УДК 811.161.1'04'371

МЕТОД ФОРМАЛЬНОГО ВЫДЕЛЕНИЯ ТЕМАТИЧЕСКИ НЕЙТРАЛЬНОЙ ЛЕКСИКИ (НА ПРИМЕРЕ СТАРОСЛАВЯНСКИХ ТЕКСТОВ)

А. А. Кретов

Воронежский государственный университет

В статье введены понятия *тематически маркированной* и *тематически нейтральной лексики*; предложен метод системного взвешивания слов по двум функциональным параметрам: прямому (частотному — *Q-параметр*) и косвенному (длина слова — *F-параметр*). Первый параметр характеризует функционирование слова в данном тексте, второй — его функционирование на продолжительном отрезке времени — настолько продолжительном, чтобы функционирование успело повлиять на длину слова; введен Индекс тематической маркированности слова (ИнТеМ), вычисляемый по формуле **ИнТеМ = Q-вес — F-вес**, где Q-вес — вес слова по частоте, а F-вес — вес слова по длине. Установлено, что в словаре-источнике более 98 % слов с отрицательным значением ИнТеМа относятся к тематически нейтральной лексике.

1. ВВЕДЕНИЕ

Лексика любого текста может быть разделена на две части: одна из них связана с темой текста (назовем ее *тематически маркированной лексикой*), другая — никак не указывает на тему текста и может встретиться в тексте любой другой тематики (назовем ее *тематически нейтральной лексикой*).

Вы деление обеих групп лексики — шаг на пути к определению содержательной (тематической) отнесенности текста.

Цель данной статьи — предложить метод формального, а следовательно — автоматизируемого, выделения тематически нейтральной лексики.

Материалом послужит «Старославянский словарь (по рукописям X—XI веков)» [1]: около 10 000 слов Авторы: Э. Благова. Р. М. Цейтлин, С. Геродес, Л. Панцерова, М. Бауэрова. Рецензенты: акад. Н. И. Толстой, докт. филол. наук А. Е. Супрун, докт. филол. наук Г. А. Хабургаев. Редакторы: Р. М. Цейтлин, Р. Вечерка, Э. Благова. — М.: Рус. яз., 1994. — 842 с.

Достоинством этого словаря является тот факт, что он описывает строго ограниченно множество текстов и содержат информацию о частоте употребления каждого из слов в тексте. К сожалению, для употребительных слов эта информация дается округленно (>100, >500, <500, >1000 и т.п.), но с этим недостатком приходится мириться — тем более, что он касается небольшой группы слов.

2. МЕТОД ФОРМАЛЬНОГО ВЫДЕЛЕНИЯ ТЕМАТИЧЕСКИ НЕЙТРАЛЬНОЙ ЛЕКСИКИ

В Научно-методическом центре компьютерной лингвистики факультета РГФ ВГУ создана электронная версия этого словаря (исполнитель — лаборант А. В. Кашкина). Поскольку служебные и дискурсивные слова являются своего рода «цементом» текста и не определяют его содержательного наполнения, в электронную версию словаря были включены только полнозначные слова: глаголы, существительные и прилагательные, а также причастия, употребляемые в функции существительного или прилагательного. В результате такого сокращения число слов уменьшилось почти на 2500.

Суть предлагаемого метода состоит в следующем. Традиционным способом выявления тематически маркированной лексики является частота словоформ (или их множеств, относящихся к одному слову-лемме). При этом предполагается, что чем чаще употребляется слово в тексте, тем важнее оно для содержания данного текста. Это справедливо только отчасти: так, например, самым употребительным словом русского языка является союз И. Очевидно, что специфики текста он не отражает. Исключив из рассмотрения служебные и дискурсивные слова, мы тем самым уже освободили частотную верхушку словаря от некоторого информационного шума.

Но есть и еще одно обстоятельство. Еще в первой половине XX века американский лингвист Дж. К. Ципф установил зависимость, су-

ществующую между частотой слова (словоформы) и его длиной: чем чаще употребляется слово, тем оно короче и наоборот. Можно сказать, что от частого употребления слова снашиваются — уменьшаются в размере (длине). Но для это требуется, чтобы слова устойчиво и продолжительно были частотными. Следовательно, если короткое слово в тексте будет частотным, это будет характеризовать его как короткое слово, а вот если длинное слово в тексте будет обладать частотой необычной для слов такой длины, то оно будет отражать специфику данного текста. Следовательно, для определения ключевых слов в данном тексте нам мало информации об их частоте или длине. Нам надо соотнести оба типа информации, а для этого необходимо найти способ сделать информацию о длине и частоте сопоставимой. Сделать это немногим проще, чем научиться определять, что длинный и зеленый крокодил длиннее, чем зеленее. И все же попробуем.

Для начала надо получить распределение слов в старославянских текстах по длине и по частоте.

(Поскольку словарь-источник дает информацию о частоте употребления всего множества форм одного слова (словоформ), представленных в словарной статье заглавной словарной формой (леммой), в данном случае нам придется пойти на некоторое огрубление текстовой реальности, приняв среднюю длину всех словоформ данного слова равной длине леммы).

Поскольку количество букв в старославянских тестах почти идеально соответствует числу звуков, длина старославянских слов может измеряться в буквах.

Поскольку сведения о частоте даются для всей словарной статьи, а в словарную статью могут входить и причастия (наряду с глаголом) и субстантивированные прилагательные (наряду с полноценными прилагательными), то количество словарных статей и, соответственно, лемм в данной таблице оказывается меньше, чем количество слов.

