A Noval Security Model for Indic Scripts - A Case Study on Telugu

Bhadri Raju MSVS msramaraju@ gmail.com
Associate Professor in CSE
S.R.K.R.Engineering College
Bhimavaram, A.P., 534 204, India

Vishnu Vardhan B mailvishnu@ yahoo.com
Professor in CSE
Indur Institute of Engg&Tech.
Siddipet, A.P., 534 204, India

Naidu G A apparaonaidug@yahoo.com
Research Scholar in CSE
JNTuniversity Kakinada.
Kakinada, A.P., 534 204, India

Pratap Reddy L prataplr@ rediffmail.com
Professor&Head of ECE
Jawaharlal Nehru Technological University
Hyderabad, A.P., 500 085, India

Vinaya Babu A dravinayababu@yahoo.com
Professor in CSE& Director,Admissions
Jawaharlal Nehru Technological University
Hyderabad, A.P., 500 085, India

Abstract

Secured communication of text information across the world is of prime importance when many languages, several alphabets and various signs (glyphs) found their existence on computing machines. Cryptography is one of the methods to attain security. Existing cryptographic systems divide the text message into words and each word into characters where character is treated as basic unit. For each character, the corresponding bit stream is generated and transformation techniques are applied on blocks of fixed length of bits or bytes. The characteristics of the language like frequency distribution may be reflected in the transformed text also. Correlation between plain text and encrypted text is to be studied from the stand point of text patterns versus symbol patterns. Frequency distribution as a parameter in the process of reverse mapping is mostly dependent on language specificity. If the language is more complex then the retrieved percentage of plain text will be less. In fact the structure and complexity of the underlying language is a multi dimensional extremely important factor when trying to assess an attacker's likelihood of success. On many occasions a large key space does not ensure that a cipher is secure. The Language complexity is to be treated as a parameter. The present work mainly focuses on the characteristics of Indic script in the form of frequency distribution of character code points with a case study on Telugu script. The evaluation is limited to 8-bit key with comparison between Latin text and Telugu text.

Keywords: Cryptography, Language Complexity, Frequency Distribution, Indic Scripts.

1. INTRODUCTION

Cryptography is one way of providing security using the process of encryption and decryption. In general any encryption and decryption scheme uses symmetric key algorithms like DES, RC5, IDEA etc, where each block of fixed size bit stream will be transformed to cipher text or asymmetric algorithms like RSA, Elliptic curve cryptography where a block of bit stream is transformed to an integer equivalent value and encryption techniques are applied. Both these types use either block cipher or stream cipher techniques for text transformation. The main parameters in these schemes are linked with algorithm and key. Providing secured communication for the data is a major and challenging task due to the primary existence of various languages with numerous sets of characters of different properties and behavior. Introduction of Unicode made it possible to represent all the characters in the world irrespective of the language in a unique way. With the increasing importance of localization, there is a need for development of international products to fit onto a region, culture and writing system using Global standards. This idea of localization can also be adopted on information Security which may support multiple languages. In this scenario the script complexity plays a vital role which needs to be considered as an additional parameter. The present paper addresses the information security issues related to Indic scripts with an emphasis on script complexity.

Many scripts of South Asia are derived from the ancient Brahmi script. Indic scripts are derivatives of a common ancestor, which contain scripts that are used for two distinct major linguistic groups, Indo-European languages in the north and Dravidian languages in the south. Linguists describe these types of writing systems as "orthographic", which means that Indic scripts are a mixture phonemic (i.e., where a basic character represents a single phoneme or a a basic unit of word distinguishing sound) and syllabic forms. When a rendering engine works on an Indic script, it usually does the processing from the level of individual syllables. A syllabic unit is a visual unit (glyph) as well. A syllable is formed around a "central" character (usually a consonant), which is known as the "base" character. Syllable is represented using the canonical structure **(C(C))CV** . The syllable may contain usually one to ten single byte character codes of machine. Work on information security till recent past is based on English Text where in there is one to one mapping between character and codes. For each character in the given document generate the bit stream. On the bit stream symmetric or asymmetric key cryptography algorithms are applied. But in today's Global village the algorithms should support data in multiple languages equally and efficiently. A simple logical conclusion is that if the script is more complex then same level of security can be achieved with smaller key size. This paper describes a noval scheme for encrypting Indic scripts with a case study on Telugu using script complexity.

