On the Cryptographic Patterns and Frequencies in Turkish Language

Conference Paper *in* Lecture Notes in Computer Science · October 2002

DOI: 10.1007/3-540-36077-8_14 · Source: DBLP

2 authors:

Mehmet E. Dalkilic
Ege University

Gökhan Dalkılıç
Dokuz Eylul University

On the Cryptographic Patterns and Frequencies in Turkish Language

Mehmet Emin Dalkılıç[1] and Gökhan Dalkılıç[2]

[1] Ege University International Computer Institute, 35100 Bornova, İzmir, Turkey
dalkilic@bornova.ege.edu.tr
http://www.ube.ege.edu.tr/~dalkilic
[2] Dokuz Eylül University, Computer Eng. Dept., 35100 Bornova, İzmir, Turkey
gokhand@bornova.ege.edu.tr
http://bornova.ege.edu.tr/~gokhand

Abstract. Although Turkish is a significant language with over 60 million native speakers, its cryptographic characteristics are relatively unknown. In this paper, some language patterns and frequencies of Turkish (such as letter frequency profile, letter contact patterns, most frequent digrams, trigrams and words, common word beginnings and endings, vowel/consonant patterns, etc.) relevant to information security, cryptography and plaintext recognition applications are presented and discussed. The data is collected from a large Turkish corpus and the usage of the data is illustrated through cryptanalysis of a monoalphabetic substitution cipher. A new vowel identification method is developed using a distinct pattern of Turkish—(almost) non-existence of double consonants at word boundaries.

1 Introduction

Securing information transmission in data communication over public channels is achieved mainly by cryptographic means. Many techniques of cryptanalysis use frequency and pattern data of the source language. The cryptographic pattern and frequency data are usually obtained by compiling statistics from a variety of source language text such as novels, magazines and newspapers. Such data is available for many languages. A good source is an Internet site "Classical Cryptography Course by Lanaki" (http://www.fortunecity.com/skyscraper/coding/379/lesson5.htm) which includes data for English, German, Chinese, Latin, Arabic, Russian, Hungarian, etc. No such data online or otherwise can be located for Turkish which is a major language used by a large number of people.

Turkish, a Ural-Altaic language, despite its being one of the major languages of the world [1], it is one of the "lesser studied languages" [2]. Even less studied are the information theoretic parameters (e.g., entropy, redundancy, index of coincidence) and cryptographic characteristics (patterns and frequencies relevant to cryptography) of the Turkish language. Earliest work (that we are aware of) on the information theoretic aspects of the Turkish is presented by Atlı [3]. Although Atlı calculated the digram entropy and gave word length and consonant/vowel probabilities, he could not go

T. Yakhno (Ed.): ADVIS 2002, LNCS 2457, pp. 144−153, 2002.

beyond the digram entropy due to insufficient computing resources. In a more recent time, Koltuksuz [4] addressed the issue of cryptanalytic parameters of Turkish where he extracted n-gram entropy, redundancy and the index of coincidence values up to n = 5. Adopting Shannon's entropy estimation approach [5], present authors empirically determined the language entropy upper-bound of Turkish as 1.34 bpc (bits per character) with a corresponding redundancy of roughly 70% [6].

Data presented in this paper is obtained from a large Turkish text corpus of size 11.5 Megabytes. The corpus --contains files that are filtered so that they consist solely of the 29 letters of the Turkish alphabet and the space, is the union of three corpora; the first one is compiled from the daily newspaper Hürriyet by Dalkılıç [7], the second one contains 24 novel samples of 22 different authors by Koltuksuz [4], and the last one consists of mostly news articles and novels collected by Diri [8].

This paper presents a variety of Turkish language patterns and frequencies (e.g., single letter frequency profile, letter contact patterns, common digrams, trigrams and frequent words, word beginnings, vowel/consonant patterns, etc.) relevant to information security, cryptography and plaintext recognition. Presented data is put to use to solve the following mono-alphabetic substitution cipher problem found in [9].

```
DLKLEĞPFÜ FLMLTU FLLĞ CBÖL ÖIBJL ĞÜNLEPĞ APEYPKÜBÜB
SIBPJP MLYLB SÇKZPA DPBNPEPFÜBP SÜĞ. CEĞLÖLYÜ ĞPZPFYCMH
PZZÜF LÖLFUBL OPİÜE. ALİV JPZYPBZÜ ĞPYBPJÜ MHZ. FCB
ÜDHNH ĞPYBPBÜB MLJELŞUBÖL.
```

Throughout the cryptanalysis example uppercase letters for ciphertext and lowercase for plaintext (typeset in Courier font) are used to improve readability.

2 Letter Frequencies

Turkish alphabet contains eight vowels {A, E, I, İ, O, Ö, U, Ü} and twenty-one consonants {B, C, Ç, D, F, G, Ğ, H, J, K, L, M, N, P, R, S, Ş, T, V, Y, Z} totaling to 29 letters. In lower case, vowels {a, e, ı, i, o, ö, u, ü} and consonants {b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z} are written as shown. {I, ı} and {İ, i} being two different letters may be confusing to many readers who are not native (Turkish) speakers.

