MODELLING OF TURKEY TURKISH WORDS BY DISCRETE MARKOV PROCESSES

Jale GÜVENTÜRK

October, 1998

Izmir Institute of Technology

The Graduate School

MODELLING OF TURKEY TURKISH WORDS
BY
DISCRETE MARKOV PROCESSES

A Thesis in

Analysis of Turkey Turkish Words

By

Jale GÜVENTÜRK

Submitted in Partial Fulfillment

of the Requirements

for the degree of

Master of Science in Computer Engineering

October, 1998

We approve the thesis of **Jale GÜVENTÜRK**

Date of Signature

....

Assistant Prof. Dr. Ahmet Hasan KOLTUKSUZ
Supervisor
Department of Computer Engineering

18.11.98

....

Prof. Dr. Sıtkı AYTAÇ
Department of Computer Engineering

18/11/1998

....

Prof. Dr. Şaban EREN
EGE UNIVERSITY
Department of Computer Engineering

18/11/'98

....

Prof. Dr. Sıtkı AYTAÇ
Head of Department

18/11/1998

PREFACE

Since the era we are going through may be named as the information era, the production of information, by all means, is the issue. What is more important is, the protection of the information produced.

Protection is needed in two cases; the first one is the desire to keep the original information from being altered by unauthorized or ill-willed people who may have intercepted it. The second is the desire to keep the information totally secret.

The best way to provide secrecy is achieved by means of *ENCRYPTION*, which is simply substituting symbols or other letters in for the original symbols or letters that make up the information. To obtain the original information, the algorithm used for encryption is needed in order to revert the document to its real form. This process is named as *DECRYPTION*.

The science of secrecy for information hiding is known as *CRYPTOLOGY*; which originates from two Greek words **cryptos**-secret, **logos**-science.

The art of trying to find out ways and algorithms to hide information is studied under a special branch of cryptology named as *CRYPTOGRAPHY*.

As well as hiding the information, there may come a time where the hidden information has to be intercepted and decrypted.[1] To do this, one has to know the encryption algorithm. If it is a time of war and the one side has intercepted an encrypted message of the other side, it may well be assumed that the encryption algorithm is not known. Then encryption is to be done by alternative methods. The act of decrypting encrypted messages, without the presence of the encryption algorithm, is studied under the *CRYPTANALYSIS* branch of cryptology.

[1] To discuss the moral and ethics involved with the act of cryptanalysis is well beyond the scope of this thesis.

In order to perform the act of cryptanalysis, one has to know primarily the language in which the original message was produced and grammar rules, letter combinations,

mostly used letters and words, etc. All of these primary concerns are called as the **cryptanalytical measures of a language**.

The objective of this thesis is to obtain the cryptanalytical measures of the words used in Turkey Turkish by Markov Processes Approach. The study is based on the "Cryptanalytical Measures of Turkey Turkish for Symmetric Cryptosystems" unpublished Ph.D. Thesis by Asistant Prof. Dr. Ahmet Hasan KOLTUKSUZ.

Although works on other languages such as English, German, French and many others have been totally completed, the only cryptanalytical study completed on Turkey Turkish language is the above mentioned Ph.D. thesis. It is hoped that this thesis will provide the necessary background and inspiration for the new comers, to the informartion theory world, to extend the studies and help to bring together a thorough cryptanalytical database for Turkey Turkish.

ACKNOWLEDGEMENTS

Even if only one name appears in the "prepared by statement ", it is obvious that without the efforts of a group of people, in any way, nothing can be accomplished by oneself alone.

Among many people I would like to thank for, the first and the most is my advisor Assistant Prof. Dr. Ahmet Hasan KOLTUKSUZ for supporting me every inch of the way, letting me build my work upon the computer codes he had produced, lending his books, and everything he has done for me. This thesis would not have existed without him.

One of the most thank you deserving person is by no doubt Res. Ass. Mete Eminağaoğlu who has given all his time to produce the code needed. Mete I know you made up the time in order to help me, when there was no time in your hands. I appreciate it tremendously.

For the understanding and time allowance they have given I would like to thank Director of İ.M.Y.O. Prof. Dr. Öcal USTA, and İ.M.Y.O. Technical Programs Department Head Prof. Dr. Nuran GÖKÇEN.

For letting me; to take the time out whenever I needed, to use the office equipment and the encouragement they have given, I would like to thank my bosses Muzaffer ÖZER, Malcolm BROOK, and all my colleagues in the office.

For being there whenever I needed, my parents İnci-Refik ŞAŞMAZ and my parents-in-law Nazan-İbrahim GÜVENTÜRK. Your making things easier for me has been well beyond my expectations. I do appreciate it very much.

My wonderful twin sister Lale ŞAŞMAZ, thank you for your efforts to find time for me during your neurosurgery residency, which you always did, and having me over your place whenever I wanted. Thank you.

The last but not the least, I would like to thank my husband Sefa GÜVENTÜRK and my lovely and adorable daughter Yıldız GÜVENTÜRK for encouraging me to start this masters degree study and supporting me all the way.

CONTENTS

	Page number
PREFACE	: ii
ACKNOWLEDGEMENTS	: iv
CONTENTS	: v
ABSRACT	: vii
CHAPTER 1 : MATERIAL AND METHOD	
1 . 1 Hardware and Software used	: 1
1 . 2 Filtration	: 1
1 . 3 Method	: 1
CHAPTER 2 : STATISTICAL BACKGROUND	: 2
2 . 1 Introduction	: 2
2 . 2 Probability Theory	: 2
2 . 2 . 1 Definition of Probability	: 2
2 . 2 . 2 Joint Probability	: 2
2 . 2 . 2 . 1 Mutual Exclusiveness	: 3
2 . 2 . 2 . 2 Independence	: 3
2 . 2 . 3 Conditional Probability	: 4
2 . 3 Random Variables	: 4
2 . 3 . 1 Definition	: 4
2 . 3 . 2 Two Random Variables	: 5
2 . 3 . 2 . 1 Joint Probability	: 5
2 . 3 . 2 . 2 Conditional Probability	: 5
2 . 4 Entropy	: 6
2 . 4 . 1 Joint Entropy	: 7
2 . 4 . 2 Conditional Entropy	: 7
2 . 4 . 2 . 1 Chain Rule	: 8
2 . 5 Stochastic Processes	: 8
2 . 5 . 1 Definition	: 8
2 . 5 . 2 Markov Processes	: 9
2 . 5 . 2 . 1 Definition	: 9
2 . 5 . 2 . 2 Mathematical Explanation	: 9
2 . 5 . 2 . 3 Order of a Markov Process	: 9
2 . 6 Examples	: 11
2 . 6 . 1 Letter Frequencies	: 11
2 . 6 . 2 Entropy	: 12

CHAPTER 3 : CRYPTANALYTICAL BACKGROUND : 13
3 . 1 Introduction : 13
3 . 2 The Need for Cryptology : 14
3 . 3 Definitions : 14
3 . 3 . 1 Cryptography : 14
3 . 3 . 1 . 1 Algorithms : 15
3 . 3 . 1 . 2 Classification of Cryptosystems : 15
3 . 3 . 1 . 3 Symmetric Cryptosystems : 16
3 . 3 . 1 . 4 Asymmetrical Cryptosystems (Public-Key Cryptosystems) : 17
3 . 3 . 2 Cryptanalysis : 18
3 . 3 . 2 .1 Cipher-text Only Attack : 19
3 . 3 . 2 . 2 Known- plaintext Attack : 19
3 . 3 . 2 . 3 Chosen-plaintext Attack : 20
3 . 3 . 2 . 4 Adaptive Chosen-plaintext Attack : 21
3 . 3 . 2 . 5 Chosen Ciphertext Attack : 21
3 . 3 . 2 . 6 Chosen Key Attack : 21
3 . 3 . 2 . 7 Rubber-hose Cryptanalysis : 22
3 . 4 Security of Algorithms : 22
3 . 4 . 1 Unconditional and Computational Security : 23
3 . 5 Complexity of an Attack : 24

CHAPTER 4 : ANALYSIS OF TURKEY TURKISH WORDS BY MARKOV PROCESSES : 25
4 . 1 Markov Processes : 25
4 . 2 Consonant - Vowel (c-v) Patterns : 26
4 . 3 Analysis : 27
4 . 4 Presentation of the Analysis : 31
4 . 5 Examples from the Analysis : 32
CHAPTER 5 : COMMENTS and FURTHER STUDIES : 33
5 . 1 Comments : 33
5 . 2 Further Studies : 33

SUMMARY : 35
ÖZET : 36
BIBLIOGRAPHY : 37
APPENDIX 1 (SELECTED EXAMPLES FROM THE ANALYSIS) : 38

ABSTRACT

This study based on the previous works of Assistant Prof. Dr. Ahmet Hasan KOLTUKSUZ, is about obtaining cryptanalytical measures of Turkey Turkish words based on Markov processes approach.

