

*МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ,
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ*

УДК 004

DOI: 10.17212/2782-2230-2026-3-41-63

**РЕАЛИЗАЦИЯ КРИПТОАНАЛИЗА ШИФРА
ПРОСТОЙ ЗАМЕНЫ С ИСПОЛЬЗОВАНИЕМ
ЗНАКОВЫХ ТРИГРАММ***

Ю.А. КОТОВ¹, А.П. БАШКИРОВ²

¹ РФ, 630073, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат физико-математических наук, доцент, доцент кафедры защиты информации. E-mail: kotov@corp.nstu.ru

² РФ, 630073, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, лаборант кафедры защиты информации. E-mail: kaf_zi@corp.nstu.ru

В работе рассматривается адаптация метода Якобсена для криптоанализа шифра простой замены с использованием знаковых триграмм применительно к русскоязычным текстам. Проведена оценка устойчивости эталона: стабильности частот и рангов наиболее частотных триграмм, полноты покрытия множества ненулевых триграмм и геометрического сходства их распределения по пластинам. Для сопоставимости частот применена нормировка, экспериментально определена оптимальная стратегия нормировки в зависимости от объема анализируемого текста. Метод реализован программно и протестирован на независимой выборке научно-популярных и художественных текстов. Сравнительный анализ с биграммной реализацией показал, что триграммный подход обеспечивает существенно меньшую ошибку идентификации (вплоть до нуля при объемах свыше 6000 символов) при приемлемых вычислительных затратах. Полученные результаты позволяют количественно оценить минимальный объем шифротекста, при котором атака на основе триграмм становится результативной, и подтверждают практическую ценность триграммного эталона для статистического криптоанализа русскоязычных текстов.

Ключевые слова: триграммы, частотный анализ, шифр простой замены, метод Якобсена, эталон частот, нормировка

* Статья получена 14 июля 2026 г.

ВВЕДЕНИЕ

Частотный анализ текстов, основанный на статистических закономерностях естественного языка, является одним из инструментов криптоанализа и реализуется через сравнение частот символов и их сочетаний – униграмм, биграмм и триграмм – с эталонными характеристиками языка [1]. Наиболее наглядно его применение демонстрируется на модели криптоанализа шифра простой замены, в котором все статистические свойства открытого текста сохраняются в шифротексте [2]. Вместе с тем статистические характеристики буквенных последовательностей, в том числе триграммы, находят применение и в смежных задачах – в частности, в задаче установления принадлежности анонимного текста конкретному автору и других задачах.

Переход от анализа отдельных символов к биграммам и триграммам существенно повышает точность криптоанализа за счет учета контекстуальных связей языка [3]. При дальнейшем увеличении длины сочетаний знаков можно перейти к использованию словарного метода, который дает меньшую ошибку [4]. Но частотные методы обладают рядом практических преимуществ перед словарными методами анализа текстов. Во-первых, из текста могут быть удалены пробелы и знаки препинания (что характерно, например, при шифровании), что делает словарный подход в таких случаях неприменимым без предварительного восстановления разбивки текста на слова. Во-вторых, частотный метод не зависит от конкретного языка: для его применения достаточно располагать эталонным распределением сочетаний знаков целевого языка.

В работе [1] использование триграмм в метаэвристических алгоритмах криптоанализа признается наиболее эффективным подходом.

С другой стороны, некоторые авторы отмечают, что при работе с русскоязычными текстами возникают дополнительные сложности, обусловленные морфологическим богатством русского языка: разнообразие словоформ ведет к большому варьированию статистики буквенных сочетаний [5]. Это может повышать требования к объему шифротекста, необходимому для получения репрезентативной триграммной статистики, поэтому зависимость успешности триграммного криптоанализа от длины шифротекста требует отдельного экспериментального исследования [5].

Применение триграммного анализа требует наличия эталона частот триграмм – набора эталонных частот встречаемости трехбуквенных последовательностей, характерных для данного языка. Для русскоязычных текстов подобный эталон биграмм был применен в методе Якобсена [2], однако аналогичного эталона триграмм в доступной литературе обнаружено не было. Настоящая работа направлена на восполнение этого пробела и сравнитель-

ную оценку эффективности триграммного подхода по отношению к биграммному.

Целью данной работы является адаптация метода Якобсена при криптоанализе шифра простой замены с использованием знаковых триграмм для русскоязычных текстов.

1. ФОРМИРОВАНИЕ ЭТАЛОНА ТРИГРАММ

1.1. ХАРАКТЕРИСТИКА ИСХОДНОГО ТЕКСТА И ПРОЦЕДУРА ПРЕДОБРАБОТКИ

Для криптоанализа шифра замены необходим эталон частот триграмм, отражающий вероятности их встречаемости в естественном языке. Готового эталона в литературе не обнаружено, поэтому был создан собственный.

В качестве корпуса использовалась «выборка 2» из работы [2]. Этот корпус уже применялся ранее, что обеспечивает корректное сравнение с результатами биграммных методов. Выборка включает 100 текстов вузовских учебных пособий разных авторов из различных областей (математика, химия, физика, машиностроение и др.). Из каждого текста случайным образом выделялись последовательные фрагменты различной длины, начинающиеся со случайного знака. Такой подход моделирует «произвольный» текст.

Русский алфавит приведен к 31 букве: мягкий и твердый знаки объединены в «Ь», «Е» и «Ё» – в «Е». Из текстов исключены пробелы, знаки препинания, цифры и прочие небуквенные символы, в статистике участвуют только буквы.

Эталон триграмм формировался суммированием всех триграмм по корпусу только для фрагментов с полным алфавитом. Количественные характеристики обработанных групп текстов приведены в табл. 1.

Таблица 1

Table 1

Характеристика групп текстов

Characteristics of groups of texts

G	D	S	K	V
1	200...2000	200	548	669 583
2	2000...10 000	2000	481	2 554 927
3	10 000...110 000	20 000	275	14 336 924
4	100 000...350 000	50 000	48	9 411 829
Итого:			1352	26 973 263

В табл. 1 столбец G содержит номер группы текстов, столбец D – диапазон объемов текстов в символах, столбец S – шаг изменения объема текстов в данном диапазоне в символах, столбец K – количество текстов, столбец V – общий объем символов текстовой группы.