В таблице 1 особого комментария требует столбец «F-вес» — функциональный вес. В столбце «Накопл.» суммируется количество слов данного интервала с количеством слов всех предшествующих интервалов, его данные используются для вычисления параметра «F-вес» по формуле, предложенной В. Т. Титовым [Титов 2004:15]:

Табл. 1

Статистическая структура распределения старославянских слов по длине

Длина	Слов	Накопл.	F-вес
2	1	1	0,9999
3	28	29	0,9962
4	339	368	0,9514
5	524	892	0,8822
6	873	**1765**	0,7668
7	1094	2859	0,6223
8	1316	4175	0,4484
9	1244	5419	0,2841
10	909	6328	0,1640
11	584	6912	0,0868
12	304	7216	0,0466
13	167	7383	0,0246
14	80	7463	0,0140
15	52	7515	0,0071
16	25	7540	0,0038
17	13	7553	0,0021
18	8	7561	0,0011
19	5	7566	0,0004
20	1	7567	0,0003
22	2	7569	0,0000

$$Pr_i = \frac{\sum_{1}^{\max} r - \sum_{1}^{i} r}{\sum_{1}^{\max} r}, \qquad (1)$$

где Σr — сумма единиц всех рангов(7569), R_{1-i} — сумма единиц от первого до данного ранга. В данном случае речь идет о рангах частот, а не о рангах слов.

Поскольку самые короткие слова в среднем — самые употребительные, расположение их в порядке возрастания длины соответствует расположению их в порядке убывания частоты. А параметр F-вес необходим для того, чтобы обеспечить сопоставимость слов по длине и частоте. В обоих случаях мы определяем системный вес слов, исходя их того, что самые употребительные (а следовательно — и самые короткие) слова — самые важные. Логика F-параметра такова: чем меньше в тексте (словаре) слов с таким или более высоким значением параметра, тем больше системный вес (значимость) данного слова и наоборот.

На рис.1 представлена зависимость числа слов от их длины.

На рис. 2 представлена динамика F-параметра.

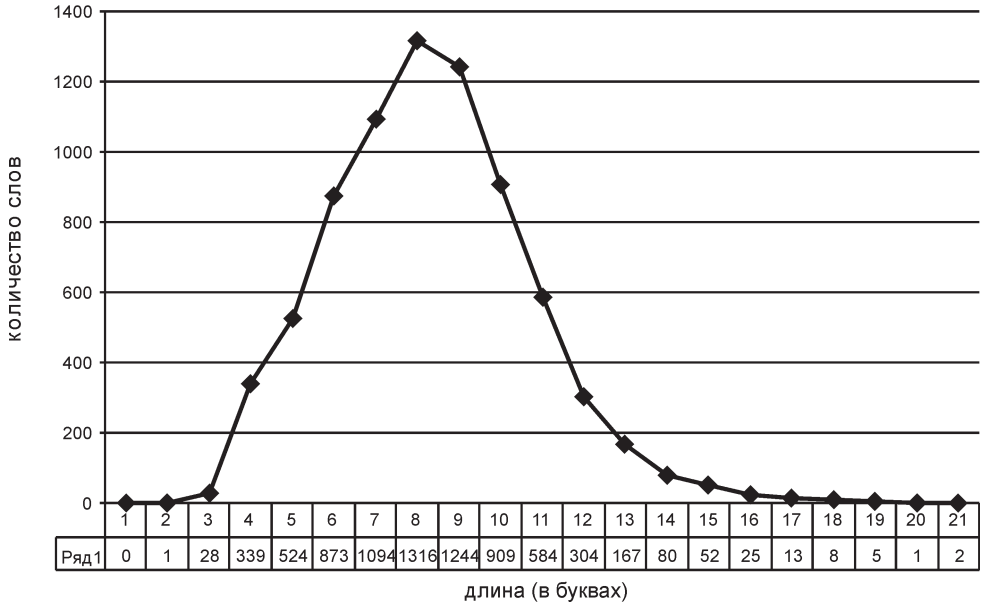

Рис. 1. Зависимость количества слов в старославянских текстах от их длины

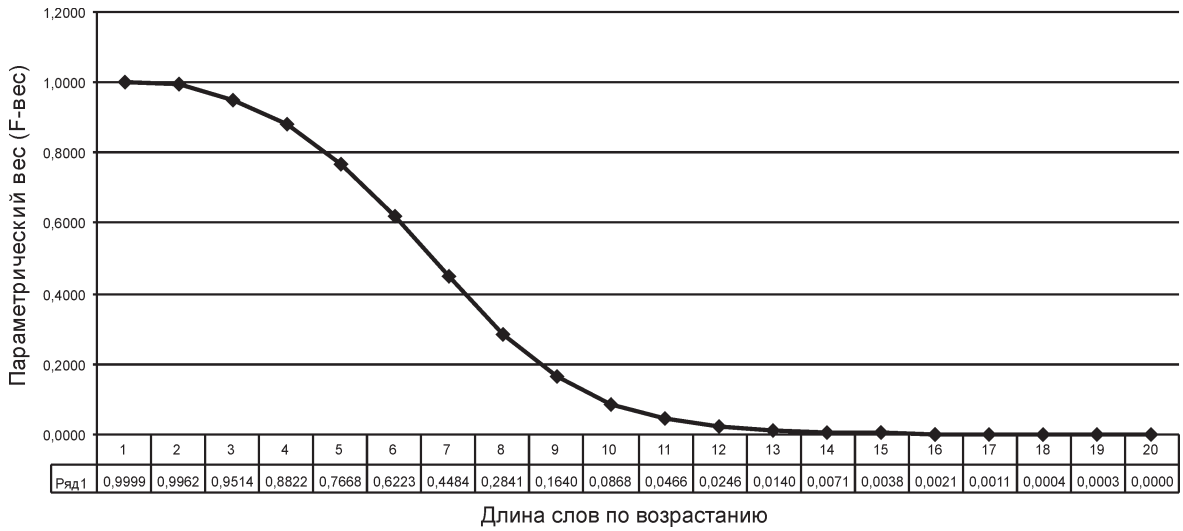

Рис. 2. Зависимость F-параметра от длины слов

Проверим, действительно ли существует зависимость между длиной слов и их частотой в тексте. Эта закономерность представлена на рис. 3.

