Crypto-Semantic Method of Text Data Protection

Saleh Ebrahim Alomar[1] Hazem (Moh'd Said) Abdel Majid Hatamleh [2]

1.AL-Ahliyya amman university,Department of Engineering, Amman Private University, Amman, Jordan

2. Computer DepartmentUniversity of Al-Balqa' Applied, Faculty of Ajloun, Jordan

Abstract.

A method of text data protection called crypto-semantic is proposed. In order to implement the method within the formally defined restrictions of the selected sphere of applied uses, it is necessary to develop a corresponding lexicographical system in the form of an applied linguistic corpus and semantically structure the information using the constructed linguistic corpus so that the encrypted text message samples present semantically plausible text fragments. Under certain conditions the method provides absolute guarantee of text data protection from confidentiality compromise.

Introduction.

A deficiency is inherent to cryptographic text data protection tools which restricts the possibilities of their use for the provision of confidentiality of particularly important confidential information, where the protection tools' cryptographic strength has to be beyond any doubts. The essence of this deficiency is that in the process of breaking any of the known cryptographic systems the cryptanalyst is able to identify the moment their work is successfully completed. This ability stems from the fact that during the cryptographic enciphering of the text the semantic content of the information being protected is, as a rule, transformed into semantically undefined set of alphabet symbols used [1]. Therefore, the fact of the successful protection system breaking is identified by the cryptanalyst at the moment of transforming the nonsense encrypted text into a semantically correct symbol sequence which as a rule reflects the original meaning of the messages. Thus, the following is true for the known text data protection methods: if a sufficiently large number of encrypted information samples (demonstrably larger than the volume of password information) is available to the cryptanalyst and the brute force attack is used, a theoretical, and in many cases practical, probability to compromise any of the known methods of symmetrical encryption of text data exists. Hence the trust for the formal definition of the cryptographic strength of the known text data protection methods is not absolute, as the known cryptographic systems do not absolutely guarantee protection against confidentiality compromise. The absolute guarantee is understood within the context of C. Shannon's canonical work [2], in which a protection system able to absolutely guarantee information confidentiality is called a perfect secrecy system.

The objective of this paper , To ensure the possibility of functioning of a cryptographic system of text data protection with any predetermined formally proven level of strength against cryptanalytic attacks, and under certain formally defined conditions to provide absolute guarantee of text information protection against confidentiality breach both during its storage in computers and its transmission through unprotected information transport environment. At the same time, protection must be absolutely guaranteed both from the theoretical and practical points of view, even in the circumstances when a sufficiently large number of encrypted information samples (demonstrably larger than the volume of password information) is available to the cryptanalyst.

Main idea. To use the semantic properties of the text data which is protected. In particular, drawing on the achievements of the lexicographical systems theory, to preliminarily linguistically process the text data samples to be encrypted so that the samples of encrypted text messages appear to be semantically plausible text fragments which are impossible for a violator to discern from true ones.

Implementation scheme

Within the formally defined restrictions of the selected sphere of applied uses, a lexicographical system is developed in the form of an applied linguistic corpus [3,4], and then, immediately before enciphering text data by any chosen method of symmetrical encryption, these data are semantically structured with the use of the constructed linguistic corpus in such a way that the samples of encrypted text messages present semantically plausible text fragments and any use of any of the possible methods of cryptanalysis result in obtaining text messages, although differing in meaning, but also semantically plausible.

Basic definitions

The information confidentiality method dealt with in this article belongs to the cryptographic methods of symmetrical encryption of text data presented in a computerized from in any of the communication languages (a natural human or formally defined artificial language) for which a corresponding lexicographical system (in particular, a linguistic corpus) has been created. This method is called a crypto-semantic method (CSM). The prerequisite of the CSM use is the following: a semantic thesaurus (i.e., duly structured semantic lexicon system) covering the subject area of the given protection method use must be incorporated into the lexicographical system.

A linguistic corpus is defined as a software tool which automatically divides the electronic text to be protected into "microcontexts" – text fragments grouped around linguistic units (in particular, words, phrases, scripts etc.) which are the interpretation objects [3,4]. A language thesaurus is a language lexicon with defined semantic relations between linguistic units. In the given case, thesaurus is a hierarchical structure of semantic lexicons defining the semantic relations between the linguistic units of the language used to display the text messages of the selected sphere of applied uses.