Table 1 shows the individual Turkish letter probabilities if space is suppressed in the text. That table also introduces the frequency ordering of Turkish as AEİNRLIKDMYUTSBOÜŞZGÇHĞVCÖPFJ.

Table 1. Normal Turkish letter frequencies (%) in decreasing order

A → 11.82	I → 5.12	T → 3.27	Z → 1.51	C → 0.97
E → 9.00	K → 4.70	S → 3.03	G → 1.32	Ö → 0.86
İ → 8.34	D → 4.63	B → 2.76	Ç → 1.19	P → 0.84
N → 7.29	M → 3.71	O → 2.47	H → 1.11	F → 0.43
R → 6.98	Y → 3.42	Ü → 1.97	Ğ → 1.07	J → 0.03
L → 6.07	U → 3.29	Ş → 1.83	V → 1.00	

2.1 Letter Groupings

Unlike letter frequencies and their order which fluctuate considerably, group frequencies are fairly constant in all languages [10]. The following are some useful Turkish letter groups where each group is arranged in decreasing frequency order.

- Vowels {A, E, İ, I, U, O, Ü, Ö} 42.9%
- High freq. consonants {N, R, L, K, D} 29.7%
- Medium freq. consonants {M, Y, T, S, B} 16.2%
- Low freq. consonants {Ş, Z, G, Ç, H, Ğ, V, C, P, F, J } 11.3%
- High freq. vowels {A, E, İ, I} 34.3%
- Highest freq. letters {A, E, İ, N, R} 43.4%
- High freq. letters {A, E, İ, N, R, L, I, K, D} 63.8%

2.2 Cryptanalysis Using Letter Patterns

Monoalphabetic substitution ciphers replace each occurrence of a plaintext letter (say `a`) with a ciphertext letter (say `H`). Table 2 shows the frequency counts for the example ciphertext given in Section 1. Clearly, an exact match between these counts and the normal plaintext letter frequencies of Table 1 is not expected. Nevertheless, it is very likely that highest frequency ciphertext letters {`P,L,B,Ü`} are substitutes for letters from the highest frequency normal letters set which is {`a,e,i,n,r`}. It is expected that the low frequency ciphertext symbols {`Ç,O,T,V,Ş,G,R`} resolve to the letters from the low frequency plaintext letters set of { `ş,z,g,ç,h,ğ,v,c, p,f,j`}.

Table 2. Letter frequencies in the example ciphertext

P → 21	E → 7	C → 4	U → 3	T → 1
L → 20	Y → 7	H → 4	D → 3	V → 1
B → 16	Z → 7	N → 3	I → 2	Ş → 1
Ü → 13	J → 5	A → 3	İ → 2	G → 0
Ğ → 9	M → 5	K → 3	Ç → 1	R → 0
F → 8	Ö → 5	S → 3	O → 1	

3 Letter Contact Patterns

Letter contact data (transition probabilities) is an important characteristic of any language because contacts define letters through their relations with one another. For instance, in Turkish vowels avoid contact, doubles are rare. These behaviors can be observed from Table 3 which holds the normal contact percentage data for Turkish. Table 3 is modeled after a similar table for English by F. B. Carter reproduced in [10]. Taking any one letter, say A: On the left, it was contacted 16% of the time by L, 10% by D, etc., and 99.43% of its total contacts on that side were to consonants. On