Как видим, начиная с длины в четыре буквы, т.е. с нормальной минимальной длины полнозначного слова в старославянском языке (три буквы — корень, одна буква — окончание), средняя частота слов данной длины последовательно убывает. Длина же от одной до трех букв — интервал служебных слов. Колебания начинаются там, где кончаются статистически надежные данные.

Рассмотрим статистическую структуру старославянской лексики по частоте появления в текстах. Соответствующие данные представлены в табл. 2.

Для удобства обработки относительные частоты были преобразованы в абсолютные посредством прибавления или вычитания единицы. Например, >200 → 201, а <200 → 199.

Значения столбца Q-вес были вычислены аналогично значениям параметра F-вес, чем мы обеспечили сопоставимость двух параметров.

Динамика Q-параметра представлена на рис. 4.

Проверим, как изменяется средняя длина слов по мере убывания их частоты. Соответствующая зависимость представлена на рис. 5.

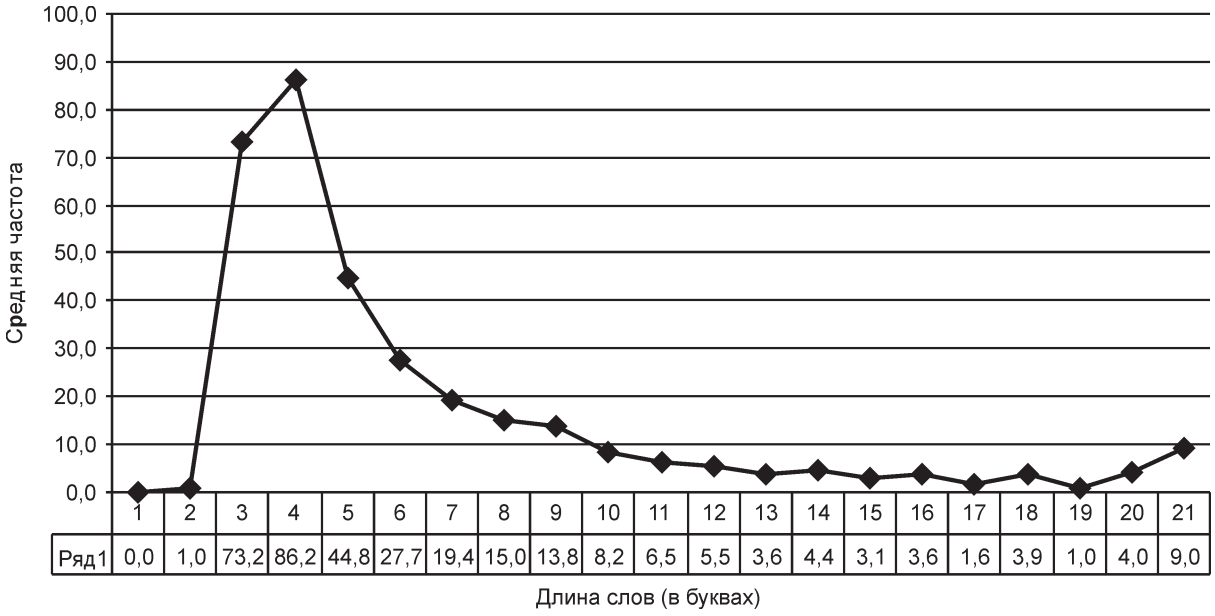

Рис. 3. Зависимость частоты слов в старославянских текстах от их длины

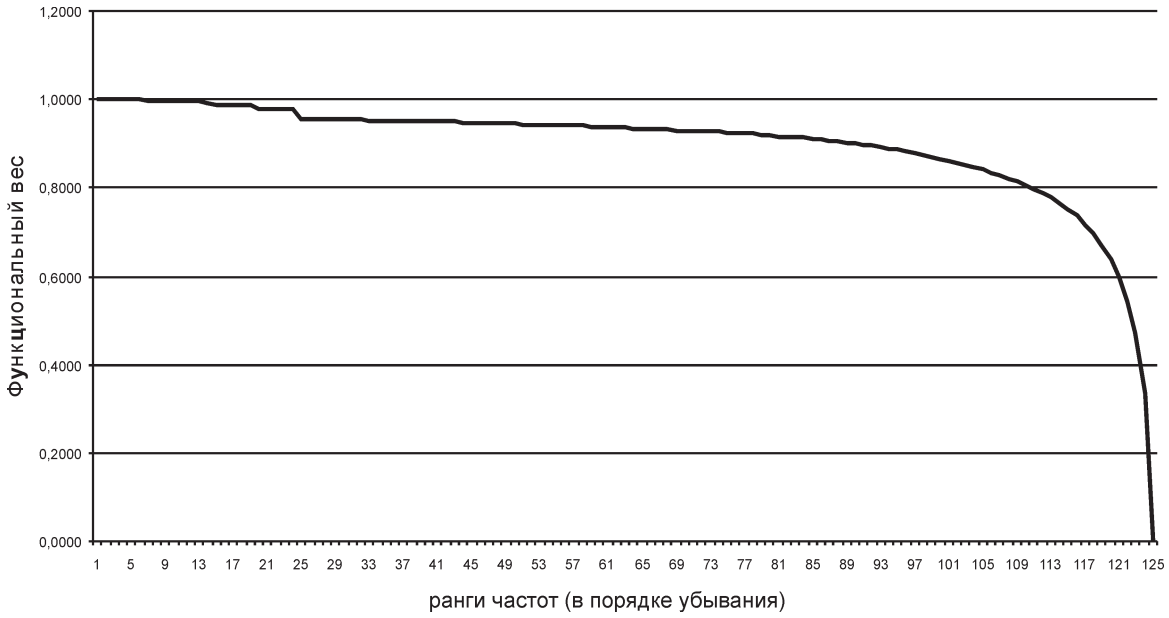

Рис. 4. Распределение старославянских слов по частоте (Q-вес)

Как свидетельствует линейный тренд, по мере убывания частот слов их средняя длина возрастает. Таким образом, данные старославянских текстов подтвердили существование зависимости, обнаруженной Дж.К.Ципфом.

Теперь у нас есть все необходимое для членения старославянской лексики на тематически маркированную и тематически нейтральную.

Для деления лексики на тематически маркированную и тематически нейтральную необходимо из значения Q-веса каждого слова вычесть значение его F-веса. Полученные значения назовем *ИНдикатором ТЕматической Маркированности* (ИнТеМ): положительные значения ИнТеМа будут характеризовать тематически маркированную лексику, отрицательные значения — тематически нейтральную лексику. При этом значения ИнТеМа будут варьироваться в интервале от +1 до -1. Таким образом, значения ИнТеМа находятся по формуле:

(2) ИнТеМ = Q-вес — F-вес.

Для содержательной проверки работы ИнТеМа выделим из таблиц 1 и 2 близкие по раз-

Таблица 2

Распределение старославянских слов по частоте появления в текстах

Частота	Слов	Накопл.	Q-вес	ДлИнт	СрДлина
7001	1	1	0,9999	4	4,0000
3301	1	2	0,9997	9	9,0000
2101	1	3	0,9996	4	4,0000
1701	2	5	0,9993	13	6,5000
1601	1	6	0,9992	7	7,0000
1401	1	7	0,9991	6	6,0000
1001	10	17	0,9978	54	5,4000
901	1	18	0,9976	6	6,0000
801	3	21	0,9972	18	6,0000
701	2	23	0,9970	11	5,5000
601	2	25	0,9967	9	4,5000
501	15	40	0,9947	80	5,3333
499	2	42	0,9945	8	4,0000
401	21	63	0,9917	80	3,8095
351	1	64	0,9915	122	122,0000
301	21	85	0,9888	5	5,0000
299	3	88	0,9884	128	6,0952
251	2	90	0,9881	16	5,3333
201	53	143	0,9811	10	5,0000
199	2	145	0,9808	343	6,4717
184	1	146	0,9807	13	6,5000
151	5	151	0,9800	4	4,0000
144	1	152	0,9799	38	7,6000
101	162	314	0,9585	9	9,0000
100	2	316	0,9582	1090	6,7284
99	3	319	0,9578	11	5,5000
98	2	321	0,9576	33	11,0000
97	6	327	0,9568	10	5,0000
96	2	329	0,9565	47	7,8333
95	3	332	0,9561	17	8,5000
94	2	334	0,9559	26	8,6667
93	3	337	0,9555	13	6,5000
92	2	339	0,9552	22	7,3333
91	2	341	0,9549	16	8,0000
90	3	344	0,9545	10	5,0000
89	4	348	0,9540	16	5,3333
88	3	351	0,9536	29	7,2500
87	6	357	0,9528	11	3,6667
86	4	361	0,9523	41	6,8333
85	2	363	0,9520	25	6,2500
84	3	366	0,9516	16	8,0000
83	1	367	0,9515	22	7,3333
82	3	370	0,9511	8	8,0000
81	3	373	0,9507	26	8,6667
80	4	377	0,9502	25	8,3333
79	3	380	0,9498	28	7,0000
78	4	384	0,9493	20	6,6667
77	6	390	0,9485	31	7,7500
76	10	400	0,9471	45	7,5000

