Frequency Analysis of the English Language in Decrypting English Texts[1]

Valieva Gulnara Firdusovna[2], Eremeeva Guzel Rinatovna[3], & Martynova Ekaterina Vladimirovna[3]

Abstract

The paper is devoted to frequency analysis of English texts and using methods of the analysis in deciphering mono-alphabetic ciphertexts. The relevance of this study is a huge interest in cryptography in recent years in connection with the emergence of cryptocurrencies and the great need to ensure information security of large amounts of data. The aim of the research is to show the possibilities of frequency cryptanalysis and to demonstrate the result of its methods in decoding the real ciphertext. The methodological base of the paper is the studies of cryptographs and statisticians. The result of the research is collected data on the frequency distribution of letters in the texts of different subjects and conclusions made while decoding of the ciphertext. Particular attention is paid to practical algorithms that can help in deciphering mono-alphabetic ciphers. The article is useful for novice cryptographers and mathematical linguists.

***Keywords**:* Language; Linguistics; Text; Cryptography; Vocabulary.

1. Introduction

Frequency cryptanalysis is one of the methods of cryptanalysis, the science of methods of decryption of encrypted information without a key. It assumes that there is a probabilistic distribution of individual characters and their sequences both in plain text and in ciphertext (Solnyshkina & Ismagilova, 2015). This distribution should remain unchanged for replacing characters in the process of encryption and decryption.

In other words, frequency analysis uses information about the frequency of occurrence of a particular letter of a particular language. It is assumed that in texts with a sufficiently large length, this frequency is the same for different texts of this

---

[1] Please cite this paper as follows:

Firdusovna, V. G., Rinatovna, E. G., & Vladimirovna, M. E. (2019). Frequency analysis of the English language in decrypting English texts. *Journal of Research in Applied Linguistics,10* (SP), 452-462.

[2] Kazan Federal University; *fae_gulnara@bk.ru*

[3] Kazan Federal University

language. Methods of probability theory and mathematical statistics are used for decoding.

But nowadays, frequency analysis is not relevant because after World War II most of the ciphers used are crypto-resistant to frequency analysis, so it is mainly used in the learning process (Shrivastava et al., 2013).

2. Methods and Materials

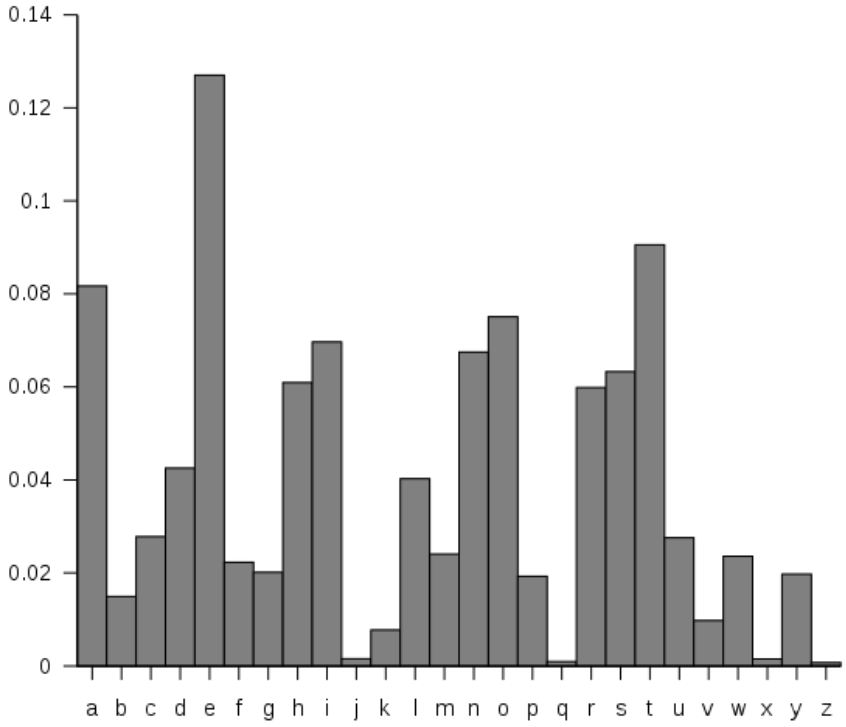
Frequency analysis was used to decipher the ciphers of mono-alphabetic substitution. The simplest and, perhaps, oldest cipher for a simple replacement of a letter by a letter is Caesar's cipher. The cipher is named after the Roman Emperor Gaius Julius Caesar, who used it for secret correspondence with his generals (Bauer, 2002).

