ISSN 1814-1196
Научный вестник НГТУ
том 78, № 1, 2020, с. 87–106

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 78, No. 1, 2020, pp. 87–106

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК519.6

DOI: 10.17212/1814-1196-2020-1-87-106

Значения некоторых частотных характеристик англоязычных текстов*

Ю.А. КОТОВ[a], О.В. САНИНА[b]

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет

[a] *kotov@corp.nstu.ru* [b] *lyalysa@gmail.com*

Для решения многих задач формального анализа текстов требуются известные значения различных частотных характеристик текстов. Это связано с тем, что математической основой такого анализа является выбор или построение критериев сравнения определенных характеристик, изменения которых в общем случае носят случайный характер. Известно, что необходимым условием для построения таких критериев и его обоснования является наличие теоретических или выборочных распределений требуемых величин. В то же время доступные значения частотных характеристик для англоязычных текстов в значительной степени являются неполными, неточными или устаревшим, что не позволяет выстраивать их анализ в соответствии с математическими требованиями. В работе приведены результаты измерений в зависимости от объемов англоязычных текстов их основных частотных характеристик: частоты появления пробела и первых двух значащих знаков, индекса совпадений, количества используемых в текстах букв, буквенных биграмм и диграмм и связанных с ними характеристик – индексов отклонения и сопряжения. Измерения проведены на двух представительных выборках из научно-технических и художественных текстов, каждая из которых включала в себя более 2100 фрагментов текстов различного объема – от 200 до 350 000 знаков. Выборки формировались случайным образом из корпуса англоязычных текстов, включавшего в себя 491 текст. Результаты представлены в виде выборочных распределений указанных частотных характеристик, содержащих среднее, минимальное и максимальное значения и соответствующее стандартное отклонение. Проведен анализ полученных распределений и их сравнение с аналогичными характеристиками русскоязычных текстов.

Ключевые слова: текст, знак, частота встречаемости, мощность алфавита, индекс совпадений, биграмма, диграмма, индекс отклонения, индекс сопряжения

ВВЕДЕНИЕ

Известные значения частотных характеристик текстов и их зависимость от объема текста необходимы при решении многих задач автоматизированной обработки текстов на естественных языках [1–11] (например, при восста-

* *Статья получена 07 февраля 2020 г.*

новлении букв текста в неизвестной кодировке, при определении авторства, размера знаковой кодировки, языка текста и т. д.). Доступные значения для англоязычных текстов в значительной степени являются неполными, неточными или устаревшими [12–18]. В этой связи был проведен вычислительный эксперимент на двух представительных выборках англоязычных текстов (художественных текстов и текстов научно-технической направленности), результаты которого представлены в работе. Рассмотрены следующие униграммные и биграммные характеристики англоязычных текстов: зависимость количества используемых букв и их сочетаний (биграмм и диграмм) от объема текста, частота пробела и первых двух значащих символов в частотном упорядочении, значения индекса совпадения и основанных на биграммных и диграммных характеристиках индексов сопряжения и отклонения.

1. ОПИСАНИЕ ВЫБОРОК, ИСПОЛЬЗУЕМЫХ ДЛЯ ИЗМЕРЕНИЙ

Измерения проведены в диапазоне $200 \le x \le 350000$ ($x \in N$ – объем текста в знаках), разбитом на 4 интервала. В каждом интервале определена своя шкала измерений, представленная в табл. 1, где K – количество текстов объема x.

Для проведения измерений был составлен корпус из 491 текста, включающий в себя работы на английском языке двух стилей: художественные (современные и классические романы различных жанров и авторов) и научные (учебники, регулярные научные журналы, труды конференций, диссертации различных областей знаний). Из корпуса сформированы две выборки – тексты 1 и 2, параметры которых приведены в табл. 1.

Таблица 1

Table 1

Характеристика выборок, используемых для измерений

Profile of the samples used for measurements

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	83	99
200	100	100	2000	100	100	110 000	78	98
400	100	100	4000	99	100	Всего 3:	535	594
600	100	100	6000	98	100	Группа 4		
800	100	100	8000	98	100	100 000	30	28
1000	100	100	10 000	98	100	150 000	28	28
1200	100	100	Всего 2:	493	500	200 000	25	27
1400	100	100	Группа 3			250 000	25	25
1600	99	100	10 000	100	100	300 000	22	24
1800	99	100	30 000	95	99	350 000	19	23
2000	99	100	50 000	91	99	Всего 4:	149	155
Всего 1:	997	1000	70 000	88	99	Итого:	2174	2249

Для первых трех интервалов в каждой выборке было выделено по 100 случайных текстовых фрагментов, для последнего интервала – 30. 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. Характеристики выборок аналогичны выборкам русскоязычных текстов 1 и 2, приведенным в [19, 20].

Исходные тексты были предварительно нормализованы: в них оставлены только буквы языка, приведенные к одному регистру ($N_A = 26$ – общее количество букв) и один знак пробела на каждое слово; другие знаки из текстов были исключены. Используемые в текстах дефис и апостроф (как, например, *state-of-the-art*, *aren't*) исключались, а соединенные ими слова или части слов объединялись в одно слово.

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

Для оценки частоты появления текстов $P(x)$, использующих не все буквы алфавита, воспользуемся формулой

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

где $K(x)$ – количество текстов объемом x; $K_1(x)$ – количество текстов объемом x, в которых используются *не все* буквы алфавита [19].

Пусть $Z(x)$ – среднее количество используемых букв алфавита в текстах $K_1(x)$. При этом если $K_1(x) = 0$, то $Z(x) = N_A = 26$.

Результаты измерений на интервале $200 \le x \le 30\,000$ приведены в табл. 2, где S.D. Z – стандартное отклонение значения $Z(x)$. На интервале $x \ge 30\,000$ $K_1(x) = 0$, $P(x) = 0$, $Z(x) = N_A = 26$. «10Т» в столбце «x» означает 10 000 и т. д.

Таблица 2

Table 2

Количество текстов, использующих не все буквы, и количество используемых в них букв алфавита *Z*

The number of texts that don't use all letters and the number of the used letters of the alphabet *Z*

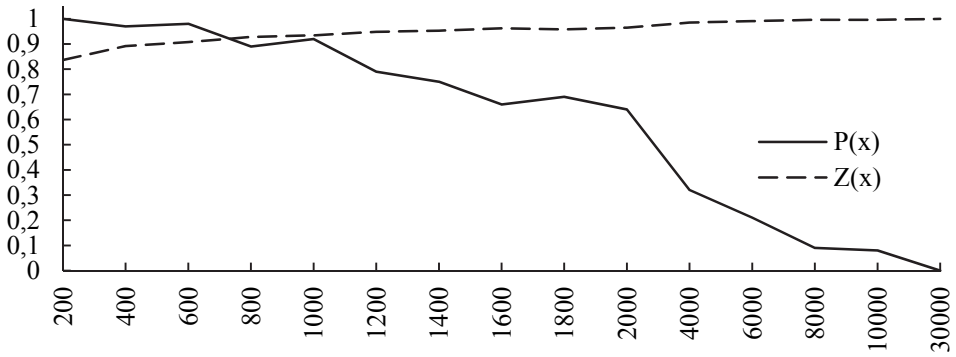
x	Тексты 1						Тексты 2					
	K	K_1	Z	мин	макс	S.D.	K	K_1	Z	мин	макс	S.D.
Группа 1												
200	100	100	21,90	18	24	1,03	100	100	21,75	19	25	1,20
400	100	98	23,24	21	26	1,04	100	97	23,18	20	26	1,23
600	100	96	23,71	22	26	0,96	100	98	23,60	21	26	1,04
800	100	93	24,04	22	26	0,98	100	89	24,12	21	26	1,17
1000	100	91	24,28	22	26	0,91	100	92	24,30	21	26	1,03
1200	100	88	24,42	22	26	0,92	100	79	24,66	23	26	0,93
1400	100	84	24,60	22	26	0,88	100	75	24,78	22	26	0,94
1600	99	80	24,75	22	26	0,89	100	66	25,02	23	26	0,86
1800	99	75	24,91	23	26	0,83	100	69	24,91	22	26	0,96
2000	99	70	25,04	23	26	0,79	100	70	24,98	23	26	0,84

