ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 6, December 2015

DECRYPTION SYSTEM OF THEMATIC TEXTS IN SPANISH USING FREQUENCY ANALYSIS INCLUDING UNIGRAMS, BIGRAMS AND TRIGRAMS

Bárbara Emma Sánchez Rinza, María del Rocío Guadalupe Morales Salgado*, Pablo León Morales. Benemérita Universidad Autónoma de Puebla, Universidad Popular Autónoma del Estado de Puebla*.
Facultad de Ciencias de la Computación, 14 Sur and Avenida San Claudio

Abstract*——*This paper describes a cryptosystem for the Spanish language. All languages have some words that are more common than others to make connections between phrases or sentences. In the Spanish language there are several types of words like prepositions, articles that are words of one, two or three frequently used letters. The decryption by syllabic frequencies is an algorithm based on Spanish grammar rules, which is done by a statistical study of frequencies of words of one, two and three most common letters. The method comprises: selecting a theme from a text of about 10,000 words, and calculate the frequencies of these words of one, two and three most used letters in the Spanish language, we will name these types of words unigrams, bigrams and trigrams and comparing with the ciphered text that has to have the same subject. For the process of encryption we will use the Vigenère algorithm to encode the encrypted text previously and subsequently it will be decoded using this technique [1].

I. INTRODUCTION

The word cryptography is a term that describes all techniques to encrypt messages or make them intelligible without resorting to a specific action.

Cryptography is based on arithmetic's: In the case of a text, consists in transforming the letters that make up the message into a series of numbers (in the form of bits since computers use the binary system) and then perform calculations with these numbers in order to:

- To modify them and make them incomprehensible, the result of this modification (the encrypted message) is called ciphered text as opposed to the initial message, called clear text.
- To make sure the recipient can decrypt them, the action of encoding a message to make it secret is called encryption and the inverse method, which is to recover the original message, is called decryption.

Cryptanalysis involves the reconstruction of an encrypted clear text message using mathematical methods. Therefore, all cryptosystems must be resistant to cryptanalysis methods. When a cryptanalysis method allows to decrypt an encrypted by using a cryptosystem message, we say that the encryption algorithm has been decoded.

II. REALIZATION OF THE CRYPTOSYSTEM

The algorithm used consisted, first in take a training text of more than 10 000 words of a specific topic and frequencies of unigrams, bigrams and trigrams was obtained (it should be emphasized that this text is not encrypted). Subsequently the same subject is encrypted by the Vigenère algorithm. Similarly the frequencies of the ciphered text were obtained. After having the two tables of frequency of the training text and the ciphered text, we will replace first by monosyllables, two-syllables and three-syllables. Subsequently a word processor yielding the percentage relationship of letters and words between the original text and decryption is used.

To implement the decryption of text by analyzing the frequencies of mono, bi- and tri-syllabic words, the process was to analyze the input ciphered text to determine the frequency of each of the letters, the frequency of the words of one-letter (unigram), two-letter words (bigram) and three-letter words (trigrams). From the training text. Once we have these frequencies, we know the most common elements in the ciphered text. Finally we proceed to replace these items with the most common in the Spanish language. The Java programming language was used for coding the algorithm. Diagrams used are shown.

Use Case Diagram

In Figure 1 it is shown a use case diagram in which the user interacts with the system. Below is a description of each use case is:

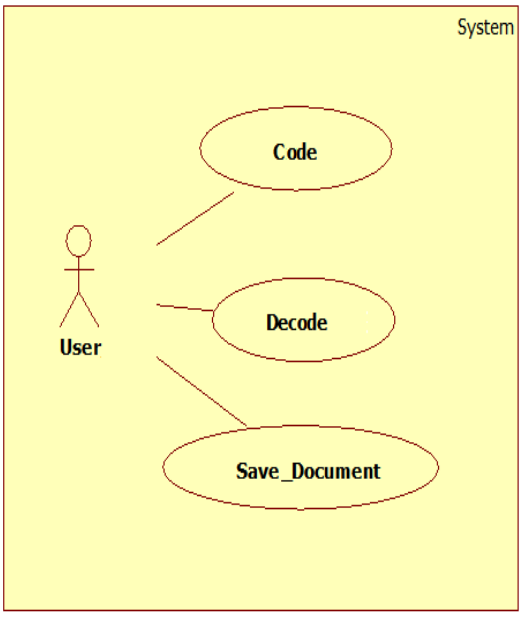

Fig 1. Use Case. Source: Own elaboration.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 6, December 2015