Initial conditions

This article demonstrates the results of creating the crypto-semantic method of text tabular data protection characterized by at least two essential advantages: the CSM, under certain circumstances, provides unrestricted cryptostrength but, at the same time, does not deteriorate the functioning quality of the applied computer systems whose information resources are to be protected. The practical implementation of this method is entirely feasible, if the lexicographic schemes theory is used for its construction and the sphere of use is limited to the protection of the data contained in the predefined table forms. This article deals with the crypto-semantic protection method used when the text information to be encrypted is presented in a table of an arbitrary type. The form of the table is predefined. No information other than that entered into the table is available.

Principle of the method.

Figure 1 shows one of the possible models of the CSM system of tabular data protection, whose function principle is based on the synchronization of pseudo random sequence generators (PRSG) located on the transmitting and receiving sides of the secure exchange channel with the help of a known ciphering key.

The model in figure 1 comprises all the basic elements of a cryptographic system [1,5]. However, beside this, an additional element is included into the flowchart – the subject area thesaurus within which it is planned to use the CSM protection system. The thesaurus is created based on the results of the subject area statistical and semantic analysis and must include all linguistic units which can potentially be included in the open text inserted into the predefined table form. With the subject area thesaurus, known password and predefined table form at hand, the administrator of the application system at the transmitting end forms the original text and sends it to the encryption engine (or program encoder). With the help of the key information, i.e. password, the PRSG is set into a definite starting state and sends the generated sequence to the encryption engine's receiving end. The enciphered text obtained as the result of the encryption engine's work is stored on the computer and (or) transmitted to the receiving side through the open communication channel. The administrator of the application system at the receiving end sends the received enciphered text to the decoding machine and with the help of the known password sets the PRSG starting state. The decoding machine produces the original text formed at the transmitting side.

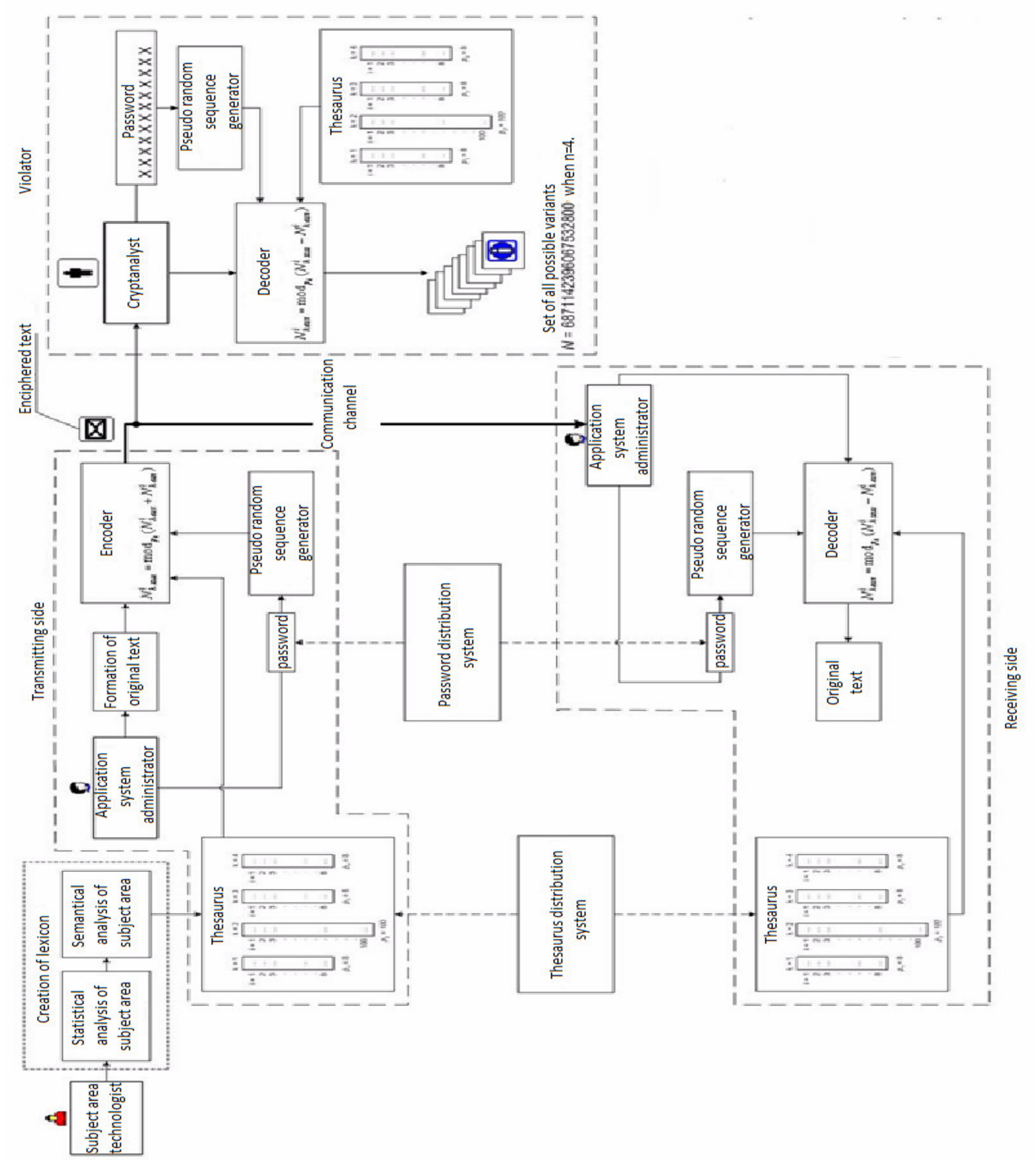
Violator
Password
XXXXXXXXXXXXXX
Pseudo random sequence generator
Thesaurus
Cryptanalyst
Decoder
Set of all possible variants
N = 6871142396067532800 when n=4.
Enciphered text
Communication channel
Transmitting side
Encoder
Formation of original text
Application system administrator
password
Pseudo random sequence generator
Password distribution system
Application system administrator
Pseudo random sequence generator
Decoder
password
Original text
Receiving side
Creation of lexicon
Statistical analysis of subject area
Semantical analysis of subject area
Subject area technologist
Thesaurus
Thesaurus distribution system
Thesaurus