Окончание табл. 2

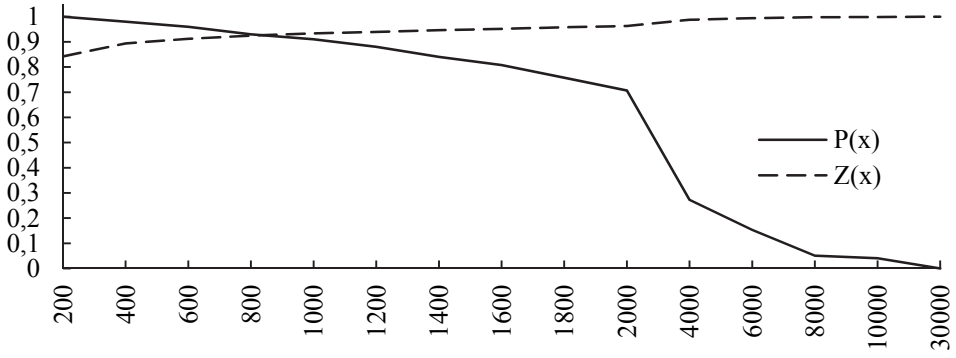
End of Tab. 2

x	Тексты 1						Тексты 2					
	K	K_1	Z	мин	макс	S.D.	K	K_1	Z	мин	макс	S.D.
Группа 2												
2000	100	70	24,88	23	26	0,93	100	64	25,08	23	26	0,85
4000	99	27	25,69	24	26	0,54	100	32	25,62	23	26	0,61
6000	98	15	25,85	25	26	0,36	100	21	25,76	24	26	0,49
8000	98	5	25,95	25	26	0,22	100	9	25,89	24	26	0,37
10Т	98	4	25,96	25	26	0,20	100	8	25,90	24	26	0,36
Группа 3												
10Т	100	5	25,95	25	26	0,22	100	10	25,88	24	26	0,38
30Т	95	0	26	26	26	0	99	0	26	26	26	0

Графики аппроксимированных значений $P(x)$ и нормированных к N_A значений $Z(x)$ для текстов 1 и 2 приведены на рис. 1. Шкала по оси X здесь и далее неравномерная.



а



б

Рис. 1. Количество текстов $P(x)$, использующих не все буквы языка, и среднее количество используемых в них букв $Z(x)$ в текстах 2 (*а*) и текстах 1 (*б*)

Fig. 1. The number of texts $P(x)$ that don't use all letters and the average number of the used letters of the alphabet $Z(x)$ in Texts 2 (*a*) and Texts 1 (*b*)

По сравнению с русским языком [19] с ростом объема текстов количество фрагментов англоязычных текстов, использующих не все буквы алфавита, сокращается медленнее. Количество таких фрагментов составляет около 70 % от их общего числа при объеме текстов в 2000 знаков. Для русскоязычных текстов такое значение достигается при объеме в 400–600 знаков [19]. При объеме текстов в 4000 знаков в русском языке такие фрагменты практически не встречаются, в то время как в английском языке их количество составляет 20…30 % от общего числа. Существенное снижение их количества в английском языке (на 40 % и 30 % для текстов 1 и 2 соответственно) приходится на увеличение объема текстов с двух до четырех тысяч знаков. При увеличении объема текстов с 200 до 2000 знаков количество таких текстов уменьшается с 100 % до примерно 70 % для обеих выборок. Начиная с точки $x = 30\,000$ знаков в англоязычных текстах используются все буквы алфавита.

Мощность используемого алфавита является одним из первых индикаторов принадлежности текста к какому-либо естественному языку, поэтому полученные оценки частоты используемых букв могут применяться в задачах по распознаванию языка текста [11, 20].

3. ОЦЕНКА ЧАСТОТЫ ПОЯВЛЕНИЯ ПРОБЕЛА В ТЕКСТАХ

Поскольку пробел является одним из разделителей слов, определение его наличия и последующая идентификация имеют важное значение при решении задач обработки текста. Известно, что в русском языке в текстах вплоть до объема $x = 1200$ знаков пробел может занимать не первое место в частотном упорядочении, а в точке $x = 1200$ количество таких текстов достигает 18 % [19]. В связи с этим необходимо оценить частоту появления не только пробела в англоязычных текстах, но и первых двух значащих символов в частотном упорядочении. Результаты таких измерений для текстов 1 приведены в табл. 3, а для текстов 2 – в табл. 4, где $C_0(x)$ – среднее количество пробелов в текстах объемом x; $C_1(x)$ – среднее количество появлений в текстах объемом x знака c_1, отличного от пробела и первого в частотном упорядочении; $C_2(x)$ – среднее количество появлений в текстах объемом x знака c_2, следующего по убыванию в частотном упорядочивании за знаком c_1; SD $C_i(x)$ – стандартное отклонение $C_i(x)$, $i = 0, 1, 2$. Важно отметить, что знаками c_1 и c_2 для каждого текста x могут быть любые буквы алфавита.

Из табл. 3 и 4 видно, что частота появления всех трех знаков является достаточно стабильной величиной, однако средняя частота появления пробела для выборки 1 больше, чем для выборки 2. Вычислив общие нормированные средние c_i, $i = 0, 1, 2$, и взяв минимальные и максимальные нормированные значения $C_i(x)$, получим оценку, представленную в табл. 5.

Таблица 3

Table 3

Частота появления пробела и первых двух отличных от пробела знаков в текстах 1

Occurrence frequency of the space and two first letters in the occurrence order different from the space character in Texts 1

x	C_0	мин	макс	S.D.	C_1	мин	макс	S.D.	C_2	мин	макс	S.D.
Группа 1												
200	36,24	25	44	3,34	20,98	16	32	3,19	16,74	13	23	2,02
400	73,46	56	86	5,28	41,48	29	54	5,37	32,05	26	42	3,32
600	110,00	90	122	6,73	61,75	47	77	6,50	47,26	38	58	4,29
800	146,72	122	165	8,41	82,88	63	106	8,70	62,29	53	74	5,00
1000	183,97	154	206	10,03	103,64	77	129	10,33	77,68	65	94	6,23
1200	221,13	187	244	11,22	124,13	96	153	11,92	92,63	79	110	7,27
1400	258,08	220	282	12,61	145,20	113	173	12,77	107,28	91	126	7,76
1600	295,15	254	319	13,97	165,36	127	193	13,93	121,62	105	145	8,81
1800	332,05	290	358	14,79	185,81	145	215	15,26	136,32	118	158	9,50
2000	369,11	323	396	16,08	206,78	165	245	16,92	151,06	129	172	9,52
Группа 2												
2T	377	322	415	18	203	156	249	17	149	122	185	13
4T	748	689	838	30	410	330	478	29	294	251	346	19
6T	1120	1028	1241	44	614	535	702	35	438	382	504	24
8T	1496	1371	1638	55	820	725	960	45	581	519	651	29
10T	1870	1710	2034	68	1025	924	1189	53	723	648	802	35
Группа 3												
10T	1853	1603	2093	72	1026	871	1170	59	724	644	835	34
30T	5557	4837	6082	209	3077	2713	3471	149	2160	1960	2390	96
50T	9293	8493	10 101	310	5121	4505	5873	242	3609	3264	3986	156
70T	13 027	11 953	14 134	432	7175	6328	8176	329	5036	4534	5589	216
90T	16 745	15 398	18 084	548	9239	8202	10 433	422	6484	5882	7160	273
110T	20 483	18 657	22 088	659	11 300	9965	12 821	514	7907	7187	8669	323
Группа 4												
100T	18 630	16 992	20 152	661	10 264	8977	11 242	470	7187	6480	7745	269
150T	27 916	25 654	30 243	1027	15 389	13655	16 570	682	10 753	9678	11749	407
200T	37 332	35 177	40 225	1257	20 382	18197	21 900	835	14 346	13 753	15178	382
250T	46 671	44 055	50 301	1555	25 459	22877	27 135	1042	17 925	17 108	18758	464
300T	55 902	52 907	60 255	1840	30 789	28402	32 380	1070	21 489	20 397	22484	617
350T	65 387	62 360	70 210	2077	35 765	33189	37 737	1274	25 026	23 829	26307	748

