ISSN 1814-1196
Научный вестник НГТУ
том 67, № 2, 2017, с. 146–162

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 67, No. 2, 2017, pp. 146–162

СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

MODERN INFORMATION TECHNOLOGIES

УДК 519.6

Значения некоторых униграммных характеристик русскоязычных текстов*

А.Ж. АБДЕНОВ[1], Ю.А. КОТОВ[2], О.В. САНИНА[3]

[1] *010000, Казахстан, г. Астана, ул. Сатлаева, 2, Евразийский национальный университет им. Л.Н. Гумилева, доктор технических наук, профессор кафедры информационных систем. E-mail: amirlan21@gmail.com*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат физико-математических наук, доцент кафедры защиты информации. E-mail: kotov@corp.nstu.ru*
[3] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, студент 4-го курса, направление «Информационная безопасность». E-mail: lyalysa@gmail.com*

Для решения ряда задач анализа текстов, особенно криптографических, необходимы известные значения определенных частотных характеристик текстов на естественном языке. В статье приведены результаты измерений в зависимости от объемов для русскоязычных текстов полноты использования букв алфавита, частоты и места в частотном упорядочивании пробела и двух следующих за ним букв, индекса совпадения. Измерения проведены на двух представительных выборках для научно-популярных и художественных текстов и текстов учебных пособий для вузов. Показано, что единственным знаком в русскоязычных текстах, который может быть идентифицирован по частоте встречаемости в тексте, является знак пробела. Получена оценка случаев, когда пробел находится не на первом месте в частотном упорядочивании знаков текста. Показано, что измерение частоты встречаемости не позволяет ответить на вопрос о наличии или отсутствии знака пробела в тексте.

Показано, что даже при малых объемах русскоязычных текстов в них используются практически все буквы алфавита. Наряду с индексом совпадения и другими характеристиками полученные значения использования букв языка в текстах различного объема могут быть использованы для отделения русскоязычных текстов от текстов на других языках. Определено среднее значение индекса совпадения для текстов, в которых используется только 31 буква русского алфавита в одном регистре, а также доверительные интервалы для различных объемов текстов, для которых не менее 95 % значений индекса для русскоязычных текстов будут находиться внутри данных интервалов.

Ключевые слова: выборка, тексты, буквы, частота встречаемости, аппроксимация, идентификация, индекс совпадения, стандартное отклонение.

DOI: 10.17212/1814-1196-2017-2-146-162

* *Статья получена 19 мая 2017 г.*

ВВЕДЕНИЕ

Спектр задач автоматизированной обработки текстов с развитием вычислительной техники постоянно расширяется и включает в себя не только традиционные задачи компьютерной лингвистики [1–8], но и задачи криптографии, распознавания и идентификации знаков (букв), цепочек знаков (слов и словосочетаний), текстов целиком [9–20]. Для формального анализа текстов на основе частотных характеристик необходимо знать средние и граничные значения требуемых характеристик в зависимости от объемов текстов, их стандартные отклонения. В силу статистического характера текстов необходимые значения можно получить лишь в результате прямых измерений. Известные в литературе значения, например [8–16], имеют, как правило, частный или частичный характер, возможно, устаревшее значение в связи с развитием языка, и в силу своей неполноты могут быть непригодны для решения прикладных задач, связанных с анализом произвольных текстов произвольного объема. В первую очередь это касается частотных характеристик самых общих свойств текстов: зависимости количества используемых букв и их сочетаний от объемов текста, места пробела в частотном упорядочивании знаков, значения индекса совпадения [9] в зависимости от объемов текстов и т. д. Отсутствие этих данных не позволяет решать более сложные задачи анализа текстов, в частности криптографические, или задачи распознавания и идентификации знаков. В работе приведены результаты измерений указанных частотных характеристик на двух представительных выборках русскоязычных текстов: научно-популярных и художественных текстов и текстов учебных пособий для вузов.

1. ОПИСАНИЕ ВЫБОРОК, ИСПОЛЬЗУЕМЫХ ДЛЯ ИЗМЕРЕНИЙ

Диапазон измерений: $200 \le x \le 350\,000$, где $x \in N$ – объем текста в знаках, разбит на четыре интервала, в каждом из которых определена своя шкала измерений, представленная в табл. 1, где K – количество текстов объема x.

Измерения были проведены на двух выборках фрагментов русскоязычных текстов: «Тексты 1» и «Тексты 2», разбитых на четыре группы для каждого интервала измерений [10, 11]. При этом в каждом из фрагментов использовались только заглавные буквы сокращенного русского алфавита: $N_A = 31$, «Е» – «Е, Ё», «Ь» – «Ь, Ъ», где N_A – общее количество букв алфавита и один знак пробела на каждое слово. Другие знаки в текстах выборок не использовались. Выборка 1 была сформирована из 100 научно-популярных и художественных текстов разных жанров и авторов; выборка 2 – из 100 текстов учебных пособий для вузов разных авторов из различных областей знаний: математика, химия, физика, машиностроение и т. д. Из текстов 1-го и 2-го типов случайным образом были выделены последовательные фрагменты различной длины. Выделение фрагментов для выборки 1 происходило после удаления из текста всех пробелов. Выделение происходило по принципу вложенности: сначала выделялись фрагменты большей длины, затем из них выделялись меньшие фрагменты последовательным удалением постоянного объема знаков. При этом начала фрагментов для одного текста совпадали. Это означает, что фрагменты разной длины для одного текста в выборке 1

включены друг в друга, а длина фрагментов совпадает со шкалой длин, представленной в эксперименте.

Таблица 1

Выборки текстов

x	K, тексты 1	K, тексты 2	x	K, тексты 1	K, тексты 2	x	K, тексты 1	K, тексты 2
Группа 1			Группа 2			90 000	50	46
200	100	100	2000	100	99	110 000	50	46
400	100	100	4000	100	100	Всего 3	300	276
600	100	100	6000	100	100	Группа 4		
800	100	100	8000	100	100	100 000	20	8
1000	100	102	10 000	100	100	150 000	20	8
1200	100	106	Всего 2	500	499	200 000	20	8
1400	100	154	Группа 3			250 000	20	8
1600	100	139	10 000	50	46	300 000	20	8
1800	100	99	30 000	50	46	350 000	20	8
2000	100	0	50 000	50	46	Всего 4	120	48
Всего 1	1000	1000	70 000	50	46	Итого	1920	1823

Фрагменты для выборки 2 выделялись со случайного знака текста с учетом пробелов, лишние из которых (более одного на одно слово) удалялись. Таким образом, в отличие от выборки 1, выборка 2 сформирована из фрагментов случайной длины, начинающихся со случайного знака текста и содержащих только один знак пробела на одно слово. Пересечения фрагментов для одного текста в выборке 2 возможны только случайно. Измерения, связанные с объемом текста, для данных фрагментов поводились для так называемого *плотного* объема – т. е. без учета пробелов (кроме оговоренных случаев), а полученные значения отнесены к ближайшим значениям сверху шкалы длин фрагментов, одинаковой со шкалой выборки 1. Это означает, что если в тексте, например, больше 200 пробелов, а шаг шкалы измерений равен 200 знакам, то полученные измерения для фрагмента точки шкалы $x = n$ будут отнесены к точке шкалы $x = n - 1$. Или, например, если шкала начинается с точки $x = 2000$ знаков, то к началу шкалы будут отнесены измерения всех фрагментов меньшего объема. Такие погрешности измерения возможны в конце первой и начале второй групп фрагментов текстов выборки 2. Но встречаются они редко и не оказывают значимого влияния на результаты измерений.

