# Methods of Cryptanalysis
# Lecture 3

Lecture Notes for a course given by Alex Biryukov
at the Weizmann Institute of Science, Israel.
Notes taken by Ilya Safro

June 1, 2000

# Lecture 3

# Methods of Cryptanalysis of Simple Ciphers

In the previous lecture we have described several popular classical encryption methods. This lecture is devoted to the analysis of these classical methods. Note that many ideas described in this section can be very useful for understanding attacks on modern ciphers as well.

## 3.1 Letter Frequencies in Cryptograms

One of the first approaches to solution of simple cryptograms is an analysis of letter frequencies. Knowing the distribution of letter frequencies and statistical properties of the language we can determine the type of the cipher that system uses and it might be very helpfull for the breaking process. In the Figure 3.1 you can see the diagram with regular English letter frequencies calculated from a sample book of size 100000 letters.

Now let us take a plaintext "Aerial reconnaissance reports enemy reinforcements estimated at battalion strength entering your sector PD Clarke" from [10] and look at the pictures of frequencies after encoding it with different ciphers :

- Figure 3.2. This is a diagram for letter frequencies of plaintext. This picture is quite similar to the Figure 3.1.
- Figure 3.3. The diagram of the transposition must be diagram with the same columns of diagram of frequencies of the plaintext. The ciphertext is

```
ANRMEMTNOENEYMAAGGRRAPRETLTY
PIIOENEIHODASRITDOEUCLSTNSAN
NRLRASFETSTSAENEOSBTEERCCNRT
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | B | C | D | E | F | G | H | I |
| 6.9 | 0.9 | 4 | 4.2 | 13.1 | 2.7 | 2 | 3 | 7.9 |
| J | K | L | M | N | O | P | Q | R |
| 0.8 | 0.9 | 3.9 | 3 | 8 | 8 | 2.2 | 1 | 8.2 |
| S | T | U | V | W | X | Y | Z | |
| 6.8 | 9 | 2.8 | 2.1 | 2.1 | 1 | 2.5 | 0.8 | |

Table 3.1: Table of frequencies of English letters

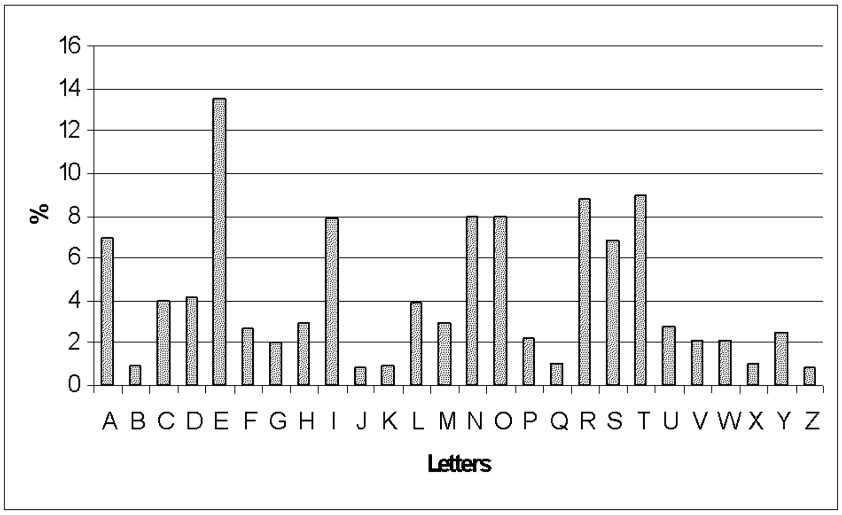


Figure 3.1: Diagram with English letters frequencies calculated from usual text.

```
ARRCKOEECITEITE
```

- Figure 3.4. The diagram of simple substitution in our case Caesar cipher. Each letter has one ciphertext equivalent. So the diagram of ciphertext frequencies is just shift right of the plaintext diagram. The ciphertext is

```
DHULDOUHFRQQDLVVDQFHUHSRUWVH
QHPBUHLQIRUFHPHQWVHVWLPDWHGD
WEDWWDOLRQVWUHQJWKHQWHULQJBR
XUVHFWRUSGFODUNHAA
```

- Figure 3.5. Polyalphabetic substitution (in our - case Vigenere) with the keyword "APRIL" makes the diagram more flat and downs the peaks. The ciphertext is

