УДК 004.056.55

ПРИМЕНЕНИЕ ЧАСТОТНОГО АНАЛИЗА ДЛЯ КРИПТОАНАЛИЗА ТРАДИЦИОННЫХ СИММЕТРИЧНЫХ КРИПТОСИСТЕМ

Жураев Г.У.

В данной статье изложен частотный криптоанализ традиционных симметричных криптосистем. Для этой цели использованы характеристики открытых текстов узбекского языка.

Ключевые слова: открытый текст, частота букв, биграмма, триграмма, частотный анализ.

Ушбу мақолада частотавий таҳлил усули асосида аъанавий симметрик криптотизимларни таҳлил қилиш масаласи баён этилган. Шу мақсадда ўзбек тили очиқ матнлари характеристикаларидан фойдаланилган.

Таянч иборалар: очиқ матн, ҳарфлар частотаси, биграмма, триграмма частотавий таҳлил.

This article is described in the issue based on the frequency analysis method are used for cryptanalysis of ordinary symmetric cryptosystem's. For this purpose, characteristic of the texts used in the Uzbek language.

Relative frequencies are established (probability of occurrence) of Cyrillic letters in Uzbek language. The bigrams and trigrams with the highest probability of the Uzbek language have been identified. The text cipher is decrypted, encrypted by a mono-alphabetic symmetric cipher using frequency analysis. For polyalphabetic symmetric ciphers, frequency cryptanalysis should be used for each alphabet of the text cipher separately.

Cryptograms derived from traditional symmetric cryptosystems can be successfully decrypted using frequency analysis. Frequency analysis is based on a consistent pattern of open text. The appearance of forbidden bi -grams (trigrams) means the incorrectness of the accepted hypotheses. Using them in cryptanalysis speeds up the process itself, as it saves a lot of unnecessary calculations. If the correct hypotheses are accepted, then after trying to replace the cipher labeling with the cipher value, the decrypted text becomes readable and word-like structures appear

Keywords: open text, frequency of letters, bigrams, trigrams, frequency analysis.

I. ВВЕДЕНИЕ

Модели открытых текстов представляют собой последовательности символов некоторого алфавита, не содержащих запрещенное сочетание букв алфавита. Для построения подобных моделей используются характеристики открытых текстов конкретного языка. Частотный криптоанализ основан именно на вероятностно-статистических закономерностях, присутствующих в открытом тексте [1-3].

II. ОСНОВНАЯ ЧАСТЬ

Как известно, в осмысленных текстах любого естественного языка различные буквы встречаются с разной частотой, при этом относительные частоты букв в различных текстах одного языка близки между собой. То же самое можно сказать и о частотах пар, троек букв открытого текста. Здесь необходимо отметить, что частота букв и их сочетания в открытом тексте зависит от длины и характера самого текста.

Криптограммы, полученные на основе традиционных симметричных криптосистем могут быть успешно дешифрованы с помощью частотного анализа [1-3]. Частотный анализ основан на устойчивой закономерности открытого текста. При этом частотный криптоанализ осуществляется на следующих двух этапах.

1. Подсчет в криптограмме частот повторений шифрообозначений букв, биграм, триграм. В достаточно длинном тексте эти частоты близки к средне-статистичным частотам букв (биграмм и триграмм) языка.

2. Принятие гипотез относительно шифрообозначений и замена шифрообозначений на шифровеличины. Появление запрещенных биграм (триграмм) означает неправильность принятых гипотез. Использование их в криптоанализе ускоряет сам процесс, поскольку избавляет от множества лишних вычислений. Если приняты правильные гипотезы, то после попытки замены шифрообозначений на шифрвеличин дешифруемый текст становится читабельным и появляется словоподобные структуры.

В узбекском языке используются 35 букв. Из них 33 служат для обозначения соответствующих звуков, а 2 буквы - ь, ъ никаким звукам не соответствуют. Алфавит узбекского языка (для удобства дальнейшего криптоаналитического исследования переведен в 32-буквенный), без учета знака пробела имеет вид, указанный в таблице 1.

Таблица 1

А	Б	В	Г	Д	Е, Ё	Ж	З	И	Й	К	Л	М	Н	О	П
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Р	С	Т	У,Ў	Ф	Х,Ҳ	Ц	Ч	Ш	Э	Ю	Я	Қ	Ғ	Ь	Ъ
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

Для изучения вероятности появления букв и буквосочетаний в открытых текстах узбекского языка были выбраны 2 произведения, написанные на узбекском языке, на основе русской кириллицы [4,5].