Выборки фрагментов для разных групп проводились независимо друг от друга. Количество фрагментов для каждой точки шкалы и общее количество фрагментов по выборкам 1 и 2 приведены в табл. 1. Выборки 1 и 2 представляют разные модели текстов. Выборка 1 представляет модель семантически связного, последовательно развивающегося (с точки зрения изложения) текста, словарь которого является наиболее общим для всех носителей языка. Такая модель дает возможность оценить в первую очередь зависимость изме-

ряемых величин от объема (длины) одного «усредненного» текста. Выборку 2 можно интерпретировать как модель «произвольного» текста. Семантика учебных пособий едина в рамках учебной дисциплины, но различается, иногда значительно, в разных разделах одного пособия. Изложение содержания, как правило, лаконично и ведется с использованием большого числа локальных сокращений. Текст перемежается большим количеством чисел, формул, таблиц и графиков, при формальном исключении которых возникают синтаксические «разрывы». Используются специальные терминологические словари. Каждый случайный фрагмент текста из учебного пособия, в котором оставлены только буквы языка, представляет собой семантический и синтаксический кластер, не всегда и не полностью грамматически правильный и понятный произвольному носителю языка.

Фрагменты выборки 2 содержат лексические погрешности, представляющие цепочки слов из одной или двух букв, а также фрагменты из одного текста, содержащего орфографические ошибки, примеры которых представлены на рис. 1. Такие погрешности получены как результат автоматизированной подготовки текстов и оставлены в текстах намеренно для оценки их влияния на результаты измерений. Количество таких фрагментов составляет приблизительно 5 % выборки 2, при этом в четвертой группе фрагменты, полученные из ошибочного текста, составляют 16,7 %. Таким образом, усреднение данных по выборке 2 дает возможность оценить зависимость измеряемых величин от объема (длины) одного «произвольного» текста, возможно содержащего лексические и орфографические ошибки.

ВЕЛИЖИНА Х ВОЗРАСТАЕТ И ДОСТИГАЕТ МГ НМ РИС КОЭФФИЕИЕНТ ЗЛАКОУЛАВЛИВАНИЯ ТОПКИ СОСТАВЛЯЕТ
И МИНИМАЛИНАЯ НАГРУЗКА ПО УСЛОВИЯМ ВЫХОДА ЖИДКОГО ЗЛАКА РАВНА НОМ ЗОЛА УНОСА ПО СРАВНЕНИЙ
С ИСХОДНОЙ ЗОЛОЙ БЕРЕЗОВСКОГО УГЛЯ ОБОГАЩАЕТСЯ САО А О И ОБЕДНЯЕТСЯ ЖТО ПРЕДОПРЕДЕЛЯЕТ
ПОВЫЗЕНИЕ ПЛАВКОСТНЫХ ХАРАКТЕРИСТИК УНОСА НА Т С Г Л А В А ЭКОЛОГИЧЕСКИ ЧИСТАИ ТЭС А Б
ВЭК А ПРИВ Т Ч С МГ НМ РИС ЗАВИСИМОСТИ КОНЕЕНТРАЕИ

_261_
СНОСТИ ДЛЯ ЛЙДЕЙ И ЖИВОТНЫХ ПРИ ПОПАДАНИИ В ДЫХАТЕЛИНЫЕ ПУТИ ЗДЕСИ СЛЕДУЕТ ОТМЕТИТИ ОЖЕНИ
НИЗКУЙ ЭФФЕКТИВНОСТИ ЗОЛОУЛАВЛИВАЙЩИХ УСТРОЙСТВ В ЭНЕРГЕТИКЕ ПРЕЖДЕ ВСЕГО ЭЛЕКТРОФИЛИТРОВ
КАК ИЗЗА ПЛОХОГО КАЖЕСТВА САМИХ УСТРОЙСТВ ТАК И ИЗЗА НИЗКОГО УРОВНЯ ИХ ЭКСПЛУАТАЕИИ
Р Р Р Р Р М Р Р Р Р Н Н М М Н Н Н Н М Г ПРИМЕЧАНИЕ Р ВЕЩЕСТВА РАСТВОРИМЫЕ В ВОДЕ БОЛЕЕ Г
НА Г РАСТВОРА М МАЛОРАСТВОРИМО В ВОДЕ МЕНЕЕ Г НА Г РАСТВОРА Н ВЕЩЕСТВО НЕРАСТВОРИМО В

Рис. 1. Пример погрешностей во фрагментах текстов выборки 2

2. ОЦЕНКА ЧАСТОТЫ ПОЯВЛЕНИЯ ТЕКСТОВ, ИСПОЛЬЗУЮЩИХ НЕ ВСЕ БУКВЫ АЛФАВИТА, И СРЕДНЕГО КОЛИЧЕСТВА ИСПОЛЬЗУЕМЫХ В НИХ БУКВ

Диапазон измерений: $200 \le x \le 350\,000$, $x \in \boldsymbol{N}$. Выборки: тексты 1 и тексты 2. Оценку частоты появления текстов $P(x)$, использующих не все буквы алфавита, проведем по формуле (1):

$$P(x) = \frac{K_1(x)}{K(x)}, \tag{1}$$

где $K(x)$ – количество текстов объемом x, $K_1(x)$ – количество текстов объемом x, в которых используются *не все* буквы алфавита.

Пусть $Z(x)$ – среднее количество букв в текстах $K_1(x)$. При этом если $K_1(x) = 0$, то $Z(x) = N_A = 31$. Тогда измерения для $x > 4000$ дают $K_1(x) = 0$, $P(x) = 0$, $Z(x) = N_A = 31$. Результаты измерений в интервале $200 \le x \le 4000$ приведены в табл. 2, где SD Z – стандартное отклонение значения $Z(x)$. Графики аппроксимированных значений $P(x)$ и нормированных к N_A значений $Z(x)$ для текстов 1 и 2 приведены на рис. 2.

