Amharic language letter frequency distribution

Samson Endale Mergissa
Crypo analysis (IP team)
Information Network Security Agency
Addis Ababa, Ethiopia
samson.endale@insa.gov.et, 4sam21@gmail.com

***Abstract*—One of the first and famous attacks on classical cryptographic algorithms is frequency analysis. It requires the letter frequency of the given language available to implement this attack. Even if an Amharic text is encrypted with a simple mono-alphabetic substitution, we couldn't break the cipher with frequency analysis because the Amharic language never been analyzed to learn the letter frequency distribution. I have tried to address this problem by collecting and analyzing a relatively large dataset of Amharic texts. This analysis and findings are done for the first time for this language and I have uncovered fascinating statistical distributions of the language.**

***Index Terms*—Amharic, frequency distribution, cryptanalysis**

I. Introduction

A. The Amharic language (አማርኛ)

Amharic is a Semitic language and the national language of Ethiopia (ኢትዮጵያ). The majority of the 25 million or so speakers of Amharic can be found in Ethiopia[5], but there are also speakers in a number of other countries, particularly Eritrea , Canada, the USA and Sweden. Amharic is written with a version of the Ge'ez script known as ፊደል (Fidel)[1]. Amharic is the second-most commonly spoken Semitic language in the world, after Arabic. Amharic is written left-to-right using a system that grew out of the Ge'ez script, called, in Ethiopian Semitic languages, Fidäl (ፊደል, "writing system", "letter", or "character" or abugida (አቡጊዳ from the first four symbols, which gave rise to the modern linguistic term abugida. It has been the working language of courts, language of trade and everyday communications, the military, and the Ethiopian Orthodox Tewahedo Church since the late 12th century and remains the official language of Ethiopia today[10]. Most of the Ethiopian Jewish communities in Ethiopia and Israel speak Amharic. In Washington DC, Amharic became one of the six non-English languages in the Language Access Act of 2004, which allows government services and education in Amharic. Furthermore, Amharic is considered a holy language by the Rastafari religion and is widely used among its followers worldwide[4]. It is the most widely spoken language in the Horn of Africa. The Amharic script is an abugida, and the graphemes of the Amharic writing system are called fidel. Each character represents a consonant+vowel sequence, but the basic shape of each character is determined by the consonant, which is modified for the vowel. This is because these fidel originally represented distinct sounds, but phonological changes merged them. The citation form for each series is the consonant+ä form, i.e. the first column of the fidel. The Amharic script is included in Unicode, and glyphs are included in fonts available with major operating systems[7].

B. Frequency analysis

Among so many cryptanalytic techniques, frequency analysis or frequency count is the most basic one other than brute-force, threat, blackmail, torture, and bribery. The frequency analysis is, in fact, the anatomy of a language. According to a book "Trattati in cifra" published in 1470 and written by Leone Battista Alberti, who is known as "Father of Western Cryptology", the aspect of cryptanalysis using frequency analysis can be traced back to Al-Kindi, who is "The Philosopher of the Arabs" and author of 290 books on medicine, astronomy, mathematics, linguistics, and music. In 1987, the Arabic scientist Al-Kindi's treatise was discovered in the Sulaimaniyyah Ottoman Archive in Istanbul and entitled "A Manuscript on Deciphering Cryptographic Message". It is believed that this manuscript is the first ever known oldest description of cryptanalysis by frequency analysis[8].

C. Other applications

Knowing the frequency distribution is essential for developing optimal encoding schemes for communication mechanism like Telegraph by Morse code. To increase the speed of the communication, the code was designed so that the length of each character in Morse is approximately inverse to its frequency of occurrence in English. Thus the most common letter in English, the letter "E", has the shortest code, a single dot[3].

Additionally, this analysis is essential for developing a keyboard layout with an efficient hand movement. Letter frequencies had a strong effect on the design of some keyboard layouts. The most frequent letters are on the bottom row of the Blickensderfer typewriter, and the home row of the Dvorak Simplified Keyboard[2][12].

Furthermore, aforementioned also heavily contribute to the field of optical character recognition and in linguistics.