Table 3. Normal (expected) contact percentages of Turkish digrams

V	**C**				**V**	**C**
00.57	99.43	C$_3$ H$_4$ N$_5$ S$_5$ T$_5$ B$_6$ R$_8$ M$_8$ Y$_9$ K$_9$ D$_{10}$ L$_{16}$	**A**	R$_{19}$ N$_{16}$ L$_8$ K$_8$ Y$_6$ M$_5$ D$_5$ S$_4$ Ş$_4$ T$_4$ Z$_3$ H$_3$ Ğ$_3$	00.81	99.19
73.18	26.82	L$_3$ M$_3$ N$_4$ R$_7$ E$_8$ İ$_{18}$ A$_{39}$	**B**	İ$_{38}$ A$_{22}$ U$_{16}$ E$_{13}$ Ь$_4$ O$_3$	98.53	01.47
68.05	31.95	Ь$_3$ U$_3$ I$_4$ R$_4$ İ$_4$ O$_{11}$ A$_{21}$ E$_{21}$ N$_{22}$	**C**	A$_{33}$ E$_{31}$ U$_{11}$ İ$_{10}$ I$_9$	96.81	03.19
77.36	22.64	K$_4$ U$_5$ R$_5$ N$_5$ Ь$_7$ E$_{13}$ A$_{17}$ İ$_{33}$	**З**	İ$_{22}$ I$_{15}$ A$_{15}$ E$_{15}$ O$_{14}$ Ь$_5$ L$_4$ M$_3$ T$_3$	88.22	11.78
35.58	64.42	M$_3$ Y$_3$ İ$_4$ L$_{10}$ E$_{12}$ A$_{16}$ R$_{18}$ N$_{26}$	**D**	E$_{27}$ A$_{24}$ İ$_{19}$ I$_{12}$ U$_9$ Ь$_5$ O$_3$	98.70	01.30
00.12	99.88	C$_4$ K$_4$ B$_4$ S$_5$ G$_5$ V$_5$ T$_6$ R$_6$ Y$_6$ N$_7$ M$_8$ D$_{15}$ L$_{16}$	**E**	R$_{22}$ N$_{17}$ L$_9$ K$_8$ T$_7$ M$_6$ D$_5$ Y$_5$ S$_5$ V$_3$ Ğ$_3$ C$_3$	00.16	99.84
88.82	11.18	U$_3$ Ь$_4$ I$_4$ O$_4$ İ$_{15}$ E$_{23}$ A$_{35}$	**F**	A$_{31}$ E$_{20}$ İ$_{12}$ I$_{11}$ T$_6$ L$_4$ R$_3$ O$_3$ U$_3$	80.83	19.17
16.57	83.43	A$_3$ İ$_3$ E$_3$ V$_6$ U$_6$ Z$_8$ Y$_{10}$ L$_{17}$ R$_{19}$ N$_{20}$	**G**	E$_{31}$ İ$_{22}$ Щ$_{18}$ Ь$_{15}$ A$_6$ U$_3$ I$_3$	98.37	01.63
99.86	00.14	Ь$_3$ Щ$_3$ O$_{10}$ U$_{11}$ İ$_{15}$ I$_{16}$ E$_{18}$ A$_{24}$	**Ğ**	İ$_{28}$ I$_{27}$ U$_{12}$ R$_8$ A$_7$ L$_7$ E$_4$ Ь$_3$	81.57	18.49
85.99	14.01	R$_3$ U$_5$ E$_9$ İ$_{11}$ A$_{59}$	**H**	A$_{42}$ E$_{18}$ İ$_{12}$ U$_4$ I$_3$ Ь$_3$ T$_3$ O$_3$	84.98	15.02
00.05	99.95	З$_3$ Ş$_4$ Y$_5$ Ğ$_6$ K$_6$ M$_6$ T$_8$ S$_9$ L$_{10}$ D$_{11}$ N$_{12}$ R$_{12}$	**I**	N$_{31}$ R$_{11}$ L$_{10}$ K$_9$ Ş$_8$ M$_7$ Y$_7$ Z$_5$ Ğ$_4$ S$_3$	00.11	98.89
00.78	99.22	З$_3$ G$_4$ Ğ$_4$ T$_6$ M$_6$ S$_6$ K$_7$ N$_8$ L$_9$ R$_{10}$ D$_{11}$ B$_{14}$	**İ**	N$_{22}$ R$_{17}$ L$_{11}$ Y$_8$ M$_7$ Ş$_6$ K$_5$ Z$_5$ S$_4$ З$_4$ T$_3$	00.47	99.53
80.88	19.12	Ь$_4$ İ$_6$ R$_{15}$ E$_{16}$ O$_{26}$ A$_{28}$	**J**	İ$_{36}$ A$_{19}$ E$_{17}$ L$_7$ U$_6$ I$_6$ D$_5$ O$_3$	86.19	13.81
79.49	20.51	Ş$_3$ L$_3$ Ь$_4$ U$_4$ O$_7$ R$_7$ İ$_{10}$ I$_{11}$ E$_{18}$ A$_{25}$	**K**	A$_{26}$ İ$_{14}$ L$_{10}$ E$_{10}$ I$_8$ T$_7$ O$_6$ U$_6$ Ь$_3$	74.87	25.13
