УДК 519.6

Значения некоторых биграммных характеристик русскоязычных текстов

Ю. А. Котов, О. В. Санина

Для решения ряда задач анализа текстов, особенно криптографических, необходимы известные значения некоторых частотных характеристик текстов на естественном языке. В статье приведены результаты измерений в зависимости от объемов для русскоязычных текстов количества используемых в них буквенных биграмм и диграмм, а также связанных характеристик, названных индексами отклонения и сопряжения. Измерения проведены на двух представительных выборках для научно-популярных и художественных текстов и текстов учебных пособий для вузов. Показано, что буквенные биграммы и диграммы образуют различные, но математически строго связанные, элементы текстов. На основе такой связи введены индексы отклонения и сопряжения, которые позволяют отличить тексты на естественном языке от случайных текстов, определять искажения в текстах или размер кодировки букв. Биграммные частотные характеристики представляются сбалансированными по стабильности, информативности и вычислительной эффективности характеристиками текстов.

Ключевые слова: выборка, тексты, буквы, частота встречаемости, биграмма, диграмма, индекс отклонения, индекс сопряжения.

1. Введение

Круг 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, например, таких как количество используемых в текстах буквенных биграмм и их зависимости от объемов текстов. Отсутствие подобных данных не позволяет решать более сложные задачи анализа текстов, в частности, криптографические или задачи распознавания и идентификации знаков текстов.

В работе приведены результаты измерений биграммных частотных характеристик на двух представительных выборках русскоязычных текстов: научно-популярных и художественных текстов и текстов учебных пособий для вузов.

2. Описание выборок, используемых для измерений

Диапазон измерений: $200 \le x \le 350000$, где $x \in \boldsymbol{N}$ – объем текста в знаках. Диапазон разбит на 4 интервала, в каждом из которых определена своя шкала измерений, представленная в табл. 1, где K – количество текстов объема x.

Таблица 1.

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			90000	50	46
200	100	100	2000	100	99	110000	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	100000	20	8
1000	100	102	10000	100	100	150000	20	8
1200	100	106	Всего 2:	500	499	200000	20	8
1400	100	154	Группа 3			250000	20	8
1600	100	139	10000	50	46	300000	20	8
1800	100	99	30000	50	46	350000	20	8
2000	100	0	50000	50	46	Всего 4:	120	48
Всего 1:	1000	1000	70000	50	46	Итого:	1920	1823

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.

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 проводились для т.н. *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 знаков, 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 измеряемых 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

3. Оценка частот буквенных биграмм и диграмм текстов

Пусть выбранная модель анализа и обработки текста содержит множество M различных лингвистических элементов (слов, букв или других), неделимых в рамках данной модели. Тогда n-граммами называются сочетания из M по n базовых лингвистических элементов [1–8].

Для сочетаний по одному элементу используется термин «униграмма». Для сочетаний из двух элементов – «биграмма». Из определения следует, что биграммами могут называться как сочетания из двух слов, так и сочетания из двух букв. Поэтому при употреблении термина «n-граммы» необходимо указывать, к какой элементной лингвистической модели он относится (знаковой, буквенной, словесной или другой).

Для сочетаний по два знаковых (буквенных) элемента – биграмм – часто можно встретить и использование термина «диграммы» в качестве синонима. Такое использование особенно распространено в работах по криптографической обработке текстов [17, 19–20]. Однако в лингвистике термином диграмма (диграф) обозначается случай, когда отдельный лингвистический элемент представляется парой *неразделяемых* элементов, в частности знаков [1–8]. Представим, например, что в некотором алфавите появилась новая буква, для обозначения которой используются две цифры, т.е. два знака. Здесь можно говорить о размере элемента – в данном случае равном двум знакам, но не о сочетании элементов, обозначаемом термином «биграмма».

Будем различать термины «биграмма» и «диграмма». Это различие особенно важно потому, что между биграммами и диграммами существует строгая математическая связь, выраженная следующим образом.