Таблица 4

Table 4

Частота появления пробела и первых двух отличных от пробела знаков в текстах 2

Occurrence frequency of the space and two first letters in the occurrence order different from the space character in Texts 2

x	C_0	мин	макс	S.D.	C_1	мин	макс	S.D.	C_2	мин	макс	S.D.
Группа 1												
200	31,04	22	41	3,58	22,28	15	33	3,47	17,15	14	22	1,74
400	60,71	40	77	6,55	44,02	33	62	6,04	34,20	27	49	4,01
600	92,47	72	113	8,56	64,56	51	83	7,23	50,30	41	62	4,54
800	124,37	97	156	10,43	86,61	69	111	9,13	65,22	52	86	6,00

Окончание табл. 4

End of Tab. 4

x	C_0	мин	макс	S.D.	C_1	мин	макс	S.D.	C_2	мин	макс	S.D.
1000	154,36	118	177	13,37	104,91	82	133	11,22	81,93	67	103	7,13
1200	185,97	144	229	17,71	126,69	94	188	14,49	98,33	79	128	9,38
1400	216,72	166	259	19,63	149,03	121	187	15,24	113,01	90	141	9,98
1600	245,38	203	290	20,27	169,64	131	208	16,78	130,43	109	162	11,26
1800	275,45	232	326	21,05	190,93	145	240	18,88	145,10	119	177	12,91
2000	310,23	246	379	26,09	211,19	156	279	19,87	161,62	137	201	14,47
Группа 2												
2T	312	251	413	31	213	171	284	23	161	136	214	14
4T	630	498	800	53	421	344	513	38	320	261	400	26
6T	939	804	1134	77	630	494	742	56	479	414	574	35
8T	1262	1074	1569	94	835	700	953	58	641	543	757	47
10T	1573	1301	1962	124	1048	745	1289	88	790	648	946	59
Группа 3												
10T	1552	1181	1827	126	1047	860	1395	82	791	666	1004	62
30T	4647	3948	5513	341	3153	2734	3592	191	2377	2060	2893	164
50T	7734	6813	9039	504	5240	4451	6338	317	3955	3482	4733	257
70T	10 841	9642	12 785	736	7308	6415	8205	384	5483	4836	6246	327
90T	13 984	12 537	16 494	940	9388	8326	10 440	509	7019	6310	8006	391
110T	17 024	15 169	19 908	1166	11 537	10 242	13 867	670	8547	7539	9747	499
Группа 4												
100T	15 502	14 087	18 188	1143	10 268	9340	11 114	485	7770	7069	8987	457
150T	23 162	20 684	27 299	1639	15 509	13 909	16 890	711	11 675	10 770	12 816	609
200T	31 048	27 581	36 173	2328	20 564	18 511	22 694	931	15 583	14 394	17 396	807
250T	38 938	35 439	44 976	2767	25 996	23 547	28 067	1062	19 367	17 861	21 297	993
300T	46 498	41 808	54 370	3579	31 118	28 168	33 379	1379	23 316	21 251	26 161	1293
350T	54 574	49 518	63 581	4146	36 216	32 786	38 902	1338	27 117	25 023	30 157	1379

Таблица 5

Table 5

Нормированные оценки частоты появления пробела и первых двух отличных от пробела знаков

Normalized estimates of the occurrence frequency of the space and two first letters in the occurrence order different from the space character

Знак	Среднее C_i	min C_i	max C_i	Среднее C_i	min C_i	max C_i
	Тексты 1			Тексты 2		
C_0, пробел	0,1855	0,1250	0,2200	0,1551	0,1000	0,2065
C_1	0,1028	0,0725	0,1600	0,1054	0,0745	0,1650
C_2	0,0744	0,0610	0,1150	0,0800	0,0643	0,1225

Из табл. 5 видно, что нормированные оценки средней частоты для C_1 и C_2 в выборках 1 и 2 близки, в то время как для пробела C_0 существенно различаются, что можно объяснить увеличением средней длины слов в текстах 2.

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

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

Тогда, используя значения $C_0(x)$ из табл. 5, получим значения по средней длине слов для англоязычных текстов, которые приведены в табл. 6 ($\max C_i - \min L$, $\min C_0 - \max L$).

Таблица 6

Table 6

Оценка средней длины слов в текстах

Estimation of the average length of words in the texts

Выборка	Среднее L	$\min L$	$\max L$
Тексты 1	4,39	3,55	7
Тексты 2	5,45	3,84	9

В текстах 2 были отмечены фрагменты, в которых пробел занимает не первое место в частотном упорядочении знаков текста. Результаты приведены в табл. 7, где $K_2(x)$ – количество таких текстов. Во всех указанных случаях на первом месте оказалась буква Е, на третью позицию пробел не перемещался. В текстах 1 таких текстовых фрагментов не было обнаружено.

Таблица 7

Table 7

Количество текстов выборки 2, в которых пробел занимает не первую позицию в частотном упорядочении знаков по убыванию

The number of texts in Sample 2, where the space character does not take the first place in the decreasing frequency order

x	200	400	600	800	1000
K	100	100	100	100	100
K_2	4	0	0	0	1

4. ОЦЕНКА ИНДЕКСА СОВПАДЕНИЯ

Оценкой формы частотного распределения знаков текста является индекс совпадения, вычисляемый по формуле

$$I_C(x) = \frac{\sum_{i=1}^{N_A} C_i(x)[C_i(x)-1]}{x(x-1)}, \qquad (3)$$

где $C_i(x)$ – число вхождений знака C_i в текст объемом x знаков.

Для уточнения средних значений индекса совпадения (3), а также определения его минимальных и максимальных значений и стандартного отклонения проведены измерения для англоязычных текстов, результаты которых представлены в табл. 8. Кусочно-линейная аппроксимация значений $I_C(x)$ из табл. 8 приведена на рис. 2.

Из рис. 2 и табл. 8 видно, что среднее значение индекса совпадения стабильно во всём диапазоне измерений и составляет 0,0651 для текстов 1 и 0,0665 для текстов 2. Наибольшие и наименьшие значения индекса 0,0854 и 0,0550 и максимальное значение стандартного отклонения 0,0052 находятся в интервале $x \leq 800$ знаков. При $x > 800$ знаков стандартное отклонение монотонно снижается, а минимальные и максимальные значения индекса совпадения приближаются к средним значениям.

Таблица 8

Table 8

Значения индекса совпадений

Values of the coincidence index

Тексты 1					Тексты 2				
x	I_C	мин	макс	S.D.	x	I_C	мин	макс	S.D.
Группа 1									
200	0,0645	0,0572	0,0820	0,0047	200	0,0659	0,0550	0,0788	0,0052
400	0,0652	0,0578	0,0753	0,0035	400	0,0674	0,0599	0,0854	0,0047
600	0,0652	0,0591	0,0722	0,0028	600	0,0669	0,0603	0,0780	0,0034
800	0,0654	0,0586	0,0731	0,0029	800	0,0670	0,0601	0,0753	0,0031
1000	0,0655	0,0584	0,0741	0,0027	1000	0,0666	0,0598	0,0770	0,0033
1200	0,0654	0,0594	0,0722	0,0025	1200	0,0668	0,0582	0,0791	0,0034
1400	0,0654	0,0600	0,0708	0,0023	1400	0,0668	0,0601	0,0733	0,0029
1600	0,0654	0,0596	0,0702	0,0022	1600	0,0671	0,0611	0,0739	0,0026
1800	0,0654	0,0597	0,0708	0,0021	1800	0,0668	0,0612	0,0726	0,0024
2000	0,0654	0,0599	0,0698	0,0020	2000	0,0668	0,0605	0,0744	0,0027
Группа 2									
2000	0,0650	0,0614	0,0703	0,0019	2000	0,0670	0,0592	0,0761	0,0029
4000	0,0651	0,0610	0,0702	0,0017	4000	0,0665	0,0613	0,0717	0,0023
6000	0,0651	0,0615	0,0699	0,0015	6000	0,0668	0,0617	0,0738	0,0025
8000	0,0651	0,0620	0,0699	0,0015	8000	0,0666	0,0622	0,0707	0,0018
10 000	0,0650	0,0624	0,0703	0,0015	10 000	0,0665	0,0589	0,0714	0,0024
Группа 3									
10 000	0,0651	0,0617	0,0700	0,0014	10 000	0,0667	0,0621	0,0771	0,0024
30 000	0,0650	0,0615	0,0679	0,0012	30 000	0,0665	0,0607	0,0712	0,0017
50 000	0,0650	0,0613	0,0680	0,0012	50 000	0,0666	0,0628	0,0721	0,0016
70 000	0,0650	0,0612	0,0678	0,0012	70 000	0,0663	0,0622	0,0695	0,0014
90 000	0,0650	0,0613	0,0678	0,0012	90 000	0,0662	0,0625	0,0706	0,0016
110 000	0,0650	0,0613	0,0678	0,0012	110 000	0,0663	0,0624	0,0710	0,0016
Группа 4									
100 000	0,0649	0,0621	0,0667	0,0011	100 000	0,0657	0,0632	0,0678	0,0012
150 000	0,0649	0,0625	0,0670	0,0011	150 000	0,0658	0,0631	0,0684	0,0013
200 000	0,0648	0,0624	0,0668	0,0010	200 000	0,0658	0,0632	0,0682	0,0012
250 000	0,0647	0,0625	0,0665	0,0009	250 000	0,0658	0,0638	0,0685	0,0014
300 000	0,0649	0,0635	0,0663	0,0008	300 000	0,0658	0,0634	0,0685	0,0014
350 000	0,0648	0,0635	0,0663	0,0008	350 000	0,0657	0,0636	0,0681	0,0013