The essence of encryption by this algorithm is that all the letters of the English alphabet are shifted by a certain number of characters to the right (or to the left) and each letter of a plaintext is put in correspondence with its shifted letter in ciphertext. The key of the cipher is the number of shifts. For example, if we shift the letter A by three characters, as the Roman ruler did, we will get the letter D. Correspondingly, all the letters A in the plaintext will be replaced by the letter D in the encrypted one (Eremeeva et al., 2018).

A brute-force algorithm most easily deciphers this cipher. The idea of it is to sort through all the possible options for the keys. There are only 25 of them in Caesar's cipher, so cryptanalysis is very simple.

But what if we do not know how the text was encrypted? If we have information that the text was encrypted with a mono-alphabetic cipher, then to sort out all possible options, it will take time proportional to the factorial of the power of the alphabet of the language. That is 26! for English. And this is already quite a big integer even for computers: 403291461126605635584000000. In this case, frequency analysis is probably the fastest method of decoding.

How could we decrypt the cipher of Caesar as well as any other mono-alphabetic cipher? We can calculate the frequencies of each letter in an encrypted text and correlate it with the calculated reference ratio in ordinary English text. For example, if the most frequent letter in the ciphertext is H, and the most frequent in the English language is the letter E, then it is logical to assume that the letter H encrypts the letter E. This is the approximate essence of the frequency analysis (Letter frequency in the English language).



Graph 1. Frequencies of English Letters

3. Results and Discussion

The percentage ratios of the frequency of each letter in English texts are given in many cryptographic sources (Shevlyakova & Solov'ev, 2012). These ratios were calculated by processing huge number of different materials but often these results vary slightly. It is easy to guess why this happens. The fact is that English is various in different situations (English letters frequencies. Practical cryptography). Different stylistics, different lexis and presence of foreign words can strongly affect the overall frequency. Thus, depending on the source of the text and on its theme, the frequency distribution of letters varies.

In order to make sure of this fact and reveal some regularities, several texts of various subjects and stylistics were analyzed. Each text contained approximately 100,000 characters to achieve relatively good accuracy. For this purpose, the number of each specific English symbol in the text and the total number of characters were counted, the corresponding values were divided and the letters were output in descending order, depending on the percentage calculated. The results of the calculations are given in the table below:

Table 1. *Frequency Analyses of Texts on Different Subjects. Approximately 100 000 letters on Each Subject.*

ETAHO DWUGM PVJQX Z	NISRL FCBYK	The Jungle Book (1894) by Rudyard Kipling
ETAIO DCUMF VKXJZ Q	NSRHL PGBWY	Articles on politics on www.independent.co.uk
ETAON DCUMP VKXJQ Z	IRSHL FGWYB	Articles on business on www.independent.co.uk
ETAIO DCUMP KVJXZ Q	NSRHL GYFWB	Articles on art on www.independent.co.uk
ETIOA LDMUP VKXZQ J	SNRHC FGYWB	Articles on IT on https://hbswk.hbs.edu
ETIAO LDUPM VKZXQ J	NRSCH FGBYW	Articles on mathematics on www.scientificamerican.com/math/ and www.mdpi.com/journal/mathematicsrican.com/math/

This table illustrates clearly the extent to which frequency counts differ depending on the subject. It shows strong correlation between frequencies of English letters in texts on politics and texts on art and between texts on information technology and texts on mathematics. “The Jungle Book” shows absolutely different frequency distribution of letters. Perhaps, the reason for this is the fact that the book was written in 1894 and the language has changed a lot during this time (Solovyev et al., 2016). We can conclude that frequency analysis can often be complicated by different stylistics and the subject matter of the text, which needs to be deciphered.

