Attacking Turkish Texts Encrypted by Homophonic Cipher

SEFIK ILKIN SERENGIL [1]   MURAT AKIN [2]
[1, 2] Institute of Science
Galatasaray University
Ciragan Cad No:36 34357 Ortakoy, Istanbul
TURKEY
[1] 09411002@ogr.gsu.edu.tr [2] murakin@gsu.edu.tr

Abstract: Homophonic cipher is developed as an alternative to substitution cipher to compose more resistant ciphertexts against to the frequency analysis attacks. Nevertheless, Attacking with taking advantage of characteristic vulnerabilities of the language is probable. In this paper, characteristic vulnerabilities of the Turkish Language for homophonic cipher are exposed and attacking approaches are illustrated.

Key-Words: Turkish n-gram Frequencies, Frequency Analysis Attacks, Homophonic Substitution, Monoalphabetic Substitution, Characteristic Vulnerabilities of the Turkish Language, Cryptoanalysis of Turkish Enciphered Texts.

1 Introduction

The idea of including homophony into cryptography thought as making stronger ciphers against frequency analysis attacks at the beginning. Homophonic cipher could be thought as extended version of substitution cipher. Homophonic cipher replaces each plaintext letter with different symbols proportional to its frequency rate. The frequency distribution of the ciphertext is manipulated and smoothed. Symbols located in ciphertext have relatively equal frequencies. Each symbol takes space of about one percent of ciphertext. That’s why, it would be securer than a substitution cipher. Initially, ciphertext could be thought to resist any potential frequency analysis attack. However, homophonic enciphered texts still contain vulnerabilities and they are indirectly weak against to frequency analysis attack. Firstly, low frequent letters would repeat in a sufficiently long ciphertexts. Secondly, it would be taken advantage of the characteristic vulnerabilities of the source language.

At this point, Turkish is one of the least studied language. Related work by Dalkılıç [1] on the cryprographic patterns and frequencies in Turkish language investigates language patterns and frequencies of Turkish. That work could contribute solving homophonic ciphers but the study hasn’t gone beyond the extraction of most frequent trigrams and contains limited information. Moreover, tetragrams and pentagrams could play key role to solve homophonic enciphered texts but these information is almost unknown for Turkish. Above all, there are not previous studies on this subject for Turkish.

In this paper, firstly high frequent n-grams while n is less than, or equal to 5 are explored and secondly useful n-grams are illustrated to analysis of homophonic ciphers for Turkish. Data presented in this article collected from the data source size of 13.4 MB and the data source consists of 120 articles of a columnist, *Çetin Altan*, from the Turkish daily newspaper Milliyet and 37 novels of 9 different authors, which are *Orhan Kemal*, *Orhan Pamuk*, *Çetin Altan*, *Aziz Nesin*, *Rıfat Ilgaz*, *Gülse Birsel*, *Ahmet Altan*, *Yılmaz Erdoğan* and *Soner Yalçın*.

2 Cryptoanalysis of Homophonic Cipher

In order to solve homophonic ciphers, making a decision of useful n-grams belongs to source language plays pivotal role. The unigrams of n-grams should have low frequencies to be determined easily in homophonic encipherd texts, whereas the n-gram itself should have high frequency to be assumed to appear in the plaintext. In other words, high frequent n-grams should consist of low frequent unigrams.

ISBN: 978-960-474-276-9

For instance, most frequent trigrams are "*lar*"(% 0,0078), "*bir*"(% 0,0067) and "*ler*"(% 0,006) in Turkish. However, "*lar*" would be expressed by 504(6x12x7) different symbols. Similarly, "*bir*" and "*ler*" would be shown by 189 and 378 symbols. Even if these trigrams are assumed to appear in ciphertext, it would almost be impossible to solve. That's why most frequent n-grams could not directly assist to solve homophonic ciphers. In contrast, the trigram of "*gör*"(% 0,001) has high frequency, too. That's why, the trigram could be assumed to appear in the plaintext. Furthermore, it also would be expressed by 7 different symbols in Turkish homophonic enciphered texts. If the trigram had compared to most frequent trigrams specified above, it could have said that detecting the trigram would be much easier.