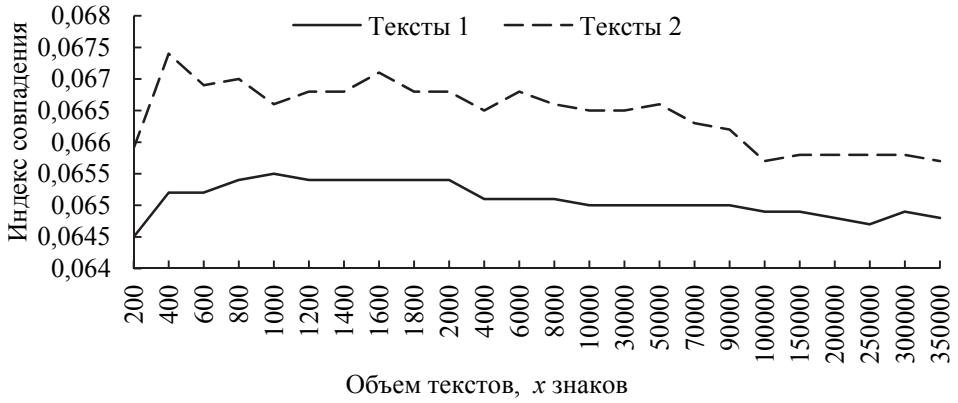


Рис. 2. Средние значения индекса совпадения для англоязычных текстов

Fig. 2. Average values of the coincidence index for English-language texts

Для использования индекса совпадения в качестве критерия для определения англоязычных текстов необходимо установить доверительные интервалы, в которые бы попадало не менее 95 % анализируемых текстов. Поскольку среднее значение индекса совпадений достаточно стабильно, разделим шкалу измерений на 4 интервала на основе значений стандартного отклонения S.D. В каждом интервале определим минимальное и максимальное значение индекса, которое уменьшим и увеличим на величину, равную двум S.D. Тогда получим следующие интервалы:

$$200 \le x < 600, \qquad 0{,}0541 < I_C(x) < 0{,}0778;$$

$$600 \le x < 1400, \qquad 0{,}0584 < I_C(x) < 0{,}0738;$$

$$1400 \le x < 30\,000, \qquad 0{,}0592 < I_C(x) < 0{,}0729;$$

$$30\,000 \le x \le 350\,000, \qquad 0{,}0613 < I_C(x) < 0{,}0700.$$

Определим количество фрагментов в текстах 1 и 2, значения индекса совпадения для которых не попадают в выделенные доверительные интервалы, и найдем их отношение к общему числу фрагментов точки шкалы измерений – $Q(x)$. Результаты измерений приведены в табл. 9.

Таблица 9

Table 9

Проверка доверительных интервалов

Verificaton of confidence intervals

Интервал	x	Q	Интервал	x	Q	Интервал	x	Q
$200 \le x < 600$	200	0,02	$1400 \le x < 30\,000$	1800	0	$30\,000 \le x < 350\,000$	70 000	0,0053
	400	0,01		2000	0,005		90 000	0,0055
$600 \le x < 1400$	600	0,02		4000	0		100 000	0
	800	0,01		6000	0,0101		150 000	0
	1000	0,015		8000	0		200 000	0
	1200	0,01		10 000	0,0051		250 000	0
$1400 \le x < 30\,000$	1400	0,015	$30\,000 \le x < 350\,000$	30 000	0,0206		300 000	0
	1600	0,01		50 000	0,0053		350 000	0

Во всех точках шкалы измерений $Q(x) < 0{,}05$, следовательно, при попадании значения индекса совпадений в выделенные интервалы в качестве начального приближения можно принять гипотезу о англоязычном тексте с доверительной вероятностью 95 % [20]. Для окончательного определения языка текста необходимо знать аналогичные данные для текстов на других языках и *не текстов* [19].

Для англоязычных текстов, использующих 26 букв алфавита в одном регистре, следует в качестве эталонного применять значение индекса $I_C(x) = 0{,}0659,$ вычисленного без учета пробела.

5. ОЦЕНКА ЧАСТОТ БУКВЕННЫХ БИГРАММ И ДИГРАММ ТЕКСТОВ

Для частотного анализа текстов на естественном языке важно знать количество используемых биграмм B для текстов разного объема. Кроме того, необходимо получить данные о количестве диграмм D_1 и D_2, на основе которых вычисляются индекс отклонения и индекс сопряжения [20]. Под биграммами будем понимать сочетания отдельных знаков текста, под диграммами – знаки, состоящие из двух отдельных знаков [20].

Индекс отклонения количества используемых в тексте биграмм и диграмм вычисляется по формуле

$$I_{BD}(x) = 1 - \frac{\min\left[K_{D1}(x),\ K_{D2}(x)\right]}{K_B(x)}. \tag{4}$$

Индекс сопряжения диграмм D_1 и D_2 вычисляется по формуле

$$I_D(x) = \frac{\min\left[K_{D1}(x),\ K_{D2}(x)\right]}{\max\left[K_{D1}(x),\ K_{D2}(x)\right]}. \tag{5}$$

Значения индексов $I_{BD}(x)$ и $I_D(x)$ позволяют оценить мощность множества D и размер используемых в тексте знаков [20].

В табл. 10 приведены значения по количеству используемых в текстах биграмм B и диграмм D_1 и D_2, а в табл. 11 – вычисленные на их основе индекс отклонения I_{BD} и сопряжения I_D для текстов 1. Соответствующие значения для текстов 2 приведены в табл. 12 и 13. Графики кусочно-линейной аппроксимации нормированных к значению 26 × 26 = 676 количеств биграмм для текстов 2 из табл. 12 и значений $I_{BD}(x)$ из табл. 13 приведены на рис. 3.