Частота	Слов	Накопл.	Q-вес	ДлИнт	СрДлина
75	4	404	0,9466	67	6,7000
74	3	407	0,9462	33	8,2500
73	2	409	0,9460	22	7,3333
72	6	415	0,9452	11	5,5000
71	3	418	0,9448	43	7,1667
70	5	423	0,9441	31	10,3333
69	6	429	0,9433	35	7,0000
68	3	432	0,9429	46	7,6667
67	8	440	0,9419	16	5,3333
66	8	448	0,9408	59	7,3750
65	6	454	0,9400	56	7,0000
64	8	462	0,9390	45	7,5000
63	2	464	0,9387	57	7,1250
62	8	472	0,9376	13	6,5000
61	5	477	0,9370	62	7,7500
60	8	485	0,9359	35	7,0000
59	7	492	0,9350	54	6,7500
58	9	501	0,9338	39	5,5714
57	7	508	0,9329	62	6,8889
56	5	513	0,9322	49	7,0000
55	4	517	0,9317	33	6,6000
54	7	524	0,9308	29	7,2500
53	8	532	0,9297	58	8,2857
52	8	540	0,9286	55	6,8750
51	7	547	0,9277	59	7,3750
50	10	557	0,9264	48	6,8571
49	10	567	0,9251	67	6,7000
48	7	574	0,9242	75	7,5000
47	8	582	0,9231	51	7,2857
46	14	596	0,9212	62	7,7500
45	15	611	0,9193	104	7,4286
44	9	620	0,9181	90	6,0000
43	13	633	0,9164	63	7,0000
42	10	643	0,9150	89	6,8462
41	6	649	0,9142	67	6,7000
40	14	663	0,9124	42	7,0000
39	27	690	0,9088	111	7,9286
38	18	708	0,9064	206	7,6296
37	10	718	0,9051	134	7,4444
36	28	746	0,9014	63	6,3000
35	16	762	0,8993	212	7,5714
34	18	780	0,8969	119	7,4375
33	18	798	0,8946	128	7,1111
32	23	821	0,8915	129	7,1667
31	24	845	0,8883	169	7,3478
30	27	872	0,8848	188	7,8333
29	39	911	0,8796	199	7,3704
28	28	939	0,8759	290	7,4359
27	21	960	0,8732	218	7,7857

Окончание табл. 2

Распределение старославянских слов по частоте появления в текстах

Частота	Слов	Накопл.	Q-вес	ДлИнт	СрДлина
26	32	992	0,8689	167	7,9524
25	38	1030	0,8639	227	7,0938
24	30	1060	0,8599	246	6,4737
23	46	1106	0,8539	239	7,9667
22	40	1146	0,8486	360	7,8261
21	47	1193	0,8424	293	7,3250
20	44	1237	0,8365	347	7,3830
19	53	1290	0,8295	340	7,7273
18	44	1334	0,8237	379	7,1509
17	65	1399	0,8151	346	7,8636
16	57	1456	0,8076	511	7,8615
15	56	1512	0,8002	447	7,8421
14	71	1583	0,7908	429	7,6607
13	71	1654	0,7814	511	7,1972
12	90	1744	0,7696	555	7,8169
11	116	1860	0,7542	709	7,8778
10	115	1975	0,7390	913	7,8707
9	156	2131	0,7184	911	7,9217
8	157	2288	0,6977	1214	7,7821
7	181	2469	0,6738	1266	8,0637
6	269	2738	0,6382	1461	8,0718
5	280	3018	0,6012	2140	7,9554
4	423	3441	0,5453	2324	8,3000
3	566	4007	0,4705	3548	8,3877
2	1014	5021	0,3365	4702	8,3074
1	2547	7568	0,0000	8752	8,6312

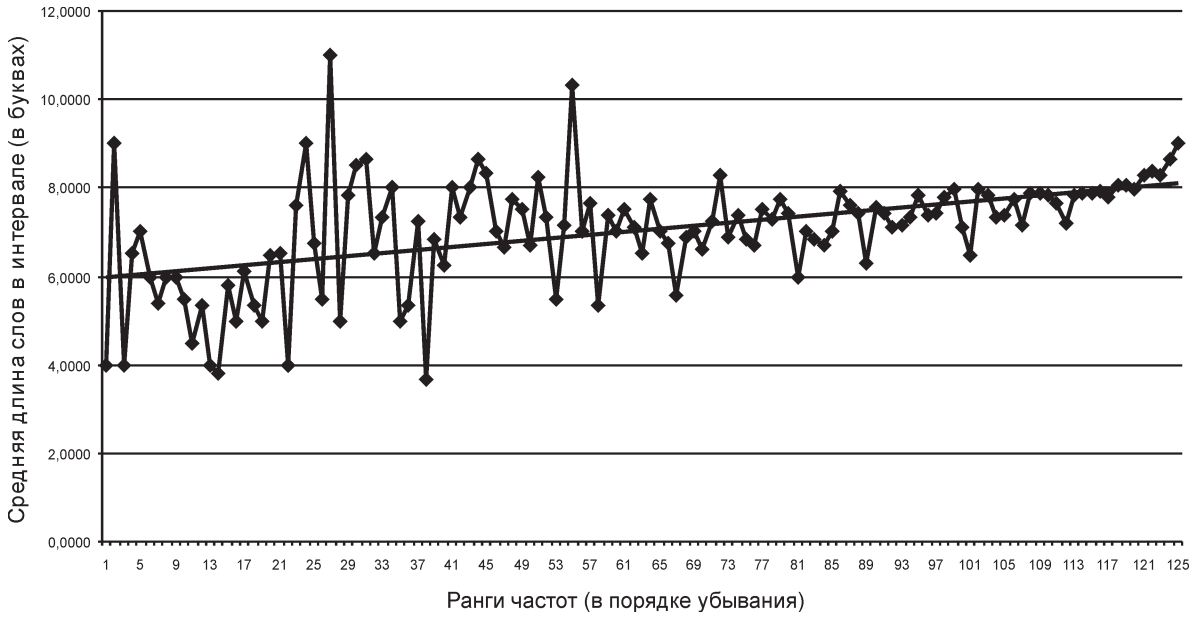

Рис. 5. Зависимость длины старославянских слов от их частоты в текстах

меру множества ядерной (т.е. обладающей максимальными параметрическими весами) лексики и наложим их друг на друга. Результаты такого наложения представлены в табл. 3.