2. LITERATURE REVIEW

Cryptanalysis is the study of a cryptographic system with an emphasis on exploring the weaknesses of the system. Different approaches of cryptanalysis in the literature use language characteristics to understand the strength of cipher system. One such approach deals with frequency statistics. Symbol occurrences in an encrypted message play a key role in the reverse mapping [1] of characters, leading to prediction of plain text. Apart from single character, relation between cipher text and plain text in terms of bigrams and trigrams also play vital role [2]. Single letter frequencies of a cryptogram are identical to that of the plaintext in transposition ciphers. In substitution systems, each plaintext letter has one cipher text equivalent. The cipher text letter frequencies may not be identical to the plaintext frequencies always, but the same count will be present in the frequency distribution as a whole. K.W. Lee et.al proposed [3] the cryptanalytic technique of enhanced frequency analysis. This technique uses the combined techniques of monogram frequencies, keyword rules and dictionary checking. The proposed three-tier approach reported to be a mechanized version of cryptanalysis of mono alphabetic simple substitution

cipher. Thomas Jakobsen proposed [4] a method for fast cryptanalysis of substitution ciphers. This method explored the knowledge of digram distribution and their mapping in the cipher text.

At present cryptanalysis activity is extended to determination of the language being used, determination of the system being used which involves character frequency distribution, searching for repeated patterns and performing statistical tests, reconstruction of the system's specific keys and reconstruction of the plain text. Recent approaches [5] in literature are being concentrated on retrieval of plain text, based on the features of the respective language. Certain language characteristics are to be identified for successful cryptanalysis. Extensive statistical analysis of frequency distribution of characters is an additive knowledge while retrieving part of plain text message.

Bárbara E. et al presented a method [6] 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 [7] 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 number 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 [8] with regard to shapes and their position in the word.

3.SECURITY MODEL

Every language has certain evaluation parameters in such a way that language primitives are used in the construction process of document. This phenomenon is used for understanding the complexity of the language. These meaningful units are the representative set of language primitives. In case of English the language primitives are represented with the help of one-to-one correspondence between characters and machine codes. Syllables are the primitives in Indic scripts and they are represented in the form of canonical structure (C(C(CV))). Machine representation of canonical structure results in a set of variable length of code points ranging from one to nine. These units are transformed with the help of crypto system. The transformation is done onto a different plane where the mapping is a reversible phenomenon. The correlation between the encrypted units and character code points is the main focus while analyzing the strength of the crypto system.

The proposed model defines meaningful units that are embedded in text documents. Text documents compose of sentences, words and primitive 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 a one-to-one mapping and it is a simple byte stream. So the present model has addressed that specificity by taking into consideration of segmentation of words into syllables and extraction of byte stream from the syllables. They will be transformed into a code point byte stream and that byte stream is again converted into bit stream which undergoes transformation similar to that of any system as presented in FIGURE 1. Analysis of this is a complex phenomena which is taken care in the present work.

A key stream is generated using efficient Random number generator. With this key stream, transformation techniques are applied on this bit stream resulting in cipher text. For decryption the cipher text is converted to bit stream which in turn is converted into code point streams. These code point streams are converted to syllables then words and sentences. The algorithm for encryption and decryption is as explained below.

