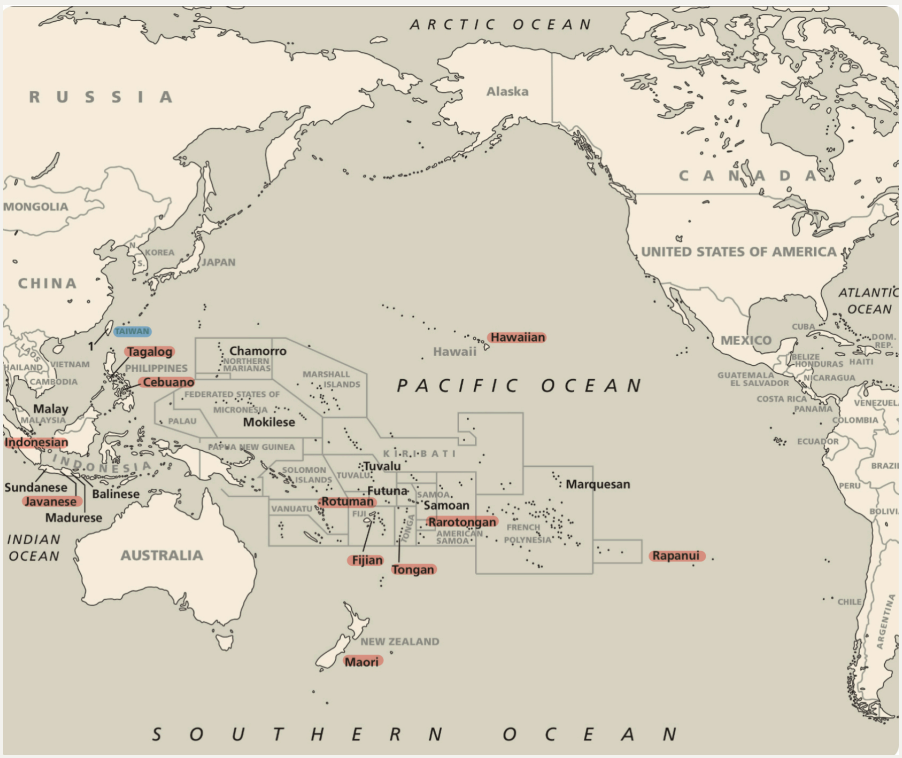
An Improved Model for Austronesian--Verification of Some Hypotheses

Introduction

This report was based on some hypotheses of Austronesian, mainly focusing on the hypothesized relations. In 1899, Wilhelm Schmidt proposed that the **Austronesian family was of the same origin as the Austroasiatic family**. In 1970, American scholar argued that the **Austronesian family shared the same origin with the Indo-European family**. Futhremore, Laurent Sagart argued that the **Austronesian family is of the same origin as the Sino-Tibetan family** and some Japanese scholars believed that the **Austronesian and Altaic languages constitute a hybrid language**，**Japanese**. One of the most famous hypothesis was that Austronesian was originated in Taiwan region. Considering the effect of Pacific Ocean current and based on the imigration route of hypothesis, 14 language samples along the spread were selected.



Moreover, Indo-European languages, Altaic languages (i.e Azerbaidzhan, kazakh) and Japanese were studied as well in order to verify the hypotheses which indicated the relations with Indo-European, Altaic.

Methodology

This was an improved model of Hierarchical Clustering, so the vairable was increased from 10 to 17, including the number (1-10) , tree, water, fire, stone, sea, hand and head, which was the communal environment of prehistorical human. By using the Hierarchical Clustering, but modifying the model with cosine distance, Pearson correlation coefficient and Euclid distance and finally comparing them together, the relations became more and more clear.

Cosine Distance

The cosine of two non-zero vectors can be derived by using the Dot Product Formula:

$$A \cdot B = |A||B|cos\theta$$

$$\text{Similarity}=cos\theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. The technique is also used to measure cohesion within clusters in the field of data mining.

