International Journal of Security and Its Applications
Vol. 4, No. 1, January, 2010

Effect of Language Complexity on Deciphering Substitution Ciphers - A Case Study on Telugu

Bhadri Raju MSVS[1], Vishnu Vardhan B[2], Naidu G A[3], Pratap Reddy L[4], and Vinaya Babu A[5]

[1]*Associate Professor in CSE,S.R.K.R.Engineering College,Bhimavaram, A.P.,India*
[2] *Professor & Head of CSE,Indur Institute of Engg&Tech..,Siddipet, A.P., India*
[3]*Research Scholar, Dept of CSE, JNTU Kakinada, Kakinada A.P., India*
[4]*Professor&Head of ECE, Jawaharlal Nehru Technological University,Hyderabad,*
[5]*Professor in CSE& Director,Admissions,JNTUniversity,Hyderabad, A.P., India*
[1]*msramaraju@ gmail.com,* [2]*mailvishnu@ yahoo.com,*
[3]*apparaonaidug@yahoo.com,* [4]*prataplr@ rediffmail.com,*[5]*dravinayababu@yahoo.com*

Abstract

Global connectivity provided praxis for data transactions. Data or information is available in different forms like text, image, audio, video etc. Security mechanisms are aimed at security algorithms with an assumption that the information is in bit stream. Human understanding in practice deals with information other than bit stream. In the transaction process of information, users allow human understandable format of data. A simple case of text data may deal with various types of scripts represented in multiple combinations of bit streams. Transformation of the basic characteristics that are embodied in the script is an interesting area to be explored in security models. Decipherment problems related to message equivocation is mainly dependent on the statistical complexity of script. The present work is aimed at analyzing the basic characteristics of a script in the form of frequency distribution of character code points. The proposed model is evaluated on Telugu script as a case study with a comparison on Latin text. The evaluation is limited to 8-bit and 16-bit key sizes.

Keywords**:** *Bit Stream, Frequency Distribution, Character Code Points, Script*

1. Introduction

Providing secured communication is a challenging task with so many languages in the world consisting of characters of different properties and behavior. Cryptography is one of the methods in which the security goals can be achieved by means of encryption and decryption. In general, such scheme uses symmetric key algorithm or asymmetric algorithm where each block of fixed/variable size bit stream will be transformed to cipher stream. They use either block cipher or stream cipher techniques for transformation. Parameters in these schemes are mainly considered to be algorithm and key. These algorithms are evaluated adequately on ASCII based Latin text. To attain greater levels of security emphasis is made on the incremental increase in the size. The reflected result can be found with increased hardware complexity. Introduction of

Unicode uniquely represent all the characters of script based languages in the world. The process of localization[1] took a gallop with increasing trends in the information exchange of language based context. This phenomena demands for security systems that are specific to script. In this scenario the characteristics of text play a vital role, which need to be considered as a parameter. Transformation of text during the cryptographic process is to be analyzed for various levels of security. A simple logical conclusion may state that if the text of a script is complex then the same level of security can be achieved with less key size also. In this paper we addressed the information security issues related to Indic scripts with an emphasis on the complexity of Telugu[2,3] which is mainly used in the southern region of India and ranks second among Indian Languages.

2. Review

Different approaches of cryptanalysis are available in the literature using language characteristics to understand the strength of cipher system. One such approach deals with frequency analysis where in the process of determining the frequency of each symbol in the encrypted message. This information is used along with knowledge of symbol frequencies in the language, to help determine which cipher text symbol maps to the respective plaintext symbol. Success may vary based on the amount of available information about the cryptosystem. In transposition systems, the letter frequencies of a cryptogram are identical to that of the plaintext. In the simplest substitution systems, each plaintext letter has one cipher text equivalent. The cipher text letter frequencies are not identical to the plaintext frequencies, but the same numbers will be present in the frequency count as a whole. K.W. Lee et.al developed [4] the cryptanalytic technique of enhanced frequency analysis using the combined techniques of monogram frequencies, keyword rules and dictionary checking. The proposed three-tier approach mechanizes the cryptanalysis of mono alphabetic simple substitution cipher. Thomas Jakobsen proposed [5] a method for fast cryptanalysis of substitution ciphers which uses the knowledge of digram distribution of the cipher text.