```
ATIQLLGVKZNCRQDSPEKPRTGWCTHV
VPMNIMTNUFZNEBVVESTJBTMPKMOA
ISIETPCQZNHKZPNVKPPNIVZTNVPW
FRHVKEOGGLNLPISP
```

And now let us introduce several basic analysis techniques that are based on statistical information leakage from the ciphertext.

## 3.2 Kasiski test

This method of cryptanalysis of a periodic Vigenere cipher (also called Caroll's Vigenere, see previous lecture for its description) was introduced in 1863 by Prussian officer Friedrich Kasiski.

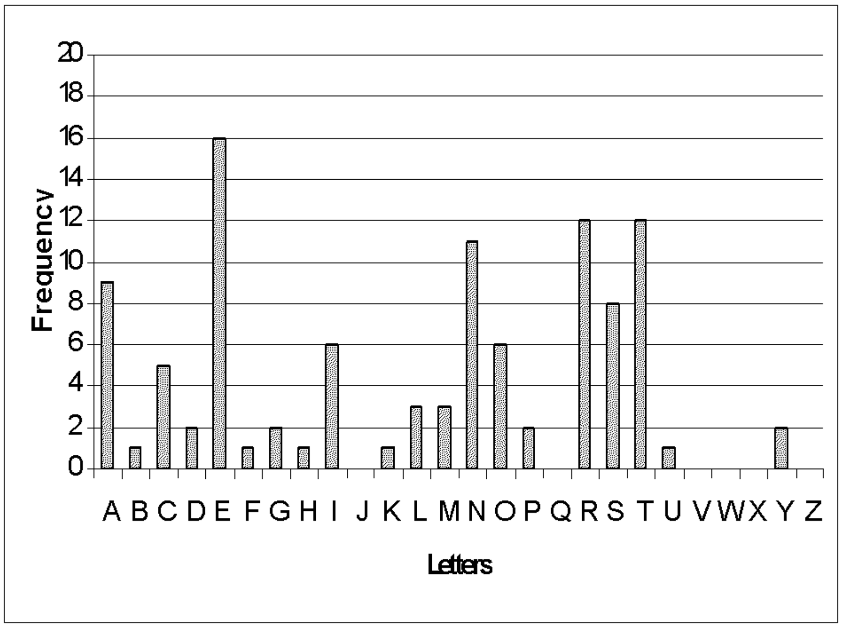


Figure 3.2: Diagram of frequencies of the plaintext

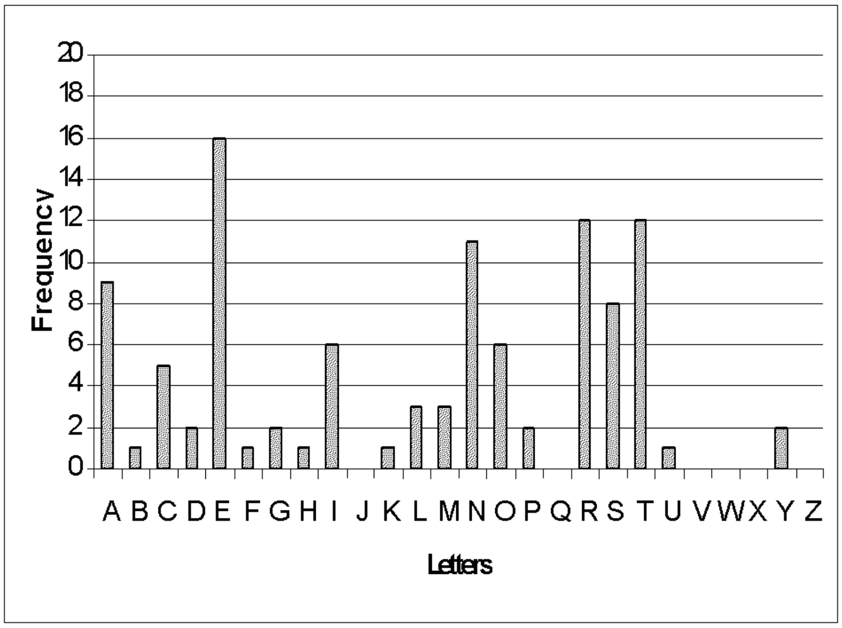


Figure 3.3: Transposition

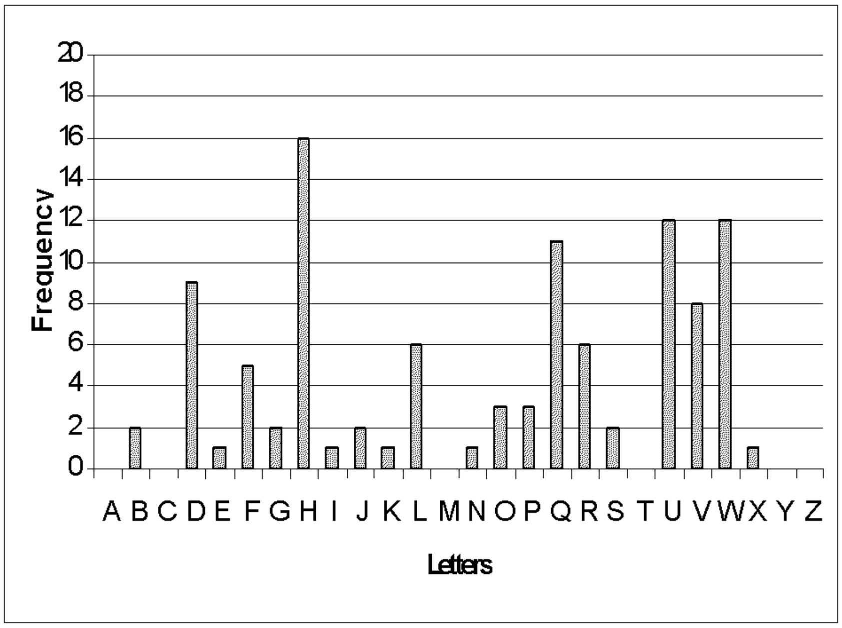


Figure 3.4: Simple substitution

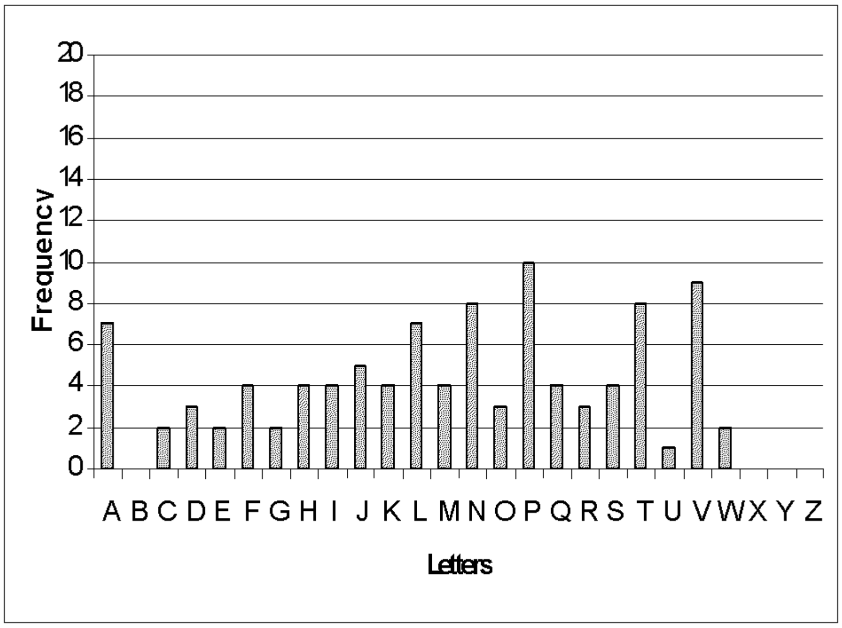


Figure 3.5: Polyalphabetic substitution (Vigenere)

Based on the following observation: "repeated portions of the plaintext encrypted with the same portion of the keyword result in identical ciphertext segments", this method represents a general case technique for analysis of periodic polyalphabetic ciphers. Kasiski's technique allows to find the length of the keyword and is based on measuring the distance between repeated n-grams in the ciphertext. Let us look at the following simple example :