ABSTRAKT

Türkiye Türkçesi kelimeleri kriptanalitik ölçütlerinin Markov yaklaşımlarıyla oluşturulmasına dayanan bu çalışmanın temelini Yrd. Doç. Dr. Ahmet Hasan KOLTUKSUZUN önceki çalışmaları temel alınarak hazırlanmıştır.

CHAPTER 1 : MATERIAL AND METHOD

Since this thesis is based on the previous works of Asst. Prof. Dr. Ahmet Hasan KOLTUKSUZ, the material and method used is exactly the same. However for the words of Turkey Turkish the Redhouse dictionary was used as explained below. [1]

1 . 1 Hardware and Software used

Cryptanalytical studies are performed on huge scale texts, and the words or more correctly different and so many combinations of the characters making up the text is numerous. Viewing the process from this side, it is an I/O bound process.

For the text formation PC based Compact Pentium MMX computer has been used and the software is Windows '95 Office based programs.

1 . 2 Filtration

The text worked on has been obtained from the Redhouse Turkish Dictionary. All the words listed in the dictionary was first transferred to the electronic environment. The ones starting with capital letters were changed to small ones. Words that are spelled exactly the same but carry different meanings were located and only one was kept, the remaining were deleted from the list. Then the idioms and phrases with more than one words were modified to appear as one single word, spaces in between being deleted. The words which are no longer in today's Turkey Turkish were removed.

1 . 3 Method

After the Redhouse dictionary has been filtered, all the words were transferred to appear in their consonant-vowel pattern. The study then was performed on these c-v patterns obtained. The conditional probabilities from 0 to 4^{th} order were obtained by the use of Markov Process Approach.

---

[1] Ahmet Hasan KOLTUKSUZ, *Simetrik Kriptosistemler için Türkiye Türkçesinin Kriptanalitik Ölçütleri,* unpublished Ph.D. thesis, Ege Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı,İzmir 1995, pp. 4 -13.

CHAPTER 2 : STATISTICAL BACKGROUND

2 . 1 Introduction

This chapter involves definitions of concepts repeatedly used throughout the following chapters and some examples to make these concepts more understandable.

2 . 2 Probability Theory

2 . 2 . 1 Definition of Probability

"*If the experiment is performed n times and the event A occurs n(A) times, then, with a high degree of certainty, the relative frequency n(A)/n of the occurrence of A is close to P(A):*

$$P(A) \cong n(A)/n$$

provided that n is sufficiently large." [1]

Using the term "sufficiently large" is a must because, in order to say the probability of observing a specific event is some number, one has to repeat the experiment endless times. To clear this expression from the definition it is necessary to develop a more explanatory and complete mathematical form. This can be summarized as :

$$P(A) = \lim_{n \to 8} \frac{n(A)}{n}$$ [2]

2 . 2 . 2 Joint Probability

Considering two events such as A and B, one can obtain the individual or sometimes called marginal probabilities using the definition given above as P(A) and P(B) .

---

[1] Athanasios Papoulis, *Probability of Random Variables, and Stochastic Processes*, p.3.
[2] ibid., p.6

If the concern is the probability of observing A and B occurring simultaneously, then a new term is needed to define this probability. To make it more clear an example could be, event A as observing 1 from a die throw and event B as observing a tale from the toss of a coin. Another example is drawing a marble twice from a bag which contains 3 red and 2 blue.

The probability of observing two or more events occurring simultaneously is named as joint probability and denoted as P(AB).

2 . 2 . 2 . 1 Mutual Exclusiveness

If any two events cannot occur simultaneously then they are said to be mutually exclusive, in notation $P(AB) = 0$.

From set theory it is known that

$$P(AB) = P(A) + P(B) - P(A+B)$$

P(A+B) is the probability of observing event A or event B. In other words observing at least one of the two events occurring (Union of events)

It can be concluded from the above that if two events are mutually exclusive then the probability of either one of them occurring is equal to the sum of their marginal probabilities. In notation ;

$$P(A+B) = P(A) + P(B)$$

2 . 2 . 2 . 2 Independence

Considering any two events, if occurrence of one of them has no effect on the occurrence of the other then the two events are called independent. In the coin toss example, the outcome of the first throw say a head will not have any influence on the second throw, chances being the same as a head or a tail.

If two events are independent then their joint probability is equal to the product of their marginal probabilities. In notation :

$$P(AB) = P(A)\,P(B)$$

2 . 2 . 3 Conditional Probability

In repeated experiments where events are dependent, if the occurrence of a certain event is known then the probability of sequential events can be determined.

The conditional probability of occurrence of event A assuming event B has occurred is denoted by P(B / A) and is written as

$$P(B / A) = P(AB) / P(B)$$

"*This result can be phrased as follows ;If one discards all the experiments in which the event B did not occur and retains only the subsequence of experiments in which B did occur, then P(A/B) equals the relative frequency of occurrence n(AB)/n(B) of the event A in that subsequence.*" [3]

2 . 3 Random Variables

2 . 3 .1 Definition

A random variable is a number assigned to every single outcome of an experiment. If it is to be viewed as a function, then it is defined as a function whose domain is a set of all experimental outcomes.

---

[3] Athanasios Papoulis, *Probability of Random Variables, and Stochastic Processes*, p.28.

One has the experiment E with the sample space S and V a subset of this sample space called events and probabilities assigned to these events. Thus a function **x** is created with domain the set S, and range a set of numbers. This function created is called a random variable if it satisfies the conditions below :

1. The set { **x** = x } is an event for every x.

2. The probabilities of P(8) and P(-8) is equal to zero.

2 . 3 . 2 Two Random Variables

In the case of two random variables, one has to consider bivariate statistics instead of marginal ones. When two random variables are concerned, it points out that one of them is the set { **x** = x } and the other is another set { **y** = y } both satisfying the conditions to be a random variable.

2 . 3 . 2 .1 Joint Probability

The joint probability of two random variables is defined as

P { **x** = x ∩ **y** = y} or simply shown as P { **x** = x , **y** = y}.

It follows from the definition that among a list of possible outcomes of two experiments the lines in which **x** = x and **y** = y appear at the same time will be counted as n(**x** , **y**) and sum of all lines being n, then joint probability in terms of relative frequency is;

$$P (x , y) = n(\mathbf{x} , \mathbf{y})/n$$

2 . 3 . 2 . 2 Conditional Probability

Conditional probability for two random variables is defined as the probability of **y** = y assuming that **x** = x and denoted as P (**y** = y / **x** = x) . Writing it in the open form :

$$P(\mathbf{y} = y / \mathbf{x} = x) = P(\mathbf{y} = y, \mathbf{x} = x) / P(\mathbf{x} = x).$$

In order to obtain this probability it is obvious that one has to know the joint probability of the two random variables as well as their marginal probabilities.

2 . 4 Entropy

Entropy is the measure of uncertainty associated with a random variable. [4]

In other words; the more one is certain that an event is to occur, the less is the entropy of that event, or the less likely it is for an event to occur, the less is its entropy. For a random variable say **x** its entropy is denoted by H(**x**) and is equal to

$$H(\mathbf{x}) = -\Sigma\, p(\mathbf{x} = x) \log p(\mathbf{x} = x)$$

or simply shown as

$$H(\mathbf{x}) = -\Sigma\, p(x) \log p(x)$$

It is known that the probability of a random variable is a number between 0 and 1. From this, it follows that entropy associated with a random variable which has only two values is also a number between 0 and 1. If $p(x) = 0$ or $p(x) = 1$ then the entropy in both cases is zero. In the first case, one is certain that the event will not occur and in the second one it is 100% that the event is to occur. In other words there is no uncertainty associated with the events, therefore their entropies are zero.

If the probability of a random variable is 0.1, then it can be concluded that the event is very less likely to occur. There is not much uncertainty. It is the same in the case where the probability is 0.9. This shows that the event is very much likely to occur, again pointing to a certain case.

[4] Pierce, J. R., *An Introduction to Information Theory*, p.13.

When the probability is 0.5, one is not certain whether the event is to occur or not. In other words the probability of the event occurring or not occurring is equal to each other. This is the case where uncertainty is maximum.

From the definition and the formulae it can be concluded that if all the events are likely to occur with the same probability then the entropy associated with the random variable is at its maximum value.