Pearson Correlation Coefficient

This coefficient is widely used to measure the degree of linear correlation between two variables

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

Euclid Distance

The ED is the distance between two variables in the Euclid Space, which is a scalar.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

The below was the sample of total languages with 17 variables, including Austronesian, Indo-European and Altaic, 22 languages in total.

```
Malay=
{'satu','dua','tiga','empat','lima','enam','tujuh','lapa
n','sembian','sepuluh','pokok','air','api','batu','laut'
,'tangan','kepala'};
```

```
Filipino=
{'isa','dalawa','tatlo','apat','lima','anim','pito','wal
o','siyam','sampu','puno','tubig','apoy','bato','dagat',
'kamay','ulo'};
Hawaiian=
{'ekahi','elua','ekolu','eha','elima','eono','ehiku','ew
alu','eiwa','umi','kumulaau','wai','ahi','pohaku','kai',
'lima','poo'};
Fijian=
{'ndua','rua','tolu','vaa','lima','ono','vitu','walu','z
iwa','tini','vunikau','wai','bukawaqa','vatu','wasawasa'
,'ligana','uluna'};
BasaJawa=
{'sichi','loro','tellu','papat','limo','enem','pitu','wo
lu','songo','sepuluh','wit','banyu','geni','watu','segar
a','tangan','sirah'};
Samoan=
{'tasi','lua','tolu','fa','lima','ono','fitu','valu','iv
a','sefulu','laau','vai','afi','maa','sami','lima','ulu'
};
Indonesian=
{'satu','dua','tiga','empat','lima','enam','tujuh','dela
pan','sembilan','sepuluh','pohon','air','api','batu','la
ut','tangan','kepola'};
Hakka=
{'yit','ngi','sam','si','ng','luk','cit','bat','giu','si
p','su','sui','fo','sak','hoi','su','teu'};
Japanese=
{'ichi','ni','san','si','go','roku','nana','hachi','kyuu
','zyuu','ki','mizu','hi','kesseki','umi','te','atama'};
Azerbaidzhan=
{'bir','iki','uc','dord','bes','alti','yeddi','sekkiz','
doqquz','on','agac','su','atas','das','deniz','el','bas'
};
Maori=
{'tahi','rua','toru','wha','rima','ono','whitu','waru','
iwa','tekau','rakau','wai','ahi','toka','moana','ringa',
'upoko'};
Rarotongan=
{'tai','rua','toru','a','rima','ono','itu','raru','iva',
'ngauru','mata','vai','manu','ivi','moana','rima','kiri'
};
Rapanui=
{'tahi','rua','toru','ha','rima','ono','hitu','vau','iva
','ahahuru','miro','bai','afi','maea','moana','rima','pu
oko'};
```

