УДК 519.6

Идентификация пробела при неизвестной знаковой кодировке в русскоязычных текстах

Ю. А. Котов, О. В. Санина

В работе рассматриваются два критерия идентификации пробела в русскоязычных текстах, представленных в неизвестной знаковой кодировке, и их совместное применение. Оба критерия основаны на сравнении распределения словника текста по длине слов с распределением Пуассона. В первом случае такое сравнение осуществляется на основе разности математического ожидания и дисперсии выборочного распределения, во втором – на основе отношения двух площадей распределения, называемого в работе индексом длины слов. Определена статистика этих критериев для текстов различного объёма и граничные условия для их применения. Проведено экспериментальное исследование погрешности использования данных критериев с найденными граничными условиями для решения задачи идентификации пробела и получена статистика для ошибок первого и второго рода.
На основе полученных данных для текстов различного объёма определены условия совместного использования данных критериев с учётом возможной миграции знака пробела в частотном упорядочивании знаков текста. Проведено экспериментальное исследование погрешности совместного использования данных критериев и получена статистика для ошибок первого и второго рода.

Ключевые слова: знак пробела, идентификация, распределение Пуассона, индекс длины слов, частота встречаемости.

1. Введение

Во многих случаях необходимым начальным условием для анализа и обработки текста является установление факта отсутствия или наличия в тексте знака пробела и его идентификация. От решения этой задачи зависят как словарные, так и частотные методы, используемые при решении задач лингвистики и криптографии [1–7]. При известной кодировке знаков текста определение наличия и идентификация в тексте знака пробела являются тривиальными. Однако в случаях применения произвольной, заранее неизвестной кодировки, называемой в криптографии «шифром простой замены», следует определить методы для решения такой задачи.

Простым способом решения задачи по идентификации пробела является подсчёт частоты встречаемости знаков, т.к. по частоте появления в тексте пробел в большинстве случаев занимает первую позицию в упорядочивании знаков по убыванию частоты встречаемости [8]. Однако, с одной стороны, такой подход не решает проблемы определения наличия в тексте пробела, а с другой стороны, в [8] было показано, что идентификация пробела по частоте встречаемости невозможна в текстах объёмом менее 1400 знаков, поскольку пробел не всегда занимает первое место в частотном упорядочивании знаков таких текстов.

Помимо частоты встречаемости существует другое измеряемое свойство пробела как знака текста, связанное с его основной функцией в текстах, заключающейся в разграничении слов. Можно принять, что пробел является единственным способом обозначения границ слов в текстах. Данное свойство пробела находит своё отражение в распределении слов словаря

текста по длине. Из ряда работ по лингвистике следует, что это распределение хорошо согласуется с распределением Пуассона [9–10].

Например, на рис. 1 представлены аппроксимированные графики дискретного распределения по длине:

1) лемм русского языка в словаре [11] (нормированных к 51646 словам);

2) словоформ без повторений для выборки из русскоязычных текстов 2 из [8, 12] (нормированных к 485099 словам);

3) словоформ без повторений для одного случайно выбранного фрагмента текста размером 150 тыс. знаков из данной выборки (нормированных к 5275 словам).

Все указанные нормировки обозначают количество уникальных слов в соответствующей выборке.

Как можно увидеть из рис. 1, форма распределения сохраняется как для уникальных словоформ в выборках различных фрагментов из текстов, так и для словаря лемм и отдельного текста. Если же в качестве ограничителя слов принять первый знак, следующий непосредственно за пробелом в частотном упорядочении, то вид распределения существенно изменится. На рис. 2 представлен график кусочно-линейной аппроксимации распределения словоформ по длине для текста (3) рис. 1, полученных в таком случае. График нормирован к 10641 «словам».

Очевидно, что в случаях произвольной кодировки текста любой знак-разделитель слов должен обладать такими же свойствами, что и знак пробела. Из примеров на рис. 1 и рис. 2 следует, что по искажению формы распределения слов словаря текста по длине можно судить о наличии в тексте пробела и идентифицировать его. Для этого в работе предложены два подхода:

1) идентификация пробела на основе оценки отклонения распределения слов текстов по длине от распределения Пуассона;

2) непосредственная оценка изменения формы такого распределения на основе введённого в работе индекса длины слов текста;

а также определены условия их совместного применения.