The study of encrypted messages is subdivided into determination of the language, reconstruction of keys and or the plaintext. Recent approaches in literature are being concentrated on retrieval of plain text based on the features of the respective language. Each language has certain characteristics [6,7] that aid in successful cryptanalysis. There are two general approaches to solve simple ciphers. One makes use of the frequency characteristics and the other uses the orderly progression of the alphabet to generate all possible decipherments from which the correct plaintext can be picked up. For example, the individual letters of any language occur with greatly varying frequencies[8]. Similar to that of single letters with typical frequency expectations, multiple letter combinations also found with varying, but predictable frequencies. Extensive statistical analysis of these frequencies are more helpful while retrieving part of plain text message.

Sujith Ravi and Kevin Knight introduced [9] a method that uses low-order letter n-gram models to solve substitution ciphers. This method is based on integer programming which performs an optimal search over the key space. Decipherment accuracy as a function of n-gram order and cipher length is reported. Empirical testing of Shannon's information theory for decipherment uncertainty is the emphasis. Thomas Jakobsen proposed [5] a method of cryptanalysis on simple substitution ciphers (both mono- and polyalphabetic), where an

initial key guess is refined through a number of iterations. The knowledge of the diagram distribution of the cipher text and the expected diagram distribution of the plaintext is reported to be necessary while refining the key. Kevin Knight et.al studied [10] a number of natural language decipherment problems adopting unsupervised learning strategy. Substitution ciphers, character code conversion, phonetic decipherment, and word-based ciphers are studied with relevance to machine translation. Substantially improvement in decipherment accuracy is reported.

Bárbara E. et al presented [11] a method for de-ciphering texts in Spanish using the probability of usage of letters in the language. The frequency of different letters is the clue to the presented de-ciphering. Bao-Chyuan et al proposed [12] a method to improve the encryption of oriental language texts with a case study on Chinese text files which are ideogram based and differ from Latin text. Moreover the numbers of characters that appear in Chinese are much larger when compared to English. The scheme proposed by Bao reported that large Chinese text can be handled more efficiently. A method for Parisian / Arabic script is proposed [13] with regard to shapes and their position in the word. In another Model stegnography is attempted [14] on Persian/Arabic Unicode based Text using the above characteristics including writing system. In the present paper the frequency characteristics of character code points are explored.

3. Security Model

Every language has certain parameters in such a way that language rules are embodied in sequence while formulating document. Complexity of script is mainly dependent on character, word and sentence formulation methods. A document with a meaningful summary can be represented as DϵSϵWϵC where 'D' is document , 'S','W' and 'C' are sentences, words and characters respectively. In case of English, 'C' is represented with the help of one-to-one correspondence of character code points in any machine, where as Indic script representation is associated with two fold phenomena. 'C' in real terms is associated with 'Syllable' which in turn represented as a set of multiple character code points. Now 'C' can be written as Syϵ C_C where 'Sy' is syllable and 'C_C' is character code point. In actual transformation, the character code points are transformed with the help of crypto system. This transformation is done onto a different plane where the mapping is a reversible phenomenon. The transformation characteristics of the meaningful units from the stand point of the frequency characteristics, is a point of interest in the present work. Generally, the frequency characteristics differ from language to language. In case of English due to the smaller size of the character set, the frequency characteristics may effectively be reflected in the transformed data. If the size of the meaningful units is large enough, then complexity of frequency characteristics is to be evaluated. In the mapping phenomena, we have attempted to understand the reflection of frequency characteristics and its impact in the crypto analysis. This is the context in which the present work is addressed.

The proposed model defines meaningful units that are embedded in text documents [9] as essential units and also treated as meaningful units in the form of character or byte stream. The byte stream is a symbolic representation of text. In case of Indic scripts this byte stream is a complex byte stream, where as in case of Latin text the byte stream is one-to-one mapping. The present model addressed this specificity by taking into consideration of words in the form of syllables and extraction of byte stream from syllables. They consist of single code point units or multiple code point units. They will be transformed into a code point byte

stream, converted into bit stream which undergoes transformation similar to that of any system. The code point streams that are derivative of syllables are converted to bit stream. A key stream is generated using efficient Random number generator. A simple 8-bit and 16-bit key stream is used in the present work. With this key stream, transformation function is applied on this bit stream resulting in cipher text as illustrated in Figure 1 and Figure 2.

