ISSN 0494-7304 0207-4435

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED

УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

745

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1986

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1986

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 745 ВЫПУСК ОСНОВАНЫ В 1893.г.

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

1986

QUANTITATIVE LINGUISTICS AND AUTOMATIC TEXT ANALYSIS

TARTU 1986

Käesolevas kogumiku "Kvantitatiivlingvistika ja tekstide automaatanalüüs" teises väljaandes on avaldatud Tartu Riikliku Ülikooli rakenduslingvistika uurimisgrupi liikmete ja väliskaastöötajate artiklid. Kogumik jätkab sarja "Töid keelestatistika alalt", mida ilmus 10 väljaannet (1976-1984).

В настоящем, втором выпуске сборника "Квантитативная лингвистика и автоматический анализ текстов" опубликованы статьи сотрудников Группы прикладной лингвистики Тартуского государственного университета и исследователей из других городов. Сборник является продолжением серии "Трудов по лингвостатистике" (10 выпусков в период 1976-1984 гг.).

This second issue of "Quantitative Linguistics and Automatic Text Analysis" consists of papers by members of the Research Group of Applied Linguistics at Tartu State University and guest authors. The collections continue the series "Papers on Linguo-Statistics" (published serially as issues of Acta et Commentationes Universitatis Tartuensis 1976 - 1984).

РАСПРЕДЕЛЕНИЯ ЛЕКСИЧЕСКИХ ЕДИНИЦ ПО ДЛИНЕ В ТЕКСТЕ И СЛОВАРЕ

П.М.Алексеев

Зная одну количественную характеристику лингвистического явления, например длину текстового слова в буквах, нельзя точно установить другую, например частоту слова известной длины. До наблюдения самые общие представления о связи между этими характеристиками получают из материалов предшествовавшего наблюдения, либо аналогичного данному, либо такого, в котором эти характеристики нашли то или иное отражение.

В случае с длинами и частотами слов (или других лингвистических единиц) можно исходить из "принципа наименьшего усилия" или "принципа экономии в языке" и полагать, что в общем чем слова короче, тем они чаще употребляются, и чем слова чаще употребляются, тем они короче. Однако сами эти принципы не могли бы быть в свое время сформулированы без неоднократных наблюдений языковых единиц в речи.

Некоторая связь между длиной и частотой слова становится очевидной уже при первом знакомстве с частотными словарями любого языка или подъязыка. Правда, там видим, что в число самых частых попадают, хотя и понемногу, слова более длинные, чем соседние, а самые короткие иногда оказываются не самыми частыми.

Не удается получать точные соответствия и тогда, когда рассматривают попарно такие характеристики, как частота слова в тексте и его частотный ранг в ЧС этого текста, как частота слова в тексте фиксированного размера и количество текстов, содержащих это слово с данной частотой, как частота слова в тексте и количество слов с данной частотой и другие.

Важность численной меры в лингвистике стала общепризнанной и не требует дополнительных аргументов. Известно также, что в каждом конкретном наблюдении величина этой меры меняется существенно или несущественно в силу и лингвистических, и нелингвистических причин. В задачи лингвостатистики как раз и входит такое наблюдение, которое, с одной стороны, позволяет внутренним свойствам системы и нормы языка, узуса и речи проявиться в унифицированных внешних условиях и, с другой стороны, при смене этих условий показывает, как они влияют на количественное отображение внутренних свойств лингвистического объекта.

Поэтому, чтобы хоть как-то судить о связи между количественным (или качественным) выражением лингвистического признака и его частотой, приходится рассматривать весь ряд частот этого признака, причем делать это много раз и на большом и разнообразном материале.

Несколько сходные проблемы имеют место в квантовой механике, где их решение соотносится с принципом неопределенности: некоторые количественные характеристики не могут быть одновременно с произвольной степенью точности установлены для данного момента и поэтому могут быть выражены только через распределения их вероятностей.

Отсюда следует важный для лингвостатистики вывод о неизбежности обращения к анализу распределений количественных характеристик лингвистических явлений. Не менее важным представляется взгляд на распределение как на количественную меру упорядоченности сложного, системного лингвистического объекта, образуемого элементами, которые сами по себе или группируясь в классы обладают различной структурной или функциональной значимостью, весом в системе (разумеется, если эту значимость удается представить численно).

Можно согласиться с пониманием распределения как модели вероятностной лингвистической системы (Тулдава, с.139), однако сегодняшний уровень работ по изучению лингвистических распределений вынуждает к более скромному определению. По-видимому, пока что распределение целесообразно понимать как количественную (не обязательно сразу вероятностную) модель лингвистической системы (опять же не обязательно сразу вероятностной). Это уточнение имеет смысл хотя бы потому, что сегодня в лингвостатистике слишком много говорится о вероятностных системах, закономерностях, распределениях, вообще о вероятностях, хотя имеют в виду по преимуществу лишь частотные закономерности, частотные распределения и оперируют пока еще частотами, получаемыми из умеренных по объему наблюдений.

В математической статистике распределением считают перечисление возможных значений случайной величины и их вероятностей, а правило, связывающее значения случайной величины и соответствующие им вероятности, называют законом распределения случайной величины. Ряд пар значений случайной величины и их вероятностей образует вариационный ряд или, по-другому, ряд распределения (Пиотровский, Бектаев, Пиотровская, 1977, с.167,

223). Можно использовать более простую формулировку и понимать под распределением перечисление значений лингвистического признака и частот (численностей) соответствующих значений. Такое определение больше соответствовало бы ситуации с наблюдением частот.

Лингвостатистик строит и изучает частотные вариационные ряды, подбирает к ним какое-либо из стандартных распределений, но нередко забывает, что если эмпирический ряд удается аппроксимировать законом, известным из теории вероятностей, то это не обязательно равнозначно переходу к вероятностям лингвистического явления. Между теоретическим, сглаживающим выборочным распределением частот и генеральным распределением "истинных" частот, т.е. вероятностей, возможна слишком большая разница, пренебречь которой было бы недопустимо. Поэтому наряду с подбором теоретических распределений к лингвистическим эмпирическим рядам следовало бы обращать внимание на то, как эти ряды получены, на различия в исходных материалах для каждого ряда, даже если эти материалы кажутся единообразными, на способы регистрации одних и тех же явлений в разных наблюдениях. Не очень удачные обобщения, излишне смелые экстраполяции ограниченных фактографических данных на генеральные совокупности как раз и возникают из-за недостаточного внимания к начальным этапам лингвостатистического описания, из-за желания поскорее придать выборочным числам значимость генеральных.

Прежде чем переходить к вероятностным обобщениям, не обойтись без рассмотрения как можно большего количества конкретных частотных рядов, без их систематизации, классификации. Уже предприняты первые попытки таких классификаций (Мартыненко, 1982; Тулдава, 1982; Алексеев, 1978, 1985).

Классическим случаем распределения количественного признака по частоте стало распределение длины лексической единицы в тексте или словаре этого текста. Объектом здесь является словоупотребление текста или словоформа словаря текста. Признаком является длина словоупотребления или словоформы, измеряемая буквами. Цифровые значения длины в буквах – это варианты, а количества словоупотреблений в тексте или словоформ в словаре для каждой длины – это частоты вариант. Простота наблюдения над таким признаком сделала его хрестоматийным примером, иллюстрирующим возможности количественных оценок в информационных измерениях словаря и текста, в стилометрии, типологии

языков и типологии текста. Его элементарность не значит, что оно дает мало интересных сведений об описываемом явлении. Более того, длина лингвистической единицы связывается с возможностями оперативной памяти человека, а расхождения в средних длинах словоупотреблений отражают типологические особенности конкретных языков, подъязыков, идиолектов (Пиотровский, Бектаев, Пиотровская, 1977, с.262-265). Длина лингвистической единицы учитывается при построении систем искусственного интеллекта, в организации памяти обучающего лингвистического автомата. Необходимы достоверные и точные сведения о распределениях лингвистических единиц, притом обязательно в двух аспектах - в текстовом и словарном. Ограничение анализа распределений только текстовыми частотами обедняет наше представление о количественной организации лингвистической системы. Не всегда подчеркивается принципиальное различие в количественном представлении на оси текста и на оси словаря; ниже будут проиллюстрированы такие различия.

Что касается распределений длин лексических единиц на словарной оси, то, как правило, данные для них извлекаются либо из "обычных", не частотных словарей и весьма редко - из частотных: в обоих случаях рассматриваются длины слов-лексем, а не словоформ. Это может быть оправдано только если считать, во-первых, что различия между длинами словоформ и слов несущественны (что справедливо лишь для языков с высокой степенью аналитизма) или, во-вторых, что в памяти человека или компьютера должны храниться не словоформы, а слова-лексемы. Если же допустить, что она содержит ни то и ни другое, но более короткие единицы, скажем основы или квазиосновы, то аргумент в пользу лексем все равно придется отвергнуть. Таким образом, количественная информация о словоформах (и словоупотреблениях) с лингвистической и лингвостатистической точек зрения более полезна, чем о словах-лексемах, кроме, возможно, отдельных случаев стилеметрии.

Вариационный ряд длина-частота слова относится по стандартной статистической классификации к распределениям количественного признака, в отличие от признаков качественного или качественного, преобразованного в порядковый (Пиотровский, Бектаев, Пиотровская, 1977, с.222 и след.; Урбах, 1964, с.II и след.). Ю.А.Тулдава считает его многообъектным "на том основании, что исчисляются разные объекты - классы слов (слова

разной длины)" (Тулдава, 1982, с.136). Однако в качестве объекта в этом случае выступает обобщенная единица одного лингвистического уровня – словоупотребление или словоформа, т.е. объект один, и признак один, а значений признака, вариант – несколько. По Г.Я.Мартыненко это распределение – многопредметное (многообъектное по Ю.А.Тулдава), структурное и двухвершинное (Мартыненко, 1982, с.117). "Структурность" противопоставляется "статусности"; первая соотносится со "строением", вторая с "поведением" слова; неясность возникает как раз из-за неразличения текстового и словарного аспектов одной и той же единицы. Тезис о наличии двух вершин в распределении слов по длине требует проверки, что и будет сделано ниже.

Предварительные рассуждения о лингвистических распределениях этого вида понадобились потому, что несмотря на их большую популярность в лингвостатистике по-прежнему остаются в тени важные вопросы их изучения, а выводы обычно основываются на незначительных по объему выборках. Далее в настоящей статье будут рассмотрены эти распределения на материале нескольких частотных словарей словоформ, составленных по выборкам от 50 тыс. до 1 млн. словоупотреблений. За исключением одного случая, оговоренного особо, длины и частоты словоформ подсчитывались по данным ЧС самим автором статьи.

Итак, распределения лексических единиц по длине естественно рассматривать там, где эти единицы фактически употребляются, то есть в самом тексте. Анализ длин словоупотреблений текста следует дополнять анализом длин словоформ в словаре этого же текста. Результаты будут надежнее, если обследовать большие по объему текстовые выборки. Таковы три условия, которым должно отвечать более или менее серьезное исследование длин лексических единиц. Удобным материалом, чем-то вроде "полуфабриката", могут послужить существующие ЧС словоформ, хотя здесь встретятся затруднения. Во-первых, если словоформы-омографы входят в ЧС раздельно, требуются определенные усилия и немалое внимание для укрупнения омографических групп и для объединения частот в таких группах, особенно если омографы рассеяны по разным парадигмам-статьям словаря и разным частотным зонам. Во-вторых, подсчеты текстовых частот и словарной "активности" длин даже по готовому ЧС отнимают много времени и достаточно трудоемки, когда приходится иметь дело не с одним, а с многими ЧС. В-третьих, и это может смутить даже то-

го, кто решил не пожалеть времени и усилий на подобную работу, большинство составленных ЧС опубликованы и поэтому доступны лишь в виде списков очень небольшого числа самых частых единиц – обычно не более I тыс., а иногда и 0,5 тыс. А это вызывает вопрос о том, в достаточной ли мере репрезентативен такой усеченный список для всего ЧС.

Можно начать исследование длин словоупотреблений и словоформ с попытки ответить на этот вопрос, для чего нужно, по-видимому, рассмотреть один и тот же ЧС целиком, затем его "верхнюю" зону и далее – ту часть, которая остается за вычетом верхней зоны. Если использовать при этом несколько ЧС одного языка, то выводы могут оказаться действительными хотя бы для этого языка и быть принятыми с некоторыми допущениями и относительно других языков.

Три ЧС английского языка содержат полные списки всех зарегистрированных в выборках словоформ и отражают употребление этих единиц в контрастных сферах языкового функционирования: в письменно-литературной речи (Kučera, Francis , I967)+, устной речи (Howes , I966) и в научно-технических текстах по электронике (составлен автором статьи). Первый базируется на выборке в I млн., второй – 250 тыс. и третий – 200 тыс. словоупотреблений. Каждый из этих ЧС рассматривается трижды – вначале целиком, затем в своей верхней зоне и, наконец, в оставшейся зоне. Раздельно строятся ряды распределения длин единиц ЧС по их частоте в тексте и по их словарной активности. Ради экономии места ниже будут представлены только графики; чтобы сделать наглядными тенденции в распределениях, их надо привести к одному масштабу – с относительными значениями частот и активности. Чтобы сохранить при этом единую размерность, относительные величины каждый раз получаются не от полного объема одной и той же выборки и одного и того же словаря этой выборки, но от суммы частот в зоне и от суммы разных единиц в этой зоне. Все эти приемы станут яснее из рассмотрения наименее громоздких в нашем случае таблиц, таблиц распределения длин словоупотреблений и словоформ в трех зонах ЧС английской устной речи (табл.I-3).

В Табл.I (весь ЧС) относительные частоты получены "обыч-

+В этом ЧС приводятся ряды распределения длин словоупотреблений и словоформ.

Таблица I

Распределения длин словоупотреблений в тексте и словоформ в словаре(английская устная речь,весь частотный словарь)+

l	Текст		Словарь	
	F	f %	M	m %
I	I6099	6,43	I7	0,I8
2	437I0	I7,45	60	0,62
3	60I08	24,00	289	2,98
4	56508	22,56	959	9,89
5	267I6	I0,67	I362	I4,04
6	I7806	7,II	I58I	I6,30
7	I3663	5,46	I590	I6,39
8	6894	2,75	I305	I3,45
9	3929	I,57	983	I0,I4
I0	2955	I,I8	703	7,25
II	II24	0,45	405	4,I8
I2	575	0,23	229	2,36
I3	I96	0,08	I0I	I,04
I4	II4	0,05	62	0,64
I5	38	0,02	28	0,29
I6	I6	0,0I	I2	0,I2
I7	9		9	0,09
I8	I		I	0,0I
I9	I		I	0,0I
20	2		I	0,0I
2I	I		I	0,0I
Итого:	250465	I00	9699	I00

ным" путем, т.е. делением абсолютных частот на объем выборки, а относительная активность – делением абсолютных значений на объем всего ЧС. В Табл.2 (верхняя зона) абсолютные частоты поделены на сумму частот в верхней зоне, абсолютные величины

+Здесь и далее l – длина словоформ (словоупотреблений), измеряемая буквами, F – абсолютная частота словоформ данной длины в тексте, M – количество разных словоформ данной длины в словаре, f и m – соответственно относительные величины текстовой частоты и словарной активности.

Таблица 2

Распределения длин словоупотреблений в тексте и словоформ в словаре (английская устная речь, верхняя зона частотного словаря)

l	Текст		Словарь	
	F	f %	M	m %
I	I6060	7,II	2	0,I9
2	43649	I9,42	25	2,38
3	59420	26,44	I07	I0,20
4	536I3	23,86	252	24,02
5	22562	I0,04	2I6	20,59
6	I3338	5,93	I65	I5,73
7	9494	4,24	I40	I3,35
8	3530	I,57	75	7,I5
9	I634	0,73	33	3,I5
I0	II34	0,50	23	2,I9
II	200	0,09	8	0,76
I2	8I		2	0,I9
I3	26		I	0,I0
Итого:	224741	I00	I049	I00

словарной активности поделены на число словоформ в этой зоне. Точно так же получены относительные величины для Табл.3 (оставшаяся часть словаря).

На рис.I приводятся все три пары рядов распределений; каждая пара для одной из зон ЧС представляет два ряда - длины словоупотреблений и длины словоформ.

Из данных Табл.I-3 и графиков на Рис.I очевидны, во-первых, явные различия в распределениях на текстовой и словарной осях (распределения I, Ia и 2, 2a), во-вторых, отсутствие заметных различий между текстовыми распределениями в верхней зоне и в полном ЧС и, в-третьих, отсутствие заметных различий между распределениями словарными в полном ЧС и оставшейся его части и текстовым в этой же части.

Первый результат лишь еще раз подтверждает, насколько важно рассматривать одну и ту же количественную характеристику лингвистической единицы (в данном случае длину) раздельно в текстовом и словарном аспектах. При всей кажущейся тривиаль-

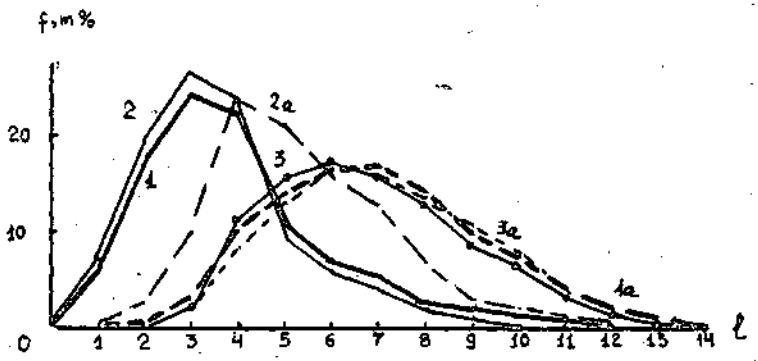


Рис.I. Распределения длин словоупотреблений в тексте и длин словоформ в словаре (английская устная речь). I – текст, Iа – словарь (весь ЧС); 2 – текст, 2а – словарь (верхняя зона); 3 – текст, 3а – словарь (оставшаяся часть ЧС).

Таблица 3

Распределения длин словоупотреблений в тексте и словоформ в словаре (английская устная речь, оставшаяся часть частотного словаря)

l	Текст		Словарь	
	F	f %	M	m %
I	39	0,I5	I5	0,I7
2	6I	0,27	35	0,40
3	688	2,67	I82	2,I0
4	2895	II,25	707	8,I7
5	4I54	I6,I5	II46	I3,25
6	4468	I7,37	I4I6	I6,37
7	4I69	I6,2I	I450	I6,76
8	3364	I3,08	I230	I4,22
9	2295	8,92	950	I0,98
I0	I82I	7,I8	680	7,86
II	924	3,59	397	4,59
I2	494	I,92	227	2,62
I3	I70	0,66	I00	I,I6
I4	II4	0,44	62	0,72
I5	38	0,I5	28	0,32
I6	I6	0,06	I2	0,I4
I7	9	0,03	9	0,I0
I8	I			
I9	I			
20	2			
2I	I			
Итого:	25724	I00	8650	I00

ности это соображение не всегда учитывается в лингвостатистических исследованиях. Второй результат уже менее тривиален и свидетельствует в пользу высказывавшегося предположения о том, что верхняя зона ЧС содержит основные лингвостатистические сведения о морфологическом строе языка. Остается добавить, что если длина текстовых словоупотреблений как раз и отражает морфологические свойства языка, то анализа верхней зоны ЧС может оказаться достаточным, чтобы получить общее, пусть и графическое, представление о его квантитативно-морфологической типологии. Третий результат выглядит вовсе неожиданным: оказывается, что если вычесть верхнюю зону ЧС, то морфологическая структура лексических единиц, отражаемая их длинами, практически совпадает на текстовой и словарной осях, а они обе совпадают со словарным представлением в рамках всего ЧС. Некоторая лингвопсихологическая интерпретация этих результатов будет сделана ниже, а сейчас предстоит обратиться к аналогичным графикам, построенным по материалам ЧС письменно-литературной и научно-технической форм английского языка (Рис.2-3).

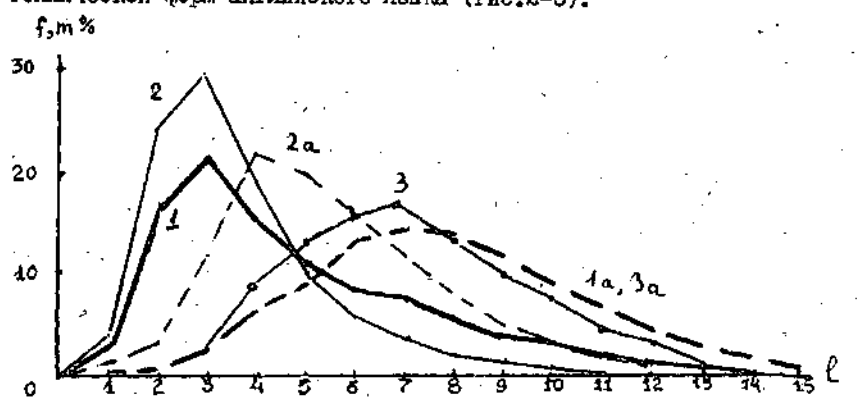


Рис.2. Распределения длин словоупотреблений в тексте и длин словоформ в словаре (английская письменно-литературная речь). I - текст, Ia - словарь (весь ЧС); 2 - текст,2a - словарь (I-я тыс. словоформ); 3 - текст; 3a - словарь (оставшаяся часть ЧС).

На Рис.2 графики Ia и 3a полностью совпадают.

Первая тысяча самых частых словоформ словаря распределена по длине в устной английской речи точно так же, как и в письменно-литературной речи. Научно-технический подъязык(электро-

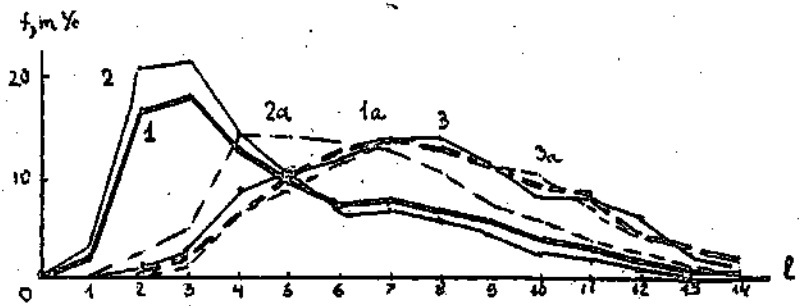


Рис.3. Распределения длин словоупотреблений в тексте и длин словоформ в словаре (тексты по электронике на английском языке). I - текст, Ia - словарь (весь ЧС); 2 - текст, 2а - словарь (I-я тыс. словоформ); 3 - текст, 3а - словарь (оставшаяся часть ЧС).

ника) нарушает это единообразие, поскольку в число самых частых единиц его ЧС неизбежно попадают тематические и, следовательно, более длинные, чем общеупотребительные. Замеченные свойства первых двух словарей следует иметь в виду при организации памяти компьютера, обучающего, информационного, а также автомата, воспринимающего и порождающего естественный текст. Чтобы проверить это наблюдение на другом, сходном материале, можно рассмотреть еще один ЧС английской устной речи (Dahl , 1979). Вывод о высокой степени близости в распределениях самых частых словоформ по длине на оси словаря подтверждается графиками на Рис.4.

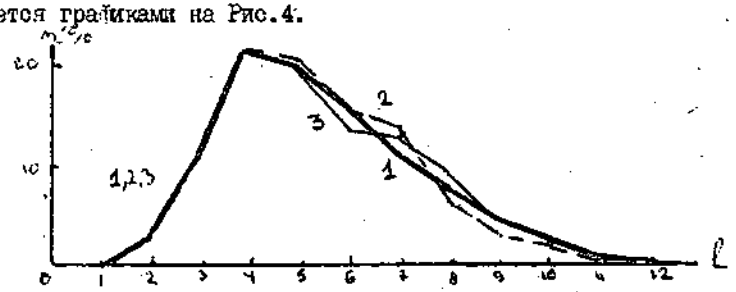


Рис.4. Распределения длин словоформ в словаре (I-я тыс.).I - письменно-литературная речь, 2 - устная речь (Howes), 3 - устная речь (Dahl). Объемы выборок равны I млн., 260 тыс. и I млн. словоупотреблений соответственно.

При сопоставлении графиков распределений словоформ словаря по длине в разных зонах ЧС обнаруживается еще одно любопытное свойство таких распределений. На Рис.5 представлены словарные распределения на материале ЧС польской драмы (Kurcz et al. , 1977). Обнаруживается практически полное совпадение процентных распределений во всем ЧС, в его зоне, оставшейся после отсечения верхней части, и в зоне с частотами, равными 1. Резко отличается от них, как и следовало ожидать, зона самых частых словоформ. Их в данном ЧС 975, и они встретились в текстах длиной 100 тыс. словоупотреблений не менее 10 раз каждая.

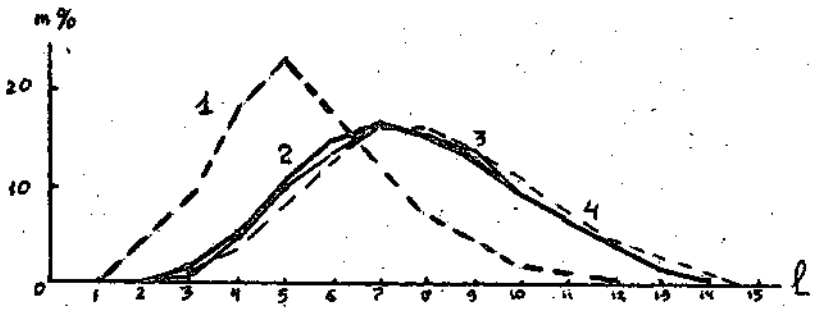


Рис.5. Распределения длин словоформ в словаре польской драмы. 1 - верхняя зона, 2 - весь ЧС, 3 - зона, оставшаяся за вычетом верхней, 4 - словоформы с частотой 1.

Отсюда очевиден вывод о том, что на структуру ЧС, отраженную длинами входящих в него словоформ, отсечение верхней зоны со сравнительно небольшим (до 1 тыс.) числом единиц, существенного влияния не оказывает. Вклад верхней и нижней зон ЧС в общую конфигурацию распределения его единиц по длине более заметен на текстовой оси (см.Рис.6).

На Рис.7 можно видеть распределения словоупотреблений по длине по данным четырех ЧС польского языка (Kurcz et al., 1974а, 1974 , 1976, 1977) в пределах верхней зоны каждого из них, а также сводный график, объединяющий частоты длин по всем четырем ЧС: верхняя зона первого плюс верхняя зона второго и т.д.

Здесь явствует четкое различие между текстами художественными (драма и художественная проза) и информационными (газета и научно-популярная литература), а внутри этих групп раз-

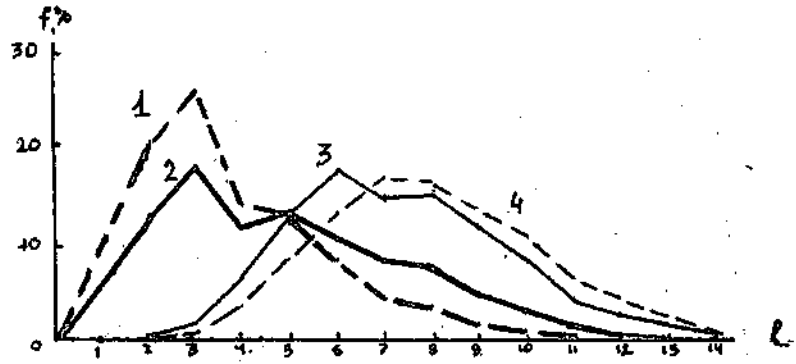


Рис.6. Распределения длин словоупотреблений в тексте польской драмы. I - верхняя зона, 2 - весь ЧС, 3 - зона, остающаяся за вычетом верхней, 4 - словоформы с частотой I.

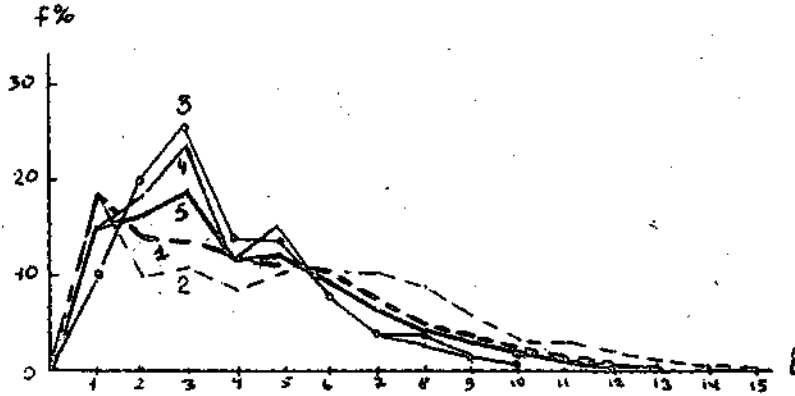


Рис.7. Распределения длин словоупотреблений в польских текстах по данным четырех ЧС. Объем выборки для каждого ЧС равен 100 тыс. словоупотреблений. I - научно-популярные тексты, 2 - мелкие газетные сообщения, 3 - драма, 4 - художественная проза, 5 - суммирующее распределение.

личия менее резки. Суммарное распределение естественным образом усредняет конфигурацию этих двух групп распределений. Не столь яркая картина различий просматривается на словарной оси,

однако и здесь очевидна близость распределений в словарях драмы и художественной прозы (Рис.8).

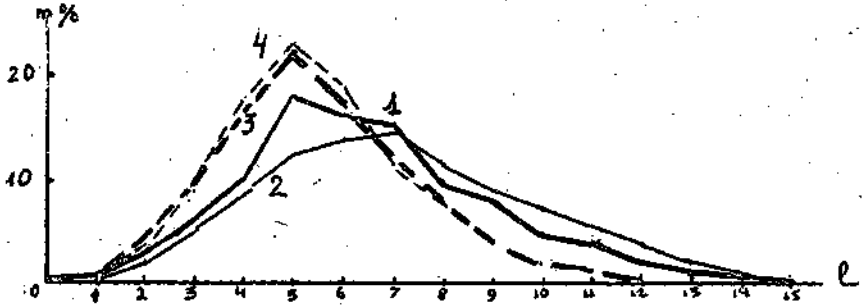


Рис.8. Распределения длин словоформ верхних зон в словарях четырех польских подъязыков. I – научно-популярные тексты, 2 – мелкие газетные сообщения, 3 – драма, 4 – художественная проза.

Материалы ЧС позволяют оценить тезис о непременном наличии двух вершин в текстовых распределениях (Мартыненко, 1980, с.116-117), а также о том, что двухвершинность относится по крайней мере к текстам на славянских языках+. Рисунки 9-10 представляют распределения на текстовой и словарной осях по данным ЧС рефератов по электроизмерительным приборам на русском языке++.

Здесь, во-первых, опять-таки очевиден некоторый изоморфизм текстовых распределений в паре верхняя зона – весь ЧС и в паре остаток ЧС – зона с частотой I. На словарной оси заметно выделяется верхняя зона, а остальные три распределения практически совпадают одно с другими. Во-вторых, текстовые распределения всего ЧС и верхней зоны имеют пилообразную форму с несколькими вершинами разной высоты, среди которых выделяются частоты длин в I, 5, 7 и 9 букв. Двухвершинность как будто сменяется многовершинностью, а это исключает рекомендацию аппроксимировать такой график двумя гауссовыми кривыми (ср. Мар-

---

+Это уточнение высказала В.И.Перебейнос на защите одной из диссертаций в ЛГУ им.А.А.Жданова в 1982 г.

++Объем выборки 105 тыс. словоупотреблений (Частотный словарь индексирования, 1974).

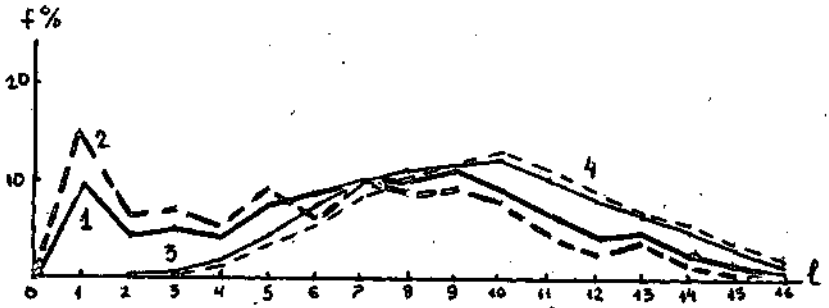


Рис.9. Распределения длин словоупотреблений в рефератах по электроизмерительным приборам на русском языке. I – весь ЧС, 2 – верхняя зона (I-я тыс.словоформ), 3 – оставшаяся зона, 4 – словоформы с частотой I.

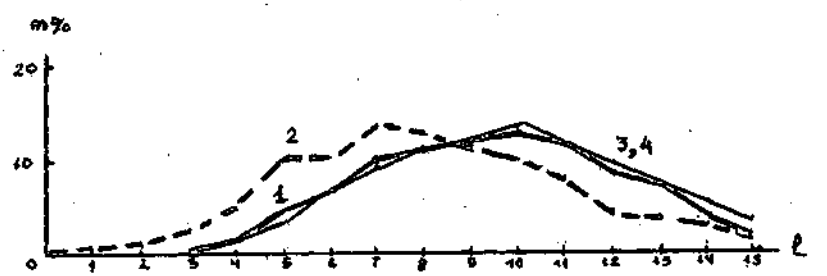


Рис.I0. Распределения длин словоформ в словаре по данным ЧС рефератов по электроизмерительным приборам на русском языке. I – весь ЧС, 2 – верхняя зона, 3 – оставшаяся зона, 4 – словоформы с частотой I.

тыненко, I980, с.II7). Чтобы проверить, не влияет ли на вид распределения то, что тексты, по которым составлен ЧС, написаны в жанре реферата, а не статьи или монографии, обратимся к другим ЧС русского языка. На рис.II-I2 представлены распределения в трех научно-технических подъязыках: электроизмерительных приборов (Частотный словарь индексирования, I974), электроники (Калинина, I968), химии полимеров (Садчикова,

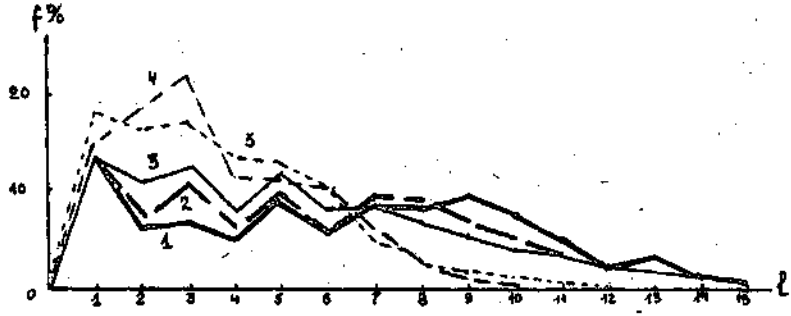


Рис.II. Распределения длин словоупотреблений в тексте по данным ЧС русского языка. I - электроизмерительные приборы; I-я тыс. словоформ; 2 - химия полимеров, I-я тыс. словоформ; 3 - электроника, I-я тыс.словоформ; 4 - устная речь, I-я тыс.словоформ; 5 - эпистолярная речь, I-я тыс.словоформ.

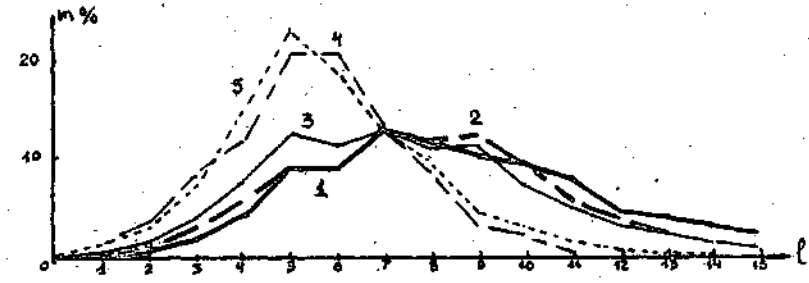


Рис.I2. Распределения длин словоформ в словаре по данным ЧС русского языка. I - электроизмерительные приборы, I-я тыс. словоформ; 2 - химия полимеров, I-я тыс.словоформ; 3 - электроника, I-я тыс.словоформ; 4 - устная речь; I-я тыс. словоформ; 5 - эпистолярная речь; I-я тыс.словоформ.

I974); в качестве контрастных привлечены данные о записях устной речи (Турко, I968) и об эпистолярной речи (Григорьева, I980)+.

+Объемы выборок соответственно равны I05, I00, 200, 50 и I00 тыс. словоупотреблений.

6. Объем выборки равен 500 тыс. словоупотреблений.

Наличие более чем одной вершины в распределениях словоупотреблений действительно имеет место в текстах на русском языке, и это относится с очевидностью к научно-техническим подъязыкам. Устная речь характеризуется одной вершиной, а эпистолярная речь, объединяющая в себе свойства устной и письменной форм языка и являющаяся чем-то вроде "письменно-разговорной" формы, занимает особое положение. Наибольшая концентрация частот, как и в устной речи, здесь приходится на словоупотребления длиной I–3 буквы; модальную частоту имеет длина I. Но, как в письменной речи, имеется тенденция к большему, чем одно, числу вершин. Можно предположить, что увеличение числа вершин типично для русской письменной речи, и это зависит от роста средних длин словоупотреблений (см. сводные данные о средних длинах в Табл.4). Пиковые значения частот в тексте приходятся на длины I, 3, 5, 7, иногда 9, т.е. на нечетные количества букв в словоупотреблении, и это последнее наблюдение относится также к словоформам на оси словаря. Польские тексты этой тенденции не проявили, кроме ЧС художественной прозы (см.Рис.7), поэтому желательно обратиться к какому-либо ЧС для другого славянского языка. Данные о словоформах и словоупотреблениях в аналогичных текстах имеются лишь в ЧС украинской художественной прозы (Частотный словник..., I98I)+. На Рис.I3 приведены графики распределений в тексте и словаре по данным этого ЧС.

Здесь очевидно наличие более чем одной вершины, а именно двух, в украинском художественно-прозаическом тексте, причем пики приходятся на четные значения частот длин в 2 и 4 буквы.

Практически убедившись в том, что верхняя зона ЧС представительна для всего ЧС тем, что она отражает общие соотношения в распределениях длин текстовых словоупотреблений, можно рассмотреть материалы других ЧС, других языков и подъязыков, теперь уже только на уровне самых частых словоформ. На Рис.I4 приводятся графики распределений длин словоупотреблений в текстах на английском (Турыгина, I968), испанском (материалы Лаборатории инженерной лингвистики ЛГПИ им.А.И.Герцена), французском (Исенин, I968), казахском (Бектаев, I975)++

+Объем выборки равен 500 тыс.словоупотреблений.

++Объемы выборок равны соответственно I00, I70, I20 и I50 тыс.словоупотреблений. Для расчетов использовались первые 500 словоформ каждого ЧС.

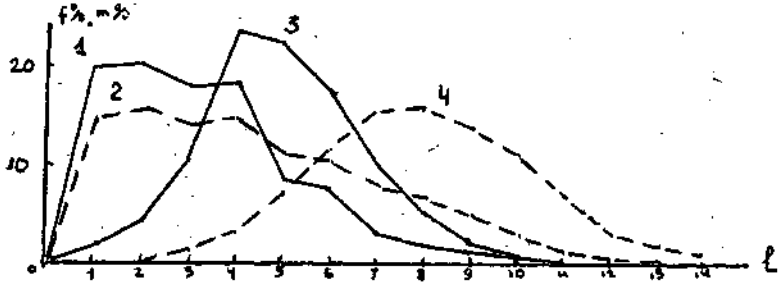


Рис.I3. Распределения длин словоупотреблений в тексте и словоформ в словаре украинской художественной прозы. I - текст, I-я тыс. словоформ; 2 - текст, весь ЧС; 3 - словарь, I-я тыс. словоформ; 4 - словарь, весь ЧС.