Пусть $\mathbf{C} = (c_1, c_2, \dots, c_m)$ – вектор-строка всех различающихся знаков, используемых в тексте, $c_i \neq c_j$ для $\forall i, j$. Иными словами, это множество встречающихся в тексте разных знаков, на котором задано некоторое, в общем случае произвольное, упорядочение. Тогда с каждым знаком текста можно однозначно связать число, соответствующее позиции знака в векторе **C**. Всего из элементов множества **C** может быть образовано m^2 пар.

Разобьем текст на два множества пар соседних знаков текста: D_1 и D_2. Множество D_1 включает пары, начинающиеся с первого знака текста, множество D_2 – пары, начинающиеся со второго знака текста, и в каждом случае происходит сдвиг по тексту на 2 знака. Неполными парами можно пренебречь или, если это необходимо, рассматривать текст в качестве циклически замкнутого.

Пусть элементами множеств D_1 и D_2 будут количества повторений каждой пары знаков (i, j) в соответствующем разбиении текста d_{ij}^k, $k = \{1,2\}$, где i – номер первого знака пары в векторе **C**, j – номер второго знака пары в векторе **C.**

Представлять множества D_1 и D_2 удобно в виде квадратных матриц $\mathbf{D}_1$ и $\mathbf{D}_2$, симметрично упорядоченных в соответствии с вектором **C**. Размерность данных матриц очевидным образом равна $m \times m$. Если некоторая пара (i, j) в разбиениях $\mathbf{D}_1$ и $\mathbf{D}_2$ отсутствует, то $d_{ij}^k = 0$.

По сути дела, в каждом из двух разбиений D_1 и D_2 мы рассматриваем пару соседних знаков как один знак, т.е. диграмму, фиксируя позицию ее появления. Сочетание знаков, т.е. биграмма, от позиции в тексте не зависит, и количество появлений биграммы b_{ij} в тексте очевидным образом равно:

$$b_{ij} = d_{ij}^1 + d_{ij}^2, \tag{1}$$

или в более общем виде (2):

$$\mathbf{B} = \mathbf{D}_1 + \mathbf{D}_2 , \tag{2}$$

где **B** – матрица биграмм, упорядоченная таким же образом, что и $\mathbf{D}_1$ и $\mathbf{D}_2$.

Кроме того, мощность множества биграмм B равна мощности объединенных множеств D_1 и D_2, т.е. каждое множество диграмм D_1 и D_2 содержит общее множество D и непересекающиеся подмножества D_1–D и D_2–D соответственно. Это открывает возможности для структурного анализа текста путем сравнения мощности множеств B, D_1 и D_2.

Для этого введем индекс отклонения количества используемых в тексте биграмм и диграмм:

$$I_{BD}(x) = 1 - \frac{\min[K_{D1}(x), K_{D2}(x)]}{K_B(x)}. \tag{3}$$

Введем также индекс сопряжения диграмм D_1 и D_2, вычисляемый по формуле:

$$I_D(x) = \frac{\min[K_{D1}(x), K_{D2}(x)]}{\max[K_{D1}(x), K_{D2}(x)]}. \quad (4)$$

Значения индексов $I_{BD}(x)$ и $I_D(x)$ позволяют при анализе текста оценить мощность множества D и реальный размер используемых в тексте знаков.

Несложно предположить, что с ростом объемов текста мощность множества D будет расти и значение $I_{BD}(x)$ будет стремиться к нулю, а значение $I_D(x)$ – к единице. При уменьшении объемов текста мощность множества D будет снижаться, множество биграмм – расщепляться на все более слабо пересекающиеся множества диграмм и значение $I_{BD}(x)$ будет стремится к 0.5 при значении $I_D(x)$, приближенном к единице.

Если же знаки текста имеют размер, равный двум, например, представлены двумя цифрами, то одно из значений $K_{D1}(x), K_{D2}(x)$ будет равно m, тогда как другое будет много больше него, т.к. для текстов на естественных языках справедливо, что $K_B(x) \gg m$. Тогда $I_D(x)$ будет стремиться к нулю независимо от объемов текста, в то время как $I_{BD}(x)$ будет стремиться к единице.