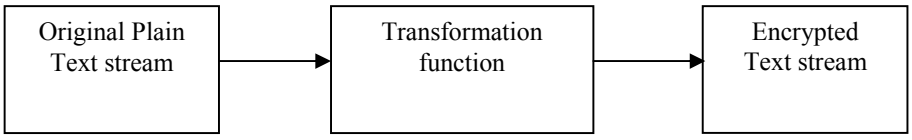

Figure 1. Encryption of Original Plain Text

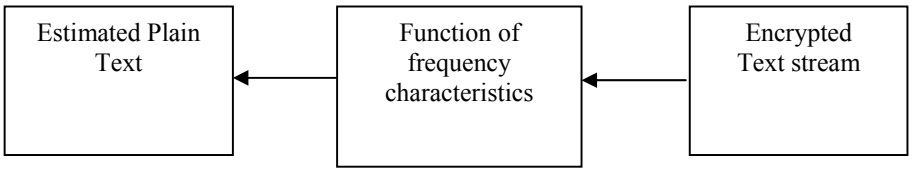

Figure 2. Estimated Plain Text using frequency characteristics

A simple plain text,encrypted text followed by the decrypted tex of English and Telugu languages are performed and presented in Figure 3.

Figure 3. Sample Plain Text, Encrypted Text and Decrypted Text in Telugu

4. Frequency Distribution of Character Code Points of Telugu Script

Basic unit of script description is found with syllable, which is defined by the canonical structure ((C)C)CV. Machine representation of this structure is composed of a set of character code points that are defined in the Unicode code chart. Even though syllables are the meaningful units of script, they are abided by the specific grammar

rules of the script, whereas the character code points in machine representation are perceived as a reflective mechanism of these grammar rules. It is necessary to understand the complex nature of the script in the utility nature of the syllables, which is dynamic in historical perspective. In the present work we are considering the machine representation of character code points and their characteristics in the form of frequency distribution as one of the information that is adopted for the crypto analysis. Many attempts are made[10] on Latin text while extracting the frequency distribution of basic alphabets in a sample of over 300,000 character code points . They demonstrated the dominance of a small set of characters in regular usage. Similar concept is extended in the present work to evaluate the characteristic nature of variable character code points that are embodied in syllables of Telugu text. A sample of 2,40,000 character code points are used for the above analysis which are mainly compiled from the present usage of text.

5. Crypto Analysis using Frequency Distribution

The proposed model is evaluated for two languages i.e. English and Telugu. The encryption algorithm is implemented on different sizes of Telugu text samples. For this process 8- bit key is generated randomly using OS based random generator. Plain text is encrypted using the proposed algorithm and randomly generated 8-bit key resulting in cipher text. The frequencies of different characters in the cipher text are extracted. Mapping is carried out between the characters of plain text and cipher text based on these frequencies. Now the characters in cipher text are replaced with the mapped characters of plain text and the percentage of the exact retrieval as compared to plain text is calculated which is illustrated in Figure 4 and Figure 5 Table1.

When English Text is considered the problems are much less because the correspondence is between the transformed text and the original text. Though the key is generated randomly, since it is fixed, the mapping function transforms it into a point in orthogonal plane. The percentage of retrieved code points using frequency distribution is calculated. If we consider Telugu script the number of character codes that exist in the original medium size text need not be the complete set. In the transformation the original medium size text need not be the complete set. In the transformation process all character codes may not exist from the original set of code points. This may lead to confusion in the crypto analysis. We adopted a thresh holding function in the crypto analysis process for reverse mapping. The percentage of plain text that can be retrieved is varying from 10% to 20% depending on the size of the plain text in case of Telugu. The same process is adopted on English text of different sizes. The percentage of plain text that can be retrieved is varying from 25% to 50% depending on the size of the plain text which is illustrated in Table1 . This percentage in case of Telugu is relatively less when compared to English which is because of the complexity involved in Telugu script. Here a study has been done based on character code point frequencies. If more syllable relations are studied then prediction is more effective.