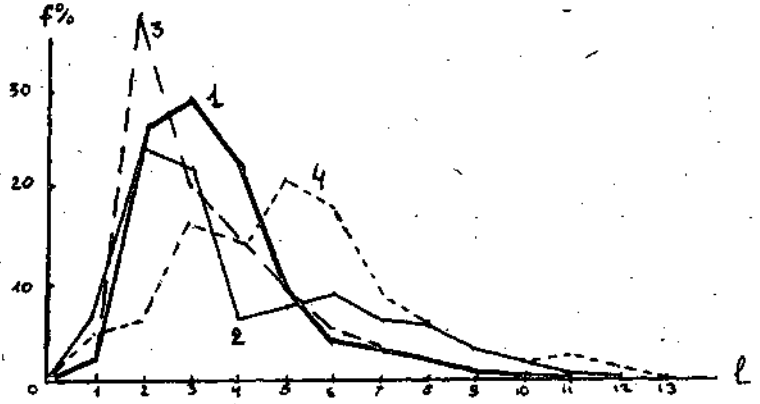


Рис.I4. Распределения длин словоупотреблений в газетных текстах по данным верхних зон ЧС. I - английский, 2 - испанский, 3 - французский, 4 - казахский языки.

Рассматривая длины словоупотреблений во французском тексте, следует учитывать, что составители использованных для настоящей статьи ЧС французских текстов выделяли записываемые через апостроф компоненты текстовой единицы и вносили их в списки как самостоятельные словоформы. Именно наличием таких высокочастотных единиц, как l',d',s',n' и др., определявшихся нами как двухбуквенные (апостроф считался буквой), объясняется резкий скачок частоты для длины 2.

Завершим рассмотрение распределений этого типа графиками на Рис.I5, которые представляют данные ЧС французского (Кочеткова, Скрелина, I968), румынского (Еван, I968), немецкого (Зореф, I97I), английского (данные автора настоящей статьи) и русского (Калинина, I968) подъязыков электроники+.

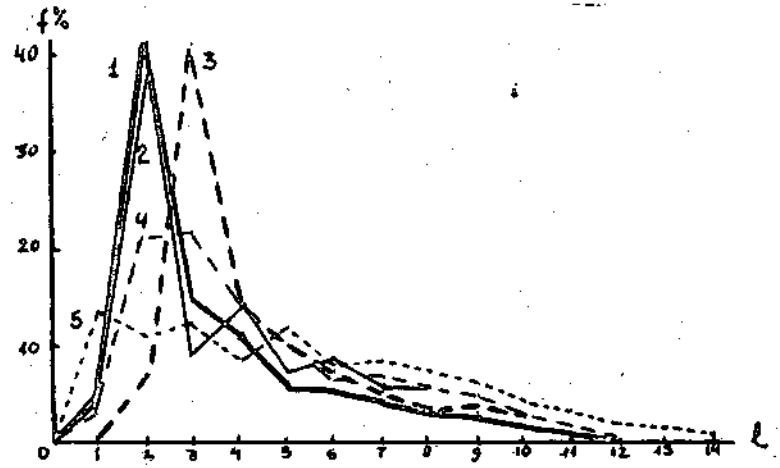


Рис.I5. Распределения длин словоупотреблений в текстах по электронике. I - французский язык, I-е 500 словоформ; 2 - румынский язык, I-е 500 словоформ; 3 - немецкий язык, I-я тыс. словоформ; 4 - английский язык, I-я тыс. словоформ; 5 - русский язык, I-я тыс. словоформ.

Наблюдения над длинами и частотами словоформ - единиц ЧС могли бы под новым углом зрения представить "вечную" проблему описания и интерпретации ранговых распределений лексических единиц, а также вопросы, связанные с обоснованием закона Ципфа, с аппроксимацией эмпирических ранговых распределений, с оценкой и истолкованием параметров аппроксимирующих кривых. До сих пор почти не рассматривались по отдельности ранговые распределения словоформ ЧС, входящих в какие-либо классы, кроме, пожалуй, нескольких опытов с построением ранговых распределений отдельно для существительных, отдельно для глаголов и

+Объемы выборок равны соответственно I00 тыс., 200 тыс., 200 тыс., 200 тыс., 200 тыс. словоупотреблений.

т.д. или отдельно для терминов и нетерминов, делавшихся в диссертациях группы "Статистика речи" (см., в частности, Лебедев, 1979; Яблонская, 1980; Григорьева, 1981)+. Основным результатом этих опытов был вывод о том, что основной вклад в ранговое распределение всех словоформ ЧС делается служебными единицами (это отражается в верхней части билогарифмического графика ранг-частота) и существительными (это отражается в протяженности графика по оси рангов, поскольку в объеме словаря наибольшее место занимают существительные, что по мере снижения частоты становится все более заметным). Рассмотрение ранговых распределений лексико-грамматических классов словоформ несомненно выходит на уровень содержательной лексико-морфологической и лексико-семантической интерпретации. Однако не меньший интерес по-прежнему связывается с более формальными основаниями группировки единиц ЧС, и распределения ранг-частота единиц, объединенных признаком длины, являются материалом для формализованного подхода.

Разумеется, статистику для этого можно получить только из полного ЧС. На Рис.16 представлены по отдельности ранговые распределения словоформ в ЧС польской драмы, имеющих длину 1, 2, 3, 4, 5 и 9 букв, а также ранговое распределение всех словоформ ЧС. Последовательно строя графики таких распределений для одного ЧС можно, как кажется, увидеть, какой график изоморфен общему и, следовательно, определить роль словоформ данной длины в том, что кривая общего рангового распределения имеет конкретный вид.

По внешнему виду на общее распределение как будто больше всего походят распределения четырехбуквенных и пятибуквенных словоформ. Такой вывод можно уточнить, сравнивая аппроксимирующие линейные графики, построенные по выражению

$$\lg F_i = \lg k N - \gamma \lg i \ ,$$

где F_i – частота i-й единицы ЧС, i – частотный ранг этой единицы, k и γ – коэффициенты закона Ципфа, N – объем выборки; в случае словоформ длины l величина N равна сумме частот словоформ, имеющих данную длину. Линейные графики представлены на Рис.17. Вместо распределений для длин 1, 2, 3 и 9, которые резко отличаются от общего распределения, помещено распределение словоформ длиной 6 букв. Без дополнительных рас-

+Диссертации защищены в ЛГУ им.А.А.Жданова.

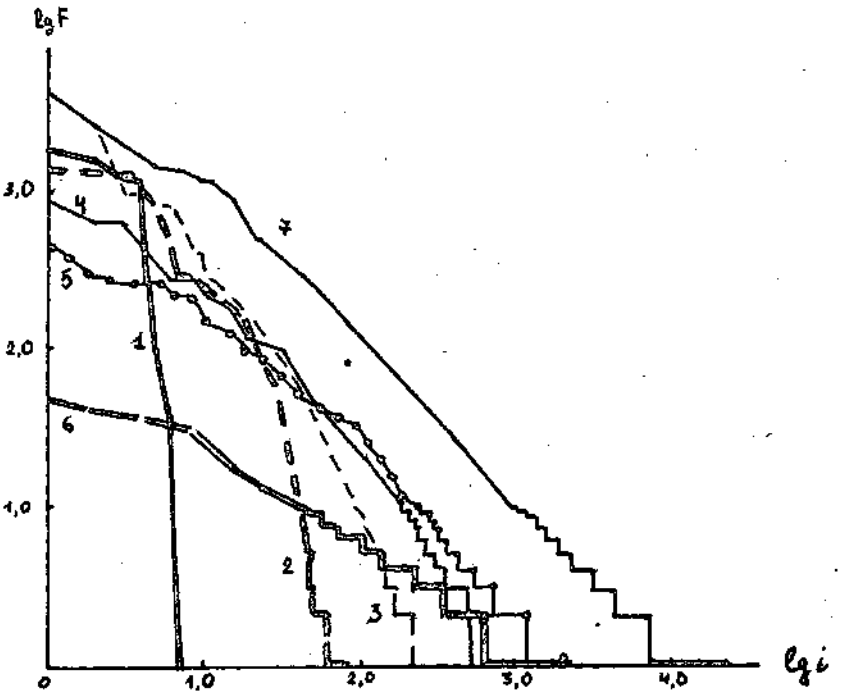


Рис.16. Ранговые распределения словоформ ЧС польской драмы. I – однобуквенные, 2 – двухбуквенные, 3 – трехбуквенные, 4 – четырехбуквенные, 5 – пятибуквенные, 6 – девятибуквенные словоформы, 7 – все словоформы ЧС. F – частота, i – ранг.

четов по оценке близости регрессий легко видеть наибольшее сходство общего графика с графиком для длины 5 букв.

Поскольку средняя длина словоупотребления в полном ЧС польской драмы равна 4,97 буквы, т.е. практически 5 буквам, это наблюдение можно расширить до вывода о том, что репрезентативным для рангового распределения всех словоформ ЧС является ранговое распределение словоформ длины, равной средней длине словоупотребления в соответствующем тексте.

В Табл.4 приведены данные о средних длинах словоупотреблений текста и словоформ словаря в различных зонах рассмотренных выше ЧС.

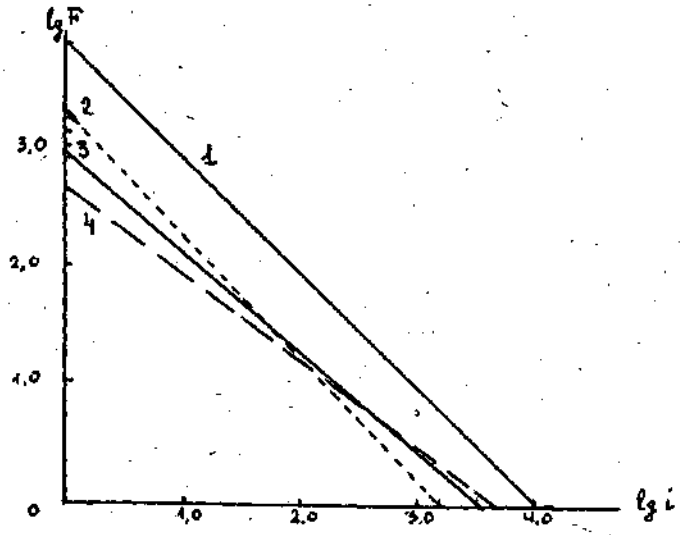


Рис. I7. Аппроксимирующие линейные графики ранговых распределений словоформ ЧС польской драмы. I – весь ЧС, 2 – четырехбуквенные, 3 – пятибуквенные, 4 – шестибуквенные словоформы.

Как будто единственный очевидный вывод из проделанного анализа состоит в том, что, как и следовало бы ожидать, распределения на текстовой и словарной осях различны в принципе. А это условие немаловажно для попытки связать среднюю длину слова с возможностями оперативной памяти человека: по-видимому, речь должна идти о длине не словоупотреблений текста, а словоформ словаря. Но средняя длина словарной словоформы резко меняется от языка к языку, а это значило бы, что и возможности памяти у носителей различных языков тоже различны. Дело все же, видимо, в длинах самых частых словоформ, а их средние действительно если не приближаются к какому-то межъязыковому стандарту, но и не превышают его. С некоторым допуском на отсутствие подтверждающих свидетельств можно предположить, что именно верхняя зона словоформ ЧС и их распределение по длинам моделируют оперативную память, тогда как менее частые словоформы следует соотносить с памятью долговременной.

В более общих терминах распределение длин словоупотреб-

Таблица 4

Средние длины словоупотреблений текста и словоформ словаря по данным частотных словарей

Язык, подъязык	Зона ЧС	Средняя длина		Язык, подъязык	Зона ЧС	Средняя длина	
		с/у	с/ф			с/у	с/ф
Польский					Весь ЧС	3,95	7,09
Научно-поп.	I тыс.	4,32	6,53	- " -	Ост.зона	6,89	7,29
Газета	I тыс.	5,08	7,20	- " -	F = I	7,66	7,66
Худ.проза	I тыс.	3,60	5,52	Устн.речь (Dahl)	I тыс.	3,62	5,56
Драма	I тыс.	3,80	5,53	Газета	0,5 тыс.	3,57	5,22
- " -	Весь ЧС	4,97	7,75	Электр-ка	I тыс.	4,57	6,69
- " -	Ост.зона	7,49	7,87	- " -	Весь ЧС	5,24	8,19
- " -	F = I	8,19	8,19	- " -	F = I	8,74	8,74
Русский				Французский			
Эл.изм.пр.	I тыс.	6,47	8,28	Газета	0,5 тыс.	3,56	5,67
- " -	Весь ЧС	7,57	I0,06	Электр-ка	0,5 тыс.	3,8I	6,55
- " -	Ост.зона	9,9I	I0,20	Румынский			
Химия пол.	I тыс.	5,80	7,98	Электр-ка	0,5 тыс.	4,I0	6,I4
Электр-ка	I тыс.	5,33	7,42	Испанский			
Устн.речь	I тыс.	3,70	5,72	Газета	2 тыс.	4,9I	7,36
Эпист.речь	I тыс.	3,75	5,63	- " -	I тыс.	4,60	6,87
Украинский				- " -	0,5 тыс.	4,24	6,47
Худ.проза	I тыс.	3,26	5,05	Немецкий			
- " -	Весь ЧС	4,82	8,36	Электр-ка	I тыс.	4,68	7,54
Английский				Казахский			
Устн.речь (Howes)	I тыс.	3,62	5,42	Газета	0,5	5,I0	5,89

лений в тексте целесообразно считать моделью реализации возможностей памяти, а распределение длин словоформ словаря – моделью устройства памяти.

Распределения длин словоупотреблений, особенно по данным верхней зоны ЧС, наглядно характеризуют типологию лингвистического объекта – языка, подъязыка, причем на текстовой оси отчетливее, чем на словарной.

Форма рангового распределения словоформ связана с распределением словоформ длины, равной средней длине словоупотребления в данном тексте и, следовательно, зависит от этой средней.

Разумеется, эти выводы требуют дополнительного обдумывания на базе расширенного материала.

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, 1975. – 75 с.

Алексеев П.М. О нелинейных формулировках закона Ципфа. – В кн.: Вопросы кибернетики. Вып.41. – М.-Л.: Научный совет по комплексной проблеме "Кибернетика" АН СССР, 1978, с.53-65.

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

Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата: Наука, 1978. – 183 с.

Григорьева А.С. О частотном словаре русской обиходной письменной речи. – В кн.: Учен.записки Тартусского университета. Вып.549. – Тарту, 1980, с.25-31.

Григорьева А.С. Статистическая структура русского эпистолярного текста (лексика частных писем). Автореф.дис. ... канд. филолог.наук. – Л., 1981. – 19 с.

Кван Л.И. Частотный словарь румынского подъязыка электроники. – В кн.: Статистика речи. – Л.: Наука, 1968, с.171-179.

Зореф М.Г. Частотный словарь немецких текстов по электронике. – В кн.: Статистика речи и автоматический анализ текста. – Л.: Наука, 1971, с.229-240.

Исенин И.А. О частотном словаре подъязыка современной французской прессы. – В кн.: Статистика речи. – Л.: Наука, 1968, с.185-190.

Калинина Е.А. Частотный словарь русского подъязыка электроники. – В кн.: Статистика речи. – Л.: Наука, с.144-150.

Кочеткова В.К., Скрелина Л.М. Частотный словарь французского подъязыка электроники. – В кн.: Статистика речи. – Л.: Наука, 1968, с.162-170.

Лебедев Б.М. Лингвостатистическое моделирование немецкого на-

учно-технического текста (подъязык телевизионной техники). Автореф.дис. ... канд.филолог.наук. - Л.: 1979. - 19 с.

Мартыненко Г.Я. Типология лингвостатистических распределений.- В кн.: Учен.записки Тартуского университета. Вып.628. - Тарту, 1982, с.103-120.

Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. - М.: Высшая школа, 1977. - 383 с.

Салчикова П.В. Английский и русский частотный словари по химии полимеров. - В кн.: Статистика речи и автоматический анализ текста. - Л.: Наука, 1974, с.295-329.

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

Турко Л.А. Частотный словарь русской разговорной речи. - В кн.: Статистика речи. - Л.: Наука, 1968, с.191-199.

Турнгина Л.А. Частотный словарь английских и американских газетных текстов. - В кн.: Статистика речи. - Л.: Наука, 1968, с.180-184.

Урбах В.Ю. Биометрические методы. - М.: Наука, 1964. - 415 с.

Частотний словник сучасної української художньої прози в двох томах. - Київ: Наукова думка, 1981. - Т. 1:864 с., Т.2: 856 с.

Частотный словарь индексирования. - Пермь: ПГУ им.А.М.Горького, 1974. 824 с.

Яблонская Н.Н. Исследование употребительности лексических и грамматических средств в немецком медицинском тексте (подъязык хирургии). Автореф.дис. ... канд.филолог.наук.- Л., 1980. - 20 с.

Dahl H. Word frequencies of spoken American English. - Essex, Conn.: Verbatim, 1979. - 348 p.

Howes D. A word count of spoken English. - In: Journal of verbal learning and verbal behavior, v 5, 1966, N° 6, p. 572-606.

Kučera H., Francis W.N. Computational analysis of present-day American English. - Providence, Rhode Isl.: Brown Univ. Press, 1967. - 424 p.

Kurcz I., Lewicki A., Sambor J., Woronczak J. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. T.I. Teksty popularnonaukowe. - Warszawa: PAN, 1974. - 858 s.

Kurcz I., Lewicki A., Sambor J., Woronczak J. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. T.II. Drobne wiadomości prasowe. - Warszawa: PAN, 1974a. - 792 s.

Kurcz I., Lewicki A., Sambor J., Woronczak J. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. T.IV. Proza artystyczna. - Warszawa: PAN, 1976. - 885 s.

Kurcz I., Lewicki A., Sambor J., Woronczak J. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. T.V. Dramat artystyczny. - Warszawa: PAN, 1977, - 632 s.

LENGTH-FREQUENCY DISTRIBUTIONS OF LEXICAL UNITS IN TEXT AND ITS VOCABULARY

Pavel M. Alekseev

S u m m a r y

Length-frequency distributions are analyzed both in text and its vocabulary; distinguishing word-token aspect from that of word-type is of prime importance for similar studies. It is shown that the upper part of a word-form frequency list contains general information of the total list being representative of the language (sublanguage) lexico-morphological structure. The remaining zone of the list shows an approximate identity of the token- and type-length distributions. Mean length of word-form as some standard value characterizing the working memory volume is offered to be treated in terms of word(form)-types as contrary to word-tokens, the former being taken from the upper zone of a frequency list, but not from the total one. A score of frequency dictionaries of the Russian, English, French, German, Polish, Ukrainian, Rumanian and Spanish languages have been used as the data sources.

ФОРМИРОВАНИЕ КОМПЛЕКСА КОНТРОЛИРУЕМЫХ УСЛОВИЙ ЛИНГВО-СТАТИСТИЧЕСКОГО ЭКСПЕРИМЕНТА

Е.В.Бахмутова

Исследования, проводимые группой "Статистика речи", показали, что закон Ципфа является важным инструментом изучения системных лингвистических объектов в типологическом аспекте (Алексеев П.М., 1983, с. 63). При этом подчеркивается важность соблюдения условий статистического эксперимента, к которым относят объем, качество, методику составления частотных словарей (Алексеев П.М., 1983, с. 35).

В данной статье рассматривается схема эксперимента, проведенного с целью выявления признаков качественной однородности специальных подъязыков. Под качественной однородностью понимается прежде всего идентичность статистического и графического "поведения" специальных подъязыков в рамках модели Ципфа. В связи с тем, что сравнительно-типологическое изучение специальных подъязыков возможно только на обширном статистическом материале, обработка которого связана с громоздкими вычислениями, особое значение приобретает разработка общей схемы лингво-статистического эксперимента, проводимого с применением ЭВМ, быстродействующих устройств ввода и вывода данных.

В самых общих чертах эксперимент заключается в наблюдении за объектом экспериментального исследования в специально контролируемых условиях (Охотников Г.Н., 1979, с. 10). Необходимость строгого соблюдения общей схемы эксперимента объясняется тем, что мы имели дело с эмпирическими данными – статистическими моделями шести специальных подъязыков английского языка, поэтому каждый раз при воспроизведении опыта данные на входе схемы имели разные числовые значения, т.е. являлись переменными. Выходные данные также менялись от опыта к опыту, поэтому "поведение" выбранных нами специальных подъязыков во многом определялось рядом факторов, влияющих на проведение эксперимента в целом. Дело в том, что при многократном повторении эксперимента для получения статистически

надежного вывода по результатам наблюдений требуется точное воспроизведение условий проведения эксперимента.

Комплекс условий нашего эксперимента формировался из лингвистических и статистических факторов, так как в качестве эмпирического материала были выбраны статистические модели лингвистических объектов – частотные словари следующих специальных подъязыков английского языка: (I) S_1 – электроники (Алексеев П.М.); (2) S_2 – судостроительных механизмов (Лукъяненков К.Ф.); (3) S_3 – квантовой электроники (Манасян Н.С.); (4) S_4 – радиообмена в авиации (Нелюбин Л.Л.); (5) S_5 – штабных документов (Нелюбин Л.Л.); (6) S_6 – антенно-фидерных устройств (Ионов А.И.). Совокупность экспериментальных объектов формировалась так, чтобы по возможности обеспечить стандартный набор условий его проведения для всех подъязыков и разработку единой программы для машинной обработки эмпирических данных.

К лингвистическим условиям эксперимента относятся:
I) принадлежность выбранных для эксперимента подъязыков к английскому языку;
2) единая методика составления частотных словарей S:
а) без учета омонимии; б) выбор словоформы в качестве единицы частотных словарей S (за исключением S_3 – частотный словарь однословных терминов).

К статистическим условиям эксперимента можно отнести переменные или константы, числовые значения которых вводились в ЭВМ для расчетов по формулам, отображающим зависимость "ранг-частота" ($F_i = KN \cdot i^{-\gamma}$, $F_i = KN \cdot (i+\rho)^{-\gamma}$, $F_i = KN \cdot i^{-\gamma - c\lg i}$):
I) так как значение коэффициента К для английского языка колеблется около 0,I (Алексеев П.М., I983, с. 4I; Нелюбин Л.Л., I983, с. 7I), то этот параметр считался константой К=0,I, что значительно упростило математическое обеспечение программ;
2) объемы выборок (N), на которых моделировались статистические структуры объектов нашего эксперимента, являются переменными (N_1 – 200 000, N_2 – 400 000, N_3 – 200 000, N_4 – I6 000, N_5 – 70 000, N_6 – I00 000). Объемы выборок $N_6 = N/N_3$-

N_2 пропорциональны и соотносятся следующим образом: 100 000 - 200 000 / 200 000 - 400 000. Предыдущие исследования показали (Лукьяненков К.Ф., 1969, с. 89), что пропорциональное соотношение длин выборок одинакового содержания на одном языке не влияет на значения параметров закона Ципфа. Таким образом, мы предположили, что данный вывод действителен и в том случае, когда пропорциональны объемы выборок из текстов разного содержания на одном языке, если тексты принадлежат подъязыкам, образующим качественно однородную группировку.

При проведении эксперимента учитывалось также единообразие подготовки и ввода исходных данных в ЭВМ. Значения логарифмов рангов и соответствующие им значения логарифмов абсолютных частот словоформ вводились после разбивки частотных словарей S на три участка: на участке I - с I по 15 ранг пошагово / под шагом понимается последовательный переход от одного ранга к другому/; на участке II - с 20 по 100 ранг через 5 шагов, с 100 по 1000 ранг через 10 шагов; на участке III - с 1000 ранга по 2000 ранг через 100 шагов, с 2000 ранга до конца частотного списка через 500 шагов (в частотных словарях S_3, S_4, S_5 участок III отсутствует).

Такая организация исходных данных соответствовала процедуре построения эмпирических и теоретических линий, являющихся графическим представлением распределения " ранг - частота". После вывода на графопостроитель отлогарифмированных эмпирических данных и вычерчивания эмпирической линии, производились вычисления коэффициентов γ, ρ по формуле $\lg F_i = \lg KN - \gamma \lg(i+\rho)$ для участков I и III и коэффициента γ по формуле $\lg F_i = \lg KN - \gamma \lg i$ для участка II. По расчетным данным вычерчивались аппроксимирующие прямые участков I, II, III. Затем по формуле $\lg F_i = \lg KN - \gamma \lg i - c \lg i^2$ рассчитывались коэффициенты γ, c для трех участков и вычерчивались теоретические кривые, аппроксимирующие эмпирическую линию. В результате эмпирическая и аппроксимирующие линии для наглядности реализовывались графопостроителем на одном листе в разном цвете. Эта процедура повторялась для каждого подъязыка S .

Все вычисления производились на универсальной микро -

процессорной ЭВМ третьего поколения SOLAR I6/65 по стандартной программе, написанной на языке ФОРТРАН. Наличие в наборе периферийных устройств графопостроителя Benson 254 позволило решить в ходе эксперимента ряд прикладных задач:

I) проверена точность построенных вручную графиков, отображающих зависимость "ранг-частота";

2) получены высокоточные машинные графики на статистическом материале, до сих пор не получившем графического представления;

3) особое внимание уделялось точности построения тех участков графиков, на которых наблюдалось наибольшее расхождение эмпирической и аппроксимирующих линий (участки I и Ш).

Все составляющие комплекса условий эксперимента можно рассматривать в качестве контролируемых переменных. Под переменной в данном случае понимается любой фактор (условие), контролируемый экспериментатором с той или иной степенью точности в ходе эксперимента. С математической точки зрения группу контролируемых переменных удобно представить в виде вектора

$$\mathcal{X} = \begin{Vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{Vmatrix}$$

Здесь x_1 – принадлежность подъязыков к английскому языку;
x_2 – единая методика составления частотных словарей;
x_3 – фиксированное значение коэффициента K = 0,I;
x_4 – пропорциональное соотношение объемов выборок;
x_5 – единообразие подготовки и ввода данных в ЭВМ;
x_6 – единое математическое и программное обеспечение.

Таким образом, группа контролируемых переменных есть вектор в шестимерном пространстве.

Векторное представление получают также выходные переменные. В проведенном эксперименте выходные переменные являются компонентами вектора в четырехмерном пространстве

$$y = \left\| \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{matrix} \right\|$$

Здесь y_1 - коэффициент γ (линейная аппроксимация)
y_2 - коэффициент ρ Расчет по формуле $\lg F_i = \lg KN - \gamma \lg(i+\rho)$

y_3 - коэффициент γ (нелинейная аппроксимация)
y_4 - коэффициент c Расчет по формуле $\lg F_i = \lg KN - \gamma \lg i - c(\lg i)^2$

Если бы специальные подъязыки имели идентичные статистические структуры, и комплекс условий воспроизводился предельно точно, то набор значений входных переменных x однозначно определял бы значения выходных переменных y

$$y = \eta(x). \qquad (1)$$

При изучении статистических закономерностей идеального случая воспроизводимости эксперимента быть не может. Для типологических исследований специальных подъязыков важное значение, помимо строгого соблюдения контролируемых условий эксперимента, имеет сведение до минимума влияния на результаты наблюдений нерегулярных (неконтролируемых) факторов. В проведенном эксперименте нерегулярные факторы составили вектор помех ξ в двухмерном пространстве с компонентами ω_1, ω_2, которыми могли быть случайные ошибки, допущенные при вводе исходных данных ω_1 и в вычислениях ω_2.

Таким образом, результаты эксперимента y связаны функциональной зависимостью не только с комплексом контролируемых условий x, но и с вектором помех ξ.

$$y = \eta(x, \xi)$$

Однако вероятность ошибки в расчетах, производимых с помощью современной электронно-вычислительной техники чрезвычайно мала $\varepsilon(\xi) \to 0$. Следовательно, ошибка в результатах наблюдения ε определяется в основном тем, что комплекс контролируемых условий не обладает идеальной воспроизводимостью. Очевидно, что на выходе схемы эксперимента появляется еще одна переменная функция $\varepsilon(x)$. Тогда функциональная зависимость (1) запишется следующим образом

$$y = \eta(x) + \varepsilon(x). \qquad (2)$$

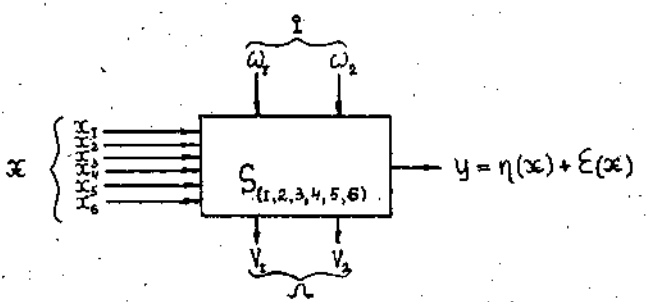
5

Модель (2) отражает структуру проведенного эксперимента.

Регулярную часть модели $\eta(\mathfrak{X})$ в статистике принято называть функцией отклика (Охотников Г.Н., 1979, с. 14), так как именно эта часть связывает функциональной зависимостью результаты наблюдения и комплекс контролируемых условий. В дальнейшем при рассмотрении результатов эксперимента мы имели в виду случай аддитивного влияния ошибки $\mathcal{E}(\mathfrak{X})$, т.е. суммировали $\mathcal{E}(\mathfrak{X})$ с функцией отклика.

Помимо группы основных выходных переменных $\mathcal{Y}$ в результате эксперимента получены неосновные переменные V_1, V_2, образующие вектор $\mathcal{N}$ в двухмерном пространстве. Компонента V_1 представлена в табличном выводе результатов вычислений значениями коэффициента A (линейная аппроксимация), являющимся промежуточным этапом расчета коэффициентов B, C (γ, ρ). Компонента V_2 соответствует коэффициенту ρ на участке П (линейная аппроксимация), значение которого не выводилось на графопостроитель.

На рис. I показана общая схема лингво-статистического эксперимента.



В нашем эксперименте функция отклика $\mathcal{Y}=\eta(\mathfrak{X})$ представлена вектором в четырехмерном пространстве. При этом разбивка эмпирической линии на три участка (I, П, Ш) привела к тому, что в каждую компоненту вектора $\mathcal{Y}$ вошло три переменных:

$$\gamma = \begin{Vmatrix} \gamma_I \\ \gamma_{II} \\ \gamma_{III} \end{Vmatrix} \quad \rho = \begin{Vmatrix} \rho_I \\ \rho_{(II)} \\ \rho_{III} \end{Vmatrix}$$ (линейная аппроксимация)

$$\gamma = \begin{Vmatrix} \gamma_I \\ \gamma_{II} \\ \gamma_{III} \end{Vmatrix} \quad C = \begin{Vmatrix} C_I \\ C_{II} \\ C_{III} \end{Vmatrix}$$ (нелинейная аппроксимация)

Таким образом, в наборе выходных данных формируется вектор второго порядка. Значениями компонент второго порядка мы оперировали при рассмотрении "поведения" подъязыков S. В результате эксперимента мы получили

$$y = \begin{Vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Vmatrix}$$ область выходных переменных первого порядка

$$y_1(\gamma) = \begin{Vmatrix} \gamma_I \\ \gamma_{II} \\ \gamma_{III} \end{Vmatrix} \quad y_3(\gamma) = \begin{Vmatrix} \gamma_I \\ \gamma_{II} \\ \gamma_{III} \end{Vmatrix}$$ область выходных переменных второго порядка

$$y_2(\rho) = \begin{Vmatrix} \rho_I \\ \rho_{(II)} \\ \rho_{III} \end{Vmatrix} \quad y_4(C) = \begin{Vmatrix} C_I \\ C_{II} \\ C_{III} \end{Vmatrix}$$

Многокомпонентный состав выходных данных, а также аналитическая и графическая формы их представления позволяют подробнее рассмотреть "поведение" специальных подъязыков в рамках модели Ципфа.

По характеру графического "поведения" статистические модели специальных подъязыков S были распределены по трем типологическим группам: I - S_1, S_2, S_6 ; II - S_4; III - S_3, S_5. В процессе сопоставления полученных аналитических данных уточнен характер связей как между выделенными типологическими группами специальных подъязыков S, так и подъязыками в группах I и III.

Результаты эксперимента дают возможность предположить, что сходство графического "поведения" экспериментальных

объектов S при максимально точном воспроизведении комплекса контролируемых условий есть внешнее проявление качественной однородности данных специальных подъязыков.

Возможно, что в ходе дальнейшего сравнительно-типологического исследования на более массивном эмпирическом материале будут выявлены новые типы графического "поведения" специальных подъязыков или расширен состав приведенных выше типологических групп.

ЛИТЕРАТУРА

Алексеев П.М. Методика квантитативной типологии текста. Л., 1983.

Лукьяненков К.Ф. Лексико-статистическое описание английского научно-технического текста с помощью электронно-вычислительной машины (подъязык судовых механизмов). КД. Минск, 1969.

Нелюбин Л.Л. Перевод и прикладная лингвистика. М., 1983.

Охотников Г.Н. Математическая статистика. Вып.I, М., 1979.

PRE-SET CONDITIONS OF A STATISTICAL EXPERIMENT IN COMPARATIVE LINGUISTICS

Yelena Bakhmutova

Summary

The article gives a thorough analysis of the pre-set conditions of an experiment carried out in the interest of comparative linguistics. The author maintains that first, extensive statistics require pre-set conditions for such experiments to ensure greater reliability of the resultant data. Second, a well-planned complex of pre-set conditions provides ample possibility for comparative studies of statistical text structures.

СЕМАНТИКО-КВАНТИТАТИВНОЕ ИССЛЕДОВАНИЕ ПОДЪЯЗЫКА
(опыт создания автоматизированной системы)

Бестужев А.К., Городецкий Б.Ю., Зайцева О.В.,
Зевахина Т.С., Кузнецов В.В., Сабурова И.Г., Эстрович Е.В.

В статье описывается особый тип автоматизированной системы обработки текста, предназначенной для семантико-квантитативного исследования подъязыка. Программное обеспечение позволяет гибко проводить формальную и семантическую инвентаризацию подъязыка с получением разнообразных словарей (частотно-тезаурусных, алфавитно-частотных, дистрибутивно-семантических, конкордансов, функционально-тематических выборок терминов и т.д.). Получаемые с помощью описываемой системы данные о подъязыке машиностроительной отрасли используются при проектировании, построении и ведении терминологического банка данных.

Постановка задачи. Участие языковедов в совершенствовании производственных и социальных процессов с применением ЭВМ выполняется сегодня в новом для лингвистики методе исследования - в моделировании языковой системы и процессов речевой деятельности. Результаты такого моделирования воплощаются в особых искусственных языках и разнообразных лингвистических алгоритмах.

Первым этапом практически любого прикладного семантического исследования является семантико-квантитативная инвентаризация подъязыка. Нужно сначала знать наличное множество семантических объектов, а уже затем решать на них те или иные прикладные задачи: будь то разработка лингвистического обеспечения автоматизированных информационных систем, обучение иностранному языку или оптимизация средств массовой коммуникации и т.д. (Ванников, 1977; Городецкий, 1973; Караулов 1981).

Полная формальная и семантическая инвентаризация подъязыка - трудоемкая и объемная работа, требующая создания специального типа автоматизированной системы обработки текстов (АСОТ), которая должна быть эффективным исследовательским инструментом первичного анализа разнообразного массива текстов и последующего моделирования подъязыка на основе первичных инвентарей. Разработкой подобного рода автоматизированных систем занимается ряд научных коллективов. [1]

сравнению с 60-70 годами эти системы реализуются на более совершенной вычислительной технике, оснащенной мощными операционными системами, базами данных и системами управления базами данных (СУБД), что позволяет не только более эффективно автоматизировать традиционные лексикографические работы по составлению разнообразных словарей и словников, но и наделить эти АСОТ новыми функциями. Так, по замыслу коллектива ученых, возглавляемых Р.Г.Пиотровским, диктантический автомат должен включать в себя комплекс программ, реализующих адаптивные семантические алгоритмы, которые могли бы не только распознавать общий смысл подаваемого на вход текста, но были бы способны осуществлять достаточно детальное распознавание смысла текста с учетом общения и потребностей абонента-собеседника (Вейнеров и др., 1978,с.6). По мнению В.М.Андрющенко, исследовательские автоматизированные лексикографические системы должны иметь встроенные процессоры естественного языка, осуществляющие морфологический и синтаксический анализ текстов (Андрющенко, 1981).

В настоящей работе описывается первая очередь автоматизированной системы, предназначенной для проведения семантико-квантитативного исследования подъязыка (СКИП).

Автоматизированная система семантико-квантитативного исследования подъязыка. При создании АСОТ-СКИП мы исходили из следующих теоретических представлений. В общем виде АСОТ можно представить как систему алгоритмов, на входе(ах) выходе которой присутствует текстовая информация. При создании моделей АСОТ можно выделить две основные задачи:

1) проблема алгоритмизации семантического анализа и синтеза текста;

2) проблема создания искусственных семантических языков (метаязыков) для записи текстовой информации.

Строение семантических языков во многом предопределяет принципиальное устройство алгоритмов анализа текста и переработки информации. Поэтому именно семантические языки составляют ядро проблемы моделирования естественного языка. Не случайно некоторые семантические языки приобретают черты алгоритмических языков, сливаясь, таким образом, с собственно алгоритмами переработки текста.

Построить модель АСОТ сразу для всего языка невозможно. Специфика каждой сферы человеческой деятельности, особенности соответству-

щего подъязыка этой сферы проявляются прежде всего в семантическом языке, разрабатываемом на множестве текстов в фиксированной области человеческого знания.

АСОТ-СКИП как инструмент разнообразных лингвистических исследований сочетает статистические методы с морфолого-синтаксическим и семантическим анализом, проводимым на начальном этапе вручную.

Приведем (рис. I) блок-схему АСОТ-СКИП как разновидности АСОТ (Городецкий, 1976, с.23-26; 1978, с. I6-I8). Разработанный формат приписывания входному тексту морфолого-синтаксической и семантической информации (он описан далее в статье) фактически представляет собой специальный экспликационный язык - ЭЯ. Сообщения на ЭЯ, результат работы блока экспликационного семантического анализа, поступают в банк информации в памяти ЭВМ (символ Р/М на рис. I обозначает сочетание машинного и ручного режимов работы, буква М - машинный режим). Блок логико-квантитативной обработки извлекает из банка информации необходимые данные в соответствии с запросом, а результаты обработки может как записывать в банк информации, так и выдавать пользователю. Ответ на запрос пользователя может быть тем или иным семантическим инвентарем, а может быть и описанием какой-либо закономерности. В настоящее время в АСОТ-СКИП отсутствует блок информационного семантического анализа. Однако он в дальнейшем может стать функционально изменяемым блоком настройки на решение конкретных лингвистических задач (фактически это сведется к выделению (вручную или полуавтоматически) дополнительной информации во входном тексте). Такой тип АСОТ может быть достаточно гибким инструментом разноцелевых исследований подъязыка.

Концепция терминологического банка данных. Создание АСОТ-СКИП проводится в рамках более крупной работы, цель которой состоит в разработке детального проекта терминологического банка данных (ТБД) в области машиностроения, методов его эксплуатации и ведения.