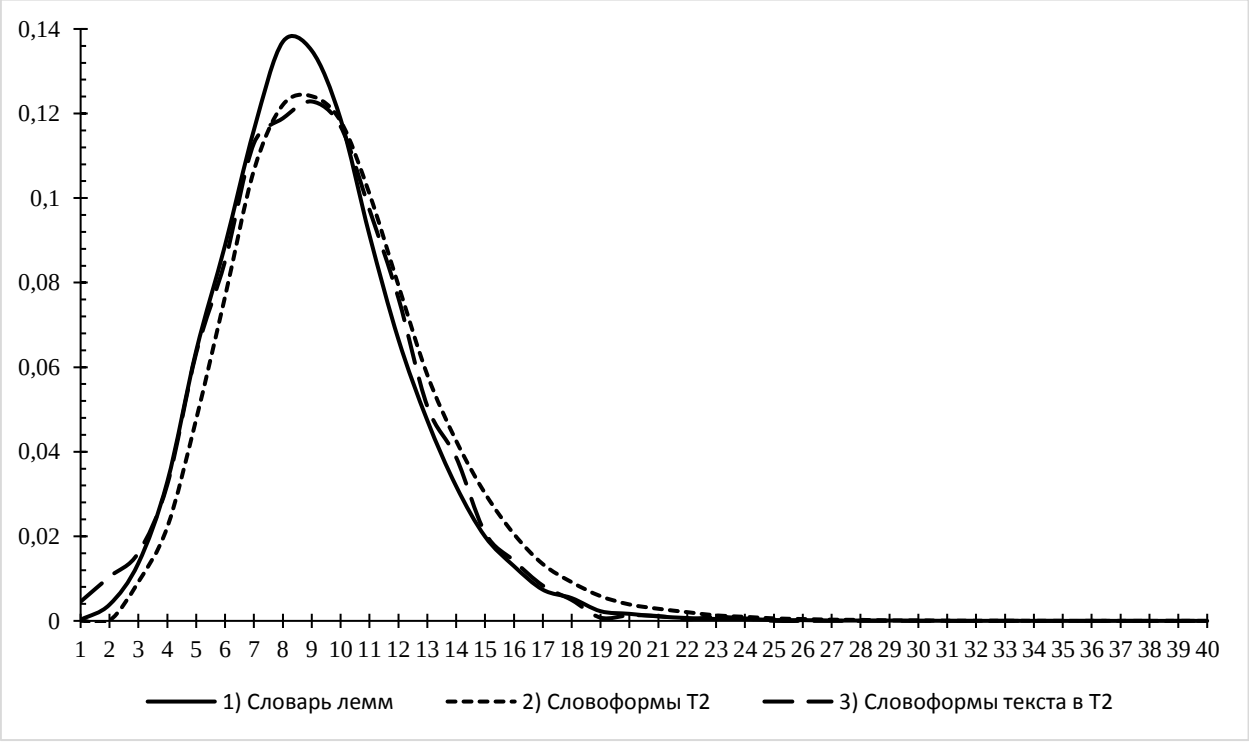

Рис. 1. Распределение по длине:
1) лемм в словаре (норм. к 51646 словам); 2) словоформ в Т2 (норм. к 485099 словам);
3) словоформ в случайном тексте размером 150 тыс. знаков (норм. к 5275 словам)

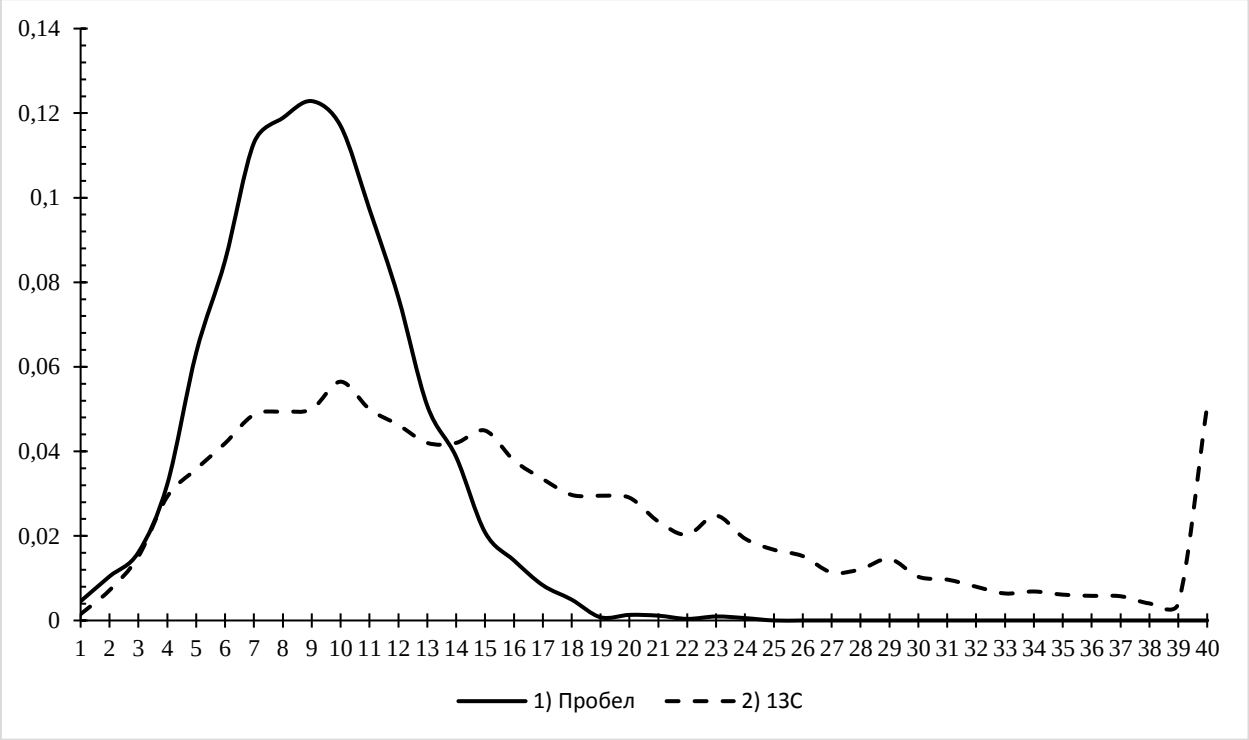

Рис. 2. Распределение по длине в тексте размером 150 тыс. знаков:
1) словоформ, разделённых пробелом (норм. к 5275 словам),
2) словоформ, разделённых знаком, идущим непосредственно за пробелом в частотном упорядочении (норм. к 10641 словам)

2. Идентификация пробела на основе распределения Пуассона

Распределение Пуассона – это распределение, при котором случайная величина ξ принимает дискретные значения $l = 1,2, \ldots$ [13], а функция вероятности имеет вид (1):

$$P(\xi = l) = \lambda^l e^{-\lambda}/l! \quad (1)$$

Важным свойством распределения Пуассона является равенство значений математического ожидания и дисперсии (2):

$$M(l) = D(l) = \lambda. \quad (2)$$

Очевидно, что на практике выполнение равенства (2) следует понимать в приближенной форме: $M(l) \approx D(l)$. Поскольку частотные характеристики пробела зависят от объёма текста [8, 12], то для того, чтобы применить распределение Пуассона к идентификации пробела в любом тексте, необходимо определить зависимость изменений в распределении от объёма текста, а также получить средние и граничные значения изменений и их разброс.

Оценку отклонения от распределения Пуассона в тексте размером x знаков проведём по формуле (3):

$$\delta_i(x) = |M_i(l) - D_i(l)|/M_i(l), \quad (3)$$

где l – длина словоформ ($l = 1,2, \ldots, 40$), $M_i(l)$ и $D_i(l)$ – математическое ожидание и дисперсия длины словоформ в тексте размером x соответственно; i принимает значение 0 для пробела и 1 для первого знака непосредственно за пробелом в частотном упорядочении по убыванию знаков текста. Под словоформой будем понимать уникальную лексему в некоторой грамматической форме без учёта количества её повторений в тексте. Чем ближе значение (3) к нулю,

тем ближе исследуемое распределение к распределению Пуассона. Отношение (3) есть ни что иное, как относительная погрешность определения математического ожидания в распределении Пуассона.

Оценку изменений значений $\delta_i(x)$, вычисляемых по формуле (3), проведем в диапазоне объёмов текстов $200 \le x \le 350000$, $x \in \boldsymbol{N}$. Используемые выборки текстов – тексты 1 и тексты 2, подробное описание которых приведено в [8, 12]. Отметим, что в текстах 1 пробелы исключены из фрагментов текстов. Для текстов 2 получим значения $\delta_i(x)$ для пробела и первого знака в частотном упорядочении, отличного от пробела, а для текстов 1 – первого знака в частотном упорядочении. Результаты измерений средних, минимальных и максимальных значений $\delta_i(x)$ и стандартного отклонения *SD* представлены в табл. 1.