Как показывает анализ открытых текстов узбекского языка две буквы – *а* и *и* появляются частотой более 10 процентов, буквы *н*, *р*, *л* - частотой более 6 процентов и буква *г* - менее 1 процента. Показатели, полученные при исследовании по [4] приведены в таблице 2. Частота встречаемости символа пробел в тексте наибольшая, чем других символов алфавита и по этой причине он не включен в таблицу 2. По результатам анализа произведений [5], также получаются аналогичные результаты. Здесь необходимо отметить, что встречаемости знаков и их сочетаний в открытом тексте зависит от длины и характера самого текста. Например, в учебных литературах посвященных криптографии частота встречаемости буквы *ф* значительно выше, чем в художественных литературах за счет использования слов таких как, *алфавит*, *аутентификация*, *идентификация*, *криптография*, *многоалфавитные шифры*, *информация*, *зашифрование*, *расшифрование*, *дешифрование* и т.д.

В большинстве случаев учет распределения букв сразу не дает открытого текста. Тогда необходимо учесть относительные частоты повторяющихся символов, т.е. биграммы. В узбекских текстах наиболее часто наблюдаются следующие повторения букв – *лл*, *мм*, *кк*, *тт*, *сс*, *дд*, *нн*, *вв*. Эти биграммы приведены в порядке убывания вероятности встречаемости. Если в шифрованном тексте имеются какие-либо повторяющиеся символы, то можно считать, что они представляют собой одну из этих пар. Среди биграммы *уу* и *ҳҳ*, имеют самые наименьшие частоты встречаемости.

Таблица 2

Символ	Вероятность	Символ	Вероятность	Символ	Вероятность
А	0,154	Б	0,03	Я	0,008
И	0,144	К	0,028	Ж	0,008
Н	0,068	С	0,027	Ф	0,006
Л	0,067	Қ	0,025	Э	0,005
Р	0,059	Ш	0,022	П	0,004
Т	0,047	В	0,016	Ғ	0,003
О	0,047	Й	0,015	Ъ	0,002
У, Ў	0,045	З	0,015	Ю	0,001
М	0,038	Е, Ё	0,019	Ц	0,001
Д	0,035	Х,Ҳ	0,018	Ь	0,001
Г	0,032	Ч	0,009		

Следующие пары букв – *ъъ*, *ьь*, *ээ*, *юю*, *яя*, *ёё*, *ўў*, *хх* и *гг* в реальных текстах узбекского языка не встречаются, по этой причине их можно отнести к запрещённым биграмм. Биграммы *ее* и *цц* в узбекских текстах встречаются

через европейские слова, например, такие как Елисеев, Менделеев, реестр, плеер, Аццолино, Пацци и т.д. А также биграмма *йй* в чистых узбекских языках не используется, но через арабские слова (например, иййака) они в реальных разговорных языках употребляются. Биграммы *ла*, *ар*, *да*, *ни*, *ин*, *ан* являются самыми часто используемыми сочетаниями различных букв.

Если в шифрованных текстах имеются пробелы между словами, тогда следуют определить слова, состоящие из одной, двух или трех букв. В этом случае использование местоимений, союзов и т.п. элементов лексикологии языка ускоряют процесс восстановления открытого текста.

В качестве примера можно рассмотреть следующий шифртекст, состоящий из 226 букв узбекского алфавита:

ТДФДҲ ТЮЭ ГСБЮЁЮЯ ТЮБ ХСУБХС - ЁСҚНЮҒЮЭ МЭ ЬҚХЮФС ЙЭФД УС ДҚДО ҒСНВСХҚСБ НМЯЮТ, ГЮҲЗ-ЬВЬЯЮИГС ПСЧГ ЎЦЗЮБСЧГФСҲ, СУУСҚСҒТЬБ МЭ ЎДЗ УС ЮҒЎЬҲЮЛГҚСБЮФС ГСЛҲЮТ, ХЦҒЬЎБСГЮЎ ХСУҚСГ УС ЕДНСБЬҚЮЎ ШСҒЮЛГЮ НДБЮИ ЯМҚЮХС ДҚЎСҲ ҲСГЮШСҚСБҲЮ НМҚФС ЎЮБЮГСЧГФСҲ ТЮБ ЭСҒЬҲХС ЛИСҒЬНХСҒЮЭ.

Для восстановления открытого текста в первую очередь составляются частоты символов (таблица 4).

