Cryptoanalysis as a method of the system approach in the algorithm development

MICHAL MUSILEK, STEPAN HUBALOVSKY
Department of Informatics
Faculty of Science
University of Hradec Kralove
Rokitanskeho 62
CZECH REPUBLIC
michal.musilek@uhk.cz, stepan.hubalovsky@uhk.cz

Abstract: The system approach is one of the education methods which can be widely applied in any subjects. The authors have found this method as suitable for training in theoretical knowledge of algorithm development and programming. Cryptoanalysis has been chosen for the presentation of the system approach in this paper. The huge advantage of Cryptoanalysis, which is a part of the curriculum in computer science, is that it enables the building of the system approach as well as interdisciplinary relations between such subjects as mother tongue, foreign languages, mathematics, history and geography. The use of the algorithm development and programming in the cryptoanalysis of the monoalphabetical substitution cipher is specifically presented in the paper.

Key-Words: Algorithms, algorithmic thinking, education, system approach, cryptoanalysis, monoalphabetical substitution cipher

1 Introduction
The algorithm development and programming belong to the main terms used in the computer science. Algorithms are encountered in all practical activities without being realized. An algorithm generally involves defining the rules and giving the sequences of steps necessary for any activities. The most common examples of algorithms in everyday life are various descriptions and giving directions. The ability to create algorithms develops logical thinking skills and imagination and is an inseparable part of study skills of prospective and undergraduate teachers specializing in “Informatics” at the Faculty of Science.

In recent years the structure of teaching of the algorithm development at the University of Hradec Kralove has significantly moved towards the use of a system approach to teaching algorithms and programming, described e.g. in [1], [2], [3], [4]. This new approach is based on certain elimination of a routine rewriting of mathematical equations and formulas using algorithms and practicing the standard algorithm, regardless of their complexity and system integration to the whole entity [5], when the exercises used are “artificial”, inconsistent and disconnected from reality.

On the other hand, the new methodology uses for training of algorithms the exercises which are based on modelling of real and idealized phenomena, processes and procedures. The methodology is intended to lead the student to problem definitions, to creation of suitable models and to implementation and realization of the model through the principles of algorithms, i.e. the student should be able to determine about the input and output data and the procedure of the transformation of the data input into the data output under the system approach. The students should be able to create an application in an appropriate programming language.

Mathematical calculus is in a “model” method necessary, but only as a secondary issue.

The paper will present an application of the system approach in the field of the cryptoanalysis, concretely in the cryptoanalysis of monoalphabetical substitution cipher.

2 Monoalphabetical substitution
The monoalphabetical substitution cipher is a cipher where one-to-one mapping is used to substitute each of the characters of the plaintext by a corresponding character of the ciphertext. A plaintext is a message before encryption. A plaintext is a normal text, e.g. in English, presented in a readable and understandable form. A ciphertext is an unreadable message after encryption. A ciphertext looks like a random jumble of letters or other characters. One way how to obtain a ciphertext from a plaintext is a character by character substitution realised according to the so-called conversion matrix. The conversion from a ciphertext back into a plaintext is called decryption. The decryption is realised by using the same conversion table in the opposite direction.
2.1 Principles of monoalphabetical substitution

A concrete example of a historic cipher is the substitution used in the Old Testament, where re-writers of the Bible wanted to leave their mark in the text. Sesah appeared instead of Babel. This encryption emerged from the substitution of the letters from the beginning of the Hebrew alphabet by letters from its end, namely the aleph by tav, bet by shin and kaf by lamed. Since the Hebrew text is often recorded only in consonants and vowels are added from the context, this cipher is known as the Atbash [6], [7].

Applying the same principle to the Roman alphabet (the international alphabet with 26 letters without diacritics is used), the following Conversion table (Table 1) can be presented:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>Y</td>
<td>X</td>
<td>W</td>
<td>V</td>
<td>U</td>
<td>T</td>
<td>S</td>
<td>R</td>
<td>Q</td>
<td>P</td>
<td>O</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 1 Conversion table

The Atbash type system has another advantage - the same conversion table can be used for both decryption and encryption. An additional advantage is that the table can be reduced by a half of the original conversion table.

