JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
31st December 2018, Volume 3, Issue 3, JSEIS, CADMEI Copyright © 2016-2018
www.jseis.org

A SOFTWARE APPLICATION OF ANCIENT SYLLABARIES TO CRYPTOGRAPHY

EVANGELOS C. PAPAKITSOS

Greek Ministry of Education, Research and Religious Affairs, Athens, Greece
Email: papakitsev@sch.gr

Academic Editor: Sundas Hanif, APCOMS, Pakistan.
Published: 31st December, 2018
Funding: The author has no funds.
Article Type: Research Article

ABSTRACT

In this paper, a novel algorithm of text encryption is presented for applications of cryptography and information security. Most historical and well-known methods of text encryption can be deciphered through the statistical analysis of the symbols' appearance frequencies in text, whenever they follow a normal pattern. The presented herein algorithm is based on a mixed frequency pattern that has been inspired by the Aegean Scripts, developed mainly in Crete more than 40 centuries ago, and by the Natural Language Processing applications for their reading or decipherment. These scripts are classified as syllabaries, because their writing symbols render syllabic phonetic values, consisting mainly of one consonant followed by one vowel. The herein cryptographic algorithm, implemented in Python programming language, performs a partial syllabic substitution (PaSyS) of the text's characters, thus being more resistant to normal statistical decipherment.

Keywords: cryptography; syllabary; Aegean scripts; information security; natural language processing;

1. INTRODUCTION

According to the Dictionary of the Spanish Royal Academy, "cryptography" is the art of writing with the help of a secret "key" or in an enigmatic manner. The Mesopotamians and the Egyptians used cryptographic methods, but firstly the Greeks and the Romans actually applied them, as warfare peoples, since there was a need for secret communications to ensure military success [1]. Creating an encrypted message ("cryptogram") by its sender is called "encryption", while recovering it to the original written form by its receiver is called "decryption". The transmission of a cryptogram follows generally the following steps:

- Original message (from sender).
- Encryption (via algorithm & key).
- Cryptogram (and its transmission).
- Decryption (via algorithm & key).
- Original message (to receiver).

The counteraction to cryptography is called "cryptanalysis", namely how the adversary ("eavesdropper") will read the original message from the cryptogram. It is worth mentioning here that deciphering a text of an unknown ancient writing language is often treated as a problem of cryptanalysis [2].

There are two main methods of encryption: the "transposition" method and the "substitution" method (for an introduction see [3, 4]). In transposition, a letter is changed its position but it retains its phonetic value, namely having the same phonetic value in both the original message and the cryptogram (like the "scytale" transposition that had been used by the Ancient Spartan Army [5]). In substitution, the symbol retains its position but its phonetic value is changed, namely having a different phonetic value in the original message than in the cryptogram. Since the proposed herein cryptographic algorithm belongs to the category of substitution encryption methods, the basic/historical algorithms of substitution cryptography will be presented next, along with their corresponding cryptanalysis techniques ("cryptology").

Corresponding Author: Evangelos C. Papakitsos
Email Address: m.hassansaad@gmail.com

2. SUBSTITUTION CRYPTOLOGY

Obviously, the adequacy of an encryption algorithm is evaluated for its resistance to cryptanalysis that is conducted by an adversary ("deciphering"). A deciphering technique corresponds to each encryption algorithm.

a) Encryption Algorithms

The main historical encryption algorithms of the substitution method are the following seven [1]:

- The alphabetical division ("Mulavediya cipher") is a simple "monoalphabetic" substitution, where the alphabet is divided into two equal parts. Each letter of the first part corresponds to a letters of the second part and vice-versa, in a random manner. Thus, each letter is replaced by its corresponding one in the pair (e.g., in the pair A-N, A is replaced by N and N by A).
- In the "homo-parallel" encryption ("Caesar cipher"), each letter of the alphabet is replaced by another letter in the cryptogram, shifted to a specific number of positions {n} in the original message, which is also the encryption key. Namely, if {n = 3}, then A is replaced by D, B by E, etc.
- In the "Polybius square", each letter of the alphabet is replaced by two other letters (or two numbers) based on the coordinates (row-column labels) of a two-dimensional key-table (e.g., K can be replaced by the BE pair).
- The "keyword encryption" uses a keyword (for example, the word DOUBLE), where the first six letters of the alphabet (from A to F) are replaced by the corresponding letter of the keyword (namely, A by D to F by E). The remaining letters of the alphabet (from G to Z) are replaced by the next letter of the alphabet after the last one of the keyword (namely, G by F) and then following the normal order (H by G, I by H, …), excluding those letters that are already in the keyword (D, O, U, B, L, E) that have been already used to substitute the first six letters (from A to F).
- In the "polyalphabetic" substitution (or "Alberti algorithm"), two alphabets of homo-parallel substitution are alternately used for each letter. Namely, the first occurrence of A in the original text is substituted (e.g.) by D (respectively, B by E, etc.), while its second occurrence is substituted by (e.g.) H (respectively, B by I, etc.). In overall, the odd occurrences of A (1st, 3rd, 5th ...) are substituted by the corresponding letter of the first substitution alphabet, while the even occurrences (2nd, 4th, 6th ...) by the corresponding letter of the second substitution alphabet.
- The "Vigenére square" is another algorithm of polyalphabetic substitution, where more than two substitution alphabets are used alternately, either by the homo-parallel substitution or by the keyword one. In the latter case, there are as many substitution alphabets as the letters of the keyword (e.g., for the keyword DOUBLE six substitution alphabets are used).
- In the "encoded alphabet" algorithm ("homophonic substitution"), a regular alphabet is not used for encryption (as in the monoalphabetic substitution), but a corresponding set of symbols instead that make up its cipher-table in an arbitrary manner.

All the previous substitution algorithms are called simple ones (either mono-alphabetic or polyalphabetic), because a single letter of the original message is substituted by one or more letters of the substitution alphabets (or one symbol of the cipher-table, in the last case). There is also a "polygraphic substitution" encryption, where the letters of the original message are substituted in larger groups, namely of two or more consecutive letters (bigrams, trigrams … n-grams or the "Playfair cipher", devised in 1854 by Charles Wheatstone [6]).

b) Deciphering Techniques

Similarly to the encryption algorithms, corresponding deciphering techniques have been also devised and used. The most basic of them, devised for reading the substitution cryptograms, are the following three [1]:

- The "frequency analysis" technique, which has been described since the 9th century by the Arab mathematician Al-Kindi [7], results from the fact that for every language there is a specific occurrence frequency of each letter in its texts [8]. Thus, the frequency of the cryptogram's letters/symbols is displayed in descending order. The first most frequent character (letter/symbol) of the cryptogram is then mapped to the first most frequent letter of the language. This cryptanalysis continues until all letters are matched and it

is usually extremely effective for cryptograms larger than 100 characters. All the monoalphabetic substitution algorithms are vulnerable to frequency analysis, since the substitution character appears in the cryptogram as many times as in the original message.

- The "maximum common divisor" technique is used to read the homo-parallel cryptograms ("Caesar cipher") for determining the displacement number {n}.
- The Babbage-Krasinski technique [9] is an attempt to discover the length of the keyword ("keyword encryption"), which also determines the number of the substitution alphabets in polyalphabetic encryption ("Vigenére square"). This deciphering technique begins by creating a list of bigrams (two consecutive characters) repeated in the cryptogram. Then, the number of characters interposed from one occurrence of a bigram to the next ("gap") is counted and the integer divisors for each gap are calculated. The length of the keyword is one of these common divisors, for example, number {6}. So, we end up with a number of six substitution alphabets. This technique is effective for every polyalphabetic encryption ("Alberti algorithm" and "Vigenére square").

Thus, for each substitution encryption algorithm, there are one or two corresponding deciphering techniques, as well.

3. PARTIAL SYLLABIC SUBSTITUTION

The presented herein encryption algorithm of substitution is inspired by the Aegean Scripts, which were used originally in Crete during the 2nd millennium BCE [10]. These ancient writing systems were syllabic ones ("syllabaries"), namely, each symbol corresponds to a syllable ("syllabogram"), usually of the consonant-vowel (CV) phonetic pattern, unlike either the consonant (C) or the vowel (V) phonetic depiction, which is usually the case of an alphabet. These syllabaries consist of the Cretan Hieroglyphics, Linear A and Linear B. In addition and on the basis of their similarity to the previous three syllabaries, the two ancient syllabaries of Cyprus, namely the Cypro-Minoan and the Cypriot Syllabary, are included in this category [11]. For all the syllabaries except the latter, no traces of their usage have been found after the 12th century BCE, while the Cypriot Syllabary remained in use until the 3rd century BCE [12].