В каких целях вводится двухзнаковая кодировка там, где может быть использована однозначковая из m знаков: для уменьшения количества применяемых знаков, введения избыточности для повышения надежности передачи или для маскировки (защиты) текста – здесь мы рассматривать не будем. Важно, что в таком случае размерность $\tilde{m}$ вектора **C** и множеств B, D_1 и D_2 может изменяться в пределах

$$\left[\sqrt{m}\,\right] \le \tilde{m} \le 2m\,, \quad (5)$$

где $\left[\sqrt{m}\,\right]$ – ближайшее сверху целое к значению $\sqrt{m}$.

Тогда при приближении $\tilde{m}$ к левой границе интервала (5) значения индексов будут вести себя так же, как и в случае однозначковой кодировки, при удалении от нее $I_D(x)$ будет стремиться к нулю, в то время как $I_{BD}(x)$ будет стремиться к единице. В любом случае одно из количеств диграмм – D_1 или D_2 – будет в точности равно m.

Например, если для русскоязычного текста $m = 32$ (включая пробел), $\tilde{m} = 6$, $\tilde{m} \times \tilde{m} = 36$, то $K_{D1}(x) = 32$, $K_{D2}(x) = K_B(x) = 36$ (это максимально возможное значение) и $I_{BD}(x) \approx 1 - 0{,}89 = 0{,}11$, $I_D(x) \approx 0{,}89$. Но уже при $\tilde{m} = 8$ и $\tilde{m} \times \tilde{m} = 64$ имеем $K_{D1}(x) = 32$, $K_{D2}(x) = K_B(x) = 64$ и $I_{BD}(x) = 1 - 0{,}5 = 0{,}5$, $I_D(x) = 0{,}5$.

Средние количества биграмм B, диграмм D_1 и D_2, минимальные и максимальные значения и стандартные отклонения для различных объемов текстов приведены в табл. 2 и 3 для текстов 1 и 2 соответственно. Средние значения $I_{BD}(x)$ и $I_D(x)$, минимальные и максимальные значения и стандартные отклонения приведены в табл. 4 и 5 для текстов 1 и 2 соответственно.

Графики кусочно-линейной аппроксимации нормированных к значению 31×31 = 931 количеств биграмм для текстов 1 из табл. 2 и значений $I_{BD}(x)$ из табл. 4 приведены на рис. 2.

Таблица 2.

x	B	min B	max B	SD B	D_1	min D_1	max D_1	SD D_1	D_2	min D_2	max D_2	SD D_2
Группа 1												
200	117.91	82	143	11.662	73.90	56	86	6.294	73.41	51	87	6.279
400	188.78	156	216	11.996	125.95	106	140	8.247	125.81	105	138	7.490
600	234.65	175	264	17.182	163.62	131	185	11.119	164.36	127	184	11.000
800	269.65	191	314	19.639	193.92	143	216	12.699	194.97	152	222	13.527
1000	296.42	193	338	23.365	217.80	150	245	15.548	219.62	157	252	16.602
1200	319.86	193	359	26.594	239.91	170	267	17.246	241.84	174	275	18.423
1400	341.94	220	388	26.176	260.19	190	288	17.504	261.60	189	302	18.751
1600	361.00	247	404	26.476	277.62	210	306	17.583	278.53	210	317	19.034
1800	376.40	254	422	26.743	292.58	216	323	17.781	293.76	214	332	19.614
2000	390.42	261	435	26.898	306.30	224	338	18.643	307.19	224	348	19.797
Группа 2												
2000	390.42	261	435	26.898	306.30	224	338	18.643	307.19	224	348	19.797
4000	481.73	355	520	24.859	398.21	306	430	19.846	398.93	312	444	21.151
6000	531.64	401	579	26.066	452.44	355	491	21.873	452.82	350	491	22.584
8000	564.95	421	610	25.912	490.19	382	532	21.845	489.36	373	527	22.690
10000	589.03	453	633	25.539	517.79	411	553	21.832	517.17	398	558	21.974
Группа 3												
10000	588.82	453	633	28.173	518.08	411	553	23.279	516.26	398	558	24.274
30000	682.10	552	721	24.922	631.06	501	662	25.216	631.32	510	670	23.666
50000	715.58	583	753	25.205	671.86	541	707	25.792	672.26	547	706	24.396
70000	735.78	601	773	24.995	695.92	566	726	24.981	696.88	570	725	24.473
90000	749.62	625	789	24.751	712.66	587	747	24.704	713.56	599	746	23.799
110000	760.98	637	807	25.394	725.80	609	763	24.976	726.06	610	766	24.811
Группа 4												
100000	761.60	736	785	14.541	723.95	695	750	16.397	726.60	701	749	13.847
150000	781.80	748	807	15.439	749.00	709	775	16.796	748.70	719	771	15.206
200000	794.85	767	818	15.140	764.15	729	788	16.812	763.85	729	785	15.376
250000	805.10	771	831	15.760	776.85	742	800	16.091	775.50	736	799	16.500
300000	812.50	772	838	16.815	786.30	744	811	17.129	784.10	740	806	17.233
350000	818.55	774	844	16.357	793.15	746	816	16.939	791.40	742	815	17.185