The proposed cryptographic model is implemented now by considering 16-bit key on Telugu using the above mentioned approach. Mapping is carried out between the characters of plain text and cipher text based on these frequencies. Now the characters in cipher text are replaced with the mapped characters of plain text and the percentage of the exact retrieval as compared to plain text is calculated which is illustrated in Table2.

The results indicate that the percentage of retrieved plain text is varying between 10- 20% whereas for 16-bit key the observed results are found in the range 1-10% only. This is an

indicative measure of language complexity with specific reference to frequency distribution of character code points.

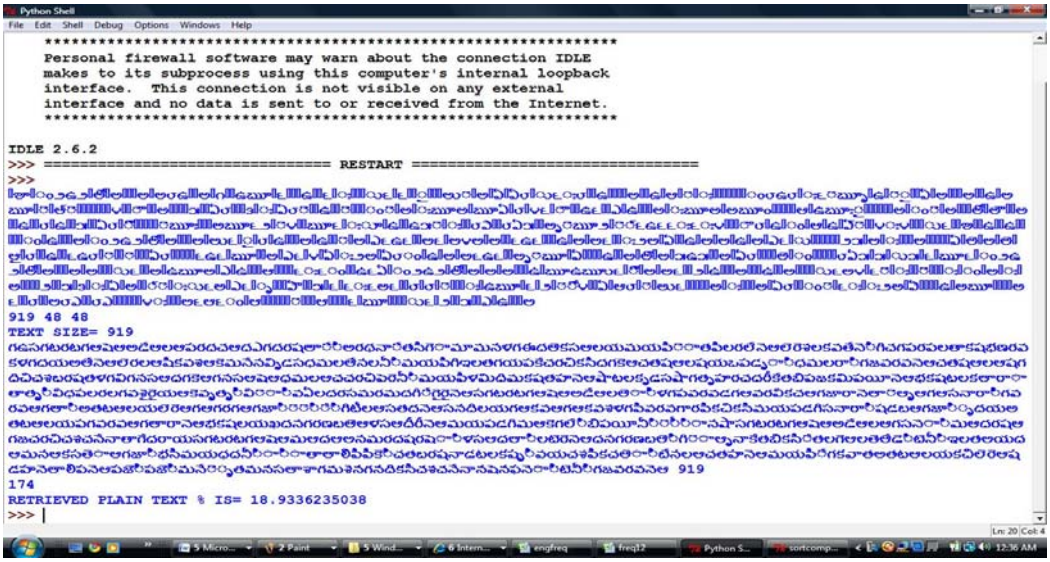

Figure 4. Retrieved Text based on Frequency distribution in Telugu

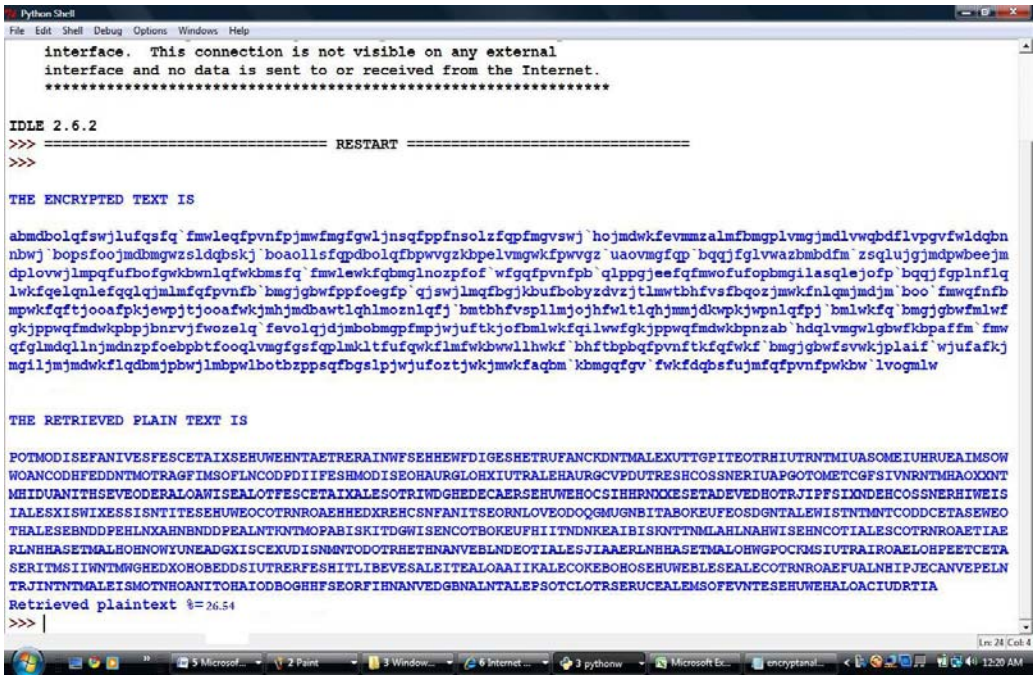

Figure 5. Retrieved Text based on Frequency distribution in English

Table 1. Percentage of retrieved character code points using frequency distribution for English and Telugu with 8-bit key

Plain Text Size Number of characters	% of character code points retrieved	
	English	Telugu
2000	24.43	20.7
4000	49.49	17.1
10000	27.12	8.5
15000	50.89	16.7
22000	41.09	15.05
35000	41.04	15.89
64000	46.81	1.15
75000	31.99	1.94

Table 2. Percentage of retrieved character code points using frequency distribution for Telugu with 8-bit and 16-bit keys

Plain Text Size Number of characters	% of character code points retrieved using	