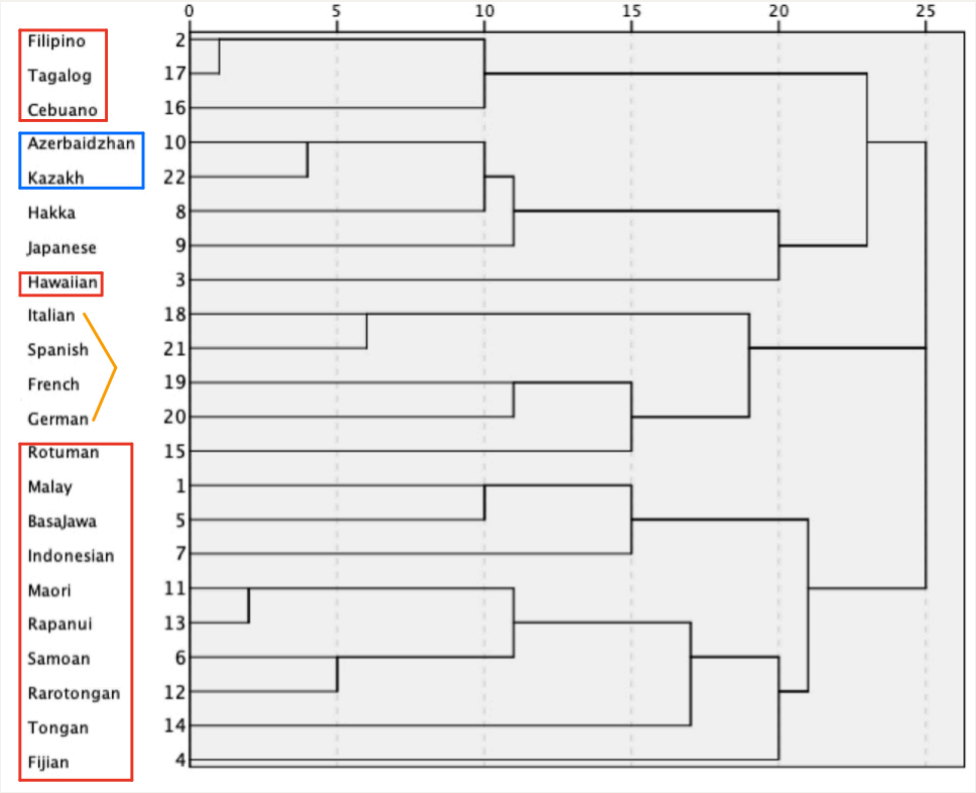
```
Tongan=
{'taha','ua','tolu','fa','nima','ono','fitu','raru','hiv
a','ehongofulu','akau','vai','afi','maka','tahi','nima',
'ulu'};
Rotuman=
{'ta','rua','folu','hake','lima','ono','hifu','valu','si
va','saghul','ai','tanu','rahi','rahi','hafu','tanu','uh
apa','mafa'};
Cebuano=
{'usa','duha','tulo','upat','lima','unum','pito','walo',
'siyam','napulo','kahoy','tubig','sunog','bato','dagat',
'kamot','ulo'};
Tagalog=
{'isa','dalawa','tatlo','apat','lima','anim','pito','wal
o','siyam','sampung','puno','tubig','apoy','bato','dagat
','kamay','ulo'};
Italian=
{'uno','due','tre','quattro','cinque','sei','sette','ott
o','nove','deici','albero','acqua','fuoco','pietra','mar
e','mano','testa'};
French=
{'un','deux','trois','quatre','cinq','six','sept','huit'
,'neuf','dix','arbre','eau','feu','pierre','mer','main',
'tete'};
German=
{'einz','zwei','drei','vier','funf','sechs','sieben','ac
ht','neun','zehn','baum','wasser','feuer','stein','meer'
,'hand','kopf'};
Spanish=
{'uno','dos','tres','cuatro','cinco','seis','seite','och
o','nueve','diez','arbol','agua','fuego','piedra','mar',
'mano','cabeza'};
Kazakh=
{'bir','eki','us','tort','bes','alti','jeti','segiz','to
giz','on','agas','sw','ot','tas','teniz','qol','basi'};
```

By digitalizing these samples, we could get the numbers instead of text.