Таблица 3.

x	B	min B	max B	SD B	D_1	min D_1	max D_1	SD D_1	D_2	min D_2	max D_2	SD D_2
Группа 1												
200	109.860	77	130	10.013	66.920	51	79	5.047	66.510	50	75	5.031
400	162.260	78	194	17.043	108.470	62	127	10.297	107.970	61	132	10.386
600	202.140	126	241	18.876	140.950	104	160	10.981	141.700	109	164	10.839
800	232.460	151	275	19.138	168.900	134	205	12.319	168.620	133	194	12.043
1000	256.402	158	303	21.024	189.539	133	225	13.853	189.706	128	221	14.907
1200	275.934	191	318	21.260	208.821	147	238	15.106	209.274	154	239	14.016
1400	296.961	204	347	24.829	228.688	171	261	17.372	227.734	170	269	16.427
1600	315.978	263	365	21.448	246.647	205	281	16.364	246.216	211	276	15.790
1800	329.000	259	369	26.093	260.000	203	293	17.953	258.939	212	295	17.841
Группа 2												
2000	331.010	250	374	23.291	261.414	211	292	16.648	259.747	201	301	16.015
4000	412.680	318	474	26.425	340.540	281	381	20.472	340.970	274	389	21.564
6000	455.680	210	522	36.240	387.300	206	436	28.063	387.490	203	437	28.907
8000	494.270	415	567	25.817	427.390	344	482	22.215	425.640	352	479	21.063

Таблица 3 (продолжение).

x	B	min B	max B	SD B	D_1	min D_1	max D_1	SD D_1	D_2	min D_2	max D_2	SD D_2
10000	510.740	364	573	34.644	446.960	323	493	29.180	446.230	316	515	29.424
Группа 3												
10000	513.304	469	555	22.875	448.804	409	485	18.984	448.717	403	485	20.243
30000	623.370	577	688	21.189	571.696	522	630	19.139	572.848	532	616	18.592
50000	656.761	553	721	25.272	610.500	525	670	21.375	611.739	517	663	23.316
70000	684.696	611	728	23.802	643.000	567	689	24.138	643.978	587	680	20.462
90000	700.891	657	745	19.470	663.109	623	708	19.567	661.652	620	712	19.631
110000	715.652	667	776	23.003	677.848	633	742	23.232	677.457	632	735	21.545
Группа 4												
100000	704.500	642	756	32.365	663.375	605	705	29.069	668.375	605	716	31.567
150000	744.750	723	778	17.152	703.750	680	733	16.308	705.375	680	739	18.090
200000	758.500	717	809	23.409	720.000	678	764	21.772	724.875	683	771	22.877
250000	770.500	745	823	23.532	735.625	713	778	19.943	739.875	715	791	23.078
300000	776.500	746	809	20.488	743.875	701	784	23.966	744.875	713	775	19.439
350000	783.625	758	817	18.234	752.875	718	792	22.206	753.625	724	784	17.832