```
Position   : 0 1 2 3 4 5 6 7 8 9 0 1 2
Plaintext  : t o b e o r n o t t o b e
Key        : m a n m a n m a n m a n m
Ciphertext : f o o q o i z o g f o o q
```

Notice, that the ciphertext has two occurrences of the letters "fooq": in positions 0 and 9 respectively. Consequently one expects the number of characters between the beginning of repeated ciphertext segments to be a multiple of keyword length. In our example the distance between repeats is 9, so this suggests a keyword of 3 or 9 letters. Ideally, it suffices to compute the greatest common divisor of the various distances between such repeated occurrences in the ciphertext. In a short text repetitions of more than 3-4 cipher letters are unlikely to occur by chance. If $m$ ciphertext repetitions are found at intervals $I_j$ $(1 \leq j \leq m)$, the period is likely to be some number that divides most of the $m$ intervals. Having a good guess for the length of the period we can check the hypothesis, that letters displaced in arithmetic progression with the step equal to the period $p$ are encrypted with the same key letter and thus represent a monoalphabetic substitution. In the next section we describe a powerful monoalphabeticity test.

## 3.3 Index of Coincidence

The Kasiski test is just a method which helps us to assume the length of the keyword. However, we must know how to "grade" the calculated assumption and how to test it. The *index of coincidence*, proposed by Friedman in 1920, is a measure of the relative frequency of letters in the ciphertext sample, which facilitates cryptanalysis of polyalphabetic ciphers by allowing determination of the period.

**Definition 3.3.1** *Let $X = x_1x_2...x_n$ be a string of symbols. The index of coincidence of $X$: $IC(X)$ is a probability that two random elements of $X$ are equal. Suppose we denote the frequencies of $A, B, ..., Z$ in $X$ by $f_0,\ f_1, ..., f_{25}$ respectively.*

$$IC(X) = \frac{Number\ of\ coincidences\ in\ X}{Number\ of\ pairs\ in\ X} = \frac{\sum_{i=0}^{25} \frac{f_i(f_i-1)}{2}}{\frac{n(n-1)}{2}} =$$

$$\frac{\sum_{i=0}^{25} f_i(f_i-1)}{n(n-1)} = \frac{\sum_{i=0}^{25} f_i^2\ -\ n}{n^2(1-\frac{1}{n})} =$$

$$\frac{\sum_{i=0}^{25} (\frac{f_i}{n})^2\ -\ \frac{1}{n}}{1-\frac{1}{n}} \approx \sum_{i=0}^{25} \left(\frac{f_i}{n}\right)^2 \xrightarrow{n \to 0} \sum_{i=0}^{25} p_i^2$$

*where $p_i$ is the probability of i-th letter.*

Using the index of coincidence we can try to detect the language of plaintext encrypted using such substitution ciphers like Vigenere and Beaufort by calculating $IC$ of parts of the ciphertext. ICs for several languages are presented in Table 3.2.

At the same time the $IC$ in random text is $\frac{1}{26} = 0.0385$ and as we see there is a solid gap between real language and random text. Note that if you solve a short text (100-200 letters) $IC$

| Language | IC |
|---|---|
| Arabic | 0.07589 |
| Danish | 0.07073 |
| Dutch | 0.07980 |
| English | 0.06689 |
| Finnish | 0.07379 |
| French | 0.07460 |
| German | 0.07666 |
| Hebrew | 0.07684 |
| Italian | 0.07380 |
| Spanish | 0.07750 |
| Russian | 0.05290 (has 30 letters) |

Table 3.2: IC of different languages.

| Number of alphabets | IC |
|---|---|
| 1 | 0.0667 |
| 2 | 0.0520 |
| 5 | 0.0440 |
| 10 | 0.0410 |
| more than 10 | 0.0380 |

Table 3.3: Number of alphabets vs. ICs.

should be calculated using $\frac{\sum_{i=0}^{25} f_i(f_i-1)}{n(n-1)}$ and not using approximate formula since the error is of the order of $\frac{1}{n}$. If the ciphersystem uses more than one alphabet the index of coincidence is going to be less and less usable. Look at the Table 3.3 for a demonstration of this.

Let us look at the example in which the attack on the Vegenere uses calculation of the $IC$. Let the ciphertext be (this example is taken from [2]) :

```
TSMVM MPPCW CZUGX HPECP RFAUE IOBQW PPIMS FXIPC TSQPK
SZNUL OPACR DDPKT SLVFW ELTKR GHIZS FNIDF ARMUE NOSKR
GDIPH WSGVL EDMCM SMWKP IYOJS TLVFA HPBJI RAQIW HLDGA
IYOU
```

Using Kasiski's method we examine the repeating trigraphs, and note the two occurrences of IYO and LVF. The IYO's are 25 letters apart and the LVF's are 55 letters apart. $gcd(25, 55) = 5$. The frequency count gives us the following data:

```
A - 7     B - 6     C - 6     D - 6
E - 5     F - 6     G - 5     H - 5
I - 10    J - 2     K - 5     L - 6
M - 8     N - 3     O - 5     P - 13
Q - 3     R - 6     S - 10    T - 5
U - 5     V - 4     W - 6     X - 2
Y - 2     Z - 3
```

Thus the $IC = 0.04066$. So it seems that at least 4 alphabets were used. So be 5 our guess about the length of keyword. Now we will split the ciphertext into five sections:

```
TMCHRIPFTSODSEGFANGWESITHRHI   from text pos. 5i i=0,1,...27.
SPZPFOPXSZPDLLHNRODSDMYLPALY   from text pos. 5i+1 i=0,1,...27.
MPUEABIIQNAPVTIIMSIGMWOVBQDO   from text pos. 5i+2 i=0,1,...27.
VCGCUQMPPUCKFKZDUKPVCKJFJIGU   from text pos. 5i+3 i=0,1,...27.
MWXPEWSCKLRTWRSFERHLMPSAIWA    from text pos. 5i+4 i=0,1,...27.
```

Let us calculate $IC$'s for all rows: $IC_{5i} = 0.04233$, $IC_{5i+1} = 0.06614$, $IC_{5i+2} = 0.05026$, $IC_{5i+3} = 0.06614$, $IC_{5i+3} = 0.04843$. As we see the most close to English $IC$s are the second and the fourth. The most frequent letter in the second row is P so we may assume "P"→"E" and to convert the entire second row into

```
HEOEUDEMHOESAAWCGDSHSBNAEPAN
```

using the same shift $mod26$ as "P"→"E" for all letters from this row. Replacing "U" by "A" in the fourth row we get:

```
BIMIAWSVVAIQLQFJAQVBIQPLFOMA
```

Now we got the following picture :

```
............................
HEOEUDEMHOESAAWCGDSHSBNAEPAN
............................
BIMIAWSVVAIQLQFJAQVBIQPLFOMA
............................
```

Usual assumption about most frequent word in English "THE" allows us to try possible positions for "T" and "E" in this structure and the first substitution gives us the plain text :

```
THE TIME HAS COME THE WALRUS SAID TO SPEAK OF MANY THINGS OF
SHOES AND SHIPS AND SEALING WAX OF CABBAGES AND KINGS AND
WHY THE SEA IS BOILING HOT AND WHETHER PIGS HAVE WINGS.
```

## 3.4 Mutual Index of Coincidence

**Definition 3.4.1** *The* **mutual index of coincidence** $MIC(x,y)$ *of two strings $x$ and $y$ is the probability that a letter in $x$ matches a letter in $y$. Let $f_0, ..., f_{25}$ be the letter frequencies in $x$ and $f_0^{'}, ..., f_{25}^{'}$ be the letter frequencies in $y$. (Obviously, this idea extends to alphabet of any size for which frequencies have been computed for the strings $x$ and $y$). Let $n$ and $n'$ be the lengths of strings $x$ and $y$, respectively.*

$$MIC(x,y) = \frac{\sum_{i=0}^{25} f_i f_i^{'}}{nn'}$$

The mutual index of coincidence $MIC(x,y)$ can be applied to the analysis of periodic alphabetic cipher assuming that we have found the key size $m$. Denote by $y_i$ $i$-th decimation of the ciphertext with step $m$ (an $i$-th column of text if ciphertext is formatted by the length of period $m$). The probability that a symbol from $y_i$ is the encryption of the $h$-th alphabetic symbol is $p_{h-k_i}$ (indices are taken modulo 26). Thus, the probability that a symbol from $y_i$ and a symbol from $y_j$ are both encryptions of the $h$-th alphabetic symbol is $p_{h-k_i} p_{h-k_j}$. Thus we should have :

$$MIC(y_i, y_j) \approx \sum_{h=0}^{25} p_{h-k_i} p_{h-k_j} = \sum_{h=0}^{25} p_h p_{h+(k_i-k_j)} = f(k_i - k_j).$$

| Relative Shift | Expected MIC |
|---|---|
| 0 | 0.065 |
| 1 | 0.039 |
| 2 | 0.032 |
| 3 | 0.034 |
| 4 | 0.044 |
| 5 | 0.033 |
| 6 | 0.036 |
| 7 | 0.039 |
| 8 | 0.034 |
| 9 | 0.034 |
| 10 | 0.038 |
| 11 | 0.045 |
| 12 | 0.039 |
| 13 | 0.043 |

Table 3.4: Expected MICs.

Observe that the value of this estimate depends only on the difference $k_i - k_j \ mod\ 26$, which we call relative shift of $y_i$ and $y_j$. Also notice that

$$\sum_{h=0}^{25} p_h p_{h+l} = \sum_{h=0}^{25} p_h p_{h-l},$$

so a relative shift of $l$ yields the same estimate of $MIC$ as does a relative shift of $26 - l$. In Table 3.4 we show the dependence between relative shifts and expected $MIC$s. The important observation is that, if the relative shift is non zero, these estimates vary between 0.031 and 0.045, which is distingushably lower for $l = 0$. We can use this observation to formulate a likely guess for $l = k_i - k_j$, the relative shift of $y_i$ and $y_j$. Finding all the relative column shifts we are left with 26 variants for the secret keyword.

## 3.5 Probable Word Method

Let us look at a well known Hangman game: a word is writtern on the desk and the players can guess it letter by letter. In our variant of this game, the equal letters are represented by equal symbols, so the pattern of the word is seen by the player:

```
- . = = * - - | . + +   or   A B C C D A A E B F F
```

By looking more carefully at this cryptogram we can see that there are three digrams with equal symbols. The most frequent double letter word-ending is LL, other frequent double letters are TT, OO, EE, SS, CC, MM. Thinking carefully about this word we get the answer: **SUCCESSFULL**.

Sometimes the attacker may be able to guess a part of the plaintext. For instance, one can know the situation in which the message was sent, its senders or the topic of the message. This is called traffic analysis. For example, if it is known that the message is between a retailer and his purchasing agent, then it is highly probable that words relating to the particular market, buying, selling, shipping, commissions, drafts and other financial matters will be included in the text. Let

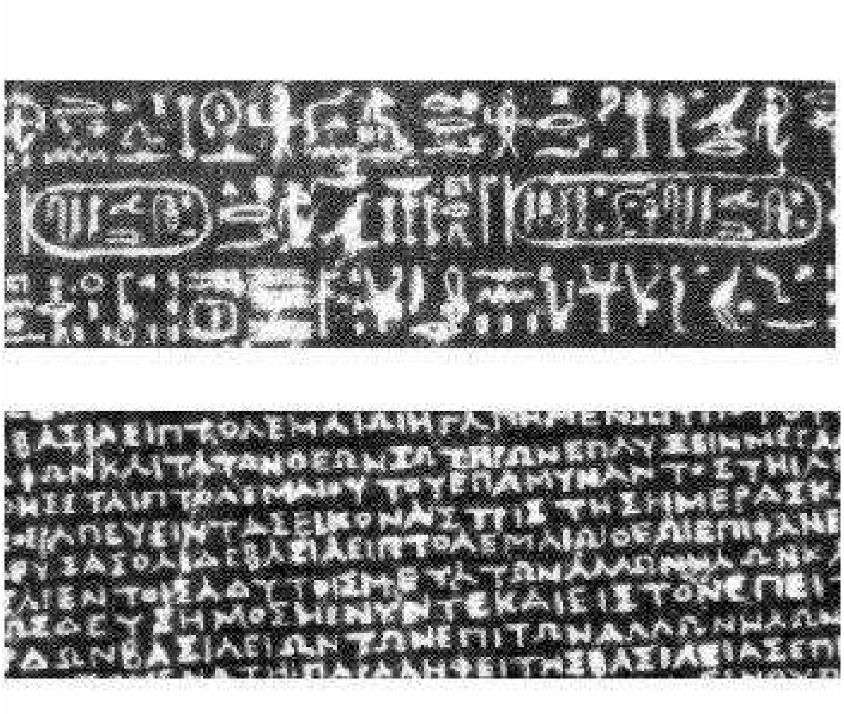

Figure 3.6: Writings on the Rosetta Stone.

us show the two well known examples from the history of Cryptanalysis, that demonstrate the application of the Probable Word Method.

In July of 1798 the French soldiers found in Egypt a block of stone which contained three distinct bands of writing. The top band and most incomplete was written in hieroglyphs, the middle was an Egyptian script called Demotic script, and the bottom was in ancient Greek. At that time the hieroglyphs were a complete mystery but the ancient Greek writing on the stone was still readable.

English physicist, Thomas Young, decided to decipher the hieroglyphics on the Rosetta stone. In hieroglyphics there are groups of symbols that are encircled with ovals (they are called cartouches). Thomas Young decided that the cartouches were proper names of kings and queens. Since Greek equivalent of text related to the ruling dynasty of Ptolemy, cartouches probably stood for Greek names 'Ptolemis' and 'Cleopatra'. Being foreign these names had to be spelled phonetically. Thomas Young correctly identified the cartouch for Ptolemy. In the Figure 3.7 you can see the part of the text on the Rosette stone which was decrypted by this method. Following these ideas French linguist Champollion was able to complete the solution of the hieroglyphics in 1828.

The second example that we would like to mention is the break of the military German ENIGMA cipher. ENIGMA was a generic name for the German machine ciphers used during WWII (next lecture we will introduce this cipher in greater detail). Germans transmitted their secret government and military messages using this machine. The first advances in breaking of this system are due to Polish cryptanalists. The effort to break the machine was reduced greately when French intelligence supplied invaluable known-plaintext information to the Polish. The later British approach also relied on a probable word or phrase also called a crib. This approach was greatly assisted by the property of the Enigma that no letter could be encrypted to itself. Given a long crib, it was

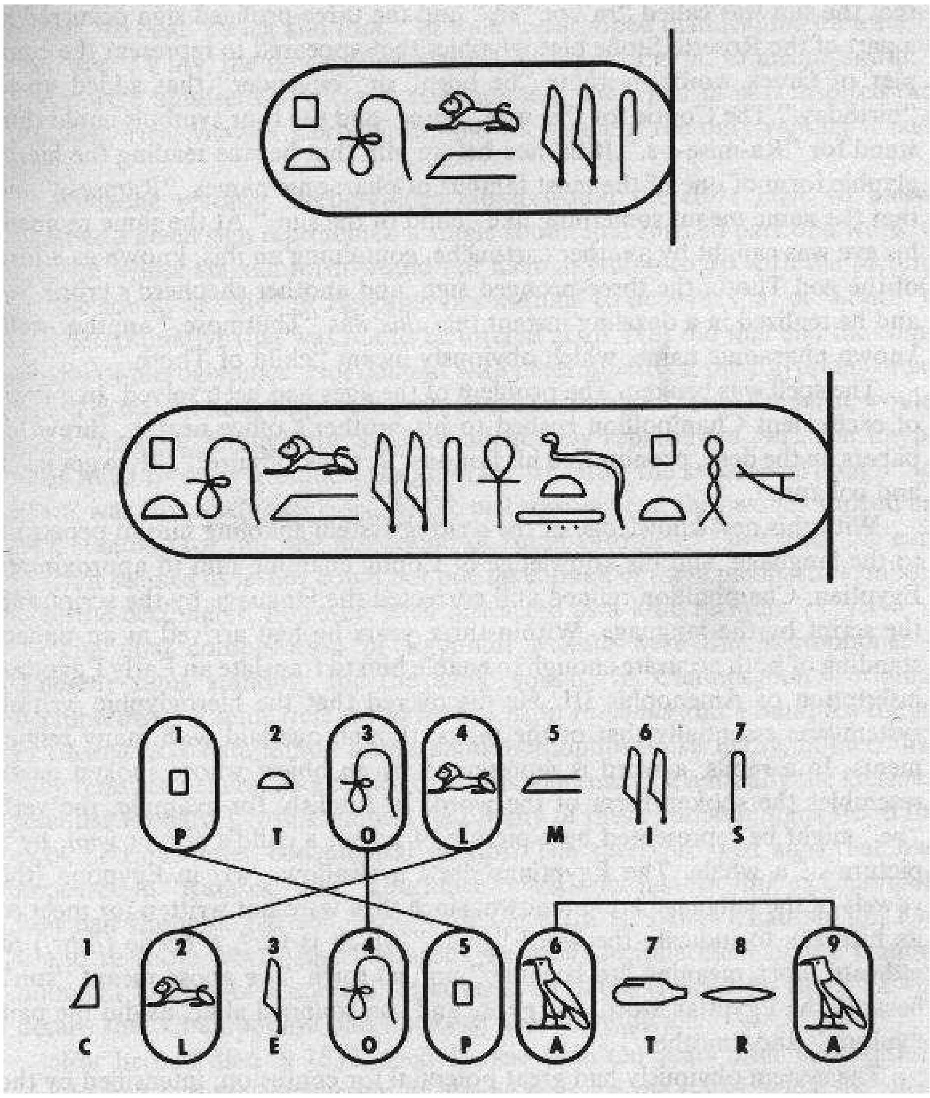