2.1 Unigram Frequencies

The unigram frequencies of the source language assesses how many symbols the letter would be expressed within homophonic cipher. Each letter would be replaced by different symbols proportional to its frequency rate.

Table 1. Turkish Unigram Frequencies and Replacing Values in Homophonic Cipher

A	%11,92	12	I	%5,114	5	R	%6,722	7
B	%2,844	3	İ	%8,6	9	S	%3,014	3
C	%0,963	1	J	%0,034	1	Ş	%1,78	2
Ç	%1,156	1	K	%4,683	5	T	%3,314	3
D	%4,706	5	L	%5,922	6	U	%3,235	3
E	%8,912	9	M	%3,752	4	Ü	%1,854	2
F	%0,461	1	N	%7,487	7	V	%0,959	1
G	%1,253	1	O	%2,476	2	Y	%3,336	3
Ğ	%1,125	1	Ö	%0,777	1	Z	%1,5	2
H	%1,212	1	P	%0,886	1			

2.2 High Frequent n-grams Consisting of Low Frequent Unigrams

Firstly, we explore Turkish n-gram frequencies and obtain a table consists of n-gram and frequency columns for each n. Then, a virtual column named as "*symbol*", which indicates how many symbols the n-gram will be expressed within homophonic cipher by the use of unigram frequencies, was created. Then, initial sorting was done with respect to the frequeny column by taking into account the first 250 records for bigrams, 1500 results for trigrams, 2500 results for tetragrams and pentagrams from the greatest to smallest. Thirdly, this new table was sorted with respect to the symbol column from the smallest to greatest. Finally, the values demonstrated in the tables obtained from this way. Since, it is needed to solve homophonic ciphers. Also, n-gram frequencies indicate frequencies in 11.371.564.

Table 2. High Frequent Bigrams Consisting of low Frequent Unigrams

GÖ	25203	1	TÜ	23620	6	ŞM	9568	8
GÜ	20124	2	UZ	15172	6	VE	49863	9
ÇO	14880	2	UŞ	14641	6	BU	44624	9
ÖZ	12477	2	YÜ	14156	6	GE	40841	9
OĞ	10648	2	BÜ	11348	6	CE	37156	9
OC	7324	2	ÜY	11253	6	Ğİ	36283	9
ĞU	18753	3	ÜS	10146	6	Gİ	35787	9
UĞ	16907	3	TO	9312	6	ST	31918	9
ÖY	14744	3	BO	9184	6	İÇ	31429	9
SÖ	10196	3	OY	8640	6	Çİ	27377	9
CU	9701	3	ÜT	7852	6	EV	25568	9
ÜZ	17636	4	SÜ	7729	6	TU	22533	9
ÜŞ	13030	4	OT	7353	6	SU	22168	9
ZÜ	7025	4	PL	7068	6	İĞ	21975	9
ŞÜ	6549	4	ĞL	7029	6	HE	21182	9
ĞI	37718	5	ÖL	6801	6	UY	20367	9
IĞ	24106	5	LG	6372	6	EC	19480	9
ÇI	18112	5	ŞU	6102	6	EĞ	19472	9
CI	10287	5	NC	21614	7	TT	19323	9
IP	10041	5	ÖR	17778	7	Hİ	18821	9
PI	8685	5	ÖN	15278	7	ÇE	17736	9
DÖ	6918	5	ĞR	7347	7	UT	13198	9
YO	61044	6	RG	6210	7	Cİ	11387	9
SO	27989	6	MÜ	11630	8	YU	11219	9
ŞT	26972	6	ÜM	11563	8	İP	10803	9

The bigram of "*gö*" consists of rare unigrams and it has a high frequency (The frequency of the most common bigram, "ar", is about %0.02). If it is seen a bigram more than one times in ciphertext and its frequency would be about %0.002(25203/11371564), it could be assumed to be "*gö*". The rest of the bigrams could contribute to solve ciphertext but their frequencies are too close. It seems better to turn back after trying to detect more symbols via other n-grams.