Таблица 2

Количество используемых букв алфавита в текстах

x	K	K_1	Z	min Z	max Z	SD Z
Тексты 1						
200	100	98	26,84	23	30	1,64
400	100	80	29,20	27	30	0,81
600	100	51	29,59	28	30	0,60
800	100	36	29,81	28	30	0,46
1000	100	24	29,79	29	30	0,41
1200	100	18	29,94	29	30	0,23
1400	100	14	29,93	29	30	0,26
1600	100	10	29,90	29	30	0,30
1800	100	7	29,86	29	30	0,35
2000	100	7	29,86	29	30	0,35
4000	100	1	30,00	30	30	0,00
Тексты 2						
200	100	99	26,80	21	30	1,79
400	100	90	28,80	25	30	1,30
600	100	70	29,23	25	30	1,00
800	100	58	29,65	27	30	0,66
1000	102	33	29,58	27	30	0,78
1200	106	39	29,72	28	30	0,55
1400	154	28	29,64	28	30	0,67
1600	139	21	29,81	29	30	0,39
1800	99	14	29,64	26	30	1,04
2000	99	12	29,92	29	30	0,28
4000	100	3	30,00	30	30	0,00

Следует отметить очень незначительное – только на 10 % – снижение среднего количества используемых в текстах букв при снижении объема текстов в 10 раз – от 4000 до 400 знаков, и снижение среднего количества используемых букв на 14 % при снижении объемов текста от 4000 до 200 знаков. Можно сделать вывод, что тексты на русском языке «неохотно расстаются» с многообразием букв алфавита, хотя число текстов, использующих не все буквы алфавита, постоянно возрастает – от нуля при $x = 4000$ знаков до 20…30 % при $x = 1000$ знаков, 50 % – при $x = 600$–800 знаков и 100 % – при $x = 200$ знаков. При $x > 4000$ знаков в текстах используются все буквы алфавита, $N_A = 31$, и нарушение этого правила свидетельствует либо об ошибочном тексте, либо о тексте на другом языке, или нетекстовом сооб-

щении. Разброс значений среднего числа используемых букв для текстов 1 возрастает при снижении объемов текста плавно и практически монотонно, для текстов 2 – с некоторыми колебаниями на интервале от 800 до 2000 знаков, что можно объяснить различиями в выборках, отмеченными ранее.

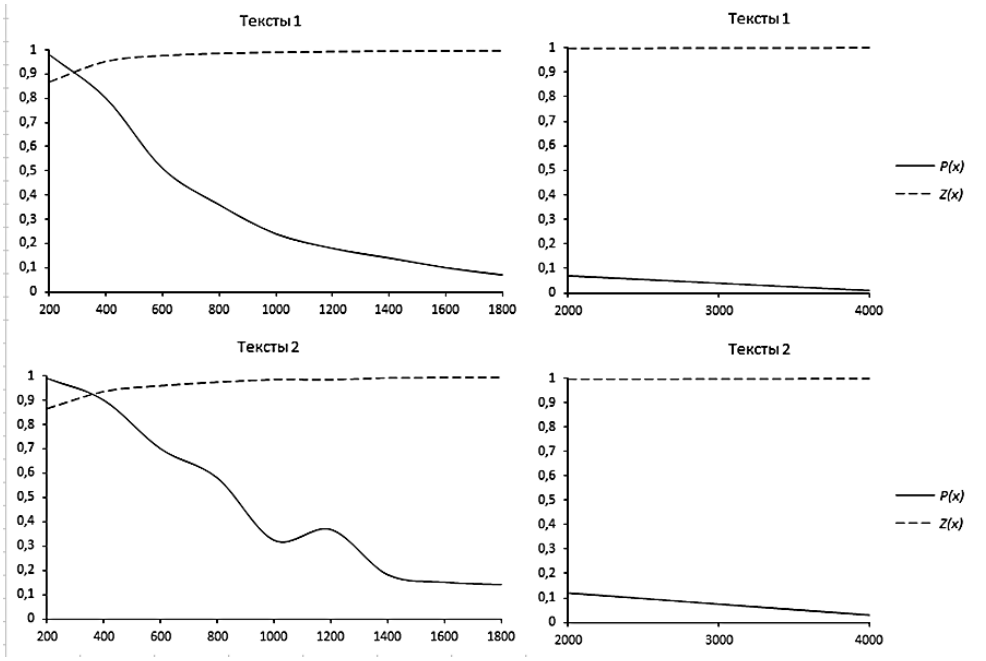

Рис. 2. Количество текстов $P(x)$, использующих не все буквы языка, и среднее количество используемых в них букв $Z(x)$ для русскоязычных текстов

3. ОЦЕНКА ЧАСТОТЫ ПОЯВЛЕНИЯ ПРОБЕЛА В ТЕКСТАХ

Наряду с буквами алфавита знак пробела является одним из важнейших знаков текста, так как разделяет его на слова. Во многих криптографических задачах определение этого знака в тексте (или его отсутствия) имеет решающее значение при частотном анализе текста. Известно, что при эталонном распределении частот знаков пробела и букв алфавита знак пробела занимает первое место в частотном упорядочивании [9]. Но для частотной идентификации пробела в произвольном тексте этой информации недостаточно. Следует определить частотные характеристики пробела в зависимости от объема текста и стандартное отклонение этих характеристик, отличие этих характеристик от подобных значений соседних в частотном упорядочении знаков.

Диапазон измерений: $200 \le x \le 350\,000$, $x \in \boldsymbol{N}$. Выборки: только тексты 2. Результаты измерений приведены в табл. 3. Для $x \ge 2000$ они округлены до целых значений. В столбце «x» T = 1000; $C_0(x)$ – среднее количество пробелов в текстах объемом x; $C_1(x)$ – среднее количество появления в текстах объемом x знака c_1, следующего по убыванию в частотном упорядочивании первым за пробелом; $C_2(x)$ – среднее количество появления в текстах объемом x знака c_2, следующего по убыванию в частотном упорядочивании вторым за

пробелом и первым за знаком c_1; SD $C_0(x)$ – стандартное отклонение $C_0(x)$. Знаки c_1 и c_2 для каждого текста x могут быть любыми буквами алфавита.

Таблица 3

Количество пробелов в текстах

x	C_0	min C_0	max C_0	SD C_0	C_1	min C_1	max C_1	SD C_1	C_2	min C_2	max C_2	SD C_2
Группа 1												
200	26,4	19	48	4,5	20,7	15	29	3,00	16,9	13	23	2,1
400	52,6	40	78	6,6	40,2	31	59	4,87	33,8	26	42	3,3
600	78,3	57	157	13,2	59,0	47	74	6,86	49,6	39	71	5,6
800	102,9	84	146	10,4	76,9	59	104	9,03	65,9	55	84	6,0
1000	129,3	101	218	16,5	95,9	73	126	11,01	82,2	65	96	6,5
1200	159,7	123	406	30,3	114,5	94	154	11,50	99,4	84	127	8,6
1400	194,2	153	296	25,3	138,9	110	192	15,89	119,2	92	161	11,5
1600	225,9	172	511	39,6	165,4	130	210	17,28	142,3	117	167	11,3
1800	259,2	191	310	22,0	187,7	140	228	16,36	161,4	136	192	12,2
Группа 2												
2T	257	213	329	19	190	156	232	16	163	136	201	13
4T	523	433	642	43	372	312	446	31	321	263	389	25
6T	778	679	899	45	555	483	743	45	482	401	608	34
8T	1055	921	1335	73	728	629	868	51	637	517	764	45
10T	1302	1081	1483	80	919	790	1170	75	801	665	933	55
Группа 3												
10T	1303	1175	1506	75	918	809	1155	70	804	716	934	49
30T	3910	3620	4374	178	2712	2380	3071	168	2373	2150	2712	130
50T	6508	5723	7523	327	4513	3885	5038	298,	3980	3588	4450	216
70T	9147	8132	10290	470	6312	5414	6931	355,	5554	4966	6299	296
90T	11765	10362	13289	548	8133	7309	9020	458	7116	6176	8300	431
110T	14370	12814	16227	632	9917	8745	11019	537,	8737	7949	10060	444
Группа 4												
100T	12787	11611	13608	551	8926	8338	9502	367	8196	7720	8791	340
150T	19298	17454	20814	873	13208	12642	13651	345	12208	11692	12920	406
200T	25852	23656	27956	1183	17561	16745	18749	557	16297	15363	17102	557
250T	32084	29152	34451	1364	22085	20221	23765	993	20265	19151	21694	977
300T	38622	35207	40844	1537	26499	24581	28355	1070	24209	22875	25841	10487
350T	44879	40762	47685	1887	30958	28712	32283	1018	28571	26750	30597	1185