Таблица 1.

	Тексты 2								Тексты 1			
x	δ_0	*min* δ_0	*max* δ_0	*SD* δ_0	δ_1	*min* δ_1	*max* δ_1	*SD* δ_1	δ_1	*min* δ_1	*max* δ_1	*SD* δ_1
Группа 1												
200	1.071	0.105	2.214	0.466	5.352	1.444	15.520	2.528	4.815	1.400	13.170	2.361
400	1.021	0.008	2.312	0.459	5.901	1.563	10.300	1.798	5.404	2.055	10.510	1.820
600	0.974	0.304	3.603	0.462	5.825	2.201	10.170	1.577	5.620	1.909	10.570	1.527
800	0.871	0.279	1.944	0.349	6.206	2.429	9.639	1.382	5.771	1.898	8.924	1.346
1000	0.875	0.223	2.167	0.350	6.107	3.732	9.280	1.175	5.871	2.166	8.396	1.249
1200	0.808	0.253	2.848	0.339	6.319	2.778	9.315	1.127	5.857	2.663	9.279	1.249
1400	0.755	0.294	2.168	0.277	6.388	3.579	9.543	1.150	5.889	3.240	8.926	1.064
1600	0.762	0.252	1.707	0.255	6.121	3.055	8.907	1.069	5.924	3.535	9.271	1.041
1800	0.686	0.304	1.513	0.261	6.228	4.248	8.822	1.136	5.909	3.701	8.399	0.920
2000	-	-	-	-	-	-	-	-	5.865	3.563	8.032	0.864
Группа 2												
2000	0.726	0.265	1.566	0.238	6.048	3.480	8.379	1.068	5.865	3.563	8.032	0.864
4000	0.629	0.276	1.317	0.216	6.105	3.980	8.256	0.934	5.659	4.176	7.213	0.657
6000	0.593	0.166	3.680	0.359	5.892	4.227	7.551	0.698	5.504	4.257	6.573	0.528
8000	0.493	0.180	0.946	0.147	5.980	3.896	7.480	0.692	5.409	4.251	6.748	0.488
10000	0.487	0.205	1.109	0.157	5.903	4.023	7.336	0.668	5.344	4.229	6.538	0.455
Группа 3												
10000	0.469	0.265	0.888	0.152	5.824	4.082	7.758	0.719	5.336	4.394	6.194	0.419
30000	0.356	0.166	0.673	0.129	5.573	4.642	6.395	0.417	4.949	4.034	5.895	0.407
50000	0.311	0.008	0.621	0.133	5.325	4.486	6.373	0.368	4.806	3.943	5.564	0.353
70000	0.309	0.100	0.849	0.139	5.281	4.643	6.076	0.322	4.691	3.965	5.461	0.323
90000	0.287	0.081	0.821	0.147	5.166	4.481	5.850	0.320	4.613	3.962	5.370	0.326
110000	0.278	0.065	0.833	0.147	5.118	4.536	5.731	0.303	4.559	3.913	5.367	0.308
Группа 4												
100000	0.285	0.130	0.518	0.114	5.252	4.691	5.691	0.294	4.664	4.166	5.377	0.334
150000	0.299	0.172	0.468	0.112	5.176	5.013	5.571	0.195	4.559	3.979	5.288	0.294
200000	0.279	0.158	0.429	0.098	5.162	4.982	5.401	0.141	4.495	3.875	5.141	0.294
250000	0.267	0.122	0.411	0.108	4.945	4.689	5.427	0.219	4.493	3.802	5.009	0.313
300000	0.264	0.138	0.416	0.102	5.115	4.764	5.471	0.222	4.509	3.921	5.113	0.295
350000	0.263	0.135	0.393	0.106	5.110	4.941	5.422	0.174	4.496	3.948	5.125	0.285

Из табл. 1 видно, что средние значения $\delta_0(x)$ для пробела с увеличением размера текста имеют тенденцию к уменьшению, в то время как значения для первого знака после пробела фактически колеблются в диапазоне $4.5 < \delta_1(x) < 6.5$.