2 . 4 . 1 Joint Entropy

If a pair of random variables is taken instead of a single one, then one has to consider their bivariate statistics instead of marginal ones. Joint entropy can be defined as the uncertainty associated with a pair of random variables. It is formulated as follows ;

$$H(\mathbf{x}, \mathbf{y}) = -\Sigma \Sigma p(x, y) \log p(x, y)$$

$H(\mathbf{x}, \mathbf{y})$ is the joint entropy of the pair of random variable $\mathbf{x}$ and $\mathbf{y}$.

$p(x, y)$ is the joint probability of the pair of random variable $\mathbf{x}$ and $\mathbf{y}$.

2 . 4 . 2 Conditional Entropy

Considering the definition of conditional probability, conditional entropy can be explained in a similar way. For a pair of random variables $\mathbf{x}$ and $\mathbf{y}$, conditional entropy of $\mathbf{y}$ known that $\mathbf{x}$ has occurred is denoted by $H(\mathbf{y} / \mathbf{x})$ and is equal to

$$H(\mathbf{y} / \mathbf{x}) = -\Sigma \Sigma p(x, y) \log p(y / x)$$

$\log p(y / x)$ is the conditional probability of y assuming $\mathbf{x} = x$

To state this in another way it can be said that $H(\mathbf{y} / \mathbf{x})$ is the uncertainty associated with **y** = y in the subsequence of experiments that **x** = x has been observed.[5]

2 . 4 . 2 .1 Chain Rule

The chain rule sates that joint and conditional entropies are related in the following way ;

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y} / \mathbf{x})$$ [6]

This equality states that joint entropy is a function of marginal entropy and conditional entropy.

2 . 5 Stochastic Processes

2 . 5 . 1 Definition

" *A stochastic Process with parameter space T and a state space E is a collection of random variables* $\{ x_t, t \varepsilon T \}$ *defined on the same probability space and taking values in E".*[7]

Depending on the parameter and state space, processes are classified as discrete or continuous. For this thesis parameters are the letters of the Turkey Turkish alphabet , where there are 29 of them starting from A ending with Z, therefore the parameter space is discrete. For the state space, it is the position of the letter in the sequence of letters forming meaningful words, and since that is countable many values it is considered discrete state space. So the process that is to be analyzed is a discrete parameter and discrete state space type stochastic process.

---

[5] Athanasios Papoulis, *Probability of Random Variables and Stochastic Processes*, p.549.

[6] The proof is in Pierce, J. R., *An Introduction to Information Theory*, p.13.

[7] Şahinoğlu, Prof. Dr. Mehmet, *Applied Stochastic Processes*, p.13.

Stochastic processes can also be defined simply as a collection of indexed events such as x_1 , x_2 , x_3 , x_4 , ...x_{n-1} , x_n . For all those events there corresponds a state defining in what stage the process is and denoted as $x_n = j$ and is pronounced as the process is in state j at time n. And for any time the process being in a certain state is assigned a probability and denoted as $p (x_n = j)$ and is called the state probability.

2 . 5 . 2 Markov Processes

2 . 5 . 2 . 1 Definition

Markov processes are a special case of stochastic processes where the next state of the process is independent of the past states, if the previous state is known. For a stochastic process to be a Markov process, it has to have a discrete state space with finite number of elements in its parameter space. To formulate this phrase mathematically :

2 . 5 . 2 . 2 Mathematical Explanation

Let T = { 0 , 1, 2 , 3..} then for all positive integers q ,

$n_1 < n_2 < n_3 < n_4 \ldots\ldots\ldots\ldots < n_{q-1} < n_q$,

$$p (x_{n1} = k_1 , x_{n2} = k_2 , x_{nq} = k_q) \text{ is equal to}$$

$$p (x_{nq} = k_q \ / \ x_{n1} = k_1 , x_{nq-1} = k_{q-1}) \, p (x_{n1} = k_1 , x_{n2} = k_2 , \ x_{nq-1} = k_{q-1})$$

or in the simplest and compact form it is equal to

$$p (x_{n1} = k_1) \, \Pi \, p (x_{nj} = k_j \setminus x_{nj-1} = k_{j-1})$$ [8]

2 . 5 . 2 . 3 Order of a Markov Process

As well as state probabilities, another important feature of the Markov Processes is

[8] The proof is in Şahinoğlu, Prof. Dr. Mehmet, *Applied Stochastic Processes*, p.19.

the state transition probabilities, i.e. the probability of the process moving to the next state knowing that it is in its present state. If the process is known to be in state j at time m , and the probability in concern is the process being in state k at time n (provided that n > m) then the transition probability of the process being at state k at time n, given that it is at state j at time m can be written as

$$p_{j,k}\ (m,n) = p\ (x_n = k / x_m = j)$$

If n = m+1 then the above equality can be rewritten as

$$p_{j,k}\ (m,m+1) = p\ (x_{m+1} = k / x_m = j) = p_{j,k}\ (m)$$

$p_{j,k}$ (m) is called the one state transition probability and as can be noted, only one present time is shown. The state to be moved is understood as the state following the present one.

The number of steps needed to be taken to reach a certain state is referred to as the order of a Markov Process. In the one state transition probability, since to reach the desired state takes only one step, the order of the process is 1. To phrase it correctly ,it is called a First Order Markov Process implying that every state is a function of only the state that it precedes and independent of all the others before that state.

So it can be concluded that the order of a Markov Process is determined by the number of steps needed to be taken (or number of states that have to be passed through) to reach the desired state. Simply the order is the difference n-m.

There is a special case named as a Zero Order Markov Process . This yields to the conclusion that in order to reach a desired state, it is not necessary to take any step , meaning the state is independent of all other states and is determined only by itself. This can be shown by using definitions of conditional probability and independence.

$p (x_n = k / x_{n-1} = j , x_{n-2} = l , ... x_1 = w)$ is equal to

$p (x_n = k) \, p (x_{n-1} = j , x_{n-2} = l , ... x_1 = w) / p (x_{n-1} = j , x_{n-2} = l , ... x_1 = w)$.

Which then is simply equal to $p (x_n = k)$.

2 . 6 Examples

2 . 6 . 1 Letter frequencies

Taking the meaningful text produced in Turkey Turkish :

"Bu tez çalışması ileride kriptanaliz konusunda çalışmak isteyenlere yardımcı olmak amacı ile hazırlanmıştır."

One can obtain letter frequencies and probabilities as follows ;

In order to obtain the frequency of letter "a" (or "A") first the total number of a's n(a), present are counted and they add up to 12. Then the number of all letters forming the sentence n(total) are counted and they are 95. The relative frequency of letter "a" , the probability of observing an "a" in this text is calculated as

$$
\begin{aligned}
p(a) &= n(a) / n(total) \\
&= 12 / 95 \\
&= 0.1263
\end{aligned}
$$

For the relative frequency of letter "m", the total number of m's n(m) are 6 therefore the probability of observing an "m" in the text is

$$
\begin{aligned}
p(m) &= n(m) / n(total) \\
&= 6 / 95 \\
&= 0.0632
\end{aligned}
$$

2 . 6 . 2 Entropy

To illustrate entropy, the word “BİLGİ” will be used as an example. First the letter frequencies are calculated and they are ;

p(B) = 0.2

p(İ) = 0.4

p(L) = 0.2

p(G) = 0.2

The entropy of the letter “İ” in the word BİLGİ is the measure of uncertainty associated with “İ”. i.e. if one letter of the word BİLGİ is chosen how uncertain one is that the chosen letter is “İ”. The probability that it is an “İ” is 0.4 , so the entropy H (İ) is,

$$H(\dot{I}) = -(0.4 \log(0.4)) = 0.52877$$

Similarly the entropy of “L” is H (L) and is equal to

$$H(L) = -((0.2 \log(0.2)) = 0.46439$$

The uncertainty associated with İ is more because the probability is close to 0.5. For the letter L, since the probability is 0.2 one can almost be sure that the selected letter is not L, therefore resulting in less uncertainty and less entropy.

2 . 6 . 2 Entropy

To illustrate entropy, the word "BİLGİ" will be used as an example. First the letter frequencies are calculated and they are ;

p(B) = 0.2

p(İ) = 0.4

p(L) = 0.2

p(G) = 0.2

The entropy of the letter "İ" in the word BİLGİ is the measure of uncertainty associated with "İ". i.e. if one letter of the word BİLGİ is chosen how uncertain one is that the chosen letter is "İ". The probability that it is an "İ" is 0.4 , so the entropy H (İ) is,

H (İ) = - (0.4 log(0.4))

= 0.52877

Similarly the entropy of "L" is H (L) and is equal to

H(L) = - ((0.2 log (0.2))

= 0.46439

The uncertainty associated with İ is more because the probability is close to 0.5. For the letter L, since the probability is 0.2 one can almost be sure that the selected letter is not L, therefore resulting in less uncertainty and less entropy.

