Submission -1
Factual analysis on Tamil Alphabets
(P.S. Prabhakar , San Jose , California)

Introduction

From wikipedia "*Tamil spoken by the Tamil people of India and Sri Lanka, and by the Tamil diaspora, Sri Lankan Moors, Chindians, and Douglas. Tamil is an official language in three countries: India, Sri Lanka and Singapore.In India, it is the official language of the Indian state of Tamil Nadu and the Union Territory of Puducherry. Furthermore, Tamil is used as one of the languages of education in Malaysia, along with English, Malay and Mandarin.Tamil is spoken by significant minorities in the four other South Indian states of Kerala, Karnataka, Andhra Pradesh and Telangana and the Union Territory of the Andaman and Nicobar Islands. It is one of the 22 scheduled languages of India.Tamil is one of the longest-surviving classical languages in the world*."

TAMIL , the one word Tamil cannot be boxed within the boundary of language, it carries a deep culture, knowledge ,pride , a long history which still cannot be traced back from where it started and how it evolved. Modern researches shows that it had evolved thousands of years ago and day by day the results keeps dragging the origin of Tamil to several years ago from its previous results..

The phrase "Survival of the Fittest" best suits for the language Tamil,It had evolved through multiple struggles and battles. It passed through multiple phases like epigraphic inscriptions, palm scripts, manuscripts, human communication , books, songs of various genres,FM,AM,CD's,DVD's,Internet. If we ask a question "Whether Tamil language was only celebrated in ancient times?",the answer is no.. Still now the language Tamil did not loose its grip on people and Tamil people did not loose its grip over the language.. Though it had taken multiple avatars like Chennai Tamil, Coimbatore Tamil , Tutucorin Tamil , Tirunelveli Tamil , Madurai Tamil, Eelam Tamil , Malaysia Tamil , Singapore Tamil it never faded away anywhere.

There are multiple facts to prove that Tamil attained its self glory all over the centuries in various ways.

- The first book printed in India in an Indian language was in Tamil and the book name is தம்பிரான் வணக்கம்..
- As per 2018 report Tamil language was used mostly in India in internet , though the number is expected to change in coming years due to high population of people who speak Hindi, we should not forget that Tamil adapted itself in all over the years through various phases and technologies.
- As per "Archaeological survey of India" more than 60 % of epigraphical inscriptions found in India were written in Tamil.The two earliest manuscripts which was registered in UNESCO from India was written in Tamil. There is no wonder why the Indian government announced Tamil ,as the first classical language in 2004.

To add few glory to modern research and analysis in Tamil ,I want to share the findings and some facts in this submission about the Tamil alphabets and its usage over the books by analyzing various books written in Tamil Language.Books taken for analysis were taken from various centuries and various authors..At the end of this submission we have answers for various facts and we can say proudly that this kind of data analysis with huge variety of books was not done in any of our Indian languages till now.

Collecting data for analysis

To come to a conclusion on factual analysis on Tamil alphabets usage in books over the period of time, it is not fair to take one or two books. For a classical language like Tamil and for the richness it has in books , we should definitely need to analyze on atleast hundreds of books. So I started collecting books from various resources and blogs in initial days. I avoided taking reference from newspapers as most of the words in newspapers were சென்னை, இரவு ,தமிழகத்தில் , கொலை ,அமைப்பு ,ஆளும், மாநில ,தேர்தல் , கணவன், காதல். Fortunately various people from all over the world had uploaded multiple Tamil books in different websites in different fonts and formats.. From all of the websites I found , I choosed a website named “Project Madurai” and I felt satisfied as the kind of books they have provided scales across various periods and in various genres. I really appreciate the effort they have taken to summarize all books in pdf and html in same format and fonts. This website contains books from various centuries, authored by various persons , books of different religions, books of different genres which scales on history, biography, grammar , drama , short stories , poems , translated books , lyrics of songs etc. I have added the list of books taken for analysis in detailed format as listed in website “Project Madurai “ at the end of submission.

Post concluding the website “Project Madurai” as a space for collecting books, using web scrapping(Web scrapping is a technique used to extract data from websites) technique all data from the website was extracted. There were total of ~450 books in 670 divisons, which was authored by more than 300 authors. The books include திருக்குறள், திருவாசகம், நாலடியார் , பழமொழி நானூறு , சித்தர் பாடல்கள் , தொல்காப்பியம் ,சிவகாமியின் சபதம் ,பெரிய புராணம்,கந்த புராணம் , நாலாயிர திவ்ய பிரபந்தம் . The collections does not include only ancient books , it includes books from 19th Century, 20th century , 21st Century written by various authors. Once I extracted all books, I got satisfied as I was able to collect few of the greatest books written in Tamil.