Важным вопросом при формировании эталона триграмм является вопрос его устойчивости как характеристики языка. Для этого следует оценить стабильность частот триграмм, полноту покрытия триграмм, стабильность «рисунка» триграмм, т. е. сходства расположения триграмм на соответствующих срезах трехмерного массива частот встречаемости триграмм при переходе от одной группы текстов к другой, а также решить вопрос нормировки эталона.

1.2. АНАЛИЗ СТАБИЛЬНОСТИ ЧАСТОТ ТРИГРАММ

Рассмотрим задачу о стабильности частот триграмм для русскоязычных текстов.

На рис. 1 представлен график распределения частот встречаемости триграмм для суммарной статистики по всем группам текстов, где для каждого значения частоты показано, сколько именно триграмм встретилося с данной частотой.

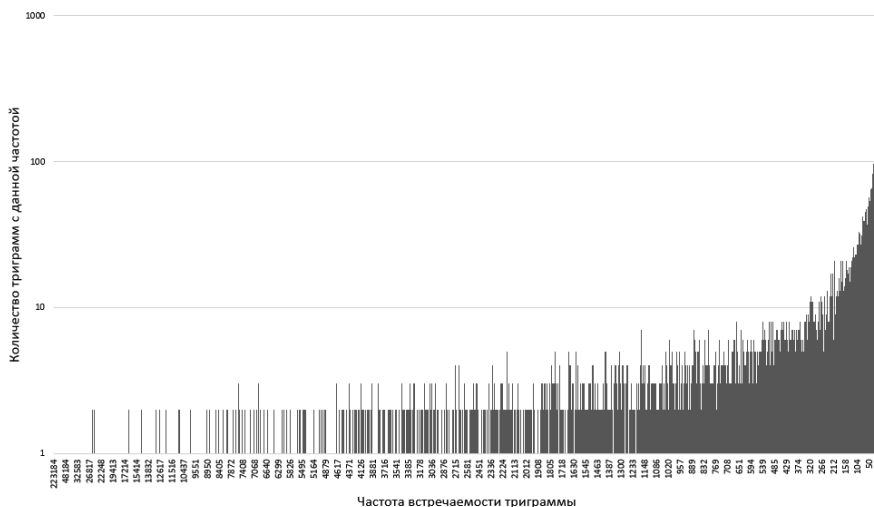


Рис. 1. Распределение частот встречаемости триграмм

Fig. 1. Frequency distribution of trigrams

Из рис. 1 видно, что в области высоких частот каждой частоте, как правило, соответствует единственная триграмма. При переходе к низким частотам одна и та же частота может относиться уже к сотням различных триграмм. Для проверки стабильности частотных характеристик рассмотрено 20 наиболее часто встречающихся триграмм по всем группам текстов. Такой выбор позволяет оперировать триграммами без частотных коллизий и исключить редкие сочетания, часто являющиеся случайными.

Далее под суммарной статистикой понимается частотная статистика триграмм, вычисленная по всему корпусу выборки 2 (без разделения на группы) объемом 26 973 263 символа.

В табл. 2 приведены 20 наиболее часто встречаемых триграмм и их частоты для групп текстов 2, 3, 4, а также для суммарной статистики.

Таблица 2

Table 2

Статистика по наиболее часто встречаемым триграммам**Statistics on the most common trigrams**

Триграмма	Суммарная статистика	Группа текстов 2	Группа текстов 3	Группа текстов 4
ЕНИ	223 184	21 640	118 048	77 984
ОСТ	169 526	15 244	90 139	60 058
НИЯ	129 043	12 155	68 613	44 952
СТВ	103 505	8801	53 200	39 055
НИЕ	99 309	9521	51 551	35 856
ПРИ	96 515	9053	51 446	33 619
СТИ	93 782	7343	48 372	36 090
ПРО	92 489	8700	50 296	31 260
ОГО	91 874	9077	48 664	31 750
НОС	90 532	8374	47 583	32 444
ЛЬН	89 514	9302	50 040	27 702
ТСЯ	85 629	7944	46 236	29 251
АНИ	79 783	7295	41 526	28 981
ЕТС	76 305	6803	40 654	26 896
ЛЕН	74 373	6615	39 389	26 800
ИЧЕ	73 415	6473	38 312	26 850
ЕНН	73 128	6539	35 499	29 274
РЕД	73 108	7037	39 330	25 050
ПОЛ	72 363	7567	39 335	23 514
ЧЕС	71 079	6155	36 582	26 626

Из табл. 2 видно, что во всех группах текстов состав топ-20 триграмм практически идентичен: триграммы ЕНИ, ОСТ, НИЯ, СТВ, НИЕ, ПРИ и другие неизменно входят в число наиболее встречаемых.

Однако важно оценить не только состав, но и порядок следования триграмм по частоте. Для этого вычислены разности рангов каждой триграммы в группе текстов и ранга в суммарной статистике. Положительная разность означает, что в группе текстов триграмма опустилась ниже, отрицательная – поднялась выше. Полученные значения (для групп текстов 1–4) визуализированы на рис. 2 в виде тепловой карты [6].

Триграмма	Группа текстов 1	Группа текстов 2	Группа текстов 3	Группа текстов 4
ЕНИ	0	0	0	0
ОСТ	0	0	0	0
НИЯ	0	0	0	0
СТВ	1	4	0	0
НИЕ	3	-1	0	1
ПРИ	0	1	0	1
СТИ	6	6	3	-2
ПРО	1	1	-1	2
ОГО	-2	-3	0	0
НОС	1	0	1	-2
ЛЬН	-7	-6	-3	4
ТСЯ	-2	-1	0	0
АНИ	-1	1	0	0
ЕТС	0	4	0	3
ЛЕН	11	4	0	4
ИЧЕ	3	5	2	2
ЕНН	1	3	5	-6
РЕД	3	-1	-1	3
ПОЛ	-4	-7	-3	3
ЧЕС	0	5	-1	0

Рис. 2. Разность рангов триграмм

Fig. 2. The difference in the ranks of trigrams

Анализ тепловой карты (рис. 2) показывает, что разности рангов в группах 3 и 4 не превышают трех-четырех позиций для большинства триграмм, а для многих равны нулю. В группе 2 разности выше (до семи), в группе 1 достигают 11, что отражает более сильные отклонения в коротких фрагментах. Тем не менее состав топ-20 триграмм практически неизменен во всех группах.