Figure1. Functional model of the CSM system of tabular data protection

The crypto-semantic method of information protection provides for the necessary use of particularly structured lexicons of the languages in which the information to be protected is displayed. Such lexicons are called the thesauri of these languages. The structure of the links between the linguistic elements of the thesauri must ensure the possibility to conceal the true semantic content of the protected information in the generated flow of the so-called plausible "semantic noise" whose role is played by the semantic content of the ciphertext.

In order to prevent the cryptanalyst from assessing the level of successfulness of the attacks on the ciphertext implemented by them, it is necessary to impart plausibility features to the ciphertext. As ciphertext plausibility features, it is expedient to choose the features of logical consistency and/or formal plausibility of its content. The condition of formal plausibility envisages the creation of a situation when:

1) the foreseen result of the ciphertext content analysis identifies a nonzero probability of events or nonzero probability of credibility of information displayed in this ciphertext;

2) in the conditions when a priori knowledge is lacking, it is possible to assume that, from the violator's point of view, the information contained in the ciphertext has approximately the same level of credibility as a lot of other information from the ensemble of possible information kinds within the given applied area.

It is clear that the higher level of formal plausibility is reached in the process of construction of the application language thesaurus, the less possibilities a violator has of successfully deciphering the intercepted text.

In contrast to the stenographic methods directed at concealing the very fact of transmitting the information to be protected, the CSM protection mechanism can be used in the conditions when the violator is aware that the object of his analysis is a ciphertext, not a plain text. However, the analysis of the ciphertext obtained through the CSM transformations denies the violator the opportunity to discern the true content of the protected information from the plausible content of the ciphertexts in the violator's disposal, as the criterion for detecting such difference is lacking. Because of this, the violator is unable to come to any well-founded conclusions as to the level of trueness of the information of the ciphertext under the condition of the absence of a priori information on the content of the protected information and, naturally, the information on the secret ciphering keys used.