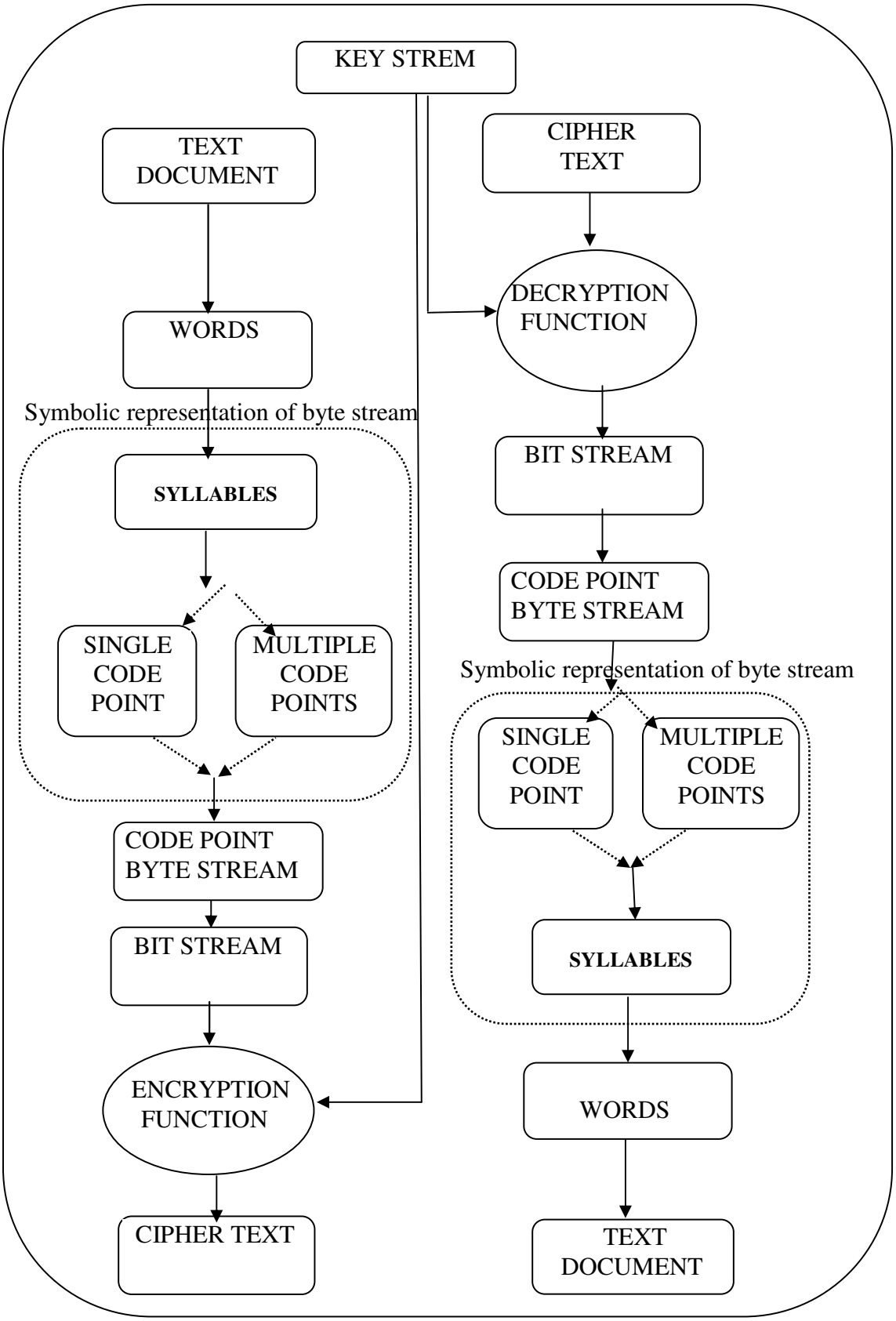

FIGURE 1. Flowchart of encryption and decryption for Indic scripts.

Algorithm for Encryption of Indic Scripts

1 : Divide the given text document into set of words.

2 : Divide each word into syllables (which is basic unit).

3 : For each syllable generate the character code point byte stream which may consists of single or multiple code points that will form that syllable.

