными отличиями по этому параметру. Выявлены некоторые статистические параметры функционирования лексики, которые представлены индексом лексической связи между анализируемыми текстами и коэффициентом лексической близости текстов. Показано,что и высокочастотные слова при использовании особых статистических приемов способны различать авторские стили. На материале группы высокочастотных слов выясняется, что среди факторов, от которых зависит частота слов, находятся следующие:особенность композиционного строения произведения, тематика, авторский стиль. Т.о. опровергается мнение, что показатели авторского стиля бывают только эмоционально окрашенные, диалектные и редкие слова.

В целом, можно оценить работу авторов рецензируемой монографии как выполненную на высоком научном уровне, затрагивающую очень важный круг теоретических и методических проблем современной квантитативной лингвистики и общего языкознания.

К числу пожеланий можно отнести то, что весомость, доказательность и универсальность делаемых в работе выводов по квантитативному анализу текстов (в т.ч. и в стилистическом аспекте) несомненно может быть усилена за счет привлечения в будущем к такого рода анализу и материала других языков, кроме украинского.

Книга может быть использована как руководство для составления и анализа частотных словарей разных типов, разных языков В т.ч. может использоваться и как учебное пособие в курсах по квантитативному анализу языка и стиля.

А.А.Поликарпов, Ю.А.Тулдава

FREQUENCY DICTIONARIES AND THE EXPERIENCE OF THEIR USE

Anatoliy Polikarpev, Juhan Tuldava

S u m m a r y

In the collective monography by V.I.Perebeinos,M.P.Muravitskaya, N.P.Darčuk the problems of compiling,analysing and practical use of frequency dictionaries have been examined. The analysis has been made on the material of the Ukrainian language. The main topics of the investigation are connected with a statistical analysis of grammar, semantics and stylistic features of text and vocabulary.

СОДЕРЖАНИЕ

Рецензия

SUMMARIES - RESÜMEES

Review

Ученые записки Тартуского государственного университета.
Выпуск 774.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ 1987.
На русском языке.
Резюме на разных языках.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор Ю. Тулдава.
Подписано к печати 05.02.1987.
МВ 01542.
Формат 60x90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 10,76. Печатных листов 11+1 вкл.
Тираж 550.
Заказ № 63.
Цена 1 руб. 50 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.