II. Background

For almost a thousand years, from 500 CE to 1400 CE, the cryptology of Western civilization stagnated. The systems used were extremely simple and more or less derivations of substitution ciphers and steganography. However the Arabs were the first to discover the importance of cryptanalysis in the 9th century CE. Till this time only cryptography existed

and of any science of cryptanalysis there was nothing. Cryptanalysis is the science of unscrambling a message without the knowledge of the key and is based on finding weaknesses in encryption methods in order to break them. The Arabs had the best conditions for inventing cryptanalysis because they had reached a high level in several disciplines, including mathematics, statistics and linguistics. Theological schools were established where the contents of the Koran were studied in detail. Theologians tried to extract a chronological order of the numerous revelations and therefore they counted the frequency of specific words in every single revelation because some words arose earlier in comparison to other words. They continued to examine the scriptures phonetically and at the level of single letters and found out, that some letters occur much more frequently than other ones and which letters go or do not go together. They realized the rarest letters in Arabic and the most common letters: the letters 'a' and 'l' are the most common in Arabic, whereas the letter 'j' appears only a tenth as frequency. This apparently innocuous observation would lead to the first great breakthrough in cryptanalysis, namely frequency analysis. It is unknown who first realized that the variation in the frequencies of letters could be exploited in order to break ciphers, but the earliest known description comes from the 9th century scientist Abū-Yūsuf Ya'qūb ibn Ishāq al-Kindī. Al-Kindi has written about 290 books on medicine, astronomy, mathematics, linguistics and music. He also is the author of 'A Manuscript on Deciphering Cryptographic Messages'. It contains detailed discussions on statistics, Arabic phonetics and Arabic syntax and describes the system of cryptanalysis in two short paragraphs: "*One way to solve an encrypted message, if we know its language, is to find a different plaintext of the same language long enough to fill one sheet or so, and then we count the occurrences of each letter. We call the most frequently occurring letter the 'first', the next most occurring letter the 'second' the following most occurring letter the 'third', and so on, until we account for all the different letters in the plaintext sample. Then we look at the ciphertext we want to solve and we also classify its symbols. We find the most occurring symbol and change it to the form of the 'first' letter of the plaintext sample, the next most common symbol is changed to the form of the 'second' letter, and the following most common symbol is changed to the form of the 'third' letter, and so on, until we account for all symbols of the cryptogram we want to solve*" [13].

III. Methodology

A. Data Gathering

The first route I took was to collect various religious texts. I successfully manage to find more than 800,000 characters suitable for analyzing. I also tried to gather some Amharic fictions, the constitution, government regulations, and government reports.

The second place I look for gathering the document was around newspapers which publish a soft-copy on their site. It was a good lead with few newspapers but most of the publisher upload the scanned version of the printed newspaper or the encoding was very bogus and I couldn't extract the Amharic letters.

Lastly, I was googling for Amharic documents by hand which was the least efficient way of tackling the problem. At first, I was looking for a search engine with a regular expression support to search for a document containing the Ethiopic Unicode block which is between 0x1200 to 0x1347. However, most of the search engines lack this functionality which was discouraging. After a lot of effort, I found a more effective method.

Analyzing the already available data, I found out a glimpse of the frequency distribution of the language with a fewer dataset. It happens to be "ን" is the most frequent character from the religious texts. The letter "ን" most likely will occur at least once in any Amharic document. This was an alternative for filtering am Amharic documents on the web. So I start searching for **"allintext:ን filetype:pdf"** on Google and the number of the result was promising. Thus I was using as a result of an "Amharic language frequency distribution" of a smaller dataset to gather a larger dataset for "Amharic language frequency distribution". This method help to gather a lot of random documents which is good for diversifying the dataset but I started to notice a few obstacles.

Most of the documents I found were not larger than 5 pages and I noticed some of the documents were Tigrigna, Ge'ez or some other Ethiopian language which use the Ge'ez script. So I start looking for a better method to query the search engines. At last, I fined tuned my search query to **"allintext:ማውጫ filetype:pdf"**. I noticed word ማውጫ usually occurs more frequently in long Amharic documents through observation and it also makes the document more likely an Amharic.

