Frequency analysis of Kinyarwanda natural language

BIZIMUNGU Theogene[1] (Assistant Lecturer)

Department of Computer Science, School of ICT, University of Rwanda

Abstract

Cryptanalysis is the study of frequency of letters in encrypted message. It can allow an attacker to break unknown message without knowing a key. There are many ways of breaking cipher text if the used cipher is known. Frequency analysis is used in substitution cipher if and only if the table of distribution of letters is known. This paper aims to present the frequency distribution of letters of Kinyarwanda natural language.

Keywords: *cryptanalysis, cipher text, cipher, frequency distribution, key*

I. INTRODUCTION

In [1], cryptography is a science of secret writing communication is an act that links two people or objects (Sender and Receiver). When a sender wants to send his/her message, the message has to be encrypted using a key and then Ciphertext is sent through insecure channel. The receiver has to decrypt the ciphertext in order to get the plaintext. The knowledge of frequency of letters can facilitate an attacker to break unkown message transmitted over insecure channel. Rwanda is a country located in East Africa; it has three official languages which are described as follows: Kinyarwanda, French and English [2]. Rwanda took Information and Communication Technology (ICT) as key factor in its development [3]. Kinyarwanda is the native languages thus it is the popular language. Kinyarwanda natural language consists of 24 letters which are the following: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, v, w, y and z. A proverb is a short pithy saying in general use, held to embody a general truth [4]. According to [5], proverbs have a great impact on the language. Proverbs of Kinyarwanda are most used daily reason why we selected them in order to see how the letter as are repeated. Greetings are an important part of the communicative competence necessary for being a member of any speech community [6].

This paper is organized as follows: after introducing this paper, section 2 contains terms and terminologies used in this paper; section 3 gives the research question of this paper; section 4 presents the methods used in order to analyze our data and finally the last section concludes this paper.

II. TERMS AND TERMINOLOGY

Sender is a person who wishes to send information

Receiver is a person or object who receives the information

Plaintext is an information a sender wishes to transmit to a receiver.

Ciphertext is a unkown message which can be transmitted after being encrypted.

Encryption is the process of converting plaintext to ciphertext using a cipher and a key

Decryption is a process of converting ciphertext into plaintext using cipher and a key

Attack is an an assault on system security that derives from an intelligent threat that is an intelligent act that is a deliberate attempt (especially in the sense of a method or technique) to evade security services and violate the security policy of a system.

Attacker is a person who is not allowed to view the information but he tries to use his techniques in order to break the ciphertext; he may read only, read and modify the message.

III. RESEARCH QUESTION

Our research aims to answer the following research question:

What and how is the frequency analysis of characters in Kinyarwanda?

IV. METHODOLOGY AND DATA ANALYSIS

In order to answer our research question we conducted the following techniques:

Literature Review has been used in order to know what has been done by others in this field.

C++ programming language has been used for writing a program in which we calculated the frequencies of letters within a given text. First of all, The distribution of letters in a language has been obtained by counting characters of different words. we summed the total count of each letter from a to z and its corresponding frequency has been computed.

We computed 38 proverbs, 102 proper nouns, 30 district names and 4 province names, 2 pages from a document and greetings.

Rwanda has 4 provinces and Kigali city which is the capital; It has 30 districts, we computed 30 districts names and 5 provinces to see the frequency of letters. Greeting a person is very important and it is a daily activity. You can greet in morning, afternoon, evening or night. We computed different greetings of Kinyarwanda.

In the figure 4.1, 38 proverbs have 936 letters; **a,i,u,n,r** are most repeated letters and c,f,v,j,p,l are rare. Letter **L** has 0 frequency.

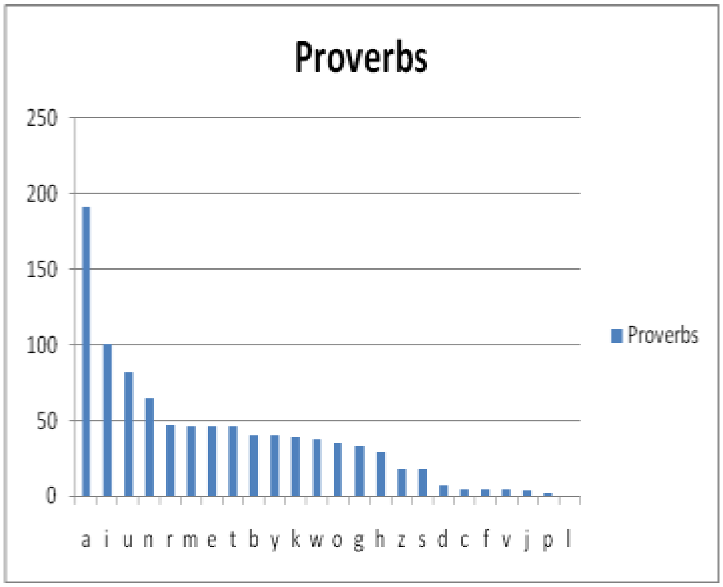

Figure 4.1 Proverbs_Frequency analysis

The Figure 4.2 presents the distribution of letters based on District names and provinces. 279 characters have been computed. **a,r,u,n,g** are most repeated whereas **p,w,d,v,l** are rare.