4 : Generate bit stream for the byte stream generated in step3.

5 : Apply Encryption technique on the bit stream generated in Step 4 and a key stream generated randomly which results in the cipher text.

6 : Repeat steps 3 to 5 for each syllable generated in step2.

Algorithm for Decryption of Indic Scripts

1 : Generate bit stream for the cipher text.

2 : Apply Decryption technique on the bit stream generated in Step 1 with a key stream generated during encryption resulting in a byte stream.

3 : Combine the bit streams of step2 to form code point byte stream.

4 : Combine the code point byte stream of step3 to form syllables

5 : Combine the syllables to form words and the words into text document

6 : Repeat step 1 through 5 for all byte streams in the cipher text.

The process of encryption and decryption applied on a sample English and Telugu document is as shown below in Figure 2 and Figure 3.

4. FREQUENCY DISTRIBUTION OF TELUGU SCRIPT

Telugu text is syllable based where syllable is the basic unit. The canonical structure defined in ISCII/Unicode is ((C)C)CV. In Telugu the first consonant forms the CV cluster and the other consonants after this cluster appear in dependent form. Basic structure [9] deals with vowels, consonants and characters with consonant + vowel sign. The other characters are coded with the help of these three groups plus special signs Virama, Anuswara and Visarga. Each syllable may have single or multiple code points and the possible groups of syllables with an example is shown in Table1.

The following example illustrates the process of transforming a syllable into code points stream. Consider the word NEWZELAND in English which can be written in Telugu as □□□□□□□□□□□□. □□□□□□□□□□□□ consists of four syllables □□□□, □□, □□□, □□. The syllable □□□□ consists of the code points of □, □, □ which are 0C28, 0C2F, 0C0A respectively. If we consider NAYAGARA in English it can be written in Telugu as □□□□□□ that consists of four syllables □, □□, □, □□ . The character □ which contains only one code point 0C28 in NAYAGARA is different from that of NEWZELAND where the meaningful relation between code points within the syllable are different.