I was also was trying to write a web crawler to index https://am.wikipedia.org but I dismiss it because like most sites, the layout of the website was a static and repeated in every page. I didn't want to "poison" my dataset with a "wrong" frequency.

B. Data preparation and clean up

Before analyzing the PDF format files, I wanted to make sure there was no duplication in the files. I first run **"sha256sum"** command in a Linux terminal and pipeline the output to **"sort"**.

```
> sha256sum * | sort
```

There were some duplicates. Therefore I used a tool called **"fdupes"** to locate and delete duplicated files.

```
\$ fdupes -d *
```

Second I converted all of the PDF files to a plain text using a tool called **"pdftotext"** because plain text files are easier to process than a PDF format. Lastly, I combined the text files into a giant plain file to shave it to my own analyzer program. The combined text contains 2,245,892 lines, 6,150,197 words and 75,143,429 letters. But I later found out only 17.2 million of the letters were a properly encoded Amharic Unicode characters.

C. Analysis

The analysis was a straightforward process. I wrote a python program which opens the documents and counts the various frequencies and dumps the result into a CSV file.

IV. Data and Results

A. Monogram frequency

Monogram frequency counts are most effective on substitution type ciphers such as the caesar cipher, substitution cipher, polybius square etc. It works because any natural language text follows a very specific frequency distribution, which is not masked by substitution ciphers[11].

Rank	Character	Count	Percentage
1	ን	738643	4.28079
2	ት	680710	3.94504
3	የ	638802	3.70216
4	በ	566363	3.28234
5	መ	556450	3.22489
6	ር	455960	2.64251
7	ው	436064	2.5272
8	ተ	427311	2.47647
9	ስ	387069	2.24325
10	ና	347901	2.01625
11	ያ	331325	1.92018
12	ም	321078	1.8608
13	አ	301741	1.74873
14	።	294046	1.70413
15	አ	287540	1.66643
16	ይ	273357	1.58423
17	ች	271122	1.57128
18	ግ	269841	1.56386
19	ገ	265020	1.53592
20	ከ	258721	1.49941
21	ረ	251049	1.45495
22	ማ	245490	1.42273
23	ል	244525	1.41714

Table I
Top 23 - Monogram frequency

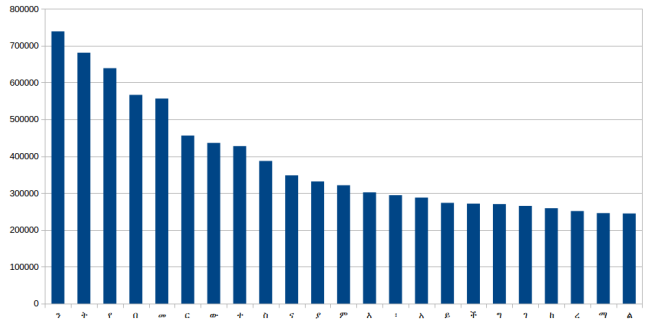


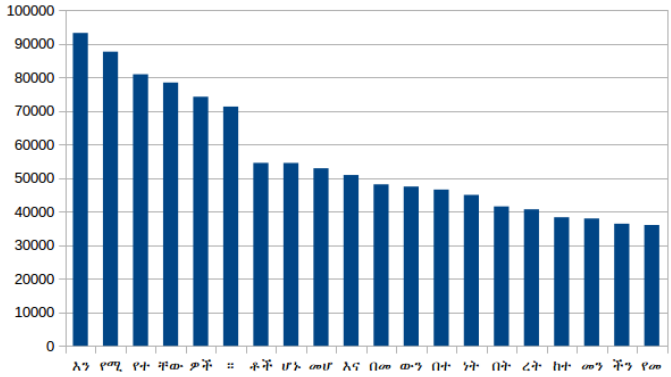
Figure 1. Top 23 - Monogram frequency

B. Bigram frequency

Bigram counts maintain the same principle as monogram counts, but instead of counting occurrences of single characters, bigram counts count the frequency of pairs of characters[11].