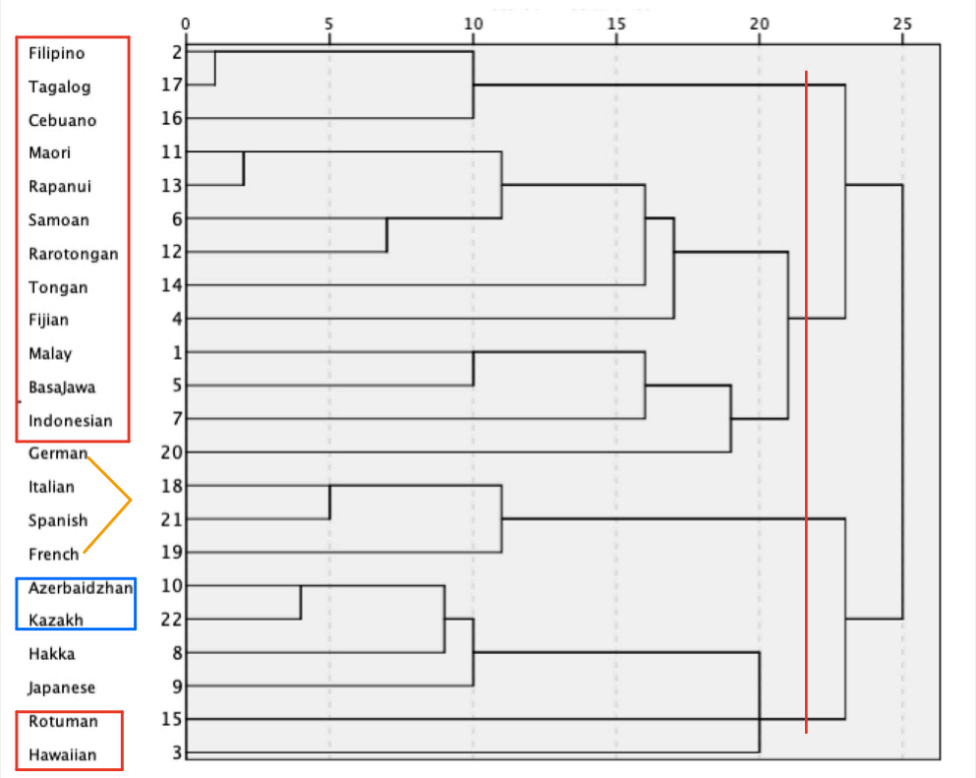
Malay	54.59815003	20.08553692	54.59815	148.413159	20.0855369	54.59815	148.413159	148.413159	403.428793	403.4287935	148.413159	54.59815	54.59815	148.413159	54.59815	54.59815	54.59815
Filipino	20.08553692	148.4131591	54.59815	54.59815	20.0855369	20.0855369	148.413159	148.413159	54.59815	148.4131591	54.59815	148.413159	54.59815	54.59815	54.59815	54.59815	54.59815
Hawaiian	148.4131591	54.59815003	148.413159	54.59815	54.59815	54.59815	148.413159	54.59815	20.0855369	20.08553692	2980.95799	20.0855369	54.59815	403.428793	20.0855369	54.59815	54.59815
Fijian	54.59815003	20.08553692	54.59815	54.59815	20.0855369	20.0855369	148.413159	148.413159	20.0855369	7.389056099	1096.63316	20.0855369	2980.95799	148.413159	1096.63316	54.59815	148.413159
Basa Jawa	148.4131591	20.08553692	54.59815	148.413159	20.0855369	54.59815	148.413159	148.413159	54.59815	403.4287935	54.59815	54.59815	54.59815	148.413159	20.0855369	54.59815	54.59815
Samoan	54.59815003	20.08553692	54.59815	20.0855369	20.0855369	20.0855369	148.413159	148.413159	54.59815	148.4131591	54.59815	54.59815	54.59815	148.413159	20.0855369	54.59815	54.59815
Indonesian	54.59815003	20.08553692	54.59815	148.413159	20.0855369	54.59815	148.413159	1096.63316	1096.63316	403.4287935	148.413159	54.59815	54.59815	148.413159	54.59815	54.59815	148.413159
Hakka	20.08553692	20.08553692	148.413159	54.59815	54.59815	20.0855369	148.413159	54.59815	20.0855369	20.08553692	54.59815	148.413159	20.0855369	54.59815	20.0855369	54.59815	20.0855369
Japanese	54.59815003	20.08553692	148.413159	54.59815	54.59815	54.59815	148.413159	54.59815	54.59815	54.59815003	54.59815	148.413159	20.0855369	403.428793	20.0855369	54.59815	54.59815
Azerbaidzhan	20.08553692	20.08553692	148.413159	20.0855369	54.59815	54.59815	54.59815	403.428793	403.428793	7.389056099	54.59815	148.413159	54.59815	148.413159	54.59815	54.59815	20.0855369
Maori	54.59815003	20.08553692	20.0855369	54.59815	20.0855369	20.0855369	148.413159	148.413159	54.59815	20.08553692	148.413159	20.0855369	54.59815	54.59815	54.59815	54.59815	148.413159
Rarotongan	20.08553692	20.08553692	20.0855369	54.59815	20.0855369	20.0855369	148.413159	148.413159	54.59815	403.4287935	54.59815	54.59815	54.59815	148.413159	54.59815	54.59815	54.59815
Rapanui	54.59815003	20.08553692	20.0855369	54.59815	20.0855369	20.0855369	148.413159	148.413159	54.59815	1096.633158	54.59815	54.59815	54.59815	148.413159	54.59815	54.59815	148.413159
Tongan	54.59815003	20.08553692	54.59815	20.0855369	20.0855369	20.0855369	148.413159	148.413159	20.0855369	8103.083928	54.59815	54.59815	54.59815	148.413159	54.59815	54.59815	54.59815
Rotuman	20.08553692	20.08553692	148.413159	54.59815	20.0855369	20.0855369	148.413159	148.413159	20.0855369	403.4287935	54.59815	54.59815	54.59815	148.413159	54.59815	7.3890561	54.59815
Cebuano	20.08553692	54.59815003	54.59815	54.59815	20.0855369	54.59815	148.413159	148.413159	54.59815	403.4287935	148.413159	148.413159	148.413159	54.59815	54.59815	54.59815	54.59815
Tagalog	20.08553692	148.4131591	54.59815	54.59815	20.0855369	20.0855369	148.413159	148.413159	54.59815	403.4287935	54.59815	148.413159	54.59815	54.59815	54.59815	54.59815	54.59815
Italian	7.389056099	20.08553692	7.3890561	403.428793	54.59815	7.3890561	20.0855369	148.413159	7.3890561	54.59815003	148.413159	148.413159	54.59815	403.428793	54.59815	7.3890561	54.59815
French	7.389056099	54.59815003	54.59815	148.413159	20.0855369	1	20.0855369	54.59815	20.0855369	20.08553692	20.0855369	54.59815	20.0855369	148.413159	7.3890561	20.0855369	20.0855369
German	20.08553692	20.08553692	20.0855369	20.0855369	20.0855369	54.59815	7.3890561	20.0855369	20.0855369	7.389056099	54.59815	7.3890561	20.0855369	20.0855369	20.0855369	1	54.59815
Spanish	7.389056099	20.08553692	7.3890561	148.413159	54.59815	7.3890561	20.0855369	54.59815	20.0855369	20.08553692	54.59815	148.413159	54.59815	403.428793	20.0855369	7.3890561	148.413159
Kazakh	20.08553692	20.08553692	148.413159	20.0855369	54.59815	54.59815	54.59815	54.59815	148.413159	7.389056099	54.59815	148.413159	54.59815	148.413159	54.59815	54.59815	54.59815