Таблица 10

Table 10

Количество биграмм и диграмм в текстах 1

The number of bigrams and digrams in Texts 1

x	B	мин	макс	S.D.	D_1	мин	макс	S.D.	D_2	мин	макс	S.D.
Группа 1												
200	104,64	88	119	6,557	63,00	53	71	3,639	62,98	54	74	3,771
400	159,91	136	179	9,046	104,88	91	118	5,507	104,57	90	116	5,752
600	195,87	175	214	8,410	135,71	121	148	5,428	135,25	121	153	5,478

Окончание табл. 10

End of Tab. 10

x	*B*	мин	макс	S.D.	D_1	мин	макс	S.D.	D_2	мин	макс	S.D.
800	222,31	199	253	9,507	160,10	141	176	7,452	159,18	143	185	7,315
1000	243,26	220	271	9,788	180,10	160	201	7,891	179,45	159	207	8,132
1200	259,29	230	286	9,569	196,52	174	216	8,079	195,63	177	223	8,211
1400	273,94	244	295	9,876	211,37	188	232	8,382	209,66	192	233	8,356
1600	285,86	256	306	10,269	223,71	202	245	8,800	222,05	203	243	8,779
1800	296,26	273	321	9,874	234,91	216	253	8,621	233,14	209	254	9,276
2000	305,31	282	331	9,790	244,60	229	265	8,704	243,14	219	265	9,211
Группа 2												
2000	305,54	279	332	11,003	243,69	220	263	9,172	243,88	225	265	9,039
4000	362,19	336	390	10,607	308,04	284	328	9,381	307,51	287	332	9,556
6000	391,10	365	419	10,139	341,54	315	362	9,845	342,04	323	374	10,148
8000	410,15	384	436	10,425	364,57	337	390	10,097	364,56	337	394	10,615
10 T	423,67	395	447	11,133	381,02	351	407	10,591	380,56	353	407	10,426
Группа 3												
10T	424,34	389	453	11,591	381,29	351	414	10,903	381,69	353	406	11,133
30T	481,00	447	504	11,464	451,08	420	474	11,173	449,30	411	472	11,165
50T	501,91	473	525	11,432	475,34	450	496	11,067	474,55	445	502	10,814
70T	513,57	489	538	11,121	489,55	465	511	10,598	488,72	463	510	11,165
90T	521,90	496	543	11,820	499,34	473	524	11,184	498,13	470	517	10,890
110T	528,55	500	549	11,785	506,73	477	532	11,044	506,31	483	527	10,961
Группа 4												
100T	530,40	507	545	9,471	505,93	483	520	8,457	506,70	489	520	8,243
150T	543,04	517	558	9,175	521,29	492	536	9,505	521,93	503	538	8,531
200T	550,68	526	568	9,557	530,24	500	547	9,680	531,20	511	548	9,428
250T	556,32	533	572	8,871	538,04	518	555	8,829	537,64	516	554	9,173
300T	560,73	537	575	9,319	543,86	521	561	8,986	543,86	521	558	9,132
350T	564,90	542	579	9,267	549,42	528	565	9,144	548,90	525	562	8,944

Таблица 11

Table 11

Значения индексов отклонения и сопряжения для текстов 1

Values of the scatter index and the conjunction index for Texts 1

x	I_{BD}	мин	макс	S.D.	I_D	мин	макс	S.D.
Группа 1								
200	0,4105	0,3371	0,4571	0,0216	0,9578	0,8611	1,0000	0,0322
400	0,3589	0,2961	0,4379	0,0230	0,9572	0,8559	1,0000	0,0305
600	0,3216	0,2692	0,3697	0,0204	0,9611	0,8963	1,0000	0,0269
800	0,2968	0,2547	0,3451	0,0186	0,9591	0,8862	1,0000	0,0280
1000	0,2747	0,2343	0,3211	0,0190	0,9631	0,8639	1,0000	0,0263
1200	0,2572	0,2174	0,3118	0,0188	0,9650	0,8585	1,0000	0,0274
1400	0,2442	0,2134	0,2821	0,0159	0,9675	0,8929	1,0000	0,0242
1600	0,2330	0,1903	0,2680	0,0158	0,9678	0,9025	1,0000	0,0235
1800	0,2235	0,1887	0,2557	0,0160	0,9667	0,8947	1,0000	0,0250
2000	0,2146	0,1759	0,2547	0,0170	0,9671	0,9091	1,0000	0,0242
Группа 2								
2000	0,2126	0,1688	0,2484	0,0151	0,9737	0,9268	1,0000	0,0174
4000	0,1604	0,1297	0,1854	0,0126	0,9763	0,9338	1,0000	0,0163
6000	0,1359	0,1133	0,1709	0,0116	0,9777	0,9207	1,0000	0,0159
8000	0,1204	0,0964	0,1527	0,0112	0,9795	0,9282	1,0000	0,0146
10T	0,1096	0,0879	0,1401	0,0105	0,9816	0,9378	1,0000	0,0151

Окончание табл. 11

End of Tab. 12

x	I_{BD}	мин	макс	S.D.	I_D	мин	макс	S.D.
Группа 3								
10T	0,1083	0,0852	0,1372	0,0107	0,9839	0,9392	1,0000	0,0127
30T	0,0704	0,0537	0,0926	0,0082	0,9865	0,9399	1,0000	0,0114
50T	0,0591	0,0410	0,0757	0,0072	0,9886	0,9632	1,0000	0,0087
70T	0,0524	0,0349	0,0694	0,0074	0,9899	0,9660	1,0000	0,0080
90T	0,0493	0,0341	0,0681	0,0073	0,9896	0,9714	1,0000	0,0072
110T	0,0467	0,0322	0,0618	0,0068	0,9896	0,9668	1,0000	0,0076
Группа 4								
100T	0,0494	0,0383	0,0663	0,0058	0,9917	0,9750	1,0000	0,0060
150T	0,0436	0,0317	0,0555	0,0061	0,9914	0,9743	1,0000	0,0068
200T	0,0406	0,0306	0,0561	0,0066	0,9910	0,9728	1,0000	0,0071
250T	0,0368	0,0274	0,0566	0,0057	0,9926	0,9656	1,0000	0,0066
300T	0,0329	0,0200	0,0438	0,0053	0,9942	0,9839	1,0000	0,0043
350T	0,0309	0,0199	0,0452	0,0059	0,9937	0,9786	1,0000	0,0053

Таблица 12

Table 12

Количество биграмм и диграмм в текстах 2

The number of bigrams and digrams in Texts 2

x	*B*	мин	макс	S.D.	D_1	мин	макс	S.D.	D_2	мин	макс	S.D.
Группа 1												
200	101,47	67	126	10,666	63,16	50	74	5,149	62,83	48	76	5,382
400	151,91	90	181	12,879	102,62	65	119	8,023	102,20	68	120	7,236
600	183,53	139	217	15,366	131,20	101	150	9,476	131,07	97	155	10,251
800	207,71	144	253	17,019	153,21	113	178	10,803	152,99	115	179	11,340
1000	225,03	162	263	20,227	169,28	117	193	13,807	168,96	129	198	13,961
1200	241,15	157	286	20,765	186,27	139	216	15,189	184,83	134	219	14,885
1400	254,57	191	294	19,295	198,74	152	226	13,915	199,03	154	228	13,090
1600	266,42	216	312	19,548	210,32	172	243	14,836	210,69	174	241	14,510
1800	275,92	218	320	20,267	219,86	167	252	15,811	220,47	168	253	14,912
2000	285,14	205	336	21,752	230,15	171	265	16,313	228,23	173	266	15,948
Группа 2												
2000	284,79	183	338	22,164	228,68	161	260	16,392	229,68	167	266	16,390
4000	339,94	294	390	22,047	289,34	248	337	18,146	289,17	254	331	18,049
6000	368,31	300	429	25,189	322,49	256	372	22,073	321,65	262	370	21,622
8000	388,85	319	445	22,617	344,69	279	397	19,448	344,43	279	392	20,942
10T	404,34	345	471	24,082	362,24	310	427	21,516	361,26	310	417	21,633
Группа 3												
10T	404,05	311	464	25,621	362,14	270	408	22,955	361,28	277	413	22,621
30T	473,89	412	539	25,516	439,48	382	494	23,007	440,00	388	498	22,774
50T	501,25	438	562	27,212	469,05	416	522	24,093	469,94	418	522	23,889
70T	519,95	458	580	25,632	488,67	434	542	22,560	489,37	441	553	22,413
90T	534,75	470	590	27,049	506,05	447	563	24,497	506,21	446	566	24,896
110T	545,41	482	664	28,596	517,47	462	635	26,748	517,99	461	647	27,469
Группа 4												
100T	547,25	504	579	21,214	517,82	480	554	20,643	518,93	479	547	18,742
150T	563,07	510	595	23,936	535,14	481	570	22,765	537,79	489	571	23,016
200T	577,82	544	611	23,196	552,63	515	581	22,687	552,52	511	584	21,702
250T	589,68	536	633	27,708	567,04	513	610	28,536	566,16	506	605	26,702
300T	598,96	547	640	26,291	574,92	522	611	25,818	577,04	526	622	26,042
350T	602,13	547	645	27,710	582,39	531	628	27,500	581,13	531	619	27,263