Description of each use case

Use Case	Description
Coding	Encryption of the original document, it can be encrypted by the user or given externally.
Decoding	Decryption of the document obtained. The user must load the document manually.
Save_Document	The user can save the document decryption. The location where it will be stored must be selected.

Class diagram

Figure 2 shows the class diagram of the system. The main class is "Encryption", which performs the cryptographic operations (training document) and decryption (document provided by the user). The "List" and "Node" classes are used for making the frequency tables. The "Practice1" and "Interface" classes are responsible for the interface display.

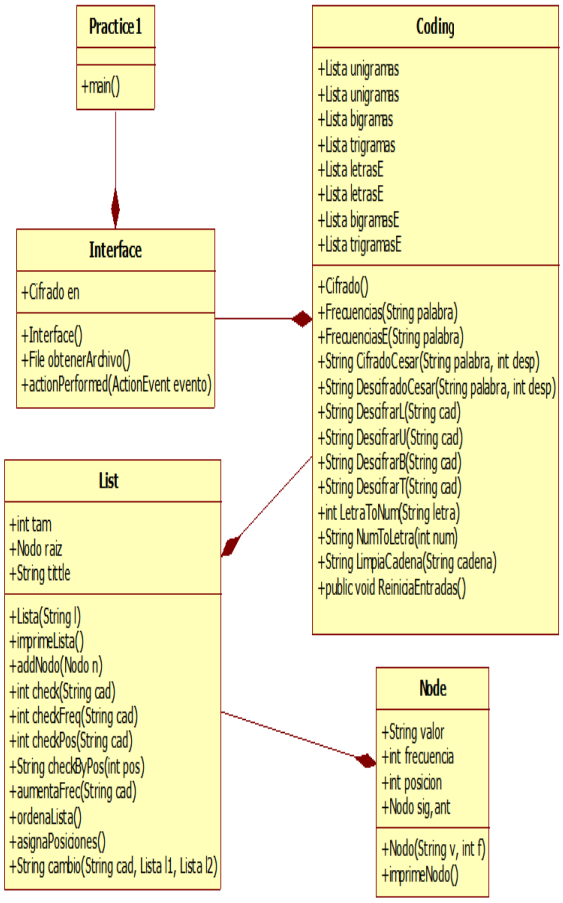

Fig 2. Class diagram. Source: Own elaboration.