Pluging in these data into SPSS, the consequeces were totaly different by using different three methods:

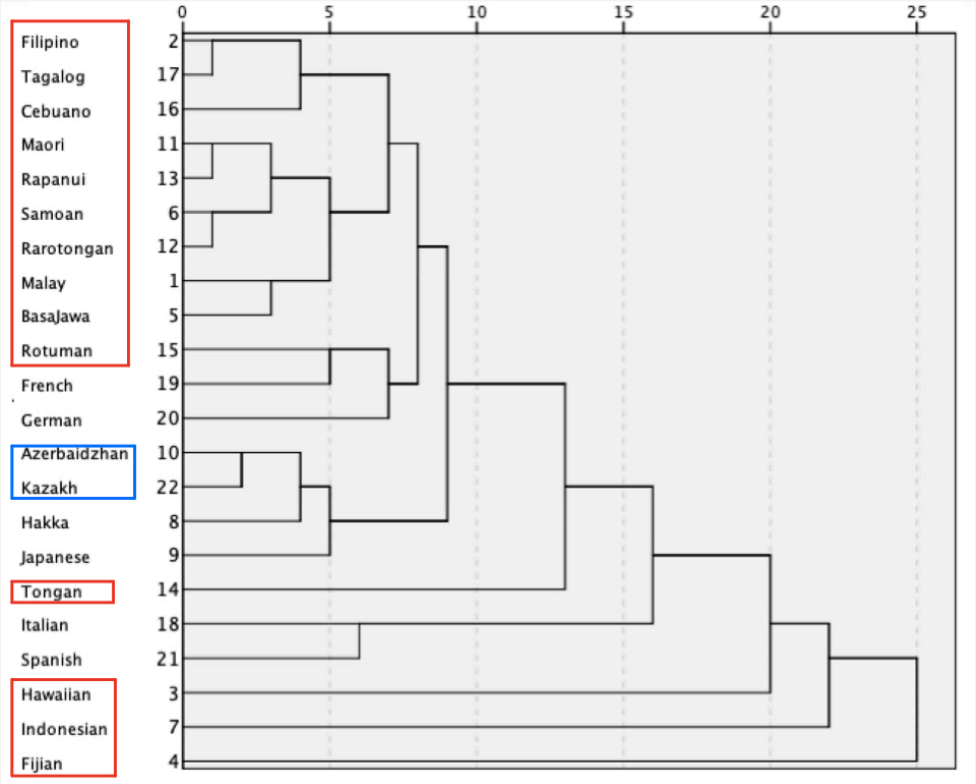
Cosine distance:



Pearson Correlation Coefficient:



Euclid Distance:



Analysis

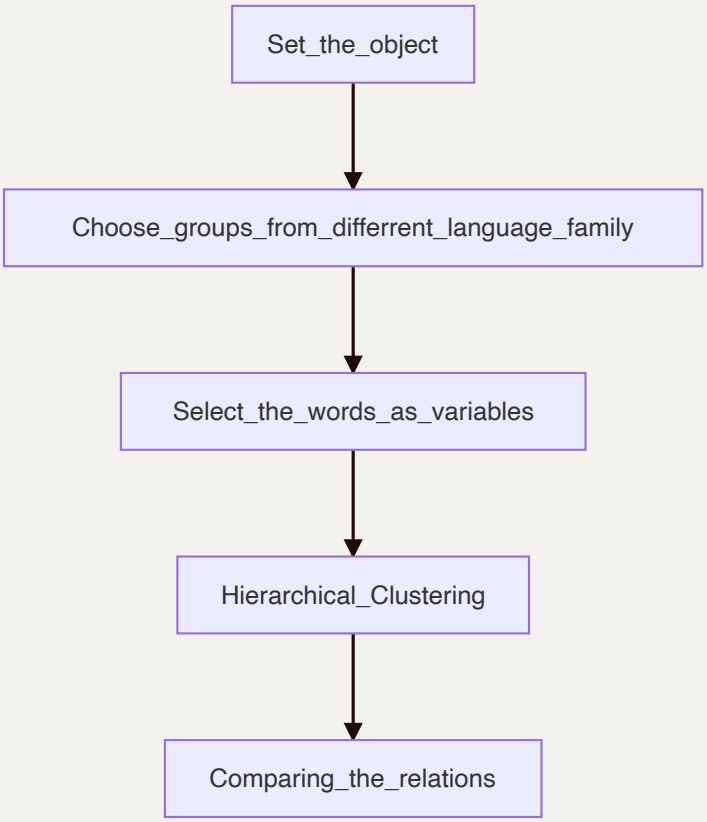
Comparing the three results, we could conclude that the hypothesis that Austronesian and Altaic languages constitute a hybrid language， Japanese, was reasonable, because all of the three graph indicated that Azerbaidzhan, Kazakh had a very closed relationship with Japanese. Geographically， since Altaic language family had some branches spreading on the East Asian, it could match the result mathematically. However, more evidence needed to be mined in order to study the how these languges constitute together, perhaps we should consider some historical factors.

For the hypothesis of Austronesian family shared the same origin with the Indo-European family, the second graph reflected this point well. However, from the graph, the Indo-European language and the group of Austronesian family were in the different branch, which means that this hypothesis was not reasonable. But this graph indicated a closed relations between Indo-European language and Altaic language.

Hakka was a language of Sino-Tibetan family, and was widely spoken in Taiwan region and Nanyang region, which was a nice sample to verify the hypotheses. To study more about the second graph, the branch of Hakka and Austronesina family share very closed relations, moreover, Rotuman and Hawaiian lies on the different group with other 12 Austronesina languages, perhaps this was due to the far distance of these two languages on the Pacific Ocean. Considering the Pacific Ocean current, it was very difficult to travel to the Hawaii and Rapanui, perhaps during this period, if other imigrants arrived these two regions, things became more hard. Lexically, similar words such as 5 (elima-Hawaii, lima--Rapanui etc) indicated the huge range of the spread of the Austronesian family.

Conclusion

In this experiment, 22 languages was studied. Mathematically, by comparing the three methods, Cosine Distance, Pearson Correlation Coefficient and Euclid Distance, the most suitable method was Pearson Correlation Coefficient. It was not surprising that most scholars supported the hypothesis that Austronesian languages originated in Taiwan region and this could be indicated from the results of this experiment clearly as well. If we want to study more about the relation between the Altaic and Austronesian languages, Japanese and Austronesian language, geographical and historical factors must be considered, since these had a huge influence on the spread of languages on the Pacific Ocean. The Hierarchical Clustering was good at study the relations between different language family, combining with the last experiment, the procedure of studying the linguistics can be carried like below:



The sample-choosing is the most important steps since it affects the results so much, and it will be more accurate if we choose the sample based on some historical and geographical factors, such as the current, immigration etc.