Если бы частоты определялись преимущественно случайными факторами, набор наиболее частотных сочетаний менялся бы хаотически. Регулярное воспроизведение одного и того же ядра высокочастотных триграмм свидетельствует о наличии устойчивых языковых закономерностей, пригодных для построения эталона.

Триграммы образуют трехмерное распределение. Двумерный срез такого распределения в той или иной плоскости назовем «пластиной».

Рассмотрены также наиболее частотные триграммы в пластинах по строкам: при фиксированной третьей букве в каждой строке (при фиксированной первой букве) находилась максимальная по частоте триграмма. Результат аналогичен предыдущему: триграммные максимумы схожи во всех группах, что подтверждает устойчивость триграммного распределения.

2. ПОЛНОТА ПОКРЫТИЯ ТРИГРАММ

Для уточнения характера распределения триграмм в русскоязычных текстах рассмотрим полноту их представления. Важным параметром статистической структуры языка является количество реально встречающихся триграмм – комбинаций из трех букв, появляющихся в тексте хотя бы один раз. Теоретически возможно 29 791 различных триграмм, однако реальный язык использует лишь их часть. Сравнение групп текстов разного объема позволяет оценить зависимость полноты покрытия от количества проанализированного материала.

Рассмотрено распределение количества ненулевых триграмм для каждой из 31 пластины. На рис. 3 представлен график количества ненулевых триграмм по пластинам для всех групп текстов. Для наглядности буквы на оси абсцисс упорядочены не по частоте встречаемости, а по числу ненулевых триграмм, в которых они занимают третье место.

На рис. 3 и далее цифрами 1–4 обозначены группы текстов (см. табл. 1), цифрой 5 – суммарная статистика. Графики показывают, что для каждой буквы число ненулевых триграмм стабильно растет от группы 1 к группе 3: чем больше текста, тем больше редких сочетаний проявляется. Исключение – группа 4, где значения ниже, чем в группе 3, хотя фрагменты длиннее, но текстов меньше (48 против 275). Это свидетельствует о том, что суммарное разнообразие триграмм при малом числе источников может быть ниже, чем при большом количестве коротких отрывков. Суммарная статистика превышает показатели любой группы, аккумулируя все триграммы.

В табл. 3 приведено количество ненулевых триграмм для каждой группы и процент от теоретического максимума (29 791).

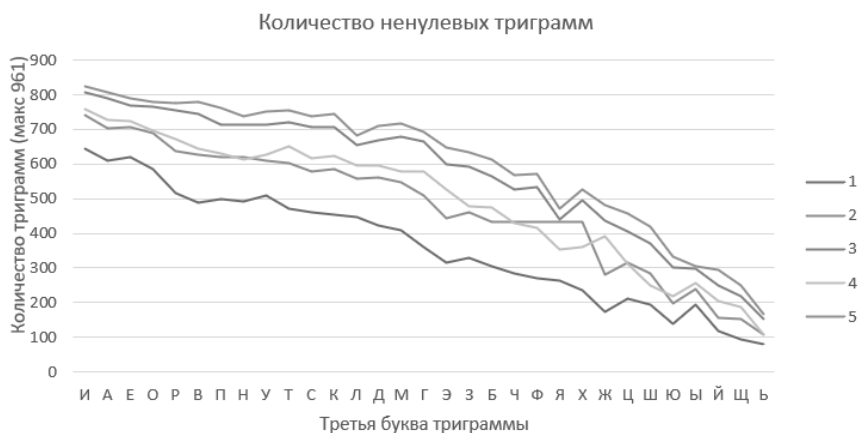


Рис. 3. Количество ненулевых триграмм по пластинам

Fig. 3. The number of non-zero trigrams on the plates

Таблица 3

Table 3

Количество ненулевых триграмм по группам текстов

Number of non-zero trigrams by text groups

Номер группы текстов	Количество ненулевых триграмм	Процент от теоретического максимума
1	11 205	37,61 %
2	14 449	48,50 %
3	17 765	59,63 %
4	15 303	51,37 %
Суммарная статистика	18 796	63,09 %

Если принять суммарную статистику за 100 %, то группа текстов 1 покрывает 59,6 % эталонного разнообразия, группа текстов 2 – 76,9 %, группа текстов 3 – 94,5 %, группа текстов 4 – 81,4 %. Таким образом, уже группа

текстов 3 обеспечивает практически полное покрытие от того, что дает суммарный эталон. Дальнейшее увеличение объема за счет добавления групп текстов 1, 2 и 4 дает прирост лишь на 5,5 %. Это свидетельствует о достижении насыщения: большинство реально существующих в языке триграмм выявляется уже при анализе текстов группы 3.

3. ПРОВЕРКА СТАБИЛЬНОСТИ ГЕОМЕТРИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Если эталон является устойчивым «усредненным портретом» языка, то его множество ненулевых триграмм должно быть близко к множествам, полученным из больших независимых выборок (группы 3 и 4). Для количественной оценки этой близости применяется бинарное сравнение, при котором учитывается только факт наличия триграммы, но не ее частота.

Для каждой группы текстов (G) и для каждой пластины (c) была поставлена в соответствие бинарная матрица $M_G^{(c)}$, где, как и в пластинах, строки соответствуют первой букве триграммы, столбцы – второй. Элемент матрицы определяется следующим образом:

$$M_G^{(1)}(i, j) = \begin{cases} 1, & \text{если триграмма } (i, j, c) \text{ встретилась хотя бы раз,} \\ 0 & \text{в противном случае,} \end{cases} \quad (1)$$

где i, j – индексы букв.