Table 3 . High Frequent Trigrams Consisting of low Frequent Unigrams

GÖZ	5755	2	GÖN	1184	7	ÇOK	9685	10
ÇOC	3833	2	ÜŞÜ	5781	8	DOĞ	4452	10
GÜV	1092	2	ÜZÜ	3851	8	DÜĞ	2915	10
HOC	1017	2	ĞÜM	964	8	KÜÇ	2321	10
GÖS	2247	3	UĞU	15363	9	ÇÜK	1907	10
GÖT	1143	3	GEÇ	7184	9	KOC	1906	10
ÜĞÜ	4160	4	HİÇ	7076	9	KÖŞ	1290	10
ÖZÜ	2872	4	SÖY	6645	9	IZC	1094	10
ÜÇÜ	2810	4	CEĞ	4410	9	HIZ	1063	10
GÜZ	2803	4	HEP	3254	9	ÜTÜ	5629	12
ÜCÜ	1607	4	GEC	3087	9	YÜZ	5561	12
HOŞ	1540	4	BÖY	3016	9	CAĞ	4672	12
KÖP	1250	5	UCU	2433	9	ÜYO	4299	12
OCU	3910	6	ÖST	2287	9	ÜYÜ	4106	12
OĞU	3155	6	UYG	1996	9	ÖZL	3266	12
TOP	2979	6	YGU	1940	9	GÜL	2851	12
SÖZ	2686	6	ÇEV	1737	9	ĞUM	2715	12
ÖTÜ	2203	6	HÇE	1719	9	HAF	2525	12
FÜS	1824	6	CEV	1278	9	ÖLÜ	2404	12
ÖLG	1038	6	SUÇ	1217	9	OĞL	2295	12
BOĞ	997	6	TUĞ	1215	9	POL	2079	12
ŞÖY	982	6	HVE	1090	9	OTO	2061	12
GÖR	12199	7	ÖPE	1043	9	HAV	2005	12
ÖNC	4016	7	EVG	962	9	BOŞ	1984	12
ÖĞR	2452	7	VGİ	954	9	ÜŞT	1937	12

Table 3 contains useful n-grams to solve ciphertext. Though the values are too close to each other, the trigram of "*gör*" and "*uğu*" could be evaluated as distinctive because of the frequency values.

Table 4 contains interesting values. The tetragram of "*cumh*" would be expressed by 12 different symbols. However, detecting the tetragram would be easy. The beginning and ending letter of the tetragram would be replaced with only 1 symbol and repeated everlastingly. Similarly, same rules are valid for tetragrams of "*ptığ*" and "*vrup*". Moreover, the tetragram of "*çocu*" and "*görü*" have a distinctive frequencies.

Table 4 . High Frequent Tetragrams Consisting of low Frequent Unigrams

GÖZÜ	1760	4	FÜSU	1797	18
GÜCÜ	628	4	GÜVE	1092	18
ÇOCU	3833	6	BUGÜ	803	18
GÖTÜ	1132	6	HUZU	764	18
OCUĞ	743	6	SOĞU	678	18
ÇOĞU	692	6	BÖLG	600	18
GÖST	2246	9	DÜĞÜ	2719	20
CUĞU	699	9	KÜÇÜ	2320	20
GÖZL	2267	12	ÜÇÜK	1876	20
FOTO	732	12	ÖZÜK	595	20
OTOĞ	683	12	VRUP	609	21
SÖZÜ	633	12	YÜZÜ	2595	24
CUMH	613	12	ÜŞTÜ	1607	24
GÜÇL	583	12	GÜLÜ	1314	24
GÖRÜ	4073	14	HOCA	1017	24
ÖRGÜ	909	14	LÜĞÜ	887	24
PTIĞ	1419	15	HİÇB	2367	27
KUVV	574	15	UYGU	1915	27
ÜĞÜM	962	16	GEÇT	1131	27
ÖZÜM	608	16	HEPS	1119	27