Для более ясного понимания общей направленности и конкретных задач обследования подъязыка машиностроения кратко остановимся на общей концепции ТБД. ТБД мыслится как ядро лингвистического обеспечения САПР и информационной технологии отрасли в целом (Городецкий, Зевакина, 1983). Очевидно, что традиционные формы ведения словарей не позволяют оперативно отслеживать высокую подвижность современных термино-

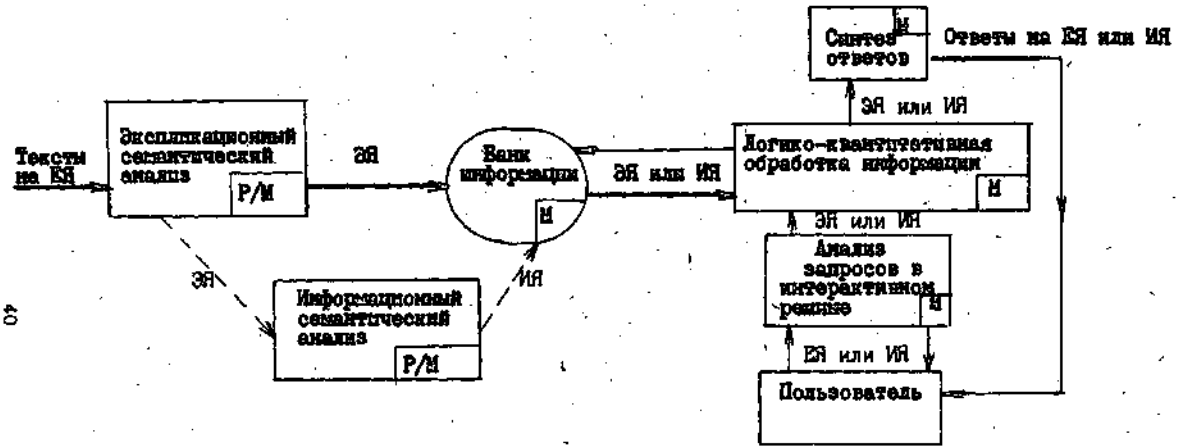


Рис.I. Общая блок-схема АСОТ-СКИП

систем, обусловленную бурным развитием науки и техники. Поэтому принципиальная установка при создании ТБД была на его машинную реализацию в виде базы данных. Учитывая возможности манипулирования данными, которые обеспечивают современные СУБД, в ТБД хранится только исходная многоаспектная терминологическая информация, которая может гибко комбинироваться в соответствии с индивидуальным запросом пользователя ТБД в разнообразные словари, тематические и дистрибутивные выборки терминов и т.д. и т.п. Как справедливо заметил В.М.Андрющенко (Андрющенко, 1984, с.7), "в автоматической лексикографии понятию типа словаря соответствует понятие режима обращения к нему: путем ограничений на выдачу словарных статей и их компонентов словарь может быть во внешней форме представлен в нужном объеме.., в нужном аспекте".

Согласно разработанной нами универсальной методике построения ТБД (Городецкий, Зевахина, 1983), словарная статья заводится на узуальный термин, т.е. термин в определенном узуальном значении. Словарная информация в нынешнем варианте словарной статьи ТБД организуется в виде (13 + ℓ) зон, где число ℓ равно числу парадигматических отношений, различаемых в данной терминосистеме. Каждая зона может рассматриваться как признак, принимающий для данного термина то или иное значение.

Перечислим принятые нами признаки термина. Первые четыре признака характеризуют план выражения термина: буквенный вид термина, его морфную запись, словоизменительную характеристику термина, синтаксический тип составного термина. Пятый признак указывает статистические характеристики термина (частотность и информативность). Последующие признаки характеризуют план содержания термина. Шестой признак задает дефиницию термина (на естественном языке и формализованном метаязыке). Признаки с седьмого по девятый описывают синтагматическую семантику: задают синтагматическую структуру семантики составного термина, критерий устойчивости составного термина, синтагматические семантические связи термина, реализуемые в текстах подъязыка. Все остальные признаки относятся к парадигматической семантике. Десятый признак указывает на наличие многозначности у термина в данном подъязыке. Признаки с одиннадцатого по тринадцатый задают отнесенность данного термина к тому или иному классу: к универсальной или специальной лексике;

к тому или иному функциональному словарю; к тому или иному понятийно-категориальному классу. Последние ℓ зон отводятся для семантических коррелятов данного термина (синонимы, гипонимы, гиперонимы и т.д.).

Очевидно, что структура словарной статьи в зависимости от конкретной функциональной ориентации ТБД может принимать самые разнообразные формы. Предлагаемый нами вариант (I3 +ℓ зон) является наиболее полным и подробным представлением словарной информации.

Разработанная нами общая концепция ТБД сходна с теми его версиями, которые развиваются, в частности, в Институте кибернетики АН УССР под руководством Э.Ф.Скороходько (Скороходько, I983), в Белорусском институте научно-технической информации (Ильина, I973), в работах Р.Ю.Кобрина (Кобрин, I985).

Исходный массив и формат экспликационной записи. Как указывалось выше, объектом исследования является подъязык машиностроения, задаваемый корпусом текстов, который на первом этапе развития системы представляет собой отраслевой классификатор технологических операций. Описание технологической операции дано в виде номинативного лексического комплекса (НЛК). Пример НЛК: обжиг гомогенизационный с нагревом в печах с окислительной атмосферой заготовок, изделий из стали. Объем массива - I57I НЛК, длина массива в словоупотреблениях - 9775. Каждому НЛК приписан I0-значный идентифицирующий (функционально-тематический) индекс, однозначно задающий место данной технологической операции в иерархической древовидной структуре, имеющей фиксированную глубину в пять уровней. На самом нижнем, пятом, уровне находятся наименования конкретных операций (НЛК), на вышестоящих уровнях расположены названия все более широких их классов. Каждый уровень кодируется двумя десятичными знаками, т.е. максимальное число элементов на уровне ограничено 99. Реально на первом уровне классификации выделено 7 основных классов технологических операций: 0I - общие операции, 02 - технический контроль, 03 - обработка давлением, 04 - термическая обработка, 05 - пайка, 06 - сборка, 07 - сварка.

Семантико-квантитативное обследование подъязыка машиностроения состоит, во-первых, в определении номенклатуры базовых узуальных терминов ТБД, и, во-вторых, в получении лингвистической инфор-

формации, необходимой для заполнения ряда зон словарных статей ТБД (в ходе его создания и ведения).

В настоящей работе представлены преимущественно те аспекты инвентаризации подъязыка, которые выполнялись автоматизированно с помощью ЭВМ.

Для того, чтобы приписать компонентам НЛК семантико-грамматическую характеристику, был разработан формат экспликационной записи: слева от каждого полнозначного слова цифрой (I/2) указывалось, является ли оно простым или сложным, справа приписывался пятизначный код, первая цифра которого обозначала часть речи, вторая - число, третья - падеж и две последние цифры - рабочую тезаурусную характеристику. Кроме этого, в слове выделялась специальным знаком основа (устойчивая графическая часть), размечались границы составных терминов. Пример размеченного НЛК: 0306010001 I пробив//ка 11104 I отверсти//я 11251 в I листов//ой 21671 I заголов//ке 11628, I детал// и 11628 на [I механическ//ом 21691 I пресс//е 11638] 38.

Остановимся более подробно на тезаурусных характеристиках, приписываемых словам. Тезаурусные характеристики отражают разбиение лексики на семантические классы. Разбиение проводилось в два этапа:

I. Отдельно для существительных (они составляют 60% от общего объема лексем).

2. Отдельно для прилагательных.

К семантическому разбиению значений мы предъявляем следующие требования:

(I) Семантические признаки не должны быть контекстно-зависимыми.

(2) Классификация каждой лексемы осуществляется по одному (основному для нее) признаку, тем самым разбиение состоит из непересекающихся классов.

(3) Инвентарь семантических признаков задается не априорно, а формируется в процессе классификации.

Источниками информации о значениях лексем служили данные о функционировании лексем в текстах подъязыка, общеязыковые и спе-

циальные словари и энциклопедии. В отдельных случаях были полезны консультации со специалистами-технологами.

В результате классификации получен семантический инвентарь, состоящий из 92 классов (существительные – 53 класса, прилагательные – 39 классов). Приведем некоторые примеры:

06. ТЕРМИЧЕСКАЯ ОБРАБОТКА	выдержка, закалка, нитроцементация, отжиг
04. ОБРАБОТКА ДАВЛЕНИЕМ	выглаживание, вырубка, высадка, гибка, дорнование
7I. ФОРМА	гофрированный, дисковый

Программное обеспечение АСОТ-СКИП. Разработка программного обеспечения АСОТ-СКИП велась в следующих направлениях:

(I) ввод НЛК, контроль формата ввода, исправление ошибок ввода, запись НЛК в базу данных;
(2) составление машинных рабочих словарей (МРС) словоформ и основ по всему массиву НЛК;
(3) составление по МРС словоформ алфавитно-частотных и словоуказательных словарей;
(4) составление по МРС основ контекстуальных словарей для основ, выступающих в роли ядра терминологического сочетания;
(5) составление алфавитно-частотного словаря служебных основ;
(6) составление инвентаря формальных моделей построения НЛК;
(7) формирование подмассивов НЛК путем произвольного комбинирования значений идентифицирующего индекса и пятизначного семантико-грамматического кода.

Программирование велось на языке Фортран IУ, ЭВМ – СМ-4, операционная система – RSX . Блок-схема программного обеспечения АСОТ-СКИП представлена на рис. 2.

Можно видеть, что в программном обеспечении реализуется конкретная разновидность АСОТ-СКИП. Так, например, совокупность МРС совместно с массивом НЛК (рис. 2) выполняют функции Банка информации (рис. I).

Раскроем более подробно содержание некоторых этапов разработки АСОТ-СКИП.

Разработка АСОТ-СКИП была начата с создания программ, обеспечивающих достаточную эффективность ввода в ЭВМ массива НЛК с учетом сложности формата экспликационных записей. Процесс ввода НЛК осуществлялся в диалоговом режиме. АСОТ-СКИП предлагала опера-

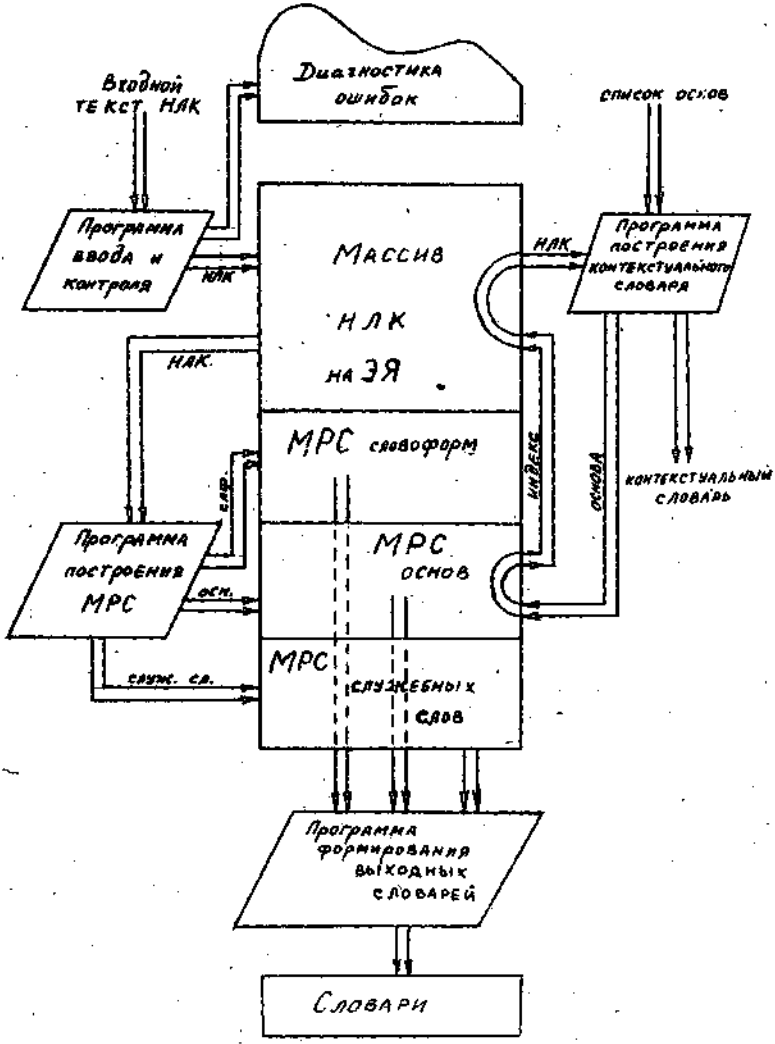


Рис.2. Блок-схема программного обеспечения АСОТ-СИПП

тору набрать текст НЛК на экране дисплея. После сообщения оператора о том, что он закончил ввод НЛК, программа осуществляла проверку НЛК на отсутствие запрещенных символов и соблюдение формата ввода НЛК. В последнем случае контролировалось, например, наличие необходимых пробелов, символов, начинающих и завершающих НЛК, проверялось количество цифр в индексе и кодах. Если ошибок в тексте НЛК не находилось, НЛК записывался на диск в базу данных, о чем сообщалось оператору. В противном случае текст НЛК выводился на дисплей, оператору сообщалось общее количество допущенных ошибок и под строкой текста специальным знаком указывалось местоположение ошибки.

Для проведения семантико-квантитативного анализа массива НЛК были разработаны программы, позволяющие составлять алфавитно-частотный МРС словоформ с указанием семантико-грамматического кода, частоты встречаемости и перечня функционально-тематических индексов НЛК, в которых он встретился. Аналогичная информация приводится в большинстве получаемых словарей.

МРС словоформ является базовым для построения ряда других словарей и формирования подмассивов НЛК (см. рис. 2). В частности, МРС основ строится без обращения к массиву НЛК, путем соответствующей обработки МРС словоформ. В свою очередь МРС основ является базой для проведения последующего анализа. Так, например, для составления частотного контекстуального словаря слов, которые могут выступать в качестве ядра терминологического словосочетания, используется МРС основ, по которому устанавливаются функционально-тематические индексы НЛК, в которых встретилась данная основа. По этим индексам из массива отбираются нужные НЛК и их тексты подвергаются дальнейшей обработке - определяются два ближайшие слова справа и слева от ядра. Данный контекстуальный словарь объединят в себе черты конкорданса, дистрибутивно-семантического и частотного словарей.

Разработан комплекс программ, позволяющий гибко формировать подмассивы НЛК путем комбинации значений функционально-тематических индексов НЛК и семантико-грамматических кодов слов. Таким образом, получаются алфавитно-частотные словари для отдельных классов (подклассов) технологических операций, словари существительных, прилагательных и т.д.

Примеры использования АСОТ-СКИП. Использование АСОТ-СКИП позволяет получить инвентарь формальных моделей построения НЛК. Формальная модель НЛК – это схема, в которой отражается его поверхностная структура с использованием буквенных символов, представляющих каждый член НЛК как определенную часть речи. Например, НЛК разрезка групповой листовой заготовки на индивидуальные по разметке на виброножницах имеет следующую формальную структуру С + П + П + Сред + НА + П + ПО + Сдат + НА + Спр.

Формальные модели построения НЛК в анализируемом массиве НЛК чрезвычайно разнообразны. Для класса "Общие операции" (213 НЛК) выявлена 101 модель построения НЛК. АСОТ-СКИП дает возможность автоматически выводить данные о продуктивности каждой модели как по всему массиву НЛК, так и по отдельным подмассивам. В подмассиве "Общие операции" наиболее продуктивны модели

С + Сред + [,] + Сред + В + Спр	–	9	НЛК
С + П + Ств + Ств	–	8	НЛК
С + Сред + ПО + Сдат + Ств	–	8	НЛК

Задача полной семантической инвентаризации подъязыка предполагает определение семантико-реляционной структуры НЛК.

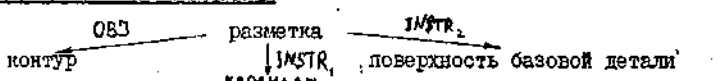
Инвентарь семантических реляций, выявленный при исследовании массива НЛК, насчитывает 34 отношения. Например:

X $\xrightarrow{\text{PH-CHAR}}$ У "У – физическое (механическое) свойство Х-а"
высокоскоростной молот

X $\xrightarrow{\text{FUNC}}$ У "У – характерный способ функционирования Х-а"
пульсирующий прижим

X $\xrightarrow{\text{MATER}}$ У "У – материал, из которого сделан Х"
трубопровод из стали

В качестве языка для представления семантико-реляционной структуры НЛК предлагается язык, основывающийся на семантической сети. Для каждого НЛК была составлена схема, отражающая его семантико-реляционную структуру. Например, разметка контура карандашом по поверхности базовой детали



Предполагается, что с учетом тезаурусной информации процедура определения семантико-реляционной структуры НЛК по его формальной модели может быть алгоритмизована.

Предварительный лексико-статистический анализ массива НЛК дал следующие результаты. Алфавитный словник насчитывает 981 разное слово (от абразивный до этикетка). Получено распределение по частям речи: существительные – 584, прилагательные – 327, причастия – 56, наречия – 9, числительные – 3, глаголы – 2. Построены словники по отдельным подмассивам НЛК, соответствующие классам, на которые разбиты технологические операции, а также словари пересечений.

Очевидно, что с помощью АСОТ-СКИП можно обрабатывать и обычные связные тексты, сопровождая каждое слово в предложении экспликационной пометой разработанного формата. Используя один и тот же формат экспликационных записей, можно произвольно менять его содержательное наполнение. Следует отметить, что АСОТ-СКИП можно при небольшой модификации использовать для обработки текстов, не снабженных специальной разметкой. В этом случае результаты инвентаризации ограничиваются составлением алфавитно-частотных словарей словоформ.

Дальнейшее развитие АСОТ-СКИП будет направлено на расширение возможностей автоматического анализа текстов на ЕЯ, сопряжение АСОТ-СКИП в одно целое с ТБД и его аналогами, повышение ее эффективности как инструмента разнообразных прикладных семантических исследований, в частности, на основе развития диалогового режима использования АСОТ-СКИП.

Л И Т Е Р А Т У Р А

Андрюшенко В.М. Автоматизированные лексикографические системы. – В кн.: Теоретические и прикладные аспекты вычислительной лингвистики. М.: Изд-во МГУ, 1981, с. 71–88.

Андрюшенко В.М. Машинный фонд русского языка. Основные компоненты. – В кн.: Квантитативная лингвистика и автоматический анализ текстов. Тарту, 1984 (Ученые зап. Тартуск. ун-та, вып. 689), с. 3–15.

Ванников Ю.В. Единая справочная лингвостатистическая система как база учебной лексикографии. – В кн.: Актуальные проблемы учебной лексикографии. М.: Русский язык, 1977, с. 71–93.

Вейнеров О.М., Кравцова И.С., Пиотровский Р.Г., Цыгарева Е.А. Лингвистические автоматы. – В кн.: Статистика речи и автоматический анализ текста. (Вопросы кибернетики, вып. 41) М.–Л.: Наука, 1978, с. 5–10.

Городецкий Б.Ю. О принципах инвентаризации семантики языка и подъязыка. - В кн.: Теоретические и экспериментальные исследования в области структурной и прикладной лингвистики. М.:Изд-во МГУ, 1973, с.20-59.

Городецкий Б.Ю. Семантические проблемы построения автоматизированных систем обработки текстовой информации. - В кн.: Вычислительная лингвистика. М.:Наука, 1976, с.16-33.

Городецкий Б.Ю. Теоретические основы прикладной семантики. Автореферат докт.дисс. - М.:МГУ, 1978.

Городецкий Б.Ю., Зевахина Т.С. К построению терминологического банка данных как компонента лингвистического обеспечения информационной системы. - В кн.:Всесоюзная конференция "Основные направления развития и совершенствования работ по стандартизации научно-технической терминологии в XI пятилетке". Тезисы докладов. М.:ВНИИКИ, 1983, с.6-8.

Городецкий Б.Ю., Зевахина Т.С., Еленевская Л.Е., Лебедева О.С., Пожарицкий И.Ф. Методика семантико-статистического анализа базового отраслевого тезауруса по геологии. - Научно-техническая информация по геологии. Экспресс-информация, вып.5. М.:ВИЭМС, 1981, с.1-5.

Ильина С.В. Возможности использования фактографических ИПС в качестве инструмента лексикографической обработки научно-технических терминов.-В кн.: Термин и слово. Горький: Изд-во Горьк. ун-та, 1973, с.75-82.

Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. - М.:Наука, 1981.

Кобрин Р.Ю. Банки данных 80-х: теория,эксперимент, внедрение. В кн.: Квантитативная лингвистика и автоматический анализ текстов. Тарту, 1985 (Ученые зап. Тартуск. ун-та, вып.711), с. 39-54.

Скороходько.Э.Ф.Семантические сети и автоматическая обработка текста. - Киев: Наукова думка, 1983.

SEMANTIC-QUANTITATIVE ANALYSIS OF A SUBLANGUAGE
(a special type of an automated system)

A.Bestuzhev, B.Gorodetskiy, O.Zaytseva, T.Zevakhina,
V.Kuznetsov, I.Saburova, E.Petrovitch

S u m m a r y

A special type of an automated text processing system for semantic-quantitative analysis of a sublanguage is presented. The software allows of formal and semantic inventorization of a sublanguage. The system is able to produce different types of dictionaries such as combined thesaurus-frequency lists, alphabetical-frequency dictionaries, distributive-semantic dictionaries, concordances, functional-thematic subsets of special terms,etc. The experimental domain of the project is machine industry. The data on the sublanguage provided by means of the automated system are used in designing, implementing and supporting the terminological data base.

ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРОБЛЕМЫ ВЫБОРКИ "ОПТИМАЛЬНОГО ОБЪЕМА"

В.Н. Бычков

Понятие выборки "оптимального объема" в статистической лингвистике имеет два значения. Во-первых, в собственно статистическом смысле оно означает минимальный объем выборки, достаточный для обеспечения вероятностно-статистической достоверности корпуса исследуемых единиц. Во-вторых, оно имеет смысл такого объема выборочной совокупности, в которой "канонический" закон Ципфа выполняется максимально точно и однозначно на всем протяжении частотно-рангового ряда

$$F_i = kN \cdot i^{-\gamma}, \qquad (1)$$

где F_i и i - частоты и соответствующие им ранги, k - максимальная относительная частота в выборочном корпусе, N - объем выборки. В данной статье выборка "оптимального объема" понимается во втором из указанных выше значений. Данный вопрос рассматривается в работах ряда отечественных авторов (см.: Фрумкина, 1964; Орлов, 1976, 1978; Алексеев, 1978).

Впервые представление об этой величине под названием собственно "оптимального объема" было введено Ципфом для теоретического и эмпирического обоснования случаев строгого выполнения частотно-рангового распределения, когда γ-параметр равняется 1 (Zipf, 1949). Практическое значение оптимального объема как специфического параметра, по мнению Ю.К. Орлова, состоит в том, что с его помощью можно описать как рост словаря, так и частотную структуру на любом произвольном объеме текста через моделирование механизма отклонения хвостов" в низкочастотной зоне частотного графика (Орлов, 1976, с. 193). Здесь выборка оптимального объема определяется как ципфовский объем, или Z - объем. По оценкам самого Ципфа, объем выборок, в которых теоретически должен выполниться закон частотно-рангового распределения с $\gamma = 1$, может иметь место не для всяких лексических совокупностей, а таких, словарь которых составляет (для английского языка) около 22 тыс. лексических единиц на выборках примерно в 200 тыс. словоупотреблений. Ципф назвал такой объем оптимальным с точки зрения психофизического принципа "наименьшего усилия". Позднее Б. Мандельбротом была предпринята попытка объяснить Z - объем и частотно-ранговое распределение в лингвостатистичес-

ких выборках в целом в терминах оптимизации кодирования (Мандельброт, 1957). Естественный язык в интерпретации Б. Мандельброта и является той системой, которая обеспечивает эту оптимизацию. Однако и в модели Б. Мандельброта парадокс "оптимального объема" не был окончательно разрешен. Дело в том, что в выборках из текстов даже на одном и том же языке $\hat{Z}$-объем не равен численно ципфовскому. Как показано, например в (Орлов, 1976), для каждого отдельного случая имеется единственное значение оптимального объема. Такая вариативность $\hat{Z}$-объема наряду с изменчивостью других параметров частотно-рангового распределения, "нежелательная" с точки зрения хрестоматийной статистики, остается до сих пор необъясненной. Это привело к тому, что сам закон Ципфа стал интерпретироваться в качестве закона не языка, а текста, так как система языка как нечто единое "не должна" порождать такие патологические явления. При этом на эмпирическом уровне в ряде случаев под ципфовским объемом эксплицитно понимается целостный текст (объемом $N = \hat{Z}$ словоупотреблений) одного автора, а также законченные части текста - главы, выделенные самим автором. В то же время выполнение закона Ципфа в так называемой "канонической" форме, то есть с параметрами выборки оптимального объема на произвольно составленных выборках объясняется чистой случайностью совпадения статистических свойств этой выборки с характеристиками $\hat{Z}$-объема (Орлов, 1976, с. 194). Более того, общий формальный генезис "правильных" ципфовских структур в таких случаях в категорической форме увязывается с целостностью и художественностью произведений.

Величина ципфовского объема, которая на частном примере Ю.К. Орлова "случайно" составила 27 тыс. словоупотреблений (с/у), действительно коррелирует с объемом небольшого текста или главы (см.: Орлов, 1976, с. 191). Но, с другой стороны, очевидно, что тексты такой или примерно такой длины составляют только частный случай речевых произведений, о художественных достоинствах которых априорно вряд ли представляется возможным говорить. Как будет показано ниже, свой так называемый "оптимальный" объем имеет любая подъязыковая выборка, но в случае сборных выборок $\hat{Z}$ - объем оказывается существенно (в 2-3 раза) больше соответствующего параметра в одно-текстовых выборках. Но такой оптимальный ципфовский объем "просто обязан" существовать хотя бы теоретически в качестве некоторого идеального в формально-математическом смысле и

устойчивого в лингвосемиотическом смысле состояния выборки.

Уточним аксиоматику Z - объема, чтобы избежать разночтений одного и того же статистического материала:

I. Выполнение закона Ципфа в его канонической форме (в Z-объеме) возможно при условии $\gamma = 1$. Если отвлечься от ограниченного по своему воздействию параметра ρ Мандельброта в

$$F_i = kN \cdot (i+\rho)^{-\gamma} \qquad (2)$$

то на билогарифмическом графике частотно-ранговое распределение предстаёт в форме прямой.

2. Исходными точками при теоретическом построении такого графика служат $F_{max} = kN$ и $i_{max} = V$ (V - объем словаря), которые задают начальные и конечные параметрические условия частотно-рангового распределения. Промежуточные значения частот и рангов могут статистически уклоняться по обе стороны от теоретической прямой, но крайние значения должны фиксироваться всегда четко и однозначно, что является одной из специфик лингвостатистических распределений, так как количественно F_{max} базируется на константности параметра k и на заданности объема выборки N, а i_{max} фиксирует объем состава V элементов выборки N. Поэтому частотно-ранговое распределение в выборке Z - объема удовлетворяет условию $\gamma = 1$, если эмпирические и теоретические частоты и ранги совпадают на экстремальных участках распределения, а также и в средней части частотного списка, где $F_{теор.} = F_{эмпир.}$ (F - средняя частота распределения).

3. Используемая лингвостатистическая модель особенно на начальном этапе исследования механизма порождения Z - объема должна иметь достаточно простую и четкую логическую структуру с тем, чтобы постоянно можно было следить за изменениями лингвистического смысла при формальных преобразованиях исходной модели. Выполнение этого условия призвано обеспечить большую воспроизводимость результатов в лингвостатистических исследованиях, в которых особенно заметна субъективность авторов при решении проблем синонимии, омонимии, формирования классов слов.

В работах (Алексеев, 1978; Бычков, 1984) развивается тезис о том, что нелинейные ципфовские модели дают лучшую аппроксимацию эмпирических распределений в лингвостатистических выборках различного объема. Далее будем исходить из предположения, что нелинейная ципфовская модель

$$F_i = k\alpha V^2 \cdot (i+\rho)^{-\gamma + \alpha \ln i} \qquad (3)$$

удовлетворяет предъявленным выше требованиям, так как все ее составляющие элементы и параметры получили достаточно определенную лингвистическую интерпретацию (см.: Бычков, 1984). Данная модель через простейшие, лингвистически обоснованные преобразования сводится к обычному частотно-ранговому виду и не противоречит ему. Рассмотрим следующие вопросы. 1. При каких условиях модель (3) удовлетворяет требованиям Z- объема? 2. Какие значения принимают параметры модели (3) в Z- объеме? 3. Можно ли осмыслить Z- объем как некоторое лингвистически значимое явление и специфическое состояние лексической системы?

В модель (3) входит функция "текст–словарь" вида

$$N = a V^{\beta} \qquad (4)$$

Ее вывод и лингвистическая интерпретация приводятся в работе (Бычков, 1983). Там показано, что эта функция дает хорошие прогнозы объемов словаря под данный объем выборки и наоборот – показывает, какой объем выборки покрывает заданный объем словаря. К числу возможных достоинств функции (4) можно отнести простоту ее математической структуры, что обеспечивает необходимые формально-математические преобразования и самой функции и ее производных. Так как $N = a V^{\beta}$, а параметр β не играет принципиальной роли, то в качестве исходной для последующих рассуждений принимаем нелинейную ципфовскую модель вида

$$F_i = k N \cdot i^{-\gamma \exp \alpha i} \qquad (5)$$

Отсюда видно, что основное условие Z-объема в данной модели возможно, если

$$\exp \alpha i = 1 , \qquad \text{то есть} \quad \alpha i = 0.$$

Очевидно, что это возможно при условии, что хотя бы один из сомножителей равен нулю, но по исходному положению частотно-ранговых распределений ранги $i \geqslant 1$. Следовательно $\alpha = 0$. Из вывода (3)

$$\alpha = \frac{\ln \gamma}{\ln i_{max}} = \frac{\ln \gamma}{\ln V} . \qquad (6)$$

Это соотношение равно нулю, если $\ln \gamma = 0$, то есть $\gamma = 1$, что возможно при одном единственном условии, что $\lg F_{max} / \lg V = 1$.

Следовательно, нелинейные частотно-ранговые модели (3) и (5) предполагают и утверждают, что при условии равенства по абсолютной величине $F_{max} = kN$ и $i_{max} = V$ частотно-ранговое распределение в билогарифмической репрезентации бу-

дет линейным с $\gamma = 1$, то есть оно становится "оптимально-ципфовским". Само по себе это достаточно тривиальное утверждение. Существенно то, что модель (5) переводится в режим Z - объема "автоматически" при условии равенства F_{max} и V. По мере того, как при последовательном увеличении исходной выборки F_{max} по абсолютной величине приближается к значению V, эта выборка все больше "начинает походить" по своим вероятносто-статистическим свойствам на выборку Z - объема и наоборот, после достижения Z - объема при дальнейшем последовательном увеличении выборки последняя все больше и больше будет отличаться по значению параметра d и от других характеристик выборки оптимального объема. В этом и состоит действие d-параметра, который выступает в качестве "автоматического параболизатора" частотно-рангового распределения. Сила его воздействия при конструировании самой модели (3) была "синхронизирована" с мерой изменчивости соотношения F_{max} и V.

В работе (Бычков, 1984) показано, что подъязыковые выборки характеризуются некоторыми константными параметрами, прежде всего k и d. Поэтому представляет теоретический и практический интерес возможность выразить величину Z-объема через константы подъязыка. Это позволило бы, с одной стороны, просто и однозначно, априорно вычислять величину оптимального объема, а с другой, более определенно увязало бы этот объем с ципфовскими моделями и их интерпретациями. Прежде чем выразить Z - объем через константы k и d, необходимо сделать уточнение следующего порядка: Z - объем можно исчислять как в единицах N словоупотреблений, так и V лексических единиц, или словоформ (с/ф). Далее их будем обозначать для выборки ципфовского объема N_z и V_z. Оба эти выражения находятся в функциональной взаимосвязи, что следует из (4).

По условиям Z - объема

$$i_{max} = V = F_{max} = kN. \qquad (7)$$

Отсюда и из (4)

$$N_z = d\,k^2 N_z^2 \;, \quad d\,k^2 N_z = 1$$

$$N_z = \frac{1}{d\,k^2} \;. \qquad (8)$$

Из (7) и (8)

$$V_z = k N_z ,$$

$$V_z = \frac{k}{a k^2} = \frac{1}{a k} . \qquad (9)$$

Следовательно, Z-объем, если его выразить соответственно в N словоупотреблений и V словоформ, получает следующий вид

$$Z\text{-объем} = \frac{1}{a k^2} \ с/у$$

$$Z\text{-объем} = \frac{1}{a k} \ с/ф$$

Разделив N_z на V_z, находим, что теоретическая абсолютная средняя частота $\bar{F}_z$ в Z-объеме равна

$$\bar{F} = \frac{1}{k} . \qquad (10)$$

Важнейшей вероятностно-статистической характеристикой является относительная средняя частота $\bar{f} = 1/V$. В Z-объеме из (8) и (9)

$$\bar{f}_z = a k . \qquad (11)$$

Так как параметры k и a являются подъязыковыми эмпирическими константами, то и их комбинации (произведение, обратная величина и т.д.) также являются константными и единственными для подъязыка в целом. Не требуется особых доказательств, что Z-объем в том виде, как он представлен в (8) и (9), также является константным выражением, комплексным характеристическим показателем выборки, не зависящим от ее объема.

Для иллюстрации возьмем английский подъязык электроники (Алексеев, 1968), в котором по данным выборки в 200 тыс. с/у $k = 0{,}0955$, $a = 0{,}000178$. Если расчеты вести в масштабе тысяч с/у и с/ф, то параметр $a = 1{,}78$.

Основные характеристики выборки Z-объема английского подъязыка электроники

	Теоретические (расчетные) величины	Имеющиеся сопоставимые эмпирические данные
N_z	62,24 тыс. с/у	60 тыс. с/у
V_z	5,9I тыс. с/ф	5,96 тыс. с/ф
$\bar{F}_z$	I0,52	I0,39
$\bar{f}_z$	0,000I69	0,000I68

В частотном словаре электроники нет эмпирических данных для выборки объемом 62,24 тыс. с/у, которая теоретичес-

ки по (8) и (9) была определена как $\bar{Z}$-объем, но и имеющиеся почти сопоставимые цифры для выборки 60 тыс. с/у показывают хорошую сходимость искомых величин и, прежде всего, таких основополагающих, как $\bar{F}$ и $\bar{f}$. Как отмечалось, в подъязыке электроники параметры k и d вычислялись предварительно по полной выборке в 200 тыс. с/у. Но эти параметры не изменит своего количественного значения, если их вычислять по данным выборок 150, 100, 50 и так далее тыс. с/у. Следовательно, $\bar{Z}$-объем, "цифровские" средние $\bar{F}_z$ и $\bar{f}_z$ и их производные также остаются неизменными, если их измерять по данным выборок различного объема для одного и того же подъязыка.

Рассмотрим далее, как соотносятся абсолютная и относительная средние частоты в лингвостатистических выборках и, прежде всего, в $\bar{Z}$-объеме. Из (10) и (11)

$$k = \frac{1}{\bar{F}_z} \quad \text{и} \quad k = \frac{\bar{f}_z}{d}$$

Отсюда

$$d = \bar{F}_z \cdot \bar{f}_z \qquad (12)$$

Можно показать, что это равенство действительно не только для $\bar{Z}$-объема, но для выборок большего и меньшего объема. Отсюда также делается понятным дополнительный формальный смысл параметра d кроме того, который был изложен в (Бычков, 1984), где отмечалось, что постоянство величины d на выборках различного объема свидетельствует об относительной однородности лексического массива в подвыборках, относящихся к одному и тому же подъязыку. Понятие относительной однородности означает допустимость включения в исходную лингвостатистическую совокупность определенного количества новых элементов по мере того, как исходная выборка увеличивается по своему объему. Параметр d задает специфическую меру на вхождение новых элементов системы, которое не нарушает общей пропорции "нового и старого" в подъязыковых выборках.

Даже априорно можно утверждать, что в различных подъязыках одного, например английского, языка параметр d может принимать различные количественные значения. Из (4) видно, что в равных по объему выборках чем больше d, тем меньше относительное разнообразие лексического массива. В силу обратной пропорциональной зависимости $\bar{Z}$-объема от величины d (8), (9) в выборках с относительно более "компактным", однообразным лексико-терминологическим составом режим $\bar{Z}$-объема

достигается быстрее, то есть на меньшем объеме выборочной совокупности. Поэтому делается понятным, почему на небольших текстах, которые тематически и, следовательно, лексически ограниченны, Z-объем меньше, чем в сборных и случайных по композиции выборках, в которых лексико-терминологическая система получает более свободное и разнообразное представление. Еще большие различия в величинах параметра d наблюдаются при сравнении выборок из текстов на различных языках, отличающихся разнообразием грамматических форм лексического материала. В этих случаях начинает играть более существенную роль параметр k, которой в подъязыках одного и того же языка является постоянной или практически постоянной величиной, но при сравнении различных языков k варьирует в широком диапазоне. Как было показано выше, k-параметр специфически связан с величиной Z-объема (N_z) и разнообразием его словаря (V_z), из (8), (9) и (10) следует, что

$$k = \frac{V_z}{N_z} = \frac{1}{F_z} \qquad (13)$$

Иначе говоря, параметр закона Ципфа k представляет собой не только максимальную относительную частоту, но и соотношение двух основных характеристик Z-объема - объема его выборки и ее словаря, а также является обратной величиной средней частоты в Z-объеме.

В целом Z-объем представляет собой меру пропорций параметров k и d конкретной лексико-семиотической системы - подъязыка. В силу относительной индифферентности k и d к объему конкретной выборки Z-объем также представляет собой инвариантный, обязательный, надситуативный параметр подъязыковой системы по отношению к другим величинам, характеризующим эту же систему в других выборках - N, V, средние и абсолютные частоты составляющих их лексических единиц, которые являются вариантными, окказиональными показателями одного, часто невоспроизводимого случая.

Назовем комбинацию статистических свойств лексико-семиотической системы (определяемую объемом выборки, составом и соотношением ее элементов и связей) состоянием системного объекта, а совокупность состояний, возможных в различных однородных выборках, областью возможных состояний объекта. Тогда различные по объему и составу лингвостатистические выборки (малые, большие и сверхбольшие) представляют собой различные статистические состояния, в которых конкретная лингвистическая система по-разному в количественном и качественном аспек-

тах проявляет свои системообразующие свойства. Известно, что эти статистические состояния описываются различными ципфовскими аппроксимациями, различающимися между собой прежде всего величиной γ-параметра. Но более существенно то, что по своей сути различия в ципфовских аппроксимациях зависят от того, насколько частотно-ранговое распределение в конкретной выборке отличается от распределения в Z-объеме. Z-объем в качестве специфического параметра распределения входит в модель (3) как обратная величина $d\,k$, а мера отличия выборки от ципфовского объема фиксируется параметром d. Именно поэтому аппроксимация эмпирических распределений с помощью нелинейной модели (3) дает существенно более точные результаты, чем при использовании "канонических" моделей.

В работе (Бычков, 1983) частоты и ранги были интерпретированы как соответственно мера функциональной значимости и семантической сложности языковых знаков, как мера их синтагматической и парадигматической сложности в рамках заданного объема лингвосемиотической системы. Так как в Z-объеме частотам количественно "зеркально" противостоят ранги, то ципфовский объем можно определить как такое состояние семиотической системы, в котором функциональный вес и синтагматическая сложность языковых знаков находятся в строго обратной пропорциональной зависимости от их семантической и парадигматической сложности. В этом смысле одни "сложностные" качества знака уравновешиваются "упрощенностью" других, а общее количество синтагматических и парадигматических связей, которое приходится в среднем на один знак лингвосемиотической системы в таком состоянии, оказывается минимальным, и в этом смысле общее состояние системы, по определению Б. Мандельброта, является более оптимальным, чем у этой же системы с большим числом элементов и их связей. В состоянии Z-объема лингвосемиотическая система, характеризующаяся минимальной средней парадигматико-синтагматической связанностью элемента с системой в целом, имеет соответственно наиболее благоприятные условия для однозначности языковых знаков. При превышении Z-объема происходит абсолютное повышение средней синтагматической и парадигматической сложности языкового знака и тем самым создаются условия для повышения уровня многозначности и омонимии элементов системы. Тогда в едином функциональном подъязыке возникают предпосылки для его разделения на относительно самостоятельные "части" (ξ совместного функционирования резко понижается).