Activity Diagrams

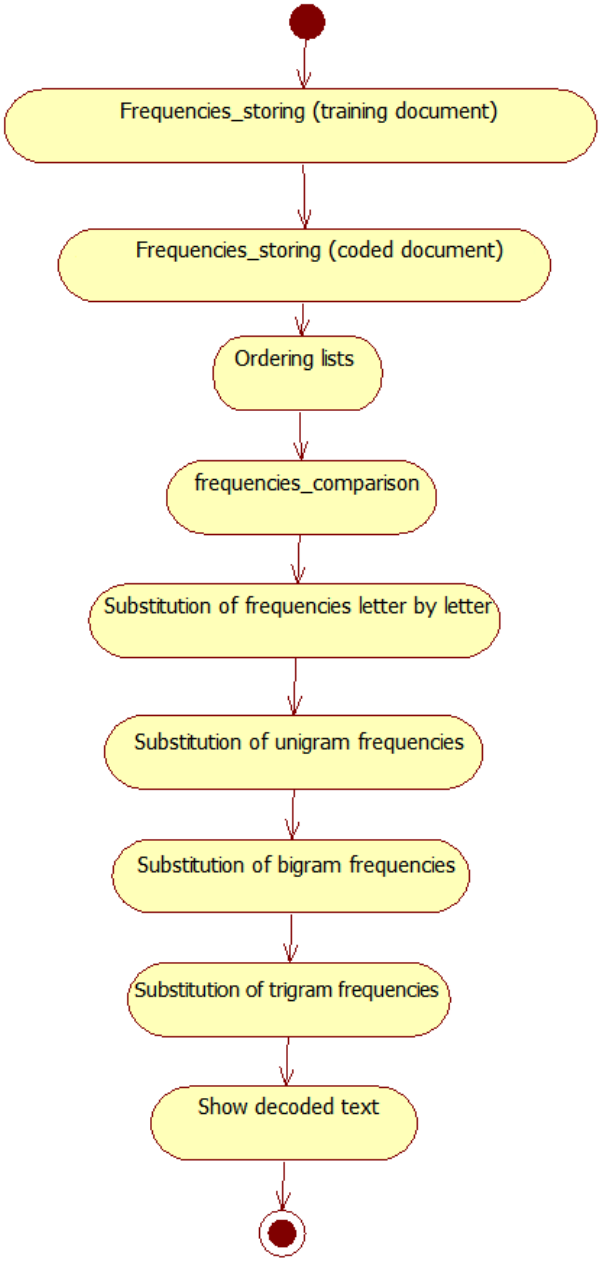

Fig 3. Activity diagrams. Source: Own elaboration.

Implementation

First, for storing frequencies four lists were created (list of frequencies of letters in Spanish, unigrams, bigrams and trigrams), we will call the first frequency of letters in general, the Spanish language some letters are more used than others as you can see in Figure 4. We have the clear text or training text and the ciphered text, the decoding program will replace the clear text list in the list of the ciphered text and save it.

Besides, the calculation of the percentage of success can be explained by: If A, B and C are three events not mutually exclusive (intersect events), that is, in such a way that A or B or C happens, or all at once (simultaneously), then the following rule applies for calculating such probability:

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 6, December 2015

P(AUBUC)= P(A)+ P(B)+ P(C)- P(A)- P(B)- P(A) + P(A)
See Figure 4

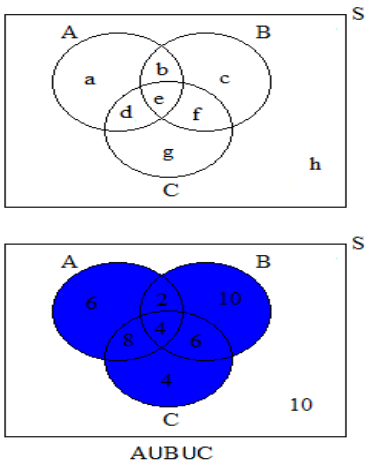

Fig 4 the probability of the sum of the 3sets

This is the mathematical back-up of what is done below

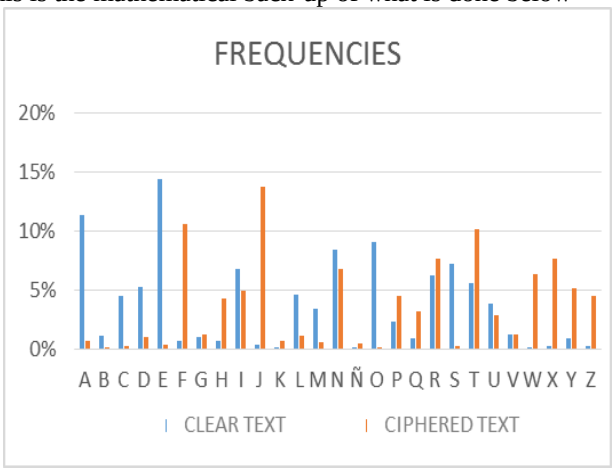

Fig 5. Frequencies of a single letter of the clear text or training text and the ciphered text.
Source: Own elaboration.

In Figure 6 we have the graph for bigrams, that is, two-letter words that are more common in Spanish from the training text.