Как видно из данных табл. 3, изменения $C_i(x)$ достаточно незначительны, и по этим значениям интервал $200 \le x \le 350\,000$ можно разделить на два отдельных интервала: $200 \le x \le 1800$ (группа 1), $1800 \le x \le 350\,000$ (группы 2–4). Вычислив общие нормированные средние cc_i, $i = 0, 1, 2$ для каждого интер-

вала, и взяв минимальные и максимальные нормированные значения $C_i(x)$ для него, получим оценки, представленные в табл. 4.

Таблица 4

Частота пробела в текстах

Интервал	Знак	Среднее cc_i	min C_i	max C_i
$200 \le x \le 1800$	c_0, пробел	0,134	0,129	0,144
	c_1	0,100	0,095	0,104
	c_2	0,085	0,082	0,090
$1800 \le x \le 350\,000$	c_0, пробел	0,130	0,128	0,132
	c_1	0,090	0,088	0,095
	c_2	0,080	0,079	0,082

Так как в текстах отсутствуют избыточные пробелы, оценку средней длины слова $L(x)$ легко получить на основе обратной к $C_0(x)$ величины по формуле (2):

$$L(x) \approx \frac{1}{C_0(x)} - 1. \qquad (2)$$

Тогда, используя для $C_0(x)$ значения из табл. 4, получим значения по средней длине слова, которые приведены в табл. 5 ($\max C_i$ – $\min L$, $\min C_0$ – $\max L$).

Таблица 5

Средняя длина слова в текстах

Интервал	Среднее L	min L	max L
$200 \le x \le 1800$	6,46	5,94	6,75
$1800 \le x \le 350\,000$	6,69	6,57	6,81

Определим также объемы и количество текстов $K_2(x)$, в которых пробел занимает не первое место в частотном упорядочивании по убыванию знаков текста, данные измерений приведены в табл. 6.

Таблица 6

Количество текстов, где пробел встречается реже, чем буквы

x	200	400	600	800	1000	1200	2000	10 000
K	100	100	100	100	102	106	100	100
K_2	18	7	9	2	2	2	1	1

На основе данных табл. 6 получим оценку погрешности идентификации пробела по первому месту в частотном упорядочении $S(x)$, вычисляемую по формуле (3):

$$S(x) = \frac{K_2(x)}{K(x)}. \qquad (3)$$

График аппроксимированных значений $S(x)$ приведен на рис. 3. Анализируя результаты измерений, можно сделать следующие выводы. На интервале $1400 \le x \le 10\,000$ отмечены только два случая, когда пробел занимает не первое место, что составляет менее 0,23 % от общего числа фрагментов (891) на данном интервале. При $x > 10\,000$ таких фрагментов нет. Более того, проведенные измерения показывают, что при $x \ge 10\,000$ знаков всегда выполняется неравенство (4):

$$\min C_0(x) > \max C_1(x) \tag{4}$$

(см. табл. 3), и, таким образом, в данном интервале пробел не может оказаться не на первом месте в частотном упорядочивании знаков текста.

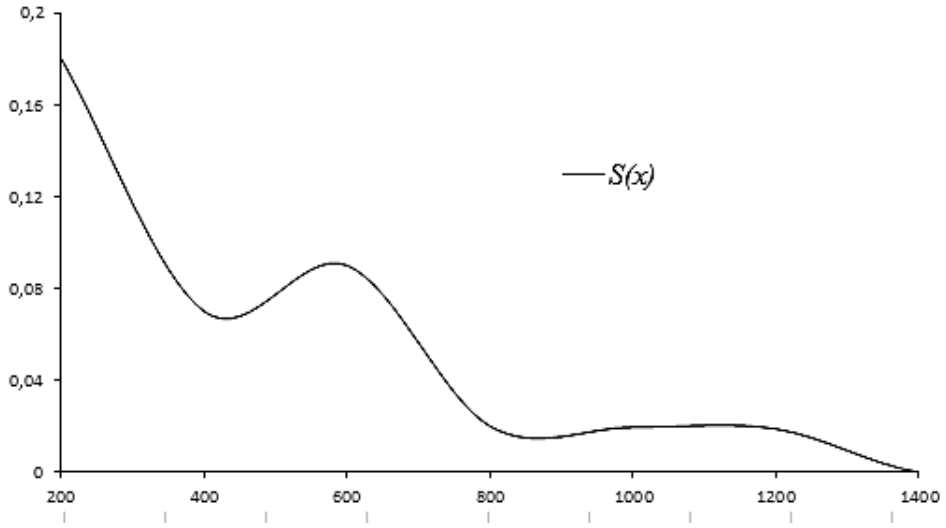

Рис. 3. Относительное количество текстов, в которых знак пробела не является первым в частотном упорядочивании

Следовательно, можно принять, что для русскоязычных текстов объемом $x > 1400$ знаков знак пробела будет всегда располагаться на первом месте (с учетом сделанных замечаний) в частотном упорядочивании (по убыванию) знаков. В интервале $700 \le x \le 1400$ знаков погрешность такого определения не превысит 5 %. При этом в большинстве случаев, когда пробел занимает не первое место, он оказался на втором месте, и только в одном случае из 42 (табл. 6), при $x = 200$, он переместился на третье место в частотном упорядочивании. Однако проведенные измерения показывают также, что знание только частотных характеристик пробела недостаточно для ответа на вопрос о его наличии в тексте.

Сравнение полученной частоты пробела c_0, приблизительно равной 0,134 (табл. 4), с эталонной частотой пробела 0,175 из распределения [9] показывает значительную избыточность пробела в обычных русскоязычных текстах, обусловленную их форматированием. Поэтому для сравнения частот двух следующих знаков c_1 и c_2 из табл. 4 с частотами букв «О» и «Е» из распределения [9] последние необходимо пересчитать с учетом частоты пробела 0,134. Пересчитанные значения частот для букв «О» и «Е» составляют 0,094 и

0,075 соответственно. Они приблизительно совпадают со средними значениями частот c_1 и c_2, но значения для $c_2 \geq 0{,}08$ оказываются выше 0,075, тогда как значение 0,094 находится между значениями 0,088 и 0,1 для средних частот c_1 для различных интервалов (табл. 4). Еще раз подчеркнем: какие буквы представляют знаки c_1 и c_2, при измерениях не анализировалось.