В качестве гипотезы можно предположить, что Z-объем (на примере подъязыка электроники и ряда других исследованных по этому параметру подъязыков - около 6 тыс. лексических единиц) является средним оптимальным числом элементов, которые образуют действительно целостную функциональную систему. Это число может быть несколько больше или меньше ципфовского объема, но в среднем подъязыки "придерживаются" именно этого объема. Этот тезис находит определенное подтверждение в лингвостатистических данных, хотя для более обоснованного вывода необходимо провести более тщательные и разносторонние исследования.

Л И Т Е Р А Т У Р А

Алексеев П.М. Частотный словарь английского подъязыка электроники. - В кн.: Статистика речи. Л.: Наука, 1968, с. 151-161.

Алексеев П.М. О нелинейных формулировках закона Ципфа. - В кн.: Вопросы кибернетики. Вып. 41. Статистика речи и автоматический анализ текста. М.-Л.: Наука, 1978, с. 53-65.

Бычков В.Н. Лингвистическая статистика и проблема эквивалентных статистических описаний (моделей). - В кн.: Структурная и прикладная лингвистика. Л.: ЛГУ, 1983, с. 72-82.

Бычков В.Н. К проблеме обобщения и интерпретации ранговых распределений в статистической лингвистике. - В кн.: Квантитативная лингвистика и автоматический анализ текстов. Тарту,: ТГУ, 1984, с. 61-71.

Мандельброт Б. О рекуррентном кодировании, ограничивающем влияние помех. - В кн.: Теория передачи сообщений. М.: 1957, с. 36-47.

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. Статистическое моделирование речевых потоков. - В кн.: Вопросы кибернетики. Вып. 41. Статистика речи и автоматический анализ текста. М.-Л.: Наука, 1978, с. 66-99.

Фрумкина Р.М. Статистические методы изучения лексики. М.: Наука, 1964, 110 с.

SOME THEORETICAL AND PRACTICAL ASPECTS OF THE SAMPLE OF OPTIMAL VOLUME

Valery N. Bychkov

S u m m a r y

The article deals with general questions of linguostatistical samples of optimal volume in the Zipfian meaning of this term when rank-frequency distributions are strictly and linearly observed. The numerical characteristics of such samples are expressed in constants and their products of a non-linear Zipfian model which secures a better fit between theoretical and empirical data. The optimal volume of a sample and its formal parameters receive a proper linguistic interpretation.

НЕКОТОРЫЕ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ БУРЯТСКОГО ТЕКСТА

(на материале произведений Х.Намсараева)

Г.А. Дырхеева

В бурятском языке язык художественной литературы является наиболее полным, глубоким и всеохватывающим воплощением литературного языка. Поэтому для первого опыта вероятностно-статистического исследования были выбраны прозаические произведения основоположника бурятского литературного языка Хоца Намсараева: рассказы (РА), повести "Цыремпия" (Ц), "Нэгэтэ hүни" (Н.h.), "Илалтын туяа" (И.Т.), "Алтан эзбэ" (А.З.), "Эдиршүүд" (ЭД), "Эжэл гурбан нүхэд" (ЭЖ), "Тэршээхэн унаган" (Т.У.), роман "Үүрэй толон" (РО).

Статистическая обработка данных в области литературы характеризуется множеством разнообразных подходов и решений. Богатые и разнообразные сведения о лексике художественного текста содержит частотный словарь. Частотный словарь по прозе Х.Намсараева составлен с помощью электронно-вычислительной техники. Предварительный, домашинный этап обработки включал сплошную индексацию текста, перфорацию и проверку ошибок перфорации. Индексация текста проводилась с учетом таких особенностей бурятского текста, как парные слова (при перфорации соединены знаком =) и определенные устойчивые словосочетания (соединены знаком ж), а также фразеологизмы. Составление словаря проходило в два этапа. I) На ЭВМ БЭСМ-4М были получены алфавитно-частотные и частотные списки словоформ (общие для всех произведений и по каждому произведению в отдельности). Схема частотного словаря сходна со схемой, принятой в группе "Статистика речи", т.е. при каждой единице списка указывается ранг, абсолютная и относительная частоты, абсолютная и относительная накопленные частоты, количество взвешенной информации и количество накопленной взвешенной

информации. В алфавитно-частотных списках к каждой единице указывалась абсолютная частота и адрес словоформы - номер страницы и строки. 2) После тщательной проверки ошибок и сведения словоформ в слова обработка была продолжена на ЭВМ ЕС-1033. В конечном варианте получен алфавитно-частотный список, упорядоченный в гнездовом виде: при каждом слове приводится перечень его зафиксированных форм с указанием частот (суммарных и отдельно по произведениям).

Поскольку подобное исследование в бурятском языкознании проводится впервые, стояла задача не только дать объективную вероятностно-статистическую модель языковых особенностей выбранных для анализа художественных произведений, но и сопоставить, по возможности, полученные статистические характеристики с имеющимися данными по другим языкам, в особенности с родственными тюркскими языками.

Общий итог машинной обработки текстов дан в таблице I.

Табл.I

№ Назв.произвед.	Объем в стр.	Длина текста - N	Объем словника - L	Кол-во парн. слов	Кол-во уст.с/соч.
I. Ра	I7I	40026	II865	I050	I44
2. Ц	I02	26I8I	8303	666	I20
3. Н.А.	80	20793	7005	249	62
4. И.Т.	85	2I69I	7I04	397	77
5. А.З.	89	22III	6705	207	73
6. ОД	I06	26418	6982	I73	99
7. Эж	36	8902	36I7	49	54
8. Т.У.	I4	249I	I507	79	3
9. Ро	393	I03254	20508	2322	295
Весь корпус:	I076	27I866	36984	5I92	927

Таким образом, весь объем обследованного текста составляет 271866 словоупотреблений, а с учетом парных слов и устойчивых словосочетаний - 277985 словоупотреблений.

Рассмотрим соотношение текст-словарь, наиболее информативное и наиболее исследованное в лингвостатистике. В табл.2 показан процент покрываемости текста группой наиболее частых словоформ и слов (данные по общему списку). Показатели таблицы подтверждают тот факт, что флективно-аналитические и флективно-синтетические языки существенно отличаются от языков агглютинативного строя (Кенесбаев С.К., Бектаев К.Б., Пиотровский Р.Г., 1969, с.3): в бурятском языке, как и в тюркских, для покрытия 50% текста необходимо более 700 самых частых словоформ, что составляет

Табл.2 Процент покрываемости текста группой наиболее частых слов и словоформ.

№	Назв.произв.	I-I0	I-50	I-I00	I-500	I-I000	I-2000
I.	Корпус слов	I5.I5	3I.45	4I.39	68.35	78.95	87.8I
2.	Корпус с/ф	7.22	I7.I7	23.76	44.68	54.99	65.I5
3.	РА	6.8	I6.7I	23.28	44.58	55.09	67.59
4.	Ц	7.94	I8.64	26.I	49.26	60.55	72.II7
5.	Н.Ь.	7.54	I9.5	27.24	50.5	6I.9	73.86
6.	И.Т.	7.8	I8.86	26.4	49.68	6I.52	73.6
7.	А.З.	8.2	20.5I	28.5	52.65	64.46	76.39
8.	ЭД	7.65	I9.8	28.2	54.I	66.25	77.97
9.	ЭЖ	8.4	2I.I	29.I3	54.62	67.89	8I.8
I0.	Т.У.	6.7	I9.I	28.I	59.57	79.65	-
II.	РО	7.44	I7.76	24,6	47.0	57.42	67.85

более I36 тыс. словоупотреблений или I.93% словника (I70 самых частых слов покрывают 50.087% текста - I36I70 словоупотреблений

и 1.44% словника) в отличии от индоевропейских языков, где 50% текста покрываются 100-150 самыми частыми словоформами. Даже небольшой по объему текст - повесть "Тэрвээхэн унаган" указывает на существенные отклонения в данном показателе: около 350 словоформ покрывают 50% текста.

Из выделенных К.Б.Бектаевым (Бектаев К.Б., 1978) типологических критериев, противополагающих агглютинативные языки флективным, на бурятском тексте наиболее явно выполняется критерий темпа роста покрываемого текста наиболее частыми словоформами: так же, как и в казахском языке, первая тысяча словоформ покрывает до 60% текста(значения выше 60% соответствуют произведениям небольшого объема). Процент же покрываемости первой 1000 слов существенно не отличается от процента покрываемости в агглютинативных и флективных языках - 78.95%. Данные таблицы подтверждают ту закономерность, что, чем меньше по объему произведение, тем выше процент покрываемости. Причем эта закономерность, судя по таблице, действует, в основном, для рангов от 1-100 до 1-2000, возможно и выше. Первые же 50 словоформ существенную зависимость от объема выборки не показывают.

Среди показателей по произведениям интересно рассмотреть данные по повести "Алтан зэбэ" и сравнить их с данными по повести "Иралтын туяа", примерно равной по объему "Алтан зэбэ", и с небольшим произведением "Эжэл гурбан нүхэд". Несмотря на то, что длина текста "Алтан зэбэ" в 2,5 раза больше длины повести "Эжэл гурбан нүхэд", её показатели ненамного отличаются от них, они находятся как бы в промежутке между показателями повести "Иралтын туяа" и "Эжэл гурбан нүхэд". Относительно высокая концентрированность словаря "Алтан зэбэ" в высокочастотной зоне говорит, видимо, о стандартном, однообразном языке данного произведения.

Интересно сравнить также строки "Рассказы" и "Корпус слово-

форм". Несмотря на существенную разницу в объеме, их данные до столбца I-500 показывают зависимость обратную отмеченной выше. Это дает нам возможность предположить, что либо рассказы, несмотря на то, что они являются ранними, первыми произведениями писателя, обладают более разнообразной лексикой, либо то, что разнообразие лексики рассказов определяется разнообразием сюжетов. Но, с другой стороны, общий частотный список, составленный по различным произведениям, тоже должен был бы отразить это разнообразие.

В современной лингвостатистике традиционным стало выявление соответствия рангового распределения закону Ципфа. Исследуемые тексты на бурятском языке не являются исключением, составленные нами частотные списки мы попытались описать с помощью закона Ципфа. На рис.I и 2 даны графики зависимостей ранг-частота для слов (общий список) и словоформ (общий список и отдельные, по произведениям). Величины параметров K и γ даны в табл.3.

Данные таблицы подтверждают ту закономерность, что "при увеличении выборок возрастают величины параметров K и γ" (Алексеев П.М., 1983, с.39). Однако, если сопоставить наши значения K и γ со значениями в казахском и других языках, то они окажутся заметно ниже. Чем это можно объяснить? Графики показывают сильное отклонение начальных эмпирических кривых от теоретических влево. Как пишет К.Б.Бектаев, это объясняется предельным синтетизмом агглютинативных языков и "небольшим количеством частых служебных слов начальных лингвистических единиц списка" (Бектаев, 1978, с.68). Коэффициент ρ, отражающий степень отклонения относительных частот самых частых лингвистических единиц от теоретических вероятностей для бурятского текста достигает 9 единиц. Значительные отклонения в зонах частых и малочастотных слов списка компенсируется увеличением средних частотных значе-

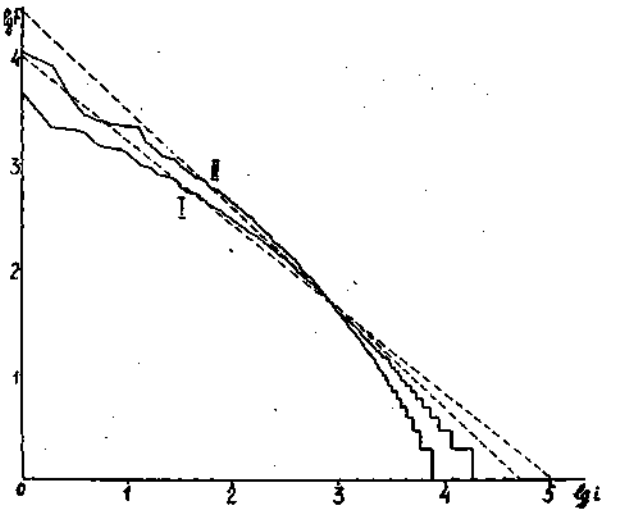


Рис.I Эмпирические и теоретические распределения словоформ (I) и слов (2)

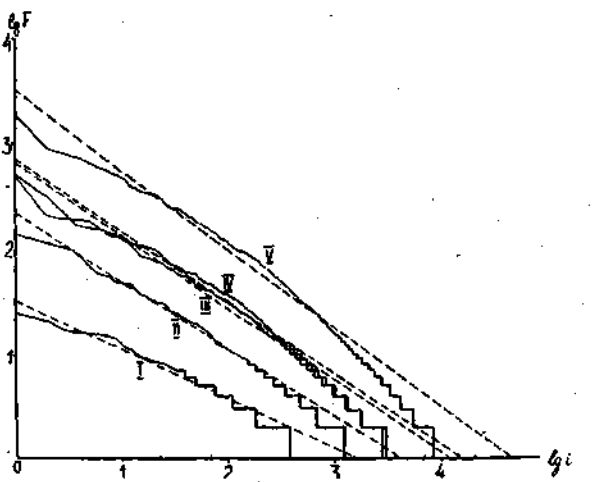


Рис. 2 Эмпирические и теоретические распределения словоформ в повестях "Тэршээхэн унаган" (I), "Эжэл гурбан нүхэд"(II), "Цыремпил" (III), "Эдиршүүд" (IУ) и в романе (У).

Табл. 3 Параметры K и γ для частотных списков слов (общий) и словоформ (общий и частные)

№	Название произв.	N	L	K	γ
I.	Корпус слов	271866	II769	0.I053	0.94I
2.	Корпус словоформ	271866	36984	0.0392	0.8
3.	РА	40026	II865	0.023	0.69
4.	Ц	26I8I	8303	0.0245	0.683
5.	Н.h.	20793	7003	0.024	0.673
6.	И.Т.	2I69I	7I04	0.0253	0.679
7.	А.З.	22III	6705	0.0272	0.688
8.	ЭД	26I8	6982	0.0268	0.68
9.	ЭЖ	8902	36I7	0.0246	0.649
I0.	Т.У.	249I	I507	0.0I35	0.477
II.	РО	I03254	20508	0.03I7	0.75I

ний: на всех графиках заметна выпуклость средних участков графиков над теоретической средней, что указывает на компактность словаря бурятского текста, его большую концентрацию в зоне средних частот. Таким образом, теоретическое значение самого частого слова у нас заметно ниже, чем в индоевропейских языках у служебных слов и соответственно меньше K, график более пологий и угол между теоретической прямой и осью абсцисс незначителен, что и отражено в коэффициенте γ.

В отличие от индоевропейских языков на распределениях бурятских единиц существенно сказывается также различное понимание слова как единицы частотного списка: на рис.1 графики слов и словоформ как бы сместились вокруг некоторой оси. Поскольку большинство наиболее частых словоформ в основном являются изменяемыми формами знаменательных слов во вспомогательных функциях, то при лемматизации частотные значения их основных форм сущест-

венно увеличиваются, хотя в то же время не достигают теоретических. Значительное сокращение словаря, увеличение количества групп равночастотных слов, перераспределение частот по всему словарю привело к тому, что, во-первых, γ намного увеличилась и, во-вторых, графики слов и словоформ сильно отличаются друг от друга.

Наилучшую аппроксимацию теоретической прямой показывает график распределения словоформ повести "Эжэл гурбан нүхэд". Кстати, его коэффициенты K и γ, при относительно меньшем объеме текста ненамного отличаются от K и γ других повестей. При сравнении произведений примерно равного объема, "Цыремпил" и "Эдиршүүд", обнаруживается нарушение указанной закономерности: при относительно меньшем объеме γ повести "Цыремпил" больше γ повести "Эдиршүүд".

Рассмотрим те зоны частотного словаря, которые показывают наибольшее отклонение от теоретической прямой: зону высокочастотных (ВЧЗ) слов и словоформ и группу *hapax legomena* (с частотой I).

200 самых частых слов, выбранных нами для рассмотрения занимают более 50% текста (52.788%), 200 первых словоформ покрывают 31.92% текста. Это еще раз подтверждает тот факт, что в зависимости от того, на базе каких элементов мы составляем словник в агглютинативном языке, существенно изменяется и структура частотного распределения, т.е., чем богаче система словоизменения в языке, тем сильнее разница в распределениях единиц между частотными списками слов и словоформ.

Среди двухсот самых частых слов знаменательные слова (см. таблицы 4 и 5) представляют большую группу - 82,5% и 78.5%. Однако при сравнении доли этих слов, например, в группе слов с F=I: 91.8% и 97.46% или с процентом во всем словаре: 94.08% и 96.6% становится ясно, что они здесь не превалируют.

На первый взгляд сразу бросается существенное преобладание среди слов 243 глагольных основ, покрывающих более 20% текста,

Табл.4 Распределение по частям речи 200 самых частых слов.

Часть речи	Кол-во в зоне ВЧЗ	Доля в ВЧЗ	Доля во всем словаре	Доля во всем тексте
Существительное	53	26.5	0.45	9.85
Прилагательное	I6	8.0	0.I36	2.585
Числительное	3	I.5	0.025	I.93
Местоимение	22	II.0	0.I87	6.074
Глагольная основа	47	23.5	0.399	20.03
Наречие	I6	8.0	0.I36	2.7I
Имя собственное	8	4.0	0.068	I.09
	I65	82.5	I.40I	44.269
Послелог	I3	6.5	0.II	2.35
Частица	I3	6.5	0.II	4.02
Союз	5	2.5	0.043	I.23
Междометие	2	I.0	0.0I7	0.35
Модальное слово	2	I.0	0.0I7	0.33
	35	I7.5	0.297	8.28
Итого:	200	I00.0	I.698	52.549

Табл.5 Распределение по частям речи 200 самых частых словоформ.

Часть речи	Кол-во в зоне ВЧЗ	Доля в ВЧЗ	Доля во всем словаре	Доля во всем тексте
Существительное	25	I2.5	0.0676	3.I44
Прилагательное	I7	8.5	0.046	2.0I2
Числительное	4	2.0	0.0I08	I.07
Местоимение	27	I3.5	0.073	4.46
Глагол	7	3.5	0.0I89	I.099
Причастие	I9	9.5	0.05I4	3.297
Деепричастие	28	I4.0	0.0757	5.407
Наречие	24	I2.0	0.0649	3.I05

I	2	3	4	5
Имя собственное	6	3.0	0.0I6	0.6095
	I57	78.5	0.4099	24.2035
Послелог	I8	9.0	0.049	2.225
Частица	I4	7.0	0.038	3.6II
Союз	5	2.5	0.0I35	I.I23
Междометие	2	I.0	0.005	0.33
Модальное слово	4	2.0	0.0I08	0.427
	43	2I.5	0.II63	7.7I6
Итого:	200	I00.0	0.5406	3I.9I95

Первые три ранга в частотном списке слов занимают глаголы гэ- 'говорить' (F=II50I), бай- 'быть, находиться' (F=8434), боло- 'становиться' (F=4I06). Естественно, что это объясняется многочисленностью их форм (глагол гэ- имеет в своем гнезде I47 словоформ, бай- - I83, боло- - I46) и многообразием выполняемых в предложениях функций: кроме прямого значения они часто используются в качестве вспомогательного глагола и служебных слов.

Количество существительных в данной группе хоть и больше, чем глаголов, однако в сумме они покрывают только 9.85% текста. Наименьший ранг - 4 - слова хүн 'человек' (гнездо из 7I словоформы), как и в частотном списке словоформ - 9, объясняется тем, что, во-первых, оно относится к обобщающим словам и, во-вторых, что оно очень часто используется в значении неопределенного местоимения 'кто-то, кто-либо', также как и второе по частоте существительное юмэн 'вещь' в значении 'что-то, что-нибудь'.

Показательным для словаря Х.Намсараева является наличие таких частых слов как колхоз 'колхоз', түрүлэгшэ 'председатель', ноён 'нойон, князь', баян 'богач, богатство', лама 'лама, буддийский монах', гулваа 'голова (в селении)'. Эти слова темати-

ческие, сюжетнозависимые, их положение в частотном списке объясняется преобладанием в отдельных произведениях.

Самым частым среди прилагательных является слово ехэ; кроме основного значения 'большой', оно очень часто употребляется как усилительное слово 'очень, сильно, весьма', поэтому в частотном списке словоформ оно стоит на втором месте после прилагательного хара 'черный'. Высокая частота прилагательного хара, как и в казахском языке, объясняется тем, что это слово обозначает не только цвет, но и такие качественные характеристики как простой, злой, темный, самый.

Послелоги - одна из многочисленных групп служебных слов, входящих в высокочастотную зону обоих списков. Однако, если в частотных словарях русского языка предлог 'в' обычно имеет самые высокие ранги, то в бурятском языке послелоги, в своей общей характеристике сходные с предлогами в русском языке, имеют максимальный ранг 29 - дээрэ 'на, над' и 32 - тээшэ 'в сторону, по направлению'.

В частотных словарях слов и словоформ всех языков группа *hapax legomena* составляет большую часть. Бурятский язык не является исключением: 18895 словоформ (51.09% от словаря словоформ и 6.95% от всего текста) и 3939 слов (33.47% от всего словаря слов и 1.45% от всего текста) составляют элементы словарей с частотой 1 (см. табл.6). Подобное соотношение, когда количество словоформ с частотой 1 составляет более 50% от объема словаря имеет место во всех отдельных списках по произведениям. При этом наблюдается стандартная закономерность: чем меньше по объему произведение, тем больше доля группы *hapax legomena*, тем больший объем текста они покрывают (см. коэффициент ξ).

Что касается использования суммарной вероятности редкоупотребляемых слов в качестве квантитативно-типологического критерия, (Бектаев, 1978, с.52), то для сравнения по объему у нас подходит

Табл.6 Статистические характеристики слов и словоформ с частотой I.

№ Назв. произв.	Кол-во слов с F=I	Доля в словаре	ξ
I. Корпус слов	3939	33.47	98.55
2. Корпус с/ф	I8895	5I.09	93.05
3. РА	7408	62.44	8I.49
4. Ц	5306	63.9	79.73
5. Н.h.	45I7	64.5	78.3
6. И.Т.	4538	63.88	79.I
7. А.З.	4I89	62.48	8I.05
8. ЭД	4I44	59.35	84.3
9. ЭЖ	2376	65.69	73.3I
I0. Т.У.	III7	74.I2	55.I6
II. РО	II953	58228	88.42

только роман, в частотном списке которого слова с частотой I и 2 составляют 72.9%, что несколько ниже указанной Бектаевым границы для агглютинативных языков.

Высокий процент "одноразовых"слов в произведениях Х.Намсараева говорит, во-первых, о богатстве грамматического оформления бурятских слов, во-вторых, косвенно о богатстве языка Х.Намсараева, об использовании им редких слов.

Однако, сравним, как рекомендуется, тексты, имеющие одинаковый объем: повести "Цыремпил" и "Эдиршүүд". Слов группы *hapax legomena* в "Эдиршүүд" существенно меньше, чем в "Цыремпиле". Кроме того, повесть "Эдиршүүд", несмотря на небольшую разницу в объеме текста, имеет меньший словарь, что, в частности, отразилось бы на показателях средней частоты и коэффициента разнообразия, также характеризующих лексическое богатство

текста, и на проценте покрываемости текста группой наиболее частых слов: повесть "Эдиршүүд" имеет словарь более концентрированный в ВЧЗ, чем "Цыремпил". Эти данные хорошо согласуются с имеющимися литературоведческими описаниями, а также с интуитивными ощущениями о лексическом преимуществе "Цыремпила".

Элементарные расчеты, которые описаны здесь, дают возможность судить некоторым образом о типологических особенностях бурятского языка и, конкретно, языка писателя. Рассмотренные параметры тесно взаимосвязаны и подтверждают мнения литературоведов о богатстве словарей отдельных произведений Х.Намсараева, а также дают дополнительные сведения для выводов о степени родства монгольских и тюркских языков.

ЛИТЕРАТУРА

Алексеев П.М. Методика квантитативной типологии текста. Л., 1983.

Бектаев К.Б. Статистико-информационная типология тюркского текста. Алма-Ата, 1978.

Кенесбаев С.К., Бектаев К.Б., Пиотровский Р.Г. О статистическом изучении тюркских текстов. - В кн.: Тезисы докладов и сообщений. Статистическое и информационное изучение тюркских текстов. Алма-Ата, 1969.

SOME STATISTICAL CHARACTERISTICS OF BURYAT TEXTS
(ON THE MATERIAL OF H.NAMSARAEV's PROSAIC WORKS)

Galina Dyrheyeva

S u m m a r y

The article presents the first attempt of statistic investigation of the Buryat text for the purpose of which the prosaic works of the classic of Buryat literature H. Namsaraev were used, with the total number of about 300,000 word-tokens. Basing on the frequency list, compiled according the computer data, there were found some characteristics, namely: 50% of Buryat text is composed of above 700 most frequent wordtypes; the values of the parameters k and γ (Zipf LAW) are considerably less than in Indoeuropean languages, while the values of ρ are considerably higher. The article presents as well a brief characteristics of the high frequency zone of vocabulary and of the group of hapax legomena with their frequency distribution in word classes.

СТАТИЧЕСКИЙ АСПЕКТ СОДЕРЖАНИЯ ТЕКСТА И ЕГО ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ

А.В. Зубов

Многоаспектные исследования различных текстов показывают, что каждый текст по своей структуре является многоплановым и противоречивым образованием. В нем можно обнаружить социальное и индивидуальное, детерминированное и случайное, обязательное и факультативное, содержательное и формальное, глубинное и поверхностное, языковое и речевое (Проблемы ..., 1983, с.43).

К тем же противоречивым особенностям текстов относится и наличие в тексте статического и динамического аспектов. Термины "статический" и "динамический" по отношению к тексту употребляются обычно в смысле "покой" и "движение" (Гальперин, 1981, с.19). Текст как последовательность лексических единиц, как некоторый результат, продукт речемыслительной деятельности находится в статическом состоянии или в состоянии покоя. Текст же в процессе его порождения, восприятия и понимания считается находящимся в движении (Новиков, 1983, с.31).

Мы рассмотрим проблему статики текста с несколько иной точки зрения. Как отмечают многие исследователи, при воспроизведении предложений, запоминаются две группы данных о предложении: информация о его семантическом содержании и информация о его синтаксической структуре. При этом семантический аспект запоминается в качестве первого шага, а синтаксический - в качестве второго (Леонтьев, 1969, с.89). В дальнейшем это положение было развито по отношению к тексту и свелось к констатации того, что в каждом тексте есть свой словарь и свой синтаксис (Кацнельсон, 1972, с.123; Купина, 1978, с.141; Проблемы..., 1983, с.56). По-разному понимаются при этом такие составляющие как

"словарь" и "синтаксис". Ближе всего нам та точка зрения, когда под словарем понимаются так называемые слова "содержание" или "несомые слова", а под синтаксисом имеется в виду грамматический строй вместе с "несущими" или служебными словами (Проблемы..., с.44). Таким образом, можно представить, что слова "содержание" представляют статику текста и являются отражением в тексте некоторого множества предметов, явлений, фактов реальной действительности, а синтаксис - динамику текста, отражающую те отношения между этими предметами, фактами, явлениями, которые устанавливает автор текста в зависимости от целевой установки, типа текста, речевого опыта и целого ряда других факторов. Наличие этих двух составляющих в тексте подтверждают и эксперименты по анализу процесса понимания текста (Лингвистические..., 1983, с. 120).

Рассмотрим подробнее, что представляет из себя статический аспект содержание текста с точки зрения его организации как единого целого.

Понятие "содержание текста" трактуется различными учеными по-разному. Например, одни в содержании текста выделяют содержательно-фабульную, содержательно-концептуальную и содержательно-подтекстовую составляющие (Гальперин, 1981, с.27). Другие в этом едином понятии подчеркивают три таких вида содержания, как доминирующее логическое содержание, предметное (тематическое) и информационное (общесмысловое) (Ванников, 1984, с.49). Первый из двух перечисленных подходов более подходит для литературоведческих рассмотрений проблемы организации текстов. Принимая второй подход, мы отнесем логическое содержание текста к его динамическому аспекту[1] и остановимся на

[1]Подробнее см., например (Зубов, 1985, с.23-29).

анализе тематического и информационного содержания текста. По существу, это – две ступени одной и той же информации – информации о составляющих ситуации, описанной в тексте. Можно согласиться с М.А.К.Хэллидеем (Halliday , 1970, 160-164) в том, что тематическая информация является основной информацией, отражающей главные субъекты, объекты, места и времена действия описываемой ситуации. Информационное же содержание, являясь второстепенным, лишь раскрывает детали ситуации, в которой приходится "действовать" главным субъектам и объектам.[1]

Как же можно формально задать такую градацию составляющих содержания текста?

Любая ситуация, отраженная в тексте, может быть рассмотрена с той или иной позиции. Помимо этого, различный отбор элементов ситуации связан с тем, что человек воспринимает действительность в условиях некоторой вероятности, зависящей от его жизненного и языкового опыта (Борель, 1964, с.81; Гальперин, 1981, с.40). Все это приводит к тому, что в процессе развертывания содержания текста автор проводит выделение элементов ситуации, приписывая им определенные значения для данной ситуации. В итоге содержание текста определяется его денотативной[1] структурой, формируемой в сознании человека (Голод, Шахнорович, 1981, с.240; Новиков, 1983, с.64-72; Шахнорович, 1979, с.11-15) и реализуется в тексте отбором конкретных слов и их распределением по тексту. Причем такой отбор осуществляется таким обра-

[1]Здесь и далее речь идет о текстах традиционной семантики, содержание которых совпадает с объективным содержанием, передаваемым семантическими структурами предложений (Шрейдер, 1976, с.154-155).

[2]В данной работе денотат понимается как некоторое неделимое отражение в сознании человека некоторого объекта внешнего мира-референта.

зом, чтобы потенциальный получатель текста мог сформировать в своем сознании то же представление об описываемом фрагменте ситуации, которое вложил в текст автор.

Отбор конкретных слов и включение их в различные типы сочетаний (т.е. конкретизация распределения слов в тексте) проводятся в зависимости от темы текста ("о ком (чем) будет текст?"). Тема определяет предметно-тематическую область словаря человека, где должен вестись поиск указанных слов содержания (Новиков, Чистякова, 1981, с.53; Чистякова, 1979, с.104). Несколько изменяя формулировку Е.Агриколы (Agricola , 1976, S.15) под темой будем понимать понятийное ядро текста, передающее содержание текста в сжатом и абстрактном виде, в форме смыслового комплекса, выраженного словесными средствами, отражающими главные действующие лица (или главные предметы, явления, факты и т.п. описания, повествования или рассуждения). Следовательно, тема - это совокупность определенных слов и словосочетаний (ср. Смит, 1980, с.336). Это как раз то инвариантное, что содержится в тексте и одинаково выделяется и автором текста и основной массой воспринимающих текста. Наличие таких инвариантных смысловых единиц подтверждается и психолингвистическими экспериментами по смысловому анализу и синтезу текстов различными испытуемыми (Апатова, 1974; Едина, 1976, Смирнов, 1966).

Что же это за слова и словосочетания?

Интуитивно ясно, что каждый текст представляет собой единое смысловое целое потому, что в нем говорится во всех его частях "об одном и том же", т.е. на всем протяжении текста действуют одни и те же субъекты и объекты, что события в относительно небольшом законченном тексте происходят в какое-то ограниченное время и в фиксированном числе мест. Но для пере-

дачи описываемой ситуации с индивидуальной точки зрения автора, с индивидуальными целями и намерениями, такие составляющие, как мы уже отмечали, могут быть неравноценны для автора. Эта неравноценность приводит к тому, что часть из них действует в большинстве микроситуаций описываемой единой ситуации, - другие же - лишь в некоторых из них. Следовательно, и среди имен текстов, представляющих упомянутые составляющие, имеется определенная иерархия, соотносимая с иерархией соответствующих денотатов Лингвистические..., 1983, с.123; Большунов, 1977; Глаголев, 1976).

Прежде всего, каждая ситуация передается в тексте конкретным автором. Именно его жизненный опыт, его восприятие мира отражено в тексте. Поэтому имена текста, отражающие личность самого автора, будут являться для нас главными опорными словами текста (ГОС).

К этой группе слов мы отнесем и те слова текста, которые передают главные действующие лица, главные объекты действия, основные места действия и времена действия. Как правило, эти слова имеют в тексте наибольшие частоты употребления и употребляются в большем числе фрагментов (абзацев) текста, образуя соответствующие "изотопические цепочки", "тематические прогрессии", "номинативные цепи", "рекуррентные цепи", "семантические анафоры", "имплицитные референции", "нанизывания" (Москальская, 1981, с.19; Проблемы..., 1983, с.49). Основная особенность таких последовательностей лексических единиц текста заключается в том, что они имеют в тексте одну и ту же предметную соотнесенность. В таком ряду могут присутствовать и полные словарные синонимы, и контекстуальные синонимы, и ассоциативные замены, и, наконец, замены местоименные (Леонтьева, 1981, с.23-25, Рубашкин, 1983, с.61-62). В нашу задачу не входит детальный анализ

способов равноценных замен в указанных семантических цепочках. Будем считать, что указанные выше 4 способа являются основными способами денотативного отождествления имен.

Такие слова в дальнейшем становятся своеобразным центром, вокруг которого формируются другие элементы, отражающие составляющие микроситуации (глаголы, прилагательные, наречия и т.п.) (Новиков, Чистякова, 1981, с.53).

Главные референты, которым в тексте соответствуют главные опорные слова, в рамках той или иной ситуации могут быть связаны с другими референтами, также являющимися относительно важными для той же ситуации, но которые являются центром не всей ситуации, а некоторых микроситуаций общей ситуации. Соответствующие им слова текста назовем второстепенными опорными словами (ВОС). Они встречаются с меньшей частотой, чем ГОС и в меньшем числе абзацев всего текста.

Существует достаточно большое число способов определения "степени важности" для содержания текста того или иного слова или словосочетания (Белоногов, Кузнецов, 1983, с.235-236; Марусенко, 1983, с.85-87; Рылова, 1973). Их можно разделить (Иванкин, 1973) на: 1) анкетные, 2) структурные, 3) словарные, 4) частотные, 5) синтаксические, 6) диалоговые.

Все эти методы имеют один существенный недостаток: они не исходят из текста как связной единицы. Например, с помощью структурных методов выделяются опорные слова на основе их вхождения в заголовок, первое предложение или в несколько отдельных предложений. Частотные методы исходят из частоты встречаемости слов в тексте. При таком подходе по существу учитывается не смысл текста, а свойства плана выражения текста (Гиндин, 1977, с.13).

Другие, существенные для нас недостатки характеризуют и иные способы поиска опорных слов текста.

Как мы отметили выше, основными критериями при выделении опорных слов текста являются для нас абсолютная частота употребления слова (с учетом всех его возможных синонимов и замен) и количество абзацев, в которых встретилось слово. Помимо этого, такой критерий не должен зависеть от общего числа слов в тексте. Наиболее удобной для нас представляется несколько измененная формула коэффициента статистической устойчивости термина (Марусенко, 1983, с.87):

$$K_{важ.} = \frac{F \cdot m}{N \cdot n}$$

В этой формуле: F - абсолютная частота слова в тексте (в нее входит суммарная частота всех типов синонимов этого слова его ассоциативных и местоименных замен), m - число абзацев, в которых встретилось слово; N - общее число слов в тексте; n - общее число абзацев в тексте.

Назовем этот критерий $K_{важ.}$ коэффициентом важности слова и определим для слов всех анализируемых текстов его критические значения $K^1_{важ.}$ и $K^2_{важ.}$, позволяющие формальным способом отделить в массе слов конкретного текста соответственно главные и второстепенные опорные слова.

Эксперименты по выделению опорных и неопорных слов научных, публицистических и поэтических текстов[1], проведенные на текстах самой различной длины (от одного абзаца до 60-ти) позволили нам выделить следующие критические значения коэффициента важности:

1) к главным опорным словам текста будет относить те слова, которые удовлетворяют требованиям, определенным в

[1]Подробнее см. (Зубов, 1985, с.172-193).

Таблица 1

Критерии отнесения слов текста к главным опорным словам

Общее число абзацев в тексте	Критерии отнесения слова к главному опорному слову	
	число абзацев, в которых встретилось слово	пределы изменения $K^{I}_{важ.}$
более 100	$\frac{1}{10}n$ и более	$\frac{(\frac{1}{10}n+1)(\frac{1}{10}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 81 до 100	$\frac{1}{8}n$ и более	$\frac{(\frac{1}{8}n+1)(\frac{1}{8}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 51 до 80	$\frac{1}{7}n$ и более	$\frac{(\frac{1}{7}n+1)(\frac{1}{7}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 31 до 50	$\frac{1}{5}n$ и более	$\frac{(\frac{1}{5}n+1)(\frac{1}{5}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 15 до 30	$\frac{1}{4}n$ и более	$\frac{(\frac{1}{4}n+1)(\frac{1}{4}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 9 до 14	более 4 и 4	$\frac{(\frac{1}{3}n+1)(\frac{1}{3}n+1)}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
от 6 до 8	более 3	$\frac{9}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
5	2 и более	$\frac{4}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
4	2 и более	$\frac{4}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
3	2 и более	$\frac{3}{N \cdot n} \leqslant K^{I}_{важ.} < 1$
2	2	$\frac{3}{N \cdot n} \leqslant K^{I}_{важ.} < 1$

Таблица 2

Критерии отнесения слов текста ко второстепенным опорным словам

Общее число абзацев в тексте	Критерии отнесения слова ко второстепенным и опорным словам	
	число абзацев, в которых встретилось слово	пределы изменения $K^2_{вал.}$
более 100	$\frac{1}{15}$ и более	$\frac{(\frac{1}{15}n)(\frac{1}{15}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{10}n+1)(\frac{1}{10}n+1)}{N \cdot n}$
от 81 до 100	$\frac{1}{12}$ и более	$\frac{(\frac{1}{12}n)(\frac{1}{12}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{8}n+1)(\frac{1}{8}n+1)}{N \cdot n}$
от 51 до 80	$\frac{1}{10}$ и более	$\frac{(\frac{1}{10}n)(\frac{1}{10}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{7}n+1)(\frac{1}{7}n+1)}{N \cdot n}$
от 31 до 50	$\frac{1}{8}$ и более	$\frac{(\frac{1}{8}n)(\frac{1}{8}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{5}n+1)(\frac{1}{5}n+1)}{N \cdot n}$
от 15 до 30	$\frac{1}{6}$ и более	$\frac{(\frac{1}{6}n)(\frac{1}{6}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{4}n+1)(\frac{1}{4}n+1)}{N \cdot n}$
от 9 до 14	$\frac{1}{4}$ и более	$\frac{(\frac{1}{4}n)(\frac{1}{4}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{(\frac{1}{3}n+1)(\frac{1}{3}n+1)}{N \cdot n}$
от 6 до 8	2 и более	$\frac{(\frac{1}{4}n)(\frac{1}{4}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{9}{N \cdot n}$
5	1 и более	$\frac{(\frac{1}{4}n)(\frac{1}{4}n)}{N \cdot n} \leqslant K^2_{вал} < \frac{5}{N \cdot n}$
4	1 и более	$\frac{2}{N \cdot n} \leqslant K^2_{вал} < \frac{4}{N \cdot n}$
3	1 и более	$\frac{2}{N \cdot n} \leqslant K^2_{вал} < \frac{3}{N \cdot n}$
2	1	$\frac{2}{N \cdot n} \leqslant K^2_{вал} < \frac{3}{N \cdot n}$

таблице 1;

2) к второстепенным опорным словам того же текста относятся те его слова, которые соответствуют требованиям, сформулированным в таблице 2.