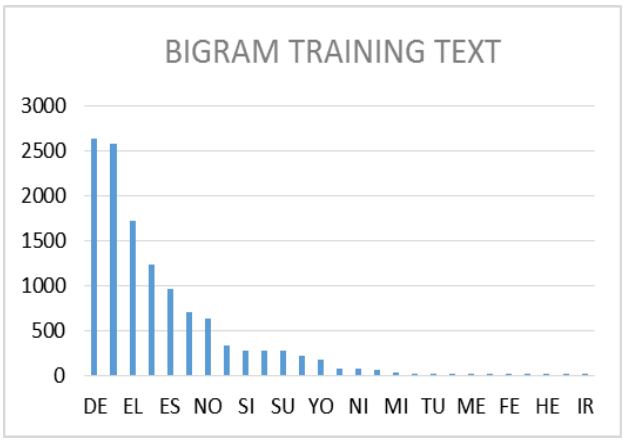

Fig 6. Frequency of the bigram from the training text.
Source: Own elaboration

The words we have in the bigram of Figure 5 are replaced by the list of bigrams taken from the ciphered text, Figure 7.

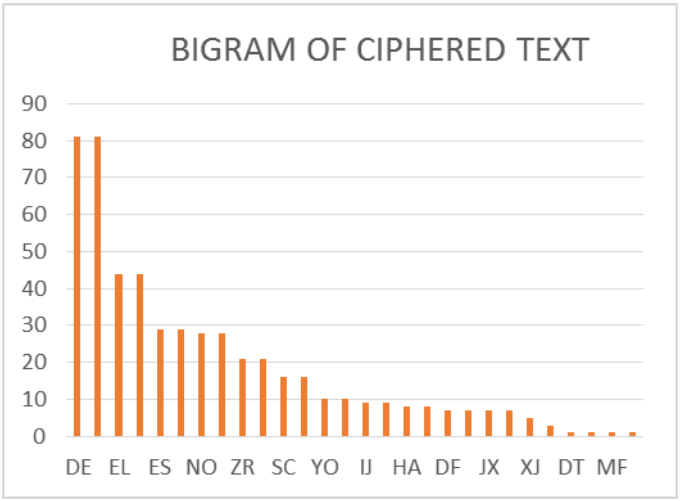

Fig 7. Bigram of the encrypted text.
Source: Own elaboration.

Below comes the trigram from the training text of 10,000 words shown in Figure 8 and will be replaced in the text that has been decoded by the list in Figure 8, the trigram of ciphered text .

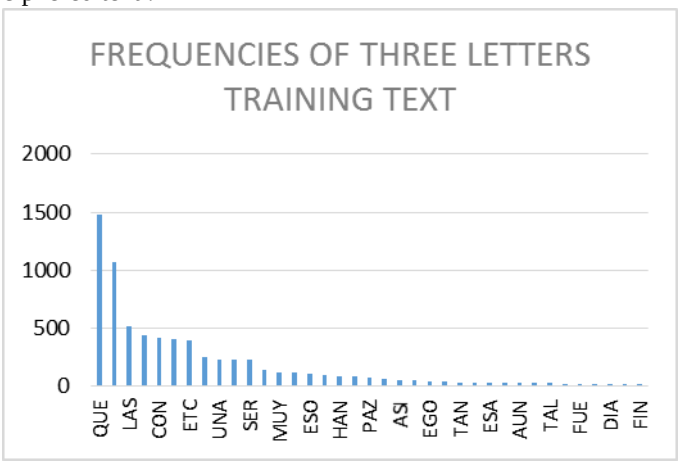

Fig 8 Trigram of the training text.
Source: Own elaboration.

On Figure 8 the words will be replaced by the ones which we have in Figure 9 as was already mentioned.