Figure 3.7: Part of the text on the Rosette Stone decrypted by this method.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 52 | 155 | 1351 | 5110 | 9987 | 17477 | 23734 | 29926 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 32380 | 30866 | 26011 | 20460 | 14938 | 9762 | 5924 | 3377 |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 1813 | 842 | 428 | 198 | 82 | 41 | 17 | 5 |

Table 3.5: Distribution of English words.

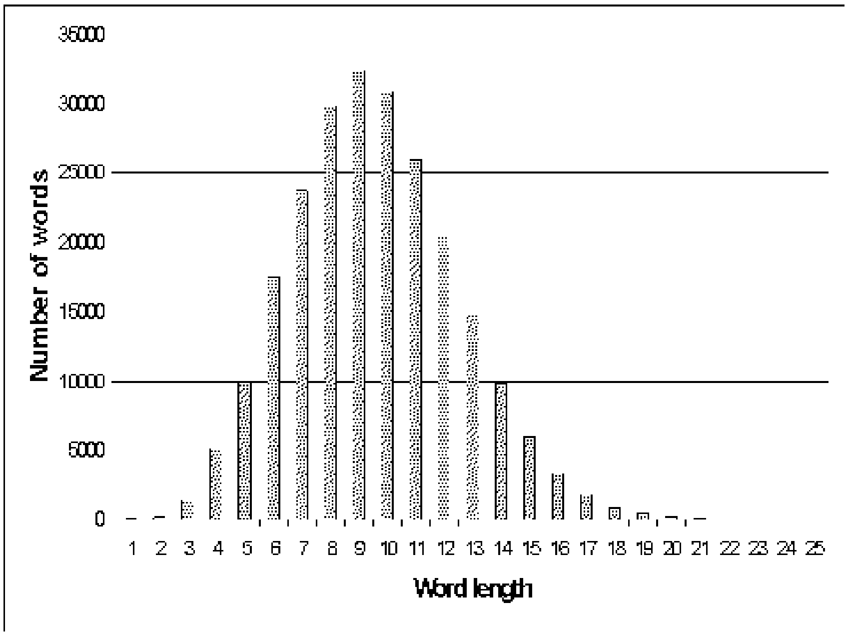


Figure 3.8: Distribution of word length in English.

therefore a simple matter to find the correct alignment of the plaintext guess and the ciphertext by searching for a position in which no letters matched between the two texts.

There are several basic points which would be useful for starting to break an encrypted message with a probable word method:

- **Stereotypes**. If the message was sent by some organization, note that all organizations tend to do things in standard ways. For example: standard formats of the messages, which often contains senders, date, security classifications, signatures, identificators.

- **Exploitation of Numbers**. Spelled out numbers are often easy to recognize when they repeat in messages.

- **Word Patterns**. Repeated words can often be recognized by their patterns of repeated letters, exactly as in the hangman game shown above.

In the Figure 3.8 (see alse Table 3.5) we show the distribution of the lengths of English words, we used a word-list with 230000 words. The dictionaries of word-patterns are also very useful [10].

## 3.6 Vowel - Consonants Relationships

This section is based on material from [10, 3]. What can we do in a situation where there is no information about the topic of the plaintext and we have no idea for a probable word ? In many cases cryptograms are so short that frequency counts are not representative. In these cases, we can use other characteristics of the language. For example those related to the digram structure of the language — letter contacts information. One of the frequently used methods is the vowel-consonant separation approach. There are 6 vowels and 20 consonants in English. In order to make words pronounceable, vowels and consonants tend to alternate. Usually there would not be a situation when more than 2-3 vowels (or consonants) are written consecutively (just look at the passage of text you are reading now). The cases in which such a situation occurred are limited to pronounceable combinations[1]. Vowels are more frequent than consonants (48% of English text are vowels) and tend to occur next to them and vice versa. Since there are more consonants than vowels in English the consonants tend to contact with smaller number of different letters than vowels. We can construct a diagram of contacts of the letters in the same way that we constructed it for letter frequencies. The *Trilateral Frequency Count* is used to record, for each letter in a cryptogram, the letter that preceeds and the letter that follows it. It would be useful to store a frequency and the number of letters that contact the current letter. For example in a Figure 3.9 we can see Trilateral Frequency Count of the following ciphertext :