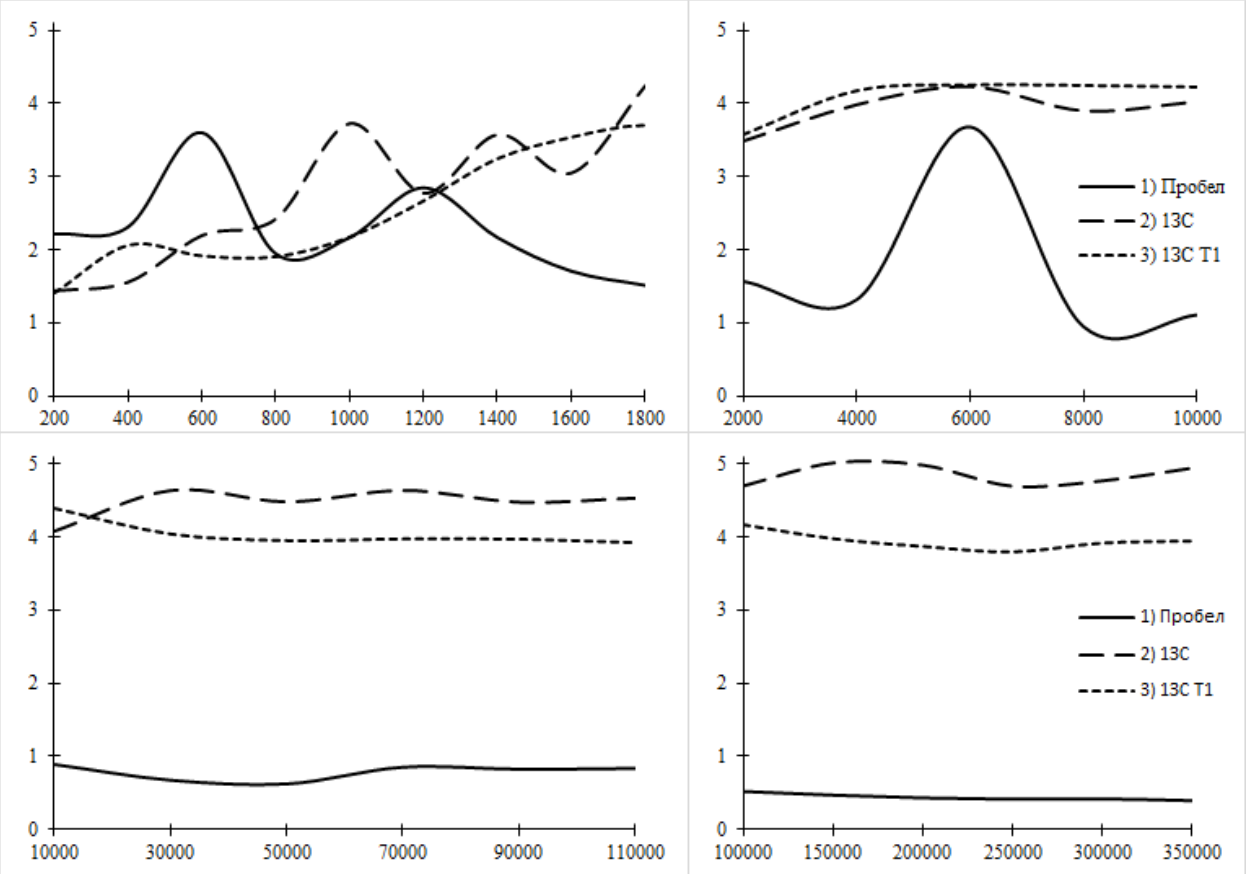

Рис. 3. Границы:
1) максимального значения для пробела и 2) минимального значения для первого значащего символа в текстах 2; 3) минимального значения для первого значащего символа в текстах 1

На рис. 3 видно, что графики *max* δ_0 и *min* δ_1 пересекаются в точке x = 1200, после чего пересечений между ними нет, следовательно, возможна безошибочная идентификация пробела. При x > 1200 наибольшее значение *max(max* δ_0 *)* = 3.68 приходится на точку x = 6000, после чего график *max* δ_0 стабилизируется в области значений, меньших единицы. Разделим шкалу объемов текстов на 2 интервала: [200; 6000] и (6000; 350000]. На втором интервале в качестве постоянного граничного значения возьмем значение 3.75, несколько меньшее среднего значения между *max(max* δ_0*)* и *min(min* δ_1) в точке x = 6000 (табл. 1). Для построения граничной прямой на первом интервале в начальной точке интервала возьмём среднее значение между $\delta_0(200)$ и $\delta_1(200)$, которое затем уточним экспериментально для минимизации ошибок; в конечной точке интервала – значение 3.75. Полученное уравнение граничной линии имеет вид (4):

$$y(x) = \begin{cases} 2.76 \cdot 10^{-4}x + 2.09, & 200 \le x \le 6000 \\ 3.75, & x > 6000 \end{cases}. \quad (4)$$

В случаях, когда оба значения $\delta_0(x)$ и $\delta_1(x)$ больше значения граничной линии (4) в точке x, будем считать, что знака пробела в тексте нет. В противном случае будем считать, что пробел в тексте есть, и в качестве знака пробела принимать тот знак, значение $\delta_i(x)$ для которого меньше либо равно значению (4) в данной точке, либо является наименьшим из двух, если оба значения $\delta_i(x)$ удовлетворяют этому условию.

Для оценки погрешности применения критерия (3) с граничными условиями (4) к решению задачи идентификации пробела на выборках текстов 1 и 2 были подсчитаны ошибки первого и второго рода. Ошибка первого рода α связана с принятием решения об отсутствии пробела в текстах 2, где пробел имеется. Ошибка второго рода β связана с принятием решения

о наличии пробела в текстах 1, не содержащих пробела. Также в эксперименте было обнаружено 8 случаев, когда оба первых знака в частотном упорядочении одновременно удовлетворяли условию (4), однако ошибки идентификации не возникло, поскольку значение отклонения (3) для пробела оказалось меньше, чем для первого значащего знака. Результаты эксперимента представлены в табл. 2, где $P_{ош}$ – частота соответствующей ошибки.

Таблица 2.

x	Тексты 2			Тексты 1			$P_{\text{ошср}}$
	N	α	$P_{\text{ош}_\alpha}$	N	β	$P_{\text{ош}_\beta}$	
200	100	3	0.03	100	10	0.1	0.065
400	100	1	0.01	100	1	0.01	0.01
600	100	2	0.02	100	1	0.01	0.015
800	100	0	0	100	1	0.01	0.005
1000	102	0	0	100	1	0.01	0.005
1200	106	1	0.0094	100	0	0	0.0047
Всего	608	7	0.0115	600	14	0.0233	0.0175

Как видно из табл. 2, в текстах объемом больше 1200 знаков ошибок идентификации нет. При объемах текстов от 800 до 1200 знаков средняя ошибка идентификации не превышает 0.5 %, при объемах текстов от 400 до 800 знаков – 1.5 %. Средняя ошибка идентификации на интервале $[200; 1200]$ знаков составляет 1.75 %, минимальная средняя ошибка – 0.47 % в точке $x = 1200$, максимальная средняя ошибка – 6.5 % в точке $x = 200$. На интервале $800 \leq x \leq 1200$ максимальный уровень значимости критерия (3) – (4) при принятии решения о *наличии* пробела составляет α = 0.01, минимальная мощность критерия: 1–β = 0.99. На интервале $200 < x < 800$ уровень значимости не превышает α = 0.02, мощность критерия: 0.99. При x = 200 уровень значимости α = 0.03, но мощность критерия – 0.9. То есть ошибка β в три раза больше ошибки α и вероятность пропустить пробел для текстов, в которых он есть, в три раза меньше вероятности найти несуществующий пробел в текстах, в которых его нет. Для таких малых текстов ошибки α и β следует поменять местами, тогда уровень значимости при принятии решения об *отсутствии* пробела в тесте будет равен 0.1, а мощность критерия составит 0.97.

3. Индекс длины слов текста

Резкое изменение формы распределения длин «слов» при использовании в качестве разделителя буквы языка, продемонстрированное на рис. 2, позволяет провести более простую, чем (3), оценку такой формы на основе отношения площадей распределения, ограниченных по длине слова, равной шестнадцати (рис. 2). Для этого определим индекс длины слов, вычисляемый по формуле:

$$I_{W_i}(x) = \frac{\sum_1^{15} N_{w_m}(x)}{\sum_{16}^{40} N_{w_m}(x) + 1}, \quad (5)$$

где N_{w_i} – количество слов размером m, $m \in 1,2,..,40$; i принимает значение 0 для пробела, 1 – для первого значащего символа, 2 – для второго значащего символа, и т.д.