Post extracting all books from website, I choosed one book to analyze initially before proceeding with other books.. The book I choosed was Tirukural written by Thiruvalluvar.Please refer below statements which was taken from wikipedia for those who are not aware about Tirukural

The Tirukkural (திருக்குறள், literally Sacred Verses), or shortly the Kural, is a ssic Tamil language text consisting of 1,330 couplets or Kurals. The text is divided into three books, each with aphoristic teachings on virtue (aram, dharma), wealth (porul, artha) and love (inbam, kama).Considered one of the great works on ethics and morality, it is known for its universality and secular nature.Its authorship is traditionally attributed to Valluvar, also known in full as Thiruvalluvar. The text has been dated variously from 300 BCE to 5th century CE. The traditional accounts describe it as the last work of the third Sangam, but linguistic analysis suggests a later date of 450 to 500 CE and that it was composed after the Sangam period.

The Kural is structured into 133 chapters, each containing 10 couplets (or kurals), for a total of 1,330 couplets.All the couplets are in kural venba metre, and all the 133 chapters have an ethical theme and are grouped into three parts, or "books"

Tirukkuṛaḷ.The first line in rural consist of 4 words and second line consists of three words.

A̲ram (28.6%)

Poruḷ (52.6%)
Inbam (18.8%)

Book I – Aṟam (அறம்): Book of Virtue (Dharma), dealing with moral values of an individual and essentials of yoga philosophy (Chapters 1-38)
Book II – Poruḷ (பொருள்): Book of Polity (Artha), dealing with socio-economic values, polity, society and administration (Chapters 39-108)
Book III – Inbam (இன்பம்): Book of Love (Kama), dealing with psychological values and love(Chapters 109-133)
"Virtue will confer heaven and wealth; what greater source of happiness can man possess?"
(Kural 31; Drew, 1840).
The book on aṟam (virtue) contains 380 verses, that of poruḷ (wealth) has 700 and that of inbam or kāmam (love) has 250. Each kural or couplet contains exactly seven words, known as cirs, with four cirs on the first line and three on the second, following the kural metre. A cir is a single or a combination of more than one Tamil word. For example, the term Thirukkural is a cir formed by combining the two words thiru and kuṟaḷ.

Of the 1,330 couplets in the text, 40 couplets relate to god, rain, ascetics, and virtue; 200 on domestic virtue; 140 on higher yet most fundamental virtue based on grace, benevolence and compassion; 250 on royalty; 100 on ministers of state; 220 on essential requirements of administration; 130 on morality, both positive and negative; and 250 on human love and passion.

Analyzing Tirukural

To analyze over Tirukural, I collected only 1330 kurals from book leaving all headings, subheadings ,titles , prefaces. In upcoming sections I am going to describe how the analysis was made , as it would help someone in some point when they involve in such activities. Please note that each individual have their own way and they can choose their own coding style.. Here I am just narrating my experience , as it would help someone by avoiding spending too much time on browsing.

At first , I wrote a program in python to extract the words from Tirukural , The program was pretty much straightforward and easy. When I executed the program it resulted me 9310 words. Once I saw the result as 9310, I know that my python program is working efficiently , as we all know each kural consists of 7 words and 7*1330 equals 9310.

Post extracting words , the next task is to extract letters.To extract letters we need to split each letter from words. For example look at below box how the word “Amma” is splitted as ['A', 'm', 'm', ‘a'] , with “,” as a separation, and four individual letters inside a list.

```
Prabhakars-MacBook-Air:~ prabhakarshanmugavel$ python3
Python 3.7.3 (default, Sep  5 2019, 17:14:41)
[Clang 11.0.0 (clang-1100.0.33.8)] on darwin
Type "help", "copyright", "credits" or "license" for more
information.
>>> a = "Amma"
>>> list(a)
['A', 'm', 'm', 'a']
>>>
```

Now if we try to perform the same task in splitting Tamil word “அம்மா”, we should get as ['அ', 'ம்', ‘மா'] as three alphabets , but instead python throws us the output like ['அ', 'ம', '்', 'ம', ‘ா'] with 5 characters. This is because python cannot understand Tamil fonts by default..

```
Prabhakars-MacBook-Air:~ prabhakarshanmugavel$ python3
Python 3.7.3 (default, Sep  5 2019, 17:14:41)
[Clang 11.0.0 (clang-1100.0.33.8)] on darwin
Type "help", "copyright", "credits" or "license" for more
information.
>>> a = "அம்மா"
>>> list(a)
['அ', 'ம', '்', 'ம', 'ா']
>>>
```

In order to overcome this , I utilized a module named open-tamil which was developed by Muthu Annamalai and his team.. Post utilizing the module, the results are yielded as expected

```
>>> from tamil import utf8
>>> word_1 = utf8.get_letters(a)
>>> word_1
['அ', 'ம்', ‘மா']
```

Here, I would take a chance to appreciate the efforts of Muthu Annamalai and his team for their efforts to develop such module in python. I am pretty sure this module is going to help lot of people to initiate various analysis in Tamil language.

Now, Lets come to our task of splitting alphabets from 9310 words which was extracted from 1330 kurals. Again I wrote a small module in python to extract alphabets from words, which yielded me 40882 count of alphabets with “ன்” being most used alphabet..