CHAPTER 3 : CRYPTOLOGICAL BACKGROUND

3 . 1 Introduction

This chapter is written to give some insight and information about cryptology; its vocabulary the uses and applications.

Cryptology originates from the two Greek words; **cryptos** meaning secret, and **logos** meaning science. Since historical times, the need to hide or protect information from unauthorized ones has been a major concern. For example; in times of war, one of the fighting sides may need to get a message to it's allies and they have to make sure that the message in concern is not to be seen or heard by its opponents. They can do this in many ways such as; having a messenger memorize a text and reveal it only to those who are authorized, use means of media and hope that no eavesdroppers are around, or to arrange the message in such a way that even if it is intercepted by others, the meaning will not be fully or thoroughly understood.

The last choice, but not the least, seems to be the best way of all. The drawbacks of the other two ways is; if a messenger is trapped by an opponent he/she may reveal the message under extraordinary circumstances such as torture or threat. The second way has its disadvantages that one can only hope for the best, but it must not be forgotten that hoping for the best brings along expecting the worst.

It is cryptology which makes it possible for people to arrange the information in such a way that, when intercepted by others, rather than the one(s) intended to receive it, does not make any sense.

3 . 2 The need for cryptology

Today cryptology is known as the science of secret and secure communication. The need for security in national or international communication, in military or diplomatic disciplines has been achieved by the secrecy that cryptology provides.

Computer systems and huge networks used in our daily lives has brought up the need to protect the information, from the unauthorized ones, and people who would destroy it. As a result of this, security has become more and more important.

Cryptology has been the answer to the need for secrecy and authentication and was mostly used by military purposes. Nowadays, if not secret but delicate information is transferred from one network to another such as health, insurance, credit card etc. records. These records are considered and should be kept private and need to be protected against alterations by unauthorized people . This privacy and protection is provided by cryptology .

3 . 3 Definitions

Cryptology or the science of secrecy is studied under two topics - Cryptography the science of ciphering and Cryptanalysis - cipher breaking.

3 . 3 . 1 Cryptography

A message is called a plaintext, or sometimes referred as the cleartext. The process of disguising a message in such a way as to hide its substance is named enciphering or encryption. An enciphered text is called a ciphertext or cryptogram. The process of converting a ciphertext into plaintext ,by all legal means is, deciphering or decryption.[1] It must be noted that deciphering is not as same as cipher breaking, which is totally a different activity.

Cryptography involves the act of producing a ciphertext - enciphering; such that when intercepted by unauthorized people it will not reveal its true meaning, and the act of obtaining the plaintext from ciphertext by legal means - deciphering. People practicing this science, cryptography, are called cryptographers.

3 . 3 . 1 . 1 Algorithms

When converting a plaintext into a ciphertext, cryptographers use different kind of algorithms which are also called keys. These algorithms are nothing but some set of transformation or substitution functions. The same holds true when a ciphertext is being deciphered. All of these algorithms used in the processes of ciphering and deciphering are called the cryptosystem.

3 . 3 . 1 . 2 Classification of Cryptosystems

Cryptosystems are classified according to the key(s) used in enciphering and deciphering.

---

[1] Bruce Schneier, *Applied Cryptography, Protocols, Algorithms, and Source Code in C*, p.1

3 . 3 . 1 . 2 . 1 Symmetric Cryptosystems

If the key used in enciphering and deciphering are the same then the cryptosystem is said to be symmetric.

To formulate this:

Message or plaintext	:	Denoted by **M** or **P** which can be stream of bits, digital image etc. or simply the message to be encrypted.
Encryption algorithm	:	Denoted by **E**.
Encryption key	:	**K**
Enciphered text	:	**C**

The encryption key used to encipher a message may be any one of the finite number of keys which make up the key space. So it is the cryptographer's choice to select the specific key to be used. This results in a new definition for the encryption function or a better way to state is:

Encryption function	:	E_k which must be a function that has an inverse.

E_k operates on **M** to produce **C**. To show this mathematically:

$$E_k(M) = C$$

can be written.

As for the decryption process a new function is needed to be defined;

Decryption algorithm : D_k

D_k will operate on the ciphertext to obtain the original message, or mathematically;

$$D_k (C) = M$$

or

$$D_k (E_k (M)) = M$$

As can be noted, the key used for both processes is the same, (k), which makes the cryptosystem symmetrical.

3 . 3 . 1 . 3 Asymmetrical Cryptosystems (Public-Key Cryptosystems)

In some cases the keys used for enciphering and deciphering are different from each other. The different key usage makes the type of the cryptosystem be called as asymmetric or public-key. With these systems, the encryption key is made public, i.e. any person has access to the encryption key (also called public key). He or she can use this key to encrypt a message, but this encrypted message can only be decrypted by the specific person who has the corresponding decryption key. This is why the decryption key is called the private key in such systems. With new definitions;

Encryption key : k_1

Decryption key : k_2

Plaintext, ciphertext, encryption and decryption functions being the same as the ones defined in the previous part, the system works as;

E_{k1} operates on M to produce the ciphertext C:

$$E_{k1} (M) = C$$

Then D_{k2} operates on C to obtain the original plaintext M.

$$D_{k2} (C) = M$$

or

$$D_{k2}\,(\,E_{k1}\,(\,M\,)\,)=M$$

3 . 3 . 2 Cryptanalysis

The whole point of cryptography is to keep the plaintext, or the key, or both, secret from eavesdroppers, intruders, interceptors or as generally called the enemies.[2] These people are assumed to have access to the communication between the sender and the receiver.

Cryptanalysis is the science of recovering the plaintext of an enciphered message without having a legal access to the key. People who are practicing this science are called cryptanalysts. A successful cryptanalyst may recover the key and then use this key to obtain the plaintext or sometimes even in the absence of any knowledge of the key may recover the plaintext.

Nowadays the science of cryptanalysis holds within itself many sciences that may seem to be irrelevant to one another such as; Probability Theory, Statistics, Information Theory, Thermodynamics, Linguistics, and Computer Sciences, which on the other hand are interconnected to each other by means of mathematical equalities. The broad field of

---

[2] Bruce Schneier, *Applied Cryptography, Protocols, Algorithms, and Source Code in C*, p.5

sciences that cryptanalysis contains in itself explains the state of the art that it has reached. Keeping in mind the aim in concern, from time to time it is named as "Black Art".[3]

Any attempted act of cryptanalysis is called an attack. Based on the assumption that the cryptanalyst has complete knowledge of the encryption algorithm used, attacks are classified into four groups; ciphertext-only, known-plaintext, chosen-plaintext, and adaptive-chosen-plaintext attacks.

3 . 3 . 2 . 1 Ciphertext-only Attack

The cryptanalyst has the ciphertexts of several messages, which all have been encrypted by the use of the same encryption algorithm. When this is the case, the cryptanalyst may choose to recover as many original plaintexts of these cryptograms or a better choice may be to deduce the key(s) which were used to encrypt the messages. Once the key(s) is(are) obtained then any message encrypted with it can easily be decrypted to obtain the corresponding original plaintex. Shortly ;

Given : $C_1 = E_k(M_1), C_2 = E_k(M_2), \ldots\ldots, C_i = E_k(M_i)$

Aim : Obtain $M_1, M_2, \ldots, M_i$; k;

or an algorithm that will produce P_{i+1} from $C_{i+1} = E_k(P_{i+1})$

3 . 3 . 2 . 2 Known-plaintext Attack

The cryptanalyst does not only has access to the ciphertext of several messages he

[3] James Bamford, *The Puzzle Palace*, pp. 5-55

also has plaintext of those several messages. What he has to do is simply find the key(s) which was used to encrypt the messages, or come up with an algorithm which will decrypt any new message encrypted with the same key(s).

Given : $M_1, C_1 = E_k(M_1), M_2, C_2 = E_k(M_2), \ldots, M_i\ C_i = E_k(M_i)$

Aim : Obtain k;

or an algorithm that will produce P_{i+1} from $C_{i+1} = E_k(P_{i+1})$

3 . 3 . 2 . 3 Chosen-plaintext Attack

The cryptanalyst does not only have access to ciphertexts and plaintexts of several messages, he also chooses the plaintext that is encrypted. This is more powerful than a known-plaintext attack because the cryptanalyst can choose specific plaintext blocks to encrypt, ones which are likely to yield more information about the key. His job is simply find the key(s) which was used to encrypt the messages, or come up with an algorithm which will decrypt any new message encrypted with the same key(s).