При этом, выбирая в качестве границы разделения площадей длину слова, равную шестнадцати, можно быть уверенным, что не менее 95 % распределения Пуассона для русскоязычных текстов будет находиться на интервале длин слов от одного до пятнадцати даже при росте средних значений, например, до десяти. Индекс (5) отвечает условиям хорошей статистики по Шеннону [6], так как вычисляется по меньшей мере в 3 раза быстрее, чем отклонение (3), при тех же исходных данных. Для его практического использования получена статистика средних,

максимальных и минимальных значений, а также стандартного отклонения, которая приведена в табл. 3 и 4.

Статистика, приведенная в табл. 3 и 4, получена на тех же выборках текстов 1 и 2, что и статистика для критерия (3), для различных знаков текста, которые могут быть приняты в качестве пробела. Диапазон измерений: $200 \le x \le 350000$, $x \in N$. В табл. 3 представлена статистика по выборке текстов 2. Приведены значения индекса (5) для пробела и первых двух соседних с ним в частотном упорядочении знаков. Важно отметить, что пробел при этом может находиться не на первом месте в частотном упорядочении.

Таблица 3.

x	I_{W_0}	$min\ I_{W_0}$	$max\ I_{W_0}$	$SD\ I_{W_0}$	I_{W_1}	$min\ I_{W_1}$	$max\ I_{W_1}$	$SD\ I_{W_1}$	I_{W_2}	$min\ I_{W_2}$	$max\ I_{W_2}$	$SD\ I_{W_2}$
Группа 1												
200	20.60	3.50	30.00	6.70	4.58	1.29	13.50	2.79	2.86	0.86	6.67	1.29
400	31.30	7.00	59.00	14.25	4.01	1.00	13.33	1.70	2.70	1.40	5.33	0.81
600	37.37	7.33	80.00	19.27	3.68	1.27	9.00	1.37	2.67	1.00	6.86	0.94
800	47.44	8.86	97.00	25.03	3.44	1.35	6.36	1.08	2.49	1.14	5.00	0.67
1000	45.94	6.30	119.00	29.38	3.31	1.48	8.82	1.07	2.52	1.29	4.71	0.71
1200	51.72	11.14	153.00	33.99	3.15	1.42	7.64	0.88	2.43	1.40	3.84	0.56
1400	57.89	11.55	171.00	39.02	3.07	1.37	5.53	0.78	2.33	0.93	4.09	0.59
1600	53.71	12.90	176.00	38.83	3.11	1.41	6.24	0.83	2.25	1.29	4.17	0.43
1800	63.57	10.63	187.00	43.80	2.98	1.32	6.13	0.79	2.23	1.26	3.38	0.45
Группа 2												
2000	54.38	12.50	177.00	38.46	3.01	1.46	5.23	0.71	2.24	0.92	3.78	0.49
4000	63.49	13.10	271.00	52.08	2.67	1.40	4.65	0.58	2.04	1.25	3.49	0.42
6000	51.65	11.17	225.00	31.27	2.50	1.78	3.55	0.39	1.96	1.20	2.91	0.33
8000	61.01	19.48	241.50	40.40	2.38	1.52	3.78	0.42	1.88	1.19	2.89	0.29
10000	50.37	14.50	269.50	30.30	2.36	1.56	3.50	0.43	1.85	1.19	2.84	0.33
Группа 3												
10000	57.03	16.58	204.33	33.72	1.79	1.15	2.92	0.38	1.46	1.03	1.91	0.20
30000	45.41	17.46	116.00	21.92	1.60	1.13	2.24	0.24	1.34	0.99	1.77	0.16
50000	46.81	21.18	378.00	51.41	1.57	1.23	2.21	0.21	1.31	0.99	1.78	0.15
70000	35.21	18.39	82.72	12.84	1.54	1.25	2.20	0.20	1.31	0.96	1.62	0.16
90000	37.25	16.63	90.06	14.84	1.50	1.14	2.08	0.20	1.25	1.02	1.60	0.13
110000	34.65	18.07	85.20	12.87	1.46	1.06	1.97	0.19	1.25	1.05	1.56	0.13
Группа 4												
100000	29.71	19.05	44.77	8.67	1.56	1.20	1.96	0.25	1.57	1.17	2.71	0.46
150000	27.30	19.95	34.15	5.01	1.38	1.13	1.59	0.14	1.42	1.21	1.69	0.12
200000	26.53	19.65	36.24	5.29	1.61	1.33	1.85	0.17	1.42	1.20	1.74	0.19
250000	25.74	17.98	35.05	6.29	1.42	1.12	1.64	0.18	1.47	1.22	1.94	0.21
300000	24.82	17.53	34.84	5.59	1.51	1.35	1.75	0.12	1.50	1.14	2.16	0.28
350000	24.47	17.14	36.51	6.31	1.39	1.11	1.56	0.15	1.38	1.22	1.51	0.10

В табл. 4 представлена статистика по выборке текстов 1. Поскольку тексты 1 не содержат пробелов, значения индекса (5) приведены для первых трёх в частотном упорядочивании знаков.

Таблица 4.