A Conversion table of monoalphabetic substitutions, however, may be more general. For example, mixing the letters of the cipher alphabet according to the password can be used; as in the following example (Table 2) where the password is based on the name of our university, from which repeated letter have been removed:

UNIVERSITY OF HADC KL (University of Hradec Králové)

Table 2 Conversion table with the password mixing letters

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>N</td>
<td>I</td>
<td>V</td>
<td>E</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>Y</td>
<td>O</td>
<td>F</td>
<td>H</td>
<td>A</td>
</tr>
</tbody>
</table>

The question is how many monoalphabetic ciphers can be created. Each cipher can be described by the table, which has in the first row the characters of the alphabet sorted in the usual alphabetical order (A, B, C, ..., Z) and in the second row there are the letters arranged in any random order. Number of different orders, i.e. permutations of the 26 letters of the alphabet is 26! = 26.25.24. ... 3.2. 1 =

\[\frac{1}{2} \times 26! = 403291461126605635584000000\]

The fact that each single substitution is relatively easy to solve, was probably discovered for the first time by Arab scholar Al-Kindi in the 9th century AD.

Al-Kindi described, in two brief paragraphs of the manuscript, the principle of the frequency analysis, which is one of the most important tools of the classical cryptoanalysis. The origin of the frequency analysis is likely to have been influenced by studying of the Qur’an, which was so thorough that it not only examined the frequency of individual words, but even the frequency of individual letters. In Arabic, the letters a and l appear very often, while the frequency of j is about ten times lower. Frequency, or percentage distribution of individual letters, however, is different in each language. The Table 3 shows the percentage distribution for English [8]:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2</td>
<td>1.5</td>
<td>2.8</td>
<td>4.3</td>
<td>12.7</td>
<td>2.2</td>
<td>2.0</td>
<td>6.1</td>
<td>7.0</td>
<td>0.2</td>
<td>0.8</td>
<td>4.0</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 3 Percentage distribution of letters in a purely English text

The letters E, T, A, O and I are the most frequent letters in English. The longer a ciphertext is and the purer English is used in the text (without a mixture of foreign words), the more helpful the frequency analysis is. The procedure of the decryption will be described in the following text of the paper.

2.2 Principles of deciphering of the monoalphabetical substitution

First, the frequency analysis of the ciphertext has to be performed. If the text is electronically available, it is possible to use the tool on the Web [4]. Letters with the highest frequency are probably E, T, A, O and I - vowels. If the ciphertext is shorter, it is probable that the sequence of letters above will be replaced. Therefore other linguistic facts, knowledge and skills have to be included in the decryption analysis:

- Consonants and vowels are in the plaintext fairly regularly alternated.
- Doubling of the letters is typical of certain sounds.
- The frequency of pairs of consecutive letters, which are called bigrams.
- The ability to estimate certain words of the ciphertext, especially when it is possible because of being aware of the topic of the ciphertext.
Note: If all characters of the ciphertext have a similar frequency without significant maximums and minimums, it would probably be a polyalphabetic cipher. In these cases the solution has to be provided by other methods, which exceed the scope of this paper.

2.3 An example of hand deciphering of monoalphabetic substitution
The process of deciphering of the ciphertext encrypted by the monoalphabetic substitution will be shown in the following example:

An English ciphertext:

```
UOMEO MUWKP UMEPX FEMPJ
WLXWE AOYKP YRMJU MEPFP
EMUXF JBUOM JMPJW LXFPP
BWNFI WPFPF FKMXS WERKW
LFSSM MYDFG JPMJK FEUWW
JMEKP JMMPY EUWKO YERPFE
EDWJB MPDFK PFSPO MFPOM
JIQKY EMKMM KFSPO WPWJMM
WUMJMJ GJFAG LMOFQ KMKKF
POWPI MKYAM KOWTY ERJSM
MLFSS MMPPA JYEYBE EPOMP
WKYE LEAGW JDPEO PPJWL
XOWAS JMMBSJ QYPWE ATMRM
PWICM KPFMW P
```

A performed frequency analysis gives the following results:

- the message has 271 characters;
- the most frequent letter is M, its occurrence is 41, the relative frequency of M is 15.1%;
- the characters in the following order are P – 12.5 %, W - 8.5% and F - 8.1%.

Based on the above mentioned, the characters can be replaced by using the percentage distribution (Table 3) by the characters E, T, A and O. The remained characters will be replaced by the dash yet.