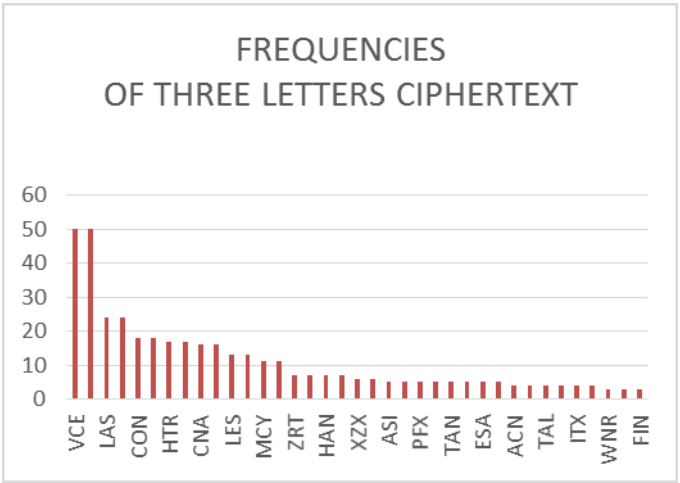

Fig 9 Trigram of cipher text.
Source: Own elaboration.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 6, December 2015

III. OBTAINED RESULTS

After implementation, testing with the ciphered text was performed, first performing the decryption of frequencies taking only into account the frequency of individual letters (Figure 10) With this configuration, a 60.2 % accuracy was obtained by comparing letter by letter the decrypted text obtained by our algorithm and the original clear text.

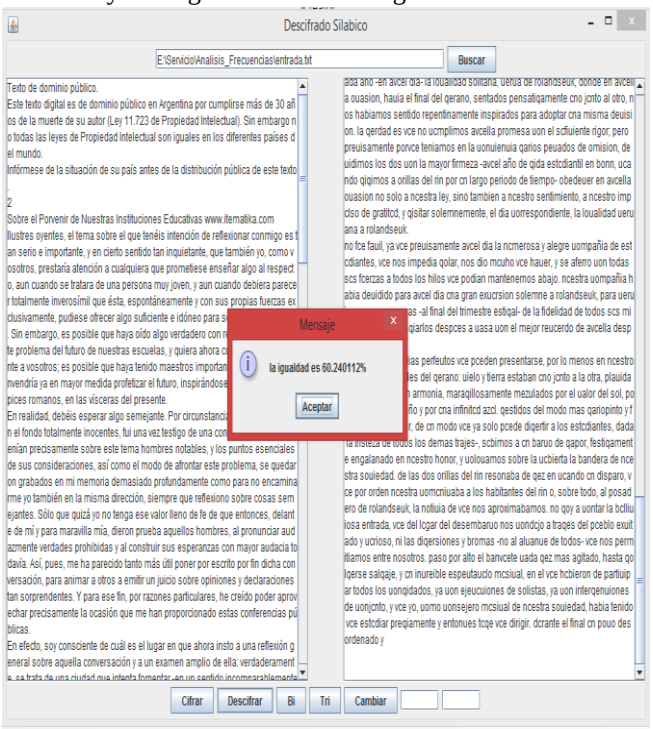

Fig 10 decoding with a single letter (unigram). Source: Own elaboration.

This result was saved and applied to the bigram to continue decoding the ciphered text and a percentage of 61.9.1% accuracy letter by letter is obtained. A slight improvement is noted referring to the understanding of the text (word comparison) and visually the text obtained is more understandable, figure 11.

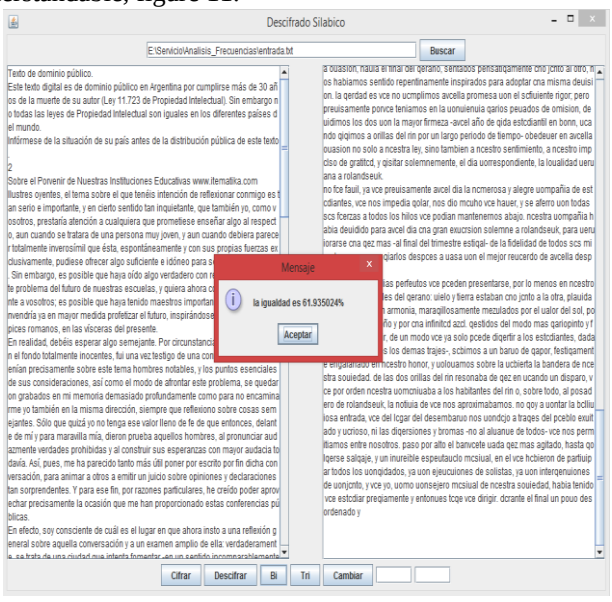

Fig 11. Decoding of the ciphered text using the bigram. Source: Own elaboration.