Given : $M_1, C_1 = E_k(M_1), M_2, C_2 = E_k(M_2), \ldots, M_i\ C_i = E_k(M_i)$

where the cryptanalyst gets to choose from $M_1, M_2, \ldots, M_i$

Aim : Obtain k;

or an algorithm that will produce P_{i+1} from $C_{i+1} = E_k(P_{i+1})$

3 . 3 . 2 . 4 Adaptive chosen-plaintext Attack

This is a special case of chosen-plaintext attack. The cryptanalyst, together with the choice of plaintext that is encrypted, can also modify his selection based on the results of the previous encryption. To make the distinction more clear; the cryptanalyst may choose a smaller part of a plaintext than he does in chosen-plaintext attack, and depending on the result of this analysis, he may choose a larger block the next time, even a larger block the following time, and so on.

Besides these, there are three other, not generally recognized, groups of attacks:

3 . 3 . 2 . 5 Chosen-ciphertext Attack

The cryptanalyst can choose different ciphertexts to be decrypted and he has access to the decrypted plaintext. An example would be that the cryptanalyst has access to the machine/system that does the decryption automatically. Then what he is to determine is, the key.

Given : C_1 ,$M_1 = D_k(C_1)$, C_2 ,$M_2 = D_k(C_2)$,. . . . . . , C_i $M_i = D_k(C_i)$

Aim : Obtain k

3 . 3 . 2 . 6 Chosen-key Attack

The name of the attack is misleading so it does not imply that the cryptanalyst can choose the key. It is described as the situation where the cryptanalyst has some knowledge about the relationship between different keys. This type of attack is strange and obscure, and also not practical.[4]

3 . 3 . 2 . 7 Rubber-hose Cryptanalysis

This is the case in which the cryptanalyst threatens, blackmails, or tortures someone who has access to the key, or even bribe him to give him the key. Getting the key by bribery is sometimes called purchase-key attack. All of these attacks, though not ethical, are all very powerful and often best way to break an algorithm.

3 . 4 Security of Algorithms

Depending on how hard it is to break an algorithm, constitutes its degree of security. Founder's of cryptographic algorithms may consider themselves safe under the following conditions:

- If the cost required to break an algorithm exceeds the value of the data that is encrypted.
- If the time required to break an algorithm is longer than the time the encrypted data should and must remain secret.
- If the amount of data encrypted with a single key is less than the data required to break the algorithm.

As for the cryptanalysis point of view; the value of the data to be obtained after deciphering, should and always be less than the cost to break the security protecting it.

---

[4] Bruce Schneier, *Applied Cryptography, Protocols, Algorithms, and Source Code in C*, p.7

According to Lars Knudsen, different catagories of breaking algorithms are classified in decreasing order of severity are: [5]

1) Total break. The cryptanalyst finds the key K, such that

$$D_k (C) = M.$$

2) Global Detection. The cryptanalyst finds an alternate algorithm, A, equivalent to $D_k (C)$ without knowing k.

3) Instance (local) deduction. The cryptanalyst finds the plaintext of an intercepted ciphertext.

4) Information deduction. The cryptanalyst gains some information about the key or plaintext, which can be a few bits of the key, some information about the form of the plaintext, etc.

3 . 4 . 1 Unconditional and Computational Security

No matter how much ciphertext a cyptanalyst intercepts, if there is not enough information for him to recover the original plaintext, the algorithm used to produce the ciphertext is considered unconditionally secure.

An algorithm is considered to be computationally secure, if it cannot be broken with any resources available [6] either existing or have a chance to be found in the future.

[5] Bruce Schneier, *Applied Cryptography, Protocols, Algorithms, and Source Code in C*, p.8

[6] This term is left to interpretation with the attacks summarized earlier.

3 . 5 Complexity of an Attack

The complexity of an attack is usually taken to be the minimum of the three factors listed below:

1) The amount of data needed as input to the attack; data complexity.
2) The time it takes to perform the attack; sometimes called the work factor but more often referred as processing complexity.
3) The amount of memory needed to perform the attack; storage requirements.

CHAPTER 4 : ANALYSIS OF TURKEY TURKISH WORDS BY DISCRETE MARKOV PROCESSES

4 . 1 Markov Processes

Using a zero order approach to obtain the probability of observing a certain word say "BİLGİ" can be explained as follows:

Each letter can be considered as a random variable and these random variables forming a function, a word. For the zero order approach all letters are independent and are not a function of the letters before them. Since this is the case, the marginal probabilities are to be considered, i.e. the letter frequencies. The probability of observing the word "BİLGİ" among all 5-letter words in a plaintext depends on the product of individual probabilities of the letters. Or :

$$p\,(BİLGİ) = p(B)\,p(İ)\,p(L)\,p(G)\,p(İ)$$

Using numerical results of a cryptanalytical measure study, the probability of occurrence of the word BİLGİ is ;

$$p\,(BİLGİ) = (0.0295)\,(0.0827)\,(0.0575)\,(0.0134)\,(0.0827)$$

$$= 1.5546\text{ E - }07$$

p(B) for example is the ratio of the total number of B's counted within the text to the total number of letters that make up the text, i.e. the relative frequency of B.

A first order Markov Process approach tells that the process is independent of past states if its present state is known. So each state is dependent on the state that it precedes. Applying this approach to the word "BİLGİ" , the letter B does not follow any letter so one has to consider its marginal probability. The letter L is followed by İ (the first İ) so the probability to be considered is the conditional probability of L known that the letter it follows is İ . So the probability of observing the word "BİLGİ" among all 5-letter words in a plaintext is :

$p\,(BİLGİ) = p\,(B)\ p_B(İ)\ p_İ(L)\ p_L\,(G)\ p_G\,(İ)$

$p_B(İ) = p\,(İ\,/\,B) = p(Bİ)\,/\,p(B)$

p(Bİ) is the ratio of total number of Bİ's found in the text to the total number of 2-letter combinations.

4 . 2 Consonant - Vowel (c-v) Patterns

Among the letters in the Turkey Turkish alphabet a, e, ı, i, o, ö, u, ü are classified as vowels, and the rest is named as consonants. To obtain the c-v pattern for example say the word "BİLGİ" one substitutes c for all consonants (B, L, G), and v for vowels (İ's).

$(BİLGİ)_{cv} = CVCCV$

For joint and conditional probabilities these patterns are used instead of the letters that make up the words. The results obtained are all based on the c-v patterns that the words or modified phrases represent.

The joint probability of the occurrence word BİLGİ of the 0 order approach is

$p(BİLGİ) = p(CVCCV) = p\,(C)\ p(V)\ P(C)\ P(C)\ p(V)$

A first order approach for the conditional probability would be ;

$p(CVCCV)_1 = p(C)\ p(V/C)\ p(C/C)\ p(C/V)$

4 . 3 Analysis

The results of the study by Ass. Prof. Dr. Ahmet Hasan KOLTUKSUZ, (the probability of 1 to 5 letter long c-v patterns) were used to obtain the conditional probabilities of c-v patterns. The overall results are turned into a table (Table 1).

The entries in bold letters are the results deducted from the original table formed by Dr. Koltuksuz's works. They are the conditional probabilities obtained by the use of marginal probabilities.

To make the concept clearer here is an example how the new entries are deducted:

The total number of vowels that were present in the bundle of text is n(v)=2.283.012, where as the total number of letters vowel or consonant added up to n(t)=5.321.885. Therefore applying the concept of relative frequency, or since the number is high enough the word probability can be used, probability of observing a v all through the text is p(v) = n(v) / n(t) = 2.283.012 / 5.321.885 = 0,428986.

To obtain the conditional probability values the procedure used is as follows;

The conditional probability of observing a v, knowing that the letter coming before that is a v, can be stated as the conditional probability of v given v and denoted as $p_v(v)$. Applying the concept of conditional probability

$p_v(v) = n(vv) / n(v)$;

n(vv) is the total of vv combinations in the bundle of text (taken from the second series of rows in the table)

n(v) is the total number of v's in the text.