Moreover, the following information is to be given:

- the group of six letters LFSSMM occurs twice in the ciphertext;
- other information from another source is that the word COFFEE appears in the message;

Based on this information, we will try to replace the characters E, T, A and O. The remaining characters will be replaced by the dash yet.

We can see that the alternation of vowels and consonants is correct, so we can try to identify other vowels’ characters. They will be marked by an asterisk and a cross:

```
--E-- E-A-T -E-T+ O-E-TR
AC+A- --*-*T -*ER- E-TTO
-E--O R--E RETRA C+TOO
-A-O- ATOTT O-E+F A---A
COFFE E**--O RTER- O--AR
RE--T REET* --A-- *--TO
--AR- ET-O- TOFT- EOT-E
R---* E-E-- -OFT- ATARE
A-EYE RO-- CE-O- -E--O
T-AT- E-*--E --A-- *--FREE
ECOFF EE*--E R***-- T-EET
A-T*-- --E-A RT-E- TTRAC
++A-F REEF* --TA-- --E-E
TA--E --TOEA T
```

We will try to replace the asterisk and the cross by vowels, a reasonable substitution seems to be Y by I and X by Y:

```
--E-- E-A-T -E-TY O-E-TR
ACYA- --I-TE I-ER- E-TTO
-E-YO R--E RE*TRA CYTOO
-A-O- ATOTT O-EYF A---A
COFFE EI--O RTER- O--AR
RE--T REETI --A-- I--TO
--AR- ET-O- TOFT- EOT-E
R---I E-E-- -OFT- ATARE
A-EYE RO-- CE-O- -E--O
T-AT- E-I-IE --A-I --FREE
ECOFF EETO- RI--I --T-EET
A-TI- --E-A RT-E- TTRAC
Y--A-F REEF* --ITA-- --E-E
TA--E --TOEA T
```

An incomplete conversion table is shown in Table 4.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>L</td>
<td>M</td>
<td>S</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>P</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
The name Tracy appears repeatedly in the plaintext, which is the name of the main character Tracy's Tiger book by William Saroyan. Let's try to use the writer's name as a password in such a way that the letters used repeatedly are deleted. We get WILAMSROYN. This ordering is in correspondence with the complete conversion table:

The encrypted plaintext can be got from the ciphertext by using of the above given table:

When he was twenty-one Tracy and his tiger went to New York, where Tracy took a job at Otto Seyfang's, a coffee importer's on Warren Street in Washington Market. Most of the other businesses of that area were produce houses, so that besides having free coffee to drink - in the Tasting Department - Tracy had free fruit and vegetables to eat.

It is a passage from the book Tracy's Tiger by William Saroyan. We can see from the example that the classic hand encryption of the monoalphabetic substitution is not an easy thing. The combination of the mathematical methods (frequency analysis) with the linguistic method (the estimation of the plaintext on the basis of the estimated words) has been used.

3 Automation of the cryptoanalysis

The following part of the paper presents a description of a possibility of using mathematical methods and algorithms for a cryptanalysis of the ciphertext.

The algorithm described in the following text, is a basis of the computer program, which automatically decrypts any ciphertext encrypted by the monoalphabatical substitution cipher. The original form of this algorithm is taken from the paper by Thomas Jacobsen [9]. Unlike this form, our algorithm was applied to a Czech text with no spaces.

The first task is to obtain a suitable reference text of the bigrams rate in the Czech language. The text from the book The Gardener's Year by Karel Capek, was used. It has about 99,000 letters. In the text we ignored the diacritics, punctuations and spaces and for each pair of consecutive letters, we incremented the square matrix element of dimension 26 (a number of international alphabet letters). Finally, the whole matrix is multiplied by 100 and divided by the number of letters of the reference text, so the value of the matrix elements now represents the percentual relative frequency of bigrams.

The developed program enables easy changing of the values of the reference matrix, so it can be used for other languages than Czech.

3.1 Algorithm

The algorithm itself was divided into three relatively independent procedures.

3.1.1 Frequency

- This procedure performs an initial setup for the conversion table. The frequency of each character in the ciphertext is detected and ciphertext characters are sorted in the descending order of frequency. That determined the sum of the rows from the reference matrix $E$ that corresponds to the
frequencies of letters in the reference text; the reference text characters are sorted in the descending order of frequency. Finally, the procedure pairs the first most frequent letter of the ciphertext with the first most frequent letter of the reference text, then the second most frequent letter of the ciphertext with the second most frequent letter of the reference text, and so on.