Таблица 4.

x	I_{BD}	min I_{BD}	max I_{BD}	SD I_{BD}	I_D	min I_D	max I_D	SD I_D
Группа 1								
200	0.388186	0.081395	0.491525	0.003248	0.949842	0.821918	1.000000	0.001389
400	0.345162	0.234568	0.407821	0.000850	0.962869	0.860465	1.000000	0.000843
600	0.313458	0.251429	0.373444	0.000680	0.963092	0.894410	1.000000	0.000680
800	0.290582	0.197628	0.347134	0.000501	0.966357	0.911215	1.000000	0.000599
1000	0.272988	0.130268	0.332117	0.000798	0.968849	0.897959	1.000000	0.000549
1200	0.256253	0.119171	0.332117	0.001109	0.972073	0.909774	1.000000	0.000441
1400	0.245942	0.075000	0.330000	0.001029	0.974036	0.913907	0.996528	0.000361
1600	0.238784	0.055363	0.312121	0.000902	0.974130	0.921136	1.000000	0.000326
1800	0.230081	0.077419	0.271318	0.000662	0.975151	0.926045	1.000000	0.000343
2000	0.223098	0.080247	0.261965	0.000570	0.976073	0.922840	1.000000	0.000307
Группа 2								
2000	0.223098	0.080247	0.261965	0.000570	0.976073	0.922840	1.000000	0.000307
4000	0.181675	0.126147	0.220833	0.000254	0.977969	0.919240	1.000000	0.000300
6000	0.155710	0.124000	0.181019	0.000130	0.983283	0.952586	1.000000	0.000160
8000	0.140310	0.114014	0.165468	0.000091	0.983236	0.947876	1.000000	0.000146
10000	0.128330	0.102389	0.151659	0.000087	0.984292	0.958106	1.000000	0.000120
Группа 3								
10000	0.128702	0.111712	0.151659	0.000090	0.983825	0.962366	1.000000	0.000103
30000	0.080095	0.063830	0.096866	0.000074	0.988384	0.962480	1.000000	0.000097
50000	0.065073	0.050546	0.083694	0.000058	0.991080	0.969914	1.000000	0.000045
70000	0.057447	0.042524	0.077241	0.000054	0.991795	0.975714	1.000000	0.000040
90000	0.053324	0.037889	0.065246	0.000038	0.990361	0.977465	1.000000	0.000038
110000	0.049950	0.035952	0.062417	0.000040	0.991892	0.977747	1.000000	0.000030
Группа 4								
100000	0.052675	0.044329	0.064644	0.000035	0.989624	0.980309	1.000000	0.000035
150000	0.045608	0.035443	0.053595	0.000030	0.992834	0.983139	1.000000	0.000028
200000	0.041850	0.034527	0.050781	0.000022	0.993761	0.984655	1.000000	0.000026
250000	0.038459	0.031566	0.047375	0.000018	0.994802	0.986094	1.000000	0.000014
300000	0.036555	0.028786	0.044280	0.000019	0.993935	0.987547	1.000000	0.000014
350000	0.035372	0.027446	0.043269	0.000023	0.993302	0.984944	1.000000	0.000020

Таблица 5.