Lets look into one example of how the most used letter from a word is calculated. Using a module collections , we can yield the results like how many times a alphabet are used in a word.. For example shown in below in word “கண்டுபிடிக்க” , “க” appeared two times and other words appear 1 time each.The result shows the same for us which is highlighted.

```
>>> from collections import Counter
>>> a = "கண்டுபிடிக்க"
>>> word_1 = utf8.get_letters(a)
>>> Counter(word_1)
Counter({'க': 2, 'ண்': 1, 'டு': 1, 'பி': 1, 'டி': 1, 'க்': 1})
>>>
```

The most used alphabet “ன்” appeared in Tirukural for 2115 times. The alphabet “ன்” appeared in 1407 lines out of 2660 lines in kural. The second most used alphabet “ம்” appeared in Tirukural for 1916 times..The most used vowel alphabet (Uyir eluthu) is “அ” which appeared for 832 times..

The most used word in Tirukural is “படும்” which appeared for 42 times and out of 42 times kural has ended 40 times with word “படும்”.. There are lot of interesting facts about words used in Tirukural , but lets not go deeper into it as this submission was completely dedicated to analysis on Tamil Alphabets..

Alphabet aayutha ezuthu “ஃ” appeared 51 times in 1330 kurals.

In a similar way , I extracted all words, alphabets and no of times the words and alphabets appeared from 450 books. The resulted value gave me some numbers which seems to be what I was expecting.. From all the books, I got more than ~8 Lakh(8,00,000) of lines excluding blank lines and ~91 Lakhs (9100000) words which includes symbols, non Tamil characters. From 91 lakhs of words ,Tamil characters was extracted using a Splunk lookup table , it resulted in ~4 Crore(40000000) alphabets excluding symbols and non Tamil characters. So the analysis was done with only Tamil alphabets(4 crore alphabets) by excluding all non Tamil characters and symbols.

I was bit exclaimed as well , because any analysis done with huge set of data will yield a better result , rather than going with less data. If we go with lesser data , the results keeps changing for each time when the sources are different..

Before we go into facts and results of analysis , let’s look into Tamil Alphabets. Here I have narrated Tamil alphabets in a simple way which can be used to teach any kids who is willing to learn Tamil. Especially I have spent some considerable amount of time in framing the Tamil_Table which comes in below sections to make the learning of Tamil alphabets in easy way.

Tamil Alphabets

Vowels (உயிர் எழுத்து)

There are 12 Vowels in Tamil Alphabets. 5 of them are Kuril(குறில்) Vowels and 7 of them are Nedil(நெடில்) vowels.

Kuril vowels are pronounced in shorter time and Nedil vowels are pronounced as twice as long as Kuril vowels.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
குறில்	நெடில்	குறில்	நெடில்	குறில்	நெடில்	குறில்	நெடில்	நெடில்	குறில்	நெடில்	நெடில்

Consonants(மெய் எழுத்து)

There are 18 consonants in Tamil Alphabets which is again classified into Vallinam (வல்லினம்), Mellinam (மெல்லினம்) and Idayinam(இடையினம்).

க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்
வல்லினம்	மெல்லினம்	வல்லினம்	மெல்லினம்	வல்லினம்	மெல்லினம்	வல்லினம்	மெல்லினம்	வல்லினம்

ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்
மெல்லினம்	இடையினம்	இடையினம்	இடையினம்	இடையினம்	இடையினம்	இடையினம்	வல்லினம்	மெல்லினம்

While pronouncing Vallinam alphabets (க், ச், ட், த், ப், ற்) , it starts from heart.
While pronouncing Mellinam alphabets (ங், ஞ், ண், ந், ம், ன்) , it starts from nose.
While pronouncing Idaiyinam alphabets (ய், ர், ல், வ், ழ், ள்) , it start from throat

Vowel-Consonants(உயிர்மெய் எழுத்து)

12 Vowels and 18 Consonants in Tamil join and produce 216 Vowel-Consonants which is responsible for more than 60 % appearance in all Tamil words.

for example

அ+க் = க
ஆ+க் = கா
இ+க் = கி
ஈ+க் = கீ

Here apart from formation of Vowels-Consonants , we have to observe one thing. Since “அ” is a Kuril eluthu it joins with “க்” and produces uyirmei Kuril eluthu(க) , similarly “ஆ” is a Nedil eluthu ,it joins with “க்” and produces uyirmei nedil eluthu(கா).In this fashion, 12 vowels and 18 consonants produce 216 vowel-consonants. Out of 216 vowel-consonants 126 Alphabets are uyirmei Nedil eluthu and 90 Alphabets are uyirmei Kuril eluthu. So totally we have 247 Tamil alphabets(12 vowels, 18 consonants, 216 vowel-consonants, 1 Aayutha eluthu)