Как мы отмечали выше, слова, входящие в каждую из двух полученных групп, неоднозначны по своему содержанию. В соответствии с предметными свойствами своих референтов они образуют группы опорных слов - субъектов, объектов, слов-мест и слов-времени. Именно они совместно с предикатами выражают взаимосвязи основных объективно существующих категорий: материи, движения, времени и пространства (Лосева, 1980, 15-51).

ГОС и ВОС текста оформляются в специальную таблицу основного статического содержания текста (ТОСС).

Например, для научного текста, приведенного в Приложении[1], ТОСС выглядит так (таблица 3).

Таблица 3

Таблица основного статического содержания научного текста Н26

Тип опорных слов	Опорные слова текста			
	субъекты		объекты	
	код	слово	код	слово
Главные-опорные слова	С01	автор	К01/01[2]	анализ
	С02	испытуемый	К01/02	исследование
	С03	группа	К02	вопрос
			К03	беседа
			К04	возможность
Второстепенные опорные слова	S01	люди	G01	выделение
			G02/01	данные
			G02/02	ответы
			G02/03	факты
			G03	качество
			G04	опыт

[1] См. ниже, стр. 46

[2] Коды, стоящие после знака "/", указывают на порядковый номер словарного или контекстуального синонима или ассоциативной замены.

Но ТОСС - это только основное содержание, тема текста. Однако для читающего текст, в зависимости от цели, с которой он читает текст, могут оказаться важными и те составляющие описанной в тексте ситуации, которые не нашли отражение в словах, зафиксированных в ТОСС. Соответствующие слова, очевидно, будут определять специфику конкретного абзаца, как основной единицы письменного текста (подробнее см.: Зубов, 1985, с.33-39).

Поэтому, рассмотрим теперь несколько подробнее проблему смысловой организации абзаца. Отмечая, что эта проблема является в целом семантико-синтаксической, исследователи в то же время подчеркивают, что синтаксические особенности этой единицы плана содержания не могут быть поняты без детального проникновения в ее лексическую структуру. Для микротемы, передаваемой абзацем, так же как и для всей темы характерны иерархия и повторяемость опорных слов и словосочетаний. Действительно, как мы отмечали выше, абзац содержит отражение некоторого фрагмента ситуации. По-существу, в абзацах отражаются те связи и отношения, которые зафиксированы у главных и второстепенных референтов ситуации с другими, менее значимыми элементами ситуации: различными типами действий и состояний ГОК и ВОК, различными местами действий, временами действий, целями, причинами действий и состояний и т.п. (Новиков, Чистякова, 1981, с.53-54). Известны различные способы представления иерархии имен, описывающих в пределах микротемы абзаца все эти многоаспектные отношения. Один из подходов связан с использованием различного рода дескрипторов. При этом выделяются дескрипторы номенклатурного характера, предикативного или модального характера, указатели отношений и некоторые другие (Маслов,1975,

с.28). При другом подходе предлагается дифференцировать слова предложений абзаца по той роли, которую выполняют их референты в ситуациях реальной действительности. По существу этот второй подход связан с фиксацией тех энциклопедических знаний, которыми владеет автор текста. Нам представляется, что этот подход является наиболее рациональным. Ведь создавая тот или иной текст, автор видит перед собой реально (или мысленно) некоторую ситуацию. Он может выделять фрагменты этих ситуаций, а в пределах фрагментов - элементы ситуации со всеми относящимися к этим элементам признаками и из связи с другими элементами той же ситуации. Такие часто повторяющиеся зависимости есть не что иное как энциклопедические знания, и поэтому проблема представления энциклопедических знаний и проблема организации словаря текста должны рассматриваться в неразрывном единстве. Существует огромное число исследований по проблемам представления знаний в системах, работающих с естественным языком (см. библиографию в работе: Клещев, 1983). Ближе всего к задачам формального описания подходит упомянутый выше "ролевой" подход, когда в представлении необходимой информации указывается слово с конкретным значением и та роль, которую выполняет это слово в общем описании микроситуации, представленной в абзаце (ср.Сильдмяэ, 1980, с.111-115, Нйм, 1980, с.87). Учитывая те роли, которые выделены в упомянутых и иных работах, и проводя анализ наших текстов по такому же принципу, мы выделили следующие роли, которые выполняют слова в предложениях абзаца (назовем их словами-конкретизаторами) (таблица 4).

Таблица 4

Роль слов-конкретизаторов в общей структуре абзаца

№№ пп	Наименование роли	Код роли	Класс слов, которые выполняют эти роли
1.	Действие	D	Глагол
2.	Состояние	C	Глагол
3.	Место	P	Существительное
4.	Время	T	Существительное
5.	Средство	S	Существительное
6.	Причина	R	Существительное
7.	Цель	L	Существительное
8.	Условие	I	Существительное
9.	Характеристика	X	Причастие, Существительное
10.	Детализация	Z	Существительное
11.	Способ	M	Существительное
12.	Адресат	A	Существительное
13.	Пример	E	Совокупность слов, относящихся к разным классам

При описании содержания абзаца после символов, обозначающих роли и коды главных и второстепенных опорных слов-субъектов, слов-объектов, слов-мест и слов-времен , указываются через двоеточие конкретные слова абзацев, которые выполняют эти роли. Если роль не относится ни к одному из перечисленных кодов, то после индекса роли ставится код 099. Если при этом используются роли "действие" и "состояние", то после кода 099 между знаками указывается "исполнитель" этой роли.

Например, для абзаца № 2 упомянутого текста Н26 таблица слов-конкретизаторов представлена на рис.1 (помня ТОСС этого текста, табл.3).

XH03 : процесс			D K02 : задавались		
D K02 : задавался			XS0I : отношении		
D C02 : оценивает	1-ое		ZS0I : родных	2-ое	
XC02 : характер	предло-		ZS0I : колхозников	предло-	
D C02 : отличается	жение		ZS0I : жителей	жение	
XC02 : черты			M099 : деревни		
XC02 : недостатки					
D C02 : может отметить-					

Рис.I
Таблицы слов-конкретизаторов абзаца №2 текста Н26

Кроме опорных слов и слов-конкретизаторов в любом тексте существует еще одна группа слов, которую мы назовем словами-наполнителями. Они "наполняют" полученное статическое содержание качественными признаками, т.е. указывают качества предметов, названных именами существительными, признаки действий, определяемых в тексте глаголами и причастиями (например, в упомянутом 2-ом абзаце текста Н26 словами-наполнителями являются: "проводимый", "собственным", "других", "положительные", "аналогичные", "знакомых"). Такая качественная детализация осуществляется путем предварительной фиксации наиболее вероятных для текстов определенного типа контактных двухсловных беспредложных словосочетаний (именных, адъективных, адвербиальных). Эти словосочетания также являются отражением в памяти человека определенных энциклопедических знаний, полученных человеком в результате собственного опыта жизни и социальных знаний, выработанных всем человечеством ("небо синее", "снег белый", "цена товара", "шапка отца", "поступить правильно" и т.п.). Такие словосочетания мы называем ЛЕ-сочетаниями (элементарными сочетаниями лингвистических единиц). Они задаются в виде

специальных списков, выделяемых статистическим путем.[1]

Таким образом, таблица основного статического содержания, списки слов-конкретизаторов к абзацам и списки ЛЕ-сочетаний, типичных для текстов определенной узкой предметной области являются формальным представлением статического аспекта содержания текста, относящегося к данной предметной области.

[1] Подробнее см. (Зубов, 1985, с.168-172)

Приложение

Научный текст Н26[1]

"Примененный нами метод исследования прост .[2]

В процессе проводимой беседы испытуемому задавался вопрос, как он оценивает свой собственный характер, чем он отличается от других людей, какие положительные черты и какие недостатки он может отметить у себя. Затем аналогичные вопросы задавались в отношении других людей - его родных, знакомых колхозников, жителей той же деревни.

Понимая ограниченные возможности таких приемов исследования, мы анализировали не столько конкретное содержание получаемых ответов и выделяемые испытуемым частные свойства, сколько самую возможность сделать предметом своего анализа свои психические качества, сознательно относясь к ним. С особенным вниманием нам хотелось бы отметить факты, которые указывали бы, что на определенных этапах развития выделение внутренних психических качеств заменяется выделением внешних обстоятельств, бытовых нужд, поступков и т.п. Мы хотели бы остановиться на сравнении тех данных, которые получены при беседе с испытуемыми разных групп, имеющих неодинаковые формы общения и разный уровень образования.

В данной серии опытов участвовало 52 испытуемых; из них 20 относились к нашей исходной группе (дехкане отдельных кишлаков, неграмотные), 15 были активными членами коллективных хозяйств, которые имели опыт в коллективном обсуждении хозяйственных вопросов, и 17 были учащимися техникумов или лицами, прошедшими хотя бы кратковременное систематическое обучение.

Основной материал был собран автором, дополнительный - В.В.Захаровой."

[1]См. (Лурия, 1974, с.150).

[2]Одной чертой в этом тексте выделены ГОС, двумя - ВОС.

Л И Т Е Р А Т У Р А

Апатова Л.И. Текст как система ориентиров в процессе понимания иноязычной речи на слух. – В кн.: Лингвистика текста. Материалы научной конференции. Часть I. М.: Наука, 1974, с. 21-27.

Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. – М.: Наука, 1983. 287с.

Большунов Я.В. Воспроизведение семантически неравноценных частей связного текста. – Вопросы психологии, 1977, № 2, с. 114-120.

Борель Эмиль Вероятность и достоверность. М.: Наука, 1964. – 119 с.

Ванников Ю.В. Начальная субкатегоризация научно-технических и официально-деловых текстов. – В кн.: Семантика текста и проблемы перевода. Сборник статей. М.: Изд. ИЯ АН СССР, 1984, с. 41-52.

Гальперин И.Р. Текст как объект лингвистического исследования. – М.: Наука, 1981. – 139 с.

Гиндин С.И. Семантика текста и различные теории информации. – Научно-техническая информация, сер. 2, 1977, № 10, с. 10-15.

Глаголев Н.В. Семантико-структурные элементы актуализации сообщения в сверхфразовом единстве. – В кн.: Проблемы синтаксической семантики. Материалы научной конференции. М.: 1976, МГПИИ им. М. Тореза, с. 104-107.

Голод В.И., Шахнарович А.М. Семантические аспекты порождения речи. – Изд. АН СССР. Серия лит. и языка, 1981, т. 40, № 3, № 3, с. 237-244.

Елина Е.Е. К проблеме "опорных единиц" текста. – В кн.: Лексико-грамматическая сочетаемость в германских языках. Выпуск I. Челябинск, 1976, с. 38-44.

Зубов А.В. Вероятностно-алгоритмическая модель порождения текста (семантико-синтаксический аспект). Дис. ... док. филол. наук. Минск: 1985. – 399 с.

Иванкин В.И. О методах выбора ключевых слов при координатном индексировании (постановка проблемы). – Научно-техническая информация, сер. 2, 1973, № 11, с. 14-18.

Кацнельсон С.Д. Типология языка и речевое мышление. – Л.: Наука, 1972. – 216 с.

Клецев А.С. Представление знаний. Методология, формализмы, организация вычислений и программная поддержка. – В кн.: Прикладная информатика. Выпуск I. М.: Финансы и статис-

тика, 1983, с. 49-84.

Кулина Н.А. Опыт системно-синтаксического анализа семантики связного текста. - В кн.: Семантика и структура предложения. Лексическая и синтаксическая семантика. Уфа: изд. Башкирск. гос. унив., 1978, с. 137-141.

Леонтьев А.А. Психолингвистические единицы и порождение речевого высказывания. М.: Наука, 1969. - 307 с.

Леонтьева Н.Н. Семантика связного текста и единицы информационного анализа. - Научно-техническая информация, сер. 2, 1981, № 1, с. 21-29.

Лингвистические вопросы алгоритмической обработки сообщений. - М.: Наука, 1983. - 246 с.

Лосева Л.М. Как строится текст. Пособие для учителей. - М.: Просвещение, 1980. - 94 с.

Лурия А.Р. Об историческом развитии познавательных процессов. - М.: Наука, 1974. - 172 с.

Марусенко М.А. О формировании словника словаря статически устойчивых научно-технических терминов. - В кн.: Структурная и прикладная лингвистика. Межвузовский сборник. Выпуск 2. Л.: ЛГУ, 1983, с. 82-89.

Маслов Б.А. Проблемы лингвистического анализа связного текста (Межфразовый уровень). Учебное пособие к спецкурсу. - Таллин: 1975. - 104 с.

Москальская О.И. Грамматика текста. - М.: Высшая школа, 1981. - 151 с.

Новиков А.И. Семантика текста и ее формализация. - М.: Наука, 1983. - 216 с.

Новиков А.И., Чистякова Г.Д. К вопросу о теме и денотате. - Известия АН СССР. Серия лит. и языка, том 40, № 1, М.: 1981, с. 48-56.

Проблемы текстуальной лингвистики. - Киев: Изд. КГУ, 1983. - 175 с.

Рубашкин В.Ш. О методах анализа связного текста (к проблеме применения семантических моделей текста в документально-фактографических ИПС). - В кн.: Вопросы информационной теории и практики. № 49, М.: ВИНИТИ, 1983, с.58-73.

Рылова Т.Н. Принципы упорядочения и группировки значимых слов текста для совершенствования критерия семантической связи между предложениями. - Научно-техническая информация. Сер. 2, № 4, 1973, с. 34-38.

Сильдмяэ И. Гносеологическая структура знаний. - В кн.: Семантика и представление знаний. Труды по искусственно-

ну интеллекту. Часть II. Тарту, 1980, с. 108-117.

Смирнов А.А. Проблемы психологии памяти. - М.: Просвещение, 1966. - 423 с.

Смит Дж.Е. Тематическая структура и тематическая сложность. - В кн.: Новое в зарубежной лингвистике. Вып. IX. М.: Прогресс, 1980, с. 333-355.

Чистякова Г.Д. Смысловая структура текста как определяющий фактор его понимания. - В кн.: Семантика, логика и интуиция в мыслительной деятельности человека. М.: Педагогика, 1979, с. 101-126.

Шахнарович А.М. Место семантического компонента в языковой способности. - В кн.: Семантика языковых единиц и текста (лингвистические и психолингвистические исследования). - М.: 1979, с. 5-18.

Шрейдер Ю.А. Текст, автор, семантика. - В кн.: Семантика и информатика. Вып. 7. М.: ВИНИТИ, 1976, с. 153-169.

Ыйы Х. Эпизоды в структуре дискурса. - В кн.: Представление знаний и моделирование процессов понимания. Новосибирск, 1980, с. 79-86.

Agricola E. Vom Text zum Thema. - Studia Grammatica XI. Probleme der Textgrammatik. - Berlin: Akademie Verlag, 1976, S. 13-27.

Halliday M.A.K. Language structure and language function. - In: J. Lyons (Ed.-) New Horizons in Linguistics. Middlesex, 1970, p. 160-164.

STATICAL ASPECT OF THE TEXT CONTENTS AND ITS FORMAL PRESENTATION

Alexandr V. Zubov

S u m m a r y

Statical aspect of the text is understood in this article as a complex of three groups of lexical units. The first group contains the most frequent words of the text (including dictionary and contextual synonyms, associative and pronominal substitutes) which are divided into primary and secondary key words according to special criteria. The second group consists of the words specific for each paragraph of the text. The third group of lexical units comprises words-modifiers set in the form of parts of two-word contact non-prepositional word combinations typical for the texts of the sublanguage considered.

НЕЙРОФИЗИОЛОГИЧЕСКИЕ ПРЕДЕЛЫ ПАМЯТИ ЧЕЛОВЕКА И БОГАТСТВА ЕГО ЛЕКСИКИ

А.Н.Лебедев

Индивидуальная память человека, вмещающая талле и запас слов, велика, но не беспредельна. Память заполняется информацией постепенно, в течение жизни, небольшими порциями, квантами. В опыте для запоминания одной порции информации, например тестовой строки из 6-8 случайно выбранных слов или цифр, требуется не меньше одной секунды. Простой расчет показывает, что человек может быть занят запоминанием чего бы то ни было не более, чем в течение одного миллиарда секунд предназначенной ему жизни. Следовательно, в памяти человека может разместиться не больше одного миллиарда квантов информации, эквивалентных тестовой строке. Следует учесть, что к тому же все вновь запоминаемое тут же начинает разрушаться сначала быстро, затем все медленнее и медленнее. И не каждая воспринимаемая порция сведений новая, и не каждый миг нашей жизни укладывается в память. Таков первый, часто китейский ориентир для расчета верхней границы памяти человека.

Словарные запасы занимают небольшую часть памяти, и все же при этом в любом языке они достаточно велики, исчисляясь порой сотнями тысяч слов. Однако словарные запасы в памяти отдельного человека, богатство его лексики обычно много меньше. Неисчерпаемы лишь комбинации из слов, тексты. Письменные тексты можно использовать для анализа механизма хранения сведений в памяти и их произвольного извлечения.

Простейшая, качественно давно известная закономерность, имеющая отношение к механизму хранения слов и их извлечения из памяти, следующая.

По мере возрастания объема текста постепенно, сначала быстро, затем все медленнее и медленнее возрастает разнообразие употребляемых в тексте слов. При этом, разумеется, отношение объема словаря к объему текста постепенно падает.

Ю.А.Тулдава развил теорию, согласно которой на каждом шаге по ходу порождения текста осуществляется выбор между "новым", т.е. ранее не появившемся словом, и "старым", уже употребленным в данном тексте/Тулдава Ю., 1980/. Эта теоретическая посылка реализована им в форме двупараметрического уравнения, выражающего зависимость объема словаря от объема текста. Формула Тулдавы оказалась более точной сравнительно с другими известными формулами, также выражающими искомую зависимость, но выведенными из других предпосылок. Исходы выбора каждого очередного слова в тексте Ю.А.Тулдава вполне обоснованно связал с ассоциативными свойствами мышления и ограниченностью человеческой памяти.

Нейрофизиологический механизм процедуры выбора был раскрыт в наших прежних работах /Забродин Ю.М., Лебедев А.Н., 1977; Лебедев А.Н., 1985/. В настоящей публикации содержится попытка связать наш прежний подход с решением задачи определения объема словаря по объему текста, количественно, с привлечением единых нейрофизиологических параметров, объяснить особенности памяти человека и указанную особенность его речи. В основе наших представлений лежат известные нейрофизиологические данные, в том числе собственные.

По нашему предположению, воспринимаемый мир отражается в структурах мозга многообразными, весьма изменчивыми пространственно-временными комбинациями из возбужденных и заторможенных центральных нейронов, в том числе нейронов- детекторов и командных нейронов /Соколов Е.Н.,1981; Хьюбел Д.,Визель Т., 1984/. Каждый нейрон, возбуждаясь, генерирует группу чаще всего из 3-4 импульсов с межимпульсными интервалами порядка 10 мс /Бехтерева Н.П.,1980/. Множество нейронов, объединяясь в микроансамбль, порождает одиночную пачку из синхронных импульсных разрядов, число которых в последовательности может быть больше числа импульсов в групповом разряде каждого одиночного нейрона. Каждая коллективная пачка это как бы буква нейронного кода. Сколько разных пачек по числу последовательных синхронных импульсов в каждой, столько и разных букв в алфавите нейронных кодов. Пачки возникают друг за другом с короткими интервалами между их передними фронтами. Цепочка из нескольких пачек порождается макроансамблем нейронов и образует кодовое слово. Сколько разных кодовых слов, столько и разных макроансамблей. Каждый образ памяти, каждый семантический квант, каждое слово речи связаны с активностью определенного макроансамбля нейронов. Таковы нейрофизиологические предпосылки для расчета верхних границ человеческой памяти и словарного запаса, в частности.

Минимально возможный интервал между импульсами внутри пачки и между передними фронтами разных пачек равен длительности одиночного нервного импульса, т.е. одной миллисекунде (Δt = I мс).

Типичный интервал на порядок больше (ρ =I0 мс). В опытах М.Н.Ливанова с регистрацией активности одновременно двух нейронов было обнаружено, что импульс (или группа их) одного из нейронов через период около I0-I3 мс сопровождается импульсацией второго нейрона /Ливанов М.Н.,I979/. Отдельные пачки и цепочки из них способны регулярно повторяться. Для человека типичны периоды повторения, равные t =I00 мс. Такова периодичность биопотенциалов в диапазоне альфа-ритма, ярко выраженного у большинства людей,особенно в состоянии покоя.

Максимально возможное число кодовых слов, при самых малых межимпульсных интервалах не превышает величины $(t/\rho)\uparrow(t/\rho)$, где стрелка служит символом возведения в степень. Основанием степени служит объем алфавита кодовых букв, а показателем - число букв в кодовом слове. Это очень большое число, и оно служит мерой иконической (в целом, сенсорной) памяти, ответственной за богатство воспринимаемых одновременно деталей, за всю полноту переживаемых при этом ощущений. Однако узоры из столь плотно упакованных импульсных последовательностей неустойчивы. Такие кодовые слова быстро, не более чем за секунду разрушаются, но не до конца. Устойчивыми к разрушению оказываются комбинации из остаточных, следовых импульсных разрядов, также образующих пачки и цепочки из них, но с интервалами на порядок большими. Разнообразие остаточных пачек по числу импульсов в каждой не больше, чем отношение t/ρ. Число пачек, т.е. кодовых букв в каждом кодовом слове также не больше указанного отношения. Одна из кодовых букв выполняет функцию пробела между кодовыми словами, поэтому разнообразие кодовых слов, хранимых в памяти, не превышает вели-

чины, заданной уравнением

$$V_{max} = \left(\frac{t}{\rho} - 1\right) \uparrow \left(\frac{t}{\rho} - 1\right), \qquad (I)$$

где стрелка по-прежнему является символом возведения в степень. Основание степени равно объему алфавита кодовых букв, а показатель – числу букв в кодовом слове.

С учетом типичных значений физиологических параметров t =100 мс и ρ =10 мс находим по формуле (I), что число всех кодовых слов, способных разместиться в памяти человека, не превышает 0,4 миллиарда, и эта оценка не противоречит первоначальной. Таков объем словаря всех образов памяти, включая образы всевозможных событий, действий, понятий, образы целей, коды мотивов и эмоциональных переживаний. Объем словаря родного языка человека, измеряемый самое большее десятками тысяч слов, представляется ничтожно малым сравнительно со всем содержимым памяти.

По формуле (I) рассчитывается максимально возможный объем кратковременной памяти человека и, следовательно, объем одной порции восприятия элементов, объем M алфавита которых известен

$$M^{H} \leq V_{max}, \qquad (2)$$

где H – искомый объем кратковременной памяти, определяющий, в частности, длину предложений, их структуру.

Скорость, с которой осуществляется выбор, например, од-

ного слова из нескольких возможных, была определена нами ранее из тех же нейрофизиологических предпосылок / Забродин Ю.М., Лебедев А.Н., 1977/. Приведем здесь конечный результат, формулу для расчета задержки, необходимой для осуществления выбора. Если V – объем словаря кодовых слов, из которых осуществляется выбор, то задержка t_V рассчитывается по уравнению

$$t_V = \frac{t^2}{2\rho}\left(1 - \frac{1-\rho/t}{V}\right)^2 \qquad (3)$$

с параметрами, заданными выше.

Значение аргумента V в формуле (3) изменчиво в случае порождения высказывания. Оно связано с контекстом высказывания и обычно невелико. Его верхний предел связан с цикличностью повторения кодовых слов.

Из-за цикличности повторения кодовых слов образы памяти, в том числе и образы слов разговорной речи имеют возможность актуализироваться либо в каждом цикле, либо через один цикл, через два цикла, через три и т.д. Вероятности их актуализации, ранжированные по величине, образуют гармонический ряд, сумма членов которого равна единице:

$$\frac{P_1}{1}; \frac{P_1}{2}; \frac{P_1}{3}; \ldots \frac{P_1}{i} \ldots \frac{P_1}{V}, \qquad (4)$$

где V – максимальный ранг, равный объему словаря актуализируемых образов; P_1 – максимально возможная вероятность актуализации образа, имеющего первый ранг (эта вероятность является строго определенной функцией от числа актуализированных образов).

Нейронные ансамбли взаимодействуют, в частности, конкурируют, вовлекая в свои ритмы разнообразные нейроны, принадлежащие к разным ансамблям, к разным кодовым словам. Образ памяти максимально актуализирован, если число нейронов, вовлеченных в ритмы этого образа, максимально. Повторяемость определенного слова в тексте тем выше, чем чаще актуализируется образ этого слова в памяти, причем указанная связь необязательно линейная.

В тексте объемом из N слов самое редкое слово из всех повторяющихся слов повторится дважды, не больше; иначе оно перейдет в класс неповторяющихся слов из данного текста. Из соотношения (4) находится ранг такого слова, равный $P_1N/2$.

Общее количество D всех повторяющихся слов находим, решая следующее уравнение:

$$D = P_1 N \sum_{i=1}^{i = P_1 N/2} \frac{1}{i} = P_1 N \left(\ln \frac{P_1 N}{2} + 0{,}577 \right). \qquad (5)$$

Это уравнение также вытекает из закономерности, выражаемой уравнением (4).

Разность $N - D$ равна числу однократно встречающихся слов, т.е. равна объему их алфавита, словаря. Суммируя найденную разность с объемом словаря $P_1 N/2$ повторяющихся слов, находим искомую формулу для расчета объема словаря V по объему текста из N слов:

$$V = N - D + \frac{P_1 N}{2} . \qquad (6)$$

Чем больше актуализированных образов, тем меньше веро-

роятность p_1 актуализации доминирующего слова, но она не может быть, по нашему предположению, меньше отношения $p_0 = p/t$. Следовательно, и объем словаря одновременно актуализированных образов памяти (необязательно осознаваемых) имеет свой предел, объективно заданный соотношением p/t. Прежде чем перейти к расчету этого предела, учтем следующие важные обстоятельства.

Прежде всего опыт показывает, что очень часто (и это чуть ли не правило) вероятность слова, занимающего первый ранг, много меньше отношения p_0, опускаясь нередко до 0,05. При этом нарушается и гармоничность распределения частот всех других слов, а именно, слова в начале и в хвосте ряда (4) имеют заниженные значения вероятности появления в тексте /Туллава Д.А., 1985/ в то время как слова, занимающие середину распределения, подчиняются гармонической зависимости. Такого рода искажения замечены многими авторами. На наш взгляд, их можно объяснить, непостоянством ранга одного и того же слова в разных участках текста и тем более в разных текстах. Кроме того следует учитывать, что разные слова речи, особенно слова-синонимы или указательные местоимения вместе с заменяемыми словами, могут порождаться актуализацией одного и того же образа памяти. Слово нельзя рассматривать как единственно возможную единицу членения речевого потока. Эта идея уже обсуждалась /Борода Ю.Г., Поликарпов А.А., 1984/. Известен также феномен неравномерности повторения слов в тексте: повторные появления то учащаются, то урежаются.

Мы учли наблюдаемые особенности, определив опытным пу-

тем следующую зависимость

$$p_1 = p_0 - k \ln \frac{N}{N_0} \qquad (7)$$

где k – коэффициент неравномерности, рассчитываемый по наблюдаемой зависимости объема словаря от объема текста; p_0 –нижний, физиологический предел вероятности актуализации самого частого образа памяти, равный отношению $p_0 = \rho/t$, где t =100 мс, ρ =10 мс; N_0 – критический объем текста, имеющего словарь, объем которого V_0 является функцией параметра p_0 в соответствии с распределением (4):

$$V_0 = p_0 N_0 .$$

Подставляя в формулу (6) значение $N = N_0$ и $p_1 = p_0$ находим критический объем текста $N_0 = \exp\left(\frac{1}{p_0} - \ln p_0 - 0{,}384\right)$ Объем словаря V_0 такого текста, равный примерно 15 600 слов, указывает на предел числа образов памяти, способных к одновременной актуализации, т.е. к взаимодействию. Вероятно, это верхняя граница словарного запаса отдельного человека, активно используемого им в высказываниях.

Достаточно двух параметров p_0 и k для прогноза объема словаря по объему текста с помощью формул (6) и (7), причем с точностью почти не уступающей известной, наилучшей на сегодняшний день модели Ю.А.Тулдава. В таблице содержатся опытные данные множества разных авторов, собранные вместе этим исследователем /Тулдава Ю.А.,1980/ и наши расчетные данные. Важно то, что опытные значения параметра p_0 оказались, как правило, равными приблизительно 0,1 для различных текстов. Такое сходство с предполагаемым отношением физиологических параметров ρ/t побуждает к дальнейшей проверке

Таблица

Зависимость объема словаря от объема текста по опытным данным разных авторов /Тулдава Ю., 1980/ и согласно теоретическому расчету по формулам (6) и (7)

Язык	Объем текста	Объем словаря		Параметры уравнений	
		В опыте	По расчету	P_0	K
I	2	3	4	5	6
Латышский	50 000	7 065	7 066	0,I035I6	0,005481
	I00 000	9 834	I0 334	"	"
	200 000	I3 389	I4 I9I	"	"
	300 000	I6 I03	I6 I02	"	"
Чешский	25 000	4 829	4 830	0,I00624	0,005656
	75 000	9 603	9 762	"	"
	I25 000	I3 056	I3 I72	"	"
	I75 000	I5 858	I5 858	"	"
Казахский	25 000	9 088	9 087	0,090377	-0,000242
	50 000	I5 047	I4 993	"	"
	I00 000	23 895	23 60I	"	"
	I50 000	29 785	29 785	"	"
Польский	I2 I72	3 434	3 434	0,I02970	0,002814
	29 787	6 I46	6 I29	"	"
	48 255	8 026	8 032	"	"
	64 5I0	9 250	9 25I	"	"

(Продолжение на следующей странице)

Таблица (продолжение)

I	2	3	4	5	6
Украинский	5 000	I 629	I 629	0,I05325	0,004I06
	I0 000	2 637	2 659	"	"
	I5 000	3 504	3 489	"	"
	20 000	4 I95	4 I96	"	"
Английский	50 495	4 87I	4 872	0,II2246	0,007391
	I00 970	6 858	6 885	"	"
	20I 966	9 470	9 437	"	"
	302 I56	II 3I4	II 30I	"	"
	403 966	I2 975	I2 973	"	"
Английский	50 000	5 399	5 399	0,I09730	0,006920
	I00 000	7 853	7 724	"	"
	I50 000	9 36I	9 334	"	"
	200 000	I0 582	I0 582	"	"
Румынский	50 000	6 785	6 785	0,I0394I	0,005997
	I00 000	I0 28I	I0 079	"	"
	I50 000	I2 477	I2 443	"	"
	200 000	I4 292	I4 292	"	"
Русский	50 000	9 463	9 464	0,095808	0,00420I
	I00 000	I4 062	I4 633	"	"
	I50 000	I7 263	I8 456	"	"
	200 000	2I 468	2I 467	"	"

Примечание: в текстах на латышском языке подсчитывались лексемы, в остальных текстах - словоформы.

гипотезы о циклических нейронных процессах как основе квантитативных закономерностей речи.

Стоит подчеркнуть, что нелинейность прироста словаря, выражаемая формулой (6) зависит как и в модели Ю.А.Тулдава, от логарифма объема текста.

Поиск параметров, имеющих одинаковое значение и одинаковый смысл в разнообразных уравнениях квантитативной лингвистики и вместе с тем в уравнениях нейрофизиологии и экспериментальной психологии, представляется весьма перспективным. Такие параметры доступны двойной и даже тройной объективной оценке по лингвистическим, психологическим и электрофизиологическим данным.

Неотложным представляется решение проблемы мельчайших семантических квантов речи и других всевозможных проявлений деятельности человека, включая его музыкальное и художественное творчество.

В заключение выражаю сердечную благодарность Анатолию Анатольевичу Поликарпову и его сотрудникам, а также Юрию Константиновичу Крылову за подробное обсуждение идей, положенных в основу нашей публикации. Я благодарен Юхану Артуровичу Тулдаве за внимание к нашей работе и ее поддержку.

ЛИТЕРАТУРА

Бехтерева Н.П. Здоровый и больной мозг человека. М.:Наука, 1980.- 208 с.

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

Забродин Ю.М., Лебедев А.Н. Психофизиология и психофизика.- М.:Наука,1977.- 288 с.

Лебедев А.Н. Кодирование информации в памяти когерентными волнами нейронной активности.- В кн.: Психофизиологические закономерности восприятия и памяти.- М.: Наука, 1985, с.6-33.

Ливанов М.Н. Роль временного фактора в деятельности нейронов коры головного мозга при явлениях обучения и следовых состояниях.- В кн.: Нейрофизиологические основы памяти.- Тбилиси: изд-во "Мецниереба", 1979, с.8-20.

Соколов Е.Н. Нейронные механизмы памяти и обучения.- М.: Наука, 1981. - 140 с.

Туллава Ю. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. - Учен. зап. Тартт. ун-та, вып. 549, Труды по лингвостатистике. Тартту, 1980, с. 113-144.

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

Хьюбел Д., Визель Т. Центральные механизмы зрения. - В кн.: Мозг. - М.: Мир, 1984, с.167-198.

NEUROPHYSIOLOGICAL LIMITS OF MAN'S MEMORY AND OF HIS VOCABULARY

Arthur N. Lebedev

The information of any kind stored within human brain is coded by chains of coherent neural multiple discharges according to author's hypothesis. These chains (or neural words) can repeat periodically. The mean period of repetitions is about t=100 ms, and critical difference of various periods within alpha band of electroencephalogram is about Δt = 10 ms.

Both parameters determine quantitatively the short (H) and longterm (V_{max}) memory span as well as the limit V_0 of individual vocabulary of a man eaqual to nearly 15 000 semantic units.

The functional interrelationship between the size N of text and size V of vocabulary is described.

The Tuldava's idea /J.Tuldava,1980/ relative the limits of human memory as one of determinant of the size of human vocabulary is succesfully proved.

ВОЗМОЖНОСТИ АВТОМАТИЧЕСКОЙ ПЕРЕРАБОТКИ ТЕКСТА С ПОМОЩЬЮ МИКРО-ЭВМ "АГАТ"

К.Я. Лепа

Одной из важнейших тенденций в развитии науки в последние десятилетия стало широкое использование электронно-вычислительных машин для решения целого ряда теоретических и прикладных задач. Возможность передать ЭВМ трудоемкие, рутинные вычисления и измерения позволяет существенно сократить сроки их решения и освобождает время исследователей для творческих проблем, повышая тем значительно продуктивность научных исследований.

И в лингвистике ЭВМ уже давно используются с большим успехом. Решение целого ряда задач лингвистической статистики связано с анализом больших текстовых массивов, ручная обработка которых требовала бы больших затрат времени. Особенно использование ЭВМ активировалось с появлением машин третьего и четвертого поколений, являющиеся более быстродействующими, обладающие большей памятью и лучше приспособленные для обработки текстовой информации (Якубайтис Т.А., 1979, с. 127); в их применении имеется уже большой опыт (Пиотровский Р.Г., 1975; Зубов А.В., 1977).

Но развитие вычислительной техники в последние десятилетия связано не только с появлением все более мощных и быстродействующих ЭВМ, но и в той же мере с распространением малой вычислительной техники, мини-ЭВМ и микро-ЭВМ. Именно эта техника должна играть важную роль при осуществлении массовой компьютеризации нашего общества (Велихов Е., Ершов А., Лавров С., Громов Г., 1985). Уже планируется ее использование в процессе школьного обучения (Дьячко А.Г., 1985; Ломов Б.Ф., 1985), все активнее она используется и в разных отраслях науки, в том числе не только в естественных науках (Баркалов Н.Б., Тихонриров А.А., 1985).

Поэтому сейчас во всей нашей стране проходит активное ознакомление научных сотрудников и преподавателей с основами информатики, вычислительной техники и программирования. В ходе таких курсов при Тартуском государственном университете возник вопрос, можно ли использовать персональные ЭВМ для автоматической обработки текстов. Уже априорно понятно, что речь здесь может идти только о текстах сравнительно малых

размеров, т.к. объем памяти и быстродействие микро-ЭВМ низки. Но в целом ряде случаев надо анализировать и тексты малых размеров, например, для студенческих курсовых работ. Так как в ТГУ все студенты при выполнении научных работ имеют свободный доступ к микро-ЭВМ типа "АГАТ", то было решено рассматривать возможности именно данной ЭВМ.

Несколько слов о технических характеристиках ЭВМ "АГАТ": емкость оперативной памяти составляет 64 килобайта, из которых пользователю для программы и данных остается свободными около 38 килобайт. Ее быстродействие - 300 000 операций в секунду (что приблизительно соответствует 1000 арифметическим действиям). Для хранения программ и данных предусмотрено использование гибких магнитических дисков с диаметром 13 см (емкость 256 килобайт). К микро-ЭВМ можно присоединить принтер. Программировать ее можно на языках высокого уровня типа БЕЙСИК или РАПИРА. Для составления программы, рассматриваемой в данной статье, был избран язык БЕЙСИК, по причине того, что он является одним из наиболее распространенных языков программирования для персональных компьютеров и обладает неплохими предпосылками для обработки текстовой информации - для этой цели имеется целый ряд специальных операторов.

Чтобы выяснить, какие возможности перечисленные технические данные открывают для обработки текстов, была поставлена задача составить программу, которая установила бы абсолютную и относительную частотность слов в вводимом в ЭВМ тексте, и выпечатала бы полученные результаты в порядке уменьшения частотности слов в тексте, в при равной абсолютной частотности аранжировала бы слова в алфавитном порядке (т.е. составила бы частотно-алфавитный словарь всех словоформ данного текста). Такая программа содержит основной элемент обработки всякой текстовой информации - сортировку данных, и может легко быть модифицирована для выдачи какой-нибудь другой элементарной информации о тексте.

Первой задачей при составлении такой программы является выбор подходящего алгоритма сортировки. При этом надо учитывать ограниченную оперативную память ЭВМ "АГАТ" и малую скорость - алгоритм должен быть достаточно быстрым, а программа, составленная на его основе, не должна занимать слишком много места в памяти, чтобы можно было ввести больше данных.

Алгоритмы сортировки являются одной из наиболее обследованных областей теории алгоритмов (см. например, Кнут Д., 1978); это и неудивительно, так как с упорядочением данных в

ЭВМ связано большинство задач обработки информации. При этом различаются 2 основных вида упорядочения информации при помощи ЭВМ - внутренняя и внешняя. Первая связана с сортировкой данных, введенных в оперативную память ЭВМ; объем обрабатываемой информации при этом, естественно, ограничивается объемом оперативной памяти. При внешней сортировке данные считываются в оперативную память и сортируются там по отдельным "порциям". Мы пока рассматриваем только внутреннюю сортировку.

Из вышеупомянутых соображений в качестве алгоритма сортировки был выбран метод Шелла, преимущества и недостатки которого подробно изложены в специальной литературе (см., например, Кнут Д., 1978; с. 105-119; Мейер Б., Бодуэн К.,1982, с. 148-152). Сущность его вкратце такова: сначала сортируются все подмассивы данного сортируемого массива, элементы которых находятся друг от друга на расстоянии определенного шага, потом шаг уменьшается и сортировка повторяется. Большая скорость данного метода по сравнению с методом простых вставок, к которому он восходит, объясняется тем, что в начале сортировки, когда упорядоченность элементов в массиве мала, рассматриваются очень короткие последовательности элементов, а в конце сортировки, когда последовательности сортируемых элементов длинные, они обладают уже значительной упорядоченностью, чем сильно уменьшается число необходимых перестановок. По словам Б. Мейера и К. Бодуэна, проводивших практические опыты с различными методами сортировки, метод Шелла является одним из наилучших для массивов от нескольких сот до нескольких тысяч (т.е., именно для текстов, которых, по-видимому, можно обрабатывать на микро-ЭВМ).

