П.М.Алексеев, Е.В.Мурмуридис
Ленинградский пединститут

К ПРОБЛЕМЕ СТАТИСТИЧЕСКОГО ОТБОРА ТЕРМИНОВ В УЧЕБНЫЕ МИНИМУМЫ

При отборе учебного материала, в частности терминологического, возникает целый ряд проблем и вопросов, решение которых требует тщательного обсуждения в процессе поисков и экспериментов. Умозрительные, оторванные от практики суждения приносят здесь мало пользы. Если не оптимальные, то близкие к ним решения могут быть подсказаны лишь неоднократным обращением к тексту, к фактическому использованию языка в реальной экстралингвистической обстановке, сопоставлением результатов многих наблюдений и их обобщением.

Практика составления учебных минимумов на лингвостатистической основе уже имеет довольно длительный опыт, и тем не менее постоянно возникают все новые проблемы и затруднения. Поэтому до "единственно правильного", универсального решения этих проблем столь же далеко, как и до создания единой, унифицированной методики обучения языку любых обучаемых в любых условиях. Можно только рассчитывать на большую или меньшую универсальность методических решений в типовых ситуациях, обобщающих множество уникальных, неизбежно различных условий.

Даже в такой, казалось бы, явно специфической и потому стандартной сфере методики, как отбор учебного терминологического материала, приходится учитывать возможность более чем одного решения. Если существует, скажем, признанное всеми определение термина как слова или словосочетания, отражающего понятие, которое входит в систему понятий той или иной предметной области, то уже использование этого определения в рабочей процедуре выделения термина, его обнаружения в тексте связано с немалыми трудностями. Наблюдатель должен хорошо знать данную предметную область и обладать необходимой лингвистической квалификацией. Эти два качества и создают то, что принято считать эрудицией и интуицией лингвиста-терминолога. Менее зависят от личных свойств наблюдателя проблемы, связанные с различными представлениями о структуре данной предметной области. Разные специалисты, разные школы могут по-разному определять одну и ту же область или в одно определение вкладывать

разное содержание. Неизбежны разногласия и при трактовке терминов как общенаучных, отраслевых, специальных, относящихся к смежным или отдаленным областям и т.д.

Поскольку ни лингвисту, даже лингвисту-терминологу, ни его консультанту не прийти к решению однозначному, не вызывающему возражений со стороны, остается сделать произвольный выбор между двумя подходами.

Первый подход состоит в том, чтобы рассматривать терминологию подъязыка, т.е. все термины, обнаруженные в текстах по тематике данной предметной области, включать в терминологию подъязыка. Второй подход состоит в том, чтобы исходить не из подъязыка (языковой подсистемы, обслуживающей данную предметную или функциональную область и отраженной именно в текстах по данной тематике в рамках данного стиля, жанра и.т.д), но из терминологической системы (подсистемы) данной предметной области.

Преподавателю иностранного языка, обучающему студентов--"предметников", первый подход должен импонировать как более реалистический и более формализуемый, чем второй. Этот второй, видимо, более приемлемый для специальных лингво-терминологических исследований, менее удобен для количественных наблюдений над текстом, особенно при ограниченных масштабах наблюдения, и более связан с "дедуктивными" описаниями лингвистических систем, с тезаурусным подходом к изучению системности лексики.

Содержание настоящей статьи связывается с первым подходом, как более соответствующим скромным возможностям одного или немногих исследователей-составителей учебных терминологических частотных словарей (ЧС). Привлекаемый здесь материал рассматривается не в плане описания терминологических систем предметных областей, а в плане описания систем подъязыков, "обслуживающих" эти предметные области.