```
Ciphertext : LBWYR YAGGB GIOYF BATGT BUUBV GKBSK TEEAT
             HBUYA YWYUF QVTWY VJVBA ATUDR TEECY DTUIG
             XYVBS TWYKN UQVYQ FQFVY FIVIG BVPST VYART
             EEAGB FIGXY VBSBN VSTWY UTUYX
```

This information can help us to detect the vowel and consonant letters, by the following rule that higher frequency plaintext letter probably would be a vowel. In cases when there are several letters with the same frequency it would be better to start with the letters with lowest frequency to determine the consonants first. For example in our ciphertext, we will pick the low frequency consonants as the starting point. The lowest frequency letters here are C, H, J, L, O, P. Their left contactors are E, T, VV, I and right side contactors are YY, BB, V and S. At this stage it is very difficult to say something, so let us continue by adding the lowest frequency letters. We first place those with a frequency of one. Next add those with a frequency of two etc. until the next frequency would represent more than 20 percent of the total, see Table 3.6.

The upper line would be the line of probable consonant letters and two colomns are left and right contactors respectively. The consonant line shows us that Y is a more frequent contactor than other letters, so the probability that Y is a vowel is very high. It seems that the letter V is looking also like a vowel, but let us proceed with the contacts of letter Y, see Table 3.7.

Now we see that letter V contacts a vowel readily, thus it is probably a frequent consonant like R, L or N that can be contactor of consonants as well. The letter W contacts Y six times and is likely consonant. We will continue by going back to the consonant line and adding W, see Tables 3.6, 3.6.

The letter T now appears as a strong candidate for a vowel. It is second only to Y in consonant contacts so far, and just as importantly, it does not contact the already selected vowel at all. We add T and its contacts to the vowel line, see Tables 3.6, 3.6.

The vowel line shows A and U as likely consonants. Constructing the whole environment of conclusions like we have done we can come to the right substitutions and decipher the message. The complete solution of this cryptogram is given in [10].

[1]like 'ou', 'ee', 'oo', 'io', 'ai', 'oi', 'ue', 'ia', 'ua', 'ui', 'you', 'ea'.

| | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency : | 8 | 13 | 1 | 2 | 6 | 8 | 9 | 1 | 5 | 1 | 3 | 1 | 0 | 2 | 1 | 1 | 4 | 3 | 5 | 13 | 9 | 12 | 5 | 3 | 18 | 0 |
| Contacts : | 7 | 12 | 2 | 4 | 4 | 6 | 8 | 2 | 5 | 1 | 6 | 1 | 0 | 4 | 2 | 2 | 4 | 4 | 5 | 12 | 9 | 11 | 3 | 2 | 12 | 0 |
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
| | YG | LW | EY | UR | TE | YB | AG | TB | GO | VV | GB | ~B | | KU | IY | VS | FV | YY | BK | AG | BU | BG | BY | GY | WR | |
| | BT | GG | | YT | EA | UQ | GB | | UG | | ST | | | BV | | | UV | DT | BT | GB | UB | QT | YY | GY | RA | |
| | ET | FA | | | TE | QQ | BI | | FV | | YN | | | | | | YF | AT | PT | KE | BY | YJ | TY | Y~ | OF | |
| | YY | TU | | | EC | QV | TT | | VG | | | | | | | | FF | | BB | AH | YF | JB | TY | | UA | |
| | BA | LV | | | TE | YI | VK | | FG | | | | | | | | | | VT | VW | TD | YB | TY | | AW | |
| | AT | KS | | | EA | BI | IX | | | | | | | | | | | | | AU | TI | QY | | | WU | |
| | YR | HU | | | | | IB | | | | | | | | | | | | | RE | NQ | FY | | | WV | |
| | EG | VA | | | | | AB | | | | | | | | | | | | | DU | YT | II | | | CD | |
| | | VS | | | | | IX | | | | | | | | | | | | | SW | TY | BP | | | XV | |
| | | GV | | | | | | | | | | | | | | | | | | SV | | TY | | | WK | |
| | | GF | | | | | | | | | | | | | | | | | | RE | | YB | | | VQ | |
| | | VS | | | | | | | | | | | | | | | | | | SW | | NS | | | VF | |
| | | SN | | | | | | | | | | | | | | | | | | UU | | | | | VA | |
| | | | | | | | | | | | | | | | | | | | | | | | | | XV | |
| | | | | | | | | | | | | | | | | | | | | | | | | | WU | |
| | | | | | | | | | | | | | | | | | | | | | | | | | UK | |

Figure 3.9: Triliteral frequency distribution

| DNKRXQCHJLOP | |
|---|---|
| E | |
| YYYYY | YYYYY |
| T | TTTT |
| B | BBB |
| VV | VVVV |
| I | |
| S | S |
| UU | U |
| | R |
| K | |
| GGG | |
| | |
| D | N |
| A | |
| FF | FF |

Table 3.6: Contactors table: stage - 2

| Y | |
|---|---|
| WWWWWW | W |
| R | R |
| A | AAA |
| O | FF |
| UU | UU |
| VVV | VVV |
| C | D |
| XX | X |
| | K |
| | Q |

Table 3.7: Contactors table: stage - 3

| W | |
|---|---|
| CHJLOP | DNKRXG |
| E | R |
| YYYYYY | YYYYYYYYY |
| TTTT | TTTT |
| BB | BBB |
| VV | VVVV |
| I | |
| S | S |
| UU | U |
| | R |
| K | |
| GGG | |
| | N |
| D | |
| A | |
| FF | FF |

Table 3.8: Contactors table: stage - 4

| Y | |
|---|---|
| WWWWWW | W |
| R | R |
| A | AAA |
| O | |
| | FF |
| UU | UU |
| VVV | VVV |
| C | |
| | D |
| XX | X |
| | K |
| | Q |

Table 3.9: Contactors table: stage - 5

| W | |
|---|---|
| CHJLOP | DNKRXG |
| E | |
| YYYYYY | YYYYYYYYY |
| TTTT | TTTT |
| BB | BBB |
| VV | VVVV |
| I | |
| S | S |
| UU | U |
| | R |
| K | |
| GGG | |
| | N |
| D | |
| A | |
| FF | FF |

Table 3.10: Contactors table: stage - 6

| YT | |
|---|---|
| WWWWWW | WWWW |
| RRR | R |
| AAAA | AAA |
| O | |
| | FF |
| UUU | UUUUU |
| VVVV | VVVV |
| C | |
| D | D |
| XX | X |
| K | K |
| | Q |
| G | G |
| | B |
| | EEE |
| | H |
| SSS | |

Table 3.11: Contactors table: stage - 7

## 3.7 Digrams analysis

In this section we will introduce a method of breaking Playfair cipher. The example is taken from [8].