Specifically, in the Aegean Scripts every common syllabogram (S_i) represents two consecutive phonemes, namely a consonant (C_i) and a vowel (V_i), having the phonetic pattern: $S_i = C_iV_i$. Thus, in an alphabet such as the Latin one, two consecutive characters (C_iV_i) are mapped to a single syllabogram (S_i), as in the following two examples:

- DA = {┣},
- PA = {╪}.

There are plenty of characters in Unicode (or UTF-8) encoding for matching a syllabic character pair of an alphabet with a CV phonetic pattern (e.g., DA, PA) to a single Unicode character. The presented algorithm of Partial Syllabic Substitution (PaSyS) is an experimental application of this idea. PaSyS has been implemented for the Greek alphabet with the Python programming language [10], according to a matching table ("key-grid"). Nevertheless, a Latin alphabet version of the key-grid will be presented herein, which is more common for clarification purposes.

The presented example of key-grid (Table 1) is named so after the original grid that was created for and because of the decipherment of Linear B script [13]. The actual key-grid consists of 7 columns and 21 lines, excluding the corresponding headings of columns and lines (in bold italics), which contain 146 substitution characters (+2 optional characters for the substitution of Space). More specifically:

- Each column corresponds to a vowel in alphabetical order, except the first one that denotes the absence of vowel, namely the substitution character of a single consonant.
- Each line corresponds to a consonant in alphabetical order, except the first one that denotes the absence of consonant, namely the substitution character of a single vowel.
- The very first cell corresponds to the substitution characters of "Space", which should be more than one in a manner of polyalphabetic substitution, for avoiding the easy recognition of word delimiters. Similarly, other punctuation marks should be also substituted. In the following examples, the space character or any punctuation mark will not be substituted for the simplicity of presentation.

Wherever there is a syllable (i.e., a sequence of two characters with a preceding consonant and a following vowel) in the text to be encrypted (e.g., BA), then it is substituted in the cryptogram by their corresponding character in the line of the consonant and the column of the vowel (i.e., "Ẅ"). Single consonants (e.g., B) are substituted by their

corresponding character of the first column (i.e., "ý"), while single vowels (e.g., A) are substituted by their corresponding character of the first line (i.e., "Ẇ"). The decryption procedure follows the reverse order.

4. RESULTS AND DISCUSSION

An example of applying PaSyS is demonstrated for the first sentence of the herein ABSTRACT (in capitalized letters):

- "IN THIS PAPER, A NOVEL ALGORITHM OF TEXT ENCRYPTION IS PRESENTED FOR APPLICATIONS OF CRYPTOGRAPHY AND INFORMATION SECURITY."

The encryption is performed through the previous key-grid (Table 1), resulting in the following cryptogram:

- "ŴÞ ÝÏñ Śćô, Ẇ rÛĤ ẂĤkŐÝŁĦ bR̂ ùĶÝ ŵÞďòĉřbÞ Ŵñ ĉőḣÞùĈ oô ẂĉĉẌḌřbÞñ bR̂ ďòĉyħẀĉľ ẂÞĈ ŴÞoôŘřbÞ ḣÁŐú."

Table. 1 Encryption key-grid

Consonants	Vowels						
		A	**E**	**I**	**O**	**U**	**Y**
	Space*	Ẃ	ŵ	Ŵ	b	Ŝ	w
B	ý	Ẅ	ĝ	Ĝ	g	Õ	ġ
C	ď	Ḍ	ā	Ā	d	Á	á
D	Ĉ	ẑ	j	J	z	Ś	ś
F	R̂	Ẓ	ō	Ō	o	Ó	ó
G	ħ	ẋ	ḱ	Ḱ	k	õ	ķ
H	Ł	Ḷ	ï	Ï	l	Ĺ	ľ
J	Ď	Ṁ	ē	Ē	m	Ê	ê
K	ẖ	Ṅ	ń	Ń	n	Ṇ	ņ
L	Ĥ	Ẋ	e	Ẍ	x	Ĵ	ĵ
M	Ħ	Ṗ	ã	Ã	p	Æ	æ
N	Þ	Ṙ	ŗ	Ŗ	r	Ŕ	ŕ
P	ĉ	Ṡ	ć	Ć	s	ň	c
Q	Ţ	Ĩ	Ĩ	ţ	Ţ	i	Ẑ
R	ô	Ẁ	ő	Ő	f	Ò	ò
S	ñ	ẋ	ḥ	Ḥ	h	Ḥ	ḥ
T	Ý	Ẏ	ù	ř	y	Ú	ú
V	Š	Û	Û	v	ü	ë	Ẅ
W	Ģ	Ġ	Ġ	ğ	Ğ	ḡ	Ḡ
X	Ķ	ķ	Ķ	ǩ	Ǩ	ķ	Ķ
Z	ě	Ċ	Ċ	č	Č	ç	Ç