В множестве, выделенном по значениям F-параметра (длина слов от 2 до 6 букв) оказалось 1763 слова, а во множестве, выделенном по значению Q-параметра (частота от 7001 до 12 появлений в тексте) — 1743.

При наложении этих множеств совпало 678 слов, т.е. 38,5% для F-ядра и 38,9% для Q-ядра.

По значению ИнТеМа члены суммарного множества распределились таким образом, как это представлено в рис. 6.

Таблица 3

Пересечение лексических ядер старославянской лексики с максимальными значениями F- и Q-весов

Слов	F+Q	F	Q
	3506	1763	1743
Общих	1356	**678**	**678**
Разных		**1085**	**1065**
% пересечения		38,45718	38,89845

Как видим, значения ИнТеМа делят всю лексику на две части: положительную и отрицательную.

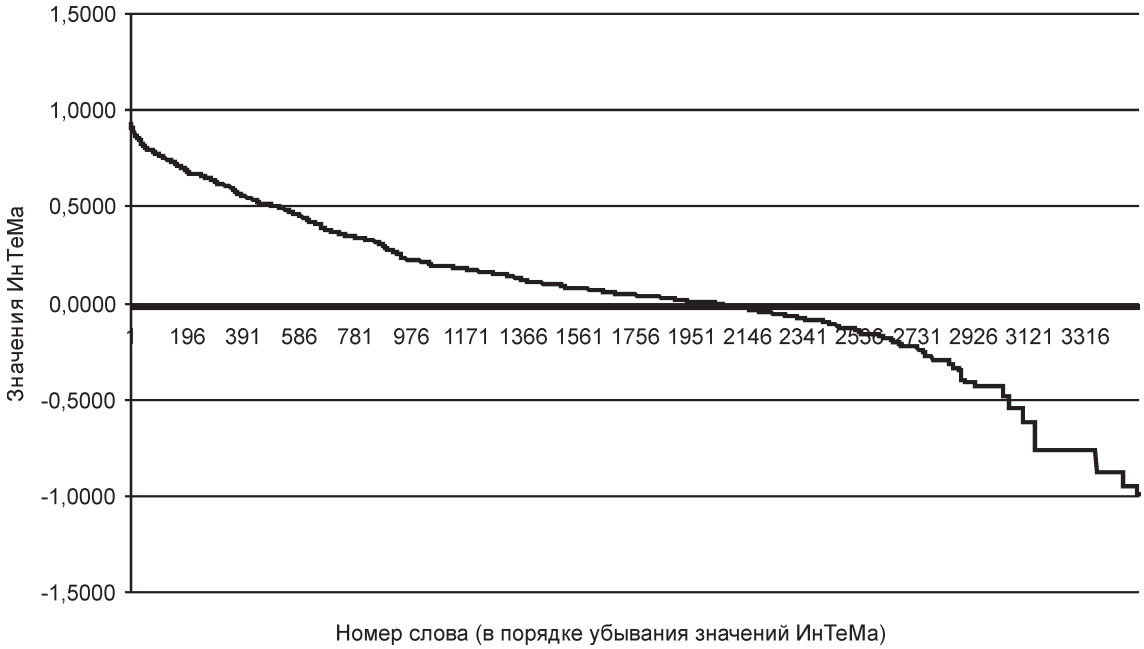

Рис. 6. Динамика ИнТеМа в подмножестве ядерной лексики старославянского языка

В положительной части мы встречаем слова *въздрадоватися, свЪдЪтельство, благословити, христосовъ, приближитися, вознеавидЪти, цЪсарьствие, благословествити, прикоснутися, кровьточай, разгнЪватися, чловеколюбець, проповЪдати, въскьснути, прЪподобьнъ, милосрьдовати, правьдьникъ, благословленъ, осквърнитися, благословление, архиепископъ, въздвигнути, безаконение, съвършитися, възвеселити, просвЪтитися, възглаголати, вьседрьжитель, свЪтильникъ, постыдЪтися, въглашение,* представленные ниже (табл. 4), и др.

Данные о наложении множеств и их членении ИнТеМом представлены ниже.

L+Q	3506
ИнТеМ+	2032
ИнТеМ-	1474
Маркированных	27
% ошибки	1,83

В отрицательной части тематически маркированными оказывается 27 слов, что составляет 1,83 % от общего числа (1474) ядерных слов с отрицательным значением ИнТеМа. Нетрудно заметить, что большую часть этих слов составляют заимствования.

Полагаем, что ошибку менее 2 % в определении тематически нейтральной лексики можно признать приемлемой, а предложенный метод определения тематически нейтральной лексики — эффективным.

Если слова с положительным значением ИнТеМа относятся к тематически маркированной лексике (в случае старославянского языка — церковно-христианской), а слова с отрицательным значением ИнТеМа относятся к тематически нейтральной лексике, то слова с нулевым (или максимально близким к нулевому) значением ИнТеМа должны представлять собой своего рода позицию нейтрализации, в которой снимается оппозиция маркированность||нейтральность.

С этой точки зрения полученные нами результаты весьма красноречивы. Максимально близкие к нулю значения ИнТеМа имеют в старославянских текстах слова РЫБА — 0,001 и ВИНО — 0,0004.

Само по себе слово *рыба* нейтрально, но в Новом Завете Христос двумя рыбами (и пятью хлебами) накормил около пяти тысяч человек, не считая женщин и детей:

[13] Но Он сказал им: вы дайте им есть. Они сказали: у нас нет более пяти хлебов и двух рыб; разве нам пойти купить пищи для всех сих людей?