На основе данного алгоритма была составлена программа, которая приведена в Приложении 1. Она состоит из следующих частей:

1) резервирование памяти и дименсионирование массива переменных;

2) ввод анализируемого текста, записанного предварительно на диск в качестве текстового файла; при этом программа сама выделяет словоформы и определяет их число;

3) алфавитная сортировка словоформ по методу Шелла с определением абсолютной частотности (величина шага определяется по формуле $h_{(k)} = (3^k - 1)/2$ (см. Кнут Д., 1978, с. 116- 118) и сохраняется в памяти ЭВМ);

4) вывод результатов на принтере с исчислением относительной

частотности в процентах (вывод происходит в два этапа - сначала определяется самая большая частотность и выводится соответствующая словоформа, а потом выводятся остальные словоформы в порядке убывания частотности).

Приводим и пример программы, записывающей исследуемый текст на диск (см. Приложение 2). И здесь в начале следует резервирование памяти и дименсионирование массива переменных. Дальше текст из блока данных считается в оперативную память ЭВМ (строки 30 - 60) и записывается на диск (строки 70-120). Строки 70, 80, 120 должны содержать имя текстового файла (в данном случае - БАЛЛАДА). В блоке данных в строке 140 должно содержаться число текстовых блоков, а с строки 150 - сами текстовые блоки (их число зависит от длины текста, с учетом того, что после каждого оператора DATA нельзя стоять больше чем <55 знаков, заключенных в кавычки).

В ходе испытания данной программы интерес представили в первую очередь две проблемы:

- тексты какого объема можно обрабатывать при помощи микро-ЭВМ;
- время работы данной программы.

На первый вопрос можно сначала дать теоретическую оценку, исходя из вышеупомянутого свободного для пользователя памяти в ЭВМ, который составил 38 000 байт. Так как для хранения одной буквы требуется объем памяти в один байт, а среднюю длину немецкого слова можно оценивать на 6-7 букв, то можно предположить, учитывая и необходимый резерв памяти для программы, что верхняя граница лежит где-то между 2500 - 3000 словами.

Для демонстрации работы программы был составлен частотно-алфавитный словарь всех словоформ баллады Ф. Шиллера "Der Ring des Polykrates". Отрывки из баллады и из выпечатки приведены в Приложениях 3 и 4. Для обработки данного текста, длина которого составляет 558 словоформ, микро-ЭВМ понадобилось около 5 минут. Но при увеличении длины текста наблюдалось неприятное обстоятельство - время обработки данных стало очень быстро увеличиваться (при тексте с длиной в 827 словоформ - около 13 минут, при тексте в 1000 словоформ больше 20 минут); т.е., выяснилось, что программа не ведет себя по закономерности, соответственно которой время обработки данных по методу Шелла зависит от числа данных по формуле c·n log n. Что дело здесь не в программе, доказывается тем, что такое же замедление наблюдается, если не увеличивать чис-

ло обрабатываемых слов, а только резервированную для данных область памяти. Если в приведенной программе, например, писать вместо приведенного варианта строки 10: 10 DIM WS(2500), F (2500), то время работы программы при сортировке тех же 558 слов увеличивается почти на 8 раз. Это объясняется принципом работы данного типа ЭВМ: они не проводят операций в области памяти, резервированной для данных, а переставляют их для этого в свободную область; если последняя мала, то число необходимых внутренних перестановок быстро растет и они занимают при DIM WS (2500), например, около 14/15 всего времени работы программы.

Конечно, положение можно немножко улучшить, если усовершенствовать программу или применить другие, более подходящие для данного типа ЭВМ методы сортировки. Так, можно в показателе абсолютной частотности, переменной F заменить вещественные числа на целые (вещественное число занимает в памяти 5 байт, целое, обозначение которого %, только 2 байта), чем при DIM F% (2000) достигается уже экономия памяти в 6 килобайт. Так как большинство времени при обработке текстов все равно уходит на ввод данных, то можно использовать программу, при которой слова вводятся отдельно и сразу переставляются на правильное место в составляемом алфавитном ряду. Немного улучшило бы положение также применение какого-нибудь метода цифровой кодировки вводимых слов.

В заключение можно констатировать, что при помощи микро-ЭВМ типа "АГАТ", по-видимому, целесообразно обрабатывать тексты, объем которых не превышает 1500 словоупотреблений. Это, конечно, немного, но этому положительно противопоставляется легкость обращения с микро-ЭВМ, которое происходит в режиме диалога: данные вводятся с клавиатуры и сразу видны на экране дисплея, ошибки могут быть сразу замечены и быстро исправлены, исправление можно производить даже после первой обработки данных и сразу повторять обработку, и, что также немаловажно, использование микро-ЭВМ не требует длительной подготовки и специальных знаний.

Приложение I

Программа обработки текста (на языке БЕЙСИК)

```
8    HIMEM: $77FF
10   DIM W$(1000),F(1000)
15   INPUT " ЧИСЛО ТЕКСТОВЫХ БЛОКОВ :";N
20   I = 1
25   PRINT CHR$(4);"OPEN  БАЛЛАДА"
30   PRINT CHR$(4);"READ  БАЛЛАДА"
40   FOR CLAUSE = 1 TO N
50   INPUT W$
60   LEFT = 0:RIGHT = 0
70   FOR J = 1 TO LEN(W$)
80   LEFT = RIGHT:IF RIGHT = 0 THEN LEFT = 1
90   IF MID$(W$,J,1) = " " THEN RIGHT = J: W$(I)=MID$(W$,LEFT,
   RIGHT - LEFT): RIGHT = RIGHT + 1:I = I + 1
100  NEXT J
110  NEXT CLAUSE
115  PRINT CHR$(4);"CLOSE  БАЛЛАДА"
120  N = I - 1
125  PRINT "ЧИСЛО СЛОВ В  ТЕКСТЕ:";N
130  REM --- ИЗЧИСЛЕНИЕ ШАГА ---
140  K = 1
150  H(K) = (3^K - 1)/2
160  IF H(K) > N/3 GOTO 190
170  K = K + 1
180 GOTO 150
190  L = K - 1
195  REM ---  СОРТИРОВКА  МЕТОДОМ ШЕЛЛА ---
200  FOR K = L TO 1 STEP -1
210  FOR I = H(K) + 1 TO N
220  B$ = W$(I)
230  FOR J = I - H(K) TO 1 STEP - H(K)
240  IF W$(J) <= B$ GOTO 270
250  W$(J+H(K)) = W$(J)
260  NEXT J
270  W$(J + H(K)) = B$
280  NEXT I
290  NEXT K
300  REM ================
310  REM ---  АБСОЛЮТНАЯ ЧАСТОТНОСТЬ ---
320  FOR I = 1 TO N
```

```
330  F(I) = 1
340  IF WS(I) = WS(I - 1) THEN F(I) = F(I - 1) + 1:F(I - 1) = 0
350  NEXT I
400  REM ---  ВЫВОД РЕЗУЛЬТАТОВ ---
410  M = 0
420  R = 1
430  FOR I = 1 TO N
440  IF F(I)  = M GOTO 470
450  M = F(I)
460  P = I
470  NEXT I
480  HTAB(1):PRINT R;".";WS(P);:HTAB(25):PRINT F(P);:HTAB(30):
   PRINT INT((F(P)/N)*100000 + 0.5)/1000
490  F(P) = 0
500  FOR J = M TO 1 STEP -1
510  FOR I = 1 TO N
520  IF F(I) = J THEN R = R + 1:HTAB(1):PRINT R;".";WS(I);:HTAB
   (25):PRINT F(I);:HTAB(30):PRINT INT((F(I)/N)*100000 + 0.5)/
   1000
530  NEXT I
540  NEXT J
550  END
```

Приложение 2

Программа, записывающая текст на диск

```
10   HIMEM: $77FF
20   DIM W$(600)
30   READ N
40   FOR I=1 TO N
50   READ W$(I)
60   NEXT I
70   PRINT CHR$(4);"OPEN  БАЛЛАДА"
80   PRINT CHR$(4);"WRITE БАЛЛАДА"
90   FOR I=1 TO N
100  PRINT W$(I)
110  NEXT I
120  PRINT CHR$(4);"CLOSE  БАЛЛАДА"
130  END
140  DATA
150  DATA
```

Приложение 3 (фрагмент поэмы)

DER RING DES POLYKRATES

Er stand auf seines Daches Zinnen,
Er schaute mit vergnügten Sinnen
Auf das beherrschte Samos hin.
"Dies alles ist mir untertänig",
Begann er zu Ägyptens König,
"Gestehe, daß ich glücklich bin."

"Du hast der Götter Gunst erfahren!
Die vormals deinesgleichen waren,
Sie zwingt jetzt deines Zepters Macht.
Doch einer lebt noch, sie zu rächen,
Dich kann mein Mund nicht glücklich sprechen,
Solang des Feindes Auge wacht."

Und eh der König noch geendet,
Da stellt sich, von Milet gesendet,
Ein Bote dem Tyrannen dar:
"Laß, Herr! des Opfers Düfte steigen,
Und mit des Lorbeers muntern Zweigen
Bekränze dir dein festlich Haar.

Getroffen sank dein Feind vom Speere,
Mich senden mit der frohen Mähre
Dein treuer Feldherr Polydor."
Und nimmt aus einem schwarzen Becken,
Noch blutig, zu der beiden Schrecken,
Ein wohlbekanntes Haupt hervor.

.....
.....
.....

Hier wendet sich der Gast mit Grausen:
"So kann ich hier nicht ferner hausen,
Mein Freund kannst du nicht weiter sein.
Die Götter wollen dein Verderben,
Fort eil ich, nicht mit dir zu sterben."
Und sprach's und schiffte schnell sich ein.

(Объем текста 558 словоупотреблений,
объем словаря 333 словоформы).

Приложение 4

Пример выпечатки результатов (частотный список словоформ)

СЛОВОФОРМА	АБСОЛ.	ОТНОС. ЧАСТОТА (%)
1. DER	19	3.405
2. UND	17	3.047
3. ICH	14	2.509
4. MIT	13	2.33
5. DEN	12	2.151
6. DEIN	8	1.434
7. DAS	7	1.254
8. DES	7	1.254
9. DIR	7	1.254
10. ER	7	1.254
11. GLUECK	7	1.254
12. NOCH	7	1.254
13. VON	7	1.254
14. ZU	7	1.254
15. IST	6	1.075
16. SIE	6	1.075
17. AUF	5	.896
18. DICH	5	.896
19. DOCH	5	.896
20. GOETTER	5	.896
21. MIR	5	.896
22. NICHT	5	.896
23. DU	5	.896
24. EIN	5	.896
25. ES	5	.896
.....		
63. WAS	2	.358
64. WIE	2	.358
65. WORT	2	.358
66. ZUM	2	.358
67. ACHT	1	.179
68. AEGYPTENS	1	.179
69. ALLEM	1	.179
.....		
331. ZWEIFELND	1	.179
332. ZWEIGEN	1	.179
333. ZWINGT	1	.179

Л И Т Е Р А Т У Р А

Баркалов Н.Б., Тихомиров А.А. Решение экономико-демографических задач на языке БЕЙСИК. – М.: Изд-во Московского университета, 1985. – 121 с.

Велихов Е., Ершов А., Лавров С., Громов Г. Персональный компьютер: перспективы близкие и далекие. – Наука и жизнь, 1985, № 10, с. 19-25.

Дьячко А.Г. Микрокомпьютеры в учебном процессе. – Вестник высшей школы, 1985, № 12, с. 70-72.

Зубов А.В. Обработка на ЕС ЭВМ текстов естественных языков. – Минск: Вышэйшая школа, 1977. – 173 с.

Кнут Д. Искусство программирования для ЭВМ, т. 3. Сортировка и поиск. – М.: Мир, 1978. – 844 с.

Ломов Б.Ф. ЭВМ и развитие человека. – Вестник высшей школы, 1985, № 12, с. 29-33.

Мейер Б., Бодуэн К. Методы программирования, т. 2. – М.: Мир, 1982. – 368 с.

Пиотровский Р.Г. Текст, машина, человек. – Л.: Наука, 1975. – 327 с.

Якубайтис Т.А. Использование ЭВМ в лингвистических исследованиях. – Вопросы языкознания, 1979, № 3, с. 127-131.

MÖGLICHKEITEN DER AUTOMATISCHEN TEXTVERARBEITUNG MIT HILFE DES PERSONALCOMPUTERS "AGAT"

Karl Lepa

R e s Ü m e e

Im Artikel werden die Möglichkeiten der automatischen Textverarbeitung mit Hilfe des in der Sowjetunion hergestellten Personalcomputers "AGAT" untersucht. Die begrenzte Leistungsfähigkeit des Computers läßt eine Analyse größerer Textmassive nicht zu; mit dem vorgestellten sehr einfachen auf der Shellschen Methode beruhenden Programm ist es aber durchaus möglich, Texte mit einer Länge bis 1500 Wörter zu analysieren (z.B. eine Häufigkeitsliste der im Text vorkommenden Wörter herzustellen).

О СПЕЦИФИКЕ РАСПРЕДЕЛЕНИЯ МНОГОЗНАЧНОСТИ ЛЕКСИЧЕСКИХ ЕДИНИЦ В КИТАЙСКОМ ЯЗЫКЕ

Н.В. Обухова

В современной лингвистике большое внимание уделяется выявлению системно-количественных характеристик многозначности лексики в языках различных типов. Практика ряда ранее выполненных в этой области исследований (Zipf, 1949; Папп, 1967; Вишнякова, 1976; Поликарпов, 1976; 1979; 1981; 1986; Крылов, Якубовская, 1977; Тулдава, 1979; Борода, Поликарпов, 1984) свидетельствует о наличии в организации лексической полисемии языков ряда закономерностей количественного характера, которые находятся в системных соотношениях с другими существенными параметрами.

Одной из наиболее фундаментальных характеристик такого плана является общий характер распределения лексических единиц в словаре по количеству присущих им значений.

Результаты процесса закрепления некоторого числа значений за определенными словами коммуникативной практикой языкового коллектива довольно удовлетворительно фиксируются в толковых словарях какого-либо языка. Таким образом, количественные характеристики словаря какого-либо языка в целом, отдельных его разделов и отдельных единиц - весьма объективный источник информации для выяснения количественных характеристик лексического состава языка.

Существующие к настоящему времени исследования по анализу полисемических распределений лексики проводились на материале узкого круга языков (английского, русского, венгерского, эстонского). Материал китайского языка, языка довольно интересного в типологическом отношении, для подобных исследований привлекается впервые.

Основным источником квантитативного анализа лексики китайского языка послужил толковый "Словарь современного китайского языка" - Сяньдай ханьюй цидянь (Пекин, 1979). Количественные характеристики лексики китайского языка, вошедшей в данный словарь, были сопоставлены с аналогичными характеристиками двух других словарей: приблизительно равного ему по объему "Китайско-русского словаря" (под редакцией И.М. Ошанина. Москва, 1952) и значительно меньшего объема "Словаря односложных слов пекинского диалекта" - Бейцзиньхуа даньиньцы

цихуэй (Пекин, 1956).

На данном материале был произведен анализ различий в характеристиках многозначности односложных лексических единиц на фонетическом и иероглифическом уровнях, а также были рассмотрены вопросы зависимости степени многозначности единиц словаря от их односложности или многосложности.

Кроме того, полученные на материале китайского языка результаты по некоторым характеристикам были сопоставлены с аналогичными данными типологически сходного вьетнамского языка и типологически несходного русского языка.[+]

Общая картина основных количественных характеристик лексики китайского языка представлена в следующей таблице:

Таблица I

Словари	Общий объем значений, покрываемых односложами	Кол-во иероглиф. односложов	Среднее кол-во знач. у иероглиф. односложов	Кол-во тонированных слогов	Среднее кол-во значений у тонированных слогов
I. Словарь современного китайского языка	I4886	8320	I,79	I273	I2,6
II. Китайско-русский словарь	I8762	8771	2,I2	I294	I4,3
III. Словарь односложных слов пекинского диалекта	5970	3444	I,73	I293	4,5

Что касается полученного в данной работе распределения групп односложных единиц иероглифического и фонетического уровней в зависимости от количества словарных значений (см. таблицы № 2 и № 3), то можно отметить, что максимально возможное количество значений у иероглифических односложов по

[+] Данные по русскому и вьетнамскому языкам были получены в дипломных работах, выполненных в I98I-I984 гг. под руководством А.А. Поликарпова на кафедре общего, сравнительно-исторического и прикладного языкознания филологического ф-та МГУ.

Таблица 2

Распределение многозначности фонетических односложов китайского языка, полученное на материале толкового словаря (Пекин, 1979) (далее ТС) и "Китайско-русского словаря" (Москва, 1952) (далее КРС).

Кол-во значений	кол-во фонет. односложов ТС	КРС	кол-во значений	кол-во фонет. односложов ТС	КРС	кол-во значений	кол-во фонет. односложов ТС	КРС
174		1	55		1	27	9	13
147		1	54		2	26	5	10
120	1		53	1	1	25	14	
103	1		52		3	24	12	13
100		1	51	1	4	23	22	18
97		1	50	2	3	22	18	19
92		1	49		3	21	18	25
90	1		48	1	2	20	15	21
89		2	47	3	5	19	25	25
85	1		46	2	3	18	30	33
84		1	45		5	17	27	28
80		1	44	3		16	35	32
78		1	43		3	15	33	32
75		1	42	1	4	14	40	30
74	1		41	2	8	13	35	32
72	1	1	40	5	2	12	34	42
71	2		39	6	11	11	44	40
68	1	1	38	3	2	10	50	35
67		2	37	4	4	9	54	50
66	1	1	36	5	8	8	54	65
65		1	35	3	3	7	69	59
63	1		34	3	14	6	71	60
62		1	33	7	18	5	60	64
59	1		32	3	8	4	87	78
58		2	31	11	13	3	90	80
57		2	30	8	10	2	95	91
56		1	29	8	6	1	91	105
			28	15	19	0*)	23	66

*) нулевое кол-во значений приписывается односложам, не имеющим самостоятельных (знаменательных) значений

Таблица 3

Распределение многозначности однослогов словарей китайского языка на иероглифическом уровне.

Кол-во значений	Толковый словарь современного кит. языка (Пекин, 1979)	Китайско-русский словарь (М., 1952)	Словарь односложных слов пекинского диалекта (Пекин, 1956)
26	1	-	-
25	1	-	-
23	1	-	-
20	1	-	-
19	2	-	-
17	1	2	1
16	4	-	-
15	10	2	-
14	7	1	-
13	5	3	-
12	9	6	-
11	17	12	1
10	34	14	2
9	38	26	8
8	57	63	8
7	86	95	14
6	165	173	34
5	206	299	62
4	394	615	153
3	649	1111	333
2	1492	2223	735
1	3808	4126	2093
0	1228	771	-

данным толкового словаря составляет 26 (в словаре Ошанина - 17), у фонетических слогов - 120 (в словаре Ошанина - 174).

Из представленных данных видно, что как среднее, так и максимальное количество значений, приходящихся на фонетический слог, значительно превосходит среднее и абсолютно максимальное количество значений иероглифических однослогов. Таким образом, из-за довольно своеобразной формы иероглифического письма в китайском языке, как ни в каком другом языке с фонетическим письмом, существует огромная разница между письменным и устным языком в плане семантической "нагруженности" письменных (иероглифических) и устно-речевых (фонетических) единиц.

На рис. № I приведены полученные на материале словарей распределения многозначности односложной лексики китайского языка на исследуемых уровнях. Что касается получившихся распределений, то представленный материал подтверждает выявленную ранее на материале английского (Винокова, 1976), эстонского (Тулдава, 1979), венгерского (Папп, 1967) языков тенденцию к систематическому уменьшению объема групп слов с данным количеством значений при движении от однозначных ко все более многозначным словам.+

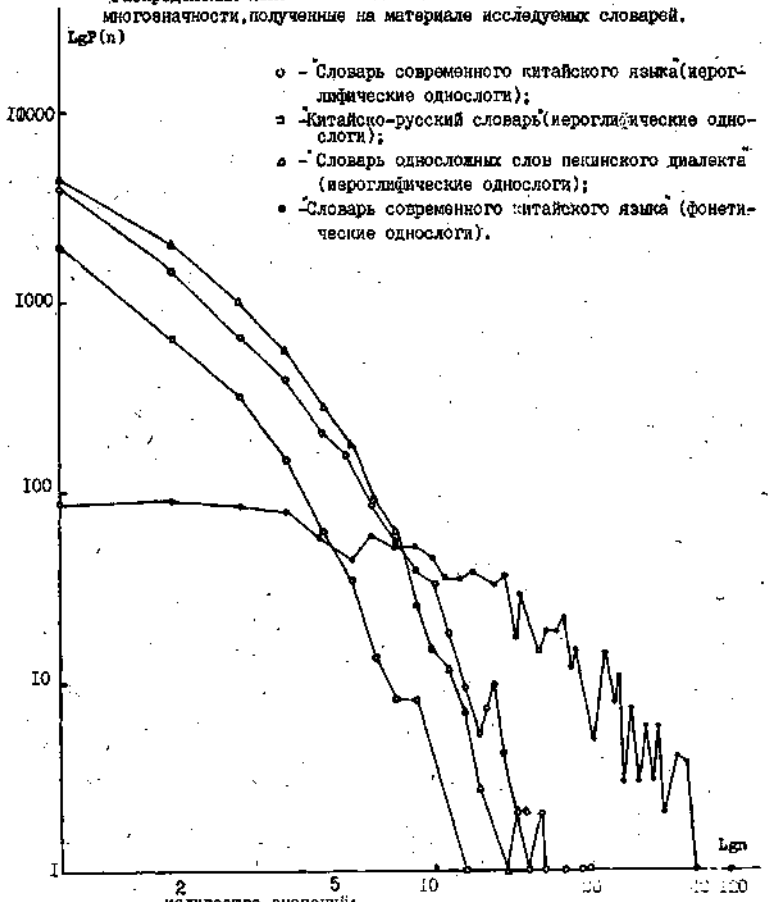
Как видно из рис. № I, распределение многозначности фонетических слогов в билогарифмическом масштабе координат по своей геометрии существенно отличается от распределения иероглифических однослогов. Последнее в большей степени может аппроксимироваться в данной системе координат прямой линией. Такое соотношение данных уровней на материале китайского языка является уникальным, поскольку подобное распределение фонетических однослогов типологически сходного вьетнамского языка (см. рис. № 2) соотносимо не с фонетическими, а скорее с иероглифическими однослогами китайского языка. Графическое соотношение распределений многозначности односложных единиц китайского языка на разных уровнях является, по сути дела, наглядным результатом явления дифференциации большей части лексических омонимов языка в иероглифической записи (80%).

На материале китайского языка при рассмотрении только односложных иероглифических единиц словаря явно прослеживается криволинейность графического распределения многознач-

+ Обсуждение вопроса о виде закона, управляющего полисемическими распределениями, содержится в работах: Папп, 1967; Крылов, Якубовская, 1977; Тулдава, 1979; Поликарпов, 1986.

Рис.№ I

Распределения лексических единиц китайского языка по степени многозначности, полученные на материале исследуемых словарей.

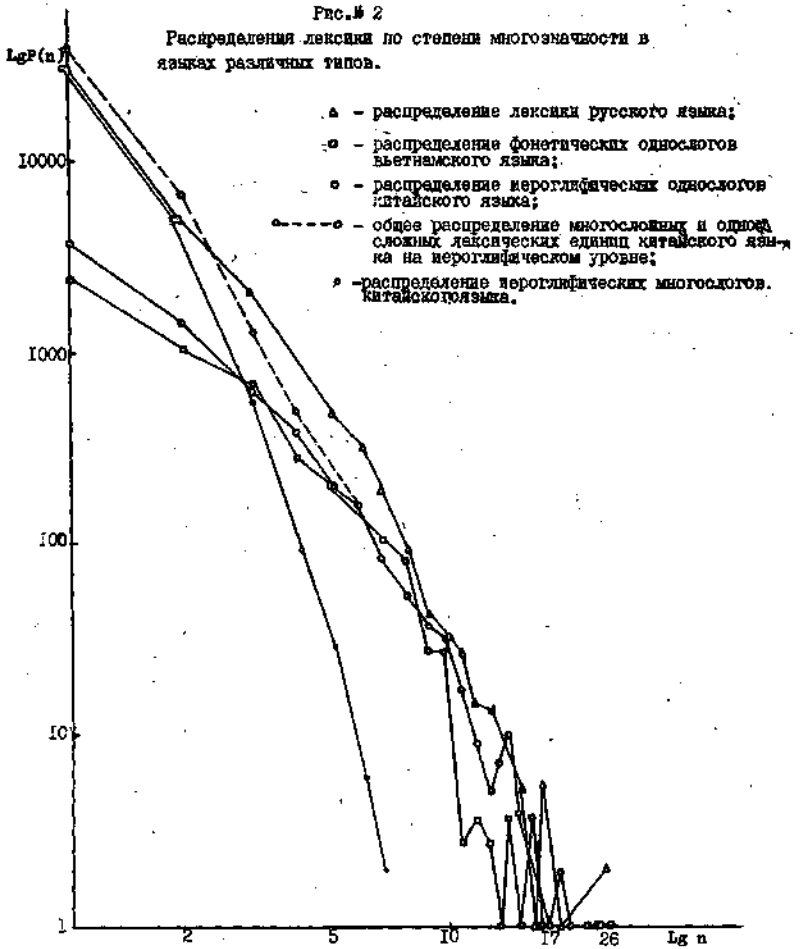


n – количество значений;

P(n) – количество лексических единиц с данным количеством (n) значений.

Рис.№ 2

Распределения лексики по степени многозначности в языках различных типов.



ности на отрезке групп иероглифов, обладающих малым количеством значений (точнее от пяти до одного). Однако графическое представление подобного распределения для иероглифических многослогов указывает на иную, чем у иероглифических однослогов, зависимость между количеством значений и количеством лексических единиц с данным количеством значений. Оно в основной своей части почти прямолинейно и вносит определяющий вклад в общую картину соотношения лексических единиц (одно- и многосложных) между собой по степени многозначности (см. рис. № 2). Наблюдаемое при этом относительное выпрямление общего распределения происходит, как видно, за счет многосложных лексических единиц, структура которых в большинстве случаев совпадает со структурой свободных словосочетаний китайского языка. Так как китайские многослоги - это по большей части не сложные слова, а устойчивые номинативные комплексы (лексемы), принадлежащие не к лексическому, а лексико-синтаксическому уровню языковой системы китайского языка, то и полученный вид полисемического распределения, видимо, зависит от указанной специфики материала.

Интересным представляется также графическое соотношение распределений многозначности иероглифических однослогов, полученных на материале разных словарей китайского языка. Как видно на рис. № 2, диаграммы распределения многозначности однослогов, полученные по толковому словарю (Пекин, 1979) и словарю пекинского диалекта (Пекин, 1956), почти параллельны, тогда как распределение, полученное по переводному словарю Ошанина (Москва, 1952), имеет более пологий наклон, т.е. относительное число многозначных иероглифических однослогов в нем меньше, чем в приблизительно равном ему по объему толковом словаре.

Как нам кажется, наблюдаемое расхождение можно объяснить различным подходом составителей данных словарей к отображению семантических структур вошедших в них лексических единиц (особенно в отношении высокополисемичных лексем), что во многом зависит от специфики самих словарей (толковый или переводной).

Угол наклона общего рисунка распределения лексических единиц какого-либо языка по степени их многозначности относительно вертикальной оси (ось P(n)) характеризует типологический статус языка. Причем, чем более аналитическим является исследуемый язык, тем больше этот угол (Поликарпов, 1986). Поэтому, как видно из сопоставления распределений лексичес-

ких единиц в китайском (на двух уровнях), вьетнамском и русском языках (на материале толкового словаря русского языка под ред. Ожегова), китайский язык в целом представляется самым аналитичным в группе анализируемых языков, что хорошо соответствует традиционным типологическим представлениям об этих языках.

В заключение подчеркнем явление существенной разницы полисемических характеристик единиц разных уровней (фонетического и иероглифического) в пределах китайского языка, что отображает, видимо, существенные (в т.ч. и типологические) различия между двумя его формами.

Л И Т Е Р А Т У Р А

Вишнякова С.М. Опыт статистического исследования многозначности слов английского языка. - В кн.: Вычислительная лингвистика. М., 1976.

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

Крылов Ю.К., Якубовская М.Д. Статистический анализ полисемии как языковой универсалии и проблема семантического тождества слова. - НТИ-2, 1977, № 4.

Папп Ф. О некоторых количественных характеристиках словарного состава языка. - "Slavica" УII, 1967.

Поликарпов А.А. Факты и закономерности аналитизации языкового строя. Автореферат канд. дис., М., 1976.

Поликарпов А.А. Элементы теоретической социолингвистики. М., 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., 1981.

Поликарпов А.А. Система в словаре и системное соотношение основных типов толковых словарей.-Тарту, 1986. (В печати).

Тулдава Ю.А. О некоторых квантитативно-системных характеристиках полисемии. - Учен. зап. Тартуского ун-та, вып.502. Linguistica XI. Тарту, 1979.

Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge (Mass), 1949.

ON THE CHARACTER OF POLYSEMY DISTRIBUTION OF LEXICAL UNITS IN CHINESE

Natalya Obukhova

S u m m a r y

The article deals with the problem of quantitative lexical polysemy in Chinese. The fundamental difference is revealed between the polysemy distribution of Chinese polysyllabic and monosyllabic character words, phonetic monosyllables and character monosyllables. These data are compared with that of Vietnamese and Russian. Some typological conclusions are made on the basis of this comparison.

НЕКОТОРЫЕ СИСТЕМНО-КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ЛЕКСИКО-СЕМАНТИЧЕСКИХ ПАРАДИГМ РАЗНЫХ ВИДОВ

Ю.Б. Сафронова

Системный подход к исследованию языка предполагает анализ взаимосвязи и взаимодействия различных его аспектов, явлений и фактов. Это касается, в частности, изучения лексико-семантических парадигм разных видов, т.е. лексико-семантических категорий, внутри которых единицы данного уровня языка связаны разного рода семантическими отношениями (Д.А. Новиков, 1981).

Лексико-семантические парадигмы разных видов образуют значения многозначного слова (его лексико-семантические варианты - ЛСВ), элементы синонимических и антонимических групп и семантических полей. Предпосылкой исследования их взаимосвязи в языке является то, что с общесемиотической точки зрения полисемия, синонимия и антонимия представляют собой частные проявления общего принципа, лежащего в основе системы языка - асимметричного дуализма языкового знака (С. Карцевский, 1965, Г.П. Мельников, 1971). Однако, несмотря на существование большого количества работ, посвященных вопросам изучения полисемии, синонимии и антонимии, исследование этих языковых явлений велось, в основном, изолированно. Взаимосвязь различных лексико-семантических категорий привлекла внимание исследователей и стала предметом самостоятельного изучения сравнительно недавно, в связи с утверждением системного подхода к изучению лексики (А.А. Уфимцева, 1968, Д.Н. Шмелев, 1973, Э.В. Кузнецова, 1982, Ю.А. Тулдава, 1979, 1980, 1980а, Л.А. Введенская, 1968, В.А. Иванова, 1979, Edmundson, Epstein, 1976). Такое изучение дает возможность по-новому подойти к выявлению некоторых общих закономерностей организации и функционирования языка в целом; оно также может оказаться полезным и с точки зрения лексикографической практики (дифференциации и детализации семантики слов в словарях, выработки критериев выделения границ синонимических групп, определения синонимов и антонимов и т.д.).

Следует отметить, что традиционно исследователей языковой синонимии, антонимии и полисемии привлекал, в основном, качественный аспект изучения этих явлений. Но качественные и количественные характеристики любого объекта взаимообуслов-

лены и составляют диалектическое единство, поэтому изучение какой-либо одной стороны явления не может дать исчерпывающих знаний о нем. Количественные характеристики помогают вскрыть специфику языка и расширяют возможности его сопоставления с системами других языков.

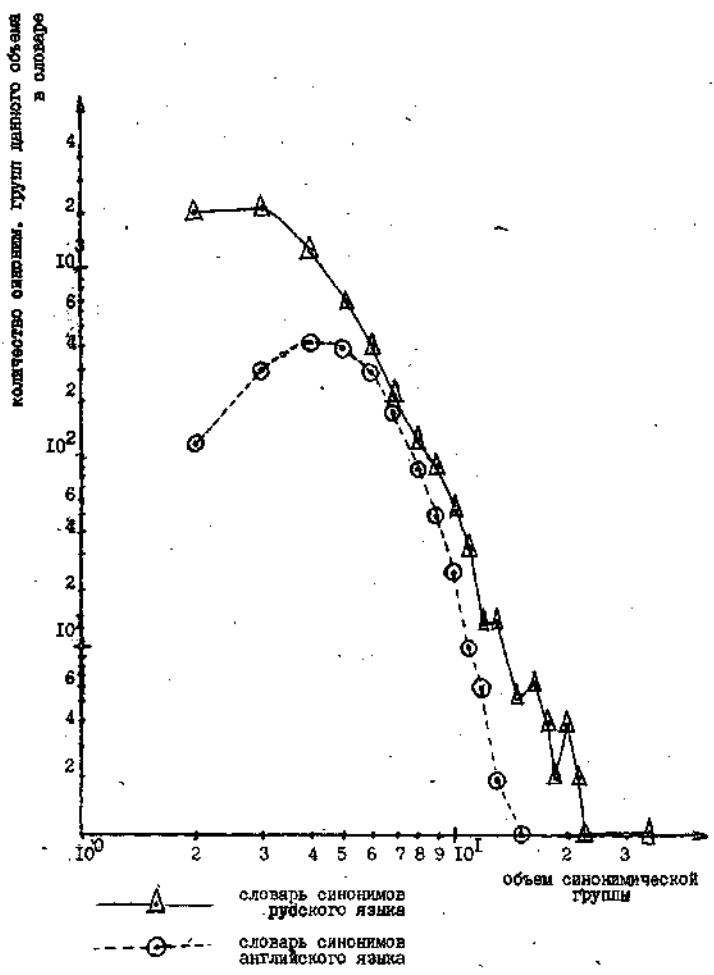
Важность изучения этой стороны языковых явлений определяет выбор метода исследования. Системность языка в целом предполагает системность его лексики. При этом необходимо отметить, что лексика представляет собой систему особого рода - вероятностную. Законы, действующие в таких системах, не являются жестко детерминирующими, а выступают как тенденции, характеризующие массовые явления (они могут не выполняться в отдельном случае, но выполняются в большинстве случаев, пробиваясь сквозь случайные отклонения и исключения). Вероятностный характер лексико-семантической системы языка предполагает возможность ее исследования посредством вероятностных, статистических методов, в частности, с помощью разного рода распределений количественных характеристик ее объектов (Ю.А. Тулдава, 1979, 1982).

Данная работа представляет собой попытку исследования в рамках системного подхода количественного аспекта соотношения и взаимозависимости синонимии, полисемии и антонимии и выявления специфики этого соотношения в английском (относительно более аналитическом) и русском (относительно более синтетическом) языках. Исследование проводилось на материале словарей синонимов английского и русского языков.[+]

Одной из важных системных характеристик синонимического словаря является распределение представленных в нем синонимических групп по их объему (т.е. по числу входящих в них лексико-семантических вариантов слов - ЛСВ). На графике представлены распределения объемов синонимических групп в словаре синонимов русского языка под ред. А.П. Евгеньевой[++] и в

[+] Идея применения вероятностных методов для исследования количественных соотношений в лексических системах разных языков успешно применялась, в частности, в работах А.А.Поликарпова (1976,1979,1981) на материале больших толковых словарей разных языков, в том числе английского и русского.

[++] Данные о распределении объемов синонимических групп в словаре синонимов русского языка под ред. А.П. Евгеньевой были получены Коломаченковой И.В. (1982).


количество синонимических групп данного объема в словаре
4
2
10
6
4
2
10^2
6
4
2
10
6
4
2
10^0
2
3
4
5
6
7
8
9
10^1
2
3
объем синонимической группы
словарь синонимов русского языка
словарь синонимов английского языка

"Новом словаре синонимов Вэбстера" (1971). Как показывает график, в словаре синонимов русского языка представлено больше синонимических групп любого объема, чем в словаре синонимов английского языка. Максимальный объем синонимической группы в словаре Вэбстера – 15 ЛСВ, в словаре синонимов русского языка – 33 ЛСВ (вдвое больше). Общее количество ЛСВ слов, охваченных отношениями синонимии в русском языке, согласно данным словаря, составляет 22427, что почти в 2,5 раза больше, чем в английском языке (в словаре синонимов Вэбстера их число равно 8377). Эти факты говорят о большем развитии лексической синонимии в русском языке по сравнению с английским.

Как показывает график, наиболее представленными в словаре синонимов английского языка являются синонимические группы, включающие 4-5 ЛСВ слов, а в словаре синонимов русского языка – синонимические группы, содержащие 2-3 ЛСВ. Выявленное соотношение является типологически значимым. Оно отражает большую дифференциацию значений слов в русском языке (более синтетическом) по сравнению с английским (более аналитическим), более детальное номинативное членение действительности лексическими средствами синтетического языка.

Примечательна сама форма графика. Сначала он показывает монотонно возрастающую зависимость, достигает "пика" (в словаре синонимов английского языка он приходится на синонимические группы, содержащие 4-5 ЛСВ, в словаре синонимов русского языка – 2-3 ЛСВ), затем наблюдается спад (а в словаре Вэбстера – даже "понятное" движение). Изменение характера функции, связанное со слабой представленностью синонимических групп малого объема в обоих словарях (в особенности в словаре синонимов английского языка) позволяет предположить, что эти словари являются неполными в отношении представленности в них определенной части лексики.+

Для изучения соотношения разных видов семантических отношений между единицами лексики (полисемии, синонимии и антонимии) мы будем пользоваться выборкой, включающей 120 си-

---

+ Возможно, что в словарях синонимов слабо представлены синонимические группы малого объема, объединяющие терминологическую лексику. Эта часть лексики, периферийная по статусу, часто оказывается слабо представленной в словарях, ориентированных на наиболее полное отражение лексических фактов языка (А.А. Поликарпов, 1986).

нонимических групп, представленных в словаре синонимов Вебстера. Выборка содержит 40 синонимических групп малого объема (2 ЛСВ), 40 синонимических групп среднего объема (4-5 ЛСВ) и 40 синонимических групп большого объема (9 - 15 ЛСВ). Общее количество ЛСВ, входящих в синонимические группы выборки равно 470.

Для анализа соотношения синонимии и полисемии было получено распределение средних количеств значений (в языке) у слов, входящих своими отдельными ЛСВ в синонимические группы выборки (данные о "языковой полисемии" этих слов), и о среднем количестве значений (ЛСВ), которыми эти слова вступают в отношения синонимии в целом ("синонимическая полисемия" этих слов):

объем синонимических групп	доминанты синонимических групп		остальные члены синонимических групп	
	языковая полисемия	синонимическая полисемия	языковая полисемия	синонимическая полисемия
2 ЛСВ	4	1,275	3,175	1,222
4-5 ЛСВ	5,675	1,675	4,079	1,424
9-15 ЛСВ	4,2	1,2	3,19	1,318

Как видно из приведенной выше таблицы, языковая полисемия слова в целом превосходит его "синонимическую" полисемию (слово вступает в отношения синонимии не всеми, а только некоторыми из своих значений). Доминанты синонимических групп характеризуются в среднем более высоким значением языковой и "синонимической" полисемии (они более многозначны, чем остальные члены синонимических групп и вступают в отношения синонимии бо́льшим числом своих ЛСВ). Этот факт соответствует лингвистическому критерию определения в качестве доминанты синонимической группы наиболее нейтрального в стилистическом отношении и наименее наделенного дополнительными смысловыми оттенками слова (по сравнению с его синонимами).

Полученные данные показывают, что слова, входящие своими ЛСВ в синонимические группы среднего объема (4-5 членов), обладают в среднем бо́льшим числом значений в языке по сравнению с другими словами; количество значений (ЛСВ), которыми они вступают в отношения синонимии, в среднем у них также будет выше, чем у других слов в выборке.