Thesaurus synthesis and structuring scheme

In view of the above-said it is possible to claim that the first step in the procedure of creating the CSM protection mechanism whose application is limited to a certain application area is reduced to performing a lexicographic task of compiling and structuring a lexicon. This lexicon, on the one hand, has to reflect the complete thesaurus (i.e., the semantic basis) of the chosen application area and, on the other hand, has to be structured in such a way as the set of the thesaurus elements is divided into certain subsets of formally plausible logically correct constructions (i.e., FSP constructions) of the selected language of displaying information.

The following generalized scheme of thesaurus synthesis and structuring for the case when it is necessary to organize secure information exchange in a certainly defined applied area of application is proposed.

1) The lexicographic task of compiling the lexicon of the application area language which comprehensively displays the thesaurus of this application area is solved. The entries of this lexicon are the linguistic units and/or constructions of the language of displaying the information to be protected. The set of the linguistic units and /or constructions in finite. Under these conditions the set of the entries of the application area language lexicon is displayed in the form of a subset in the set of the entries of the lexicon of the language used for human communication.

2) The set of the entries of the application area language lexicon is divided (on the basis of the application area language semantic analysis) into n subsets, each of them comprising elements which are formally semantically related (i.e., each subset in the lexicon must comprise FSP elements of only one of their kind). Under the condition of such lexicon division each of the resulting subsets comprises only those linguistic units (or linguistic constructions) between which uniquely defined semantic interrelation exists. Such interrelation can be defined (but not necessarily) with the help of a more abstract linguistic unit and/or construction taken from a definition dictionary of the same language. Each subset of FSP elements within the chosen application area must be complete, i.e. comprise all linguistic units or linguistic constructions of the given lexicon which are in the FSP relations with the elements of this subset. When the lexicon is divided in this way, under the designation n the number of different FSP kinds in the chosen application area thesaurus in understood.

3) Each established kind of formal semantic connection between the lexicon elements is allocated an unambiguous (unique) definition in the form of a linguistic unit or linguistic construction of the chosen language in which the information to be protected is displayed. The number of such definitions must equal n. In essence, each of the definitions is an identifier of a corresponded FSP element subset (or an identifier of the kind of formal semantic connection between the lexicon elements).

4) The thesaurus of the chosen application area language is synthesized in the form of two unambiguously interrelated lexicons: one lexicon comprises the identifiers of the FSP kinds, and the other lexicon consists of N sublexicons, each of which is a structured complete subset of FSP elements.

Thus, having a thesaurus for a chosen subject area structured in the way described above, it is possible to organize a secure information exchange on the basis of the CSM protection method. In order to do this, all parties of such exchange must use the same thesaurus.

In the case of enciphering the data of the defined table form according to the specifications of any CSM algorithm, the encrypted tabular data will present a random sequence of formally plausible logically correct constructions (i.e., FSP constructions) of the language used, whose analysis does not allow the cryptanalyst to form any sound view as to the content of the analyzed text.

Below, the functioning of the CSM protection mechanism and the corresponding procedures of enciphering and deciphering the information to be protected that realize this mechanism on the basis of the PRSG use are explored.

Enciphering procedure is based on the substitution of the true element of the plain sequence of linguistic units for the masking element of the secured sequence. Suppose that the thesaurus consists of n sublexicons, where n is the number of columns in the defined table form. Each sublexicon is used to process the data from one column and therefore the number of thesaurus sublexicions used for enciphering/deciphering equals the number of columns of the defined table form. Choose a sublexicon whose elements are of the FSP kind coinciding with the FSP kind of the first element of the input plain sequences of linguistic units. Suppose this is a kth sublexicon from among n sublexicons composing a thesaurus. All the elements of this sublexicon are numbered. Suppose also that the location of the first true element of the input sequence has a sequence number in the structure of the chosen sublexicon Nor, and the first random number provided by the PRSG equals Nran, where 0 ≤ Nran ≤ pk , pk being the dimension (i.e., the total number of words) of the kth sublexicon. Then the sequence number of the location in the structure of the kth sublexicon of a “non-true” (i.e. masking) language element placed as the first element in the output enciphered sequence instead of the true first language element, Nenc will equal :

$$N^{i}_{k\,\mathrm{enc}} = \mathrm{mod}_{\,p_k}\,(N^{i}_{k\,\mathrm{or}} + \mathrm{mod}_{\,p_k}\,(N^{i}_{k\,\mathrm{ran}}) + p_k)\,. \qquad (1)$$