	8-bit key	16-bit key
2000	20.7	4.4
4000	17.1	2.05
10000	8.5	7.3
15000	16.7	4.47
22000	15.05	9.97
35000	15.89	4.31
64000	1.15	0.8
75000	1.94	1.2

6. Conclusions

Shannon's model of perfect security is a point of interest for the researchers across the world. Statistical evaluation of Ideal message equivocation model is the main goal. In this process the context of "message" be termed as "script" dependent text. In the world of multi lingual data, every script possesses different complexity levels. In the present work, an attempt is made to analyze the text based crypto model using frequency distribution of character code points as a parameter with specific study on Indic scripts. The encryption and decryption process is tested in comparison with English and also on Telugu with different key sizes. Evaluation of the model is carried out with the help of frequency distribution as one of the prominent characteristic of text. Crypto analysis is carried out on both the languages with 8-bit key and the percentage of matches in the reverse transformation is presented. The mapping for English text ranges from 23 % to 50 % where as for Telugu it ranges from 10% to 20%. The analysis is extended to 16-bit key size on Telugu text . The mapping for 16-bit is observed in the range of 1% to 10%. If the text complexities are considered for each script, greater levels of security are observed with smaller key sizes. Evaluation of the proposed model on other Indic scripts of the same nature is in progress.