```
''This message was received by an intercept station in
Scotland. The frequency and format indicate that it is a
most urgent message from one of our agents who landed a
week ago in Norway. His controllers have been unable to
read it. Although it clearly uses his backup cipher, the
Playfair, the keys assigned to him do not work. We cannot
reach him before his normal scheduled transmission in two
weeks, so we urgently request that you attempt to decrypt
this and let us know the contents. In case it helps, he is
carrying materials to assist a previously dropped team in
their work regarding the Norsk Hydro facility at Rjukan. His
recognition code should appear in the message: It is
"beware ice weasels." If he is operating under duress, he
should include the phrase "red penguin frenzy." He will use
"STOP" between sentences and "END" at the end.''

Received message:
VYTES YEDLU TERVL FNVUH DWARD LCFFB SDEWN PXKIC
FTREO LKALZ YLSLT OBKEV LYARM KRBOD NALDY PLAET
OLQAD FHSFZ WNAID SMURU OLHRY LLOTW FYLDI CVLUS
VSSFZ YLUNF FXLKT GBCDO BFALE WRPFY WLHUL DARLI
TFLAB FFZCY FUUFB GXXXX
```

Now let us analyze the situation. We know from the received text that each sentence ends with STOP. One can see that no sequence of two digrams was repeated in the ciphertext, so the division of the word STOP was done like -S and P-. We look at frequent digrams and see that OL, AR, and LD all appeared three times. Now let us look at the gaps between sequences to see if any gave correct room for the code "Red penguin frenzy" or the second phrase "Beware ice weasels" (both are 16 letters) and the final P of the first stop and the initial S of the last stop, i.e., 18 digrams. Only OL fits this condition. If the phrase "B EW AR EI CE WE AS EL S" have been used we should spot a reversed digram in the ciphertext. Since it is not present — the first phrase "red penguin frenzy" must have been used. Now if we will start to construct the square using simple guesses about the words in plaintext we will get the answer : "most urgent stop all members of glider team killed stop in contact with norsk hydro informant stop red penguin frenzy stop do not send followup team until I give coordinates and time for safe landing zone end" and the cipher square is based on a key phrase "final victory" transcribed in a spiral form. It looks like this:

```
f i n a l

h k m p v

s x z q c

e w u s t

d b y r o
```

based on a key phrase "final victory" transcribed in a spiral form.

## 3.8 Anagramming

Cryptanalyst can easily determine whether a ciphersystem is a transposition, just by matching the ciphertext letter frequencies to the frequencies of letters in English. Transposition ciphers can be solved by **anagramming** — the process of restoring a disarranged set of letters into their original positions. This could be achieved with the help of tables of frequencies of di- and tri- gramms. Note some useful language properties of English :

- The vowel-consonant pairs are the most common — no high frequency pair has two vowels.
- Letters that occur in many different pairs are probably vowels.
- Consonants except for N and T, occur most frequently with vowels.
- If XY and YX both occur, one letter is likely to be a vowel.

Assume that periodic transposition of period $p$ was applied to the text, or similarly, that several messages of equal length $p$ were transposed by the same key. Then we place strings of $p$ letters below each other and cut the columns of the resulting matrix. Correct order of columns will result in a correct text emerging in all the rows simultaneously.

Let us look, for example, at the grille from 3.12 the homework 1. One of the ways of solving such grille is by anagramming with the $180^o$ rotation of the same grille. Assume that the text of the grille begins with "THE", we see that as a result the grille must have the following holes (4, 1), (2, 1), (2, 3) or (2, 6). These holes when the grille is rotated by $180^o$ give us "HEV" or "HAV", and considering the next turn of the grille we receive "HEVE" or "HAVE". Picking the more probable "HAVE" we notice that most probably it should be "I HAVE", "WE HAVE", "THEY HAVE". Indeed there is a close "I" in position (4, 5), reading in reverse we have "THE L" and we know 4 grill holes out of 9. See the complete solution of this grille in [9].

## 3.9 Improbable Word Method

This method bears some similarity to the Probable Word Method but here instead of looking for probable words we try to filter out the *improbable* words till only few possibilities are left. Consider the following example of progressive polyphonic substitution. In these two sentences we know that X stands for E; Q for T; J for O; Z for A and K for I :

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | S | E | I | T | I | R |
| 2 | H | T | E | I | N | E |
| 3 | C | L | E | S | T | E |
| 4 | A | C | I | E | I | W |
| 5 | H | T | V | E | E | A |
| 6 | E | D | V | E | K | R |

Table 3.12: Grille from Homework 1.

```
Qjx jxxj xkxx qjzq kxqx xxxq jzk qj kj xxxqjqjkxz.

Zxk kx jjj zqx zqkxx qxzkkxz qjkz qjqjjzj kq kzz xjq zj kzk zq zxx!
```

Our task is to decipher these two lines of text. Let us look now at the word "xxxqjqjkxz". If we will generate (for example by some program) a list of words, using the dictionary, that might be substituted by "xxxqjqjkxz" we can detect the illegal words and remove them from this list. Using table of 3.8 we see that there are 32380 9-letter words in a 230000 word list. We can filter the words from this list based on our partial knowledge of the polyphones. For example in the first 3 letters 'xxx' there has to be at least one vowel and it can be only 'e' or 'u' or 'y'. This simple observation leaves us only with 174 candidate words. From these, for example, **reproduced**, and **surrounded** are not valid because the last D implies D→Z and there is no other Z in "xxxqjqjkxz", so there can not be another D in the word. Similarly in **performing** the first R implies R→X while the seconds R implies R→Q. The word **responding** fails because the two N's. So really the only choice that were left with for "xxxqjqjkxz" is **everything** [9]. Using the similar methods of Probable and Improbable Words we receive the final answer :

```
The only five that were left had to do everything.

And if you are still reading this through it was not so bad at all!
```

# Bibliography


[1] A. Menezes, P. van Oorschot, S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1997.

[2] Dorothy Elizabeth Robling Denning, *Cryptography and Data Security*, Addison-Wesley Publishing Company, Inc. 1983.

[3] Helen Gaines, *The Cryptanalysis (a study of ciphers and their solution)*, Dover Publications, INC. New York, 1956.

[4] D. Kahn, *The Codebreakers*, Macmillan, 1967.

[5] Martin Gardner, *Codes, Ciphers and Secret writing*, Dover Publications Inc. 1972.

[6] Man Young Rhee, *Cryptography and Secure Communications*, McGraw-Hill Series on Computer Communications 1994.

[7] Bruce Schneier, *Applied Cryptography*, John Wiley and Sons Inc. 1994.

[8] *Decoding nazi secrets*, `http://www.pbs.org/wgbh/nova/decoding`

[9] *Solutions for Homework 1 from course homepage*,

[10] *Army Field Manual (Cryptanalysis)*,
`http://www.und.nodak.edu/org/crypto/crypto/army.field.manual/`