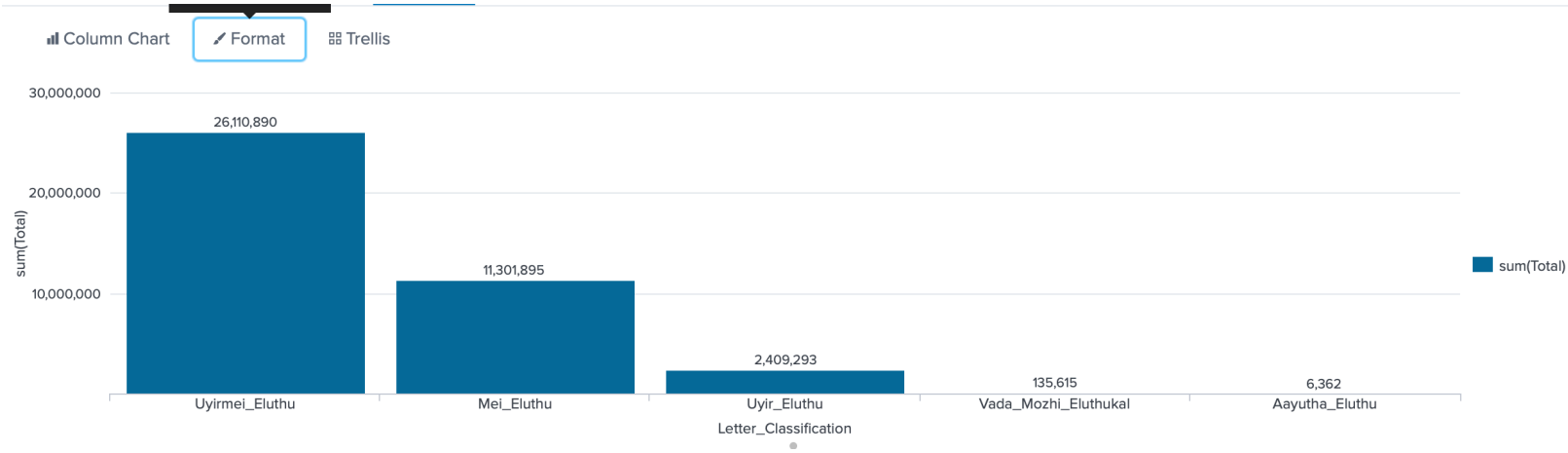
		Kuril	Nedil	Kuril	Nedil	Kuril	Nedil	Kuril	Nedil	Nedil	Kuril	Nedil	Nedil	
	Uyir Eluthu ->	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ	
	Mei Eluthu ↓													
vallinam	க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ	uyirmei_vallinam
mellinam	ங்	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ	uyirmei_mellinam
vallinam	ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ	uyirmei_vallinam
mellinam	ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ	uyirmei_mellinam
vallinam	ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ	uyirmei_vallinam
mellinam	ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ	uyirmei_mellinam
vallinam	த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ	uyirmei_vallinam
mellinam	ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ	uyirmei_mellinam
vallinam	ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ	uyirmei_vallinam
mellinam	ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ	uyirmei_mellinam
idaiyinam	ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ	uyirmei_idaiyinam
idaiyinam	ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ	uyirmei_idaiyinam
idaiyinam	ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ	uyirmei_idaiyinam
idaiyinam	வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ	uyirmei_idaiyinam
idaiyinam	ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ	uyirmei_idaiyinam
idaiyinam	ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ	uyirmei_idaiyinam
vallinam	ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ	uyirmei_vallinam
mellinam	ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ	uyirmei_mellinam
		U_Kuril	U_Nedil	U_Kuril	U_Nedil	U_Kuril	U_Nedil	U_Kuril	U_Nedil	U_Nedil	U_Kuril	U_Nedil	U_Nedil	
				U_kuril=Uyirmei_Kuril					U_Nedil = Uyirmei_Nedil					
						Tamil_Easy_Table								

Aayutha Eluthu (ஃ) is a special alphabet in Tamil , Apart from 247 Tamil Alphabets, there are many other Grantha alphabets (ஜ்,ஸ்,ஷ்,ஹ்,க்ஷ் ,ஸ்ரீ)which are being used in Tamil often with the joint of Tamil vowels.

Analytical Results

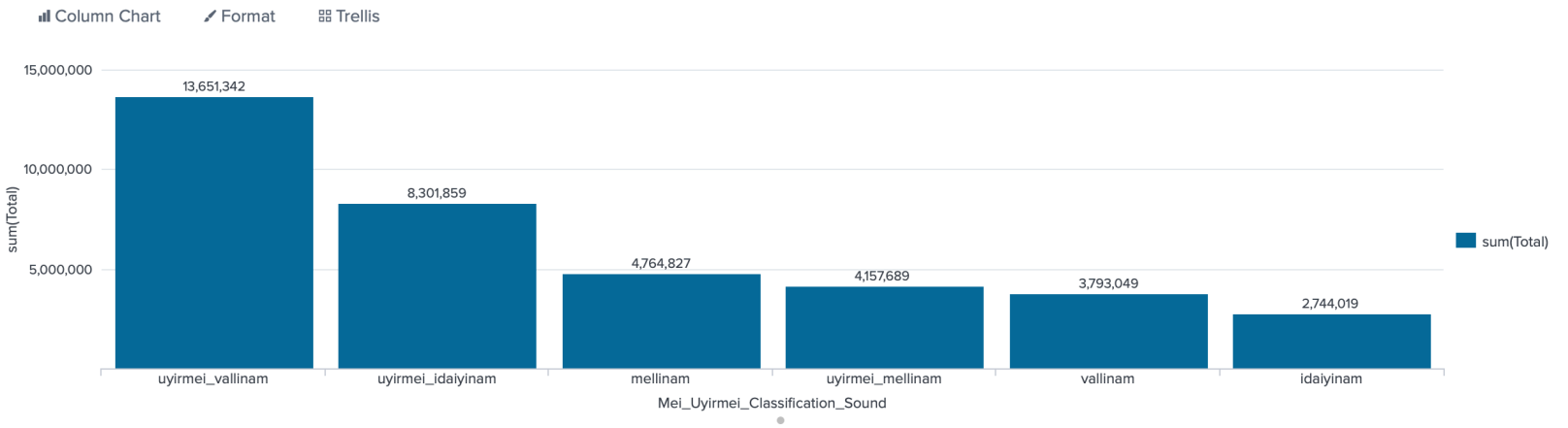
As we had already discussed previous sections of how a word and alphabets were extracted from books using python, lets move to analyze the results by using the output yielded from python program. Here an analytical tool named “Splunk” is used to analyze the numbers and plot those analysis into graphs to understand in a easier way..