	Character	Count		Character	Count
1	እን	93271	11	በመ	48123
2	የሚ	87682	12	ውን	47487
3	የተ	80939	13	በተ	46551
4	ቸው	78480	14	ነት	44970
5	ዎች	74229	15	በት	41547
6	።	71312	16	ረት	40688
7	ቶች	54518	17	ከተ	38304
8	ሆኑ	54498	18	መን	37955
9	መሆ	52918	19	ችን	36395
10	እና	50904	20	የመ	36001

Table II
Top 20 - Bigram frequency

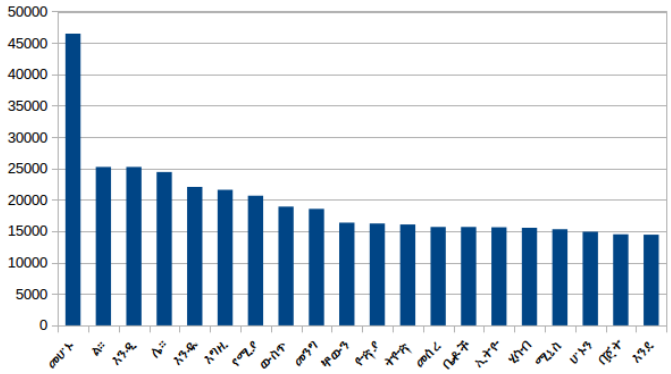


C. Trigram frequency

Just as bigram counts count the frequency of pairs of characters, trigram counts count the frequency of triple characters[11].

	Character	Count		Character	Count
1	መሆኑ	46465	11	ዮጵያ	16220
2	ል ።	25242	12	ትዮጵ	16074
3	እንዲ	25239	13	መሰረ	15673
4	ሌ ።	24421	14	ቤቶች	15665
5	እንዱ	22040	15	ኢትዮ	15617
6	እግዚ	21588	16	ሂሳብ	15541
7	የሚያ	20648	17	ሚኒስ	15304
8	ውስጭ	18908	18	ሆኑን	14916
9	መንግ	18554	19	በጀት	14504
10	ቸውን	16357	20	እንደ	14436

Table III
Top 20 - Trigram frequency

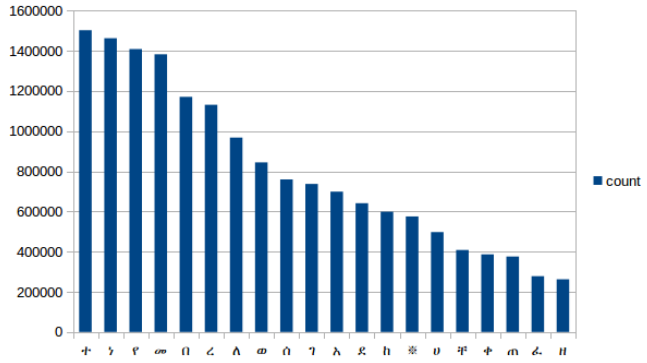


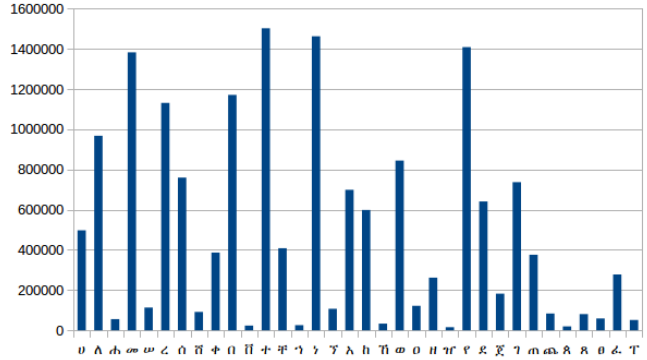
D. Consonant group (’ቤት’) frequency

Consonant group frequency is the frequency of the 34 Amharic consonant variants which is one row from the Ge‘ez script consonants cluster. All of the ’ሀ’, ’ሁ’, ’ሂ’, ’ሃ’ ... variants are grouped together and analyzed[9].