- The procedure creates matrix $D$ of relative frequencies of the bigrams of the ciphertext and evaluates the compliance with the reference text by using an evaluation function:

$$f = \sum_{i=1}^{26} \sum_{j=1}^{26} |D_{ij} - E_{ij}| \quad (1)$$

### 3.1.2 Bigrams

- Step by step, 2 columns and 2 rows in the matrix $D$ are exchanged and thus the matrix $D'$ is obtained. The columns and rows are exchanged in the order of the frequencies of the letters in the ciphertext from the most frequent to the least frequent:
  - the column corresponding to the order of the first exchanged character is replaced by the column corresponding to the second exchanged character;
  - then the row corresponding to the order of the first exchanged character is replaced by the row corresponding to the second exchanged character;

- The substitution of the columns and rows of the matrix corresponds to the replacement of the letters in the second row of the conversion table. First, the neighbouring letters are replaced, then the first and the third letters are replaced, then the second and the fourth letters are replaced and so, then the first and the fourth letters are replaced, then the second and the fifth letters are replaced and so on, until finally the first letter is exchanged by the last one.

- After each substitution a new matrix $D'$ is obtained, and the evaluation of the compliance of the relative frequency of the bigrams in the ciphertext and the reference text is obtained using the evaluation function:

$$f' = \sum_{i=1}^{26} \sum_{j=1}^{26} |D'_{ij} - E_{ij}| \quad (2)$$

- After each substitution the values $f$ and $f'$ are compared.

If $f' > f$, the procedure continues with the next pair of the letters in order. However, if $f' < f$, the procedure immediately stops the process of the letter substitution and the exchange in the conversion table is proposed, which will improve the compliance (lowering the value of evaluation function $f$). If the program exhausts all the possible exchanges of the characters, without any condition $f' < f$, the message is notified to the operator. If the ciphertext is not fully deciphered, the operator can propose a manual replacement.

### 3.1.3 Exchange

- As not always the procedure "Bigrams" suggests an optimal exchange, especially for short texts (less than 500 characters), it is possible to perform the required exchange manually, to set it directly into the conversion table based on the reading of a partly deciphered text and estimating of the content of messages and expected words.

- Whether the exchange is proposed automatically or manually, the procedure "Exchange" will create a new matrix $D$ of relative frequencies of the bigrams of the ciphertext, and it will provide a new assessment of compliance with the reference text using the evaluation function (2).

From the description of the above procedures it is obvious that the procedure "Frequency" is run only once at the beginning of the program to set the appropriate initial conditions. The procedures "Bigrams" and "Exchange" are run alternately, or instead of run the procedure "Bigrams", a manual exchange of characters in the second row of the conversion table is suggested based on the reading of the text and partly decrypted ciphertext and based on the estimation of the words contained in the plaintext that we get.

That algorithm is very reliable and for sufficiently long texts (with a length of at least 500 characters) it usually operates completely automatically without any need of manual correction. Each exchange in the conversion table has to be run by a button press of the operator. The possibility of the manual intervention in the process
of solving is suitable for the decryption of short messages.

3.2 Computer simulation of the algorithm
The algorithm, described above, has been assembled in JavaScript programming language. Figure 1 shows the user interface of the application. The detailed description of the program as well as the programming code exceeds the scope of this article and can be found in [4].

The application in JavaScript verified the functionality of the presented algorithm.

Fig. 1 User interface of the application for automatic deciphering of monoalphabetical substitution

4 Conclusion
There are various approaches how to provide training in algorithms, how to introduce and develop basic algorithmic thinking of students. The paper offered one of the kinds of the teaching / learning strategy by using the system approach. The system approach can be set as the default paradigm for a wide integration of the principles of the algorithm development into education. The paper emphasizes the fact that the algorithm is neither a mere mathematics, nor a mere programming.

Cryptoanalysis has been chosen for the presentation of the application of the system approach, concretely cryptoanalysis of monoalphabetical substitution.

References:

Acknowledgements
This research has been partially supported by the Specific research project of the Faculty of Science of University of Hradec Kralove No. 2129.