Матрица разности для двух групп текстов A и B определяется по формуле

$$D_{A,B}^{(c)} = M_A^{(c)} - M_B^{(c)}. \quad (2)$$

Элементы $D_{A,B}^{(c)}(i, j)$ могут принимать три значения:

- 1) 0 – триграмма (i, j, c) присутствует либо отсутствует в обеих группах текстов;
- 2) 1 – триграмма есть только в группе текстов A ;
- 3) -1 – триграмма есть только в группе текстов B .

Далее для каждой пластины подсчитывается количество уникальных триграмм по формулам:

$$S_1^{(c)} = \sum \left(D_{A,B}^{(c)}(i, j) = 1 \right), \quad S_{-1}^{(c)} = \sum \left(D_{A,B}^{(c)}(i, j) = -1 \right), \quad (3)$$

где $S_1^{(c)}$ – количество триграмм, уникальных для группы текстов A ; $S_{-1}^{(c)}$ – количество триграмм, уникальных для группы текстов B .

Абсолютные величины S_{-1} и S_1 зависят от общего числа ненулевых триграмм в каждой группе. Чтобы получить характеристику, позволяющую сравнивать различные объемы текстов, вычисляется значение по формуле

$$R_{A,B}^{(c)} = \frac{S_1^{(c)} + |S_{-1}^{(c)}|}{N_A^{(c)} + N_B^{(c)}}, \quad (4)$$

где $N_A^{(c)} = \sum \left(M_A^{(c)}(i, j) \right)$ – общее количество ненулевых триграмм в пластине c для группы текстов A .

Если $R = 0$, множества ненулевых триграмм полностью совпадают; если $R = 1$ – множества не пересекаются. Промежуточные значения отражают долю уникальных триграмм в объединенном множестве.

Рассматривались попарные сравнения независимых групп текстов и сравнение каждой группы с суммарной статистикой. Попарное сравнение выполнялось с группой 3, обладающей наибольшим объемом (более 14 млн символов). На рис. 4 представлены графики значений R для различных сочетаний групп текстов; по оси абсцисс отложены буквы пластин.

На графике, представленном на рис. 4, видно, что для всех пластин значения R уменьшаются от пары (3 + 1) к паре (3 + 4). Например, для буквы О: $0,1376 \rightarrow 0,0646 \rightarrow 0,0562$; для буквы Ж: $0,4525 \rightarrow 0,2580 \rightarrow 0,1325$. Это указывает на то, что с ростом объема текста множество ненулевых триграмм стремится стать идентичным. Для самых часто встречаемых букв даже в группе текстов 1 значения R составляют $0,12 \dots 0,13$, т. е. $87 \dots 88\%$ триграмм совпадают с группой текстов 3. Для редких букв R может достигать $0,3 \dots 0,4$ даже в группе текстов 4, что говорит о высокой вариативности сочетаемости редких букв.

Для оценки того, за счет какой группы текстов происходит расхождение, на рис. 5 показан вклад каждой группы текстов (1 и 3) отдельно для сравнения их друг с другом.



Рис. 4. Результат сравнения для ненормированных данных

Fig. 4. Comparison result for non-normalized data



Рис. 5. Сравнение вклада групп текстов по отдельности

Fig. 5. Comparing the contribution of groups of texts individually

Из рис. 5 видно, что триграмм, присутствующих только в группе текстов 3, на порядок больше. Это говорит об отсутствии многих реально существующих сочетаний в малых выборках небольших фрагментов текстов.

Дополнительно было проведено сравнение каждой группы текстов с суммарной статистикой. Результаты для ненормированных данных приведены на рис. 6.



Рис. 6. Сравнение с суммарной статистикой

Fig. 6. Comparison with summary statistics

Сравнение с суммарной статистикой дает оптимальную оценку сходства: любая триграмма группы входит в суммарную статистику. Поэтому R здесь – доля триграмм эталона, отсутствующих в данной группе. Для группы 1 она составляет в среднем 13...45 % (в зависимости от пластины), для группы 3 – менее 1...2 %.

Для каждой группы вычислена разность R при сравнении с суммарной статистикой и с группой 3; разности просуммированы по всем пластинам. Для групп 2 и 4 получено уменьшение R , что указывает на большую близость суммарной статистики к этим группам, чем у самой большой независимой группы. Для группы 1, напротив, значение R увеличилось из-за малого объема фрагментов. В целом результаты подтверждают, что суммарная статистика наиболее полно отражает языковые закономерности и является оправданным выбором в качестве эталона.

3.1. НОРМИРОВКА ДАННЫХ

Корпус текстов для построения эталона имеет объем более 26 млн символов. При таком объеме в эталоне появляются очень редкие триграммы – результат случайных сочетаний, специфических терминов или опечаток. Половина всех триграмм в суммарной статистике имеет частоту встречаемости более 12 735. Большое число редких триграмм затрудняет оценку различия при сопоставлении эталона с анализируемым текстом ограниченного объема, поскольку они практически отсутствуют в последнем.

Для приведения частот к сопоставимому виду применена нормировка к объему 5000 символов по формуле

$$f_{\text{норм}} \approx \frac{x}{N} 5000, \quad (5)$$

где x – количество вхождений конкретной триграммы в анализируемом корпусе; N – общее число символов (букв) в этом корпусе.

Полученное значение округляется до ближайшего целого.

При нормировке сохраняется порядок: если частота одной триграммы была выше другой, после нормировки она не станет меньше, но может сравняться. Нормировка статистики по всем текстам выполнялась после суммирования данных по группам.

После нормировки и округления возникает вопрос: не теряется ли важная информация при отбрасывании или округлении редких триграмм? Для ответа необходимо проверить, насколько сильно различается распределение триграмм в разных группах текстов и как влияет нормировка с округлением на степень их сходства.

С использованием нормированных частот для каждой группы рассчитан коэффициент расхождения R по формуле (4). На рис. 7 приведены распределения ненулевых триграмм для нормированных и ненормированных данных.

Как видно из графиков на рис. 7, нормировка с округлением занижает коэффициент расхождения R , причем сильнее всего для пар групп с изначально большими различиями. Например, при сравнении группы 2 и группы 3 нормировка уменьшает R гораздо заметнее, чем при сравнении группы 3 и группы 4.