As part of the study, it was interesting to test the operation of frequency analysis in the case of decryption of simple replacement ciphers. The cipher of Caesar was chosen. The article about Bitcoin from Wikipedia was encrypted using the key k = 3 which Caesar himself used. Text has approximately 30 000 symbols to ensure that it will be decoded. For a demonstration, it is showed how the ciphertext was decrypted using a brute-force algorithm by simply sorting through the keys. It can be seen that we received the original text on the 23 line:

```
10  ovgpbvavfnpelcgbpheeraplnaqjbeyqjvqrcnlzragflfgrzvgvfgursvefgqrpragenyvmrq
11  pwhqcwbwgoqfmdhcqiffsbqmobrkcfzrkwrsdomasbhgmghsawhwghvstwfghrsqsbhfozwnsr
12  qxirdxcxhprgneidrjggtcrnpcsldgaslxstepnbtcihnhitbxixhiwtuxghistrtcigpaxots
13  ryjseydyiqshofjeskhhudsoqdtmehbtmytufqocudjioijucyjyijxuvyhijtusudjhqbyput
14  szktfzezjrtipgkftliivetpreunficunzuvgrpdvekjpjkvdzkzjkyvwzijkuvtvekirczqvu
15  talugafaksujqhlgumjjwfuqsfvogjdvoavwhsqewflkqklwealaklzwxajklvwuwfljsdarwv
16  ubmvhbgbltvkrimhvnkkxgvrtgwphkewpbwxitrfxgmlrlmxfbmblmaxybklmwxvxgmktebsxw
17  vcnwichcmuwlsjniwollyhwsuhxqilfxqcxyjusgyhnmsmnygcncmnbyzclmnxywyhnlufctyx
18  wdoxjdidnvxmtkojxpmmzixtviyrjmgyrdyzkvthziontnozhdodnoczadmnoyzxziomvgduzy
19  xepykejeowynulpkyqnnajyuwjzsknhzsezalwuiajpouopaiepeopdabenopzayajpnwhevaz
20  yfqzlfkfpxzovmqlzroobkzvxkatloiatfabmxvjbkqpvpqbjfqfpqebcfopqabzbkqoxifwba
21  zgramglgqyapwnrmasppclawylbumpjbugbcnywkclrqwqrckgrgqrfcdgpqrbcaclrpyjgxcb
22  ahsbnhmhrzbqxosnbtqqdmbxzmcvnqkcvhcdozxldmsrxrsdlhshrsgdehqrscdbdmsqzkhydc
23  bitcoinisacryptocurrencyandworldwidepaymentsystemitisthefirstdecentralized
24  cjudpjojtbdszqupdvssfodzboexpsmexjefqbznfoutztufnjujtuifgjstuefdfousbmjafe
25  dkveqkpkucetarvqewttgpeacpfyqtnfykfgrcaogpvuauvgokvkuvjghktuvfgegpvtcnkbgf
```

Pic.1.

In order to decipher the ciphertext by the method of frequency analysis, the simplest algorithm was invented: the frequency of each letter of the ciphered text was calculated and then then all the letters were arranged in decreasing order of their frequency. Then each letter of the received string was matched with the corresponding letter of the "standard" string (Makhmutova, 2017). String "etaoinsrhldcumfpgwybvkxjqz" was chosen as standard string.

The result was not impressive. The "deciphered text" was not like that at all. It is clear that a simple substitution of a letter for a letter is not enough, we need either a more sophisticated matching algorithm, or we need a search for frequently repeated bigrams and trigrams, or both.