Анализ стандартных отклонений табл. 4 показывает колебания разброса средних частот C_i (x), $i = 0{,}1{,}2$ для пробела, знаков c_1 и c_2 соответственно. Их частота и амплитуда увеличиваются по мере уменьшения объема текста. При этом наибольшие колебания отмечаются для разброса среднего значения пробела. При $x \geq 6000$ знаков относительный разброс среднего значения частоты пробела начинает успокаиваться, и при $x \geq 10\,000$ его можно считать практически постоянным. На этом же интервале также снижаются и реже пересекаются колебания разброса средних частот знаков c_1, c_2, однако эти колебания не затухают вплоть до верхней границы в 350 000 знаков.

На основе проведенных измерений можно сделать вывод о проблематичности определения того, является ли данный знак пробелом, только по частоте его появления в тексте. Частично это можно сделать для текстов, формализованных как в выборке 2 и объемом $x \geq 4000$ путем сравнения количества используемых в тексте знаков с мощностью алфавита языка N_A (табл. 2, рис. 2). Как уже отмечалось, в таких текстах должны использоваться все буквы языка, и появление дополнительного знака может быть интерпретировано как появление знака пробела. Однако данный критерий не представляется удовлетворительным. В целом для решения задачи наличия либо отсутствия в тексте пробела частоты его появления недостаточно, и должны быть использованы и другие частотные характеристики.

4. ОЦЕНКА ИНДЕКСА СОВПАДЕНИЯ

Оценкой формы частотного распределения знаков текста является индекс совпадения [9], вычисляемый по формуле (5):

$$I_c(x) = \frac{\sum_{1}^{NA} C_i[C_i - 1]}{x(x-1)}, \tag{5}$$

где $NA = N_A$, C_i (x) – число вхождений знака c_i в текст объемом x знаков.

Несложно увидеть, что индекс совпадения представляет собой отношение суммы квадратов к квадрату суммы, и если текст состоит только из разных знаков, он равен нулю, а если текст состоит из одного повторяющегося знака, то индекс совпадения равен единице. При одинаковом объеме текста x индекс совпадения тем больше, чем меньше число используемых в нем знаков ($a^2 > (a - b)^2 + b^2$, $a > b$), и, таким образом, зависит от мощности используемого алфавита. Индекс совпадения дает интегральную оценку формы частотного распределения знаков текста в зависимости от мощности алфавита и имеет довольно стабильное значение для текстов различного объема. Это позволяет, в частности, определять принадлежность текста тому или иному языку путем сравнения расчетного значения индекса совпадения с эталонным значением. В работе [9] приведены данные по значениям индекса совпадения

для текстов на некоторых языках, в частности для русского языка 0,0529 и английского языка 0,0662. Приблизительно оценить значение индекса совпадения можно на основе суммы квадратов вероятностей, в качестве которых можно взять оценки эталонных частот, пересчитанные с учетом исключения пробела [9]. Тогда для русского языка получим $I_c(x) \approx 0{,}0556$, для английского – $I_c(x) \approx 0{,}0677$. И если для английского языка такое приближение можно в некоторой степени считать удовлетворительным, то разница в значениях индекса для русского языка показывает значительное расхождение. Возможно, это связано с тем, что при определении значений индексов совпадений, приведенных в работе [9], отдельно учитывались заглавные и строчные буквы языков, но явно это не указано. Кроме того, по данным исследования [9], невозможно оценить минимальные и максимальные значения индекса совпадения и стандартные отклонения, необходимые для его практического применения. Для уточнения средних значений индекса совпадения (5), а также определения его минимальных и максимальных значений и стандартного отклонения проведены измерения индекса совпадения для русскоязычных текстов, результаты которых представлены в табл. 7.

Таблица 7

Значения индекса совпадения для текстов разного объема

Тексты 1					Тексты 2				
x	I_C	min I_C	max I_C	SD I_C	x	I_C	min I_C	max I_C	SD I_C
Группа 1									
200	0,05831	0,04709	0,06844	0,00440	200	0,05792	0,04799	0,07276	0,00515
400	0,05764	0,04981	0,07008	0,00352	400	0,05938	0,04979	0,07449	0,00446
600	0,05752	0,04929	0,06501	0,00307	600	0,05883	0,05111	0,07840	0,00408
800	0,05725	0,05013	0,06425	0,00261	800	0,05857	0,05177	0,07200	0,00365
1000	0,05729	0,05005	0,06864	0,00271	1000	0,05856	0,05233	0,07021	0,00306
1200	0,05712	0,04993	0,06572	0,00245	1200	0,05852	0,05231	0,06947	0,00270
1400	0,05702	0,05061	0,06339	0,00230	1400	0,05820	0,05071	0,06461	0,00263
1600	0,05695	0,05145	0,06175	0,00221	1600	0,05844	0,05250	0,06508	0,00238
1800	0,05693	0,05226	0,06220	0,00211	1800	0,05829	0,05072	0,07059	0,00272
2000	0,05695	0,05202	0,06263	0,00219	–	–	–	–	–
Группа 2									
2000	0,05695	0,05202	0,06263	0,00219	2000	0,05833	0,05199	0,06470	0,00239
4000	0,05684	0,05169	0,06079	0,00168	4000	0,05831	0,05204	0,06323	0,00237
6000	0,05679	0,05184	0,06061	0,00142	6000	0,05813	0,05365	0,06513	0,00200
8000	0,05681	0,05220	0,06006	0,00133	8000	0,05758	0,05253	0,06156	0,00173
10000	0,05681	0,05254	0,05993	0,00127	10000	0,05788	0,05376	0,06483	0,00170
Группа 3									
10000	0,05698	0,05386	0,05993	0,00118	10000	0,05801	0,05530	0,06320	0,00187
30000	0,05676	0,05401	0,05933	0,00111	30000	0,05753	0,05346	0,06149	0,00138
50000	0,05666	0,05445	0,05909	0,00099	50000	0,05763	0,05529	0,06054	0,00121
70000	0,05666	0,05462	0,05890	0,00097	70000	0,05740	0,05537	0,06006	0,00115
90000	0,05665	0,05481	0,05888	0,00094	90000	0,05744	0,05498	0,05998	0,00111
110000	0,05664	0,05491	0,05900	0,00091	110000	0,05748	0,05501	0,06002	0,00116