Таблица 13

Table 13

Значения индексов отклонения и сопряжения для текстов 2

Values of the scatter index and the conjunction index for Texts 2

x	I_{BD}	мин	макс	S.D.	I_D	мин	макс	S.D.
Группа 1								
200	0,3917	0,1882	0,4597	0,0429	0,9520	0,8169	1,0000	0,0363
400	0,3369	0,2547	0,4012	0,0283	0,9643	0,8818	1,0000	0,0285
600	0,3001	0,1818	0,3682	0,0282	0,9570	0,8784	1,0000	0,0267
800	0,2740	0,2114	0,3365	0,0268	0,9676	0,8758	1,0000	0,0261
1000	0,2612	0,1629	0,3190	0,0244	0,9644	0,8720	1,0000	0,0273
1200	0,2422	0,1465	0,2993	0,0203	0,9685	0,8900	1,0000	0,0233
1400	0,2299	0,1401	0,2833	0,0237	0,9698	0,8973	1,0000	0,0213
1600	0,2229	0,1721	0,2695	0,0186	0,9665	0,8698	1,0000	0,0236
1800	0,2134	0,1498	0,2645	0,0195	0,9710	0,9144	1,0000	0,0210
2000	0,2079	0,1478	0,2548	0,0206	0,9698	0,8870	1,0000	0,0222
Группа 2								
2000	0,2069	0,1202	0,2470	0,0220	0,9697	0,9076	1,0000	0,0218
4000	0,1589	0,1246	0,1952	0,0144	0,9767	0,9147	1,0000	0,0180
6000	0,1340	0,1118	0,1738	0,0120	0,9806	0,9307	1,0000	0,0126
8000	0,1227	0,0914	0,1655	0,0141	0,9802	0,9186	1,0000	0,0162
10T	0,1140	0,0867	0,1584	0,0133	0,9807	0,9271	1,0000	0,0146
Группа 3								
10T	0,1118	0,0909	0,1382	0,0114	0,9844	0,9429	1,0000	0,0124
30T	0,0783	0,0541	0,1096	0,0114	0,9864	0,9400	1,0000	0,0119
50T	0,0698	0,0490	0,0885	0,0084	0,9860	0,9616	1,0000	0,0094
70T	0,0652	0,0435	0,0951	0,0102	0,9876	0,9623	1,0000	0,0087
90T	0,0580	0,0336	0,0844	0,0077	0,9904	0,9627	1,0000	0,0080
110T	0,0553	0,0350	0,0810	0,0084	0,9903	0,9580	1,0000	0,0076
Группа 4								
100T	0,0591	0,0373	0,0870	0,0104	0,9865	0,9650	0,9981	0,0100
150T	0,0520	0,0375	0,0675	0,0074	0,9900	0,9527	1,0000	0,0088
200T	0,0479	0,0337	0,0701	0,0078	0,9912	0,9755	1,0000	0,0068
250T	0,0439	0,0278	0,0560	0,0083	0,9902	0,9712	1,0000	0,0070
300T	0,0431	0,0311	0,0576	0,0069	0,9903	0,9731	0,9983	0,0069
350T	0,0373	0,0274	0,0515	0,0063	0,9929	0,9753	1,0000	0,0065

Из данных табл. 10 и 12 можно увидеть, что количество биграмм в текстах 1 меньше практически во всём диапазоне шкалы измерений, чем в текстах 2. В то же время значения индекса отклонения для текстов 1 больше, чем для текстов 2, также практически на всем диапазоне шкалы измерений. Следовательно, аппроксимированные значения для текстов 1 будут вести себя аналогично графикам рис. 3, но находиться «внутри» них.

Поведение графика средней частоты появления биграмм для английских текстов (рис. 3) схоже с аналогичным графиком для русских текстов [20]. Так, при объеме текстов в 2 тысячи знаков количество используемых биграмм и в русском, и в английском языке достигает 30 %, а при объеме в 30 тысяч знаков – 70 % от максимально возможного числа биграмм N_A^2. Отличие состоит

в $K_{B\max} = N_A^2 - K_{BZ}$, где K_{BZ} – количество запретных биграмм, т. е. сочетаний знаков, которые недопустимы по правилам языка [21, 22]. Русский язык использует 87,8 % всех биграмм [20] при объеме текстов в 350 тысяч знаков, в то время как английский язык – 95,4 % при том же объеме.

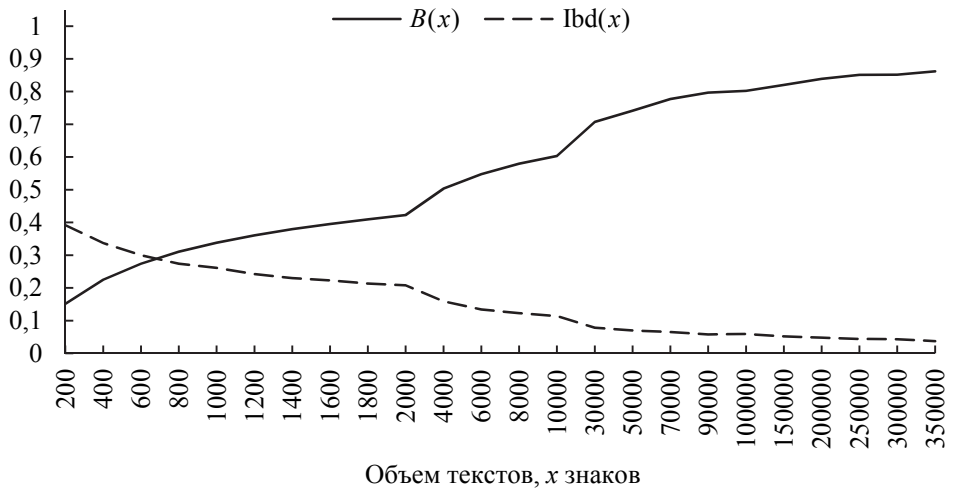


Рис. 3. Средняя частота появления биграмм и средние значения индекса отклонения для текстов 2

Fig. 3. Average occurrence frequency of bigrams and average values of the scatter index for Texts 2

Как и в случае с русскоязычными текстами, можно выделить шесть основных интервалов: $30\,000 \le x \le 350\,000$; $10\,000 \le x < 30\,000$; $4000 \le x < 10\,000$; $2000 \le x < 4000$; $800 \le x < 2000$; $200 \le x < 800$, на первых пяти из которых изменения в количестве биграмм практически линейны, на шестом начинается быстрое нелинейное уменьшение количества биграмм и их расщепление на диграммы.

ЗАКЛЮЧЕНИЕ

Материалы работы показывают сильную зависимость значений частотных характеристик текста от его объема. Эта зависимость такова, что ее нельзя игнорировать при построении точных методов анализа и моделирования текстов и их формальной оценке. Особенно это относится к задачам защиты информации, таким как стеганография, криптография и криптоанализ, идентификация / аутентификация текстов и их авторов. Например, сравнивая полученные в статье данные для англоязычных текстов с аналогичными данными для русскоязычных текстов [19, 20], можно получить ответы на следующие вопросы.

1. При объеме текста 4000 знаков для русскоязычных текстов, 30 000 знаков для англоязычных в нем будут использоваться все буквы алфавита. 2. При объеме текста более 1000 знаков можно различить англоязычные и русскоязычные тексты по числу используемых в них букв. 3. Ни при каком

объеме текста нельзя различить англоязычные и русскоязычные тексты по индексу совпадения, так как диапазоны значений индекса для таких текстов существенно пересекаются. То есть индекс совпадения можно использовать для распознавания текста и не текста, но не для различения текстов на английском и русском языках. 4. При объеме текста более 10 000 знаков можно различить англоязычные и русскоязычные тексты по количеству используемых биграмм.