Обратимся еще к одному распределению, содержащему данные о среднем количестве антонимов у слов, входящих своими

ЛСВ в синонимические группы разного объема:

объем синонимической группы	среднее количество антонимов у слов, входящих своими ЛСВ в синонимические группы данного объема
2 ЛСВ	0,85
4- 5 ЛСВ	1,95
9-15 ЛСВ	1,5

Данные этого распределения показывают ту же тенденцию. Среднее число антонимов у членов синонимических групп среднего объема (4–5 ЛСВ) будет больше, чем у других слов в выборке.

Выявленные факты говорят о существовании зависимости между способностью слов вступать в отношения синонимии и другими семасиологическими характеристиками слова, в частности, способностью вступать в антонимические противопоставления и иметь то или иное количество значений.

Выявив существование определенных корреляций между наиболее сильными семантическими связями (синонимическими, антонимическими, связями значений многозначного слова), организующими лексико-семантические парадигмы низших уровней, было бы интересно исследовать соотношения, существующие между синонимическими группами и семантическими полями тезауруса. Семантическое поле представляет собой лексико-семантическую парадигму (категорию) более высокого уровня. Характер связи между элементами синонимической группы и семантического поля одинаков (смысловое сходство), но различна степень интенсивности этой связи: в семантическом поле эта связь слабее, чем в синонимической группе.

Выявление структуры семантического поля в тезаурусе и соотнесение составляющих его структурных единиц с лексико-семантическими парадигмами более низкого уровня показало, что минимальная структурная единица семантического поля в тезаурусе (группа слов, выделенная внутри семантического поля знаком ";") приблизительно соответствует синонимической группе, объединяющей соответствующие единицы в словаре синонимов (исследования проводились на материале словаря синонимов Вэбстера и Тезауруса Роже).

Однако между объемами синонимических групп и соответствующих им структурных единиц семантических полей могут наблюдаться расхождения. Степень такого расхождения ("приращe-

ние синонимической группы в семантическом поле*) определяется как среднее значение разности между числом элементов, входящих в минимальную структурную единицу семантического поля, но не входящих в соответствующую ей синонимическую группу, и числом членов этой синонимической группы, не вошедших в минимальную структурную единицу внутри семантического поля. Значение этой характеристики изменяется в зависимостиот объема синонимической группы:

объем синонимической группы	приращение синонимической группы в семантическом поле
2 ЛСВ	4,9
4- 5 ЛСВ	8,5
9-15 ЛСВ	0,45

Таблица показывает, что в среднем минимальное объединение элементов внутри семантического поля по своему объему превосходит соответствующую синонимическую группу ("приращение" имеет положительное значение). При этом наиболее интенсивно до объема минимальной структурной единицы семантического поля будут расширяться синонимические группы среднего объема (4-5 ЛСВ), элементы которых оказываются в наибольшей степени охваченными разными видами семантических отношений в языке. Наиболее интенсивное расширение объема этих синонимических групп в семантическом поле тезауруса говорит о том, что члены этих синонимических групп в смысловом и понятийном отношении связаны с наиболее широким кругом слов по сравнению с другими лексическими единицами, входящими в выборку, что подкрепляет предположение о наибольшей активности этой группы слов в отношении вступления в семантические отношения разных видов.

Таковы, вкратце, результаты и выводы данной работы. Необходимо отметить, что они во многом носят предварительный характер. Но, несмотря на это, можно надеяться, что постановка этих проблем и поиск путей их решения будут способствовать дальнейшему развитию теории и практики лексикологии, лексикографии и типологии языков.

Л И Т Е Р А Т У Р А

Введенская Л.А. О взаимодействии синонимии и антонимии. – В кн.: Вопросы лексики и фразеологии современного русского языка. Ростов-на-Дону, 1968.

Евгеньева А.П. Словарь синонимов. М., 1975.

Иванова В.А. Антонимия в русском языке. Кишинев, 1982.

Иванова В.А. Синонимо-антонимические блоки. – В кн.: Вопросы грамматики и лексикологии русского языка. Кишинев, 1979.

Карцевский С. Об асимметричном дуализме языкового знака. – В кн.: Звегинцев В.А. История языкознания XIX–XX вв. в очерках и извлечениях. Ч. 2. М., 1965.

Коломаченкова И.В. Некоторые закономерности организации лексической синонимии (по данным словаря синонимов русского языка под ред. А.П. Евгеньевой). Дипломная работа, МГУ, филол. ф-т, кафедра ОСиПЯ, 1982.

Кузнецова Э.В. Лексикология русского языка. М., 1982.

Мельников Г.П. О типах дуализма языкового знака. – ФН, 1971, № 5.

Новиков Л.А. Семантика русского языка. М., 1981.

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., 1981.

Поликарпов А.А. Система в словаре и системное соотношение основных типов толковых словарей. – В наст. сб.

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

Поликарпов А.А. Элементы теоретической социолингвистики. Некоторые предпосылки, результаты и перспективы причинного подхода в общей семиотике и языкознании. М., 1979.

Тулдава Ю.А. Квантитативное исследование генетического состава эстонского словаря. – УЗ Тартуского ГУ, вып. 628, Тарту, 1982.

Тулдава Ю.А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста. – УЗ Тартуского ГУ, вып. 549, Тарту, 1980 (в тексте: 1980а).

Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (I). – УЗ Тартуского ГУ, вып. 544, Тарту, 1980.

Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (3). - УЗ Тартуского ГУ, вып. 619, Тарту, 1982.

Уфимцева А.А. Слово в лексико-семантической системе языка. М., 1968.

Шмелев Д.Н. Проблемы семантического анализа лексики. М., 1973.

Edmundson H.P., Epstein M.N. Research on Synonymy and Antonymy. A model and its representation. - In: Papers in Computitional Linguistics. Budapest, 1976.

Roget's International Thesaurus. New-York, 1977.

Webster's New Dictionary of Synonyms. Springfild (Mass), USA, 1973.

QUELQUES CHARACTERISTIQUES QUANTITATIVES DES CATHEGORIES LEXICO-SEMANTIQUES DES SORTES DIFFERENTS

Julie Safronova

R é s u m é

Comme le lexique est un systhème, on peut l'étudier par les méthodes du calcul des probabilités. Dans cette article il s'agit du problème d'emploi des méthodes quantitatives aux études de synonymie, polysémie et antonymie. Les resultats obtenus montrent la correlation des rapports sémantiques des sortes differents (synonymiques, antonymiques et polysémiques), qui ont lieu entre les éléments du systhème lexique.

О ЧАСТОТНОМ СПЕКТРЕ ЛЕКСИКИ ТЕКСТА

Ю.А. Тулдава

Статья является второй в серии статей автора, посвященных актуальным проблемам частотной структуры текста. В предыдущей статье (Тулдава Ю.А., 1985) были рассмотрены ранговые распределения лексики и различные варианты закона Ципфа. В настоящей статье подвергается анализу частотный спектр лексики и рассматриваются возможности его аналитического описания.

Частотный спектр. Упорядоченный ряд численностей слов с данной частотой образует "спектральное" распределение частот, или частотный спектр (именуемый в квантитативной лингвистике также частотно-лексическим или просто лексическим спектром). Частотный спектр отражает тот же основной принцип концентрации и рассеяния лексических единиц, который был обнаружен при рассмотрении рангового распределения лексики. Концентрация единиц появляется здесь в области редких частот: в словаре и тексте слова с частотой $F = 1$ образуют наиболее многочисленную группу, затем следуют группы слов с частотами $F = 2$, $F = 3$ и т.д., вплоть до рассеянных групп высокочастотных слов. На графике такое распределение напоминает гиперболу (рис. 1), причем обнаруживается крутой спад при переходе от $F = 1$ до $F = 2$, т.е. одноразовые слова заметно превосходят по численности двухразовые слова. Например, по данным ЧС лексем авторской речи художественной прозы эстонского языка (по тексту объемом $N = 100\,000$) одноразовых слов в словаре 59,3 %, а двухразовых слов – 14,0 %; в ЧС словоформ соответствующие проценты – 70,8 и 12,6 (Tuldava J., 1977, с. 167-168).

При обработке данных ЧС авторской речи эстонской художественной прозы были также получены данные о частотном спектре отдельных подвыборок – текстов объемом $N = 5000$. При сравнении частотных спектров больших и малых выборок отчетливо выявляется разница в распределении частот (табл. 1). При увеличении текста уменьшается доля одноразовых слов в словаре, и соответственно увеличивается доля более частых слов ($F > 1$). В тексте же уменьшается доля всех малочастотных слов, но зато сильно возрастает удельный вес частых слов, в частности покрытие текста словоформами, имеющими частоту

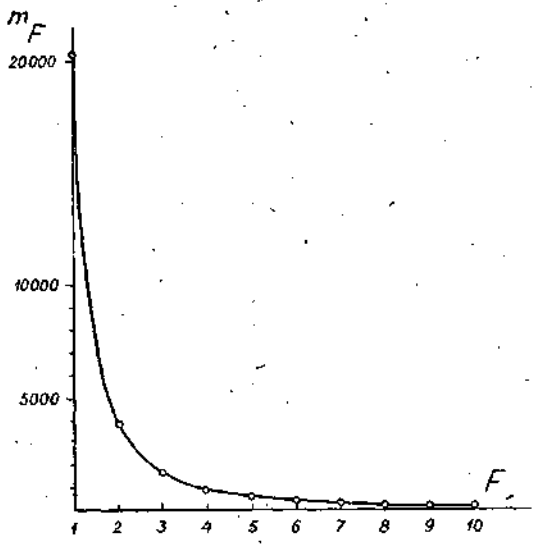


Рис. I. Связь между частотой (F) и числом слов с данной частотой (m_F) по данным ЧС словоформ эстонского языка.

$F > 10$, составляет 51,5 % в большом тексте (N = 100 000) и только 23,7 % в малом тексте (N = 5000). Значительные изменения в спектральном распределении частот слов можно наблюдать и при представлении данных в кумулятивной форме, например, по данным ЧС английского языка Кучеры и Фрэнсиса (Kučera H., Francis W.N., 1967) (см. табл. 2). И здесь наблюдается сильное уменьшение доли одноразовых слов в тексте по мере увеличения текста (от N = 2000 до N = 1 000 000): 28,3 – 19,2 – 7,0 – 4,5 – 2,2 (%), в то время как доля частых слов постоянно увеличивается. Аналогичную картину можно наблюдать в текстах русского языка (см., например, Григорьева А.С., 1981).

Некоторые ученые полагают, что при увеличении текста наступает момент, когда частотный спектр, в частности доля одноразовых слов, стабилизируется (например, Williams C.B., 1970, с. 103). Однако приведенные опытные данные говорят о

том, что при увеличении текста удельный вес редких слов неуклонно падает. Теоретически можно даже предположить, что в очень большом тексте одноразовые, двухразовые слова составляют ничтожную долю. Действие механизма порождения частотного спектра Р.Г. Пиотровский (1975, с. 108) объясняет следующим образом: при последовательном увеличении текста все большее число "нуль-частотных" слов, т.е. слов, реально бытующих в лексике исследуемого языка, но не попавших в выборку, попадает в выборочную совокупность и фиксируется в ЧС. Одновременно увеличивается частота некоторых редких слов, которые передвигаются в средние зоны ЧС. На характер изменения состава частотного спектра может повлиять также переход слов из средней зоны в зону высокочастотных слов и наоборот (хотя такие передвижения оказываются менее обычными при увеличении объема однородного текста). В общем наблюдается закономерность "неравномерного перехода" (Muller Ch., 1976, с. 144): всякое подмножество m_F, т.е. число слов с частотой F ($F = 1, 2, \ldots, n$), имеет тенденцию при увеличении текста "выигрывать" больше, чем "терять". Это означает, например, что переходы из группы m_F в m_{F+1} более многочисленны, чем переходы из группы m_{F+1} в m_{F+2}.

При сравнении частотных спектров разных текстов можно пользоваться оценкой энтропии (Shannon C.E., 1948), представляющей собой меру "неопределенности", в данном случае – меру равномерности распределения частот слов в словаре или тексте. С одной стороны, энтропия большого словаря, как правило, опережает энтропию небольшого словаря, благодаря тому, что равномерность в распределении частот с увеличением текста увеличивается (концентрация редких слов падает). С другой стороны, при сравнении текстов одинаковой длины мера энтропии может служить показателем "богатства" лексики, которому соответствует меньшее значение энтропии (в последнем случае распределение частот менее равномерно из-за большей доли одноразовых слов).

Частотный спектр может дать представление о покрываемости текста словами разных частот. Такие данные имеют значение при стилометрическом анализе текстов и при типологическом исследовании языков (Бектаев К.Б., 1978). При сравнении текстов одинакового объема (N = 100 000) эстонского, русского и английского языков (табл. 3) можно заметить, что одноразовые и другие редкие словоформы ($F \leq 5$) покривают в эстонском словаре и тексте значительно бо́льшую долю, чем в

Таблица I

Частотные спектры (в %) по данным ЧС словоформ эстонского языка на основе текстов разного объема: N = 5000 (I) и N = 100 000 (II)

F	Словарь		Текст	
	I	II	I	II
I	78,3	70,8	42,2	21,8
2	10,9	12,6	11,7	7,8
3	4,1	5,3	6,5	4,9
4	1,9	2,6	4,2	3,2
5	1,1	1,7	3,0	2,6
6-10	2,2	3,5	8,7	8,2
>10	1,5	3,5	23,7	51,5
Всего (%)	100,0	100,0	100,0	100,0

Таблица 2

Кумулятивные частотные спектры (в %) по данным ЧС английского языка Кучера и Фрэнсиса

F	Словарь				
	N = 2000	N = 10051	N = 101566	N = 253538	N = 1014232
I	69,9	64,2	51,6	48,0	44,7
I-5	94,1	92,2	83,4	80,1	75,4
I-10	97,3	96,6	91,5	88,3	84,3
> 20	0,5	1,7	3,9	6,0	9,5
	Текст				
I	28,3	19,2	7,0	4,5	2,2
I-5	54,8	41,7	19,4	13,2	6,7
I-10	64,8	51,3	27,6	19,0	10,0
> 20	25,1	41,3	63,4	73,1	85,4

русском и английском языках. Это объясняется большей синтетичностью эстонского языка (наличие большого количества словоизменительных форм). Разница между русским и английским языками в данном примере небольшая, хотя ощущается большая концентрация редких словоформ в русском тексте. Здесь сказывается то обстоятельство, что сравниваются тексты разных подъязыков: русские тексты по технике (которая беднее по лексике) и английские тексты по литературе. Таким образом, кумулятивный частотный спектр (покрываемость) может различать и разные стили языка. Сравнение данных эстонского и английского языков (при близких жанрах) выявляет действительную разницу в структуре языков: например, одноразовые словоформы составляют в эстонском тексте 22 %, а в английском тексте такого же объема только 7 %.

Таблица 3

Покрываемость текста словоформами (в %) в текстах одинакового объема (N = 100 000) в эстонском, русском (Калинина Е.А., 1968) и английском (Kučera H., Francis W.N., 1967, с. 327-329) языках

F	Словарь			Текст		
	Эст.	Рус.	Англ.	Эст.	Рус.	Англ.
I	70,8	45,3	51,6	21,8	6,4	7,0
I-5	93,1	79,8	83,5	40,3	20,3	19,4
I-10	96,6	89,3	91,5	48,5	30,5	27,6
I-100	99,7	99,3	99,4	73,0	67,7	53,2

Одноразовые слова (т. наз. hapax legomena - греч.), выявляемые при составлении частотных словарей, представляют собой особую группу "редких" слов. В эту группу входят разнообразные неологизмы, окказионализмы, диалектизмы, архаизмы, иностранные слова и др. наряду с обычными словами, которые случайно могут оказаться в группе редких слов и при дальнейшем увеличении выборки могут перейти в группу среднечастотных слов. По данным ЧС авторской речи эстонской художественной прозы группа одноразовых слов включает в основном существительные (45 % всех одноразовых лексем в словаре текста объемом N = 100 000 словоупотреблений), глаголы (30 %), прилагательные (12 %) и наречия (10 %). Особо надо отметить, что редкие слова составляют важный резерв обогащения словаря.

Можно, например, сравнить состав одноразовых слов в упомянутом ЧС с составом нормативного ортологического словаря (ОС-76: Õigekeelsussõnaraamat, 1976). Оказывается, что около 20 % одноразовых слов ЧС не представлены в ОС-76. Среди них много интересных новообразований, которые в дальнейшем могут войти во всеобщее употребление.

Связь с законом Ципфа. Что касается аналитического выражения спектрального распределения частот слов, т.е. частотного спектра, то надо иметь в виду, что это распределение "органически" связано с ранговым распределением (они образуют две взаимосвязанные половины общей частотной структуры текста; подробнее см. Тулдава Ю.Х., 1985). Следовательно, сфера действия закона Ципфа должна распространяться и на спектральное распределение. Если исходить из того, что ранговая форма данного распределения частот слов описывается формулой закона Ципфа с поправкой Мандельброта:

$$F_i = C\,(i+B)^{-\gamma}, \qquad (1)$$

где F_i – частота слова с рангом i; C, B и γ – параметры, то спектральным аналогом рангового распределения является (Хайтун С.Д., 1983, с. 161)

$$m_F = c\,F^{-(1+\alpha)}, \qquad (2)$$

где m_F – число слов с частотой F; c и α – параметры, причем $\alpha = 1/\gamma$ (γ из формулы (1)) и $c = \alpha(L-1)/(F_{min}^{-\alpha} - F_{max}^{-\alpha})$, где L – объем словаря. Так как в распределениях частот слов $F_{min}^{-\alpha} = 1$ и $F_{max}^{-\alpha}$ минимально мало, причем $L \gg 1$, то практически можно считать, что $c \approx \alpha L$. Соотношением (2) описывается ряд ситуаций, встречающихся в разных отраслях знаний; оно известно также под именами "закон Парето", "закон Уиллиса-Юла" и др. (см. Пиотровский Р.Г., 1975, с. 98).

При $\gamma = 1$ формула принимает вид

$$m_F = c\,F^{-2}. \qquad (3)$$

Такой представил себе эту зависимость первоначально сам Ципф, хотя он позднее предложил уточненный вариант (Zipf G.K., 1949):

$$m_F = c\,(F^2 - 0{,}25)^{-1} \qquad (4)$$

Были еще другие попытки более точно аппроксимировать спектральное распределение прибавлением добавочных параметров, например, по следующей формуле (Krallmann D., 1966, с. 88):

$$m_F = c F^{-k} b^F, \qquad (5)$$

где c, b и k – параметры.

Основываясь на том или другом варианте закона Ципфа, еще другие авторы вывели свои формулы для аналитического выражения частотного спектра (Орлов Ю.К., 1976; Арапов М.В. и др. 1975; Крылов Ю.К., 1982). Исходя из предположения о связи некоторых особенностей порождения речи с пространственно-временной организацией периодических процессов головного мозга, А.Н. Лебедев (1983) выводит особую формулу для вычисления частотного спектра. Свои формулы предлагали еще Б. Мандельброт (Mandelbrot B., 1961), Б. Брукс (Brookes B.C., 1982) и др.

Все упомянутые формулы с большей или меньшей точностью описывают спектральное распределение частот слов в словаре и тексте.

По теоретическим соображениям можно предположить, что соответствие между параметром γ рангового распределения, т.е. из формулы (1), и параметром α из формулы (2) достигается наилучшим образом в том случае, если исходить из значений параметра γ, установленных для средней или хвостовой части рангового распределения. Эти части рангового распределения соответствуют средней и начальной частям спектрального распределения частот слов. Как известно, параметр γ (выражающий уклон линии рангового распределения в билогарифмическом масштабе) обычно меняет свое значение в соответствии с отклонениями в начальной и хвостовой частях рангового распределения. В общих чертах можно говорить о трех "стадиях" распределения, которым соответствуют параметры $\gamma_1, \gamma_2, \gamma_3$ (для начальной, средней и хвостовой части). Для проверки мы взяли данные частотной структуры законченного художественного произведения (по ЧС лексем I-го тома романа эстонского классика А.Х. Таммсааре "Правда и право" – "Tõde ja õigus"; см. Villup A., 1978; полную частотную структуру см. в табл. 4). Значения параметра γ для рангового распределения были вычислены в следующих диапазонах:

Таблица 4

Частотная структура эстонского текста по данным ЧС лексем I-го тома романа "Правда и право" ("Tõde ja õigus") А.Х. Таммсааре. Объем текста: N = 160 356, объем словаря: L = 8228.

i	F	n	i	F	n
I	7168	I	34	730	I
2	5210	I	35	728	I
3	4268	I	36	705	I
4	3715	I	37	691	I
5	2807	I	38	682	I
6	2492	I	39	633	I
7	2470	I	40	593	I
8	2268	I	41	585	I
9	2251	I	42	581	I
I0	I759	I	43	564	I
II	I730	I	44	554	I
I2	I382	I	45	540	I
I3	I329	I	46	532	I
I4	I284	I	47	526	I
I5	I209	I	48	510	I
I6	II6A	I	49	503	I
I7	II33	I	50	491	I
I8	I090	I	5I	484	I
I9	I046	I	52	477	I
20	I046	I	53	470	I
2I	976	I	54	469	I
22	97I	I	55	460	I
23	96I	I	56	450	I
24	928	I	57	446	I
25	9I0	I	58	443	I
26	897	I	59	442	I
27-28	832	2	60	437	I
29	82I	I	6I	436	I
30	790	I	62	435	I
3I	780	I	63	433	I
32-33	754	2	64	405	I

i	F	n	i	F	n
65	397	I	II0	245	I
66	396	I	III	235	I
67	395	I	II2	226	I
68	394	I	II3-II4	225	2
69	39I	I	II5	223	I
70	384	I	II6	220	I
7I	383	I	II7	2I8	I
72	379	I	II8	2I7	I
73	373	I	II9	2I5	I
74	372	I	I20	2I3	I
75	369	I	I2I	2I0	I
76	367	I	I22-I23	208	2
77	36I	I	I24	207	I
78	357	I	I25	206	I
79	345	I	I26	20I	I
80	34I	I	I27	200	I
8I	328	I	I28	I98	I
82-83	32I	2	I29	I97	I
84	3I8	I	I30	I95	I
85	3I7	I	I3I	I94	I
86	308	I	I32	I93	I
87-88	305	2	I33-I34	I92	2
89-90	300	2	I35	I86	I
91-92	296	2	I36	I85	I
93	284	I	I37-I39	I84	3
94	280	I	I40	I83	I
95-96	274	2	I4I	I82	I
97	269	I	I42	I80	I
98	264	I	I43	I79	I
99-I0I	265	3	I44	I77	I
I02	263	I	I45-I46	I76	2
I03	260	I	I47	I74	I
I04	253	I	I48-I49	I73	2
I05	252	I	I50	I7I	I
I06	25I	I	I5I-I52	I68	2
I07	248	I	I53-I54	I67	2
I08	246	I	I55-I56	I63	2
I09	245	I	I57	I62	I

i	F	n
158-160	161	3
161	159	1
162	158	1
163-166	157	4
167	156	1
168	155	1
169-171	153	3
172-173	150	2
174-175	149	2
176	148	1
177-178	147	2
179-180	145	2
181	142	1
182	141	1
183-185	140	3
186-187	139	2
188-189	138	2
190	137	1
191-193	136	3
194	135	1
195-196	133	2
197	132	1
198-199	131	2
200-201	129	2
202	128	1
203-204	127	2
205-207	126	3
208	124	1
209	123	1
210-211	122	2
212-213	121	2
214	120	1
215	119	1
216	118	1
217	117	1
218	116	1
219-221	114	3

i	F	n
222-223	113	2
224-227	112	4
228-229	111	2
230	110	1
231	109	1
232	107	1
233-234	106	2
235-238	105	4
239	104	1
240-243	103	4
244-246	102	3
247-251	101	5
252	100	1
253-254	99	2
255	97	1
256-258	96	3
259-262	95	4
263	94	1
264-265	93	2
266-267	92	2
268	91	1
269-273	90	5
274	89	1
275	88	1
276-282	87	7
283-284	86	2
285-286	85	2
287-288	84	2
289-293	83	5
294-297	82	4
298-301	81	4
302-303	80	2
304-306	79	3
307-310	78	4
311-312	77	2
313-318	76	6
319-321	75	3

i	F	n
322-325	74	4
326-330	73	5
331-334	72	4
335-340	71	6
341-348	70	8
349-353	69	5
354-356	68	3
357-360	67	4
361	66	1
362-364	65	3
365-368	64	4
369-371	63	3
372-378	62	7
379-385	61	7
386-389	60	4
390-395	59	6
396-397	58	2
398-403	57	6
404-405	56	2
406-411	55	6
412-418	54	7
419-428	53	10
429-435	52	7
436-444	51	9
445-448	50	4
449-458	49	10
459-465	48	7
466-469	47	4
470-475	46	6
476-488	45	13
489-501	44	13
502-514	43	13
515-529	42	15
530-545	41	16
546-552	40	7
553-560	39	8
561-570	38	10
571-585	37	15

i	F	n
586-599	36	14
600-614	35	15
615-631	34	17
632-651	33	20
652-659	32	8
660-679	31	20
680-694	30	15
695-711	29	17
712-719	28	8
720-737	27	18
738-755	26	18
756-774	25	19
775-804	24	30
805-825	23	21
826-850	22	25
851-865	21	15
866-899	20	34
900-929	19	30
930-970	18	41
971-994	17	24
995-1042	16	48
1043-1095	15	53
1096-1151	14	56
1152-1190	13	39
1191-1246	12	56
1247-1311	11	65
1312-1396	10	85
1397-1506	9	110
1507-1630	8	124
1631-1805	7	175
1806-2024	6	219
2025-2321	5	297
2322-2762	4	441
2763-3375	3	613
3376-4591	2	1216
4592-8228	1	3637

$i = 1 \div 30$ - $\gamma_1 = 0{,}7$

$i = 30 \div 1500$ - $\gamma_2 = 1{,}1$

$i > 1500$ - $\gamma_3 = 1{,}4$

Проверка показала, что для установления хорошего соответствия формул (1) и (2) наилучшим образом подходит γ_3. В данном случае при $\gamma_3 = 1{,}4$, $\alpha = 1/\gamma_3 = 0{,}7$ и показатель степени в формуле (2) $1 + \alpha = \beta = 1{,}7$. Соответствие между теоретическими и эмпирическими данными достаточно хорошее для частотного спектра от $F = 1$ до $F = 100$ (см. табл. 5). В этот диапазон попадает практически 97 % всех слов словаря данного текста. Все же наблюдается отклонение в начальной части частотного спектра, если исходить из требования, что параметр $c \approx \alpha L$. Аппроксимация лучше, если найти значение параметра c опытным путем (при сохранении значения показателя степени $\beta = 1 + \alpha$, где $\alpha = 1/\gamma_3$). Еще лучшее соответствие получается добавлением поправочного коэффициента d (аналогично поправке Мандельброта), и формула спектрального распределения принимает окончательный вид:

$$m_F = c\,(F + d)^{-\beta}, \qquad (6)$$

где c, d и β - параметры (см. табл. 5). Аналогичные результаты мы получили на материале ЧС русского языка (см. табл. 6).

Таким образом, можно сделать вывод, что как ранговое, так и спектральное распределения частот слов подчиняются одной и той же закономерности, выражаемой в первом приближении обычной степенной функцией с отрицательным показателем степени (в общей форме: $y = ax^{-b}$) и более точной модифицированной функцией с добавлением поправочного коэффициента (в общей форме: $y = a(x+d)^{-b}$). Как было показано в работе (Тулдава Ю.А., 1985, с. 96), степенная функция раскрывается в дифференциальной записи как закон "постоянного относительного роста (или убывания)":

$$\frac{dy/y}{dx/x} = b$$

(в данном случае $b < 0$), широко известный во многих областях науки как универсальный закон, охватывающий широкий круг явлений материального мира (по форме к нему относится, например, "аллометрический" закон роста в биологии). В содержательном смысле это означает, что относительный прирост или убывание функции (например, численности слов с данной

Таблица 5

Частотный спектр ($F \leq 100$) по данным ЧС лексем романа А.Х. Таммсаare (эст.). N = 160 356, L = 8228. Наблюдаемое и ожидаемое количество слов m_F с частотой F. Вычисления по формулам:

I - $m_F = c F^{-\beta}$ $(c = \alpha L;\ \beta = 1+\alpha;\ \alpha = 1/\gamma_3)$

II - $m_F = c(F+d)^{-\beta}$ (аппроксимация)

III - $m_{F+1} = m_F \frac{(a+F-1)}{(x+F)}$ (Уэринт-Хердан)

F	m_F наблюд.	m_F ожид. I	ожид. II	ожид. III
1	3637	5700	3617	(3637)
2	1216	1754	1295	1436
3	613	881	676	763
4	441	540	420	472
5	297	370	288	320
6	219	271	211	231
7	175	208	162	174
8	124	166	129	136
9	110	136	105	109
10	85	114	87	89
...				
15	53	57	43	41
20	34	35	26	23
30	15	18	13	10
40	7	10	7	7
50	4	7	5	5
80	3	3	2	3
100	2	2	2	2
Параметры:		c = 5700 β = 1,7	c = 5800 β = 1,8 d = 0,3	a = 1,35 x = 2,42

Таблица 6

Частотный спектр ($F \leq 600$) по данным ЧС лексем русского языка (1977). Объем текста N = I 056 382; объем словаря L = 39 268. Наблюдаемое и ожидаемое количество слов m_F с частотой F; вычисления по формулам:

I: $m_F = c F^{-\beta}$; II: $m_F = c(F+d)^{-\beta}$

III: $m_{F+1} = m_F \frac{(a+F-1)}{(x+F)}$ (Уэрминг-Хердан).

F	m_F набл.	m_F ожид. I	II	III
I	I3379	I7000	I3068	(I3379)
2	5746	6OI0	577I	6690
3	3364	3272	3368	3983
4	2243	2I25	2253	2638
5	I68I	I52I	I634	I872
6	I279	II57	I25I	I395
7	977	9I8	995	I078
8	84I	75I	8I5	857
9	7I3	630	682	698
I0	595	538	580	578
I5	286	293	3II	275
20	200	I90	I99	I59
30	I09	I04	I05	70
40	60	67	67	40
50	45	48	47	30
60	30	37	35	
70	3I	29	28	
80	26	24	22	
90	I5	20	I9	
I00	I4*	I7	I6	
I50	7	9	8	
200	4	6	5	
300	3	3	3	
400	2	2	2	
500	2	2	I	
600	I	I	I	
Параметры:		c = I7000 β = I,5	c = 25000 β = I,6 d = 0,5	a = 2,08 x = 3,I6

* Для частот $F \geq$ I00 взяты средние значения m_F, например, для F = I00 вычислено среднее значение m_F в промежутке F = 98 ÷ I02.

частотой) пропорционален относительному приросту аргумента (например, частоте слов). Небольшие отклонения от такой простой и естественной зависимости объясняются разными, в т.ч. лингвистическими причинами.

Модель Уэринга-Хердана. Модель частотного спектра Уэринга-Хердана (Waring-Herdan) представляет собой другую возможность описания спектрального распределения частот слов (Herdan G., 1964; Muller Ch., 1968; см. также Тулдава Ю., 1971, с. 215-218). Здесь исходят из представления о том, что спектральное распределение частот слов образует монотонно убывающий ряд $m_1, m_2, \ldots, m_n$ (т.е. число слов с частотой 1, частотой 2 и т.д.), который определяется двумя параметрами a и x:

$$m_{F+1} = m_F \frac{(a+F-1)}{(x+F)}. \qquad (7)$$

Эта модель соответствует закономерности постоянного увеличения отношения m_{F+1} / m_F.* Для практического применения модели нужно предварительно знать объем текста (N), объем словаря (L) и число одноразовых слов (m_1).

В нашем примере (по ЧС романа А.Х. Таммсаare, см. табл. 4) исходными данными являются N = 160 358, L = 8228, m_1 = 3637. Нужно вычислить следующие промежуточные величины:

$$M = L/N = 8228/160356 = 0{,}0513;$$
$$p_1 = m_1/L = 3637/8228 = 0{,}442;$$
$$Q = (1-p_1)^{-1} = 0{,}558^{-1} = 1{,}792.$$

Затем на их основе вычисляются параметры a и x:

$$a = (Q-M-1)^{-1} = (1{,}792-0{,}0513-1)^{-1} = 1{,}35;$$
$$x = aQ = 1{,}35 \cdot 1{,}792 = 2{,}42.$$

Например, $$m_2 = m_1 \frac{(a+1-1)}{(x+1)} = m_1 \frac{a}{(x+1)} =$$

$$= 3637 \cdot \frac{1{,}35}{3{,}42} = 1436;$$

$$m_3 = m_2 \frac{(a+1)}{(x+2)} = 1436 \cdot \frac{2{,}35}{4{,}42} = 763{,}5;$$

---

* Частным случаем распределения Уэринга-Хердана является распределение Юла (Yule G.U., 1944; см. Patil G.G., Joshi S.W., 1968). - В формуле (7) F - упорядоченные частоты слов ($F = 1, 2, \ldots, n$).

$$m_4 = m_3 \frac{(a+2)}{(x+3)}$$ и т.д.

Сравнение эмпирических и теоретических данных показывает хорошее соответствие (см. табл. 5). Считается, что модель Уэринга–Хердана работает хорошо лишь на выборках умеренного объема ($N < 200\,000$). Можно попробовать применить модель на ЧС русского языка (1977), где $N = 1\,056\,382$, $L = 39\,268$ и $m_1 = 13\,379$. Вычисления дают: $M = 0{,}0371$, $Q = 1{,}517$ и параметры $a = 2{,}08$ и $x = 3{,}16$. Соответствие эмпирических и теоретических данных достаточно хорошее в начальной части частотного спектра, примерно до m_{15}, но в дальнейшем модель дает сильно сниженные оценки спектра (табл. 6).

Модель Уэринга–Хердана замечательна тем, что она показывает переходы от m_1 до m_2, от m_2 до m_3 и т.д. Французская исследовательница Дольфэн (Dolphin, 1974; цит. по: Muller Ch., 1976) высказала гипотезу, что по этой модели можно будет определить и переход назад, от m_1 до m_0, причем под m_0 подразумеваются "нуль-частотные" слова, т.е. слова, которые (по предположению) относятся к лексикону автора или авторов, но которые в данном тексте не были использованы. Дольфэн предлагает на этот случай особый метод вычисления параметров a и x (см. Muller Ch., 1976, с.143):

$$a = \frac{\rho_1 + \bar{F}(q_1 - \rho_1)}{\rho_1 \bar{F} - 1} ;$$

$$x = \frac{a + \rho_1}{q_1} ;$$

где $\rho_1 = m_1 / L$; $q_1 = 1 - \rho_1$; $\bar{F} = N/L$.

Из формулы (7) следует, что $m_1 = m_0 \frac{(a-1)}{x}$ и

$$m_0 = m_1 : \frac{(a-1)}{x} . \qquad (8)$$

По данным романа А.Х. Таммсааре $a = 2{,}7$ и $x = 5{,}63$. Следовательно:

$$m_0 = 3637 : \frac{1{,}7}{5{,}63} = 12\,045.$$

Весь запас слов автора (в пределах данной тематики или жанра) определяется в объеме $L + m_0 = 6228 + 12\,045 \approx$ $\approx$ 20 000.

Наши предварительные опыты показывают, что метод Дольфэн можно использовать при стилометрическом анализе выборок приблизительно одинакового объема, причем показатель m_0 следует истолковать как относительную характеристику индивидуального стиля.

Модель логнормального распределения. Особый интерес представляет то обстоятельство, что спектральное распределение частот слов можно описывать также логарифмически нормальным распределением (см., например, Herdan G., 1960; Howes D., 1964; Carroll J.B., 1967; Williams C.B., 1970).

Логарифмически нормальное (логнормальное) распределение представляет собой распределение случайной величины x, логарифм которой $\ln x$ (не имеет значения, применяются ли десятичные или натуральные логарифмы) подчинен закону нормального распределения.

Логнормальное распределение определяется двумя параметрами: средней $\mu = \overline{\ln x}$ и среднеквадратичным отклонением логарифмов $\sigma = \sigma_{\ln x}$. Функция плотности логнормального распределения $\rho(x)$ имеет вид:

$$\rho(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \qquad (9)$$

Одним из наиболее эффективных и простых способов проверки логнормальности распределения является графическое расположение точек кумулятивных значений эмпирического распределения в логарифмически-вероятностных координатах, где одна из осей распределена по функции нормального распределения, а другая имеет логарифмическую шкалу. Если случайная величина x имеет логарифмически нормальное распределение, то графическое изображение эмпирической функции $\ln x$ будет представлять собой прямую.

В данном случае за случайную величину мы принимаем частоту слова F и вычисляем соответствующие накопленные вероятности (относительные численности) $\rho^*(F)$ отдельно для словаря и текста. Для построения логарифмически-вероятностной системы координат нужно найти квантили нормального распределения $\Phi(z)$. Квантили нормального распределения находятся в специальных таблицах (например, Ван дер Варден Б.Л., 1960; Tiit E., 1971). По материалу ЧС романа А.Х. Таммсааре эти данные приводятся в табл. 7, и на этой основе строится график (рис. 2). Как видно, на графике действительно наблюдается прямая линия в логарифмически-вероятностных

Таблица 7

Распределение частот слов (интегральное спектральное распределение) по данным ЧС лексем I-го тома романа А.Х. Таммсааре "Правда и право" (зст.): F - частота слова, p_c^* - покрытие словаря, p_T^* - покрытие текста, z - квантили нормального распределения. N = 160 356, L = 8228, m_1 = 3637, F_{max} = 7168.

F	p_c^*	z	p_T^*	z
I	0,442	- 0,I5	0,023	-2,00
2	0,590	0,23	0,038	-I,77
3	0,664	0,42	0,049	-I,65
4	0,7I8	0,58	0,060	-I,55
5	0,754	0,69	0,070	-I,48
6	0,78I	0,78	0,078	-I,42
7	0,802	0,85	0,085	-I,37
8	0,8I7	0,90	0,092	-I,33
9	0,830	0,95	0,098	-I,29
I0	0,84I	I,0	0,I03	-I,26
I5	0,874	I,I5	0,I25	-I,I5
20	0,895	I,25	0,I44	-I,06
25	0,908	I,33	0,I60	-0,99
30	0,9I8	I,39	0,I73	-0,94
40	0,934	I,5I	0,203	-0,84
50	0,946	I,6I	0,23I	-0,74
60	0,953	I,67	0,25I	-0,67
70	0,959	I,73	0,267	-0,62
80	0,963	I,79	0,285	-0,57
90	0,967	I,84	0,302	-0,52
I00	0,970	I,88	0,3I5	-0,48
I50	0,979	2,03	0,374	-0,32
200	0,985	2,I7	0,425	-0,I9
300	0,989	2,29	0,477	-0,06
400	0,992	2,4I	0,530	0,08
500	0,994	2,5I	0,578	0,20
600	0,995	2,58	0,607	0,27
I000	0,997	2,80	0,708	0,55
2000	0,9988	3,05	0,786	0,79
2800	0,9994	3,25	0,863	I,09
3700	0,9995	3,30	0,896	I,26
4300	0,9996	3,35	0,920	I,4I
5200	0,9998	3,55	0,955	I,70
7I68	I,0	-	I,0	-

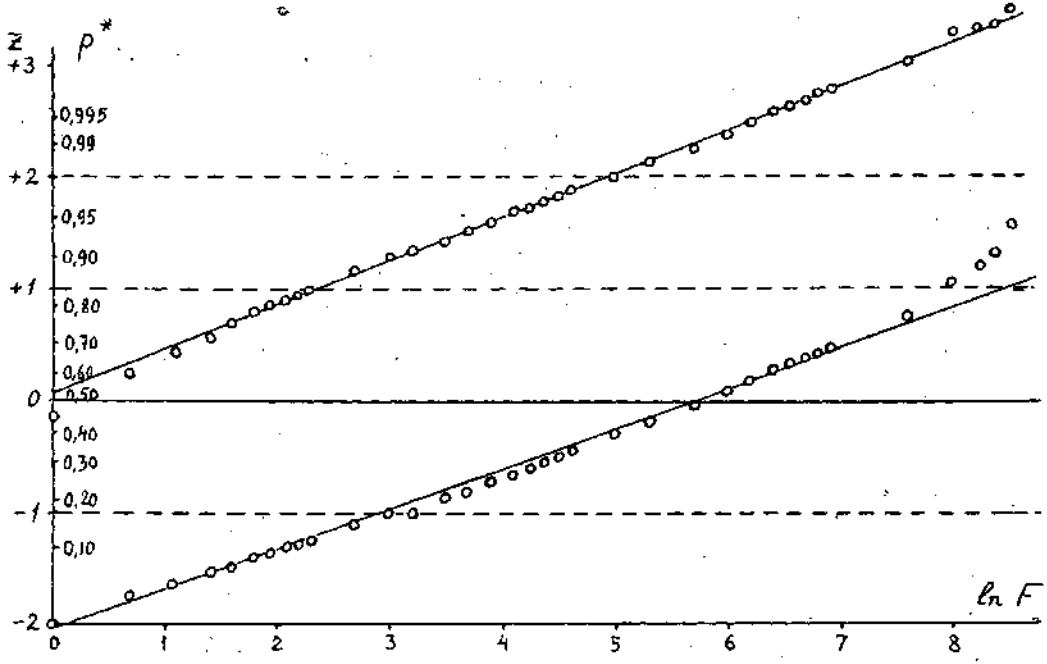


Рис. 2. Проверка на логнормальность спектрального распределения частот слов (частотного спектра) по данным ЧС лексем романа А.Х. Таммсааре (эст.). По абсциссе логарифмический масштаб, по ординате квантили нормального распределения (z) и соответствующее покрытие словаря или текста (p^*).