При расширении или углублении лингвостатистического анализа отраслевых подъязыков с целью классифицировать их лексику на термины и нетермины возникает еще одна проблема. Не так давно казалось, что главное - это составить полный перечень разных слов, употребляемых в тексте или совокупности текстов той или иной тематики. Если в процессе составления такого перечня подсчитывать, сколько раз употребляется каждое слово в тексте, в

результате можно получить ЧС слов этих текстов. По готовому ЧС как будто не слишком трудно определить, какие из них термины, а какие нет, особенно если тексты написаны на темы очень ограниченного круга. Так поступали и поступают почти все составители отраслевых ЧС - минимумов. Однако такие ЧС начинают составлять не всегда с несколькими целями, но, скажем, с конкретной основной целью - изучить именно терминологическую лексику, не рассматривая вовсе общеупотребительную. Тогда становится очевидным, что в первом случае (выделения терминов из готового ЧС всех слов текста) утрачиваются важные сведения. Это, во-первых, сведения о тех словах, которые в одном и том же тексте были использованы в разных значениях - в терминологических и нетерминологических, и, во-вторых, сведения о тех словах, которые, не будучи терминологичными в "изолированном" употреблении, становятся такими в составе терминологического словосочетания. Поэтому решение здесь видится в следующем. Если предстоит получить по возможности свободный от указанных недостатков ЧС терминов подъязыка, надо регистрировать в текстах те слова и словосочетания, которые использованы в нем в терминологическом значении. Это повышает ответственность наблюдателя и вынуждает его чаще консультироваться со специалистами, справочниками, учебниками и другими источниками информации. Хотя некоторые издержки неизбежны и здесь, ошибок все же должно быть меньше, чем в первом случае.

Следует отметить еще одно отличие результатов, получаемых в первом случае (если исходить из подъязыка), от результатов, которые можно получить во втором случае (если исходить из терминосистемы предметной области). Именно в силу того, что "подъязыковой" подход ориентирован на учебные цели, при нем допускаются решения, неприемлемые с точки зрения ортодоксального терминолога. В число единиц ЧС могут попасть слова, которые едва ли можно считать терминами, но которые являются своеобразными маркерами научно-технического стиля в целом или какого-то конкретного подъязыка. Например, в число самых частых слов английских научных подъязыков входят слова fig.(рис.),result (результат), obtain (получать), а слово let (в значении пусть, допустим, что) является одним из самых частых в английском математическом тексте. Поэтому, говоря об учебном терминологическом минимуме, не обя-

зательно иметь в виду словарь терминов в строгом понимании, и это все-таки не снижает применимости такого минимума для учебных целей. Не следует, однако, завышать его пригодность для чисто терминологических исследований. Разумнее просто предполагать, что терминолог сможет найти в нем полезный для себя материал.

При планировании работы по составлению ЧС-минимума терминологической лексики приходится учитывать одно важное свойство лексикостатистической организации текста. Сумма частот текстовых терминоупотреблений может не превышать половины всех словоупотреблений и обычно лежит в пределах 30-40% общей длины текста. В словаре этого же текста, т.е. среди разных слов, использованных в тексте, эта доля может оказаться выше, но не на много, как это показано в табл. I.

Таблица I

Доля терминов среди словоупотреблений текста и разных слов словаря этого текста по данным ЧС английских подъязыков электроники, квантовых генераторов, математики, биологии и психологии*

Доля терминов, %	Подъязык				
	Электроника	Квантовые генераторы	Математика	Биология	Психология
В тексте	40,56	32,I7	32,0I	33,65	38,40
В словаре	58,77	-	53,6I	58,46	34,52

* Используются цифровые данные, извлеченные из ЧС - минимумов [8; 7; 6; 4; 5;] . В каждом случае объем выборки равен 200 тыс. словоупотреблений.

Приведенный пример наглядно показывает, какой экономии можно достичь, исключая из подсчетов 40-60% используемых в тексте словарных единиц, на которые приходится 60-70% всех словоупотреблений. Это прежде всего служебная лексика и затем общеупотребительная. Однако энтузиазм относительно такой экономии не может не сдерживаться тем, что частоты основного

корпуса терминологического ЧС невысоки, как это показывают данные табл. 2. А это значит, что для получения так называемых "достоверных" частот объем текстовой выборки, измеряемый в словоили терминоупотреблениях, следует существенно увеличивать.