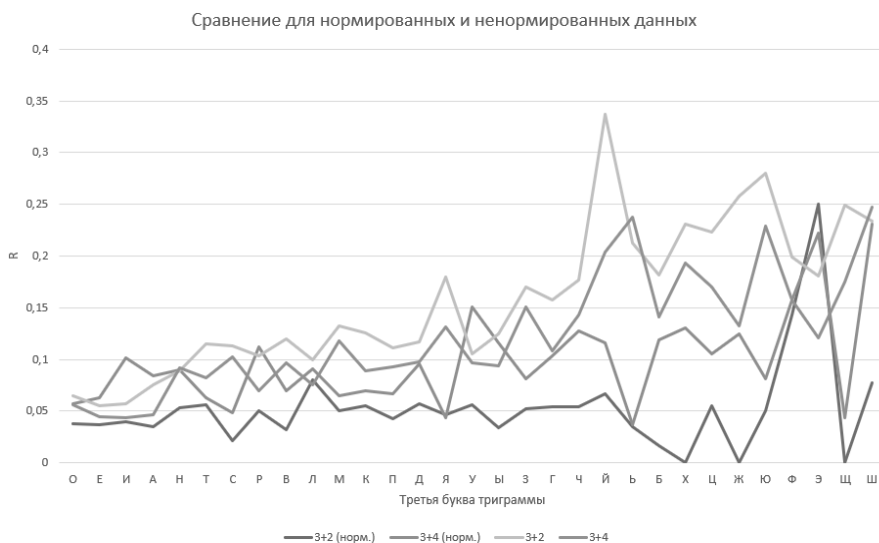


Рис. 7. Сравнение нормированных и ненормированных данных

Fig. 7. Comparison of normalized and non-normalized data

Окончательно влияние нормировки рассмотрим в разделе 4.

4. РЕАЛИЗАЦИЯ МЕТОДА ЯКОБСЕНА КРИПТОАНАЛИЗА ШИФРА ПРОСТОЙ ЗАМЕНЫ НА ОСНОВЕ ЗНАКОВЫХ ТРИГРАММ

4.1. АДАПТАЦИЯ МЕТОДА ЯКОБСЕНА ДЛЯ ТРИГРАММ

Формально N -граммой на алфавите A , состоящем из конечного числа символов, называется строка из N символов, принадлежащих алфавиту A [7].

Модель триграмм базируется на наборах:

$$G_i = \{g_{i1}, g_{i2}, g_{i3}, \theta(g_{i1}, g_{i2}, g_{i3})\}, \quad (6)$$

где $g_{i1} \in A_g$, $g_{i2} \in A_g$, $g_{i3} \in A_g$ – символы из словарного алфавита A_g , (g_{i1}, g_{i2}, g_{i3}) – частота встречаемости триграммы, т. е. количество повторений трехсимвольной цепочки (g_{i1}, g_{i2}, g_{i3}) . В анализируемом тексте частоты

триграмм обычно называют частотными характеристиками, поскольку их легко привести к вероятностной форме нормированием к общему числу символов или на число определенных символов. На практике чаще используют абсолютные частоты (число повторов) [8].

При идентификации знаков текста на основе триграммного распределения частот применяется обобщение метода Якобсена, изначально предложенного для биграмм [2]. Алгоритм основан на минимизации целевой функции вида

$$W(T) = \left(\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N |t_{ijk} - b_{ijk}| \right) - \delta, \quad (7)$$

где t_{ijk} и b_{ijk} – значения частот триграмм анализируемого текста и эталона соответственно; δ – поправка для устранения двойного учета триграмм; N – размер алфавита.

Поиск минимального значения $W(T)$ выполняется процедурой упорядочивания знаков с переменным шагом. Начальный порядок алфавита задается сортировкой по убыванию частот отдельных знаков в анализируемом тексте, эталонный порядок – аналогичной сортировкой частот в эталоне. Далее последовательно просматриваются пары знаков: сначала соседние, затем отстоящие на один, два знака и более, вплоть до пары из первого и последнего (максимальный шаг $N - 1$). Множество пар с одинаковым расстоянием между индексами образует уровень просмотра, а индекс начального элемента пары – ее место. Просмотр на каждом уровне начинается с первого символа.

Для каждой пары вычисляются W_0 при текущем порядке и W_1 после условной перестановки этих знаков. Перестановка влечет одновременную перестановку соответствующих слоев во всех трех измерениях куба частот. Если выполняется условие

$$W_0 > W_1, \quad (8)$$

то пара переставляется окончательно, и просмотр возвращается к нулевому уровню (шаг = 1) с самого начала. В противном случае знаки возвращаются на исходные позиции, и просмотр продолжается в прежнем порядке.

Процесс повторяется, пока при полном проходе по всем уровням не будет выполнено ни одной перестановки. Таким образом, метод Якобсена для триграмм представляет собой сортировку с переменным шагом, адаптированную к минимизации трехмерного распределения частот, и является естественным обобщением двумерного варианта.

5. СРАВНЕНИЕ БИГРАММНОЙ И ТРИГРАММНОЙ РЕАЛИЗАЦИИ МЕТОДА ЯКОБСЕНА ДЛЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Сформированный по выборке 2 эталон триграмм требует экспериментальной проверки на реальных текстах. Для этого использована «выборка 1» из работы [2], тексты которой не участвовали в формировании эталона. Выборка 1 отличается составом и способом формирования фрагментов, что позволяет оценить универсальность эталона для языка.

Выборка 1 составлена из научно-популярных и художественных текстов различных жанров и авторов. В отличие от выборки 2 (случайные фрагменты), здесь применен принцип вложенности: после удаления пробелов выделялись фрагменты большей длины, из которых затем формировались меньшие фрагменты путем последовательного удаления постоянного объема знаков. Начала всех фрагментов одного текста совпадали. Это позволяет оценить зависимость погрешности метода от объема текста.