Expression (1) is true when the dimension of the used sublexicon coincides with the spread of random sequences generated by the PRSG (spread being the difference between the maximum and minimum value of the pseudo random numbers in the sequence). In the general case, when the PRSG spread equals the dimension of the largest sublexicon, the enciphering equation, which is the formula for determining the sequence number of the masking word location of the masking work in the kth sublexicon of the thesaurus, has the following form:

$$N_{\mathrm{enc}} = \mathrm{mod}_{\,p_k}\,(N_{\mathrm{or}} + \mathrm{mod}_{\,p_k}\,(N_{\mathrm{ran}}))\,, \qquad (2)$$

where Nor is the sequence number of the location in the kth sublexicon of the ith true language element taken from the original plain sequence of linguistic units to be enciphered, where i is the sequence number of this element’s location in the plain sequence;

Nenc is the sequence number of the location in the kth sublexicon of the masking element put in the ith position (instead of the true element) of the output enciphered linguistic unit sequence;

Nran is a pseudo random whole number generated by the PRSG at the ith stage of generating for the enciphering of the ith true element having the sequence number Nor;

pk is the dimension of the kth sublexicon;

k is the sequence number of the sublexicon in the application area language thesaurus whose elements are connected by a common FSP with the ith language element of the original plain sequence.

Equation for deciphering the information enciphered with the help of any CSM mechanism is the formula for determining the sequence number of the location in the chosen sublexicon (the sublexicon is chosen from the n set of sublexicons of the used thesaurus based on the FSP of their elements with the elements of the original enciphered sequence) of the ith true linguistic unit placed in the output deciphered sequence instead of the masking linguistic unit taken from the original enciphered sequence.

The deciphering equation has the following form:

$$N_{or} = \mathrm{mod}_{\,p_k}\,(N_{\mathrm{enc}} - \mathrm{mod}_{\,p_k}\,(N_{ran}) + p_k)\,, \qquad (3)$$

where Nor is the sequence number of the linguistic unit (according to the numeration in the kth sublexicon) identical to the true unit displayed in the original plain sequence of linguistic units;

Nenc is the sequence number of the enciphered linguistic unit (according to the numeration in the kth sublexicon) taken from the enciphered sequence being processed, at the current stage of deciphering;

Nran is a pseudo random number generated by the PRSG for deciphering the enciphered unit having the sequence number Nenc;

pk is the dimension (i.e., the total number of words) of the kth sublexicon;

k is the sequence number of the sublexicon chosen at the current stage of the deciphering procedure.

Formula (3) is true if the spread of the PRSG equals the dimension of the sublexicon having the most language units among the set of all sublexicons of the thesaurus used.

Provision of randomness of the substitution procedure. It is clear that during one password lifetime under standard conditions the substitution mechanism described above does not ensure the randomness of changes, as by isolating the ciphertext elements spaced apart precisely for one period of PRSG generation, it is possible to isolate the groups of ciphertext elements within which these elements are substituted for the same thesaurus elements [6]. This, in its turn, allows to expect the successfulness of cryptoattacks based on the statistical analysis of the occurrence of elements in intercepted cyphertexts. Therefore, the condition of plausibility of language structures on the transmitting end of the encoder is not sufficient. In such cases it is necessary to ensure the randomness of the procedure of substituting true language structures for structures taken from the subject area thesaurus lexicons. The substitution randomness can be ensured if a table is inserted into the PRSG scheme to store the data on its current state at the moment the communication session was terminated. In this case, the next communication session begins with the PRSG generating a sequence not from the initial state, but from the place where the generation of the pseudo random sequence was finished at the previous session. And if during one PRSG generation period its randomness is ensured (which is totally possible in practice) and under the conditions when the number of lines in any one table form is significantly less than the PRSG generation period, the enciphering system breaking scheme on the basis of the probability analysis of the frequency of occurrence of language structures on the transmitting end of the encoder will not be effective.