x	I_{W_1}	min I_{W_1}	max I_{W_1}	SD I_{W_1}	I_{W_2}	min I_{W_2}	max I_{W_2}	SD I_{W_2}	I_{W_3}	min I_{W_3}	max I_{W_3}	SD I_{W_3}
Группа 1												
200	5.05	1.67	19.00	2.43	3.83	1.17	16.00	2.14	2.77	1.17	9.00	1.31
400	4.94	2.00	12.33	1.88	3.55	1.25	9.00	1.28	2.86	1.45	6.00	0.93
600	4.82	2.47	15.25	1.86	3.52	1.69	6.63	0.97	2.69	1.17	4.78	0.69
800	4.67	2.10	16.20	1.76	3.35	1.56	6.75	0.89	2.70	1.35	4.38	0.62
1000	4.57	2.10	11.67	1.58	3.33	1.11	6.22	0.75	2.68	0.29	5.08	0.60
1200	4.39	2.39	10.00	1.27	3.29	1.11	5.82	0.66	2.68	1.63	4.39	0.57
1400	4.38	2.36	9.71	1.09	3.21	1.24	5.13	0.69	2.68	1.43	3.86	0.53
1600	4.41	2.39	9.63	1.11	3.14	1.35	4.75	0.55	2.65	1.72	3.68	0.49
1800	4.31	2.14	7.67	1.00	3.13	1.41	5.08	0.56	2.63	1.70	3.84	0.46
2000	4.35	2.36	8.00	1.03	3.10	1.48	4.42	0.50	2.60	1.65	3.66	0.39
Группа 2												
2000	4.35	2.36	8.00	1.03	3.10	1.48	4.42	0.50	2.60	1.65	3.66	0.39
4000	4.13	2.44	6.52	0.75	2.89	1.98	4.17	0.41	2.44	1.55	3.18	0.30
6000	4.04	2.61	5.37	0.60	2.76	1.95	4.08	0.36	2.40	1.48	2.99	0.27
8000	3.96	2.55	5.12	0.55	2.67	1.98	3.89	0.31	2.34	1.57	2.91	0.23
10000	3.89	2.51	4.98	0.51	2.60	1.92	3.36	0.29	2.26	1.63	2.93	0.22
Группа 3												
10000	3.96	2.51	4.98	0.51	2.60	1.92	3.16	0.25	2.25	1.63	2.93	0.23
30000	3.57	2.00	4.53	0.48	2.35	1.63	2.97	0.26	2.05	1.37	2.44	0.20
50000	3.36	2.04	4.13	0.42	2.23	1.49	2.92	0.24	1.96	1.54	2.38	0.17
70000	3.26	2.07	3.96	0.40	2.19	1.38	2.78	0.23	1.88	1.39	2.31	0.17
90000	3.17	2.09	3.80	0.38	2.14	1.36	2.72	0.23	1.84	1.33	2.35	0.17
110000	3.11	2.05	3.74	0.35	2.09	1.30	2.52	0.21	1.80	1.30	2.27	0.17
Группа 4												
100000	3.12	2.08	3.69	0.38	2.18	1.80	2.70	0.20	1.86	1.50	2.32	0.18
150000	2.99	1.93	3.45	0.31	2.05	1.81	2.34	0.13	1.79	1.56	2.15	0.14
200000	2.89	1.90	3.24	0.30	1.97	1.75	2.23	0.12	1.75	1.34	2.10	0.16
250000	2.88	1.84	3.23	0.32	1.98	1.78	2.16	0.10	1.77	1.29	2.07	0.17
300000	2.91	1.88	3.28	0.33	1.99	1.68	2.18	0.13	1.78	1.28	2.12	0.17
350000	2.90	1.85	3.26	0.33	1.98	1.71	2.20	0.12	1.76	1.30	2.12	0.16

На рис. 4 представлен аппроксимированный график минимального значения индекса (5) для пробела, максимальных значений индекса (5) для первого и второго знака, отличных от пробела. Как видно на рис. 4, в табл. 3 и 4, минимальное значение индекса (5) для пробела по мере увеличения объёма текста сначала с хаотичными колебаниями возрастает до значения 19.5 в точке x = 8000 знаков, а затем колебания стабилизируются около значения 18 и становятся более плавными и менее значительными. В то же время максимальные значения индекса (5) для других знаков изменяются незначительно – сначала с колебаниями убывают, но уже с точки x = 1400 колебания прекращаются и максимальные значения (5) практически гладко и стабильно снижаются с ростом объемов текстов до значения, приблизительно равного двум и почти постоянного начиная с x = 100000.

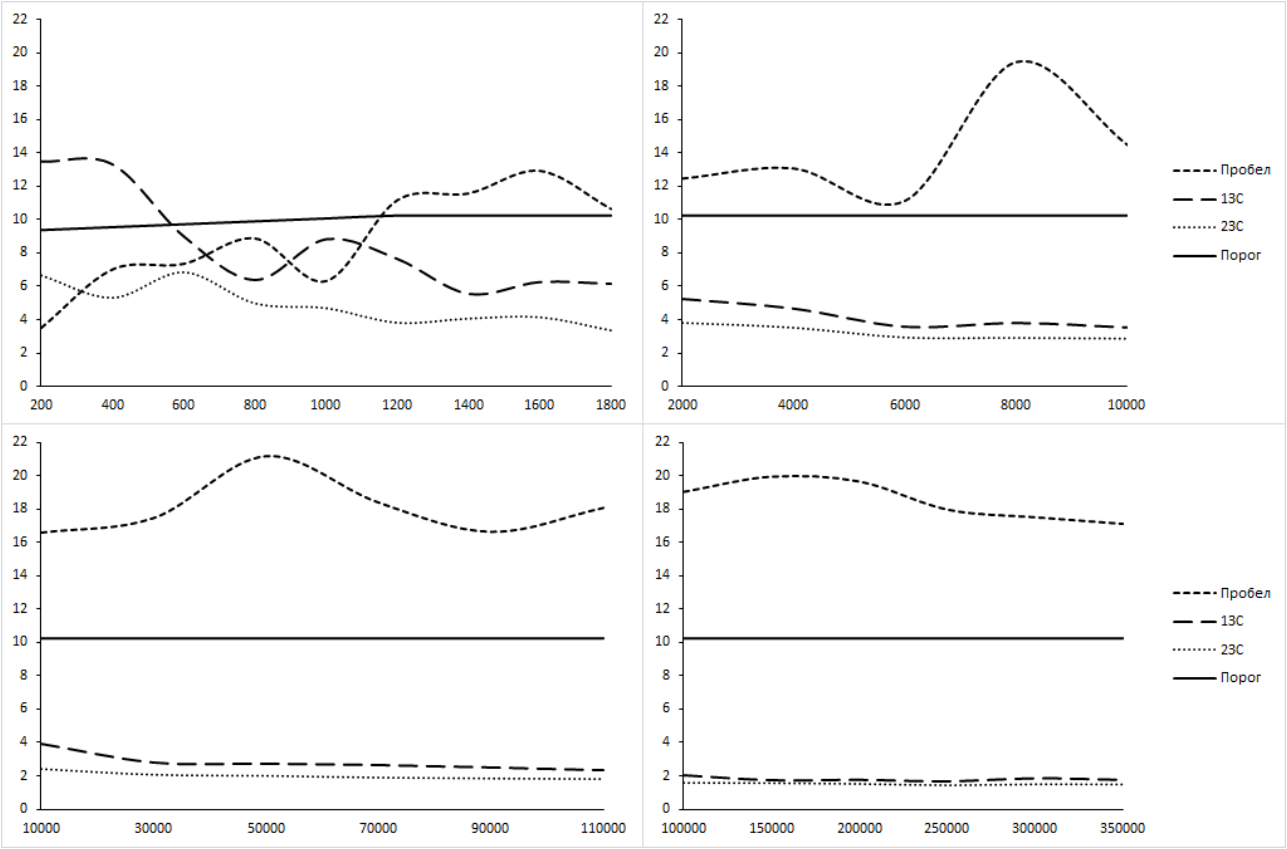

Рис. 4. Значения индекса длины слов: минимальное для пробела, максимальное для 1-го и 2-го значащих символов и пороговое

На рис. 4 видно, что графики минимальных значений (5) для пробела и максимальных значений для первого значащего символа несколько раз пересекаются в различных точках на интервале $x = [200; 1200)$. Начиная с точки x = 1200 пересечений больше нет и возможна однозначная идентификация пробела без ошибок.