x	I_{BD}	min I_{BD}	max I_{BD}	SD I_{BD}	I_D	min I_D	max I_D	SD I_D
Группа 1								
200	0.402668	0.220779	0.457627	0.001271	0.961906	0.850000	1.000000	0.000959
400	0.345155	0.153846	0.407821	0.000995	0.959946	0.848214	1.000000	0.000931
600	0.310796	0.087302	0.385892	0.001320	0.966470	0.882353	1.000000	0.000703
800	0.285401	0.119205	0.352941	0.000711	0.966006	0.890909	1.000000	0.000634
1000	0.271066	0.189873	0.313433	0.000370	0.969890	0.900474	1.000000	0.000479
1200	0.253757	0.063107	0.306397	0.000881	0.967901	0.900901	1.000000	0.000635
1400	0.241754	0.062802	0.290735	0.000785	0.970644	0.905473	1.000000	0.000384
1600	0.230289	0.152381	0.276786	0.000423	0.973183	0.911392	1.000000	0.000401
1800	0.221595	0.111969	0.263456	0.000625	0.971724	0.910314	1.000000	0.000449
Группа 2								
2000	0.223135	0.112000	0.261589	0.000386	0.972371	0.913621	1.000000	0.000408
4000	0.182707	0.138365	0.219400	0.000241	0.979124	0.941667	1.000000	0.000243
6000	0.156549	0.033333	0.194896	0.000307	0.982399	0.930295	1.000000	0.000187
8000	0.145742	0.099156	0.171429	0.000153	0.979802	0.940503	1.000000	0.000207
10000	0.132717	0.090909	0.167897	0.000153	0.983348	0.946121	1.000000	0.000162
Группа 3								
10000	0.133132	0.110664	0.164815	0.000105	0.983039	0.949474	1.000000	0.000140
30000	0.087707	0.070288	0.111801	0.000090	0.987411	0.971380	1.000000	0.000072
50000	0.074475	0.055807	0.096215	0.000065	0.989133	0.971519	1.000000	0.000066
70000	0.065740	0.053125	0.086384	0.000048	0.988195	0.962963	1.000000	0.000084
90000	0.060161	0.037681	0.085137	0.000077	0.989034	0.973602	1.000000	0.000055
110000	0.058639	0.043860	0.078804	0.000055	0.988365	0.969118	1.000000	0.000073
Группа 4								
100000	0.060250	0.050139	0.067460	0.000033	0.988537	0.978626	1.000000	0.000055
150000	0.057727	0.046512	0.068365	0.000031	0.992102	0.985816	1.000000	0.000025
200000	0.051230	0.043941	0.055624	0.000016	0.992295	0.984828	1.000000	0.000031
250000	0.046808	0.038660	0.054678	0.000031	0.990946	0.983565	0.998661	0.000026
300000	0.045285	0.034392	0.060322	0.000052	0.991967	0.983170	0.995951	0.000017
350000	0.042183	0.033877	0.052770	0.000027	0.993049	0.985274	0.998676	0.000017

Из данных табл. 2 и 3 несложно увидеть, что количество биграмм для текстов 2 меньше во всем диапазоне шкалы измерений, чем для текстов 1. В то же время значения индекса отклонения для текстов 2 больше, чем для текстов 1, также во всем диапазоне шкалы измерений. Т.е. аппроксимированные значения для текстов 2 будут вести себя аналогично графикам рис. 2, но находиться «внутри» них.

Анализируя полученные результаты, следует отметить, что при изменении объемов текстов количество используемых в них биграмм изменяется очень медленно. Так, при изменении объема текста от 3500 до 350000 знаков сам объем вырастает в 100 раз, тогда как количество используемых в текстах биграмм увеличивается только в 1.7 – 1.9 раз.

Можно выделить шесть основных интервалов: $30000 \le x \le 350000$, $10000 \le x < 30000$, $4000 \le x < 10000$, $2000 \le x < 4000$, $800 \le x < 2000$, $200 \le x < 800$, на первых пяти из которых изменения практически линейны, на шестом начинается быстрое нелинейное уменьшение количества биграмм и их расщепление на диграммы.

Наибольший средний прирост количества биграмм в текстах наблюдается при объемах от 400 до 30000 знаков, и в дальнейшем происходит их медленное насыщение до $K_{Bmax} = N_A^2 - K_{BZ}$, где K_{BZ} – количество запретных биграмм, т.е. сочетаний знаков, которые недопустимы по правилам языка.

Надо отметить, что K_{BZ} для русскоязычных текстов невелико, особенно в случаях, когда биграммы определяются для текстов без учета пробелов, начинающихся и заканчивающихся случайным образом (т.е. не с полного слова) и содержащих аббревиатуры и лексические погрешности. Поэтому в общем случае запретными биграммами можно пренебречь.