[14] Ибо их было около пяти тысяч человек. Но Он сказал ученикам Своим: рассадите их рядами по пятидесяти.

[15] И сделали так, и рассадили всех.

[16] Он же, взяв пять хлебов и две ***рыбы*** *и воззрев на небо, благословил их, преломил и дал ученикам, чтобы раздать народу.*

Таблица 4

Слово	Часть речи	Значение	Q	F	Q-вес	F-вес	Q-F-вес
въздрадоватися	глсов.	1) возрадоваться, обрадоваться	97	14	0,957	0,014	0,943
свЪдЪтельство	сущ.	1) свидетельство	69	13	0,944	0,025	0,919
благословити	гл.	1) благословить/благословлять	101	12	0,959	0,047	0,912
христосовъ	прил.	1) Христов, Христа	101	12	0,959	0,047	0,912
приближитися	глсов.	1) приблизиться	99	12	0,958	0,047	0,911
възненавидЪти	глсов.	1) возненавидеть	54	13	0,931	0,025	0,906
цЪсарьствие	сущ.	1) царствование, царская власть	301	11	0,989	0,087	0,902
благословествити	гл.	1) благословить	36	16	0,902	0,004	0,898
прикоснутися	глсов.	1) прикоснуться	67	12	0,942	0,047	0,895
кровьточай	прич.	1) страдающий кровотечением	201	11	0,981	0,087	0,894
разгнЪватися	глсов.	1) рассердиться, разгневаться	52	12	0,929	0,047	0,882
человеколюбець	сущ.	1) тот, кто любит людей, отличается человеколюбием	34	13	0,897	0,025	0,873
проповЪдати	глнесов.	1) проповедовать, провозглашать, возвещать	101	11	0,959	0,087	0,872
въскрьснути	глсов.	1) встать	101	11	0,959	0,087	0,872
прЪподобнъ	прил.	1) преподобный	99	11	0,958	0,087	0,871
милосрьдовати	глнесов.	1) проявлять милосердие, сострадание, милость	32	13	0,892	0,025	0,867
правьдьникъ	сущ.	1) праведник	84	11	0,952	0,087	0,865
благословленъ	прил.	1) благословенный	31	13	0,889	0,025	0,864
оскврьнитися	глвозвр.	1) осквернитьcя	39	12	0,909	0,047	0,863
благословление	сущ.	1) благословение	28	14	0,876	0,014	0,862
архиепискупъ	сущ.	1) архиепископ	38	12	0,907	0,047	0,860
въздвигнути	глсов.	1) поднять	75	11	0,947	0,087	0,860
безаконение	сущ.	1) беззаконие, проступок	71	11	0,945	0,087	0,858
съврьшитися	глвозвр.	1) произойти, совершиться, завершиться, закончиться	71	11	0,945	0,087	0,858
възвеселити	глсов.	1) обрадовать, порадовать	69	11	0,944	0,087	0,857
просвЪтитися	глвозвр.	1) засиять	36	12	0,902	0,047	0,855
възглаголати	глсов.	1) начать говорить, заговорить	35	12	0,900	0,047	0,853
вьседрьжитель	сущ.	1) вседержатель	28	13	0,876	0,025	0,852
свЪтильникъ	сущ.	1) светильник, свеча, лампа	61	11	0,937	0,087	0,850
постыдЪтися	глсов.	1) устыдиться, быть пристыженным	60	11	0,936	0,087	0,849
възглашение	сущ.	1) литург. возглас	58	11	0,934	0,087	0,847

[17] *И ели, и насытились все; и оставшихся у них кусков набрано двенадцать коробов.*

(Лука, Гл. 9) [3, С.75].

Кроме того, для древних христиан изображение рыбы было символом Иисуса Христа. «В Китае, Индии и некоторых других ареалах Р.[ыба] символизирует новое рождение... Не случайна в этом отношении «рыбная» метафорика Иисуса Христа, прослеживаемая на формальном уровне (греческое слово ιχθύς, «рыба», расшифровывалось как аббревиатура греческой формулы «Иисус Христос, божий сын, спаситель»), так и по существу [ср. Р. Как символ веры, чистоты, девы Марии, а также крещения, причастия (где она заменяется хлебом и вином; в том же ряду стоит евангельский мотив насыщения Р. и хлебами]; Иисус Христос иногда называется в ранней христианской литературе «Рыбой» (ср. этот образ в катакомбном искусстве), а христиане «рыбаками» [4].

Что касается *вина*, то это то вино, в которое Христос превратил воду и которое назвал кровью своей:

[26] *И когда они ели, Иисус взял хлеб и, благословив, преломил и, раздавая ученикам, сказал: примите, ядите: сие есть Тело Мое.*