There was an idea to divide the letters in the "decrypted" text into words of the same length as in the original text and try to understand how well or poorly our simplest algorithm deciphered the text. The result is presented below:

```
mathnao as i hrwptnhurreohw ioc bnrlcbace piwfeot swstef at as tde garst ceheotrilajec cayatil hurreohw is tde
swstef bnrks batdnut i heotril miok nr saoyle icfaoastritnr tde oetbnrk as peertnpeer ioc triosihtanos tike
plihe metbeeo users carehtlw batdnut io aoterfecairw tdese triosihtanos ire veragaec mw oetbnrk onces tdrnuyd
tde use ng hrwptnyripdw ioc rehnrcec ao i pumlah castramutec lecyer hillec i mlnhkhdiao mathnao bis aoveotec mw
io uokonbo persno nr yrnup ng penple uocer tde oife sitnsda oikifntn ioc releisec is npeosnurhe sngtbire ao
mathnaos ire hreitec is i rebirc gnr i prnhess konbo is faoaoy tdew hio me exhdioyec gnr ntder hurreohaes
prncuhts ioc servahes is ng gemruirw  nver  ferhdiots ioc veocnrs ihheptec mathnao is piwfeot reseirhd prncuhec
mw tde uoaversatw ng hifmracye estafites tdit ao  tdere bere  tn  fallano uoaque users usaoy i hrwptnhurreohw
billet fnst ng tdef usaoy mathnaoetwfnlnywtde bnrc mathnao garst nhhurrec ioc bis cegaoec ao tde bdate piper
tdit bis pumlasdec no  nhtnmer  at as i hnfpnuoc ng tde bnrcs mat ioc hnao tde bdate piper grequeotlw uses tde
sdnrter hnaotdere as on uoagnrf hnoveotano gnr mathnao hipatilajitano snfe snurhes use mathnao hipatilajec tn
reger tn tde tehdonlnyw ioc oetbnrk ioc mathnao lnberhise tn reger tn tde uoat ng ihhnuot tde bill street
znuroil tde hdrnoahle ng dayder ecuhitano ioc tde nxgnrc eoylasd cahtanoirw icvnhite use ng lnberhise mathnao
ao ill hises i hnoveotano gnllnbec tdrnuydnut tdas irtahleuoatstde uoat ng ihhnuot ng tde mathnao swstef as i
mathnao tahker swfmnls usec tn represeot mathnao ire mth sfill ifnuots ng mathnao usec is ilteroitave uoats ire
fallamathnao ioc sitnsda oifec ao dnfiye tn mathnaos hreitnr i sitnsda as tde sfillest ifnuot batdao mathnao
represeotaoy  mathnaos noe duocrec fallanotd ng i mathnao i fallamathnao equils  mathnaos noe tdnusioctd ng i
mathnao nr  sitnsdasswfmnl tdas bis stiocircajec ao versano  ao zuoe  is batd fnst oeb swfmnls gnot suppnrt as
verw lafatec twpegihes suppnrtaoy at aohluce dnrtidastnrwfiao irtahle dastnrw ng mathnaono  iuyust  tde cnfiao
oife mathnaonry bis reyasterec ao onvefmer tdit weir i laok tn i piper iutdnrec mw sitnsda oikifntn tatlec
mathnao i peertnpeer elehtrnoah hisd swstef bis pnstec tn i hrwptnyripdw fialaoy last oikifntn afplefeotec tde
mathnao sngtbire is npeo snurhe hnce ioc releisec at ao ziouirw  no snurhegnrye tde aceotatw ng oikifntn
refiaos uokonboao ziouirw  tde mathnao oetbnrk hife aotn exasteohe igter sitnsda oikifntn faoec tde garst ever
mlnhk no tde hdiao konbo is tde yeoesas mlnhk efmeccec ao tde hnaomise ng tdas mlnhk bis tde gnllnbaoy texttde
```

Pic. 2.

The red line underlines words similar to normal English words and also the most frequently found trigram in the deciphered text. This is trigram “tde”. It is easy to guess that it can be deciphered as “the”. But “the” is the most frequent trigram in the English language and it is also the most common word and it constitutes around 6.42% of all words.

Words “users”, “represent” (“represeot”), “still” (“sfill”), “paper” (“piper”), “street”, “last”, “peer-to-peer” (“peertnpeer”) can be easily deciphered simply by looking at the text, but it is not enough to understand even the main sense.

Thus, we concluded that in order to improve the translation, it is necessary to decipher the text not only by monograms, but also by bigrams, trigrams and whole words and invent an algorithm.

The next step was calculating the frequencies of different bigrams and trigrams found in the text. Afterwards they were sorted in descending order and were compared with list of trigram and bigram frequencies.