Объемы фрагментов распределены по четырем группам, аналогичным группам выборки 2:

- 1) группа 5 – фрагменты 400...1800 знаков с шагом 200 (655 фрагментов);
- 2) группа 6 – 2000...10 000 знаков с шагом 2000 (492 фрагмента);
- 3) группа 7 – 30 000...110 000 знаков с шагом 20 000 (300 фрагментов);
- 4) группа 8 – 100 000...350 000 знаков с шагом 50 000 (120 фрагментов).

Все фрагменты приведены к алфавиту из 31 буквы; пробелы и небуквенные символы удалены.

5.1. СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ БИГРАММ И ТРИГРАММ

Для оценки эффективности триграммного подхода проведено сравнение с методом Якобсена на биграммах на одних и тех же текстах выборки.

Использовалась ошибка полной идентификации O_1 – доля фрагментов, содержащих хотя бы один неправильно идентифицированный символ, относительно всех фрагментов данной длины [2].

На рис. 10 показана зависимость O_1 от объема текста для биграммного метода [2] и триграммного метода с разными вариантами нормировки.

На рис. 10 видно, что на малых объемах все методы дают высокие значения O_1 , но триграммный метод показывает несколько меньшую ошибку. Начиная с 800 символов триграммный метод с оптимальной нормировкой заметно превосходит биграммный: при 2000 символов O_1 снижается в 6,5 раза. С 6000 символов триграммный метод с оптимальной нормировкой достигает нулевой ошибки O_1 , нормировка к 15 000 обеспечивает наименьшую долю

ошибочных фрагментов на всём диапазоне. Нормировка к 5000 на малых и средних объемах близка к оптимальной, но свыше 6000 символов сохраняет 2...6 % ошибочных фрагментов.

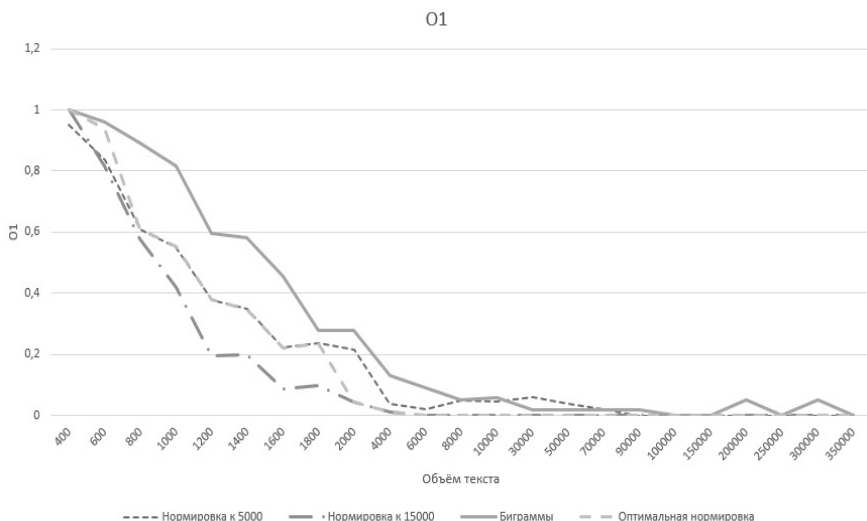


Рис. 8. График ошибок O1

Fig. 8. Error graph O1

Для оценки погрешности идентификации символов использовалась ошибка O2 [2] – среднее число неправильно идентифицированных знаков на один ошибочный фрагмент, нормированное на размер алфавита. Величина O2 вычислялась по формуле (9).

$$O2 = \frac{M}{N \cdot 31}, \quad (9)$$

где M – суммарное количество ошибочно идентифицированных символов для всех текстов данного объема; N – количество ошибочных текстов данного объема; 31 – размер алфавита.

На рис. 11 представлен график зависимости O2 от объема текста для биграммного [2] и триграммного методов с различными вариантами нормировки.

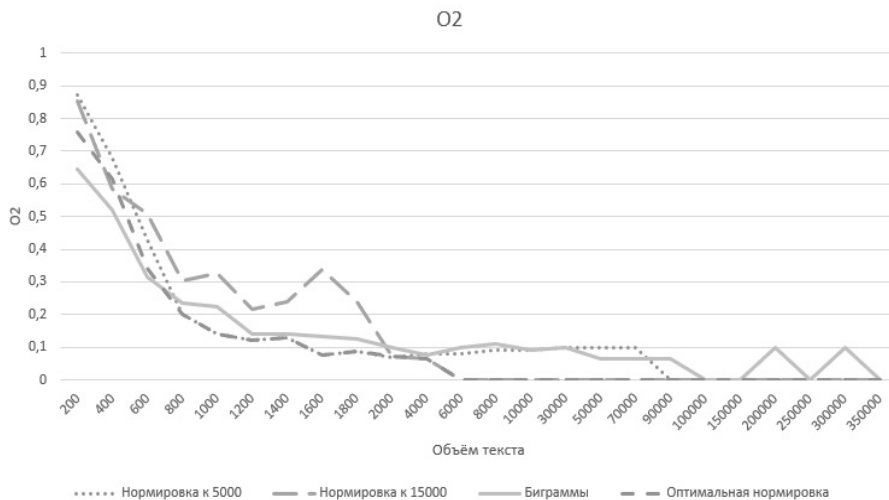


Рис. 9. График распределения ошибок O2

Fig. 9. Graph of O2 error distribution

Из рис. 11 видно, что в диапазоне 200...600 символов ошибка O2 для биграммного метода меньше на 8...16 %, чем для триграммного с оптимальной нормировкой. С 800 до 6000 символов соотношение меняется: триграммный метод дает ошибку O2 на 20...30 % ниже. Начиная с 6000 символов триграммный метод с оптимальной нормировкой достигает нулевого уровня O2, тогда как у биграммного остаются ненулевые значения. Ошибка снижается до нуля лишь к 90 000 символам.

Несмотря на более высокие затраты памяти и времени, снижение ошибки идентификации делает триграммный подход привлекательным для криптоанализа шифра простой замены при работе с текстами объемом более 800 символов.