Полученные в статье результаты дополняют аналогичные данные для русскоязычных текстов [19, 20]. Их сравнительный анализ позволил построить и реализовать систему критериев распознавания с высокой точностью русскоязычных и англоязычных текстов, представленных в произвольных знаковых кодировках [23].

Следует отметить также, что форма представления выборочных распределений частотных характеристик текстов в зависимости от их объемов, предложенная в статье, является необходимой и достаточной для получения известными математическими методами производных и дополнительных значений, которые могут понадобиться при том или ином формальном анализе текста. Это может быть, например, объединение средних, минимальных, максимальных значений и стандартного отклонения при необходимости получения точечных оценок; расчет относительных значений, стандартной ошибки, аппроксимация и интерполяция значений частотных характеристик в зависимости от объемов рассматриваемых текстов; получение интегральных значений и определение граничных значений различных критериев и т. д.

Английский язык является самым распространенным в мире языком. В силу этого определенные в статье зависимости частотных характеристик текстов от их объемов являются актуальными при решении широкого круга задач по анализу и обработке текстовой информации, известных в информатике под общим названием *text mining*.

СПИСОК ЛИТЕРАТУРЫ

1. *Котов Ю.А.* Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. – 2016. – № 1. – С. 181–197. – DOI: 10.15622/sp.44.11.

2. *Blondeau C.*, *Nyberg K.* Joint data and key distribution of simple, multiple, and multidimensional linear cryptanalysis test statistic and its impact to data complexity // Designs, Codes and Cryptography. – 2017. – Vol. 82, N 1. – P. 319–349. – DOI: 10.1007/s10623-016-0268-6.

3. *Williams H.* Applying statistical language recognition techniques in the ciphertext-only cryptanalysis of enigma // Cryptologia. – 2000. – Vol. 24, N 1. – P. 4–17. – DOI: 10.1080/0161-110091888745.

4. Authorship attribution on bengali literature using stylometric features and neural network / A. Islam, M. Kabir, S. Islam, A. Tasnim // Proceedings 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT 2018). – Dhaka, Bangladesh, 2018. – P. 360–363. – DOI: 10.1109/CEEICT.2018.8628106.

5. *Digamberrao K.S., Prasad R.S.* Author identification on literature in different languages: a systematic survey // Proceedings 2018 International Conference On Advances in Communication and Computing Technology (ICACCT). – Sangamner, 2018. – P. 174–181. – DOI: 10.1109/ICACCT.2018.8529635.

6. A review on playfair substitution cipher and frequency analysis attack on play-fair / N. Sharma, H. Meghwal, M. Mehta, T. Kumar // Proceedings 2nd International Conference on Trends in Electronics and Informatics (ICOEI). – Tirunelveli, 2018. – P. 1–9. – DOI: 10.1109/ICOEI.2018.8553837.

7. *Yang N., Ma-li A.D.* Modifying keyboard layout to reduce finger-travel distance // Proceedings 28th International Conference on Tools with Artificial Intelligence (ICTAI). – San Jose, CA, 2016. – P. 165–168. – DOI: 10.1109/ICTAI.2016.0034.

8. *Noraset T., Demeter D., Downey D.* Controlling global statistics in recurrent neural network text generation // Proceedings of the AAAI Conference on Artificial Intelligence. – North America, 2018. – P. 5333–5341.

9. Recurrent convolutional neural networks for text classification / S. Lai, L. Xu, K. Liu, J. Zhao // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. – North America, 2015. – P. 2267–2273.

10. *Behmer L., Crump M.* Crunching big data with finger tips: how typists tune their performance toward the statistics of natural language // Big Data in Cognitive Science. – Abindgon, UK: Taylor & Francis Group, 2017.

11. *Kotov Yu., Sanina O.* Criteria and algorithm for the Russian language text recognition based on the frequency characteristics set // 2018 XIV International Scientific-Technical Conference on Actual problems of electronic instrument engineering (APEIE): proceedings. – Novosibirsk, 2018. – P. 175–179. – DOI: 10.1109/APEIE.2018.8545877.

12. *Bourne Ch.P., Ford D.F.* A study of the statistics of letters in English words // Information and Control. – 1961. – Vol. 4, iss. 1. – P. 48–67. – DOI: 10.1016/S0019-9958(61)80036-3.

13. Interplay of bigram frequency and orthographic neighborhood statistics in language membership decision / Y. Oganian, M. Conrad, A. Aryani, H.R. Heekeren, K. Spalek // Bilingualism: Language and Cognition. – 2015. – Vol. 19, N 3. – P. 578–596. – DOI: 10.1017/S1366728915000292.

14. *Jones M.N., Mewhort D.J.K.* Case-sensitive letter and bigram frequency counts from large-scale English corpora // Behavior Research Methods, Instruments, & Computers. – 2004. – Vol. 36, N 3. – P. 388–396. – DOI: 10.3758/BF03195586.

15. *Rawlinson G.E.* Bigram frequency counts and anagram lists // Quaterly Journal of Experimental Psychology. – 1976. – Vol. 28, iss. 1. – P. 125–142. – DOI: 10.1080/14640747608400546.

16. *Rubinstein-Salzedo S.* The Vigenère Cipher // Cryptography. – Cham: Springer, 2018. – P. 41–54. – DOI: 10.1007/978-3-319-94818-8_5.

17. Analysis of four historical ciphers against known plaintext frequency statistical attack / C.W. Chuah, V.L. Samylingam, I. Darmawan, P.S.S. Palaniappan, C.F. Mohd Foozy, S.N. Ramli, J. Alawatugod // International Journal of Integrated Engineering. – 2018. – Vol. 10. – P. 183–192. – DOI: 10.30880/ijie.2018.10.06.026.

18. *Rajput N.K., Ahuja B., Riyal M.K.* A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita // Pramana. – 2019. – Vol. 92, N 4. – P. 60. – DOI: 10.1007/s12043-018-1709-8.

19. *Абденов А.Ж., Котов Ю.А., Санина О.В.* Значения некоторых униграммных характеристик русскоязычных текстов // Научный вестник НГТУ. – 2017. – № 2 (67). – С. 146–162. – DOI: 10.17212/1814-1196-2017-2-146-162.

20. *Котов Ю.А., Санина О.В.* Значения некоторых биграммных характеристик русскоязычных текстов // Вестник СибГУТИ. – 2017. – № 4. – С. 24–34.

21. Kesteren R. van, Dijkstra T. Smedt K. de. Markedness effects in Norwegian–English bilinguals: task-dependent use of language- specific letters and bigrams // The Quarterly Journal of Experimental Psychology. – 2012. – Vol. 65, N 11. – P. 2129–2154. – DOI: 10.1080/17470218.2012.679946.

22. Syllables and bigrams: orthographic redundancy and syllabic units affect visual word recognition at different processing levels / M. Conrad, M. Carreiras, S. Tamm, A.M. Jacobs // Journal of Experimental Psychology: Human Perception and Performance. – 2009. – Vol. 35, N 2. – P. 461–479. – DOI: 10.1037/a0013480.

23. *Kotov Y.A., Sanina O.V.* Recognition of English and Russian-language texts based on frequency characteristics // Proceedings of the 14 International Forum on Strategic Technology (IFOST 2019) – Tomsk, 2019. – P. 202–205.