So $p_v(v)$ = 105.043 / 2.283.012 = 0,046011

Table 1. Conditional probabilities for the Markov Orders and Related Entropies of Turkey Turkish Words

Known	frequency	probability	**given : v**	**given : c**	sum
v	2283012	0,428986	**0,046011**	**0,953989**	1,0000
c	3038873	0,571014	**0,716703**	**0,283297**	1,0000
Sum	5321885	1,000000			

Known	frequency	probability	**given : v**	**given : c**	sum
vv	105043	0,019738	**0,009444**	**0,990556**	1,0000
vc	2177969	0,409248	**0,611186**	**0,388814**	1,0000
cv	2177969	0,409248	**0,047774**	**0,952226**	1,0000
cc	860903	0,161767	**0,983646**	**0,016354**	1,0000
Sum	5321884	1,000000			

Known	frequency	probability	**given : v**	**given : c**	sum
vvv	992	0,000186	**0,003024**	**0,996976**	1,0000
vvc	104051	0,019552	**0,542955**	**0,457045**	1,0000
vcv	1331144	0,250127	**0,045489**	**0,954511**	1,0000
vcc	846824	0,159121	**0,983639**	**0,016361**	1,0000
cvv	104051	0,019552	**0,009505**	**0,990495**	1,0000
cvc	2073918	0,389696	**0,614609**	**0,385390**	1,0000
ccv	846824	0,159121	**0,051366**	**0,948634**	1,0000
ccc	14079	0,002645	**0,984090**	**0,015910**	1,0000
Sum	5321883	1,000000			

Table 1. Conditional probabilities for the Markov Orders and Related Entropies of Turkey Turkish Words (Continued)

Known	frequency	probability	**given : v**	**given : c**	sum
vvvv	3	0,000001	**0,000000**	**1,000000**	1,0000
vvvc	989	0,000186	**0,618807**	**0,381193**	1,0000
vvcv	56495	0,010616	**0,044022**	**0,955978**	1,0000
vvcc	47556	0,008936	**0,978573**	**0,021427**	1,0000
vcvv	60553	0,011378	**0,008967**	**0,991033**	1,0000
vcvc	1270591	0,238748	**0,597392**	**0,402607**	1,0000
vccv	832969	0,156518	**0,051530**	**0,948470**	1,0000
vccc	13855	0,002603	**0,985565**	**0,014435**	1,0000
cvvv	989	0,000186	**0,003033**	**0,996967**	1,0000
cvvc	103062	0,019366	**0,542227**	**0,457773**	1,0000
cvcv	1274649	0,239511	**0,045555**	**0,954445**	1,0000
cvcc	799268	0,150185	**0,983940**	**0,016060**	1,0000
ccvv	43498	0,008173	**0,010253**	**0,989747**	1,0000
ccvc	803326	0,150948	**0,641842**	**0,358158**	1,0000
cccv	13855	0,002603	**0,041501**	**0,958499**	1,0000
cccc	224	0,000042	**0,892857**	**0,107143**	1,0000
Sum	5321882	1,000000			

The number of words that the analysis was performed over adds up to 21,395. To make the method clear, some examples are listed below:

One letter words

word	:	o
c-v pattern	:	v
number of letters	:	1
Markov order	:	0
p(v)	:	0.428986

Two letter words

word : et

c-v pattern : vc

number of letters : 2

Markov order : 0

$p(v)p(c)$: : 0.048986 * 0.571014 = 0.244957

Markov order : 1

$p(v)p_v(c)$: 0.428986 * 0.953989 = 0.409248

Three letter words

word : aba

c-v pattern : vcv

number of letters : 3

Markov order : 0

$p(v)p(c)p(v)$: : 0.105083

Markov order : 1

$p(v)p_v(c)p_c(v)$: 0.233309

Markov order : 2

$p(vc)p_{vc}(v)$: 0.250127

Four letter words

word : baca

c-v pattern : cvcv

number of letters : 4

Markov order : 0

$p(c)p(v)p(c)p(v)$: 0.600004

Markov order : 1

$p(c)p_c(v)p_v(c)p_c(v)$: 0.279813

Markov order : 2

$p(cv)p_{cv}(c)p_{vc}(v)$: 0.238177

Markov order : 3

$p(cvc)p_{cvc}(v)$: 0.239511

Five letter words

word	:	engin
c-v pattern	:	vccvc
number of letters	:	5
Markov order	:	0
$p(v)p(c)p(c)p(v)p(c)$	:	0.034263
Markov order	:	1
$p(v)p_v(c)p_c(c)p_c(v)p_v(c)$	:	0.079270
Markov order	:	2
$p(vc)p_{vc}(c)p_{cc}(v)p_{cv}(c)$	:	0.149042
Markov order	:	3
$p(vcc)p_{vcc}(v)p_{ccv}(c)$	:	0.148478
Markov order	:	4
$p(vccv)p_{vccv}(c)$	:	0.148453

4 . 4 Presentation of the analysis

The analysis was performed over 21,395 words of different word lengths. Assuming the average word length is 4 letters, in order to display the results, in the standard form chosen requires 7 lines;

1 line for the word itself

1 line for its c-v pattern

1 line for the word length

4 lines for 0 to 3rd order approach results and relevant entropic values.

The fourth chapter with a rough estimate would be consisting of 149,765 lines. A standard A4 sheet can take up to almost 50 lines with the text written in Arial 10. To display the results of this specific analysis one copy requires 3,000 standard A4 sheets, which is equal to 6 A4 packs of paper.

Due to the time that it would take to print out, the finance the process demands, and the fact that 5 copies are to be made, the results are compressed with the use of WinZip software and then saved on a 3½ floppy disk.

4 . 5 Examples from the Analysis

Since it was not efficient to display all the results, some examples were selected from some letters to give some idea about the conditional probabilities of c-v patterns. They are presented in Appendix 1.

CHAPTER 5 : COMMENTS and FURTHER STUDIES

5 . 1 Comments

If the conditional probability and conditional entropy values are analyzed, it can be seen that starting from 2^{nd} order and on, these probability and entropy values are very close to each other and in most cases they are exactly the same. (i.e. through 0, 1^{st}, and 2^{nd} orders the probability changes, then stays almost the same through 3^{rd} and 4^{th} orders.)

This can be interpreted as follows;

No matter how many characters are in a word, if one knows the first three letters then to make up the whole word from that three letters can be easily done. Knowing the 4^{th} or the 5^{th} or more letters of the word do not provide more information than knowing the first three letters.

Reviewing the structure of Turkey Turkish words, the above conclusion is no surprise. In order to pronounce a vowel, one needs a consonant before or after that vowel. It is very very rare to see two vowels and/or three consonants sequentially.

With these results, a cryptanalyst may easily obtain the c-v pattern of an enciphered text. If this study is to be done on letter basis instead of c-v pattern basis, then a cryptanalyst may easily differentiate between the letters. For example if he comes across a three letter word say b?l, looking at the entropy values he can decide which one is most probable; bil, bul, bal, bol, böl etc.

5 . 2 Further Studies

Although cryptanalytical measure studies in most other languages (English, German, French etc.) have been completed, works on Turkey Turkish have started just recently.

These results may not be meaningful all by themselves but they are obtained by using the results of the previous study.

The results obtained from the analysis in this thesis may well constitute basis for another study for example analyzing the words in character format instead of c-v patterns.

If all these results(previous, these, and future ones) are to be stored in a database, which can perform queries and useful sorts, then all these numerical values would have a meaning.

SUMMARY

The Redhouse Turkish Dictionary was transferred to the electronic environment and then gone under a filtration process. The filtration process involved all words starting with capital letters being replaced with small case ones, spaces between idioms and two or more word phrases being deleted to make them appear as a single word, and words that are spelled exactly the same but carry different meanings were eliminated so that only a single one was left in the sample space.

Cryptanalytical measures of Turkey Turkish words converted into their corresponding c-v patterns were obtained by Markov processes approach. These measures were obtained for 0, 1st, 2nd, 3rd, and 4th degree approaches each. For each word available in the sample space and/or dictionary;

1) The word itself
2) It's c-v pattern
3) Word length
4) Its c-v pattern's conditional probability starting from 0 order to n_{th} order (n=wordlength-1)
5) For each order the corresponding entropy values were calculated.

The number of words analyzed is 21,395.

ÖZET

Redhouse Türkçe sözlüğü elektronik ortama aktarılmış ve daha sonra saflaştırma işlemine tabii tutulmuştur. Saflaştırma işleminin aşamaları sırasıyla şöyledir; büyük harfle başlayan kelimeler küçük harflerle değiştirilmiş, deyimler ve birden fazla sözcükten oluşan isimler arasındaki boşluklar silinmiş, aynı şekilde yazılan fakat farklı anlamlar taşıyan kelimelerden yanlızca bir tanesi kalmak üzere diğerleri örnek uzayından silinmiş, son olarak günümüz Türkçesi'nde kullanılmayan sözcükler elimine edilmiştir.

Türkiye Türkçesi'nde kullanılan kelimelerin kriptanalitik ölçütleri ayrık Markov yaklaşımlarıyla belirlenmiştir. Bu ölçütler sırasıyla 0, 1., 2., 3. ve 4. derece yaklaşımlarla elde edilmiştir. Örnek uzayında ve/veya sözlükte yer alan tüm kelimeler için;

1) Kelimenin kendisi
2) Sesli-sessiz deseni
3) Kelime uzunluğu
4) 0 ile n arası yaklaşımların her biri için sesli-sessiz deseninin koşullu olasılık (n=kelime uzunluğu -1)
5) Her derece için karşılık gelen entropi değerleri belirlenmiştir.