Для определения граничных значений разобьём шкалу объемов текстов на два интервала: $200 \le x < 1200$ и $1200 \le x \le 350000$. На втором интервале в качестве постоянного граничного значения возьмем значение 10.25, несколько меньшее среднего значения между *max(max* I_{W1}) и *min(min* I_{W0}) на интервале. На первом интервале в начальной точке возьмём среднее значение между $I_{W0}(200)$ и $I_{W1}(200)$, которое затем уточним экспериментально для минимизации ошибок; в конечной точке – значение 10.25.

Полученное уравнение граничной линии имеет вид:

$$y(x) = \begin{cases} 0.00087x + 9.206, & 200 < x < 1200 \\ 10.25, & x \ge 1200 \end{cases}. \qquad (6)$$

Для оценки погрешности применения критерия (5) с граничным условием (6) были подсчитаны ошибки первого и второго рода. Ошибка первого рода α – установление факта отсутствия пробела в текстах 2, где пробел имеется, т.е. случаи, когда I_{w_0} и I_{w_1} не превышали значения прямой в точке *x*, а также неверная идентификация пробела в текстах 2, т.е. случаи, когда значения I_{w_0} и I_{w_1} превышали граничное, но I_{w_1} оказывался больше или равен I_{w_0} – было обнаружено 2 таких текста объемом *x = 200* знаков. Ошибка второго рода β – установление факта наличия пробела в текстах 1, не содержащих пробела, т.е. случаи, когда I_{w_1} превышало граничные значения (6) в точке *x*. Результаты эксперимента представлены в табл. 5, где $P_{ош}$ – частота соответствующей ошибки.

Таблица 5.

x	Тексты 2			Тексты 1			$P_{\text{ош}_{\text{ср}}}$
	N	α	$P_{\text{ош}_\alpha}$	N	β	$P_{\text{ош}_\beta}$	
200	100	6	0.06	100	5	0.05	0.055
400	100	2	0.02	100	3	0.03	0.025
600	100	6	0.06	100	3	0.03	0.045
800	100	1	0.01	100	1	0.01	0.01
1000	102	1	0.01	100	2	0.02	0.015
1200	106	0	0	100	0	0	0
Всего	608	16	0.0267	600	14	0.0233	0.025

Как видно из табл. 5, на участке $x = [1200; 350000]$ ошибок нет. Средняя ошибка идентификации пробела на интервале $[200; 1000]$ составила 2.5 %, минимальная средняя ошибка – 1 % в точке $x = 800$, максимальная средняя ошибка – 5.5 % в точке $x = 200$. Средние ошибки α и β можно считать приблизительно равными. На интервале $800 \le x \le 1200$ максимальный уровень значимости критерия (5) – (6) при принятии решения о *наличии* пробела составляет $\alpha = 0.01$, минимальная мощность критерия: 0.98. На интервале $200 \le x < 800$ уровень значимости не превышает $\alpha = 0.06$, мощность критерия: 0.95 при $x = 200$ и 0.97 – в остальных случаях.

4. Совместное использование

Критерий (3) с граничными условиями (4) на основе распределения Пуассона показывает в полтора-два раза лучший результат, чем индекс (5) с условием (6) на отрезке $x = [200; 1200)$. Для текстов объемом от 1200 знаков и выше оба критерия показывают одинаковые результаты, но относительная простота вычисления индекса (5) по сравнению с критерием (3) делают первый более предпочтительным. Представляется целесообразным сопряжение критериев (3) и (5) по объемам текстов: для текстов объемом $x = [200; 1200)$ знаков использовать критерий (3), для текстов объемом $x \ge 1200$ знаков – индекс (5).

Кроме того, учтём, что при объемах текстов $x \le 1400$ знаков пробел может занимать не первое место в частотном упорядочивании знаков по убыванию, а переходить на вторую и даже третью позицию [8] (при получении статистики, изложенной в предыдущих двух разделах, этот факт не учитывался и данные получались для конкретных известных знаков – например, для текстов 1 было подсчитано количество ошибок второго рода для первого значащего символа, хотя значение критерия для него не всегда оказывалось наименьшим). Поэтому, прежде чем применить критерий (3) или (5), сначала должны быть подсчитаны значения критерия:

1) для первых трёх знаков в частотном упорядочивании – при объемах текстов $x = 200$ знаков;
2) для первых двух знаков – при объемах текстов $200 < x < 10000$ знаков;
3) для первого знака – при $x \ge 10000$ знаков;

а затем из них выбран наименьший. Выбор данных границ обоснован в работе [8].

Данная методика была проверена на двух вариантах контрольной выборки, составленной по аналогии с выборками [8, 12]. Выборка содержит 2400 фрагментов текстов учебных пособий и художественной литературы, размер которых варьируется от 200 до 350000 знаков. Фрагменты содержат только заглавные буквы сокращенного русского алфавита: $N_A = 31$, «Е» – «Е,Ё», «Ь» – «Ь,Ъ», где N_A – общее количество букв алфавита. Избыточные пробелы – более одного на слово – из фрагментов выборки исключены. Вариант 1 выборки включает пробелы, в варианте 2 пробелы исключены. Количество фрагментов K для каждой точки шкалы измерений и общее количество фрагментов по контрольным выборкам 1 и 2 приведены в табл. 6.

Таблица 6.

x	K	x	K	x	K	x	K	x	K
Группа 1		1400	143	6000	100	50000	100	150000	50
200	100	1600	118	8000	100	70000	100	200000	50
400	100	1800	101	10000	100	90000	100	250000	50
600	100	Всего 1:	1000	Всего 2:	500	110000	100	300000	79
800	100	Группа 2		Группа 3		Всего 3:	600	350000	21
1000	107	2000	100	10000	100	Группа 4		Всего 4:	300
1200	131	4000	100	30000	100	100000	50	Итого:	2400

Результаты совместного использования критериев (3) и (5) с учетом возможной миграции пробела в частотном упорядочивании для текстов контрольных выборок приведены в табл. 7.