Graphical representation on classification of Alphabets

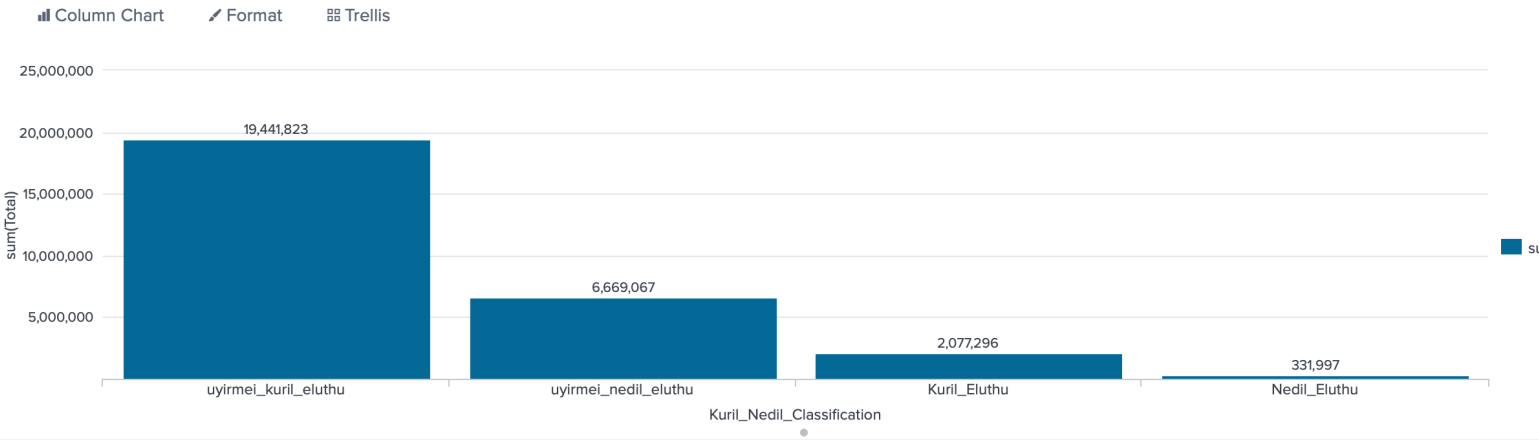


Uyirmei Eluthukal accounts for 26110890 (~65.4%) out of ~4Crore alphabets taken for analysis Mei Eluthukal accounts for 11301895(~28.3%) ,Uyir Eluthukal accounts for 2409293(~6.02 %) Vada Mozhi Eluthukal accounts for 135615(~0.3%) ,And Aayutha Eluthukal accounts for 6362(~0.01 %)

Graphical representation on Mei and Uyirmei Alphabet Classification



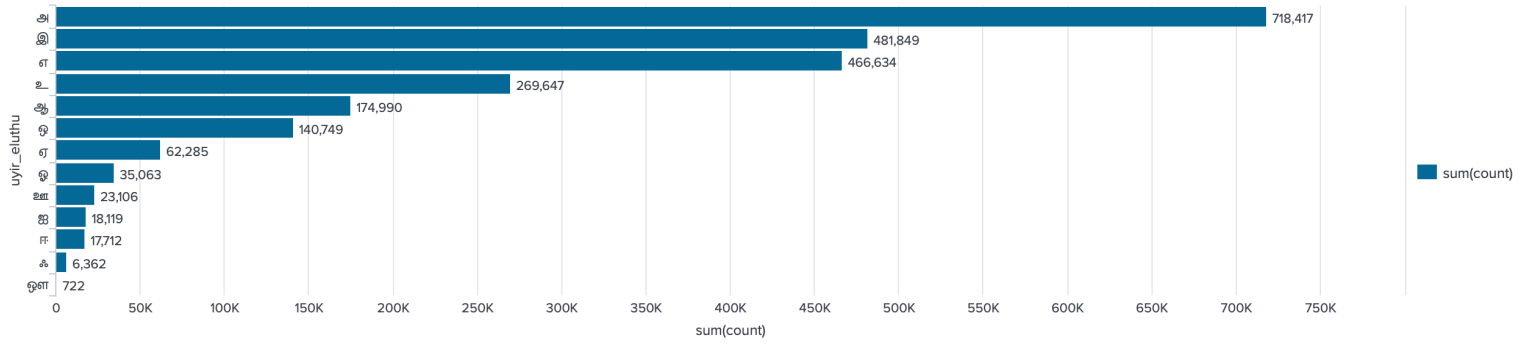
There is a general saying among Tamil professors that uyirmei vallinam , uyirmei idayinam and uyirmei mellinam alphabets are most used than vallinam,mellinam,idayinam. But from above plot we can understand that in mellinam category mellinam was most used compare to uyirmei mellinam.. The reason for this variation is the alphabets “ ன்” and “ம்” which belongs to mellinam are heavily used alphabets in Tamil. Also when we look into Kuril Nedil classification