	Character	Count		Character	Count
1	ተ	1502138	11	አ	698619
2	ነ	1462331	12	ደ	640813
3	የ	1408562	13	ከ	598522
4	መ	1382417	14	ሀ	497119
5	በ	1170744	15	ቸ	408120
6	ረ	1131021	16	ቀ	386470
7	ለ	967577	17	ጨ	375543
8	ወ	844325	18	ፈ	277775
9	ሰ	759615	19	ዘ	261978
10	ገ	736809	20	ጀ	182281

Table IV
CONSONANT GROUP (’ቤት’) FREQUENCY



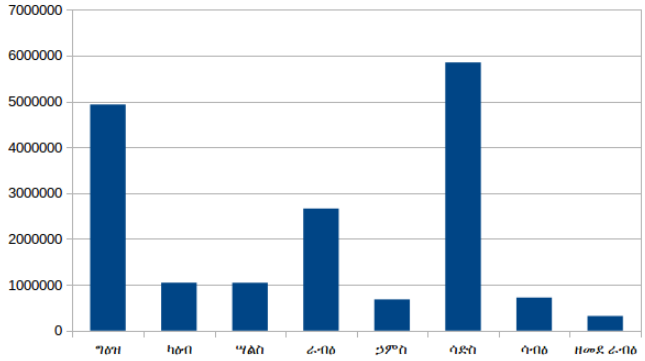


E. Vowel group (’መደብ’) frequency

Vowel group frequency is the frequency of the vowel variants which will modify each consonant (i.e one column from the Ge‘ez script consonants cluster). All of the ’ሳድስ’ variants (’ህ’, ’ል’, ’ሕ’, ’ም’ ...) are grouped together and analyzed[9].

	Character	መደብ	Count
1	6	ሳድስ	5848229
2	1	ግዕዝ	4933207
3	4	ራብዕ	2661791
4	2	ካዕብ	1046439
5	3	ሣልስ	1044916
6	7	ሳብዕ	720252
7	5	ኃምስ	680979
8	8	ዘመደ ራብዕ	318991

Table V
TOP 20 - VOWEL GROUP (’መደብ’) FREQUENCY

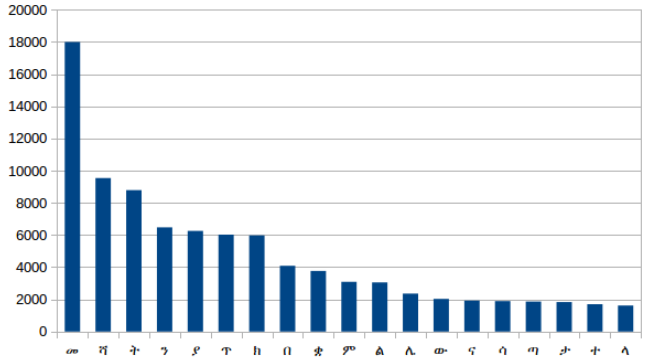


F. Double letter frequency

A double-letter word is one that contains consecutive letters that are the same, such as rabbit, deer or kitten. For each letter of the alphabet, this are the most common double letters in Amharic.

	Character	Count		Character	Count
1	።	71312	11	ም	3091
2	መ	17992	12	ል	3056
3	ሸ	9540	13	ሌ	2362
4	ት	8785	14	ው	2031
5	ን	6474	15	ና	1925
6	ያ	6255	16	ሳ	1891
7	ጭ	6024	17	ጫ	1867
8	ክ	5975	18	ታ	1836
9	በ	4090	19	ተ	1701
10	ቁ	3769	20	ላ	1616

Table VI
TOP 20 - DOUBLE LETTER FREQUENCY

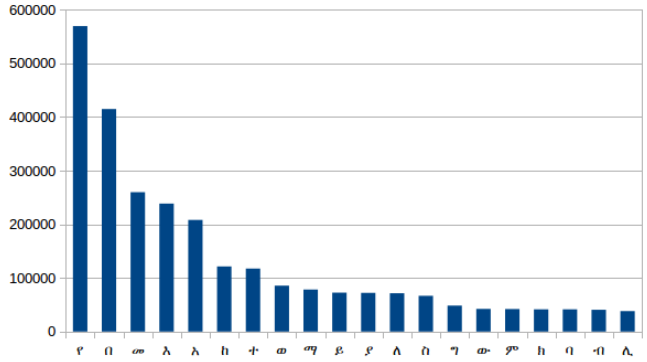


G. First letter frequency

First letter frequency is the frequency of a given Amharic language first letter from a given word.