[27] *И, взяв чашу и благодарив, подал им и сказал: пейте из нее все,*

Таблица 5

Слово	Часть речи	Значение	Q	F	Q-вес	F-вес	Q-F
адъ	сущ.	1) ад, пекло	76	3	0,947	0,996	–0,049
мvра	сущ.	1) миро	51	4	0,928	0,951	–0,024
адовъ	прил.	1) адов, адский	22	5	0,849	0,882	–0,033
геона	сущ.	1) ад, пекло, геенна	22	5	0,849	0,882	–0,033
мощи	сущ.	1) мощи	14	4	0,791	0,951	–0,160
хризма	сущ.	1) миро	11	6	0,755	0,767	–0,012
манъна	сущНескл.	1) манна (небесная)	11	6	0,755	0,767	–0,012
оплатъ	сущ.	1) Святые дары	10	6	0,740	0,767	–0,027
мьша	сущ.	1) месса (церковная служба, литургия)	8	4	0,699	0,951	–0,253
масть	сущ.	1) миро 2) масло, жир	6	5	0,639	0,882	–0,243
идолъ	сущ.	1) идол	5	5	0,602	0,882	–0,280
адьскъ	прил.	1) адский	4	6	0,547	0,767	–0,220
богыни	сущ.	1) богиня	4	6	0,547	0,767	–0,220
хитонъ	сущ.	1) хитон (нижняя одежда)	4	6	0,547	0,767	–0,220
кадило	сущ.	1) ладан, фимиам	3	6	0,472	0,767	–0,295
облашь	прилСубст.	1) мирянин	3	6	0,472	0,767	–0,295
чьтьць	сущ.	1) чтец (в церкви)	3	6	0,472	0,767	–0,295
раискъ	прил.	1) райский	3	6	0,472	0,767	–0,295
амбонъ	сущ.	1) амвон (возвышение в церкви для проповедующего)	2	6	0,338	0,767	–0,428
попъ	сущ.	1) священник	2	4	0,338	0,951	–0,613
родъ	сущ.	1) ад <>rod" ogn'nyi геенна огненная, ад	2	4	0,338	0,951	–0,613
диво	сущ.	1) чудо	1	4	0,003	0,951	–0,949
елхи	сущ.	1) молитва	1	4	0,003	0,951	–0,949
фелонь	сущ.	1) фелонь (верхняя одежда)	1	6	0,003	0,767	–0,764
лития	сущ.	1) лития (молебен за упокой души)	1	5	0,003	0,882	–0,879
мънихъ	сущ.	1) монах	1	6	0,003	0,767	–0,764
урарь	сущ.	1) орарь (часть дьяконского облачения)	1	5	0,003	0,882	–0,879

[28] *ибо сие есть Кровь Моя Нового Завета, за многих изливаемая во оставление грехов.*
[29] *Сказываю же вам, что отныне не буду пить от плода сего виноградного до того дня, когда буду пить с вами новое* ***вино*** *в Царстве Отца Моего.*

(Матф. Гл. 26) [3, С. 33].

«Продолжение (с сохранением тех же словесных формул) и переосмысление ближневосточной традиции, сказавшейся также в гностических книгах, обнаруживается в христианской мифологии (в словах Иисуса Христа, взявшего чашу В.[ина] и сказавшего: «сие есть кровь моя», Матф. 26, 28)». [5, С.236].

Таким образом, в словах *рыба* и *вино*, действительно, происходит нейтрализация оппозиции «тематически маркированная||немаркированная лексика», поскольку, с одной стороны *рыба* и *вино* не представляют собой ничего специфически церковного или христианского, а с другой стороны, в новозаветных текстах эти слова пронизаны несомненной христианской символикой.

3. ЗАКЛЮЧЕНИЕ

Подведем итоги. В статье введены понятия *тематически маркированной* и *тематически нейтральной лексики*; предложен метод системного взвешивания слов по двум функциональным параметрам: прямому (частотному — *Q-параметр*) и косвенному (длина слова — *F-параметр*). Первый параметр характеризует функционирование слова «здесь и сейчас» (в данном тексте) второй — его функционирование на продолжительном отрезке времени — настолько продолжительном, чтобы функционирование успело найти свое отражение в форме слова. Каждое из слов в тексте по каждому из функциональных параметров получает вес, измеряющийся числом в интервале от 1 до 0.

На материале старославянских текстов, нашедших отражение в словаре [1] подтверждено наличие зависимости частоты слов от их длины и длины слов от их частоты.

Предложен способ измерения тематической маркированности слов в тексте и соответствующее понятие ИнТеМ (индекс тематической

маркированности слова), вычисляемый по формуле **ИнТеМ = Q-вес — F-вес**, где Q-вес — параметрический вес слова по его частоте, а F-вес — параметрический вес слова по длине. Установлено, что более 98 % слов с отрицательным значением ИнТеМа относятся к тематически нейтральной лексике.

Ближайшей перспективой разработки проблематики статьи является проверка предложенного инструментария на материале «Словаря языка Пушкина», что связано с проблемой формального выделения идиолектно маркированной лексики.

ЛИТЕРАТУРА

1. Старославянский словарь (по рукописям X — XI веков): около 10 000 слов. Авторы: Э. Благова. Р. М. Цейтлин, С. Геродес, Л. Панцерова, М. Бауэрова. — М.: Рус. яз., 1994. — 842 с.

2. *Титов В.Т.* Частная квантитативная лексикология романских языков: Монография / В. Т. Титов; Воронеж. гос. ун-т. — Воронеж: Изд-во Воронеж. гос. ун-та, 2004, С. 15.

3. Библия. Книги Священного Писания Ветхого и Нового Завета. Канонические. В Русском Переводе с Параллельными местами, М.: Изд-ние Всесоюзного Совета Евангельских христиан-баптистов, 1968, С. 75. (Перепечатано с Синодального издания).

4. *Топоров В.Н.* Рыба // Мифы народов мира. Энциклопедия в 2-х т. Т. 2, М.: Советская энциклопедия, 1988, С. 393.

5. *Иванов В.В.* Вино // Мифы народов мира. Энциклопедия в 2-х т. Т. 1, М.: Советская энциклопедия, 1987, С. 236.

Статья принята к опубликованию
25 декабря 2006 г.