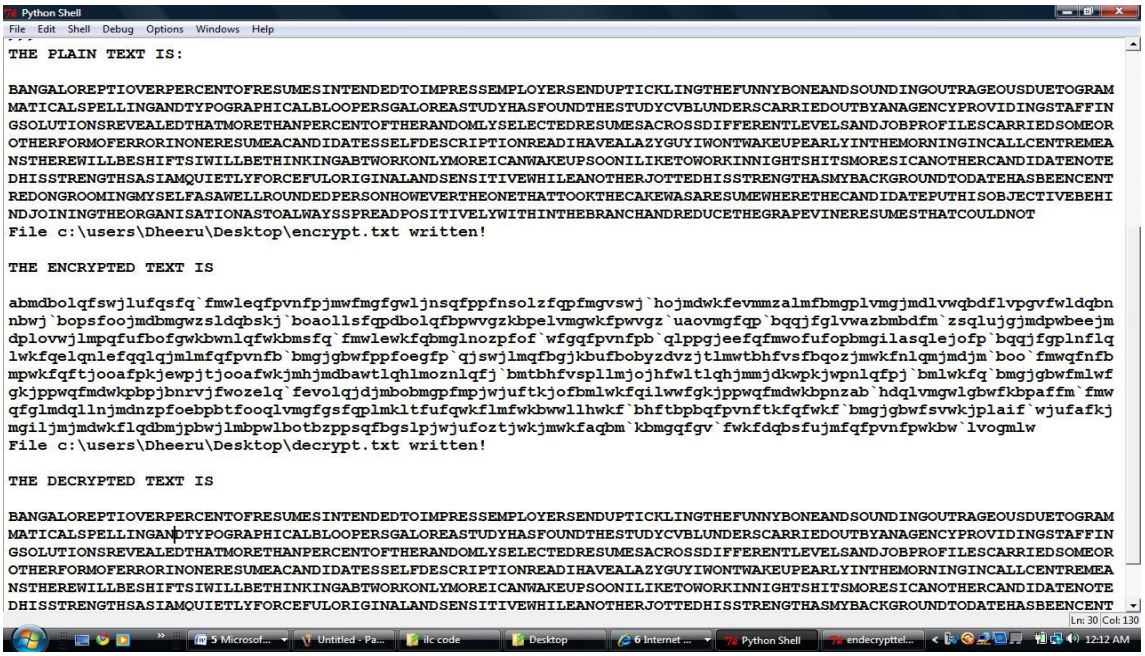

FIGURE 2 : Sample Plain Text, Encrypted Text and Decrypted Text in English

FIGURE3 : Sample Plain Text, Encrypted Text and Decrypted Text in Telugu

Syllable	Example	Code points in Unicode	Code Points
C	□	\U0C15	Single code point
V	□	\U0C05	
CV	□□	\U0C30 \U0C3E	Multiple code points
CCV	□□□□	\U0C24 \U0C4D \U0C15 \U0C3E	
CCCV	□□□□□	\U0C15 \U0C4D \U0C37 \U0C4D \U0C2E	
Dead Consonants	□□	\U0C17 0C4D	

TABLE-1 : Unicode Code Points for Telugu Syllables

Syllable is a complex structure in Indic scripts. The abstract entities(Character code points) are grouped under the influence of grammar rules with specific relation among them resulting into a syllable. Logical combinations of syllables are reported [9] to be as excessive of more than seven hundred thousands. Mapping of these syllables will lead to complex definition of transformation plane. In the present paper we addressed the machine representation units (character code points) for the purpose of analysis. Statistical behaviour of the character code points is limited to frequency distribution and the same is adopted for cryptanalysis. For a simple text like English, when a Substitution Cipher is used with a fixed random key, each specific letter of the alphabet is replaced by the same substituted letter, no matter where it appears in the text. If the frequency of the letters in a message is reflected in the form of a table then the frequencies for the cipher text show the same imbalance but with the frequencies distributed differently amongst the letters. By comparing these frequencies, a cryptanalyst might reasonably guess which alphabet in cipher text maps to the corresponding alphabet in plain text.

The statistical behavior of all characters in English expressed as a percentage of the letters in a sample of over 300,000 characters is evaluated in [10]. They show, quite clearly, that English text is likely to be dominated by a very small number of letters. When text in Telugu is considered, the following Table 2 shows the frequencies expressed as a percentage of the character code points of the alphabet in a sample of over 2,400,000 characters taken from passages from numerous newspapers, novels, stories, songs, sports and literature etc. The reason for certain frequencies in column1 of above table to be zero is that they are the deprecated characters in the usage of the language. The zero frequencies in column 3 represent the numbers from 0 to 9 in Telugu language which are not used in colloquial language. An interesting phenomenon is observed in the frequency distribution of character code points. The highest frequency of 1% among vowels is associated with the vowel □ \U 0C05. All other vowels are observed with the frequency less than or equal to 0.5%. Among consonants the highest frequency of 6.2% is associated with the consonant □. Only four consonants are observed with frequency greater than 4%. Among vowel signs, only three of them are observed with frequency around 7%. This phenomena is more associated with CV Core which are reported [9] with 54% in the syllable structure. The Nasal symbol □ is observed with 4.7% frequency and the highest frequency of 8.5% is associated with Halant □. It is quite interesting to know that Halant is not treated as a syllable at all. However the significant roll of Halant is observed in the conjunct formations of syllables. The statistical behavior of these code points are adopted for the cryptanalysis as described in section 5.