	Character	Count		Character	Count
1	የ	569324	11	ያ	72231
2	በ	414816	12	ለ	71432
3	መ	259743	13	ስ	66667
4	እ	238565	14	ግ	48456
5	አ	208080	15	ው	42183
6	ከ	121484	16	ም	42005
7	ተ	117446	17	ክ	41459
8	ወ	85590	18	ባ	41416
9	ማ	78348	19	ብ	40596
10	ይ	72597	20	ሊ	38199

Table VII
TOP 20 - FIRST LETTER FREQUENCY

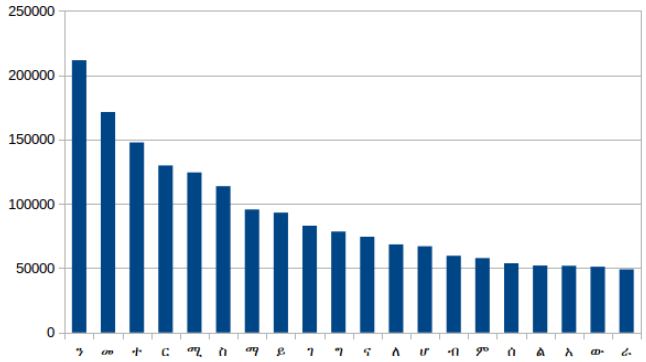


H. Second letter frequency

Second letter frequency is the frequency of a given Amharic language second letter from a given word.

	Character	Count		Character	Count
1	ን	211523	11	ና	74411
2	መ	171288	12	ለ	68440
3	ተ	147704	13	ሆ	66992
4	ር	129813	14	ብ	59655
5	ሚ	124359	15	ም	57804
6	ስ	113693	16	ሰ	53831
7	ማ	95597	17	ል	52029
8	ይ	93176	18	አ	51903
9	ገ	82967	19	ው	51116
10	ግ	78444	20	ራ	49058

Table VIII
TOP 20 - SECOND LETTER FREQUENCY

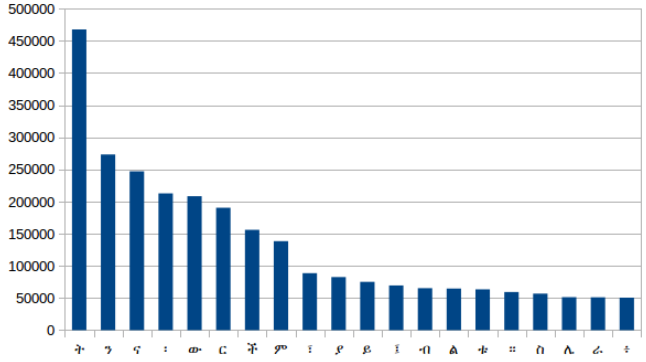


I. Last letter frequency

Last letter frequency is the frequency of a given Amharic language last letter from a given word.