Analiz edilen toplam sözcük sayısı 21,395 tanedir.

BIBLIOGRAPHY

BAMFORD, James, *The Puzzle Palace*, Houghton - Mifflin Co., Boston, U.S.A., 1982

KOLTUKSUZ, Ahmet Hasan,*Simetrik Kriptosistemler için Türkiye Türkçesinin Kriptanalitik Ölçütleri*, unpublİshed Ph.D. thesis, Ege Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İzmir 1995.

MADRON, Thomas W., *Network Security in the 90's*, John Wiley & Sons, U.S. 1992.

PAPOULIS, Athanasios, *Probability, Random Variables and Stochastic Processes*, Mcgraw-Hill, U.S. 1991.

PIERCE, J. R., *An Introduction to Information Theory*, 2nd ed, Dover Pub., Inc, New York, U.S.A., 1980.

SCHNEIER, Bruce, *Applied Cryptography, Protocols, Algorithms, and Source Code in C.*, John Wiley & Sons, U.S. 1995

SHANNON, C. E. , "Communication Theory of Secrecy Systems", *Bell System Technical Journal*, vol. 28, October 1949.

SHANNON, C. E. , "Prediction and Entropy of Printed English", *Bell System Technical Journal*, Jan 1951.

SIMONDS, Fred, *Network Security*, John Wiley & Sons, U.S 1992.

ŞAHİNOĞLU, Prof. Dr. Mehmet, *Applied Stochastic Processes*, 1. Baskı Set Ofset Ltd. Sti., Ankara, Türkiye, 1992.

APPENDIX 1. Selected examples from the analysis

Word	:	aba
Pattern	:	VCV
Word Length	:	3
Order	Markov	Entropy
0th.	0.105083	0.341562
1st.	0.293309	0.519012
2nd.	0.250127	0.500071

Word	:	abaküs
Pattern	:	VCVCVC
Word Length	:	6
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.191316	0.456475
2nd.	0.138616	0.395171
3rd.	0.140062	0.397197
4th.	0.136129	0.391637

Word	:	abalı
Pattern	:	VCVCV
Word Length	:	5
Order	Markov	Entropy
0th.	0.025741	0.135906
1st.	0.200543	0.464862
2nd.	0.145571	0.404717
3rd.	0.146737	0.406270
4th.	0.142626	0.400735

Word	:	abandone
Pattern	:	VCVCCVCV
Word Length	:	8
Order	Markov	Entropy
0th.	0.003600	0.029227
1st.	0.038845	0.182032
2nd.	0.053014	0.224647
3rd.	0.052769	0.223959
4^{th}.	0.057576	0.237120

Word	:	abanmak
Pattern	:	VCVCCVC
Word Length	:	7
Order	Markov	Entropy
0th.	0.008393	0.057883
1st.	0.054199	0.227940
2nd.	0.086740	0.305946
3rd.	0.085857	0.304099
4th.	0.089704	0.312052

Word	:abanoz	
Pattern	:VCVCVC	
Word Length	:6	
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.191316	0.456475
2nd.	0.138616	0.395171
3rd.	0.140062	0.397197
4th.	0.136129	0.391637

Word :abartı		
Pattern :VCVCCV		
Word Length :6		
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.056813	0.235072
2nd.	0.091092	0.314863
3rd.	0.090506	0.313680
4th.	0.094578	0.321787

Word :abartılı		
Pattern :VCVCCVCV		
Word Length :8		
Order	Markov	Entropy
0th.	0.003600	0.029227
1st.	0.038845	0.182032
2nd.	0.053014	0.224647
3rd.	0.052769	0.223959
4th.	0.057576	0.237120

Word :abartılmak		
Pattern :VCVCCVCCVC		
Word Length :10		
Order	Markov	Entropy
0th.	0.001174	0.011428
1st.	0.010498	0.069013
2nd.	0.031589	0.157455
3rd.	0.030875	0.154914
4th.	0.029983	0.151707

Word	:bulgu	
Pattern	:CVCCV	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.079270	0.289897
2nd.	0.149042	0.409300
3rd.	0.147728	0.407579
4th.	0.147773	0.407639

Word	:bulgur	
Pattern	:CVCCVC	
Word Length	:6	
Order	Markov	Entropy
0th.	0.019565	0.111041
1st.	0.075623	0.281698
2nd.	0.141921	0.399769
3rd.	0.140140	0.397305
4th.	0.140158	0.397331

Word	:bulmaca	
Pattern	:CVCCVCV	
Word Length	:7	
Order	Markov	Entropy
0th.	0.008393	0.057883
1st.	0.054199	0.227939
2nd.	0.086740	0.305946
3rd.	0.086131	0.304673
4th.	0.089959	0.312571

Word	:bulmak	
Pattern	:CVCCVC	
Word Length	:6	
Order	Markov	Entropy
0th.	0.019565	0.111041
1st.	0.075623	0.281698
2nd.	0.141921	0.399769
3rd.	0.140140	0.397305
4th.	0.140158	0.397331

Word	:buluğ	
Pattern	:CVCVC	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.266939	0.508630
2nd.	0.226798	0.485466
3rd.	0.228616	0.486724
4th.	0.228600	0.486713

Word	:bulunç	
Pattern	:CVCVCC	
Word Length	:6	
Order	Markov	Entropy
0th.	0.019565	0.111041
1st.	0.075623	0.281698
2nd.	0.088182	0.308935
3rd.	0.088106	0.308778
4th.	0.092036	0.316756

Word :bulundurmak
Pattern :CVCVCCVCCVC
Word Length :11
Order Markov Entropy
0th. 0.000670 0.007067
1st. 0.010015 0.066518
2nd. 0.030080 0.152057
3rd. 0.029565 0.150189
4th. 0.028709 0.147057

Word :bulunmak
Pattern :CVCVCCVC
Word Length :8
Order Markov Entropy
0th. 0.004793 0.036926
1st. 0.051705 0.220965
2nd. 0.082596 0.297163
3rd. 0.082213 0.296336
4th. 0.085891 0.304171

Word :bulunmaz
Pattern :CVCVCCVC
Word Length :8
Order Markov Entropy
0th. 0.004793 0.036926
1st. 0.051705 0.220965
2nd. 0.082596 0.297163
3rd. 0.082213 0.296336
4th. 0.085891 0.304171

Word :buluntu
Pattern :CVCVCCV
Word Length :7
Order Markov Entropy
0th. 0.008393 0.057883
1st. 0.054199 0.227939
2nd. 0.086740 0.305946
3rd. 0.086665 0.305789
4th. 0.090558 0.313785

Word :enzim
Pattern :VCCVC
Word Length :5
Order Markov Entropy
0th. 0.034263 0.166765
1st. 0.079270 0.289898
2nd. 0.149042 0.409300
3rd. 0.148478 0.408564
4th. 0.148453 0.408531

Word :epey
Pattern :VCVC
Word Length :4
Order Markov Entropy
0th. 0.060004 0.243544
1st. 0.279814 0.514147
2nd. 0.238177 0.493001
3rd. 0.238749 0.493359

Word :epeyce	
Pattern :VCVCCV	
Word Length :6	
Order	Markov Entropy
0th.	0.014698 0.089487
1st.	0.056813 0.235072
2nd.	0.091092 0.314863
3rd.	0.090506 0.313680
4th.	0.094578 0.321787

Word :epik	
Pattern :VCVC	
Word Length :4	
Order	Markov Entropy
0th.	0.060004 0.243544
1st.	0.279814 0.514147
2nd.	0.238177 0.493001
3rd.	0.238749 0.493359

Word :eprimek	
Pattern :VCCVCVC	
Word Length :7	
Order	Markov Entropy
0th.	0.008393 0.057883
1st.	0.054199 0.227940
2nd.	0.086740 0.305946
3rd.	0.087105 0.306705
4th.	0.090943 0.314561

Word :er
Pattern :VC
Word Length :2
Order Markov Entropy
0th. 0.244957 0.497116
1st. 0.409248 0.527501

Word :erat
Pattern :VCVC
Word Length :4
Order Markov Entropy
0th. 0.060004 0.243544
1st. 0.279814 0.514147
2nd. 0.238177 0.493001
3rd. 0.238749 0.493359

Word :erbap
Pattern :VCCVC
Word Length :5
Order Markov Entropy
0th. 0.034263 0.166765
1st. 0.079270 0.289898
2nd. 0.149042 0.409300
3rd. 0.148478 0.408564
4th. 0.148453 0.408531