The results obtained in Figure 11 are saved and the algorithm is applied again for trigrams. Here 65.8 % is obtained, which turns into a more legible message, and with knowledge in Spanish it can be read. Figure 12 .

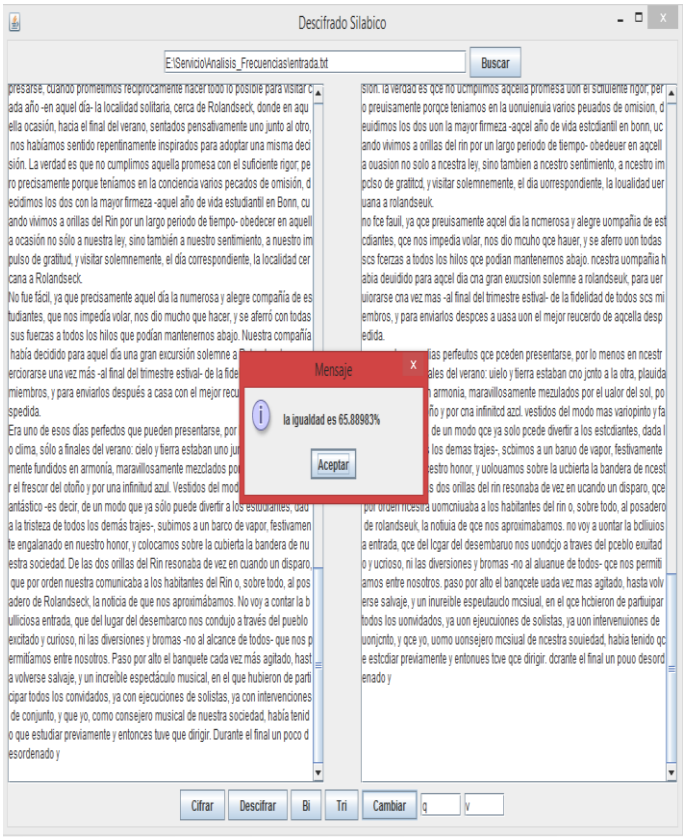

Fig 12. Decoding of the ciphered text using unigrams, bigrams and trigrams. Source: Own elaboration.

IV. FUTURE WORK

The system presents great opportunities for future versions, making it possible to work with special characters, such as accents, signs, and capital letters, as the decrypted text is fully returned in lowercase. In addition to this you can add a smart algorithm that generates a search for similar words, not existing words, and based on the theme, similarity of the word and existing repetition, look up for the appropriate word and change it.

V. CONCLUSIONS

With this algorithm using frequency analysis for the decryption of texts, acceptable results were obtained, although it harder work must be done to improve the results. Applying the algorithm to a ciphered text, a great visually understanding of the text is obtained. But it only serves to decode texts whose mathematical functions are overyectics [2].

This type of frequency decoding algorithms only get good results with running encryptions, or those algorithms that yield the same amount of letters as the clear text and its encryption.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 6, December 2015

REFERENCES

[1] Sánchez, B. Bigurra, Diana. et all. De-Encryption of a text in Spanish using probability and statistics. 18th International Conference on Electronics, Communications and Computers: isbn 07695 3120 2 march 2008.

[2] Sánchez, B. Cruz, S. Cesar decryption algorithm, but the method of frequency points in the Spanish language. International Journal of Engineering and Innovative Technology, vol 3, issue 5 November 2013 ISSN 2277375.

AUTHOR BIOGRAPHY

Bárbara Emma Sánchez Rinza Bachelor in Physics, Master Degree in Optics, Doctor's Degree in Optics. She has written 44 chapters of books, 34 national and international paper, 13 memoirs. She has participated in 105 conferences in different forums. She has directed 31 Bachelor Thesis and 6 Master Thesis.