Как видно из таблицы 4 чаще всего встречаются буквы С (38 раз) и Ю (28 раз). Из этого можно предположить, что С=*а* и Ю=*и*. Здесь будем учитывать еще фактор того, что в тексте встречается пробелы и это даёт возможность определить некоторые короткие слова. Например, после замены С=*а*, слово УС заменяется на У*а*. Сочетания У*а* наверняка означает союз *ва*, т.е. символ У необходимо заменить на букву *в*. Тогда с учетом этого и выше сказанного предположения можно получить следующее первые приближения открытого текста:

ТДФДҲ Т*и*Э Г*а*Б*и*Ё*и*Я Т*и*Б Х*а*вБХ*а* - Ё*а*ҚН*и*Ғ*и*Э МЭ ЬҚХ*и*Ф*а* ЙЭФД *ва* ДҚДО Ғ*а*НВ*а*ХҚ*а*Б НМЯ*и*Т, Г*и*ҲЗ-ЬВЬЯ*и*ИГ*а* П*а*ЧГ ЎЦЗ*и*Б*а*ЧГФ*а*Ҳ, *авва*Қ*а*ҒТЬБ МЭ ЎДЗ *ва* *и*ҒЎЬҲ*и*ЛГҚ*а*Б*и*Ф*а* Г*а*ЛҲ*и*Т, ХЦҒЬЎБ*а*Г*и*Ў Х*а*вҚ*а*Г *ва* ЕДН*а*БЬҚ*и*Ў Ш*а*Ғ*и*ЛГ*и* НДБ*и*И ЯМҚ*и*Х*а* ДҚЎ*а*Ҳ Ҳ*а*Г*и*Ш*а*Қ*а*БҲ*и* НМҚФ*а* Ў*и*Б*и*Г*а*ЧГФ*а*Ҳ Т*и*Б Э*а*ҒЬҲХ*а* ЛИ*а*ҒЬНХ*а*Ғ*и*Э.

Таблица 4

Символы	Т	Д	Ф	Ҳ	Ю	Э	Г	С	Б	Ё
Количество	7	9	7	10	28	7	13	38	13	2
Символы	Я	Х	У	Қ	Н	Ғ	М	Ь	Й	О
Количество	4	9	7	12	7	9	5	9	1	1
Символы	В	З	И	Ч	Ў	Ц	Л	Е	Ш	
Количество	2	3	3	3	8	2	4	1	2	

Теперь, обратим внимание на слово *аввaҚaҒТЬБ*. В современном узбекском языке обычно слово, такие как *аввалгидек*, *аввало*, *авваламбор* начинаются с сочетаний *авва,* в целом их немного. Слово *аввaҚaҒТЬБ* попробуем заменить с одной из них. Второе слово *аввало* явно не соответствует по количеству букв. После замены Ю=*и* в сочетание *аввaҚaҒТЬБ* буква *и* не появилась. Поэтому остаётся только одна замена для данного слова: *авваламбор*. Отсюда можно сделать вывод, что Қ=*л*, Ғ=*м*, Т=*б*, Ь=*о*, Б=*р*. С учетом эти замены, второе приближение открытого текста имеет вид:

*б*ДФДҲ *би*Э Г*ариЁи*Я *бир* Х*авр*Х*а* - Ё*ал*Н*ими*Э МЭ *ол*Х*и*Ф*а* ЙЭФД *ва* Д*л*ДО *ма*НВ*а*Х*лар* НМЯ*иб*, Г*и*ҲЗ - *о*В*о*Я*и*ИГ*а* П*а*ЧГ ЎЦЗ*ира*ЧГФ*а*Ҳ, *авваламбор* МЭ ЎДЗ *ва им*Ў*о*Ҳ*и*ЛГ*лари*Ф*а* Г*а*ЛҲ*иб*, ХЦ*мо*Ў*ра*Г*и*Ў Х*авла*Г *ва* ЕДН*ароли*Ў Ш*ами*ЛГ*и* НД*ри*И ЯМ*ли*Х*а* Д*л*Ў*а*Ҳ Ҳ*а*Г*и*Ш*алар*Ҳ*и* НМ*л*Ф*а* Ў*ири*Г*а*ЧГФ*а*Ҳ *бир* Э*амо*ҲХ*а* ЛИ*амо*НХ*ами*Э.

В узбекском языке слова с участием триграмм *авр* немного: *наврўз, бакалавр давра, давр*. Сравнивая слова Х*авр*Х*а* с этими словами можно заметить, что слова Х*авр*Х*а* означает *даврда*, т.е. Х=*д*. Тогда можно предположить, что Х*авла*Г будет означат *давлат*, т.е. имеет место замены Г=*т*.

Среди трёхсимвольных разговорных слов, начинающимся с *би* часто встречаются *биз* и *бир*. С учетом этого слова *би*Э можно принимать, как *биз*. Так, как Э не может заменить букву р (на выше была замена Б=*р*). Тогда МЭ=М*з* означает *ўз*, так как буква М не может быть заменена на *и* или *о* (выше были замены Ю=*и* Ь=*о*).