Word	:erbaş	
Pattern	:VCCVC	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.079270	0.289898
2nd.	0.149042	0.409300
3rd.	0.148478	0.408564
4th.	0.148453	0.408531

Pattern	:VCCVCV	
Word Length	:6	
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.056813	0.235072
2nd.	0.091092	0.314863
3rd.	0.091256	0.315192
4th.	0.095283	0.323166

Word	:içtüzük	
Pattern	:VCCVCVC	
Word Length	:7	
Order	Markov	Entropy
0th.	0.008393	0.057883
1st.	0.054199	0.227940
2nd.	0.086740	0.305946
3rd.	0.087105	0.306705
4th.	0.090943	0.314561

Word :içyağı		
Pattern :VCCVCV		
Word Length :6		
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.056813	0.235072
2nd.	0.091092	0.314863
3rd.	0.091256	0.315192
4th.	0.095283	0.323166

Word :içyapı		
Pattern :VCCVCV		
Word Length :6		
Order	Markov	Entropy
0th.	0.014698	0.089487
1st.	0.056813	0.235072
2nd.	0.091092	0.314863
3rd.	0.091256	0.315192
4th.	0.095283	0.323166

Word :içyüz		
Pattern :VCCVC		
Word Length :5		
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.079270	0.289898
2nd.	0.149042	0.409300
3rd.	0.148478	0.408564
4th.	0.148453	0.408531

Word	:idam	
Pattern	:VCVC	
Word Length	:4	
Order	Markov	Entropy
0th.	0.060004	0.243544
1st.	0.279814	0.514147
2nd.	0.238177	0.493001
3rd.	0.238749	0.493359

Word	:idame	
Pattern	:VCVCV	
Word Length	:5	
Order	Markov	Entropy
0th.	0.025741	0.135906
1st.	0.200543	0.464862
2nd.	0.145571	0.404717
3rd.	0.146737	0.406270
4th.	0.142626	0.400735

Word	:idamlık	
Pattern	:VCVCCVC	
Word Length	:7	
Order	Markov	Entropy
0th.	0.008393	0.057883
1st.	0.054199	0.227940
2nd.	0.086740	0.305946
3rd.	0.085857	0.304099
4th.	0.089704	0.312052

Word :idare		
Pattern :VCVCV		
Word Length :5		
Order	Markov	Entropy
0th.	0.025741	0.135906
1st.	0.200543	0.464862
2nd.	0.145571	0.404717
3rd.	0.146737	0.406270
4th.	0.142626	0.400735

Word :idareci		
Pattern :VCVCVCV		
Word Length :7		
Order	Markov	Entropy
0th.	0.006305	0.046087
1st.	0.137117	0.393049
2nd.	0.084720	0.301701
3rd.	0.086084	0.304574
4th.	0.081322	0.294403

Word :idareli		
Pattern :VCVCVCV		
Word Length :7		
Order	Markov	Entropy
0th.	0.006305	0.046087
1st.	0.137117	0.393049
2nd.	0.084720	0.301701
3rd.	0.086084	0.304574
4th.	0.081322	0.294403

Word :idareten		
Pattern :VCVCVCVC		
Word Length :8		
Order	Markov	Entropy
0th.	0.003600	0.029227
1st.	0.130808	0.383853
2nd.	0.080673	0.292985
3rd.	0.082168	0.296238
4th.	0.077618	0.286213

Word :kolaylanmak		
Pattern :CVCVCCVCCVC		
Word Length :11		
Order	Markov	Entropy
0th.	0.000670	0.007067
1st.	0.010015	0.066518
2nd.	0.030080	0.152057
3rd.	0.029565	0.150189
4th.	0.028709	0.147057

Word :kolaylaşmak		
Pattern :CVCVCCVCCVC		
Word Length :11		
Order	Markov	Entropy
0th.	0.000670	0.007067
1st.	0.010015	0.066518
2nd.	0.030080	0.152057
3rd.	0.029565	0.150189
4th.	0.028709	0.147057

Word	:kolaylaştırıcı
Pattern	:CVCVCCVCCVCVCV
Word Length	:14
Order Markov Entropy	
0th. 0.000070 0.000972	
1st. 0.004908 0.037646	
2nd. 0.010700 0.070043	
3rd. 0.010660 0.069840	
4th. 0.010506 0.069054	

Word	:kolaylaştırmak
Pattern	:CVCVCCVCCVCCVC
Word Length	:14
Order Markov Entropy	
0th. 0.000094 0.001255	
1st. 0.001940 0.017478	
2nd. 0.010955 0.071341	
3rd. 0.010632 0.069697	
4th. 0.009596 0.064325	

Word	:kolaylık
Pattern	:CVCVCCVC
Word Length	:8
Order Markov Entropy	
0th. 0.004793 0.036926	
1st. 0.051705 0.220965	
2nd. 0.082596 0.297163	
3rd. 0.082213 0.296336	
4th. 0.085891 0.304171	

Word	:kolböreği
Pattern	:CVCCVCVCV
Word Length	:9
Order Markov Entropy	
0th. 0.002056 0.018351	
1st. 0.037057 0.176175	
2nd. 0.050482 0.217480	
3rd. 0.050529 0.217615	
4th. 0.051293 0.219795	

Word	:kolcu
Pattern	:CVCCV
Word Length	:5
Order Markov Entropy	
0th. 0.034263 0.166765	
1st. 0.079270 0.289897	
2nd. 0.149042 0.409300	
3rd. 0.147728 0.407579	
4th. 0.147773 0.407639	

Word	:kolçak
Pattern	:CVCCVC
Word Length	:6
Order Markov Entropy	
0th. 0.019565 0.111041	
1st. 0.075623 0.281698	
2nd. 0.141921 0.399769	
3rd. 0.140140 0.397305	
4th. 0.140158 0.397331	

Word	:koldaş	
Pattern	:CVCCVC	
Word Length	:6	
Order	Markov	Entropy
0th.	0.019565	0.111041
1st.	0.075623	0.281698
2nd.	0.141921	0.399769
3rd.	0.140140	0.397305
4th.	0.140158	0.397331

Word	:kolej	
Pattern	:CVCVC	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.266939	0.508630
2nd.	0.226798	0.485466
3rd.	0.228616	0.486724
4th.	0.228600	0.486713

Word	:sağır	
Pattern	:CVCVC	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.266939	0.508630
2nd.	0.226798	0.485466
3rd.	0.228616	0.486724
4th.	0.228600	0.486713

Word	:sağırlaşmak	
Pattern	:CVCVCCVCCVC	
Word Length	:11	
Order	Markov	Entropy
0th.	0.000670	0.007067
1st.	0.010015	0.066518
2nd.	0.030080	0.152057
3rd.	0.029565	0.150189
4th.	0.028709	0.147057

Word	:sağırlık	
Pattern	:CVCVCCVC	
Word Length	:8	
Order	Markov	Entropy
0th.	0.004793	0.036926
1st.	0.051705	0.220965
2nd.	0.082596	0.297163
3rd.	0.082213	0.296336
4th.	0.085891	0.304171

Word	:sağiç	
Pattern	:CVCVC	
Word Length	:5	
Order	Markov	Entropy
0th.	0.034263	0.166765
1st.	0.266939	0.508630
2nd.	0.226798	0.485466
3rd.	0.228616	0.486724
4th.	0.228600	0.486713

Word :sağlam
Pattern :CVCCVC
Word Length :6
Order Markov Entropy
0th. 0.019565 0.111041
1st. 0.075623 0.281698
2nd. 0.141921 0.399769
3rd. 0.140140 0.397305
4th. 0.140158 0.397331

Word :sağlama
Pattern :CVCCVCV
Word Length :7
Order Markov Entropy
0th. 0.008393 0.057883
1st. 0.054199 0.227939
2nd. 0.086740 0.305946
3rd. 0.086131 0.304673
4th. 0.089959 0.312571

Word :sağlamak
Pattern :CVCCVCVC
Word Length :8
Order Markov Entropy
0th. 0.004793 0.036926
1st. 0.051705 0.220965
2nd. 0.082596 0.297163
3rd. 0.082213 0.296336
4th. 0.085861 0.304108

Word :sağlamlamak
Pattern :CVCCVCCVCVC
Word Length :11

Order	Markov	Entropy
0th.	0.000670	0.007067
1st.	0.010015	0.066518
2nd.	0.030080	0.152057
3rd.	0.029565	0.150189
4th.	0.028699	0.147020

Word :sağlamlaşmak
Pattern :CVCCVCCVCCVC
Word Length :12

Order	Markov	Entropy
0th.	0.000383	0.004345
1st.	0.002837	0.024007
2nd.	0.018823	0.107881
3rd.	0.018123	0.104860
4^{th}.	0.015659	0.093903