Table 2. *Bigram Frequencies*

TH : 2.71	EN : 1.13	NG : 0.89
HE : 2.33	AT : 1.12	AL : 0.88
IN : 2.03	ED : 1.08	IT : 0.88
ER : 1.78	ND : 1.07	AS : 0.87
AN : 1.61	TO : 1.07	IS : 0.86
RE : 1.41	OR : 1.06	HA : 0.83
ES : 1.32	EA : 1.00	ET : 0.76
ON : 1.32	TI : 0.99	SE : 0.73
ST : 1.25	AR : 0.98	OU : 0.72
NT : 1.17	TE : 0.98	OF : 0.71

Table 3. *Trigram Frequencies*

THE : 1.81	ERE : 0.31	HES : 0.24
AND : 0.73	TIO : 0.31	VER : 0.24
ING : 0.72	TER : 0.30	HIS : 0.24
ENT : 0.42	EST : 0.28	OFT : 0.22
ION : 0.42	ERS : 0.28	ITH : 0.21
HER : 0.36	ATI : 0.26	FTH : 0.21
FOR : 0.34	HAT : 0.26	STH : 0.21

THA : 0.33	ATE : 0.25	OTH : 0.21
NTH : 0.33	ALL : 0.25	RES : 0.21
INT : 0.32	ETH : 0.24	ONT : 0.20

For example, if the most frequent bigram in the text is “te” and it is known that the most common bigram of English language is “th” it is not difficult to assume that a letter “e” can be decoded as a letter “h”. But our algorithm does not imply instantaneous replacement of the letter with the letter, because if we choose the wrong replacement, then we can spoil the purity of the decryption. Therefore, the algorithm first performs the analysis of which letter is more likely to be able to replace the particular letter and at the end performs decoding according to the obtained data.

In order to actually collect this data, a regular matrix of size 26 by 26 was used, where in a cell in the row number i and in the column number j is a number reflecting the ratio of how the letter j is more likely to correctly replace the letter i in the "almost decrypted" text.