6. МЕСТО РАБОТЫ В КОМПЛЕКСНОЙ ЗАЩИТЕ ОБЪЕКТА ИНФОРМАТИЗАЦИИ

Объект информатизации – совокупность информационных ресурсов, средств и систем обработки информации, а также помещений и объектов, в которых они установлены или которые предназначены для конфиденциальных переговоров [9]. Его защита обеспечивается комплексом мер: правовой,

технической, криптографической и физической [10]. Каждый вид защиты решает свой круг задач и по отдельности недостаточен.

Криптографическая подсистема обеспечивает конфиденциальность и целостность информации с помощью алгоритмов шифрования в СКЗИ. Стойкость алгоритмов подтверждается методами криптоанализа [11], включая статистические – анализ распределения символов шифротекста и тесты на случайность. Чем ближе распределение к равновероятному, тем выше стойкость [12]. Шифр простой замены по современным меркам недостаточно стойкий [11]: энтропия языка сохраняется в шифротексте, и статистические закономерности остаются доступны. Поэтому он служит учебной моделью статистического криптоанализа. Сформированный эталон частот триграмм и адаптированный метод Якобсена позволяют исследовать уязвимость шифра простой замены, оценить минимальный объем шифротекста для результативной атаки и сравнить триграммный и биграммный подходы. Триграммы применяются и в других задачах информационной безопасности. В работе [13] на фрагменте художественной литературы ошибка идентификации автора методом триграмм составила 23,8 % против 34,2 % для биграмм, подтверждая их большую информативность. Эталон частот триграмм – универсальный инструмент статистического анализа русскоязычных текстов.

ЗАКЛЮЧЕНИЕ

В работе решена задача адаптации метода Якобсена для криптоанализа шифра простой замены с использованием знаковых триграмм применительно к русскоязычным текстам. Основные результаты:

- сформирован эталон частот триграмм на основе репрезентативной выборки текстов общим объемом более 26 млн символов. Установлено, что состав топ-20 наиболее частотных триграмм стабилен для различных групп текстов, а группа текстов 3 (10 000...110 000 символов) покрывает 94,5 % эталонного разнообразия, что подтверждает достижение насыщения статистических характеристик языка;
- адаптирован метод Якобсена для триграммных распределений, и разработана его программная реализация. Экспериментально определена оптимальная стратегия нормировки частот в зависимости от объема анализируемого текста, позволяющая минимизировать ошибку идентификации знаков;
- проведен сравнительный анализ с биграммной реализацией метода. Триграммный подход обеспечивает существенно меньшую ошибку идентификации: для текстов объемом свыше 6000 символов ошибка O1 снижается до нуля, тогда как биграммный метод в этом диапазоне сохраняет ненулевые

значения вплоть до 90 000 символов. Ошибка O2 для триграммного метода в диапазоне 800...6000 символов на 20...30 % ниже по сравнению с биграммным подходом;

- количественно оценен минимальный объем шифротекста (около 6000 символов), при котором атака на основе триграммного эталона становится результативной, что подтверждает практическую ценность разработанного подхода для статистического криптоанализа русскоязычных текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Sabonchi A.K.S., Akay B.* A survey on the metaheuristics for cryptanalysis of substitution and transposition ciphers // *Computer Systems Science & Engineering*. – 2021. – Vol. 39 (1). – P. 87–104. – DOI: 10.32604/csse.2021.05365.
2. *Котов Ю.А.* Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // *Труды СПИИРАН*. – 2017. – № 1 (50). – С. 190–208. – EDN XVLOWN.
3. *Васильев Е.М., Гусев К.Ю.* Анализ избыточности русскоязычного текста // *Вестник Воронежского государственного технического университета*. – 2010. – Т. 6, № 8. – С. 101–104.
4. *Котов Ю.А., Рассказов А.Д., Таронов Р.П.* Алгоритм идентификации простой замены букв текста по словарю // *Научные вести*. – 2019. – № 6 (11). – С. 201–212. – EDN SVJULP.
5. *Холоденко А.Б.* О построении статистических языковых моделей для систем распознавания русской речи // *Интеллектуальные системы*. – 2001. – Т. 6, № 1–4. – С. 381–394. – EDN YWAXTV.
6. *Munzner T.* Visualization analysis and design. – Boca Raton: CRC Press, 2015. – 404 p. – ISBN 978-1-4665-0893-4.
7. Алгоритм применения N-грамм для корректировки результатов распознавания / Т.В. Манжиков, О.А. Славин, И.А. Фараджев, И.М. Янишевский // *Вестник Российского фонда фундаментальных исследований*. – 2016. – № 4 (92). – С. 116–123. – DOI: 10.22204/2410-4639-2016-092-04-116-123.
8. *Котов Ю.А.* Криптографические методы защиты информации: шифры: учебное пособие. – Новосибирск: Изд-во НГТУ, 2016. – 60 с.
9. ГОСТ Р 51275–2006. Защита информации. Объект информатизации. Факторы, воздействующие на информацию. Общие положения. – М.: Стандартинформ, 2007.
10. ГОСТ Р 50922–2006. Защита информации. Основные термины и определения. – М.: Стандартинформ, 2008.

11. *Карпов А.В., Ишмуратов Р.А.* Введение в криптографию: учебное пособие. – Казань: Казан. ун-т, 2024. – 128 с.

12. Оценка стойкости криптографических шифров с помощью анализа шифртекстов / Т.С. Байшоланов, Ж.М. Алимжанова, Н. Байшолан, К.Е. Кубаев, К.С. Байшоланова // Известия Национальной академии наук Республики Казахстан. Серия физико-математическая. – 2022. – № 1. – С. 26–33. – DOI: 10.32014/2022.2518-1726.113.

13. *Кислицына М.Ю., Орлов Ю.Н.* Статистический анализ полного корпуса художественной литературы на русском языке и распознавание автора // Препринты ИПМ им. М.В. Келдыша. – 2024. – № 17. – С. 1–24. – EDN NIPYEE. – DOI: 10.20948/prepr-2024-17.