References

[1] Adam Stone: Internationalizing the Internet. J. Internet Computing. 3, 2003, pp. 11-12
[2] Pratap Reddy, L.,:A New Scheme for Information Interchange in Telugu through Computer Networks: Doctoral Thesis. JNTU,Hyderabad, India, 2001
[3] Vishnu Vardhan, B., :Analysis of N Gram Model on Telugu Document Classification: Doctoral Thesis. JNTU,Hyderabad, India,2007
[4] Lee K.W., C.E. Teh, Y.L. Tan :Decrypting English Text Using Enhanced Frequency Analysis. In: National Seminar on Science, Technology and Social Sciences 2006 (Ui TM-STSS 2006). pp. 1-7
[5] Jakobsen, T.: A fast Method for Cryptanalysis of Subsittution Ciphers. J. *Cryptologia,* Volume 19, Issue 3, 1995, pp.265-274.
[6] Bauer F,L., : Decrypted secrets-Methods and Maxims of Cryptology, Springer ,2007
[7] Menezes A. J. P., : Handbook of Applied Cryptography. CRC Press,2001
[8] De Canniere, C.; Biryukov, A.; Preneel, : An Introduction to Block Cipher Cryptanalysis. J. IEEE,Volume 94, Issue 2 ,2006
[9] Sujith Ravi and Kevin Knight :Attacking Letter Substitution Ciphers with Integer Programming, J.Cryptologia, 33:4, Oct 2009, pp.321 — 334.
[10] Kevin Knight, Anish Nair, Nishit Rathod, Kenji Yamada:Unsupervised Analysis for Decipherment Problems, Proc. of ACL-COLING, 2006
[11] Bárbara , E., Sánchez Rinza, Diana Alejandra Bigurra Zavala, Alonso Corona Chavez, : De-encryption of a text in spanish using probability and statistics. In:18th IEEE International Conference on Electronics, Communication and Computers, 2008, pp 75-77.
[12] Bao-Chyuan Guan, Ray-I Chang, Yung Chung Wei, Chia-Ling Hu, Yu-Lin Chiu, : An encryption scheme for large Chinese texts . In : IEEE 37th Annual 2003 International Carnahan Conference on Security Technology, pp 564- 568. Taipei, Taiwan.
[13] M.H. Shirali-Shahreza , M. Shirali-Shahreza, : Steganography in Persian and Arabic Unicode Texts Using Pseudo-Space and Pseudo-Connection Characters. J. Theoretical and Applied Information Technology (JATIT). 8, 682-687(2008)
[14] M.H. Shirali-Shahreza and M. Shirali-Shahreza, .: A New Approach to Persian/Arabic Text Steganography. In : 5th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2006)., Honolulu, HI, USA.,p310-315.

Authors

Bhadri Raju MSVS received M.Tech. Degree in Computer Science and Engineering from Birla Institute of Technology, Ranchi in 2001. He is working as Associate Professor in Department of Computer Science and Engineering at S.R.K.R.Engineering college, Bhimavaram, Affliated to Andhra University,Visakhapatnam. He has a Teaching experience of over 14 Years. He is pursuing research at JNTuniversity, Hyderabad. His current research focuses on Information security and Language Technologies .

Dr. Vishnu Vardhan B received M.Tech. Degree in Computer Science and Engineering from Birla Institute of Technology, Ranchi in 2001 and Ph.D. in Computer Science from JNTUniversity, Hyderabad in 2008. He is working as Professor in the Department of Computer Science and Engineering at Indur Institute of Engineering and Technology, Siddipet, Affliated to JNTUniversity, Hyderabad. He has a Teaching experience of over 14 Years. His current research focuses on Information Retrieval and Language Technologies . He published more than 10 papers at National and International Journals and conferences.

Naidu G A received the B.Tech. Degree in Computer Science and Engineering from Nagarjuna University in 1997 and M.Tech. Degree in Computer Science and Engineering from AndhraUniversity, Visakhapatnam in 2001. He has a Teaching experience of over 8Years. He is pursuing research at JNTuniversity. His current research focuses on Information security, Computer Networks and Language Technologies . He published five papers in international conferences and Journals.

Dr Pratap Reddy L received B.Tech degree in E.C.E. from Andhra University,Visakhapatnam and M.Tech degree in E.C.E. from Regional Engineering College, Warangal. He received his Ph.D. from JNTUniversity, Hyderabad . He is currently Professor and Head of the Department of Electronics and Communication at JNTU college of Engineering, JNTuniversity, Hyderabad. His research interests include Information Retrieval, Information Security, Character Recognition and Language Technologies. He published more than 40 papers at several National and International Conferences and Journals.

Dr Vinaya Babu A received B.Tech degree in E.C.E. from Osmania University, Hyderabad and M.Tech degree in E.C.E., M.Tech in C.S.E. He received his Ph.D. from JNTUniversity, Hyderabad. He is currently Professor of C.S.E. and Director, Admissions JNTuniversity, Hyderabad. Previously he held several positions as Head of the Department of C.S.E., Director, SCDE at JNTU, Hyderabad. His research interests include Information Retrieval, Compiler Design, Information Security, and Language Technologies. He published more than 40 papers at several National and International Conferences and Journals.