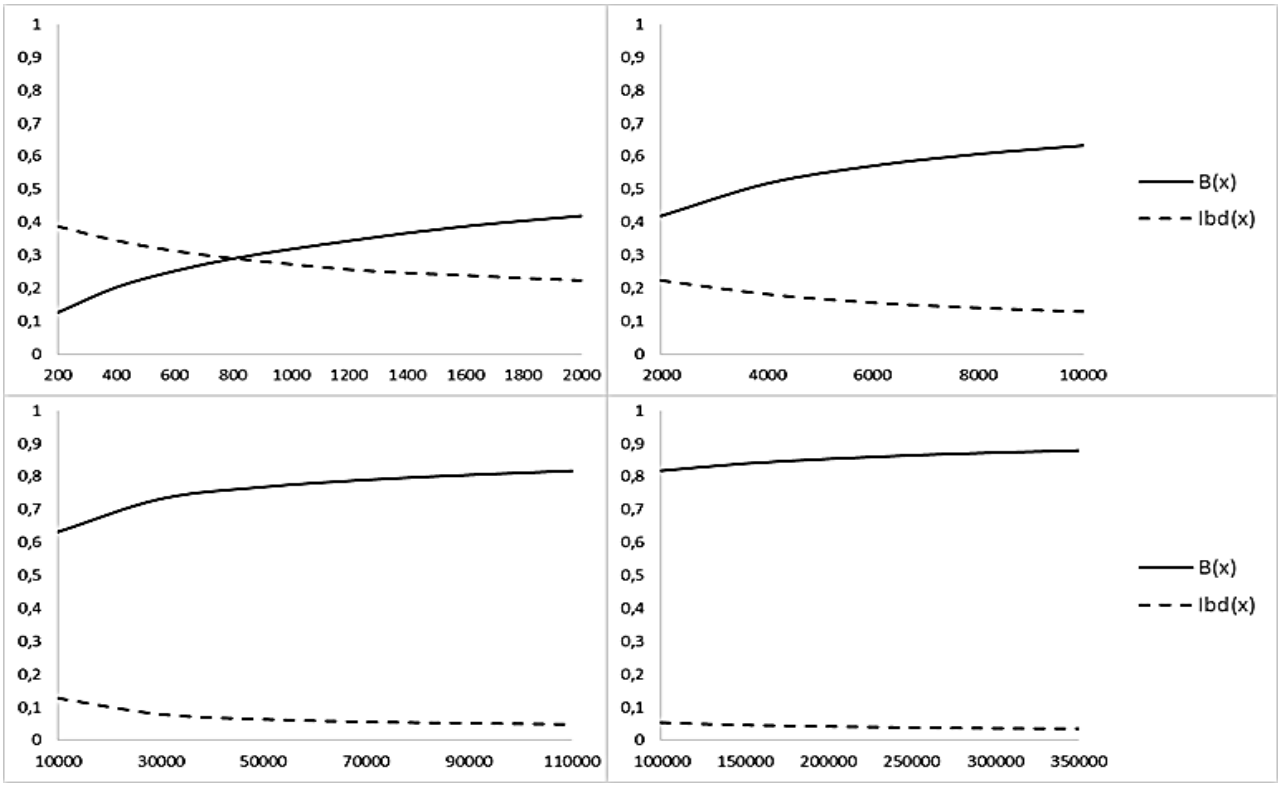

Рис. 2. Количество биграмм и значения индекса отклонения для текстов 1

4. Заключение

Максимальное количество буквенных биграмм, используемых в русскоязычных текстах, при объемах текстов в 350000 знаков достигает 817–844 единиц. Количество биграмм при изменениях объемов текстов изменяется очень медленно. Так, например, при снижении объема текста от 350000 до 4000 знаков, то есть почти на два порядка, среднее число биграмм снижается только в 1.7 – 1.9 раз. Биграммы и диграммы образуют различные, но математически строго связанные, элементы текстов. На основе такой связи можно отличить тексты на естественном языке от случайных текстов, определять искажения в текстах или размер кодировки букв.