graph , we can clearly understand that Kuril eluthukal dominates more than Nedil eluthukal , though Kuril eluthu account for only 5 alphabets , whereas Nedil eluthukal accounts for 7 alphabets. Similarly uyirmei kuril eluthukal accounts for 90 alphabets and uyirmei nedil eluthukal accounts for 126 alphabets

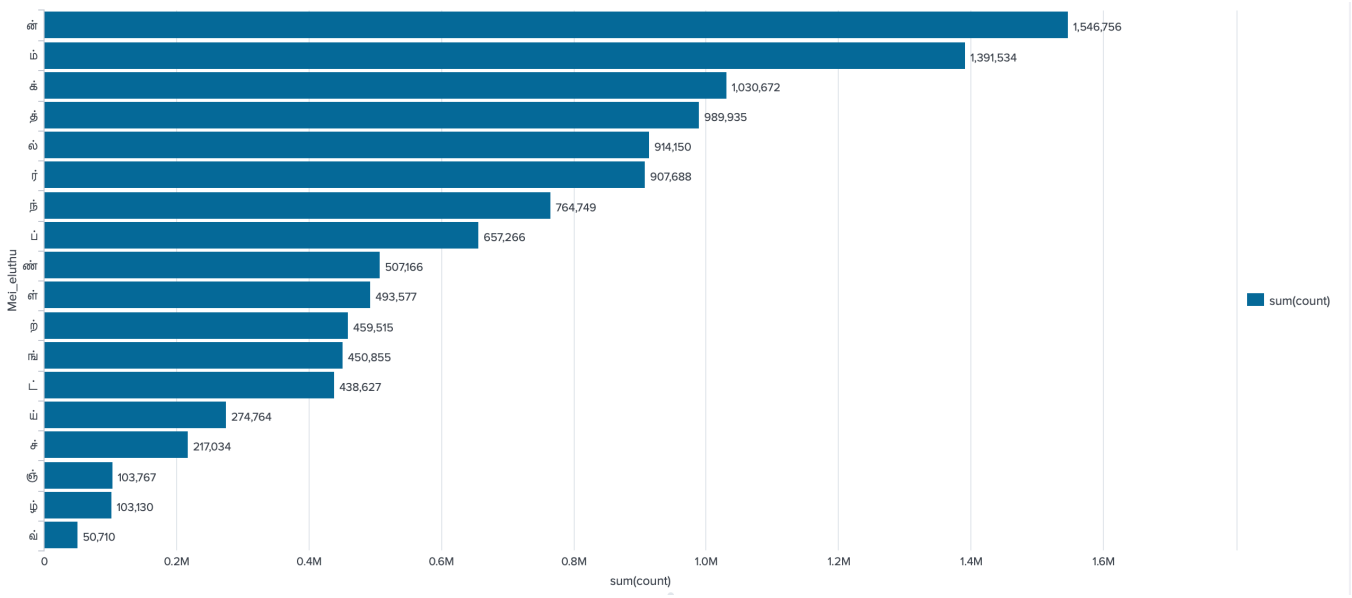
Most used Vowels to least(Uyir Eluthukal)

In this graph we have accounted “ஃ” as well for analyzing, there is also a general saying that “ஃ” is least used alphabet in this category. But by detailed analysis we were able to find that “ஔ” is least used and it is even used far less than any Vowels(Uyir Eluthukal)