5. CRYPTO ANALYSIS USING FREQUENCY DISTRIBUTION

The proposed cryptographic model is tested initially on two languages i.e. English and Telugu . The encryption algorithm is implemented on text of different sizes in Telugu. For this process a key is generated randomly using a OS based random generator. The plain text is encrypted using the proposed algorithm and randomly generated key resulting in cipher text. The frequencies of different characters in the cipher text are calculated and the results are tabulated. Mapping is done 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 plain text retrieved is calculated which is illustrated in Figure 4 and Figure 5. 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 distinct point in the orthogonal plane. On many occasions for large text size almost all characters are present. Even for a medium sized text this is true because of less number of characters that exist. More over because of one-to-one mapping predictability is more. The percentage of retrieved code points is calculated using frequency distribution. If we consider Telugu script the number of character codes that exist in the original text need not be the complete set. Even though the mapping function takes care of one to one correspondence, 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 threshholding function in the crypto analysis process for reverse mapping. The percentage of plain text that can be retrieved is observed in the range 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.

□	0	□	0.2	□	0.0	□	0.4	□	0.5
□	4.7	□	0.1	□	0.4	□	3.2	□	1.9
□	0	□	0.0	□	3.5	□	0.7	□	0.1
□	1.0	□	4.4	□	0.2	□	0.6	□	8.5
□	0.5	□	0.1	□	2.7	□	2.6	□	0
□	0.3	□	1.9	□	0.5	□	0.5	□	0
□	0.2	□	0.1	□	6.2	□	6.8	□	0
□	0.3	□	0.0	□	3.2	□	7.8	□	0
□	0.1	□	2.3	□	0.1	□	1.3	□	0
□	0.0	□	0.1	□	0.7	□	6.6	□	0
□	0	□	0.7	□	0.5	□	0.8	□	0
□	0	□	0.0	□	2.7	□	0.2	□	0
□	0	□	0.0	□	2.1	□	0.0	□	0
□	0.3	□	1.9	□	5.3	□	1.3	□	0
□	0.1	□	0.1	□	0.0	□	2.2	-	-
□	0.1	□	1.9	□	4.7	□	0.4	-	-

TABLE-2 : Frequency distribution of character code points of Telugu script

The percentage of plain text that is retrieved varied in the range from 25% to 50% depending on the size of the plain text which is illustrated in Table3 . This result in case of Telugu is relatively less when compared to English which is due to large amount of complexity in Telugu script.

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

TABLE3: Percentage of retrieved character code points using frequency distribution

From the above table, it is easy to infer that cryptanalysis of text of complex languages like Telugu is much more difficult . On an average the percentage of plain text retrieved in case of English is 39.11 where as in case of Telugu it is only 12.13%. Then the larger key size applicable to Latin text can be reduced in case of complex languages like Telugu even by providing greater level of security. The percentage of plain text retrieved is not linear with text size because a proper threshold function is required to map cipher text symbols to corresponding plain text symbols for which the work is in progress.