Таблица 2

Частоты самых частых терминов и нетерминов в английских подъязыках электроники и биологии

№ п/п (ранги)	Частота в подъязыке			
	Электроника		Биология	
	Нетермин	Термин	Нетермин	Термин
I	I9I58	875	I7208	I2I2
2	I027I	836	I0075	646
3	8870	736	87I6	50I
4	505I	665	5699	472
5	4937	64I	5603	426
6	4432	586	5536	409
7	2330	542	2I58	366
8	I952	536	I996	3I9
9	I884	534	I728	3I7
I0	I722	437	I652	307
ИТОГО	60607	6388	60371	4975

Особую остроту приобретает вопрос об объеме выборки для получения достоверных сведений о частотах терминологических словосочетаний. Здесь вмешивается еще одна лингвостатическая закономерность, которая в общем формулируется так: чем короче лингвистическая единица, тем она чаще встречается, и наоборот. Это можно в достаточно простой и наглядной форме видеть на материале табл. 3-4. Отсюда следует вывод о том, что для "достоверной" статистики длинных единиц требуется выборка в общем настолько бо́льшая, насколько эти единицы длиннее коротких, при том же уровне достоверности.

Таблица 3

Распределение частот терминологических словосочетаний по длине (количеству слов) в английском подъязыке электроники

Количество слов	Доля в тексте	Доля в словаре
2	65,72	26,00
3	28,10	40,00
4	5,34	22,00
5	0,58	8,00
6	0,28	4,00
ИТОГО	100	100

В контексте настоящей статьи, по-видимому, нет необходимости вдаваться в дискуссию относительно того важного раздела лингвостатистики, привлекающего все большее внимание, который концентрирует в себе вопросы анализа распределений. Целесообразнее было бы пока ограничиться самыми общими рассуждениями и несколькими примерами в связи с проблематикой отбора терминологических учебных минимумов.

Распределение, будучи в общем случае способом и результатом разбиения, группировки, упорядочения совокупности лингвистических явлений, может рассматриваться и как отображение упорядоченности такой совокупности. Если далее за совокупностью видеть системный объект, тогда распределение можно считать отображением, моделью сложного системного лингвистического объекта, которым является текст, группа текстов, подъязык, функциональный стиль, язык с образующими их словарем и грамматикой. Распределение моделирует упорядоченность внутри графемной, морфемной, лексической, терминологической и других подсистем такого системного объекта. Сами такие распределения будут тогда называться лингвистическими распределениями [1].

Лингвистические распределения можно рассматривать исходя из признака, на основании которого строится ряд распределения. В примере табл. 4 признаком является длина текстового словоупотребления или словоформы словаря. Количественное выражение признака –

[1] О возможных классификациях лингвистических распределений см., напр. [3].

Таблица 4

Распределение словоупотреблений текста и словоформ словаря по длине (ЧС английского подъязыка электроники)

Длина в буквах	Доля, %	
	в тексте	в словаре
I	2,46	0,04
2	I7,07	0,59
3	I8,08	2,24
4	I3,07	7,2I
5	9,94	9,57
6	7,42	II,88
7	8,3I	I3,8I
8	7,26	I2,78
9	6,I4	II,7I
I0	3,97	I0,00
II	3,00	6,62
I2	I,64	4,82
I3	0,73	3,3I
I4	0,45	2,09
I5	0,24	I,36
I6-39	0,22	I,97
ИТОГО	I00	I00

это конкретные его значения (I,2,3,4 и т.д. буквы), или по-другому, варианты. Величины, проставленные в колонках "доля", - это численности (частоты) значений, вариант признака. Ряд, составленный из пар вариант признака и частот вариант, называется вариационным рядом, рядом распределения. В табл. 4 представлены два вариационных ряда. Один состоит из вариант от I до 39 и частот этих вариант в тексте; второй ряд состоит из тех же вариант и их частот в словаре.