Table 5. High Frequent Pentagrams Consisting of low Frequent Unigrams

ÇOCUĞ	748	6	GÖRDÜ	2408	70
FOTOĞ	683	12	ÖRDÜĞ	981	70
OCUĞU	666	18	ĞUMUZ	568	72
GÖZÜK	593	20	OTOBÜ	548	72
GÖRÜŞ	892	28	PTIĞI	1418	75
GÖZÜN	472	28	DÜĞÜM	803	80
ŞOFÖR	394	28	HÜKÜM	425	80
ÇOCUK	3085	30	GÖSTE	2246	81
CUMHU	612	36	UYGUS	633	81
GÖTÜR	1125	42	UVVET	543	81
ÖRGÜT	802	42	YGUSU	539	81
GÖRÜY	540	42	OTOĞR	679	84
GÖVDE	381	45	ÖRÜYO	540	84
GÖLGE	433	54	ÖTÜRÜ	484	84
ÜĞÜNÜ	914	56	GÖRÜL	414	84
ÜŞÜNC	679	56	SOĞUK	581	90
GÖRMÜ	446	56	MÜŞTÜ	807	96
ÖZÜNÜ	441	56	GÜLÜM	750	96
HÜZÜN	390	56	ÖLÜMÜ	702	96
GÖREV	1101	63	GÖRÜN	1592	98

The challengest n-gram seems to be a member of pentagrams. The pentagram of "*çocuğ*" would be expressed by 6 different symbols. More interestingly, 3 letters of the tetragram, "*ç,c,ğ*" would be repeated permanently in the ciphertext because each letter would be replaced with only 1 symbol. It would be easier to detect rest of the letters of the tetragram, "*o,u*", if the other letters are solved. Similarly, the pentagram of "*fotoğ*" is a interesting n-gram too. Whereas, first and last letter of the pentagram have about %1 frequency. Furthermore, the pentagrams of "*çocuk*" and "*gördü*" have a distinctive frequency. Another point that shouldn't be ignored is both the pentagrams of "*çocuğ*" and "*çocuk*" consisting of the distinctive tetragram of "*çocu*".

Distinctive n-grams exist as seen. It seems more meaningful to begin with looking for the bigram of "*gö*" first and attempting to solve pentagrams and tetragrams second. If it could be detected pentagrams or tetragrams in the ciphertext, it provides significant advantage in the rest of the process. Even if, these tetragrams and pentagrams don't appear in plaintext, distinctive bigrams and trigrams would most probably help to go ahead.

3 Conclusion

We have presented a novel method of exposing vulnerabilities of a historical encryption method for a specific language with taking advantage of its characteristic vulnerabilities.

Although the encryption method contains vulnerabilities for Turkish, it could clearly be said that the method is stronger than a classical substitution cipher. Moreover, it is needed to have a too long ciphertext to cryptoanalysis. If it is haven a long enough and uniform distributed ciphertext, distinctive n-grams would most probably contribute to detect vast majority of the letters of the alphabet. All in all, the method still maintains its resistance today against frequency analysis attacks if short ciphertexts have haven.

References:

[1] Dalkılıç, M. E., Dalkılıç, G. On the Cryptographic Patterns and Frequencies in Turkish Language, *Proceedings of the International Conference on Advances in Information Systems*, Vol. 2457, 2002, pp. 144-153.

[2] Singh, S., *The Code Book: The Secret History of Codes and Code-breaking*, Anchor, 2000.

[3] Stahl, F. A., A homophonic cipher for computational cryptography, *Proceedings of the National Computer Conference*, Vol. 42, 1973, pp. 565-568.