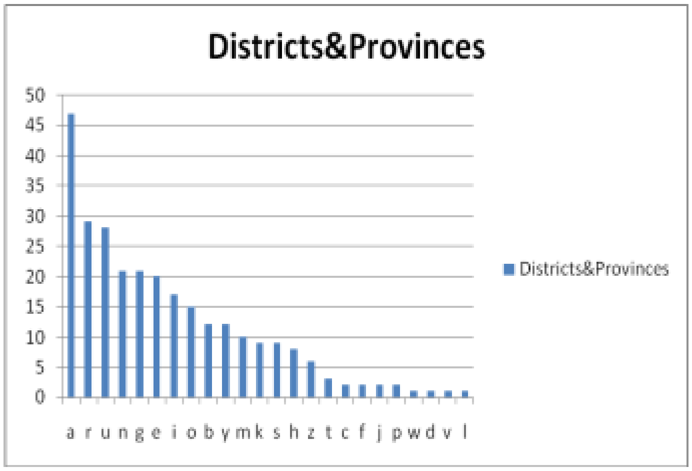

Figure 4.2 District_Frequency

Figure 4.3 displays the frequency of letters based on text selected in e-document[7]; 717 letters have been computed; **a,i,u,n** and **r** are the most common letters and **f,j,v,p** and **l** are the least used letters.

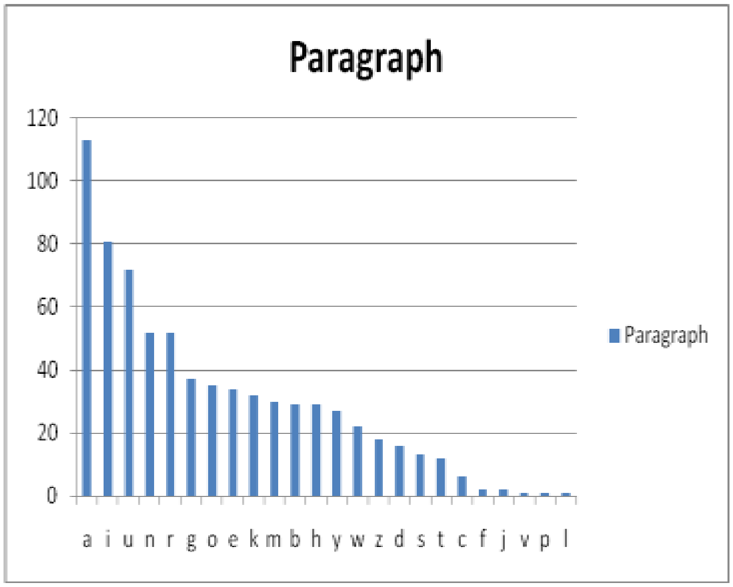

Figure 4.3 Text_Frequency letter

Figure 4.4 presents the distribution of letters from 102 propernames; 888 characters have been tested; **a,i,n,m** and **u** are most repeated and **v,f,p,c** and **l** have the low frequencies.

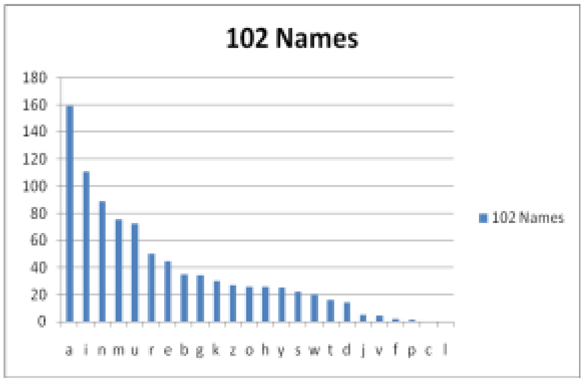

Figure 4.4 Name_Frequency

The Figure 4.5 consists of the distribution of letters based greetings. 77 characters have been analyzed. **a,r,m,i** and **u** have high frquencies and **c,f,v,p** and **l** have the low frequencies.