В табл. 2 приведен фрагмент распределения другого типа, рангового. Если обнаруженные в тексте слова расположить в порядке убывания их частот, получим ЧС в его собственно частотном оформлении, т.е. частотный список. Такой список можно считать вариацион-

Таблица 5

Ранговое распределение частот терминов в английском подъязыке электроники (объем выборки 200 тыс. словоупотреблений)

Ранг i	Частота F	lg i	lg F
I	875	0	2,9420
2	836	0,30I0	2,9222
3	736	0,477I	2,8669
4	665	0,602I	2,8228
5	64I	0,6990	2,8069
I0	437	I,0000	2,6405
20	35I	I,30I0	2,5453
30	296	I,477I	2,47I3
50	2I8	I,6990	2,3385
75	I8I	I,875I	2,2577
I00	I52	2,0000	2,I8I8
250	72	2,3979	I,8573
500	37	2,6990	I,5682
750	22	2,875I	I,3424
I000	I4	3,0000	I,I46I
I2I9-I295	I0	3,0860-3,II23	I,0000
I296-I374	9	3,II26-3,I380	0,9542
I375-I469	8	3,I383-3,I670	0,903I
I470-I588	7	3,I673-3,2009	0,845I
I589-I708	6	3,20II-3,2325	0,7782
I709-I858	5	3,2327-3,2690	0,6990
I859-2076	4	3,2692-3,3I72	0,602I
2077-2383	3	3,3I74-3,377I	0,477I
2384-2867	2	3,3773-3,4574	0,30I0
2868-4I36	I	3,4576-3,6I66	0

ным рядом качественного признака: каждое из слов списка - это значение (варианта), а количество случаев употребления данного слова в тексте - это частота варианты. Если слова списка заменить их порядковыми номерами по убыванию частот, результатом будет порядковый (ранговый) вариационный ряд, а распределение качественного признака преобразуется в ранговое распределение.

О лингвистических ранговых распределениях пишут много; интерес к ним не ослабевает, поскольку в них видят одну из наиболее обобщенных моделей упорядоченности текста и словаря. Табл.5 иллюстрирует ранговое распределение частот терминов по данным ЧС английского подъязыка электроники. Этот ряд дается с интервалами ради экономии места: перечисление всех частот заняло бы в данном случае 213 строк. Основная цифровая информация содержится в начале ряда (самые частые единицы ЧС) и в его конце (количества самых редких единиц с частотами от 1 и примерно до 10). Между этими двумя участками, т.е. в средней зоне распределения, частоты следуют одна за другой в общем монотонно, без скачков. Для терминов такая монотонность частот характерна уже в первом десятке (см. выше, табл.2), что существенно отличает их распределение от распределения всех слов или словоформ ЧС.

В табл.5 две правые колонки содержат величины логарифмов частот и рангов, которые потребуются для того, чтобы представить ранговое распределение в более компактном и обозримом виде, чем это позволяет сделать табличная форма, т.е. в виде графика. Логарифмический масштаб используется прежде всего для придания графическому представлению именно этой компактности. Если последнюю цифру в колонке рангов и первую в колонке частот выразить в миллиметрах, то в натуральном масштабе график распределения по данным табл.5 занял бы площадь приблизительно четыре метра на метр (3,619 $м^2$). При уменьшении этой площади до размера, приемлемого для иллюстрации статьи в настоящем сборнике, никакой тенденции в распределении обнаружить не удалось бы. Поэтому логарифмический масштаб предпочтителен даже по этим соображениям, не считая других.

На рис.1 в графическом виде приведены пять ранговых распределений: распределения однословных терминов в подъязыках электроники, математики и психологии, терминологических словосочетаний и всех словоформ в подъязыке электроники; язык английский.