Таблица 7.

x	Вариант 1			Вариант 2			$P_{\text{ош}_{\text{ср}}}$
	N	α	$P_{\text{ош}_\alpha}$	N	β	$P_{\text{ош}_\beta}$	
200	100	1	0.01	100	10	0.1	0.055
400	100	1	0.01	100	7	0.07	0.04
600	100	0	0	100	0	0	0
800	100	0	0	100	1	0.01	0.005
1000	107	0	0	107	0	0	0
1200	131	0	0	131	0	0	0
1400	143	0	0	143	1	0.007	0.0035
1600	118	0	0	118	0	0	0
1800	101	1	0.01	101	0	0	0.005
Всего	1000	3	0.003	1000	19	0.019	0.011

Как видно из табл. 7, на участке $x = [2000; 350000]$ ошибок нет. Средняя ошибка идентификации пробела на интервале $x = [200; 1800]$ составила 1.1 %, минимальная средняя ошибка равна нулю сразу в нескольких точках этого интервала, максимальная средняя ошибка – 5.5 % в точке $x = 200$.

При изменении границы сопряжения критериев с 1200 на 2000 знаков ошибки в точках x = 1400 и x = 1800 исчезают.

Таким образом, при совместном использовании рекомендуется применять критерий (3) на интервале [200; 2000) знаков, а критерий (5) – при $x \geq 2000$ знаков.

5. Заключение

Для определения наличия и идентификации пробела в русскоязычных текстах можно применить два подхода, основанных на сравнении распределения словника текста по длине слов с распределением Пуассона: сравнение математического ожидания и дисперсии распределения и сравнение площадей распределения на основе индекса длины слов. Первый подход дает более точное решение задачи: для текстов объемом от 1600 знаков и больше можно ожидать абсолютно точного решения; для текстов от 600 до 1600 знаков уровень значимости при принятии решения об *отсутствии* пробела составляет 0.01, мощность критерия: 0.99; для текстов от 200 до 600 уровень значимости – 0.1; мощность – 0.97. С другой стороны, индекс длины слов вычисляется как минимум в 3 раза быстрее, чем отклонение математического ожидания от дисперсии распределения Пуассона, и отвечает условиям хорошей статистики по Шеннону. С его использованием, как и в первом случае, можно ожидать абсолютно точного решения, но

для текстов объемом от 2000 знаков и больше. Для текстов меньшего объема ошибки первого и второго рода, в отличие от первого подхода, приблизительно равны, что позволяет в равной степени ставить вопрос как о наличии, так и об отсутствии пробела в текстах. С учетом этого рекомендуется использовать оба подхода, сопрягая их для текстов разного объема.

Приведённые в статье результаты могут быть использованы для определения наличия и идентификации пробела в русскоязычных текстах, представленных в неизвестной знаковой кодировке, определения языка текста и при решении других задач формального анализа текстов.

Литература

1. *Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A.* Sentiment in short strength detection informal text // Journal of the American Society for Information Science and Technology. 2010. V. 61, № 12. P. 2544–2558.
2. *Bowker L.* Computer-aided Translation Technology: A Practical Introduction Front Cover. University of Ottawa Press, 2002. 185 p.
3. *Ferrer-i-Cancho R., Elvevag B.* Random texts do not exhibit the real Zipf's law-like rank distribution // PLoS One. 2010. V. 5, № 3. P. 1–10.
4. *Котов Ю. А.* Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. 2016. № 1. С. 181–197.
5. *Котов Ю. А.* Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // Труды СПИИРАН. 2017. № 1 (50). С. 190–208.
6. *Shannon C.* Communication theory of secrecy systems // Bell System Technical Journal. 1949. V. 28, № 4. P. 656–715.
7. *Жданов О. Н., Куденкова И. А.* Криптоанализ классических шифров. Красноярск: Изд-во Сиб. гос. аэрокосм. ун-та им. акад. М. Ф. Решетнева. 2008. 107 с.
8. *Абденов А. Ж., Котов Ю. А., Санина О. В.* Значения некоторых униграммных характеристик русскоязычных текстов // Научный вестник НГТУ. 2017. № 2 (67). С. 146–162.
9. *Воевудский Д. С., Тушавин В. А.* Статистическая обработка лингвистических данных нидерландско-русских словарей // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2013. № 1. С. 169–176.
10. *Smith R. D.* Distinct word length frequencies: distributions and symbol entropies // Glottometrics. 2012. V. 23. P. 7–22.
11. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материале Национального корпуса русского языка). М.: Азбуковник, 2009. 923 с.
12. *Котов Ю. А., Санина О. В.* Значения некоторых биграммных характеристик русскоязычных текстов // Вестник СибГУТИ. 2017. № 4 (40). С. 24–34.
13. *Попов В. А.* Теория вероятностей. Часть 2. Случайные величины. Казань: Казанский университет, 2013. 45 с.

Статья поступила в редакцию 11.09.2018.

Котов Юрий Алексеевич
к.ф-м.н., доцент кафедры защиты информации НГТУ (630073, Новосибирск, пр-т К. Маркса, 20), email: kotov@corp.nstu.ru.

Санина Ольга Валерьевна
магистрант кафедры защиты информации НГТУ, email: lyalysa@gmail.com.

Space character identification in Russian language texts with unknown encoding

Yu. Kotov, O. Sanina

The paper considers two criteria and their joint use for identifying spaces in Russian language texts with unknown encoding. The criteria are based on comparing the word length distribution in a text glossary with the Poisson distribution. The first criterion estimates the difference between the expected value and variance of samples' distribution. The second criterion calculates the relation between two distribution squares called words length index. The measurements were made to estimate criteria statistics for texts varied in size and determine their critical values as well as terms of use. The accuracy of using the criteria with the determined critical values to solve the problem of space identification were calculated, and the statistics of type I and type II errors was obtained.
Based on the obtained statistics for texts of various sizes, the terms for sharing these criteria are determined with regard to possible change in space's place in the frequencies ordering of text characters. The accuracy of the criteria sharing was calculated, and the statistics of type I and type II errors was obtained.

Keywords: space character, identification, Poisson distribution, words' length index, frequency.