Окончание табл. 7

Тексты 1					Тексты 2				
x	I_C	min I_C	max I_C	SD I_C	x	I_C	min I_C	max I_C	SD I_C
Группа 4									
100000	0,05698	0,05548	0,05882	0,00097	100000	0,05817	0,05686	0,06029	0,00126
150000	0,05700	0,05547	0,05960	0,00102	150000	0,05785	0,05678	0,05949	0,00104
200000	0,05702	0,05547	0,05968	0,00106	200000	0,05759	0,05599	0,05897	0,00096
250000	0,05703	0,05555	0,05985	0,00111	250000	0,05772	0,05643	0,05941	0,00085
300000	0,05701	0,05561	0,05971	0,00109	300000	0,05778	0,05592	0,05911	0,00093
350000	0,05698	0,05570	0,05999	0,00107	350000	0,05782	0,05608	0,05932	0,00099

График аппроксимированных значений $I_c(x)$ из табл. 7 приведен на рис. 4. Как можно увидеть из данных табл. 7 и графика на рис. 4, среднее значение индекса совпадения довольно стабильно для текстов различного объема, особенно для текстов 1 при $x \geq 1400$ знаков и текстов 2 при $x \geq 8000$ знаков. Общее среднее значение индекса совпадения во всем диапазоне измерений $200 \leq x \leq 350\,000$ для текстов 1 составляет 0,0570, для текстов 2 – 0,0581, при этом разница между значениями увеличивается при снижении объемов текста от $x < 8000$ знаков. Наименьшие и наибольшие значения индексов совпадения 0,0471 и 0,0784 и максимальное значение стандартного отклонения 0,0051 достигаются при уменьшении объемов текста от $x \leq 600$ знаков. При этом, как видно из табл. 7, с ростом x стандартное отклонение монотонно снижается и минимальные и максимальные значения индекса совпадения приближаются к средним значениям. Однако и при значениях $x \leq 600$ стандартное отклонение значения индекса совпадения относительно невелико. Отмеченные минимальные и максимальные значения для такого объема связаны в основном с уменьшением количества используемых знаков (рис. 2) и/или с упомянутыми ранее погрешностями и ошибками в текстах и составляют единичные наблюдения.

Для достижения заданного уровня значимости при определении языка текста на основе индекса совпадения необходимо в первую очередь установить доверительные интервалы. На основе данных табл. 7 разделим шкалу измерений на четыре интервала и предположим, что для получения уровня значимости решения не менее 95 % для текстов 1 и 2 одновременно границы интервалов должны определяться на основе минимального и максимального значений среднего и удвоенного стандартного отклонения. Получим четыре интервала:

1) $200 \leq x < 800$, $0{,}0495 < I_c(x) < 0{,}0683$;
2) $800 \leq x < 8000$, $0{,}05187 < I_c(x) < 0{,}06587$;
3) $8000 \leq x < 100\,000$, $0{,}05415 < I_c(x) < 0{,}06175$;
4) $100\,000 \leq x \leq 350\,000$, $0{,}05481 < I_c(x) < 0{,}06069$.

Проверим предположение экспериментально, подсчитав отношение количества фрагментов для текстов 1 и 2, значения индекса совпадения для которых не попадают в выделенный доверительный интервал, к общему числу фрагментов точки шкалы измерений $Q(x)$. Результаты измерений приведены в табл. 8.

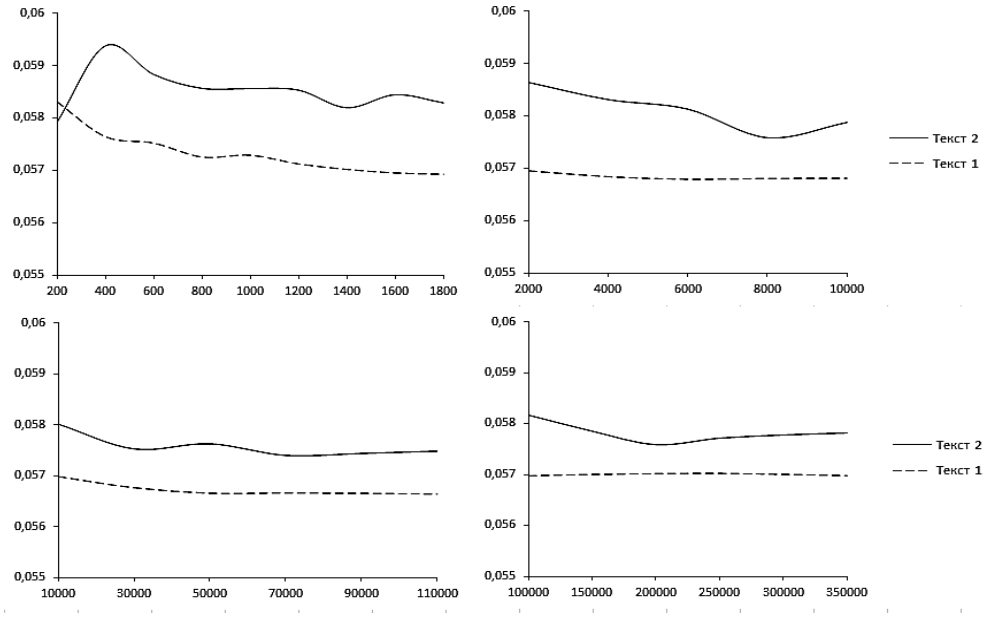

Рис. 4. Значения индекса совпадения $I_c(x)$ для русскоязычных текстов

Таблица 8

Выход из доверительных интервалов для индекса совпадения

Интервал	x	Q	Интервал	x	Q	Интервал	x	Q
$200 \le x < 800$	200	0,0450	$800 \le x < 8000$	1800	0,0101	$8000 \le x < 100\,000$	70 000	0
	400	0,0300		2000	0,0050		90 000	0
	600	0,0150		4000	0,0050	$100\,000 \le x \le 350\,000$	100 000	0
$800 \le x < 8000$	800	0,0350		6000	0,0050		150 000	0
	1000	0,0297	$8000 \le x < 100\,000$	8000	0,0450		200 000	0
	1200	0,0097		10 000	0,0313		250 000	0
	1400	0,0079		30 000	0,0208		300 000	0
	1600	0,0042		50 000	0		350 000	0

Как видно из данных табл. 8, во всех точках шкалы измерений $Q(x) < 0{,}05$. Следовательно, при попадании значения $I_c(x)$ в выделенные интервалы можно в качестве начального приближения принять гипотезу о русскоязычном тексте с уровнем значимости 95 %. Однако для определения языка текста только этого недостаточно. Необходимо знать аналогичные данные для текстов на других языках и *не текстов*, от которых необходимо отделить русскоязычные тексты. При этом даже данных о распределении значений индекса совпадения и их попадании в доверительный интервал в общем случае недостаточно в связи с пересечением таких распределений. К анализу необходимо привлекать и другие характеристики текстов. Например, данные о количестве используемых в тексте знаков (табл. 2, рис. 2) и иные критерии.