Котов Юрий Алексеевич, кандидат физико-математических наук, доцент, доцент кафедры защиты информации Новосибирского государственного технического университета. Основное направление научных исследований – информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Имеет более 35 публикаций. E-mail: kotov@corp.nstu.ru

Санина Ольга Валерьевна, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Основное направление научных

исследований – криптография на основе теории сложности, в частности безопасность протоколов обмена ключами. Имеет более 10 публикаций. E-mail: lyalysa@gmail.com

Kotov Yuri A., Ph.D. (Phys. & Math), associate professor at the Department of Information Security, Novosibirsk State Technical University. Research interest includes information and computer security, cryptography and cryptanalysis, and software for computer systems. He is the author of more than 35 scientific papers. E-mail: kotov@corp.nstu.ru

Sanina Olga V., graduate student at the Department of Computer Engineering, Novosibirsk State Technical University. Research interest includes complexity-based cryptography, in particular security of key exchange protocols. She is the author of more than 10 scientific papers. E-mail: lyalysa@gmail.com

DOI: 10.17212/1814-1196-2020-1-87-106

Values of some frequency characteristics in english-language texts[*]

Yu.A. KOTOV[a], *O.V. SANINA*[b]

Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation

[a] *kotov@corp.nstu.ru* [b] *lyalysa@gmail.com*

Abstract

Well-known values of various frequency characteristics are required to solve a number of problems in the formal analysis of texts since the analysis is based mathematically on the choice or the establishment of criteria for comparing the specific characteristics that change randomly. Theoretical or sample distributions of the required variables are known to be necessary for establishing and proving the criteria. At the same time, it is not possible to analyse frequency characteristics according to the mathematical requirements since the currently available values of them are not complete, accurate or up-to-date. The paper provides measuring results of the basic frequency characteristics depending on sizes of English-language texts. The occurrence frequency of spaces and two first letters in the occurrence order, the coincidence index, and the number of letters being used in a text are taken into account. In addition, the number of used bigrams and digrams, as well as related characteristics namely the scatter index and the conjunction index are studied. The measurements have been taken for two representative samples of fiction and non-fiction texts. Each of the samples consists of 2100 fragments of texts varying in size from 200 up to 350 000 characters. The fragments and texts for the samples have been selected randomly from an English corpus with 491 texts. The results are presented as sample distributions of the specified frequency characteristics with calculated average, minimal and maximal values, as well as the corresponding standard deviation. The distributions obtained are analysed and compared with the same characteristics of Russian-language texts.

Keywords: text, character, occurrence frequency, alphabet capacity, coincidence index, bigram, digram, scatter index, conjunction index

REFERENCES

1. Kotov Yu.A. Determinirovannaya identifikatsiya bukvennykh bigramm v russkoyazychnykh tekstakh [Determinate identification of Russian text letter bigrams]. *Trudy SPIIRAN = SPIIRAS Proceedings*, 2016, iss. 44, pp. 181–197. DOI: 10.15622/sp.44.11. (In Russian).

[*] *Received 07 February 2020.*

2. Blondeau C., Nyberg C. Joint data and key distribution of simple, multiple, and multidimensional linear cryptanalysis test statistic and its impact to data complexity. *Designs, Codes and Cryptography*, 2017, vol. 82, no. 1, pp. 319–349. DOI: 10.1007/s10623-016-0268-6.

3. Williams H. Applying statistical language recognition techniques in the ciphertext-only cryptanalysis of enigma. *Cryptologia*, 2000, vol. 24, no. 1, pp. 4–17. DOI: 10.1080/0161-110091888745.

4. Islam A., Kabir M., Islam S., Tasnim A. Authorship attribution on bengali literature using stylometric features and neural network. *Proceedings 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT 2018)*, Dhaka, Bangladesh, 2019, pp. 360–363. DOI: 10.1109/CEEICT.2018.8628106.

5. Digamberrao K.S., Prasad R.S. Author identification on literature in different languages: a systematic survey. *Proceedings 2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, Sangamner, 2018, pp. 174–181. DOI: 10.1109/ICACCT.2018.8529635.

6. Sharma N., Meghwal H., Mehta V., Kumar T. A review on playfair substitution cipher and frequency analysis attack on playfair. *Proceedings 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, 2018, pp. 1–9. DOI: 10.1109/ICOEI.2018.8553837.

7. Yang N., Mali A.D. Modifying keyboard layout to reduce finger-travel distance. *Proceedings 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA, 2016, pp. 165–168. DOI: 10.1109/ICTAI.2016.0034.

8. Noraset T., Demeter D., Downey D. Controlling global statistics in recurrent neural network text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, North America, 2018, pp. 5333–5341.

9. Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, North America, 2015, pp. 2267–2273.

10. Behmer L., Crump M. Crunching big data with finger tips: how typists tune their performance toward the statistics of natural language. *Big Data in Cognitive Science*. Abindgon, UK, Taylor & Francis Group, 2017.

11. Kotov Yu., Sanina O. Criteria and algorithm for the Russian language text recognition based on the frequency characteristics set. *2018 XIV International Scientific-Technical Conference on Actual problems of electronic instrument engineering (APEIE): proceedings*, Novosibirsk, 2018, pp. 175–179. DOI: 10.1109/APEIE.2018.8545877.

12. Bourne Ch.P., Ford D.F. A study of the statistics of letters in English words. *Information and Control*, 1961, vol. 4, iss. 1, pp. 48–67. DOI: 10.1016/S0019-9958(61)80036-3.

13. Oganian Y., Conrad M., Aryani A., Heekeren H.R., Spalek K. Interplay of bigram frequency and orthographic neighborhood statistics in language membership decision. *Bilingualism: Language and Cognition*, 2015, vol. 19, no. 3, pp. 578–596. DOI: 10.1017/S1366728915000292.

14. Jones M.N., Mewhort D.J.K. Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior Research Methods, Instruments, & Computers*, 2004, vol. 36, no. 3, pp. 388–396. DOI: 10.3758/BF03195586.

15. Rawlinson G.E. Bigram frequency counts and anagram lists. *Quarterly Journal of Experimental Psychology*, 1976, vol. 28, iss. 1, pp. 125–142. DOI: 10.1080/14640747608400546.

16. Rubinstein-Salzedo S. The Vigenère Cipher. *Cryptography*. Cham, Springer, 2018, pp. 41–54. DOI: 10.1007/978-3-319-94818-8_5.

17. Chuah C.W., Samylingam V., Darmawan I., Palaniappan P.S.S., Foozy Mohd C.F., Ramli S.N., Alawatugod J. Analysis of four historical ciphers against known plaintext frequency statistical attack. *International Journal of Integrated Engineering*, 2018, vol. 10, pp. 183–192. DOI: 10.30880/ijie.2018.10.06.026.

18. Rajput N.K., Ahuja B., Riyal M.K. A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita. *Pramana*, 2019, vol. 92, no. 4, pp. 60. DOI: 10.1007/s12043-018-1709-8.

19. Abdenov A.Zh., Kotov Yu.A., Sanina O.V. Znacheniya nekotorykh unigrammnykh kharakteristik russkoyazychnykh tekstov [Values of some unigram frequency characteristics of Russian language texts]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Science bulletin of the Novosibirsk state technical university*, 2017, no. 2 (67), pp. 146–162. DOI: 10.17212/1814-1196-2017-2-146-162.

20. Kotov Yu.A., Sanina O.V. Znacheniya nekotorykh bigrammnykh kharakteristik russkoyazychnykh tekstov [Importance of some bigram characteristics for Russian language texts]. *Vestnik SibGUTI*, 2017, no. 4, pp. 24–34. (In Russian).

21. Kesteren R. van, Dijkstra T. Smedt K. de. Markedness effects in Norwegian–English bilinguals: task-dependent use of language- specific letters and bigrams. *The Quarterly Journal of Experimental Psychology*, 2012, vol. 65, no. 11, pp. 2129–2154. DOI: 10.1080/17470218.2012.679946.

22. Conrad M., Carreiras M., Tamm S., Jacobs A.M. Syllables and bigrams: orthographic redundancy and syllabic units affect visual word recognition at different processing levels. *Journal of Experimental Psychology: Human Perception and Performance*, 2009, vol. 35, no. 2, pp. 461–479. DOI: 10.1037/a0013480.

23. Kotov Y.A., Sanina O.V. Recognition of English and Russian-language texts based on frequency characteristics. *Proceedings of the 14 International Forum on Strategic Technology (IFOST 2019)*, Tomsk, 2019, pp. 202–205.

Для цитирования:

Котов Ю.А., Санина О.В. Значения некоторых частотных характеристик англоязычных текстов // Научный вестник НГТУ. – 2020. – № 1 (78). – С. 87–106. – DOI: 10.17212/1814-1196-2020-1-87-106.

For citation:

Kotov Yu.A., Sanina O.V. Znacheniya nekotorykh chastotnykh kharakteristik angloyazychnykh tekstov [Values of some frequency characteristics in english-language texts]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Science bulletin of the Novosibirsk state technical university*, 2020, no. 1 (78), pp. 87–106. DOI: 10.17212/1814-1196-2020-1-87-106.