57.54	42.46	T$_3$ Ş$_3$ L$_3$ U$_4$ Y$_5$ R$_5$ I$_6$ K$_6$ N$_7$ O$_8$ E$_{10}$ İ$_{12}$ A$_{14}$	**L**	A$_{29}$ E$_{24}$ İ$_{11}$ I$_8$ D$_5$ M$_5$ U$_4$ L$_3$	79.93	20.07
63.72	36.28	Ş$_3$ Ь$_3$ T$_4$ N$_4$ U$_7$ R$_8$ I$_8$ L$_9$ E$_{13}$ İ$_{14}$ A$_{17}$	**M**	A$_{29}$ E$_{22}$ İ$_{15}$ I$_{11}$ U$_7$ L$_4$ Ь$_4$ D$_3$	87.83	12.17
97.89	02.11	O$_5$ Ь$_6$ U$_7$ I$_{17}$ E$_{17}$ İ$_{20}$ A$_{23}$	**N**	D$_{17}$ E$_{13}$ İ$_{12}$ I$_{12}$ A$_{10}$ L$_9$ U$_6$ C$_4$ M$_3$	57.15	42.85
00.69	99.31	B$_4$ T$_5$ D$_8$ З$_8$ S$_{13}$ K$_{14}$ Y$_{33}$	**O**	R$_{28}$ L$_{21}$ N$_{16}$ K$_9$ Ğ$_4$ C$_4$ Y$_4$	00.25	99.75
00.09	99.91	З$_3$ Y$_4$ B$_9$ K$_{10}$ D$_{12}$ S$_{16}$ G$_{41}$	**Щ**	R$_{20}$ N$_{19}$ Y$_{17}$ Z$_{15}$ L$_7$ Ğ$_4$ T$_4$ S$_3$	00.02	99.98
90.62	09.38	Ь$_3$ R$_3$ U$_6$ E$_{10}$ O$_{10}$ I$_{11}$ İ$_{12}$ A$_{37}$	**P**	A$_{31}$ E$_{12}$ I$_{11}$ L$_{10}$ T$_6$ O$_6$ R$_5$ İ$_5$ M$_4$ S$_3$	67.27	32.73
93.20	06.80	Щ$_3$ Ь$_4$ U$_6$ I$_6$ O$_{10}$ İ$_{16}$ E$_{22}$ A$_{28}$	**R**	A$_{16}$ İ$_{14}$ I$_{12}$ D$_{11}$ E$_9$ L$_6$ U$_6$ M$_5$ K$_4$ T$_4$ S$_3$	60.19	39.81
72.68	27.32	M$_3$K$_4$ Ь$_4$ N$_5$ I$_5$ U$_5$ R$_7$ İ$_{15}$ E$_{18}$ A$_{23}$	**S**	A$_{18}$ İ$_{16}$ I$_{16}$ E$_{13}$ T$_9$ O$_7$ U$_7$ Щ$_3$ Ь$_3$	83.26	16.74
94.07	05.93	O$_3$ R$_4$ Ь$_8$ E$_8$ U$_9$ I$_{18}$ İ$_{23}$ A$_{26}$	**Ş**	T$_{16}$ I$_{14}$ A$_{13}$ L$_{11}$ İ$_{11}$ E$_9$ M$_7$ K$_6$ U$_4$ Ь$_4$	56.60	43.40
49.41	50.59	L$_3$ U$_4$ T$_5$ İ$_6$ R$_8$ Ş$_9$ S$_{10}$ K$_{10}$ A$_{15}$ E$_{18}$	**T**	A$_{19}$ E$_{16}$ İ$_{15}$ I$_{14}$ Ь$_8$ U$_6$ L$_5$ T$_4$ M$_4$ O$_3$	82.03	17.97
00.18	99.82	C$_3$ Y$_4$ Ğ$_4$ T$_6$ S$_6$ M$_7$ K$_7$ L$_8$ R$_9$ N$_{10}$ D$_{13}$ B$_{14}$	**U**	N$_{20}$ R$_{15}$ L$_{10}$ M$_9$ Y$_7$ K$_6$ Ş$_5$ Z$_5$ Ğ$_5$ T$_4$ S$_4$	00.62	99.38
00.02	99.98	З$_3$ Ş$_3$ Z$_3$ S$_5$ B$_5$ K$_6$ M$_6$ N$_6$ L$_6$ R$_7$ Y$_8$ G$_{11}$ D$_{13}$ T$_{13}$	**Ь**	N$_{23}$ R$_{15}$ Z$_9$ K$_8$ L$_7$ Ş$_7$ Y$_6$ M$_6$ S$_5$	00.12	99.88
88.31	11.69	Ь$_3$ İ$_4$ U$_5$ A$_{28}$ E$_{41}$	**V**	E$_{43}$ A$_{24}$ İ$_8$ R$_5$ U$_5$ L$_5$	82.06	17.94
94.24	05.76	O$_4$ Ь$_5$ Щ$_6$ U$_8$ I$_{10}$ E$_{15}$ İ$_{22}$ A$_{25}$	**Y**	A$_{29}$ O$_{17}$ E$_{15}$ L$_8$ I$_7$ İ$_5$ Ь$_4$ U$_3$ D$_3$	81.82	18.18
96.32	03.68	E$_7$ Щ$_9$ U$_{10}$ Ь$_{11}$ I$_{15}$ A$_{21}$ İ$_{21}$	**Z**	A$_{19}$ E$_{18}$ L$_{14}$ İ$_{11}$ I$_{11}$ D$_7$ Ь$_6$ U$_4$ M$_3$	69.70	30.30