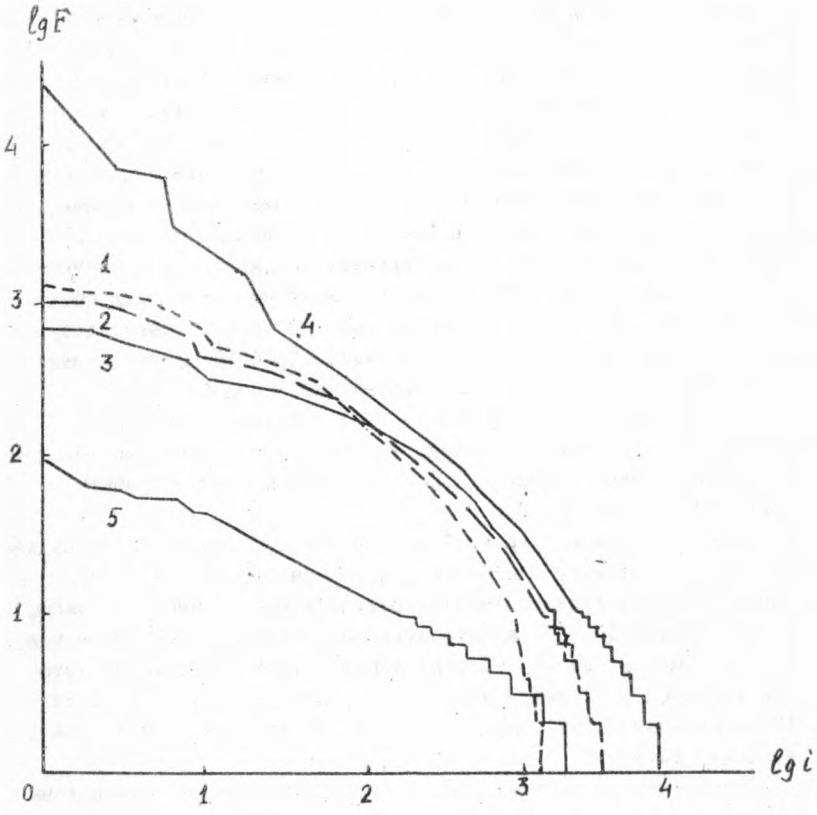

Рис. I. Ранговые распределения однословных терминов в английских подъязыках математики (I), электроники (2) и психологии (3). График 4 представляет словоформы, а график 5 - терминосочетания подъязыка электроники

Здесь нетрудно видеть, что распределения однословных терминов принимают четко выраженную нелинейную форму и резко отличаются от распределений словоформ и терминосочетаний. Последние два можно описывать линейным "ципфовским" графиком (об этом см. ниже). Первые три характерны тем, что, во-первых, частоты "соседствующих" терминов мало отличаются одна от другой, и, во-вторых, на малые частоты приходится меньше разных единиц словаря (ср. длину "ступенек" в нижней зоне распределения). Это придает ранговому распределению терминов выпуклость, нелинейность. С этой точки зрения из трех подъязыков заметно выделяется подъязык математики. При равных размерах выборки (по 200 тыс. словоупотреблений) в текстах по электронике оказалось 4136 разных терминов, по психологии - 3279, а по математике - всего 1662.

Отсюда следует важный вывод: если задать определенный объем терминологического словаря, то его можно извлечь из текстов по математике общей длиной в 2,5 раза меньшей, чем потребуется для минимума по электронике.

Однако, если ориентироваться не на объем словаря, а на заданную частоту, считая ее критерием для отбора терминов в минимум, то здесь ситуация будет интерпретироваться уже по-иному. Скажем, с частотой более 10 в ЧС электроники зафиксировано 1218 терминов, в ЧС математики - 799 и в ЧС психологии - 1374 термина. Но зато на 799 терминов приходится 95% всех терминоупотреблений текста, на 1218 терминов - 89% и на 1374 - 81%. Более подробные сведения приведены в табл.6.

Относительные величины (%) в словарных колонках определены от объемов всего словаря терминов, а в текстовых - от длины текста, измеряемой числом терминоупотреблений. На редкую зону (частоты 1-10) приходится 70% терминов подъязыка электроники, 52% - математики, 58% - психологии; на текстовой оси соответствующие величины равны 10%, 4% и 9%.

Возвращаясь к затронутому выше вопросу о линейности/нелинейности ранговых распределений, следует добавить, что кривизна логарифмического графика вызвана в общем случае "замедленным приростом" словаря. Если в разных текстовых выборках одинаковой длины зафиксированы словари разных объемов, то это объясняется различиями в качестве самих выборок, текстов. Такие различия могут быть языковыми, подъязыковыми, авторскими, тематическими и т.д.