координатах как в отношении словаря, так и в отношении текста, причем согласно теории допускается небольшое отклонение в области больших частот (при $\rho^* > 0{,}95$). Однако небольшое отклонение наблюдается и в начальной части распределения, т.е. в области малых значений F. Можно предположить, что такое отклонение является в большой степени артефактом, порожденным невоспроизводимым "обрезанием" выборки ("обрезая" нуль-частотные слова, т.е. $F = 0$). По существу мы имеем дело с т.н. усеченным логнормальным распределением, так как спектральное распределение частот слов с необходимостью начинается с единицы ($F \geq 1$), а полное логнормальное распределение начинается с нуля.

Для устранения эффекта усечения иногда используется модификация логнормального распределения (см. Четыркин Е.М., Калихман И.Л., 1982, с. 121):

$$p(x) = \frac{1}{(x-k)\,\sigma\sqrt{2\pi}}\, e^{-\frac{[(\ln x - k) - \mu]^2}{2\sigma^2}}, \quad (10)$$

включающая еще один параметр k и представляющая собой более гибкую модель для описания усеченного логнормального распределения. Здесь, в частности, x изменяется не от 0 до ∞, а от k до ∞. В случае спектрального распределения частот слов $k = 1$. Эта поправка должна улучшать соответствие теоретических данных эмпирическим.

Значения параметров логнормального распределения (μ и σ) вычисляются обычно методом максимума правдоподобия или методом моментов (см. Липкин М.И., 1972, с. 77 и след.). Однако нахождение параметров усеченного логнормального распределения практически много сложнее, чем нахождение параметров полного распределения. Вполне можно довольствоваться графической проверкой, которая в нашем опыте показала безусловное сходство эмпирического распределения с логнормальным распределением.

Можно сделать вывод, что спектральное распределение частот слов при известных условиях (при "усеченной" форме распределения) может быть аппроксимировано логнормальным распределением. Как известно, логнормальное распределение используется для описания совокупного действия многих случайных факторов, когда их влияние на изменение конечного результата примерно пропорционально изменению самих факторов (Четыркин Е.М., Калихман И.Л., 1982, с. 120), т.е. для опи-

сания определенного рода вероятностных процессов. При такой интерпретации распределение частот слов (частотный спектр) может быть результатом некоторого вероятностного процесса, действующего при порождении речи. Такого мнения придерживаются многие исследователи, причем считается, что логнормальность распределения частот слов отражает "присущий естественному языку принцип оптимального кодирования информации" (Herdan G., 1964, с. 61-62). Отмечается также, что усеченное логнормальное распределение является "средством приближения к закону Ципфа" (Бычков В.Н., 1983, с. 81).

Однако имеются и возражения против использования логнормального распределения для аппроксимации спектрального распределения частот слов. Считается, что аппроксимация некорректна в силу того, что в данного типа лингвистическом распределении наблюдается зависимость моментов (средней и др.) от объема выборки, а это противоречит натуре "гауссовых" распределений, к которым относится и логнормальное распределение. Даже если частотный спектр конечной выборки удовлетворительно описывается логнормальным распределением, он от этого не делается "гауссовым", так как не сходится к гауссовому распределению с ростом объема выборки до бесконечности (подробнее см. Хайтун С.Д., 1983, с. 81 и 184-185). Таким образом, при такой точке зрения имеется формальное препятствие для объяснения сущности частотного спектра лексики на основе закона логнормального распределения.

Заключение. Спектральное распределение частот слов ("частотный спектр"), представляющее собой, наряду с ранговым распределением, одну из взаимосвязанных сторон общей (лексико-)частотной структуры текста, позволяет с квантитативно-системной точки зрения адекватно описывать структуру лексики в любом тексте естественного языка и в соответствующем словаре. Частотный спектр имеет (как и ранговое распределение) динамический характер, и его состав изменяется с изменением объема текста, причем это изменение подчиняется закономерности "неравномерного перехода". В общей сложности при увеличении текста относительная доля редких, в частности однократных слов, неуклонно уменьшается. Частотный спектр может рассматриваться как на уровне словаря, так и на уровне текста, и в обоих случаях он может быть представлен в дифференциальной (некумулятивной) или в интегральной (кумулятивной) форме. При известных условиях частотный спектр может служить

объективной типологической характеристикой языков или подъязыков, а также стиледифференцирующей характеристикой индивидуальных текстов.

Связь частотного спектра лексики с ранговым распределением отчетливо проявляется при аналитическом описании обоих типов распределений с помощью той или другой функции. Существует множество вариантов аналитического описания частотного спектра на основе закона Ципфа. В статье рассматриваются эти возможности и делается вывод, что как ранговое, так и спектральное распределение лексики соответствующим образом описываются одинаковой по форме степенной функцией, представляющей закон Ципфа и по существу выражающей принцип пропорциональности относительных приростов сопоставляемых переменных (закон постоянного относительного роста). Проведенный анализ показал также, что наиболее естественно совершается переход от ранговой формы распределения к спектральной, если исходить из ципфовского параметра γ, вычисленного для малочастотной зоны рангового распределения.

В статье рассматривались также некоторые другие возможности аналитического описания частотного спектра частот слов, более подробно подвергались анализу метод (модель) Уэринга-Хердана и метод аппроксимации с помощью логнормального распределения. Анализ выявил сильные, а также - слабые стороны этих методов.

Можно было констатировать, что все рассмотренные методы позволяют вполне удовлетворительно описывать спектральное распределение частот слов. Такая множественность решений не должна удивлять. В науке уже давно известно (и по мере накопления опыта это знание распространяется и в квантитативной лингвистике), что математическими моделями изучаемых явлений могут служить различные функции, задаваемые уравнениями прямой, экспоненты, гиперболы, логисты и т.д., а все дело в правомерности принятой системы исходных постулатов по отношению к конкретному явлению (см., например, Добров Г.М., 1969, с. 155). Нужен анализ процесса или явления по существу, по внутренней логике или по физической сущности с целью формирования адекватного представления об изучаемом явлении и его использования при выводе общего вида искомой модели или при интерпретации этой модели. Таким образом, предпочтение одной или другой модели (при их равноправном формальном соответствии эмпирическим данным) зависит от содержательного анализа конкретной проблемы. В нашем конкретном случае можно,

например, предпочесть модель Ципфа по той причине, что она выражает очень простую и естественную зависимость между переменными (пропорциональность относительных приростов) или что она в некотором смысле указывает на связь с деятельностью мозга (гипотеза А.Н. Лебедева). Известные отклонения от точного соответствия с законом Ципфа могут получить лингвистическое или другое объяснение. Можно предпочесть или принять в качестве возможной альтернативы модель Уэринга-Хердана по той причине, что здесь явно просматривается связь с закономерностью постоянного увеличения отношения m_{F+1}/m_F, наблюдаемого на практике, причем есть возможность учитывать наличие т.н. нуль-частотных слов. Принятие модели логнормального распределения позволяет рассматривать порождение речи как вероятностный процесс, что может служить основанием для определенных выводов о природе языка и т.д. Разные гипотезы могут быть приняты или не приняты другими исследователями, но в конце концов правильность вывода конкретной модели проверяется практикой.

Важно отметить то, что большинство задач, в том числе лингвистических задач, при математическом моделировании не решается однозначно, а сводится к получению ряда альтернатив. Недаром отмечается в современной теории анализа данных (см. Мостеллер Ф., Тьюки Дж., 1982), что "мысль о том, что существует одна-единственная регрессия, которую мы могли бы подобрать по заданному множеству данных, часто ложна" (с. 17). Следует также иметь в виду, что математическая модель всегда может быть лишь некоторым приближением, и на практике никогда не бывает распределений, в точности удовлетворяющих формуле, хотя есть как умозрительные, так и фактические основания считать, что многие эмпирические распределения должны ею аппроксимироваться (там же, с. 24). Цитируемые авторы называют "наивным оптимизмом" и "карикатурой" усердное применение критериев значимости при сравнении эмпирических распределений с теоретическими, так как это означало бы "полный уход от действительности с ее неопределенностью" (с. 42). Во многих работах последнего времени подчеркивается важность общей формы представления данных и ценность рисунков при попытках установить какую-то тенденцию или соответствие между теоретическими и эмпирическими распределениями. Делается даже парадоксальный вывод, что лучшей моделью является не та, которая полнее всего согласуется с наличными данными, а та, которая более обоснована с содер-

кательной или теоретической точки зрения (см., например, Техника и практика, 1971, с. 278). По-видимому, приходится считаться с этими тенденциями в науке и в квантитативной лингвистике, особенно когда речь идет о теоретических моделях, претендующих на объяснение изучаемых явлений.

ЛИТЕРАТУРА

Арапов М.В., Ефимова Е.Н. Понятие лексической структуры текста. – Научно-техническая информация. Серия 2. М., 1975, № 6, с. 3-7.

Бектаев К.Б. Статистико-информационная типология тюркского текста. – Алма-Ата: Изд-во Наука Каз. ССР, 1978.

Бичков В.Н. Лингвистическая статистика и проблема эквивалентности статистических описаний (моделей). – В кн.: Структурная и прикладная лингвистика. Вып. 2. – Л.: Изд-во Ленингр. ун-та, 1983, с. 72-82.

Ван дер Варден Б.Л. Математическая статистика./Перев. с нем., М., 1960.

Григорьева А.С. Статистическая структура русского эпистолярного текста (лексика частных писем). Канд. дисс. Л., 1981.

Добров Г.М. Прогнозирование науки и техники. – М.: Наука 1969.

Крылов Ю.К. Об одной парадигме лингвостатистических распределений. – Ученые записки Тартуского ун-та, вып. 628. Тарту, 1982, с. 80-102.

Лебедев А.Н. Закономерности повторения слов в речи. – Психологический журнал, т. 4. М., 1983, № 5, с. 11-22.

Липкин М.И. Кривые распределения в экономических исследованиях. – М.: Статистика, 1972.

Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1./ Перев. с англ. – М.: Финансы и статистика, 1982.

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

Пиотровский Р.Г. Текст, машина, человек. – Л.: Наука, 1975.

Теория и практика прогнозирования развития науки и техники в странах СЭВ. – М.: Экономика, 1971.

Тулдава Ю. Статистический метод сравнения лексического состава двух текстов. – Linguistica IV. Tartu: Изд-во Тарт. ун-та, 1971, с. 163-198.

Тулдава Ю.А. Частотная структура текста и закон Ципфа.– Ученые записки Тартуского ун-та, вып. 711. Тарту, 1985, с. 93-116.

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

Частотный словарь русского языка./ Под ред. Л.Н. Засориной. – М.: Русский язык, 1977.

Четыркин Е.М., Калихман И.Л. Вероятность и статистика. – М.: Финансы и статистика, 1982.

Brookes B.C. Quantitative Analysis in the Humanities: the Advantage of Ranking Techniques. – In: Studies on Zipf's Law./Ed. by H. Guiter and M.V. Arapov. Bochum, Brockmeyer, 1982, pp. 65-115.

Carroll J.B. On Sampling From a Lognormal Model of Word-frequency Distribution. – In: Computational Analysis of Present-day American English by H. Kučera and W.N. Francis. Providence, R. I., 1967, pp. 406-424.

Dolphin ___ Méthodes de la statistique linguistique; le vocabulaire fantastique de Malpertuis. Strasbourg, 1974. (Mimeogr.)

Herdan G. Type-Token Mathematics. – The Hague: Mouton, 1960.

Herdan G. Quantitative Linguistics. – London: Butterworths, 1964.

Howes D. Application of the Word-frequency Concept to Aphasia. – In: Ciba Foundation Symposium on Disorders of Language./ Ed. by A.V.S. de Reuck and Maeve O'Connor. London, 1964, pp. 47-75.

Krallmann D. Statistische Methoden in der stilistischen Textanalyse. Bonn, 1966.

Kučera H., Francis W.N. Computational Analysis of Present-day American English. – Providence, R. I.: Brown University Press, 1967.

Mandelbrot B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse. – In: Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics. Vol. XII. Providence, R. I., 1961.

Muller Ch. Initiation à la statistique linguistique. – Paris: Larousse, 1968.

Muller Ch. Some Recent Contributions to Statistical Linguistics. – In: Statistical Methods in Linguistics 1976. Stockholm: Skriptor, 1976, pp. 136-147.

Patil G.G., Joshi S.W. A Dictionary and Bibliography of Discrete Distributions. - Edinburgh: Oliver & Boyd, 1968.

Shannon C.E. A Mathematical Theory of Communication. - Bell System Technical Journal. Vol. 27, 1948, pp. 379-423, 623-656.

Tilt E. Matemaatiline statistika I. - Tartu: TRÜ, 1971.

Tuldava J. Sagedussõnastik leksikostatistiliise uurimise objektina. - Töid keelestatistika alalt II. TRÜ Toimetised, vihik 413. Tartu, 1977, lk. 141-171.

Villup A. A.H. Tammsaare romaani "Tõde ja õigus" I kõite autori- ja tegelaskõne sagedussõnastik. - Rmt.: Töid keelestatistika alalt III. TRÜ toimetised, vihik 446. Tartu, 1978, lk. 5-106.

Williams C.B. Style and Vocabulary; Numerical Studies. - London: Griffin, 1970.

Yule G.U. The Statistical Study of Literary Vocabulary. - Cambridge (Mass.): Cambridge University Press, 1944.

Zipf G.K. Human Behavior and the Principle of Least Effort. - Cambridge (Mass.): Addison-Wesly Press, 1949.

Õigekeelsussõnaraamat. / Toimetanud R. Kull ja E. Raiet. - Tallinn: Valgus, 1976.

THE FREQUENCY SPECTRUM OF TEXT AND VOCABULARY

Juhan Tuldava

S u m m a r y

The "spectral" distribution of word-frequencies, called the frequency (or lexical) spectrum of the vocabulary of a given text, is the counterpart of the rank distribution of word-frequencies and is closely connected with it. The spectrum has dynamic character, viz.it changes in such a manner that with increasing text size the rate of rare words gradually diminishes. The frequency spectrum of word forms may be used as an objective typological feature of languages and sublanguages and it may also be used as a characteristic of individual style. The article deals further with the possibilities of the analytical description of a frequency spectrum on the basis of Zipf's law, the model of Waring-Herden, and the lognormal distribution. The theoretical analysis is illustrated by examples from Estonian, English, and Russian.

ПРОБЛЕМЫ КВАНТИТАТИВНОГО АНАЛИЗА ТЕКСТА

(по материалам семинаров межвузовской проблемной группы "Текст как объект междисциплинарного исследования". Вороново, 8-12.10.85; Тарту, 27-31.01.86)

М.Г. Борода, В.А. Долинский

Проблемы комплексного исследования текста как специфического объекта сложной природы вызывают в настоящее время все возрастающий интерес. Характерной чертой нового этапа в развитии этих исследований является не только широкое применение количественно-математических методов, но и попытки связать полученные результаты (а подчас - и саму постановку задачи) с гипотезами и рассуждениями о качественных характеристиках текста - его художественной формы, жанра, языка, и т.д. Важным проявлением этой тенденции является, в частности, наметившийся в квантитативной лингвистике переход от "чисто статистических" моделей к моделям, где порождающие принципы связаны с гипотезами психологического характера о закономерностях восприятия и генерирования текста. Не менее характерным для нового этапа в развитии анализа текста является и интерес к его квантитативному изучению со стороны различных специалистов, включение в орбиту этих исследований текстов нелингвистической природы - музыкальных текстов, "текстов" живописи, а также особых текстов типа ассоциативных потоков, и т.п. При этом наблюдается тенденция к интеграции этих исследований к поискам на этой интегративной основе общих принципов порождения текста как связного целого, к определенному единству программ и комплексности. Естественным в этих условиях является образование рабочих групп по наиболее актуальным проблемам квантитативного анализа текста.

Одной из таких групп стала образованная в начале 1985 года по инициативе ряда исследователей, ведущих разработки в области квантитативной лингвистики и анализа текста, межвузовская проблемная группа "Текст как объект междисциплинарного исследования: качественные и количественные закономерности его организации". В состав группы вошли представители различных специальностей - лингвисты, математики, музыковеды, психологи - и ВУЗов Москвы, Ленинграда, Тарту, Тбилиси,

Еревана, Киева, Минска, Риги и др. городов. Целью группы является содействие прогрессу в комплексном – квалитативно-квантитативном – исследовании текста, изучении общих и специфических (связанных с конкретным языком, жанром, стилем, наконец, с целевой направленностью текста) принципов и закономерностей его организации, построении математических моделей его порождения и, в перспективе, – формирование исследовательских программ в этих областях. Работа группы – являющейся исследовательским коллективом с "открытой структурой" – организуется бюро в составе: Ю.А. Тулдава (Тартуский гос. университет), председатель оргбюро; М.Г. Бородa (Тбилисская гос. консерватория); В.К. Детловс (Латвийский гос. университет); А.А. Поликарпов (Московский гос. университет). На проводимых раз в полгода семинарах группы заслушиваются доклады ее участников и других приглашенных специалистов, и обсуждаются актуальные проблемы комплексного анализа текста. В настоящем сообщении дан обзор результатов первых двух семинаров группы: I-го, проведенного в октябре 1985 г. в пос. Воронове на базе МГУ им. Ломоносова, и 2-го, проведенного в январе 1986 г. на базе Тартуского государственного университета.

Организатором первого семинара группы явилась кафедра общего, сравнительно-исторического и прикладного языкознания МГУ.

На пленарном заседании были заслушаны доклады П.М.Алексеева (ЛГПИ им. А.И. Герцена) и Ю.А. Тулдава (Тартуский гос. университет).

В докладе П.М. Алексеева "О знаковости текста и уровнях его описания в квантитативной лингвистике" были проанализированы особенности естественных знаковых систем и представлена многоуровневая схема обобщения лингвистической информации. Доклад Ю.А. Тулдава "Теоретические вопросы статистической организации текста" ознакомил с различными концепциями и методиками, существующими в квантитативной лингвистике. Докладчиком были проанализированы более десятка аналитических формул ранговых распределений – вариантов закона Ципфа. Была подчеркнута важность поисков онтологических обоснований этого закона. Значительное внимание в докладе было уделено проблеме унификации терминологии, связанной с квантитативно-лингвистическим анализом текста.

Далее работа семинара была продолжена по трем направлениям. Первое из них – "Общие проблемы статистической органи-

зации лексики" - открылось докладом Г.Я. Мартыненко "О гауссовых и негауссовых распределениях в филологической науке". Автор показал, что возведение закона Ципфа в ранг "парадигмы лингвистических распределений" не обосновано. Неустойчивость параметров и рост дисперсии с ростом объема выборки наблюдается также и во многих природных явлениях. Характерным свойством распределений такого рода является различие элементов по их функции в тексте.

В докладе Ю.К. Крылова "Порождение текста как слабостационарный случайный процесс" было показано, что порождение текста описывается, в терминах теории случайных процессов, авторегрессией 2-3-го порядка - независимо от объема выборки (в первом приближении). В докладе Ю.К. Орлова "О приложениях обобщенного закона Ципфа-Мандельброта" описывалась построенная автором на базе модели В. Калинина модель статистически однородного квазитекста, позволяющая строить гипотезу о совокупности как набор математических описаний частотных спектров и словарных запасов на выборках (текстах) разного объема. Последнее дает возможность сравнения статистической структуры текстов различного объема путем "приведения" их к одному объему (называемому автором "объемом Ципфа"). В.Е. Остапенко в докладе "Квантитативные модели текста" рассмотрел методику отбора терминологической лексики для использования ее в информационно-поисковой системе анализа лексической структуры текстов. Анализ ранговых, дисперсионных и параболических распределений, отображающих рост словаря при нарастании объема текста, позволил автору классифицировать почти половину из 1000 наиболее употребительных в данном корпусе лексем.

Моделированию частотных характеристик различных языковых единиц на базе логнормального закона были посвящены доклады М.В. Овсянникова и Н.С. Манасян. Л.М. Сутягина в докладе "О горизонтальных распределениях" ознакомила с опытом таксономического анализа частотных характеристик по выборкам минимального объема.

С большим интересом участники семинара прослушали доклад А.Н. Лебедева "Некоторые онтологические основания закона Ципфа". Исходя из гипотезы о кодировании образов памяти пакетами когерентных волн нейронной активности и предположения о примерном равенстве относительной частоты слов в речи относительной частоте активаций образов этих слов, автор предложил оригинальную формулу зависимости частоты F_i слова

в тексте от его ранга i (при обычном в лингвостатистической практике ранжировании слов по невозрастанию их частоты): $F_i = Rq^{-1}\ln(1-\frac{q}{i})$, где R -критическая разность фаз между кодирующими волнами нейронных разрядов, q - диапазон колебаний ранга слова.

Часть представленных на семинаре докладов группировалась по проблематике "Теоретические и прикладные аспекты количественного анализа лексики". В докладе А.А. Поликарпова "Полисемические спектры" были представлены результаты исследования рангово-полисемических и спектрально- полисемических распределений в обширном корпусе толковых словарей русского и английского языков. Была показана зависимость параметров этих распределений от типа словаря и степени аналитичности данного языка. В докладе Е.А. Ильюшиной "Анализ системной организации художественного текста методом ранговых распределений" была показана возможность использования случайных чисел для сопоставительного анализа ранговых структур языковых единиц разного уровня. М.А. Марусенко в докладе "Оптимизационный подход к анализу сложных лингвистических объектов" остановился на нерешенных проблемах атрибуции текстов. Автором была предложена пятичленная классификация лингвостатистических признаков текста и методика измерения признакового расстояния между текстами в евклидовом пространстве.

С интересными результатами в русле квантитативного анализа музыкального текста познакомил доклад В.К. Петловса, посвященный актуальной проблеме ладовой организации мелодии. На материале народной и профессиональной музыки автором была показана связь ладовой организации с направлением звуковысотного движения мелодии (более подробно эта проблематика была развернута автором в его докладе на 2-м семинаре группы - см. ниже).

Заседания в секции "Нетрадиционная проблематика в квантитативной лингвистике" открылись докладом А.А. Поликарпова "Асимметричный дуализм языкового знака: системно-количественный аспект". Развивая идеи С.О. Карцевского о системной взаимообусловленности синонимии и полисемии, автор высказал гипотезу, что в языках с более высокой полисемичностью лексических единиц синонимические их отношения развиты в меньшей степени. Параметры синонимических и полисемических распределений оказываются, в определенном смысле, взаимно-дополнительными. Рангово-полисемические распределения в словарях английского языка рассматривались в докладе А.В. Малова "Опыт

системного анализа количественных характеристик лексики английского языка". Докладчиком был предложен ряд аналитических формул, аппроксимирующих исследованные распределения.

В.А. Долинский в докладе "Ассоциативный эксперимент: распределение реакций как функция частоты и полисемии стимула" рассказал о результатах исследования ранговых распределений слов-реакций, полученных в экспериментах по свободному ассоциированию, проводившихся с русскими и американцами. Как было показано в докладе, вид рангового распределения ассоциаций зависит от частотной и полисемической характеристик слова-стимула и описывается гиперболическим законом Ципфа. Примененный математический аппарат позволил "развести" параметры частоты и полисемии и выявить сложный характер влияния каждого из этих факторов.

В докладе В.П. Белянина "Психолингвистический подход к типологии художественного текста" была предложена классификация текстов беллетристики, основанная на психологическом типе отношения личности к миру, определяющем выбор тех или иных лексико-семантических групп в высказывании.

Доклад Тань-Аошуан "Синтаксическое формление текста аморфного языка и языковая конвенция" был посвящен трудностям гносеологического характера, возникающим при формализации некоторых малоизученных семантико-синтаксических характеристик китайского языка.

На семинаре были также заслушаны доклады Ю.Б.Сафроновой "Системно-количественные характеристики синонимии, полисемии и антонимии в английском словаре синонимов"; О.В. Бубуевой "Системно-количественные закономерности соотношения полисемических характеристик слов в тексте и словаре (на примере английского языка)"; А.М. Карапетянца "К вопросу анализа ритмической организации китайского общения и текста"; О.В. Баракина "Восприятие текста личностью и общественной группой".

В ходе работы семинара состоялась дискуссия за круглым столом по проблеме "Закон Ципфа на современном этапе развития лингвистики". Было сосредоточено внимание на ряде актуальных вопросов. Каково онтологическое содержание закона Ципфа? Каковы критерии его выполнимости? В чем сходство и различие наблюдаемых в природе и обществе явлений, которые описываются этим законом. В выступлениях подчеркивалось, что накопление и уточнение математических формул, описывающих лингвистическую реальность, сколь бы важным оно ни казалось

само по себе, не может являться самоцелью.

В целом, работа I-го семинара группы показала целесообразность и естественность интеграции усилий специалистов различных профилей в изучении квалитативных и квантитативных аспектов организации текста.

2-й семинар группы был организован на базе группы прикладной лингвистики Тартуского гос. университета. В семинаре приняли участие как члены группы "Текст как объект междисциплинарного исследования", так и другие приглашенные специалисты. На семинаре было заслушано около 40 докладов по проблемам статистической организации текста, лингвистических распределений, диалога, структурной организации словаря и текста, классификации и автоматического анализа текстов и проблемам квантитативного музыковедения. Тематика ряда докладов развивала темы сообщений, сделанных авторами на I-м семинаре группы в Вороново.

На пленарных заседаниях были заслушаны доклады Ю.М. Лотмана о трех функциях текста (передача устойчивой информации, порождение новой информации, культурная память), Р.Г. Пиотровского о конфликтных взаимоотношениях естественного языка и языка ЭВМ при автоматической обработке текста и о возможных путях их преодоления или "обхода", опирающегося на более глубокий, чем это осуществляется в настоящее время, анализ коммуникативно-семиотических аспектов текста, А.Н. Лебедева о предложенной им психофизиологической модели квантитативной организации текста, базирующейся на представлениях о периодических процессах как основе памяти и слове как наборе последовательных пачек нейронных импульсов, А.В. Зубова о статических и динамических компонентах текста и возможностях их формализованного анализа, И.А. Чернова и С.Т. Золяна о семантике текста и проблеме семиировой достижимости. Ряд поставленных авторами проблем обсуждался далее на секционных заседаниях.

На секции "Лингвистические распределения" был сделан постановочный доклад Ю.А. Тулдава "Об интерпретации лингвистических распределений". В нем был очерчен общий подход к анализу лингвистического распределения как специфического объекта, показана значимость различения одно-, много-объектного и комплексного распределений и рассмотрены три уровня анализа распределения: по форме, по содержанию и по происхождению (генетический аспект). Особое внимание было уделено важной в квантитативном анализе текста проблеме количествен-

ного оценивания лингвистического объекта, приписывания ему количественной характеристики. Как было показано в докладе, недооценка содержательного аспекта этого оценивания может быть источником серьезных ошибок в квантитативном анализе организации текста. На секции были также заслушаны доклады Г.Г. Сильницкого о корреляционных соотношениях между глагольными признаками в тексте и их значимости для индивидуальности текста, и Н.С. Манасян о комплексном анализе лингвистических распределений на материале текстов программного обеспечения для системы ЕС ЭВМ на русском языке.

Секция "Классификация и автоматический анализ текстов" была посвящена проблемам автоматической переработки текстов на микро-ЭВМ (доклад К.Я. Лепа), новым подходам к анализу "трудности" и "читабельности" учебных текстов (доклады Я.А. Шикка и Л.И. Васильченко, Х.П. Лийва, К.Э. Соомере), лингвистическим проблемам моделирования логико-смысловой структуры английского научно-технического текста на ЭВМ (доклад А.А. Лариной), автоматическим методам установления авторства текста (доклад Г.В. Ермоленко), классификационным задачам квантитативной лингвистики (Г.Я. Мартыненко) и проблеме классификации текстов с помощью факторного анализа (доклад Ю.А. Тулдава), проблемам квантитативного подхода к анализу ассоциативных потоков у здоровых и душевнобольных (доклад М.Г. Бороды и В.Э. Пашковского). В докладах обращала на себя внимание общность в постановке проблем, стремление найти новые подходы к их решению. Так, в докладе Ю.А. Тулдава была показана эффективность предложенного автором метода факторного анализа существенных квантитативно-лингвистических признаков текста в задачах классификации художественных текстов. В докладе Г.В. Ермоленко продемонстрирована существенная роль высокочастотной и низкочастотной лексики в стилистической (авторской) индивидуальности текста, важность учета обоих этих лексических слоев в задачах атрибуции текста. Новые возможности квантитативного подхода к оценке читабельности и трудности учебного текста были показаны в докладах Х.П. Лийва, Я.А. Шикка и Л.И. Васильченко, эффективный подход к изучению методики и теста восстановления текста с пропусками был предложен в докладе К.Э. Соомере. В докладе К.Я. Лепа были описаны существенные ограничения, характеризующие обработку текстов на микро-ЭВМ типа "Агат", в докладе А.А. Лариной - отмечена возможность моделирования логико-смысловой структуры английского научного текста на ЭВМ на базе анализа

его абзацной структуры. В докладе М.Г. Бороди и В.З. Пашковского была показана значимость, и предложен метод, анализа ритмических характеристик ассоциативного потока (связанных с образованием цепочек слов одинаковой длины или с одинаковым местом ударения, и т.п.) в диагностических задачах и в исследовании процесса генерирования текста.

Принципиальным проблемам квантитативной организации текста были посвящены две секции – "Статистическая организация текста" и "Структурная организация словаря и текста". На первой из секций активно обсуждались различные альтернативные математические модели лексической (статистико-лексической) структуры текста, предложенные М.В. Араповым, Ю.К. Крыловым и Ю.К. Орловым. Существенным качеством двух первых подходов, как отмечалось в обсуждении докладов на секции, является стремление авторов связать систему своих гипотез о тексте и механизмах, порождающих наблюдаемые в нем количественные закономерности, с некоторыми интуитивно естественными общими к а ч е с т в е н н ы м и принципами (принцип распределения слов по ячейкам классификации и построенная на его основе модель – М.В. Арапов; принцип симметрии функциональных свойств элементов связного текста и языка – Ю.К.Крылов). Как было отмечено в выступлениях участников секции, тот факт, что лексические распределения (например, распределение Ципфа) оказываются в ы в о д и м ы м и в этих моделях из некоторых простых принципов, а не постулированными, делает их ценными как в онтологическом плане, так и в плане построения адекватной д и н а м и ч е с к о й модели связного текста – в особенности, художественного. Помимо названных докладов, на секции были обсуждены доклад В.Н. Бычкова о проблеме выборки "оптимального объема" (понимаемого автором как объем, на котором текст выполняет закон Ципфа-Мандельброта в канонической форме) и подходе к ее решению на базе нелинейной ципфовской модели, С.И. Гиндина о разработанной им многоцелевой системе количественных параметров текста, И.Ш. Надарейшвили о статистической структуре "Витязя в тигровой шкуре" Ш. Руставели.

На секции "Структурная организация словаря и текста" рассматривались, в целом, проблемы квантитативного исследования с и с т е м н о с т и их организации. Проблемы эти были подробно обсуждены в докладе А.А. Поликарпова, где рассматривались полученные автором результаты исследований полисемических распределений по данным толковых словарей и бы-

ло показано существование параметров этих распределений, чувствительных к типологическому различию языков и, с другой стороны, типу толкового словаря. Рассмотренные докладчиком вопросы затрагивали актуальнейшие в квантитативном анализе текста проблемы структурных единиц и критериев их различения (классификации). Вторая из этих проблем составила тему доклада М.Д. Якубовской, где рассматривался вопрос о зависимости частотных характеристик единицы текста от определяющих эту единицу классификационных условий. В коллективном докладе В.И. Перебейнос, Т.А. Грязнухиной, М.П. Муравицкой, Н.П. Дарчук (докл.), Л.И. Комаровой обсуждались результаты исследования закономерностей структурной организации реферативного текста, в докладе Л.В. Орловой рассматривались условия формирования типов сверхфразовых единств, характерных для научных текстов. Проблеме крупных структурных единиц текста и их квантитативному анализу был посвящен доклад И.А. Болдак об абзаце научного текста как репрезентанте его тематической организации. Проблемы связи квантитативных характеристик распределения текстовых единиц и некоторых качественных характеристик языка и текста рассматривались в докладах С.В. Райтар ("Количественные оценки контекстуальной реализации эстонского слова") и С.П. Клявинь ("Лингвостатистическое сопоставление функциональных стилей"). Обсуждение всех докладов на секции "Структурная организация словаря и текста" показало большую значимость дальнейшей разработки проблемы естественных структурных единиц текста и естественных критериев их классификации, и возможность базировать решение этих проблем на массивных и целенаправленных исследованиях квантитативной организации текста.

Актуальным проблемам квантитативной теории диалога была посвящена специальная секция. На ней были рассмотрены проблемы оптимальной организации современных лингвистических процессоров и, в связи с этим, проблемы типологической классификации диалога (доклад Б.Ю. Городецкого), построения формализованной модели динамического поведения участников коммуникативного взаимодействия (доклад В.М. Сергеева и М.А.Сиверцева), проблема общих принципов моделирования диалога с ЭВМ на естественном языке (доклад Х.Я. Ыйма, доклад М.З.Койт и Т.А. Роосмаа), проблема анализа диалога как специфического вида текста (доклад Г.В. Андрусенко). Проведенное обсуждение докладов на секции показало связь проблем исследования диалога с общей проблематикой квалитативно-квантитативного ана-

лиза текста, перспективность сотрудничества, обмена идеями, методами, исследователей в этих областях.

Значительный интерес участников семинара вызвала работа секции музыковедения. Заслушанные на ней доклады, проведенные обсуждения показали разносторонность исследовательских интересов в области анализа музыкального языка и текста, внимание к фундаментальным проблемам в этой области. Так, доклад В.К. Детловса был посвящен проблеме количественного анализа лада - важнейшего элемента музыкального языка. Автор показал на широком стилистическом материале связь смены направления в мелодии с теми или иными конкретными ступенями лада, роль ладовой функциональности как "режиссера поведения" звуковысотности; при этом обнаруженные автором явления носят характер метастилистической, языковой закономерности. Проблема соотношения "языкового" и "текстового" на уровне n-звуковых последовательностей (своеобразных аналогов n-грамм) рассматривалась в докладе И.В. Бахмутовой, В.Д. Гусева и Т.Н. Титковой, в котором описывались методы машинного распознавания "бытующих" и "оригинальных" интонаций (под которыми авторы понимали названные n-граммы), а также - предложенное авторами упорядочение композиторов по степени использования ими "бытующих" интонаций. "Языковые" аспекты музыкального мышления рассматривались также в докладе М.Г. Бороды, посвященном анализу общих закономерностей, стилистических характеристик и принципов эволюции распределений длин выделенных автором мелодических единиц типа "микромотива" и крупного мотива. Автором была предложена гипотеза об аналитизации языка европейской музыки последних трех столетий и о существовании внутри этого процесса "аналитиаационно-синтетизационных волн", характеризующих развитие и смену музыкальных стилей. Значительный интерес представило сообщение Я. Росса и И. Рэйтел об автоматической нотации одноголосных мелодий. В рамках предложенного авторами подхода эта актуальнейшая для квантитативного музыковедения проблема решается методами, близкими методам выделения частоты основного тона в речи. Помимо обсуждения докладов, предметом серьезной дискуссии на секции музыковедения были вопросы сегментации музыкального текста, выделения естественных и строго определенных музыкальных единиц, проблема естественных критериев классификации, а также проблемы применения в исследованиях музыкального текста и языка методов количественной лингвистики.

По завершении работы секций семинара состоялось обсуждение основных его докладов, а также дискуссия за круглым столом по нерешенным проблемам, связанным с исследованием и интерпретацией закона Ципфа. Дискуссия эта, явившаяся продолжением начатой по этой же проблеме на I-м семинаре группы, выявила необходимость фундаментальной проработки проблем, связанных с уровнем исследования частотных закономерностей (частотной организации) текста - конкретно, с рангом исследуемых структурных единиц и с типом критерия их классификации (различения). Дискуссия по проблемам "ципфовских" закономерностей показала также крайнюю актуальность работы по унификации квантитативно-лингвистической терминологии. Наконец, проведенные обсуждения показали важность специальных исследований частотной организации текстов нелингвистической природы для понимания принципов, порождающих закономерности типа закона Ципфа и, в более общем плане, - для понимания принципов квантитативной организации текста.

На заключительном заседании семинара был принят выработанный оргбюро группы Итоговый документ, в котором были обобщены основные результаты семинара, определены рекомендации направлений дальнейшей работы группы.

PROBLEMS OF QUANTITATIVE TEXT ANALYSIS

M. Boroda, V. Dolinskiy

S u m m a r y

Recent years have seen the emergence of a new approach to text analysis which aims to link up quantitative and qualitative text characteristics through psychological models of text generation, integration of studies of linguistic and nonlinguistic (musical, art, etc.) texts, etc. To promote the approach, a research group - "Text as an object of interdisciplinary study" - has been set up, with members from various cities of the USSR. The group holds seminars twice a year. A survey is given of the first two seminars (October 1985 and January 1986), sponsored by the Moscow and Tartu State Universities, respectively. Both seminars centered round problems of interpreting Zipf's Law.

The organizing committee includes J. Tuldava (chairman, Tartu State University), M. Boroda (Tbilisi Conservatoire), V. Detlovs (Riga), A. Polikarpov (Moscow).

С О Д Е Р Ж А Н И Е

Хроника:

SUMMARIES - RESUMEES - RÉSUMÉS

Survey:

Ученые записки Тартуского государственного университета.
Выпуск 745.
КВАНТИТАТИВНАЯ ЛИНГВИСТИКА И АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ.

На русском языке.
Резюме на разных языках.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликоола, 18.
Ответственный редактор Ю. Тулдава.
Подписано к печати 16.09.1986.
МВ 07945.
Формат 60х90/16.
Бумага писчая.
Машинопись. Ротапринт.
Учетно-издательских листов 10,76. Печатных листов 11,0.
Тираж 550.
Заказ № 647.
Цена 1 руб. 60 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Тähe, 78.
6 - 2