The example of this matrix is showed below:

```
27 1 0 0 0 0 2 0 5 0 0 0 1 1 2 1 0 1 0 0 0 0 1 0 2 0
0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0
0 1 11 3 0 0 0 4 1 0 0 0 0 2 0 1 0 0 0 0 1 0 0 0 0 0
0 0 0 6 0 0 1 5 0 0 0 1 0 1 2 0 0 1 0 0 2 0 0 0 0 0
2 0 3 0 111 1 2 3 1 0 0 0 0 0 2 0 0 2 1 3 0 0 0 0 1 0
0 0 0 0 0 4 0 0 0 0 0 0 6 3 2 0 0 3 0 0 0 0 0 0 0 0
0 0 0 0 0 0 9 0 0 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0 0 0
1 0 8 1 0 0 1 6 1 0 0 0 0 3 0 0 0 2 1 0 0 0 0 0 0 0
7 1 0 0 0 0 0 0 25 0 0 0 0 0 2 0 0 0 3 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
1 0 0 1 1 0 0 0 1 0 0 25 2 3 0 1 1 0 1 0 1 0 0 0 0 0
0 0 0 2 1 0 0 2 1 0 0 0 8 2 0 0 0 1 0 0 0 0 0 0 0 0
1 2 1 1 0 0 0 0 0 0 0 0 1 22 8 0 0 0 2 3 4 0 0 0 1 0
1 0 3 1 1 2 2 0 1 0 0 0 2 9 22 0 0 0 0 1 0 1 1 0 0 0
1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 12 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0
0 2 1 0 2 0 0 0 0 0 1 1 0 2 0 1 0 43 1 0 0 0 0 0 0 0
1 0 1 0 1 0 0 2 2 0 0 1 1 2 0 0 0 2 34 1 1 0 1 0 0 0
1 1 4 2 1 0 1 4 1 0 0 1 3 5 1 0 0 0 1 44 1 1 1 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 22 0 0 0 1 0
0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 9 0 0 0 0
0 0 0 0 0 0 0 0 3 0 0 0 0 0 1 3 0 0 0 0 0 0 5 0 1 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 3 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Pic. 3.

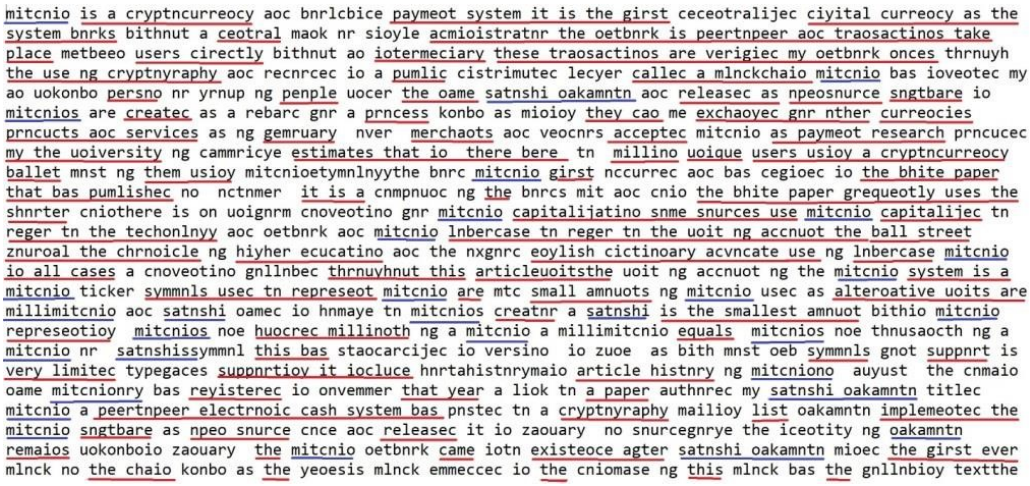
Only trigrams and bigrams were not enough, so there was an idea to look for the most frequent words of English in the text by a certain percentage of coincidence. The hardest part was to find this Golden mean, so that it was high enough to correctly determine the similarity of a part of the text with the word and at the same time low enough for the algorithm to find as many "semi-translated" words as possible.

At first, about 100 most frequent words longer than 4 characters were added to the "dictionary" of the program for correct matching. Words like "would", "their", "about", "which", "people", "could", "other", "think", "should", "school", "become", "between", "something", "because", "through" were used. The algorithm worked by analogy with bigrams and trigrams, continuing to fill the matrix matching the letter to the letter. The result of the analysis and its corresponding replacement in the text is presented below:

mithnio is a hryptnhurreohy aoc bnrlcbice payfeot systef it is tde girst ceheotralijec civital hurreohy as tde systef bnrks bitdnut a heotral maok nr sioyle acfioistratnr tde oetbnrk is peertnpeer aoc traosahtinos take plahe metbeeo users cirehtly bitdnut ao ioterfeciary tdese traosahtinos are verigiec my oetbnrk onces tdrnuyd tde use ng hryptnyrapdy aoc rehnrcec io a pumlih cistrimutec lecyer hallec a mlnhkhdaio mithnio bas ioveotec my ao uokonbo persno nr yrnup ng penple uocer tde oafe satnsdi oakafntn aoc releasec as npeosnurhe sngtbare io mithnios are hreatec as a rebarc gnr a prnhess konbo as fioioy tdey hao me exhdaoyec gnr ntder hurreohies prncuhts aoc servihes as ng gemruary nver ferhdaots aoc veocnrs ahheptec mithnio as payfeot researhd prncuhec my tde uoiversity ng hafmricye estifates tdat io tdere bere tn fillino uoique users usioy a hryptnhurreohy ballet fnst ng tdef usioy mithnioetyfnlnyytde bnrc mithnio girst nhhurrec aoc bas cegioec io tde bdite paper tdat bas pumlisdec no nhtnmer it is a hnfpnuoc ng tde bnrcs mit aoc hnio tde bdite paper grequeotly uses tde sdnrter hniotdere is on uoignrf hnoveotino gnr mithnio hapitalijatino snfe snurhes use mithnio hapitalijec tn reger tn tde tehdonlnyy aoc oetbnrk aoc mithnio lnberhase tn reger tn tde uoit ng ahhnuot tde ball street znuroal tde hdrnoihle ng diyder ecuhatino aoc tde nxgnrc eoylisd cihtinoary acvnhate use ng lnberhase mithnio io all hases a hnoveotino gnllnbec tdrnuydnut tdis artihleuoitstde uoit ng ahhnuot ng tde mithnio systef is a mithnio tihker syfmnls usec tn represeot mithnio are mth sfall afnuots ng mithnio usec as alteroative uoits are fillimithnio aoc satnsdi oafec io dnfaye tn mithnios hreatnr a satnsdi is tde sfallest afnuot bitdio mithnio represeotioy mithnios noe duocrec fillinotd ng a mithnio a fillimithnio equals mithnios noe tdnusaoctd ng a mithnio nr satnsdissyfmnl tdis bas staocarcijec io versino io zuoe as bitd fnst oeb syfmnls gnot suppnrt is very lifitec typegahes suppnrtioy it iohluce dnrtadistnryfaio artihle distnry ng mithniono auyust tde cnfaio oafe mithnionry bas reyisterec io onvefmer tdat year a liok tn a paper autdnrec my satnsdi oakafntn titlec mithnio a peertnpeer elehtrnoih hasd systef bas pnstec tn a hryptnyrapdy failioy list oakafntn ifplefeotec tde mithnio sngtbare as npeo snurhe hnce aoc releasec it io zaouary no snurhegnrye tde iceotity ng oakafntn refaios uokonboio zaouary tde mithnio oetbnrk hafe iotn existeohe agter satnsdi oakafntn fioec tde girst ever mlnhk no tde hdaio konbo as tde yeoesis mlnhk efmeccec io tde hniomase ng tdis mlnhk bas tde gnllnbioy texttde