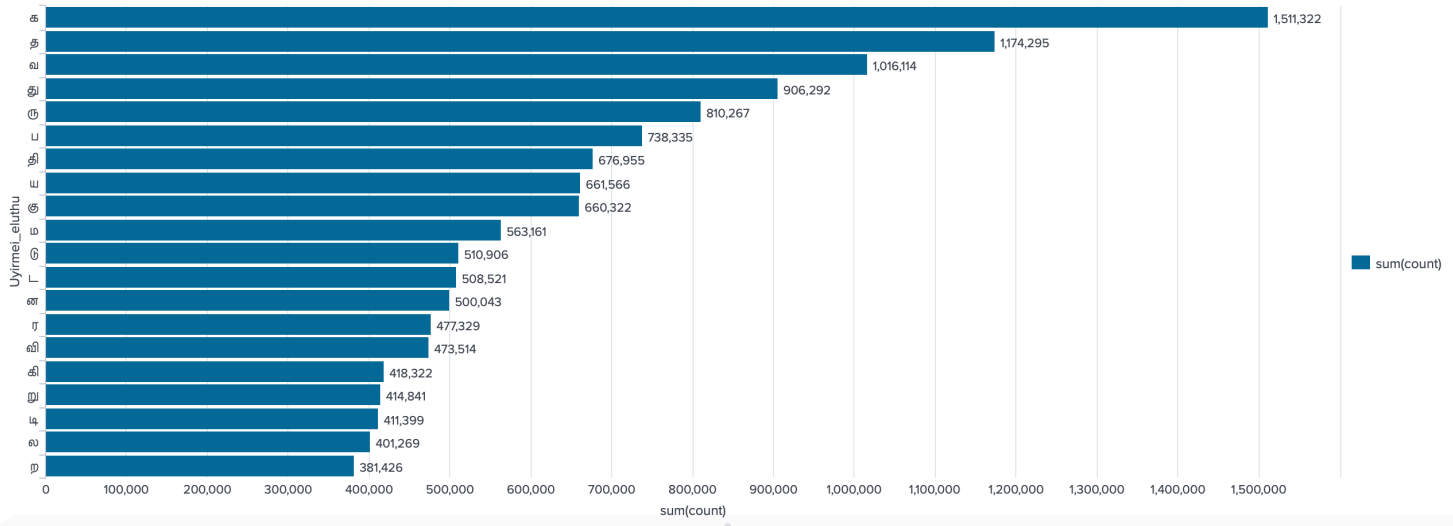
Most used Alphabet on Consonants to least(Mei_Eluthukal)

The Alphabet “ன்” is not only the most used alphabet in consonants category , but it also holds the position of most used alphabet in Tamil across all categories.

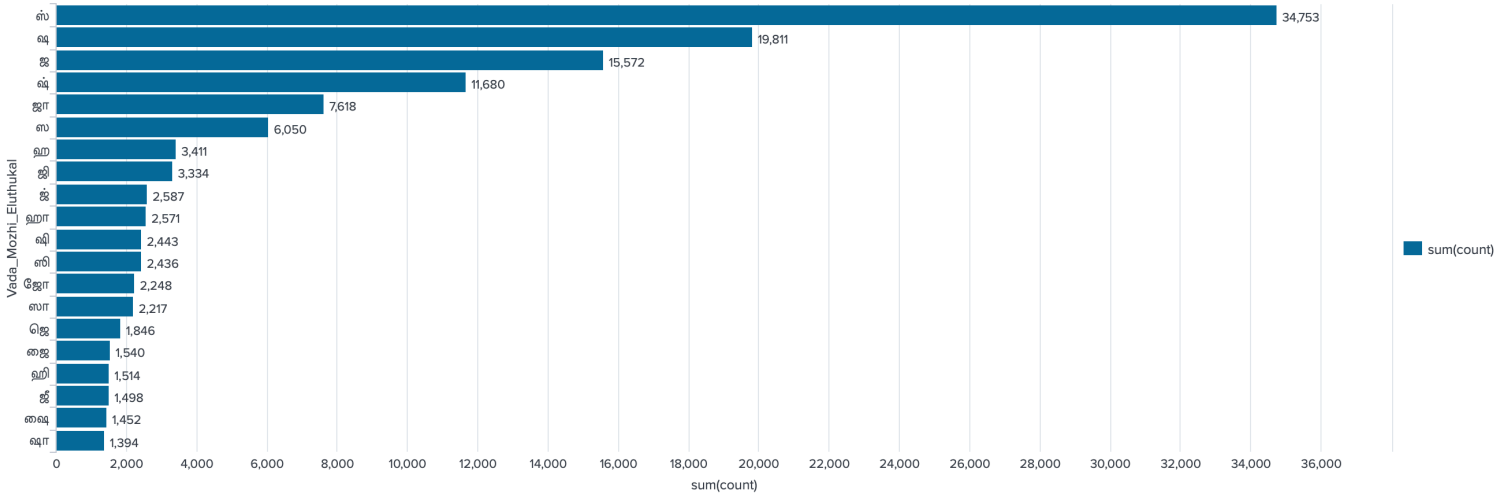


Top 20 Most used Vowel-Consonants (Uyirmei)