the right, it was contacted 19% of the time by R, and 0.81% of the time by vowels. Note that the table only contains contacts with a frequency of 3% or more. The most marked characteristics of Turkish letter contacts are *vowels do not contact vowels on either side* and *doubles are rare*. Since no consonant avoids vowel contact, vowels and consonants are very much distinguishable. Only significant doubles are TT and LL for consonants and AA for vowels. Doubles for vowels are so rare that even AA did not make into the Table 3.

3.1 Most Frequent Digrams and Trigrams

In Turkish, among 29^2 digrams about one third and among the 29^3 trigrams about one tenth constitute the 96% of the total usage. In Table 4, none of the first 100 digrams contains doubles. Fifty of them are in the form consonant-vowel (**CV**), 42 are in the form vowel-consonant (**VC**) and only 8 are in the form (**CC**). Usage of the first ten digrams in Table 4 sums up to 16.9%, first 50 to 47.9% and first 100 to 69.5%.

Table 4. The 100 Most Frequent Digrams in Turkish (Frequencies in 100,000)

AR 2273	Dİ 1021	NI 703	OL 586	AŞ 500	BE 433	KI 350
LA 2013	ND 980	AY 698	Sİ 578	NL 496	KE 424	RU 349
AN 1891	RA 976	YO 686	LI 576	TI 494	EY 421	Ğİ 347
ER 1822	AL 974	EK 683	RE 566	EM 494	ES 411	AZ 343
İN 1674	AK 967	RD 681	SI 565	ÜN 492	IK 407	İS 343
LE 1640	İL 870	TA 670	Mİ 564	DU 487	RL 393	Gİ 342
DE 1475	Rİ 860	AM 638	TE 562	GE 480	MI 392	ĞI 340
EN 1408	ME 785	DI 637	ET 560	AT 479	İK 379	AH 338
IN 1377	Lİ 782	SA 624	İM 541	SE 457	CA 379	YL 324
DA 1311	OR 782	İY 619	Tİ 537	ED 452	LD 362	ÜR 319
İR 1282	NE 738	Kİ 618	HA 528	UR 452	CE 361	
Bİ 1253	RI 733	UN 606	AS 527	ON 452	NU 359	
KA 1155	BA 718	NA 602	BU 516	KL 447	IŞ 355	
YA 1135	Nİ 716	AD 592	VE 508	IL 438	İZ 353	
MA 1044	EL 710	YE 588	IR 503	İŞ 434	LM 353	

Observe that Turkish letter contact data (Table 3) is not symmetric. For instance BU is in the table but its reverse UB is not. High frequency digrams (Table 4) with rare reverses (i.e., reverses are not in Table 3) e.g. ND, OR, RD, DI, OL, BU, NL, TI, ÜN, DU, GE, ON, RL, LD, YL and high frequency digrams with reverses which are also high frequency digrams (i.e., both the digram and its reverse are in Table 4) e.g., AR, LA, AN, ER, İN, LE, DE, EN, IN, DA, İR, KA, YA, MA, RA, AL, AK, İL, Rİ, ME, Lİ, NE, Rİ, EL, AY, EK, TA, AM, SA, UN, NA, AD, YE, Sİ, LI, RE, Mİ, TE, İM, HA, AS, IR, SE, ED, UR, IL are useful for distinguishing letters from each other.

Table 5 shows that none of the most common Turkish trigrams contain two vowels or three consonants in sequence; that is, no **VVV**, **VVC**, **CVV**, or **CCC** patterns. Most common trigram patterns in Table 5 are **VCV** (46 times), followed by **CVC** (35 times). **CCV** and **VCC** are seen 11 and 8 times, respectively. The first 10, 50 and 100 trigrams represent, respectively, 7.2%, 19.0%, and 28.7% of total usage.

Table 5. The 100 Most Frequent Trigrams in Turkish (Frequencies in 100,000)

LAR 1237	ANI 362	ESİ 283	İLİ 245	RLA 216	AĞI 194	RDI 169
BİR 952	AMA 357	NIN 280	BAŞ 243	MİŞ 213	ORD 194	SON 168
LER 949	RIN 345	YLE 277	ARD 242	YAN 212	GEL 194	ILA 167
ERİ 764	NLA 338	ADI 273	NİN 239	ECE 209	MAN 192	BEN 166
ARI 757	DAN 338	İYO 271	RDU 231	AYI 207	ACA 192	CAK 165
YOR 643	IND 336	ELE 271	MIŞ 229	LMA 207	ÖYL 191	İRİ 163
ARA 521	EDİ 326	İNE 266	OLA 227	İĞİ 207	KAD 187	EYE 163
NDA 482	ADA 321	SİN 265	IĞI 226	EDE 206	ERD 183	AŞI 162
İNİ 432	AYA 316	ANL 263	EĞİ 223	TAN 205	ORU 178	ÇIK 160
INI 428	KAR 299	KLA 262	EME 223	NDİ 204	RAK 177	KAN 159
ASI 387	ALA 298	ERE 262	INA 222	KAL 204	DİY 172	
DEN 383	LAN 296	ALI 258	ANA 220	ONU 201	KLE 171	
NDE 383	ENİ 294	ELİ 256	KEN 218	UNU 200	VER 170	
RİN 372	SIN 294	İYE 255	İÇİ 217	END 199	EMİ 169	
İLE 367	İND 291	BİL 246	IYO 217	ÇİN 198	GÖR 169	

3.2 Cryptanalysis Using Letter Contact Patterns

Vowels usually distinguish themselves from consonants by not contacting each other. Table 6 shows the letter contact frequencies for the six most frequent ciphertext letters {P,L,B,Ü,Ğ,F}. Highest frequency letters {P,L} do not contact each other and almost certainly are vowels. Letters {B,Ğ} both contact {P,L}, and thus they are consonants. Letter Ü does not contact either of the {P,L} and it is likely to be a vowel. Letter F must be a consonant because it contacts two (presumable) vowels {L,Ü}. When the table is extended to include all ciphertext letters, it is determined that {P,L,Ü,C,H,I,Ç} are very likely to be vowels.