Таким образом, будем осуществлять замены Х=*д*, Г=*т*, Э=з и М=. Тогда для третьего приближения открытого текста можно иметь:

*б*ДФДҲ *биз тари*Ёи*Я бир даврда* - Ё*ал*Н*имиз ўз олди*Ф*а* ЙзФД *ва* Д*л*ДО *ма*НВ*адлар* Н*ў*Я*иб*, *ти*ҲЗ - *о*В*о*Я*и*И*та* П*а*Ч*т* ЎЦЗ*ира*Ч*т*Ф*а*Ҳ, *авваламбор ўз* ЎДЗ *ва им*Ў*о*Ҳ*и*Л*тлари*Ф*а та*ЛҲ*иб*, д*Ц*м*о*Ў*рати*Ў *давлат ва* ЕДН*ароли*Ў Ш*ами*Л*ти* НД*ри*И Я*ўлида* Д*л*Ў*а*Ҳ Ҳ*ати*Ш*алар*Ҳ*и* Н*ў*л*Ф*а* Ў*ирита*Ч*т*Ф*а*Ҳ *бир замо*Ҳ*да* ЛИ*амо*Н*дамиз*.

Здесь используя вероятных слов можно восстановить некоторые целые слова. Например, Я*ўлида* – *йўлида* (Я=*й*), *тари*Ёи*Я* – *тарихий* (Ё=*х*), Ё*ал*Н*имиз* –*халқимиз* (Н=*қ*), д*Ц*м*о*Ў*рати*Ў – де*мократик* (Ц=*е*, Ў=*к*). С учетом этих замен четвёртое приближение открытого текста имеет вид:

*б*ДФДҲ *биз тарихий бир даврда - халқимиз ўз олди*Ф*а* Й*з*ФД *ва* Д*л*ДО *мақ*В*адлар қўйиб*, *ти*ҲЗ - *о*В*ойли*И*та* П*а*Ч*т* кЕЗ*ира*Ч*т*Ф*а*Ҳ, *авваламбор ўз* кДЗ *ва имко*Ҳ*и*Л*тлари*Ф*а та*ЛҲ*иб*, *демократик давлат ва* ЕД*қаролик* Ш*ами*Л*ти* к*Дриш йўлида* Д*лка*Ҳ Ҳ*ати*Ш*алар*Ҳ*и* қ*ў*л*Ф*а* к*ирита*Ч*т*Ф*а*Ҳ *бир замо*Ҳ*да* ЛИ*амоқдамиз*.

Теперь, обратим внимание на слова *мақВадлар*, отсюда видно, что необходимо осуществлять замену В=*с*. Тогда слов *оВойли*И*та* и ЛИ*амоқдамиз* превращаются соответственно на *осойли*И*та* и Л*шамоқдамиз*. В этом случае

имеют место замены И=*ш* и Л=*я*. Далее к*Дриш означает қуриш* (т.е. Д=*у*), а также *замоҲда* соответствует на *замонда* (т.е. Ҳ=*н*). С учетом этих замен в окончательном виде можно получить следующий открытый текст:

Бугун биз тарихий бир даврда - халқимиз ўз олдига эзгу ва улуғ мақсадлар қўйиб, тинч-осойишта ҳаёт кечираётган, аввалambор ўз куч ва имкониятларига таяниб, демократик давлат ва фуқаролик жамияти қуриш йўлида улкан натижаларни қўлга киритаётган бир замонда яшамоқдамиз (И.Каримов. Юксак маънавият - енгилмас куч.).

III. ЗАКЛЮЧЕНИЕ

Установлены относительные частоты (вероятности появления) букв кириллицы на узбекском языке.

Определены биграммы и триграммы, имеющие наибольшую вероятность узбекского языка.

Дешифрован шифртекст, зашифрованный моноалфавитным симметричным шифром с помощью частотного анализа.

Для полиалфавитных симметричных шифров частотный криптоанализ следует применять для каждого алфавита шифр текста отдельно.

ЛИТЕРАТУРА

[1] Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В. Основы криптографии: Учебное пособие. – М.: Гелиос АРВ, 2002. 480 с.

[2] Бабаш А.В., Шанкин Г.П. Криптография. – М.: СОЛОН-ПРЕСС, 2007. – 512 с.

[3] Жданов О.Н., Куденкова И.А. Криптоанализ классических шифров. – Красноярск, 2008. 107 с.

[4] Каримов И.А. Ўзбекистон XXI аср бўсағасида: хавфсизликка таҳдид, барқарорлик шартлари ва тараққиёт кафолатлари. – Т.: Ўзбекистон, 1997.

[5] Орипов А. Йиллар армони: Шеърлар ва достонлар. -Т.: Адабиёт ва санъат нашриёти, 1987. -592 б.