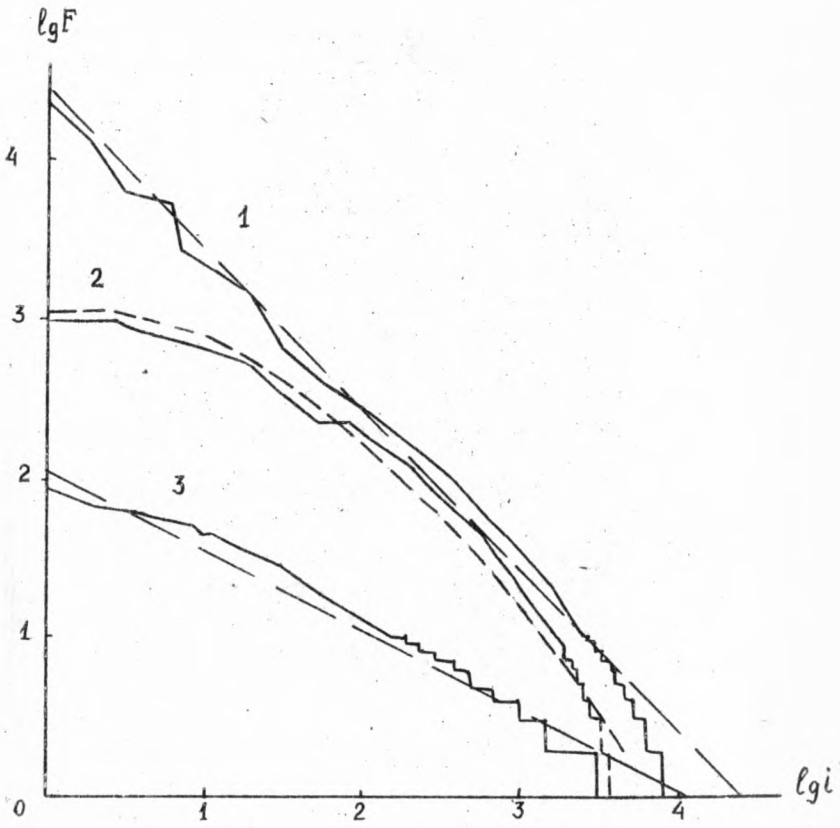

Рис. 2. Эмпирические и теоретические графики ранговых распределений в английском подъязыке электроники: 1 - словоформы, 2 - однословные термины, 3 - терминологические словосочетания

Таблица 6

Распределения словарных и текстовых частот редкоупотребительных терминов в английских подъязыках

Частота	Электроника				Математика				Психология			
	Словарь		текст		словарь		текст		словарь		текст	
	Абс.	%	Абс.	%	Абс.	%	Абс.	%	Абс.	%	Абс.	%
I	I269	30,68	I269	I,59	437	26,29	437	0,68	542	I6,53	542	0,7I
2	484	II,70	968	I,2I	74	4,45	I48	0,23	35I	I0,70	702	0,9I
3	307	7,42	92I	I,I5	58	3,49	I74	0,27	235	7,I7	705	0,92
4	2I8	5,27	872	I,09	69	4,I5	276	0,43	I8I	5,52	724	0,94
5	I50	3,63	750	0,94	54	3,25	270	0,42	I3I	4,00	655	0,85
6	I20	2,90	720	0,90	32	I,96	I92	0,30	II3	3,45	678	0,88
7	II9	2,88	833	I,04	39	2,35	273	0,43	88	2,68	6I6	0,80
8	95	2,30	760	0,95	45	2,7I	360	0,56	96	2,93	768	I,00
9	79	I,9I	7II	0,89	24	I,44	2I6	0,34	66	2,0I	594	0,77
I0	77	I,86	770	0,96	3I	I,87	3I0	0,48	I02	3,II	I020	I,33
ИТОГО:	29I8	70,55	8574	I0,74	863	5I,92	2656	4,I5	I905	58,I0	7004	9,I2
Весь словарь, текст	4I36	I00	7984I	I00	I662	I00	64027	I00	3279	I00	76795	I00