При использовании индекса совпадения для определения языка текста следует помнить, что он дает оценку формы частотного распределения знаков текста и не зависит от их расположения в тексте. Индекс совпадения может иметь одинаковые значения и для открытого текста, и для текста, зашифрован-

ного, например, перестановкой, а также и случайного набора знаков. Следовательно, вывод о языке текста только на основе значения индекса совпадения, без учета контекста задачи, может оказаться безосновательным. Индекс совпадения дает необходимую, но недостаточную информацию для определения языка текста. Индекс совпадения можно применять также для определения размеров ключа некоторых шифров, как это, например, показано в работе [9]. С другой стороны, чувствительность индекса совпадения к ошибкам и погрешностям в тексте может быть использована для локализации участков с ошибками при достаточном объеме текста. Для русскоязычных текстов, использующих 31 букву алфавита в одном регистре, следует в качестве эталонного применять значение индекса $I_c(x) = 0{,}0576$, вычисленного без учета пробела.

ЗАКЛЮЧЕНИЕ

Единственным знаком в русскоязычных текстах, который может быть идентифицирован по частоте встречаемости в тексте, является знак пробела, если он в тексте есть. Для текстов, из которых удалены избыточные пробелы, при минимальном объеме текста в 200 знаков ошибка такой идентификации не превысит 18 %, а суммарная ошибка для произвольных объемов текстов в диапазоне от 200 до 1400 знаков не превысит 5 %. При больших объемах текста погрешность такой идентификации будет равна нулю.

Даже при малых объемах русскоязычных текстов в них используются практически все буквы алфавита. Все буквы используются в текстах объемом от 4000 знаков и выше, однако при снижении объема текстов на порядок – до 400 знаков – среднее количество используемых в них букв уменьшается только до 29, а минимальное – до 27. В текстах объемом 200 знаков соответствующие значения составляют 27 и 23 буквы соответственно.

Средняя длина слова для объемов текста от 200 до 1800 знаков составляет 6,46 буквы, для объемов текста свыше 1800 знаков – 6,69 буквы. Среднее значение индекса совпадения для тестов, в которых используются только 31 буква русского алфавита в одном регистре, составляет 0,0576. При использовании для различных объемов текстов x следующих доверительных интервалов: (1) $200 \le x < 800$, $0{,}0495 < I_c(x) < 0{,}0683$; (2) $800 \le x < 8000$, $0{,}05187 < I_c(x) < 0{,}06587$; (3) $8000 \le x < 100\,000$, $0{,}05415 < I_c(x) < 0{,}06175$; (4) $100\,000 \le x \le 350\,000$, $0{,}05481 < I_c(x) < 0{,}06069$; не менее 95 % значений индекса для русскоязычных текстов будут находиться внутри данных интервалов.

Приведенные в статье результаты измерений ряда частотных характеристик русскоязычных текстов позволяют формализовать решения задач по определению наличия в текстах пробела, идентификации знака пробела и языка текстового сообщения, идентификации знаков сообщения, а также будут полезны при решении других задач формального анализа текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Соснина Е.П.* Введение в прикладную лингвистику. – Ульяновск: Изд-во: УлГТУ, 2012.

2. *Sidorov G.* Syntactic dependency based N-grams in rule based automatic English as second language grammar correction // International Journal of Computational Linguistics and Applications. – 2013. – Vol. 4, N 2. – P. 169–188.

3. Syntactic N-grams as machine learning features for natural language processing / G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández // Expert Systems with Applications. – 2013. – Vol. 41, N 3. – P. 853–860.

4. *Нокель М.А.* Метод учета структуры биграмм в тематических моделях // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2014. – № 4. – С. 89–97.

5. *Васильев Е.М., Жданова Д.В.* Диахроническое исследование энтропии графем русского письма // Вестник Воронежского государственного технического университета. – 2010. – Т. 6, № 4. – С. 138–140.

6. *Васильев Е.М., Гусев К.Ю.* Анализ избыточности русскоязычного текста // Вестник Воронежского государственного технического университета. – 2010. – Т. 6, № 8. – С. 101–104.

7. *Губарев В.В.* Введение в теоретическую информатику. – Новосибирск: Изд-во НГТУ, 2014. – 420 с.

8. *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М.: Азбуковник, 2009. – 923 с.

9. *Жданов О.Н., Куденкова И.А.* Криптоанализ классических шифров. – Красноярск: Изд-во Сиб. гос. аэрокосм. ун-та им. М.Ф. Решетнева, 2008. – 107 с.

10. *Котов Ю.А.* Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. – 2016. – № 1 (44). – С. 181–197.

11. *Котов Ю.А.* Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // Труды СПИИРАН. – 2017. – № 1 (50). – С. 190–208.

12. Развитие криптографических методов и средств защиты информации / Л.К. Бабенко, Е.А. Ищукова, Е.А. Маро, И.Д. Сидоров, П.П. Кравченко // Известия ЮФУ. Технические науки. – 2012. – № 4. – С. 40–50.

13. *Бабенко Л.К., Ищукова Е.А.* Анализ симметричных криптосистем // Известия ЮФУ. Технические науки. – 2012. – № 12. – С. 136–147.

14. Введение в теоретико-числовые методы криптографии / М.М. Глухов, И.А. Круглов, А.Б. Пичкур, А.В. Чертмушкин. – СПб.: Лань, 2011. – 400 с.

15. *Минеев М.П., Чубариков В.Н.* Лекции по арифметическим вопросам криптографии. – М.: Попечительский совет Механико-математического факультета МГУ им. М.В. Ломоносова, 2010. – 186 с.

16. *SambasivaRao Baragada, Satyanarayana Reddy P.* A survey of cryptanalytic works based on Genetic Algorithms // International Journal of Emerging Trends & Technology in Computer Science. – 2013. – Vol. 2, iss. 5. – P. 18–22.

17. *Amrit Pal Singh, Pal S.K., Bhatia M.P.S.* The firefly algorithm and application in cryptanalysis of monoalphabetic substitution ciphers // American Journal of Computer Science and Engineering Survey. – 2013. – Vol. 1, N 1. – P. 33–52.

18. *Морозенко В.В., Плешкова И.Ю.* О применении генетического алгоритма для криптоанализа шифра Тритемия–Белазо–Виженера // Современные проблемы науки и образования. – 2014. – № 2. – С. 1–11.

19. *Aditi Bhateja, Shailender Kumar, Ashok K. Bhateja*. Cryptanalysis of vigenere cipher using particle swarm optimization with Markov chain random walk // International Journal on Computer Science and Engineering. – 2013. – Vol. 5, no. 5. – P. 422–429.

20. *Mohan M., Kavitha Devi M.K., Jeevan Prakash V.* Security analysis and modification of classical encryption scheme // Indian Journal of Science and Technology. – 2015. – Vol. 8, no. 8. – P. 542–548.

Абденов Амирза Жакенович, профессор, доктор технических наук, профессор кафедры информационных систем Евразийского национального университета им. Л.Н. Гу-милева, г. Астана, Казахстан. Основное направление научных исследований – информационная и компьютерная безопасность. Имеет более 65 публикаций. E-mail: amirlan21@ gmail.com.