Table 6. Letter frequencies in the example ciphertext

	P	L	B	Ü	Ğ	F
P	--	--	3	--	1	3
L	--	1	1	--	1	1
B	4	1	--	2	--	--
Ü	--	--	4	--	1	1
Ğ	4	1	--	1	--	--
F	--	2	--	2	--	--

Now, let us focus on the two doubles LL in FLLĞ and ZZ in PZZÜF. These doubles may be associated to the only significant doubles of Turkish {tt,ll,aa} i.e., {L,Z} → {t,l,a}. Due to frequency and vowel analysis, L is a strong candidate for a, we may temporarily[1] bind L→a. This makes letter P the strongest candidate for letter e i.e., P→e. Letter Ü is a vowel and it is either one of {ı,i }. BÜB is a repeated trigram with the '121' pattern and in Table 5, the only '121' patterned trigrams

[1] Cryptanalysis is mostly a trial and error process and all bindings are temporary.

with starting and ending with a consonant (i.e., **CVC**) are NIN and NİN. Thus, we may bind letter B→n and stop here to continue later.

4 Word Patterns

Primary vowel harmony rule is that all vowels of a Turkish word are either back vowels {A,I,U,O} or front vowels {E,İ,Ü,Ö}. *Secondary vowel harmony rule* states that (i) when the first vowel is flat {A,E,I,İ}, the following vowels are also flat e.g., BAKIRCI, İSTEK, (ii) when the first vowel is rounded {U,O,Ü,Ö}, the subsequent vowels are either high and rounded {U,Ü} or low and flat {A,E}, and (iii) low and rounded vowels {O,Ö} can only be in the first syllable of a word. *Last Phoneme Rule* is that Turkish words do not end in the consonants {B,C,D,G}. Each of these rules has exceptions. Using a root word lexicon, Gungor [11] determined that only 58.8% of the words obey the primary vowel harmony rule. The secondary vowel harmony rule is obeyed by 72.2%. The most obeyed rule is the last phoneme rule with 99.3%.

The average word length of Turkish is 6.1 letters about 30% more than that of English. Words with 3 to 8 letters represent over 60% of total usage in Turkish text.

4.1 Common Words, Beginnings, and Endings

When word boundaries are not suppressed in ciphertext, frequent word beginnings, endings and common words provide a wealth of information. The most frequent 100 Turkish words and the most common 50 n-grams for each of the other categories together with their percentage in total usage are given below.

Common words BİR VE BU DE DA NE O GİBİ İÇİN ÇOK SONRA DAHA Kİ KADAR BEN HER DİYE DEDİ AMA HİÇ YA İLE EN VAR TÜRKİYE Mİ İKİ DEĞİL GÜN BÜYÜK BÖYLE NİN MI IN ZAMAN İN İÇİNDE OLAN BİLE OLARAK ŞİMDİ KENDİ BÜTÜN YOK NASIL ŞEY SEN BAŞKA ONUN BANA ÖNCE NIN İYİ ONU DOĞRU BENİM ÖYLE BENİ HEM HEMEN YENİ FAKAT BİZİM KÜÇÜK ARTIK İLK OLDUĞUNU ŞU KADIN KARŞI TÜRK OLDUĞU İŞTE ÇOCUK SON BİZ VARDI OLDU AYNI ADAM ANCAK OLUR ONA BİRAZ TEK BEY ESKİ YIL BUNU TAM İNSAN GÖRE UZUN İSE GÜZEL YİNE KIZ BİRİ ÇÜNKÜ GECE (23%)

Note that any n-gram enclosed within a pair of punctuation mark(s) and space(s) is counted as a word. For instance "BAKAN'IN" (minister's) is taken as two words "BAKAN" and "IN". The only one-letter word in Turkish is O(it/he/she). The list contains many non-content words such as BİR (one), VE (and), BU (this), DE DA (too/also), NE (what), Kİ (that/who/which), AMA (but), İLE (with) and fewer conceptual words such as TÜRKİYE (Turkey), ZAMAN (time), İNSAN (human), GÜZEL (beautiful).