FIGURE4: Retrieved Text based on Frequency distribution in English

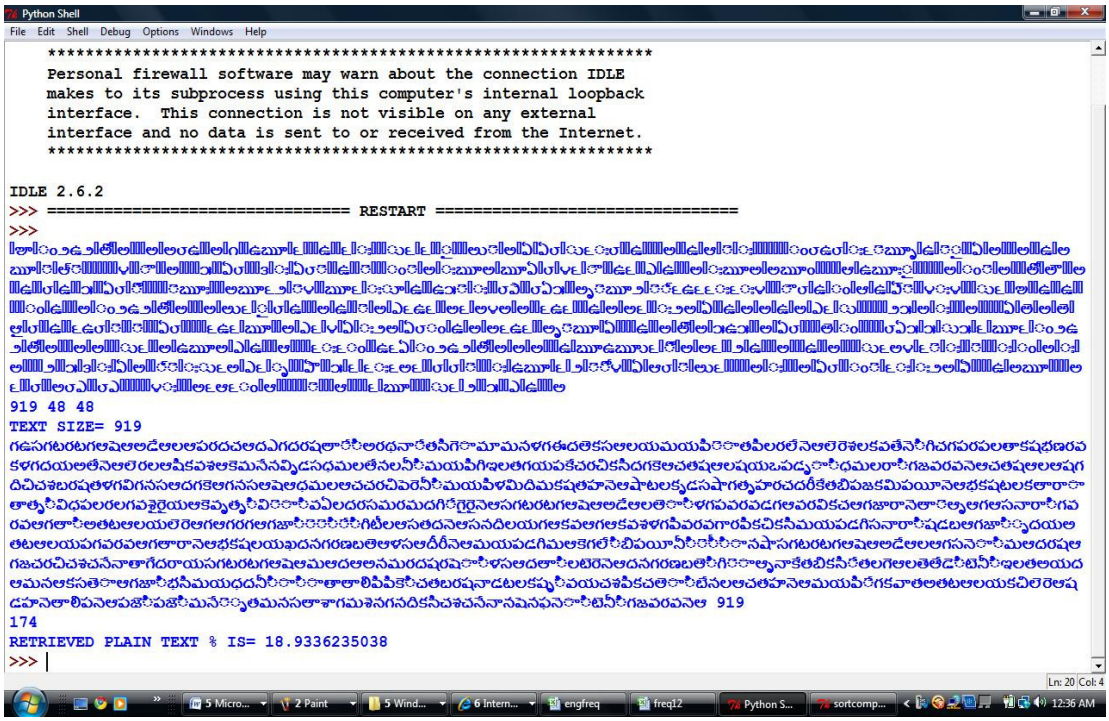

FIGURE5: Retrieved Text based on Frequency distribution in Telugu

6. CONCLUSIONS

In the present work an attempt is made to provide a frame work for security model with specificity on Indic scripts. In this context syllable formation is considered as the basis for the present analysis where the canonical structure is emphasized. These syllables are later transformed into code point byte stream. The encryption and decryption process is tested and compared with English text. An extensive evaluation of character code points and their frequency distribution is carried out on a large sample set. The cryptanalysis of the model is carried out with the help of frequency distribution of character code points as a parameter of interest. From the observed results it is easy to infer that the reverse mapping is more complex in case of Indic scripts (specific reference to Telugu). The statistical behavior of syllables and their influence on security model is in progress.

7. REFERENCES

1. Antoine Joux. *"Algorithmic Cryptanalysis"*, CRC Press (2009)
2. Mark Stamp. *"Information security Principles and Practice",* JohnWiley & Sons. Inc (2006)
3. 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 pp. 1-7 (2006)
4. Jakobsen, T., "*A fast Method for Cryptanalysis of Subsittution Ciphers*". Cryptologia, XIX, 3: 265-274(1995)
5. Majdi Al-qdah & Lin Yi Hui "*Simple Encryption/Decryption Application"* International Journal of Computer Science and Security, Volume 1 : Issue 1(2007)
6. Bárbara E. Sánchez Rinza, Diana Alejandra Bigurra Zavala, Alonso Corona Chavez, "*De-encryption of a text in spanish using probability and statistics"* In the Proceedings of the 18th IEEE Conference on Electronics, Communications and Computers,pp 75-77 (2008)
7. Bao-Chyuan Guan; Ray-I Chang; Yung Chung Wei; Chia-Ling Hu; Yu-Lin Chiu, *"An encryption scheme for large Chinese texts"* , In Proceedings of the IEEE 37th Annual 2003 International Carnahan Conference on Security Technology , Taipei, Taiwan, ROC, pp 564- 568. (2003)

8. M.H. Shirali-Shahreza , M. Shirali-Shahreza, "*Steganography in Persian and Arabic Unicode Texts Using Pseudo-Space and Pseudo-Connection Characters*". Theoretical and Applied Information Technology (JATIT). 8, pp 682-687(2008)
9. Pratap Reddy, L.,: "A New Scheme for Information Interchange in Telugu through Computer Networks " : Doctoral Thesis. JNTU,Hyderabad, India, (2001)
10. H. J. Beker and F. C. Piper.*" The Protection of Communication"* Cipher Systems(2002)