	Character	Count		Character	Count
1	ት	467777	11	ይ	75082
2	ን	273331	12	፤	69485
3	ና	247121	13	ብ	65238
4	፡	212763	14	ል	64701
5	ው	208305	15	ቱ	63441
6	ር	190347	16	።	59351
7	ች	156041	17	ስ	56831
8	ም	138486	18	ሉ	51317
9	፣	88674	19	ራ	51164
10	ያ	82511	20	÷	50495

Table IX
TOP 20 - LAST LETTER FREQUENCY

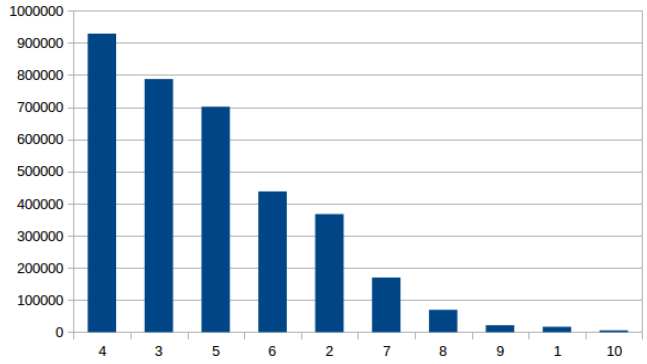


J. Word length frequency

The frequency of the Amharic language words length statistics.

	Length	Count
1	4	928666
2	3	787376
3	5	701324
4	6	437831
5	2	367253
6	7	170080
7	8	69592
8	9	21743
9	1	16734
10	10	5602

Table X
TOP 20 - WORD LENGTH FREQUENCY



V. DISCUSSION

The Amharic language is the second most spoken Semitic language and the official language is Ethiopia which the frequency distribution never been analyzed to acquire its frequency distribution. My analysis has uncovered a lot of statistics which will be instrumental in a lot of fields including but not limited to cryptanalysis, linguistics, and optical character recognition. These findings are self-explanatory but I would discuss some observations I found interesting.

I noticed the top 23 frequent monograms occurs more than 50 percent which means the more than half of every Amharic text is composed of only 23 characters.

The decline of the word separator ':' ('ሁለት ነጥብ') and the use of space like English was also noticeable. Only a few texts I found from religious sites contain texts with the word separator ':'. But still, it's the most frequent punctuation in an Amharic text followed by '፣' (comma), '፤' (semicolon) and ' ። '. Further, I have noticed writers chose to type two consecutive word separator (':') instead of the ' ። ' full stop (period) which has it's own single glyph in Unicode with (0x1362) code point.

Another statics I noticed was that despite the letter ('ን') is the most frequent letter, it's doesn't frequently occur in the first letter of a given word like the letter ('የ'). On the other hand, 83% of the letter ('የ') occurred in the leading (first) position of the word from the total occurrence.

VI. CONCLUSION

In conclusion, anyone can use this results to break an Amharic plain text encrypted with cryptographic algorithms which don't adequately hide the language statistical distribution of the letters. These findings should also be used to develop an efficient keyboard layout which is optimized for speed, number of required key combination and hand movements. The results will also be valuable for researchers and experts developing an Amharic optical character recognition software using this data to prioritize and optimize their algorithm to the most frequent letters. They can also predict when challenged using this result. The possibility is innumerable to use this knowledge, I hope it will be useful for all who finds it.

VII. LIMITATIONS OF THE STUDY

The first limitations I was faced was the lack of a large dataset of processable Amharic text, Which took me more than half of my time doing this analysis. When some texts were found, most of them were just a scanned version of the printed book or newspaper which made it hard for processing.

The lack of publicly available (open source) optical character recognition(OCR) for the Amharic language made it hard to analyze the scanned version of the printed book or newspaper. If OCR was available it was possible to analyze additional datasets which were stored as an image.

One of the analyses I was planning to do was the root word (ስርወ ቃል) frequency which will give a great insight into the most common words we use in the language. Analysing an Amharic word into their constituent morphemes (meaningful parts) is a challenging by itself and out of this paper scope. So I used 'HORNMORPHO 2.5'[6] python package developed by Michael Gasser of Indiana University, School of Informatics and Computing. HORNMORPHO is a Python program that analyzes Amharic, Oromo, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the word's grammatical structure. HORNMORPHO is very slow and requires a high computational capability to analyze millions of words. So, I have excluded root word frequency from this paper.