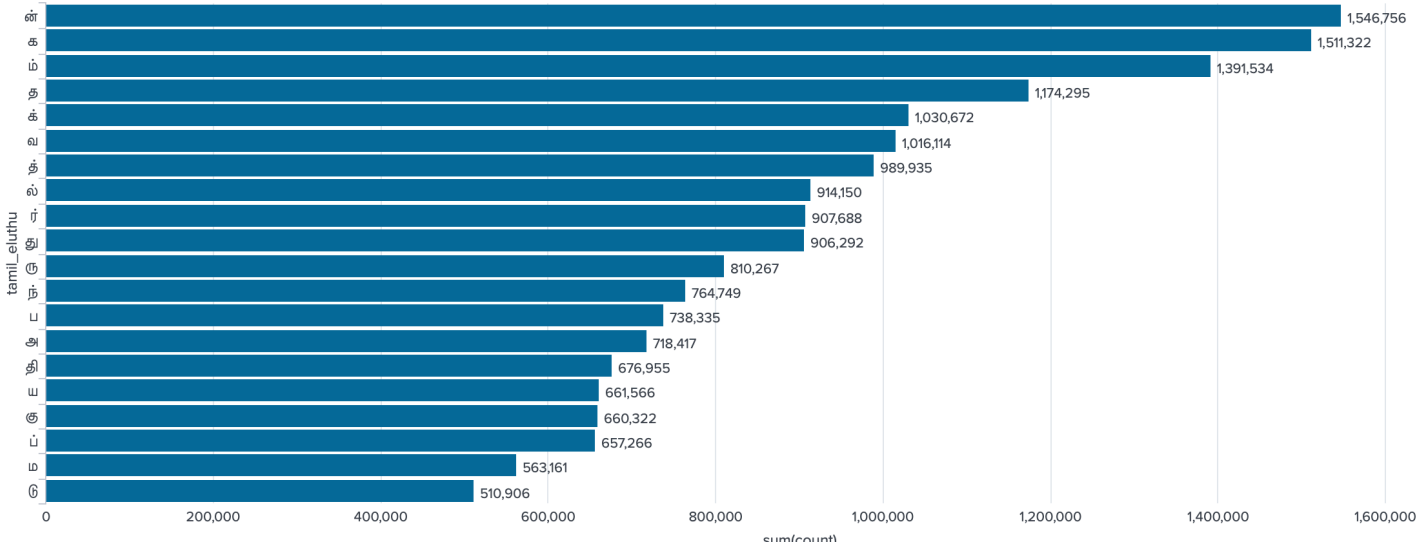
Since there are 216 letters , I had plotted graph only on top 20 Alphabets. But in my GitHub page I have added all the results.



Most used non Tamil alphabets



Top 20 Tamil Alphabets



Facts observed from Tirukural.

- Total number of words from 1330 Kurals = 9310 words
- Total number of letters from 9310 words = 40882 Alphabets
- Out of 9310 words , 6338 words are unique
- “படும்” word appeared for 42 times and this word is the most repeated word in tirukural
- Out of 42 times , kural ends 40 times with word “படும்”
- Most used alphabet in tirukural is "ன்” and it appeared 2115 times.
- “ன்” appeared in 1407 lines out of 2660 lines in kural
- “ம்” is the second most used alphabet in Tirukural and it appeared 1916 times
- The most used vowel alphabet (Uyir eluthu) is “அ” which appeared for 832 times
- Aayutha ezuthu “ஃ” appeared 51 times in 1330 kurals

Key points to remember

- Number of books taken for analysis = ~450
- Number of authors who authored the books = ~300
- Number of words extracted from books = 91 lakhs
- Out of 91 Lakhs words ,~19 lakhs are unique words. (You all may be wondering how its possible to get 17 lakhs of unique Tamil words, whereas there are only close to ~3-4 lakhs of identified words in tamil, for example the word “லிங்கம்” is used as லிங்கத்தின், லிங்கத்திற்கு, லிங்கத்தை, லிங்கன், லிங்கனம், லிங்கப்பையர், லிங்க ,லிங்கபூசை by various authors and I could find that there are 130 ways of using word “லிங்கம்”. We have extracted all such words without modifying any alphabets, hence we have identified 17 lakhs of unique Tamil words from 450 books)
- Number of Tamil Alphabets extracted from 91 Lakhs of words = ~4 Crores
- Uyirmei Eluthukal accounts for 26110890 (~65.4%) out of ~4Crore alphabets

- Uyirmei vallinam alphabets are most used among uyimei eluthukal
- Mei Eluthukal accounts for 11301895(~28.3%)
- Uyir Eluthukal accounts for 2409293(~6.02 %)
- Vada Mozhi Eluthukal accounts for 135615(~0.3%)
- Aayutha Eluthukal accounts for 6362(~0.01 %)
- Most used alphabet is “ன்” and it appeared 1546756 times out of 4 crore alphabets extracted ,which is close to 3.9%.
- Most used Vowel is “அ” and it appeared 718417 times which is 1.7%
- Least used vowel is “ஔ” and it appeared only 722 times
- Most used non Tamil alphabet is “ஸ்” and it appeared 34753 times which is 0.08 %
- Least used Tamil alphabets are “ஙெ” , “ஙோ”, “ஙௌ”, “ஜௌ”, “ணௌ” and “றௌ” which appeared for 1 time.
- Most used word is “என்று” which appeared 60910 times out of 17 Lakh words

All the results and sources taken for this analysis will be updated in my github page constantly. Please refer to it and reach me @ prabhakar6372@icloud.com for any corrections.

Sources

Books are extracted from “https://www.projectmadurai.org/pmworks.html"

Tamil python module taken from “https://pypi.org/project/Open-Tamil/"

Github url “https://github.com/prabhakar6372/tamil_alphabet_analysis"

Technologies used
1) Bash script was used for Web scrapping
2) Python was used to write program and extract letters and words
3) Splunk was used to analyze the data
4) Github was used to upload the results and maintain the sources