Биграммные частотные характеристики представляются наиболее сбалансированными по стабильности, информативности и вычислительной эффективности характеристиками текстов.

Приведенные в статье результаты измерений биграммных 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.

Литература

1. *Соснина Е. П.* Введение в прикладную лингвистику. Ульяновск: Изд-во:УлГТУ, 2012.
2. *Grigori Sidorov.* Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction // International Journal of Computational Linguistics and Applications. 2013. V. 4, № 2. P. 169–188.
3. *Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández.* Syntactic N-grams as Machine Learning Features for Natural Language Processing // Expert Systems with Applications. 2013. V. 41, № 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. № 4. С. 1–3.
6. *Васильев Е. М., Гусев К. Ю.* Анализ избыточности русскоязычного текста // Вестник Воронежского государственного технического университета. 2010. № 8. С. 1–4.
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 материале 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. акад. М. Ф. Решетнева, 2008. 107 с.
10. *Котов Ю. А.* Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. 2016. № 1. С. 181–197.
11. *Котов Ю. А.* Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // Труды СПИИРАН. 2017. № 1. С. 190–208.
12. *Бабенко Л. К., Ищукова Е. А., Маро Е. А., Сидоров И. Д., Кравченко П. П.* Развитие криптографических методов и средств защиты информации // 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. *Глухов М. М., Круглов И. А., Пичкур А. Б., Черёмушкин А. В.* Введение в теоретико-числовые методы криптографии. СПб.: Лань, 2011. 400 с.
15. *Минеев М. П., Чубариков В. Н.* Лекции по арифметическим вопросам криптографии. М.: Изд-во «Попечительский совет Механико-математического факультета МГУ им. М. В. Ломоносова», 2010. 186 с.
16. *Sambasiva Rao Baragada, P. Satyanarayana Reddy.* A Survey of Cryptanalytic Works Based on Genetic Algorithms // International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 2013. V. 2, № 5. P. 18–22.
17. *Amrit Pal Singh, Dr. S K. Pal and Dr. M P S Bhatia.* The Firefly Algorithm and Application in Cryptanalysis of Monoalphabetic Substitution Ciphers // American Journal of Computer Science and Engineering Survey. 2013. V. 1, № 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: электронный научный журнал. 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 (IJCSE). 2013. V. 5, № 5. P. 422–429.
20. *Maya Mohan, M. K. Kavitha Devi, V. Jeevan Prakash.* Security Analysis and Modification of Classical Encryption Scheme // Indian Journal of Science and Technology. 2015. V. 8, № 8. P. 542–548.

Статья поступила в редакцию 22.05.2017

Котов Юрий Алексеевич
к.ф-м.н., доцент кафедры защиты информации Новосибирского государственного технического университета, email: kotov@corp.nstu.ru.

Санина Ольга Валерьевна
студент кафедры защиты информации Новосибирского государственного технического университета, email: lyalysa@gmail.com.

Importance of some bigram characteristics for Russian language texts

Yu. Kotov, O. Sanina

To solve a number of text analysis problems, especially cryptographic, the known values of some frequency characteristics of natural language texts are required. The paper provides measuring results, depending on Russian language texts sizes: the number of alphabetic bigrams and digrams used in texts, as well as connected characteristics named as scatter index and conjunction index. Measurements were taken for two samples: the first sample includes nonfiction and fiction, the second one consists of university study guides.
This paper represents alphabetic bigrams and digrams to form different but mathematically strictly related texts elements. Based on this relationship, the scatter index and the conjunction index are introduced enabling to distinguish texts in natural language from random texts, to determine misrepresentations in texts and the size of letter encoding. Bigrams frequency characteristics are submitted to be the most balanced characteristics according to stability, informativity, and computing efficiency.

Keywords: sample, texts, letters, frequency, bigram, digram, scatter index, conjunction index.