Digram word endings. AN EN İN AR IN DA ER DE Dİ AK LE Nİ NA NE NI İM DI Rİ RI OR EK YE RA DU UN YA Kİ İR LA IM Lİ Sİ IK IR LI ET TI Tİ CE SI UM IŞ RE ĞI İZ İK İŞ MA IZ Bİ (73.1%)

Trigram word endings. LAR DAN LER DEN YOR ARI INI NDA İNİ ERİ İNE INA NDE NIN NİN RDU YLE MIŞ AYA ASI MİŞ RAK IĞI RIN CAK ESİ RDI ARA İYE NRA MAK MEK TAN İĞİ DAR RİN EYE MAN LIK RUM UNU ADA RDİ ADI KEN DIR TEN DİR LİK YLA (41.6%)

Digram word beginning . Bİ KA YA DE BA BU GE VE OL DA HA SA BE GÖ SO KO TA Gİ SE NE HE AL GÜ YE AN Dİ İÇ KE Kİ AR TE ÇO DÜ KU İN VA İS ME KI DO PA ON İL ÇA DU YO MA TÜ ÇI Mİ (67.3%)

Trigram word beginnings. BİR BAŞ İÇİ KAR GEL GÖR SON BEN OLA KAD YAP BİL KAL VER KEN ÇIK DEĞ VAR GÜN YAN GİB İST BAK DİY TÜR HER ARA OLM DED ÇOK DÜŞ DAH BUN GER OLD YER KON GEÇ PAR DUR KUR BİZ ANL ÇOC YAR YIL BUL SEN OLU YOK (30.6%)

A careful observation of Turkish word endings and beginnings given above reveal a distinct feature of Turkish; *the first two and the last two letters of a word contain a vowel*. In other words, (almost) no Turkish word starts or ends with a consonant-consonant (**CC**) or vowel-vowel (**VV**) pattern. Very few words (about 2%), mostly foreign origin e.g. TREN (train), KREDİ (credit), RİNG (ring) do not obey this rule. A vowel identification method is developed for Turkish using this *"no CC or VV patterns at word boundaries rule"* and presented in the next subsection.

4.2 A New Vowel Identification Method for Turkish

When spacing is not suppressed in a ciphertext for a mono-alphabetic cipher the following technique can be employed to distinguish vowels from consonants.

First, make a list of digram word beginnings and endings. Let us call it the *PairList*. Then pick a pair containing a high frequency (in the ciphertext) letter. Remove the pair from the *PairList*. Then, create two empty lists *List1* and *List2* and put one letter of the pair to *List1* and the other to the *List2*. Next, repeat the following steps until all elements in *List1* and *List2* are marked (processed), (i) pick and mark first unmarked element (say X) in *List1* or *List2*. In the *PairList* find each pair in the form XY or YX and put Y to the list that X does not belong to and remove that pair from the *PairList*.

At the end of the process, remove duplicates from both lists and smaller list will be (very likely) vowels and the other list will contain consonants. For those few words that do not obey the "*no CC or VV patterns at word boundaries*" rule may cause a letter to end up in both lists. If that is the case, get two counts: separately count the number of times the letter contacts to the members of *List1* and *List2*. Since a contact between the members of a list indicates either **CC** or **VV** pattern, if one count is dominant remove the letter from that list. For example if letter X is seen many times with the elements of *List1* and few times with the elements of *List2*, remove it from *List1*, and keep it in *List2*. If no count is dominant, remove it from both lists.

At the end, the *PairList* may contain pairs whose letters not placed in either list because they do not make contact (at word beginnings or endings) to any other letter in lists (*List1* and *List2*). In such situations, again count number of contacts to each list's elements for both letters of the left behind pairs, this time using all contacts, not only the digrams at word beginning and endings. Then, using these counts determine whether they fit in the vowel list or the consonant list.

Let us illustrate this method for our ongoing example: *PairList* = {DL,FL,CB, ÖI,ĞÜ,AP,SI,ML,SÇ,DP,SÜ,CE,ĞP,PZ,LÖ,OP,AL,JP,ĞP,MH,FC,ÜD, FÜ,TU,LĞ,ÖL,JL,PĞ,ÜB,JP,LB,PA,BP,ÜĞ,YÜ,ÜF,BL,ÜE,İV,ZÜ,JÜ, HZ,CB,NH}. First, pick the DL pair and create *List1* ={D}, *List2* ={L}. Next, pick D from *List1* and process pairs containing D i.e., DP and ÜD resulting in *List1*={D*}, *List2* ={L,P,Ü} where D* means D has been processed. Then, pick L from *List2* and process pairs containing L e.g., FL,ML,AL,… producing *List1*={D*,F,M,A,Ö,Ğ, J,B}, *List2*={L,P,Ü}, and continue until all letters in both lists are processed. Final lists are formed as *List1*={D*,F,M,A,Ö,Ğ,J,O,Z,B,E,Y,S}, *List2*={L,P,Ü, C,I,H,Ç}. Since *List2* is shorter, it contains vowels.

Pairs {İV} and {TU} are left in the *Pairlist*. In the ciphertext, İ occurs twice and contacts P, Ü, L and they are all vowels. Thus, İ must be a consonant and V must be a vowel. Similar analysis adds T and U to the consonant and vowel lists respectively.