Но если различаются единицы ЧС, то и распределения будут тоже каждый раз другими. На рис. 2 представлены выделенные из рис. I три распределения в одном подъязыке - английском подъязыке электроники. Распределения единиц разных уровней - словоформ, однословных терминов и терминосочетаний естественным образом отличаются одно от других. На рис. 2 приведены также теоретические, сглаживающие (аппроксимирующие) графики для этих распределений - графики линейные для словоформ и терминосочетаний и нелинейные для однословных терминов. Процедуры построения аппроксимирующих ранговых распределений подробно описаны в [I] , а обсуждение проблемы линейности и нелинейности ранговых распределений- в [2].

Несмотря на то, что представленные и расмотренные выше статистические данные о терминах свидетельствуют о не очень высокой частоте их употребления в текстах не слишком малого объема (200 тыс. словоупотреблений в каждом случае), это еще не означает недостаточности статистического метода в отборе учебного материала. Это означает лишь, насколько серьезна и ответственна такая работа, когда она выполняется с опорой на статистику. Сильных сторон у критерия частности больше, чем слабых, и он пока остается самым объективным средством отделить главное от второстепенного. Его следует совершенствовать, но не браковать только потому, что не хватает сил, времени и терпения для последовательного и строгого его применения.

ЛИТЕРАТУРА

I. Алексеев П.М. Методика квантитативной типологии текста: Учеб. пособие. Л., 1983.

2. Алексеев П.М. О нелинейных формулировках закона Ципфа// Вопросы кибернетики. Вып. 41: Статистика речи и автоматический анализ текста Науч. совет по комплекс.проблеме "Кибернетика" АН СССР, М.; Л., 1983.

3. Алексеев П.М. Об уровнях лингвистического анализа и о знаковости текста // Инженерная лингвистика и романо-германское языкознание: Сб. науч. тр. Л., 1985.

4. Лексико-терминологические материалы для чтения текстов по биологии на английском языке: Частотный минимум / Сост. Л.Г.Берзиньш. Л., 1983.

5. Лексико-терминологические материалы для чтения текстов по психологии на английском языке: Частотный минимум / Сост. Г.В.Басовская, А.В.Вербицкий. Л., 1980.

6. Учебные терминологические материалы для чтения текстов по математике на английском языке: Частотный минимум/ Сост. Л.М.Сутягина. Л., 1982.

7. Частотный англо-русский словарь-минимум по квантовым генераторам / Сост. Н.С.Манасян. М., 1983.

8. Частотный англо-русский словарь-минимум по электронике / Сост. П.М.Алексеев. М., 1971.

Л.М.Сутягина
Уральский университет

ФОРМИРОВАНИЕ ВЫБОРОЧНОГО КОРПУСА ТЕКСТОВ ПРИ СОСТАВЛЕНИИ ЧАСТОТНОГО СЛОВАРЯ

(возможный алгоритм построения выборки)

В статье "Формирование выборочного корпуса текстов при составлении частотного словаря (качественная сторона формирования выборки)" [I], где обсуждались некоторые проблемы, связанные с качественной стороной формирования выборки для составления частотного словаря (ЧС), были сделаны следующие основные выводы:

I. Один из самых простых способов приблизить структуру выборки к структуре генеральной совокупности - это ограничить исследование как можно более узким подъязыком, т.е. выборка должна быть направленной;

2. Трудоемкую процедуру отбора текстов для составления ЧС с помощью жеребьевки или других аналогичных методов можно заменить направленным отбором ввиду практически случайного распределения лексических единиц (ЛЕ) в самих текстах;

3. К формированию узконаправленной выборки для составления ЧС целесообразно приступать лишь после того, как выделено "ядро" журналов и авторов в данной области знаний и определен по специальным таблицам полупериод жизни публикаций в этой области.

Достаточно ли выполнения этих условий для того, чтобы приступить непосредственно к составлению ЧС? Очевидно, нет. Более или менее ясно, какие тексты нужно включить в выборку, но не ясно, сколько нужно обследовать их для получения надежного