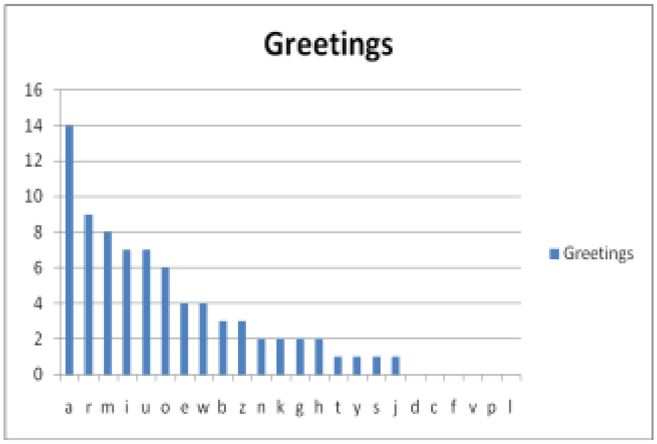

Figure 4.5 Greeting Frequency

V. CONCLUSION

Since Kinyarwanda has 24 letters, cryptanalysis can be applied on Z_{24} instead of using Z_{26} as English or French. In Table 4.1, 2897 characters have been computed; the total has been obtained by adding count of each letter on proverbs, Districts and Greetings. **Tot=** $\sum_{k=0}^{n} x^{k}$ where **x** represents count of a letter in proverbs, districts or greetings, **k** represents 0 to 4 because there are 5 parameters(proverbs, Districts & Directions, Paragraph, Proper Names, Greetings), and **n** is the highest value. The frequency has been computed as follows **f=**$\frac{1}{n}x$ where x represents the total for each character and n represents the general total of all letters. Finally **%=f*100** where **%** represent a percentage of every character.

In Table 4.1, 2897 characters have been computed,character **a** represents 18.1 %, **i** represents 10.9% , **u** is 9 %, **n** is 7.5 % and **r** is 6.5%. **a, i,u,n** and **r** cheracters represent 52%. j,c,f,v,p and l are not most repeated, they are rare.

Table 5.1: Letter distribution

No	**Letter**	**Proverbs**	**Dist&Prov**	**Paragraph**	**102 Names**	**Greetings**	**Total**	**%**
1	A	191	47	113	159	14	524	18.1
2	I	100	17	81	111	7	316	10.9
3	U	82	28	72	72	7	261	9.0
4	N	65	21	52	89	2	229	7.9
5	R	47	29	52	50	9	187	6.5
6	M	46	10	30	75	8	169	5.8
7	E	46	20	34	45	4	149	5.1
8	G	33	21	37	34	2	127	4.4
9	B	40	12	29	35	3	119	4.1
10	O	35	15	35	26	6	117	4.0
11	K	39	9	32	30	2	112	3.9
12	Y	40	12	27	25	1	105	3.6
13	H	29	8	29	26	2	94	3.2
14	W	37	1	22	20	4	84	2.9
15	T	46	3	12	16	1	78	2.7
16	Z	18	6	18	27	3	72	2.5
17	S	18	9	13	22	1	63	2.2
18	D	7	1	16	14	0	38	1.3
19	J	3	2	2	5	1	13	0.4
20	C	4	2	6	0	0	12	0.4
21	F	4	2	2	2	0	10	0.3
22	V	4	1	1	4	0	10	0.3
23	P	2	2	1	1	0	6	0.2
24	L	0	1	1	0	0	2	0.1
Total		936	279	717	888	77	2897	

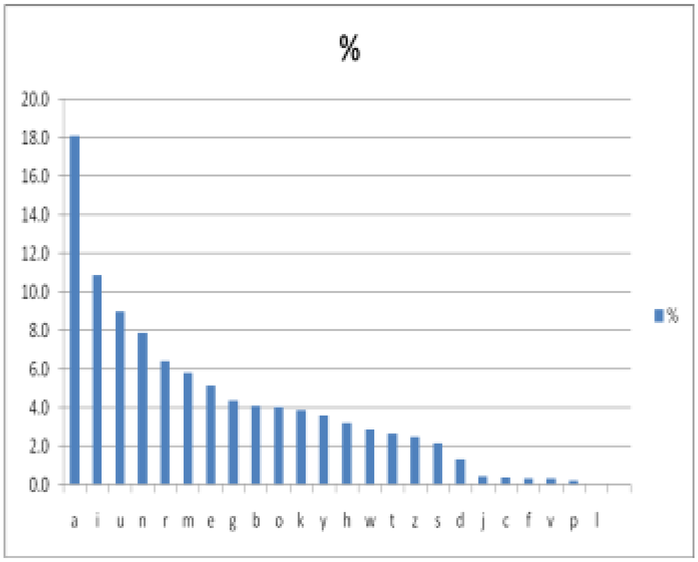

Figure 5.1: Frequency of letters

REFERENCES

[1] *Mihir Bellare and Phillip Rogaway. Introduction to modern cryptography. Notes, 1996-2004*

[2] *http://www.gov.rw/home/geography visited on 10/10/2015*

[3] Rwanda National ICT Strategy and Plan,NICI III, 2010-2015

[4] *Judy Pearsall and Bill Trumble, "Oxford English Reference Dictionary",New Delhi, 2008*

[5] *Outi Lauhakangas,"use of proverbs and narrative though",2007*

[6] ***Alessandro Duranti,*** *usniversal and culture-specific properties of greetings, university of California,1977*

[7] http://ralc.gov.rw visited on 10 october 2015