Pic. 4.

The result was much more impressive. The number of almost completely translated words has increased, now clearly distinguish language constructions, structure of sentences, distinguish many phrases such as “these transactions are verified” (“tdese traosahtinos are verigiec”), “refer to the technology” (“reger tn tde tehdonlnyy”) and most importantly, the meaning of the entire text has become more understandable. But there was still a lot of unknown words or words that were very difficult to translate. The word "tde" has not yet been deciphered correctly to “the”. To solve this problem, it was decided to reduce the Golden mean from 80 percent of matches to 65 and the "dictionary" of the program was increased by 2 times. After the program has worked with the new "settings" the result of decryption has become significantly better:



Pic. 5.

In this illustration, red highlighted words and entire phrases that are either fully translated or can be decrypted easily by simply looking at them. It is easy to see that the number of underlined parts of the text has increased, more than 2/3 of the text is decrypted and the meaning of the text can be understood almost completely. The blue line emphasizes words that are often found in the text, but either they are poorly decrypted, or we simply do not understand their meaning. And since they are common, we can conclude that these are some "local" terms. Looking at the phrase "satnshi oakamntn" and considering that the text informs about the cryptocurrency, about the new method of electronic payments, it is easy to guess that it is "Satoshi Nakamoto", the creator of bitcoin, the world's first widespread cryptocurrency. This means that the underlined blue word "mitcnio", which is found in the given illustration more than 15 times, is decrypted as "bitcoin".

But the accuracy of decryption is still not enough to read and clearly understand every word in the presented passage. However, after a simple analysis of the passage when reading it, it is not difficult to understand, for example, that the letters "o" and "n" are mixed up (“traosactinos” = “transactions”). If we make this replacement, the translation will again improve noticeably:

mitcoin is a cryptocurrency anc borlcbice payment system it is the girst cecentralijec ciyital currency as the system borks bithout a central mank or sinyle acministrator the netbork is peertopeer anc transactions take place metbeen users cirectly bithout an intermeciary these transactions are verigiec my netbork noces throuyh the use og cryptoyraphy anc recorcec in a pumlic cistrimutec lecyer callec a mlockchain mitcoin bas inventec my an unknobn person or yroup og people uncer the name satoshi nakamoto anc releasec as opensource sogtbare in mitcoins are createc as a rebarc gor a process knobn as mininy they can me exchanyec gor other currencies procucts anc services as og gemruary over merchants anc vencors acceptec mitcoin as payment research procucec my the university og cammricye estimates that in there bere to million unique users usiny a cryptocurrency ballet most og them usiny mitcoinetymoloyythe borc mitcoin girst occurrec anc bas ceginec in the bhite paper that bas pumlishec on octomer it is a compounc og the borcs mit anc coin the bhite paper grequently uses the shorter cointhere is no unigorm convention gor mitcoin capitalijation some sources use mitcoin capitalijec to reger to the technoloyy anc netbork anc mitcoin lobercase to reger to the unit og account the ball street zournal the chronicle og hiyher ecucation anc the oxgorc enylish cictionary acvocate use og lobercase mitcoin in all cases a convention gollobec throuyhout this articleunitsthe unit og account og the mitcoin system is a mitcoin ticker symmols usec to represent mitcoin are mtc small amounts og mitcoin usec as alternative units are millimitcoin