*Space should be better substituted by at least two characters.

It can be observed that the original text is not encrypted evenly, since in some cases a substitution character from the key-grid replaces a single letter of the original text, while in other cases it replaces a syllabic pair of letters. Consequently, the main features of PaSyS can be summarized and discussed below, especially regarding its resistance to the previously presented deciphering techniques:

- In the case of frequency analysis decipherment, the frequency of each letter of the original text is drastically altered, since in the cryptogram, an A (for example) sometimes occurs alone but mostly along with the 20 different consonants of the key-grid (C_iA). So do the consonants along with the six vowels of the key-grid.
- In the case of maximum common divisor decipherment, the PaSyS cryptogram is not based on the displacement encryption.
- In the case of Babbage-Krasinski decipherment, the original text is not evenly substituted by bigrams, since only the combination C_iV_i is used and not the rest of them (i.e., V_iC_i, C_iC_i and V_iV_i). Thus, two consecutive characters of the cryptogram sometimes correspond to two letters of the original text (C_iV_i) and sometimes to three ($V_iC_iV_i$, $C_iV_iV_i$, $C_iC_iV_i$, $C_iV_iC_i$).

Finally, special care should be taken for the encryption of word delimiters (space and punctuation marks), either by substituting or by deleting them. The decipherment of polygraphic encryption can be facilitated by the n-grams statistics (e.g., for the Greek language see [14]). Equivalently, there are word distribution frequencies (e.g., for the Greek language see [15]) that may support an adversary in deciphering a cryptogram. In this particular case, a single alone character of the cryptogram (e.g., " Ẃ ") may correspond to 12 English words [16]:

- {"a", "I", "by", "do", "go", "hi", "lo", "me", "my", "no", "to", "we"}.

Although an attempt to decipher a large text by using the above mapping may lead to combinatorial explosion, the relevant positioning of the particular word and the frequency of its occurrence may lead to a manageable deciphering process. Thus, word delimiters should be either difficult to determine by using polyalphabetic substitution for them or deleted altogether.

5. CONCLUSION

In conclusion, PaSyS algorithm of text encryption has been presented for applications of cryptography and information security. This algorithm has been inspired by the Aegean Scripts, developed mainly in Crete (later on in Cyprus as well) more than 40 centuries ago, and by the Natural Language Processing applications for their reading [17] or decipherment [18]. These scripts are classified as syllabaries, because their writing symbols render syllabic phonetic values, consisting mainly of one consonant followed by one vowel (C_iV_i). Accordingly, PaSyS performs a partial syllabic substitution of the original message, only for those syllables that conform to this pattern (C_iV_i), through a corresponding key-grid (Table 1). In this key-grid, every single consonant or vowel and every pair of consonant-vowel (C_iV_i) is arbitrarily mapped to a Unicode (or UTF-8) character. In this respect, PaSyS is an encryption algorithm of:

- semi-polygraphic substitution, since not all of the bigrams (groups of two letters) are substituted;
- a combination of polyalphabetic and homophonic substitution, since an irregular set of symbols, taken from many Unicode alphabets, is used in the key-grid for the substitution.

The above features create a mixed frequency pattern that make a PaSyS cryptogram more resistant to normal statistical decipherment. Most historical and well-known methods of encryption can be deciphered through the statistical analysis of the symbols' occurrence frequencies in text, whenever they follow a normal substitution pattern. For an even more resistant encryption to deciphering techniques, a second phase of encryption may follow right after the first phase of PaSyS application, based on modern mathematical methods of the computer era [19, 20]. PaSyS has been originally implemented in Python programming language for texts in Greek [10], while an English version has been demonstrated herein.