Котов Юрий Алексеевич, кандидат физико-математических наук, доцент кафедры защиты информации. E-mail: kotov@corp.nstu.ru

Башикиров Александр Петрович, лаборант кафедры защиты информации Новосибирского государственного технического университета. E-mail: kaf_zi@corp.nstu.ru

DOI: 10.17212/2782-2230-2026-3-41-63

Implementation of cryptanalysis of a simple substitution cipher using sign-based trigrams*

Yu.A. Kotov¹, A.P. Bashkirov²

¹*Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Information Security. E-mail: kotov@corp.nstu.ru*

²*Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, laboratory assistant of the Department of Information Security. E-mail: kaf_zi@corp.nstu.ru*

This paper examines the adaptation of the Jacobsen method-using sign-based trigrams-for the cryptanalysis of simple substitution ciphers applied to Russian-language texts. The study evaluates the robustness of the reference model by assessing the stability of the frequencies and ranks of the most common trigrams, the coverage of the set of non-zero trigrams, and the geometric similarity of their distribution across the cipher segments. Normalization was applied to ensure frequency comparability, and the optimal normalization strategy was experi-

* Received 14 July 2026.

mentally determined based on the volume of the text being analyzed. The method was implemented in software and tested on an independent dataset comprising popular science and literary texts. A comparative analysis with a bigram-based implementation demonstrated that the trigram approach yields a significantly lower identification error rate-dropping to zero for texts exceeding 6,000 characters-while maintaining acceptable computational costs. These results allow for a quantitative assessment of the minimum ciphertext volume required for a successful trigram-based attack and confirm the practical value of the trigram reference model for the statistical cryptanalysis of Russian-language texts.

Keywords: trigrams, frequency analysis, simple substitution cipher, Jacobsen's method, frequency reference, normalization

REFERENCES

1. Sabonchi A.K.S., Akay B. A survey on the metaheuristics for cryptanalysis of substitution and transposition ciphers. *Computer Systems Science & Engineering*, 2021, vol. 39 (1), pp. 87–104. DOI: 10.32604/csse.2021.05365.
2. Kotov Yu.A. Approximatsiya 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 (50), pp. 190–208.
3. Vasilyev E.M., Gusev K.Yu. Analiz izbytochnosti russkoyazychnogo teksta [The analysis of redundancy the Russian-speaking text]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta = Bulletin of Voronezh State Technical University*, 2010, vol. 6, no. 8, pp. 101–104.
4. Kotov Yu.A., Rasskazov A.D., Taronov R.P. Algoritm identifikatsii prostoi zameny bukv teksta po slovaryu [Algorithm of identification one-to-one substitution text characters on dictionary]. *Nauchnye vesti = Scientific News*, 2019, no. 6 (11), pp. 201–212.
5. Kholodenko A.B. O postroenii statisticheskikh yazykovykh modelei dlya sistem raspoznavaniya russkoi rechi [On the construction of statistical language models for Russian speech recognition systems]. *Intellektual'nye sistemy = Intelligent Systems*, 2001, vol. 6, no. 1–4, pp. 381–394.
6. Munzner T. *Visualization analysis and design*. Boca Raton, CRC Press, 2015. 404 p. ISBN 978-1-4665-0893-4.
7. Manzhikov T.V., Slavin O.A., Faradzhev I.A., Janiszewski I.M. Algoritm primeneniya N-gramm dlya korrekcirovki rezul'tatov raspoznavaniya [N-grams algorithm application for the correction of recognition results]. *Vestnik Rossiiskogo fonda fundamental'nykh issledovaniy = Russian Foundation for Basic Research Journal*, 2016, no. 4 (92), pp. 116–123. DOI: 10.22204/2410-4639-2016-092-04-116-123.

8. Kotov Yu.A. *Kriptograficheskie metody zashchity informatsii. Shifry* [Cryptographic methods of information protection. Ciphers]. Novosibirsk, NSTU Publ., 2016. 60 p.
9. GOST R 51275–2006. *Zashchita informatsii. Ob"ekt informatizatsii. Faktory, vozdeistvuyushchie na informatsiyu. Obshchie polozheniya* [State standard R 51275–2006. Protection of information. Object of informatisation. Factors influencing the information. General principles]. Moscow, Standartinform Publ., 2007.
10. GOST R 50922–2006. *Zashchita informatsii. Osnovnye terminy i opredeleniya* [State standard R 50922–2006. Protection of information. Basic terms and definitions]. Moscow, Standartinform Publ., 2008.
11. Karpov A.V., Ishmuratov R.A. *Vvedenie v kriptografiyu* [Introduction to cryptography]. Kazan, Kazan University Publ., 2024. 128 p.
12. Baisholanov T., Alimzhanova Zh., Baisholan N., Kubaev K., Baisholanova K. Otsenka stoikosti kriptograficheskikh shifrov s pomoshch'yu analiza shifrttekstov [Evaluation of the strength of cryptographic ciphers using ciphertext analysis]. *Izvestiya Natsional'noi akademii nauk Respubliki Kazakhstan. Seriya fiziko-matematicheskaya = News of the Academy of Sciences of the Republic of Kazakhstan. Series Physico-Mathematical*, 2022, no. 1, pp. 26–33. DOI: 10.32014/2022.2518-1726.113.
13. Kislitsyna M.Yu., Orlov Yu.N. Statisticheskii analiz polnogo korpusa khudozhestvennoi literatury na russkom yazyke i raspoznavanie avtora [Statistical analysis of the complete corpus of fiction in Russian and recognition of the author]. *Preprinty IPM im. M.V. Keldysha = Keldysh Institute preprints*, 2024, no. 17, pp. 1–24. DOI: 10.20948/prepr-2024-17.

Для цитирования:

Котов Ю.А., Башкиров А.П. Реализация криптоанализа шифра простой замены с использованием знаковых триграмм // Безопасность цифровых технологий. – 2026. – № 3 (122). – С. 41–63. – DOI: 10.17212/2782-2230-2026-3-41-63.

For citation:

Kotov Yu.A., Bashkirov A.P. Realizatsiya kriptoolnalyza shifra prostoi zameny s ispol'zovaniem znakovykh trigramm [Implementation of cryptanalysis of a simple substitution cipher using sign-based trigrams]. *Bezopasnost' tsifrovyykh tekhnologii = Digital Technology Security*, 2026, no. 3 (122), pp. 41–63. DOI: 10.17212/2782-2230-2026-3-41-63.