4.3 Cryptanalysis Using Word Patterns

We temporarily marked {L,P,Ü,C,I,H,Ç,U,V} as vowels and bound L→a, P→e, B→n, and Ü→{ı,i}. The *primary vowel harmony* rule gives us a way to distinguish between ı and i: If Ü→ı association is correct, Ü will coexist in many ciphertext words with L→a, otherwise Ü→i is true and Ü will be seen together with P→e in many ciphertext words. Since {P→e,Ü→i} seen together in eight words while {L→a,Ü→ı} in only three words, the likely option is Ü→i.

Let us concentrate on the next two highest frequency letters Ğ and F which appear in a rare pattern ĞLLF → ?aa?. There are not too many four letter words with the unusual pattern ?aa?. There are only 7 matching words; faal, maaş, naaş, saat, vaat, vaaz, zaaf. Except the word saat each candidate contains a letter from the low frequency consonants group {f,ş,v,z}. Thus, it is likely that F→s, and Ğ→t. At this point there are many openings to explore.

-a-a-tesi sa-a-- saat –n-a ––n-a ti-a-et -e—-e-inin --
ne-e –a-an ----e- -en-e-esine **–it.** ––ta-a-i **te-es----**
e—-is a-as-na –e-i-. –a-- -e-—-n-i **te-ne**-i ---. **s-n**
i---- **te-nenin** –a—-a—-n-a.

The partial words **-a-a-tesi, –it**, **s-n**, **te-nenin** can easily be identified as Pazartesi(Monday), git (go), son (last), and teknenin (yatch's). Furthermore, since we have already identified t, and a, only remaining candidate for l is Z. Putting all this together, we have the partial decryption:

pazartesi sa-a— saat on-a ––n-a ti-aret –erkezinin g-
ne-e –akan g-zle- pen-eresine git. Orta-aki telesko—-

```
ellis a-as-na –evir. –a-- -elkenli tekne-i –-l. son
ip--- teknenin –a-ra—-n-a.
```

Completion of the decryption is left to the curios reader. Full decryption reveals that the ciphertext contains a single substitution error. Can you find it?

5 Conclusion

We have presented some Turkish language patterns and frequency data compiled from a large text corpus. The data presented here is relevant not only to the classical cryptology but also to the modern cryptology due to its potential use in automated plaintext recognition and language identification.

We have also demonstrated two things; first, the data's usage on a complete cryptanalysis example and the second, new insight can be attained through careful and systematic study of language patterns. We have discovered a distinct pattern of Turkish language and used it to develop a new approach for vowel identification.

What we could not address due to the limited space are (i) the fluctuations of the data for short text lengths, and (ii) the application of the data to other cipher types, especially substitution ciphers without word boundaries and the transposition ciphers.

Our future work plan includes the investigation of n-gram *versatility* (the number of different words in which the n-gram appears), and *positional frequency* of n-grams.

References

1. Schultz, T. and Waibel, A.: Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. Proceedings of EuroSpeech'97 (1997)
2. Oflazer, K.: Developing a Morphological Analyzer for Turkish. NATO ASI on Language Engineering for Lesser-studied Languages, Ankara, Turkey (2000)
3. Atlı, E.: Yazılı Türkçede Bazı Enformatik Bulgular. Uyg. Bilimlerde Sayısal Elektronik Hesap Makinalarının Kullanılması Ulusal Sempozyum, Ankara, Turkey (1972) 409-425
4. Koltuksuz, A.: Simetrik Kriptosistemler için Türkiye Türkçesinin Kriptanalitik Ölçütleri. PhD. Dissertation, Computer Eng. Dept., Ege University, İzmir, Turkey (1995)
5. Shannon, C.E.: Prediction and Entropy of Printed English. Bell System Technical Journal. Vol. 30 no 1 (1951) 50-64
6. Dalkilic, M.E. and Dalkilic, G.: On the Entropy, Redundancy and Compression of Contemporary Printed Turkish. Proc. of. Intl. Symp. on Comp. and Info. Sciences (2000) 60-67
7. Dalkilic, G.: Günümüz Yazılı Türkçesinin İstatistiksel Özellikleri ve Bir Metin Sıkıştırma Uygulaması. Master Thesis, Int'l. Computer Inst., Ege Univ., İzmir, Turkey (2001)
8. Diri, B.: A Text Compression System Based on the Morphology of Turkish Language. Proc. of International Symposium on Computer and Information Sciences (2000) 12-23.
9. Shasha, D.:Dr. Ecco'nun Şaşırtıcı Serüvenleri (Turkish translation: The Puzzling Adventures of Dr. Ecco.) TUBITAK Popüler Bilim Kitapları 24 (1996)
10. Gaines, H. F.: Cryptanalysis. Dover, New York (1956)
11. Güngör, T.: Computer Processing of Turkish: Morphological and Lexical Investigation PhD. Dissertation, Computer Eng Dept., Boğaziçi Univ., İstanbul, Turkey (1995)