REFERENCES

1. Gómez, J., *Mathematicians, spies and pirates of computer science: Coding and cryptography, The world is mathematics*. 2011, Athens: 4π [in Greek].
2. Papakitsos, E.C., *Natural language and Computational Mathematics*. 2013, Athens: Technoglossia [in Greek].
3. Rivest, R.L., *Cryptography*, in *Handbook of Theoretical Computer Science*, J. Van Leeuwen, Editor. 1990, New York: Elsevier.
4. Biggs, N., *Codes: An introduction to information communication and cryptography*. 2008, London: Springer.
5. Ashchenko, V.V., *Cryptography: an introduction*. 2002, American Mathematical Society.
6. Marks, L., *Between silk and cyanide*. 1998, New York: The Free Press.
7. Al-Kadi, I.A., *The origins of cryptology: The Arab contributions*. Cryptologia, 1992. **16**(2): p. 97–126.
8. Mikros, G., N. Hatzigeorgiu and G. Carayannis, *Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus*. Journal of Quantitative Linguistics, 2005. **12**(2-3): p. 167-184.
9. Singh, S., *The Code Book: The Science of Secrecy from Ancient Egypt to quantum cryptography*. 1999, London: Fourth Estate.
10. Papakitsos, E.C., *Software documentation for syllabic compression of Greek texts with experimental encryption in Python programming language*. 2017, Athens: NLG [in Greek].
11. Davis, B., *Introduction to the Aegean Pre-Alphabetic Scripts*. KUBABA, 2010. **1**: p. 38-61.
12. Karali, M., *The Cypriot Syllabary*, in *A History of Ancient Greek from the Beginnings to Late Antiquity*, A.-F. Christidis, Editor. 2007, Cambridge University Press, p. 239–242.
13. Hooker, J.T., *Introduction to Linear B* (2nd reprint). 2011, Athens: MIET [translated into Greek].

14. Touratzidis, L., *Transformations of natural language and applications in correction of spelling and Greek stenotype*. 1991, Doctoral thesis: National Technical University of Athens [in Greek].
15. Papakitsos, E.C., *Contribution to the morphological processing of Modern Greek: Functional Decomposition – Cartesian Lexicon*. 2000, Doctoral thesis: National & Kapodistrian University of Athens [in Greek].
16. Stavropoulos, D.N. and A.S. Hornby, *Oxford English-Greek Learner's Dictionary* (4th reprint). 1988, Oxford University Press.
17. Kontogianni, A., C. Papamichail and E.C. Papakitsos, *Educational software for learning Linear B*, Proceedings of the *9th Conference on Informatics in Education* (CIE2017). 2017, University of Piraeus, p. 423-433 [in Greek].
18. Chorozoglou, G., N. Koukis and E.C. Papakitsos, *An application of software engineering for investigating the language of Phaistos Disk*. Academic Journal of Software Engineering and Its Applications, 2017. **1**(1): p. 1-10.
19. Zwicke, A., *An Introduction to Modern Cryptosystems - GIAC Version 1.4b* (option 1). 2003, Bethesda, MD: SANS Institute.
20. Bellare, M. and P. Rogaway, *Introduction to modern cryptography*. 2005, California: San Diego and Davis.

AUTHOR PROFILE

Dr. Evangelos C. Papakitsos received his BSc in Physics (1987) from National & Kapodistrian University of Athens, Greece. He received his MSc degree in Information Systems (1992) from The University of Liverpool, UK and his post-graduate diploma in Documentation (2000) from Athens University of Economics and Business, Greece. He received his Doctoral degree in Computer Science (2000) from National & Kapodistrian University of Athens, Greece. He has also received a post-graduate certificate in Career Counselling (2008) from Panteion University of Social and Political Sciences, Greece. His post-doctoral research was in Natural Language Processing (2000-2004) at the National & Kapodistrian University of Athens, Greece. He has served as an Adjunct Professor in four universities of Athens, Greece, teaching both under-graduate and post-graduate courses. Currently, he is serving as a Career Counsellor in Greek Secondary Education and he has been selected as teaching/research staff at the Department of Industrial Design & Production Engineering of the University of West Attica, Greece, waiting to be appointed soon. His research interests are in natural language processing, archaeolinguistics, applied Physics and power engineering. Finally, he is a reviewer in seven international journals, publication organizations and national conferences.