Area of application. The CSM protection mechanisms may be implemented in the class on any substitution ciphers. In particular, for enciphering text information any kinds of polyalphabetic substitution schemes (like, for example, Vigenère, Beaufort etc.) can be used, which do not mask the frequency of occurrence of alphabet symbols on the transmitting end of the encoder and thus are hack-sensitive when the ciphertext probability analysis is used. It is also possible to use different variants based on the use of pseudo random sequence generators (PRSG). Under standard conditions such protection systems are also hack-sensitive, as they do not ensure the randomness of the substitution of the output text sequence symbols when the analyzed ciphertext volumes significantly exceed the period of the generated PRSG sequences. However, the PRSG constructed with due consideration of the above, ensures the randomness of substituting the plain text symbols by the symbols taken from the subject area thesaurus and thus ensures the strength against attacks based on the probability analysis of cryptograms.

Main results and conclusions.

1) A method of text data protection called a crypto-semantic method (CSM) is proposed. In order to implement the method within the formally defined restrictions of the selected sphere of applied uses, it is necessary to develop a corresponding lexicographical system in the form of an applied linguistic corpus and semantically structure the information using the constructed linguistic corpus so that the encrypted text message

samples present semantically plausible text fragments. Under certain conditions the method provides absolute guarantee of text data protection from confidentiality compromise.

2) The peculiarities of the crypto-semantic protection method implementation are considered for the case when the text information to be enciphered is presented in a table of an arbitrary type. The form of the table is predefined. No information other than that entered into the table is available. A functional model of the CSM system of tabular data protection is presented. The applicable equations of encryption and decryption of information are provided.

3) The prerequisite of the CSM use is the inclusion in the lexicographical system of a semantic thesaurus (i.e., particularly structured system of semantic lexicons) embracing the subject area of the use of this protection method. A thesaurus is a necessary element in the mechanism of the original plain table text transformation into encrypted table text having the features of formal semantic plausibility (FSP). It should be noted that the index of the similarity measure with the original (true) text must be a random variable whose specific value is unknown to the cryptanalyst. The thesaurus must comprise all possible kinds of linguistic constructions used to build (synthesize) original plain texts entered into the defined table form. No requirements to the thesaurus secrecy are specified. This thesaurus must be available to all potential secure exchange parties and used in the enciphering processes on the transmitting side (the one that generates information) and in the deciphering processes on the receiving (reading) side.

4) A scheme has been developed of thesaurus synthesis and structuring for the case when it is necessary to organize secure information exchange in a certainly defined applied area of application. This development made the organization of secure information exchange on the basis of the proposed CSM mechanism possible.

5) The main characteristic of the crypto-semantic methods of information encryption is connected with the specific features of thesaurus structuring. Firstly, a thesaurus must adequately reflect the application area of the CSM protection system use, as any information which may in principle require protection must be fully reflected in the thesaurus of the constructed CSM system. Secondly, a thesaurus must be specifically structured in such a way as to provide the possibility to obscure the content of the information to be protected in the generated flow of plausible semantic noise.

References

1. Schneier B., Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd ed. New York // John Wiley and Sons, 1996.

2. Шнайер Брюс. Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си – Москва (Російська Федерація):Издательство ТРИУМФ, 2003. –816 с.

3. Шеннон К.Э. «Теория связи в секретных системах». В кн.: Шеннон К.Э. Работы по теории информации и кибернетике. М.: Иностранная литература. 1963, с. 332-402, -829 с.

4. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. – Київ (Україна): Довіра, 2005. -471 с.

5. McEnery, Tony and Wilson, Andrew. Corpus Linguistics: An Introduction. 2nd Edition, Edinburgh: Edinburgh University Press, 2001. – P. 2 – 3.

6. Математичні основи криптоаналізу [Текст]: навч. посібник / С.О.Сушко, Г.В. Кузнецов, Л.Я. Фомичова, А.В. Корабльов. – Дніпропетровськ (Україна): Національний гірничий університет, 2010. - 465 с.

7. A statistical Test Suite for the Validation of Random Number Generators and Pseudo Random Number Generators for Cryptographic Applications [Text]: NIST Special Publication 800-22 Rev1. – Gaithersburg, Maryland: NIST, 2008. – 153 p.