I also have faced another limitation when I conduct the word length frequency. I have stated in the methodology section, I have used 'pdftotext' conversion tool to convert the documents to a plain file. In some of the texts, I have noticed the tool was not recognizing the space between words and it was combined them into a single word. For example, I have found 50 character length word 'ገምጋሚባልሙያዎችጋርከመቅረባቸዉበፊትበቅድሚያስፈላጊዉሁሉ' which is understandable and can be divided into a properly separated words 'ገምጋሚ ባለሙያዎች ጋር ከመቅረባቸዉ በፊት በቅድሚያ ስፈላጊዉ ሁሉ' manually with ease but hard to automate and integrate into the analysis process. Thus I have removed the larger word length occurrences.

Lastly, I want to point out that more than 50 percent of the text I found was religiously or politically biased. Most of the available processable text out there is biased which made my trigram frequency to not look native.

VIII. Recommendations

I want to urge researchers in the relevant scientific community with a bigger dataset to use the tools I developed, to analyze for better quality results. Whom that with a large computational power use HORNMORPHO[6] package to investigate the root word frequency which holds a valuable statistics.

IX. Acknowledgements

I would like to thank Ermias, Tariku, and Mulat from crypto analysis team for advising, reviewing, and supporting me write this paper. Special thanks for Yilikal Argaw for his support to gather a lot of publicly available Amharic documents.

References

[1] Simon Ager. *Amharic alphabet, pronunciation and language*. URL: https://www.omniglot.com/writing/amharic.htm. (accessed: 30.06.2018).

[2] Nick Baker. *Why do we all use Qwerty keyboards?* URL: https://www.bbc.com/news/technology-10925456. (accessed: 13.07.2018).

[3] Russell W. Burns. *Communications: An International History of the Formative Years*. London: Institution of Electrical Engineers, 2004, p. 84. URL: https://books.google.com.et/books?id=7eUUy8-VvwoC. (accessed: 30.06.2018).

[4] District of Columbia Office of Human Rights. *Language Access Act Fact Sheet*. Tech. rep. 2007. URL: https://ohr.dc.gov/sites/default/files/dc/sites/ohr/publication/attachments/LAAFactSheet-English.pdf. (accessed: 30.06.2018).

[5] Population Census Commission. *Population and Housing Census 2007 Report, National*. Report. Central Statistical Agency, 2010, p. 91. URL: http://www.csa.gov.et/census-report/complete-report/census-2007?download=189:national-statistical. (accessed: 30.06.2018).

[6] Michael Gasser. "HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya". In: *HornMorpho* (2011). URL: ftp://ftp.cs.indiana.edu/pub/gasser/hltd11.pdf. (accessed: 30.06.2018).

[7] Grover Hudson. *The World's Major Languages*. London: Routledge, 2009, pp. 594–617.

[8] Y.L. Tan K.W. Lee C.E. Teh. "Decrypting English Text using enhanced frequency Analysis". In: *Proceedings of National Seminar on Science, Technology and Social Sciences*. Melaka, Malaysia, 2006, pp. 1–7. URL: https://pdfs.semanticscholar.org/5013/213db36162f6e5ae6d2f1c72fe8240adb567.pdf. (accessed: 30.06.2018).

[9] Lawrence Lo. *Ancient Scripts: Ethiopic*. URL: http://www.ancientscripts.com/ethiopic.html. (accessed: 13.07.2018).

[10] Ronny Meyer. "Amharic as lingua franca in Ethiopia". In: *Lissan* XX, No. I/II (2006), pp. 118–131. URL: https://www.academia.edu/5514187/Amharic_as_lingua_franca_in_Ethiopia. (accessed: 30.06.2018).

[11] *Monogram, Bigram and Trigram frequency counts*. URL: http://practicalcryptography.com/cryptanalysis/text-characterisation/monogram-bigram-and-trigram-frequency-counts/. (accessed: 13.07.2018).

[12] Jan Noyes. "The QWERTY keyboard: a review". In: *International Journal of Man-Machine Studies* 18, Issue 3 (1983), pp. 265–281.

[13] Sabrina Schönhart and Armin Müller. *The Breakthrough of Frequency Analysis*. URL: http://cs-exhibitions.uni-klu.ac.at/index.php?id=279. (accessed: 13.07.2018).