Котов Юрий Алексеевич, кандидат физико-математических наук, доцент кафедры защиты информации Новосибирского государственного технического университета. Основные направления научных исследований: информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Имеет более 29 публикаций. E-mail: kotov@corp.nstu.ru

Санина Ольга Валерьевна, студент Новосибирского государственного технического университета. Основное направление научных исследований – информационная безопасность, криптография. E-mail: lyalysa@gmail.com

Values of some unigram frequency characteristics of Russian language texts *

A.Zh. ABDENOV[1], *Yu.A. KOTOV* [2], *O.V. SANINA*[3]

[1] *EA National University, 2, Satlaev Street, Astana, 010000, Kazakhstan, D. Sc. (Phys. & Math), professor. E-mail: amirlan21@gmail.com*
[2] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, Ph.D. (Phys. & Math), associate professor. E-mail: kotov@corp.nstu.ru*
[3] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, student. E-mail: lyalysa@gmail.com*

To solve a number of problems related to analyzing texts, especially cryptographic ones, some known frequency characteristics values of natural language texts are required. Based on the Russian language text size the paper provides some measuring results of the following characteristics: the level of alphabet characters usage, the frequency of occurrence and position of a space and two following it characters in occurrence ordering, and the coincidence index of texts Two representative samples are studied: the first sample includes nonfiction and fiction texts and the second one consists of university study guides. The paper shows that a space is the only character to be identified by its frequency of occurrence in texts. Cases when the space character does not take the first position in character occurrence ordering in texts of different sizes are analyzed. It is shown that measuring the frequency of occurrence does not allow answering the question about the presence or absence of a space character in the text.

Even in low-sized Russian language texts, almost all alphabet characters are used. Along with the coincidence index and other characteristics, the obtained values of characters usage in texts of different sizes can be used to distinguish Russian texts from texts in other languages. The average value of the coincidence index for texts with only 31 letters in one letter case as well as the index confidence interval containing no less than 95% of index values for texts of different sizes are given.

Keywords: sample, texts, character, occurrence frequency, approximation, identification, coincidence index, standard deviation

DOI: 10.17212/1814-1196-2017-2-146-162

REFERENCES

1. Sosnina E.P. *Vvedenie v prikladnuyu lingvistiku* [Introduction in applied linguistics]. Ul'yanovsk, UlGTU Publ., 2012.

2. Sidorov G. Syntactic dependency based N-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications*, 2013, vol. 4, no. 2, pp. 169–188.

3. Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., Chanona-Hernández L. Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 2013, vol. 41, no. 3, pp. 853–860.

4. Nokel M.A. Metod ucheta struktury bigramm v tematicheskikh modelyakh [Method of accounting structure bigrams in thematic models]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyi analiz i informatsionnye tekhnologii – Proceedings of Voronezh State University. Series: Systems analysis and information technologies*, 2014, no. 4, pp. 89–97.

5. Vasil'ev E.M., Zhdanova D.V. Diakhronicheskoe issledovanie entropii grafem russkogo pis'ma [Diachronic study of the entropy of the graphemes of the Russian writing]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta – The Bulletin of Voronezh State Technical University*, 2010, vol. 6, no. 4, pp. 138–140.

* *Received 19 May 2017.*

6. Vasil'ev E.M., Gusev K.Yu. Analiz izbytochnosti russkoyazychnogo teksta [Redundancy analysis of Russian text]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta – The Bulletin of Voronezh State Technical University*, 2010, vol. 6, no. 8. pp. 101–104.

7. Gubarev V.V. *Vvedenie v teoreticheskuyu informatiku* [Introduction to theoretical informatics]. Novosibirsk, NSTU Publ., 2014. 420 p.

8. Lyashevskaya O.N., Sharov S.A. *Chastotnyi slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency dictionary of modern Russian language (materials Russian National Corpus)]. Moscow, Azbukovnik Publ., 2009. 923 p.

9. Zhdanov O.N., Kudenkova I.A. *Kriptoanaliz klassicheskikh shifrov* [Cryptanalysis of classical ciphers]. Krasnoyarsk, Siberian State Aerospace University Publ., 2008. 107 p.

10. Kotov Yu.A. Determinirovannaya identifikatsiya bukvennykh bigramm v russkoyazychnykh tekstakh [Determinate identification of Russian text letter bigrams]. *Trudy SPIIRAN – SPIIRAS Proceedings*, 2016, no. 1 (44), pp. 181–197.

11. Kotov Yu.A. Approksimatsiya raspredelenii chastot bukvennykh bigramm teksta dlya identifikatsii bukv [Approximation of distributions of text characters bigrams frequencies for alphabetic characters identification]. *Trudy SPIIRAN – SPIIRAS Proceedings*, 2017, no. 1 (500, pp. 190–208.

12. Babenko L.K., Ishchukova E.A., Maro E.A., Sidorov I.D., Kravchenko P.P. Razvitie kriptograficheskikh metodov i sredstv zashchity informatsii [Development of cryptographic methods and information security tools]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki – Izvestiya Southem Federal University. Engineering sciences*, 2012, no. 4, pp. 40–50.

13. Babenko L.K., Ishchukova E.A. Analiz simmetrichnykh kriptosistem [Analysis of symmetric cryptosystems]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki – Izvestiya Southem Federal University. Engineering sciences*, 2012, no. 12, pp. 136–147.

14. Glukhov M.M., Kruglov I.A., Pichkur A.B., Chertmushkin A.V. *Vvedenie v teoretiko-chislovye metody kriptografii* [Introduction to number-theoretic methods in cryptography]. St. Petersburg, Lan' Publ., 2011. 400 p.

15. Mineev M.P., Chubarikov V.N. *Lektsii po arifmeticheskim voprosam kriptografii* [Lectures on arithmetic cryptography]. Moscow, Popechitel'skii sovet Mekhaniko-matematicheskogo fakul'teta MGU im. M.V. Lomonosova Publ., 2010. 186 p.

16. SambasivaRao Baragada, Satyanarayana Reddy P. A survey of cryptanalytic works based on Genetic Algorithms. *International Journal of Emerging Trends & Technology in Computer Science*, 2013, vol. 2, iss. 5, pp. 18–22.

17. *Amrit Pal Singh, Pal S.K., Bhatia M.P.S.* The firefly algorithm and application in cryptanalysis of monoalphabetic substitution ciphers. *American Journal of Computer Science and Engineering Survey*, 2013, vol. 1, no. 1, pp. 33–52.

18. Morozenko V.V., Pleshkova I.Yu. O primenenii geneticheskogo algoritma dlya kriptoanaliza shifra Tritemiya–Belazo–Vizhenera [On the application of a genetic algorithm for cryptanalysis of the cipher Triteria-Belazo-Vigenère]. *Sovremennye problemy nauki i obrazovaniya – Modern problems of science and education*, 2014, no. 2, pp. 1–11.

19. Aditi Bhateja, Shailender Kumar, Ashok K. Bhateja. Cryptanalysis of vigenere cipher using particle swarm optimization with Markov chain random walk. *International Journal on Computer Science and Engineering*, 2013, vol. 5, no. 5, pp. 422–429.

20. Mohan M., Kavitha Devi M.K., Jeevan Prakash V. Security analysis and modification of classical encryption scheme. *Indian Journal of Science and Technology*, 2015, vol. 8, no. 8, pp. 542–548.