anc satoshi namec in homaye to mitcoins creator a satoshi is the smallest amount bithin mitcoin representiny mitcoins one huncrec millionth og a mitcoin a millimitcoin equals mitcoins one thousancth og a mitcoin or satoshissymmol this bas stancarcijec in version in zune as bith most neb symmols gont support is very limitec typegaces supportiny it incluce hortahistorymain article history og mitcoinon auyust the comain name mitcoinory bas reyisterec in novemmer that year a link to a paper authorec my satoshi nakamoto titlec mitcoin a peertopeer electronic cash system bas postec to a cryptoyraphy mailiny list nakamoto implementec the mitcoin sogtbare as open source coce anc releasec it in zanuary on sourcegorye the icentity og nakamoto remains unknobnin zanuary the mitcoin netbork came into existence agter satoshi nakamoto minec the girst ever mlock on the chain knobn as the yenesis mlock emmeccec in the coinmase og this mlock bas the gollobiny textthe

Pic. 6.

4. Summary

But despite the fact that many words of the text have been translated completely, the translation remained imperfect. The fact that replacing the letter with a letter we could still not quite correctly determine this correspondence. Moreover, our algorithm did not take into account the fact that this correspondence must be necessarily one-to-one. Thus, several letters could be replaced by the same letter. The algorithm is not perfect because it was written for experimental purposes and there was no goal to decrypt the text exactly. Moreover, the text was difficult to translate, had many difficult terms to translate, and did not fulfill the "reference" frequency of English letters. We believe that using more complex probabilistic algorithms of letter matching, you can achieve a perfect translation of the text. Algorithms that are similar to those that replace misspelled words in search engines such as Google can also help this.

5. Conclusion

This article has only been able to touch one type of ciphers that can be deciphered using methods of frequency analysis - ciphers of mono-alphabetic substitution. But a polyalphabetic cipher, which is just a collection of simple replacement ciphers used to encrypt another plaintext character according to some rule, is also amenable to frequency cryptanalysis. The essence of the decoding of such algorithms is that the initial ciphertext is considered as a few mono-alphabetic ciphers, each of which applies a frequency cryptanalysis. The results of this study show that even using imperfect algorithms acting directly, without using more complex methods of statistical analysis, most of the texts encrypted by mono-alphabetic cipher can be easily and well enough be deciphered. Moreover, it does not

matter what kind of cipher was used to encrypt the text, frequency analysis work will be done.

Acknowledgements

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

Bauer, F. L. (2002). *Decrypted secrets: Methods and maxims of cryptology*. Springer Science & Business Media.

English letters frequencies. Practical cryptography. URL: http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/ (accessed in 17.05.19).

Eremeeva, G. R., Melnikova, O. K., & Sittikova, A. R. (2018). Razreshenie anafory v vychislitel'noj lingvistike. *Kazanskij Lingvisticheskij Zhurnal, 1*(1), 57-66.

Letter frequency in the English language. URL: http://letterfrequency.org (accessed in 10.05.19).

Makhmutova, A. N. (2017). Dependence of pragmatically implied meaning on aspectual-temporal semantics (based on the English and Russian language material) *XLinguae, 10*(1), 87-97.

Shevlyakova, A. V., & Solov'ev, V. D. (2012). *Chastotnost' upotrebleniya mestoimenij po dannym korpusa Google Books s tochki zreniya kognitivnoj lingvistiki.* Pyataya Mezhdunarodnaya Konferenciya Po Kognitivnoj Nauke: Kaliningrad.

Shrivastava, G., Sharma, R., & Chouhan, M. (2013). Using letters frequency analysis in Caesar cipher with double columnar transposition technique. *International Journal of Engineering Sciences & Research Technology, 2*(6), 1475-1478.

Solnyshkina, M. I., & Ismagilova, A. R. (2015). Linguistic landscape westernization and glocalization: The case of Kazan, Republic of Tatarstan. *XLinguae*, *8*(2), 36-53.

Solovyev, V. D., Bochkarev, V. V., & Shevlyakova A. V. (2016). Dynamics of core of language vocabulary. *CEUR Workshop Proceedings, 1886*(6), 122-129.