The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

University of Nottingham

Formal Report

Summer research internship

Phylogenetic tree construction for Austronesian

Author:
Mingdong He

Student Number:
20126521

September 2, 2020

Contents

1 Abstract 2

2 Introduction 2

3 Method 2

4 Discussion 8

5 Conclusion 9

6 Acknowledgement 9

1 Abstract

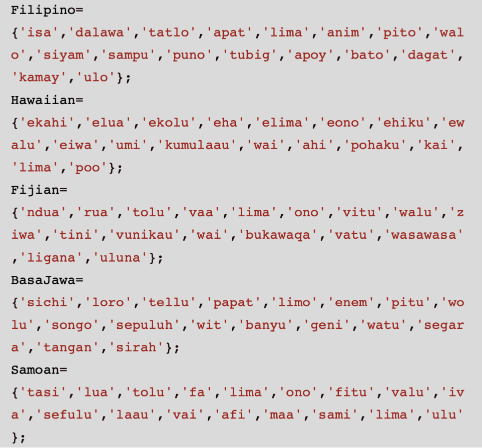
Languages are evolving like species [1]. Biologists analyze the heritable features to study the links between species, such as analyzing DNA by DNA Sequence Model or building phylogenetic trees. While linguisticians study the languages based on word, grammar and phonemes so similar method could be used. Hypothesis shows that Austronesian originates from Southern China[2]. This report mainly focuses on the basic method of phylogenetic tree construction for Austronesian based on analyzing the language distance and historical factors are also considered. The same method is also used for Indo-European language family as an illustration.

2 Introduction

The basic idea of study method is the comparison between linguistics and biology, compare languages and species, words and genes, evolution of languages and species etc. Due to the large amount of words, it is necessary to cause huge inaccuracy. In order to consider a more effective way to choose our sample, we could consider that primitive humans would face some of the same environmental characteristics, such as trees, stones, and the expression of numbers, and we could choose some common words for our study, based on this idea, the American linguistician, Morris. Swadesh has created a core vocabulary that contains more than 200 words that he believes all languages should contain. Thus Swadesh list [4] could be used as our sample. Moreover, Levenshtein distance or edit distance is used to measure the distance between two words, which is the minimal step for converting one word to another by adding, deleting or swapping.

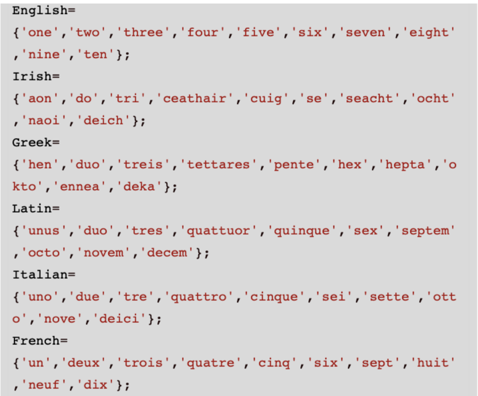
3 Method

We first build a language matrix based on Levenshtein distance and wrote a MATLAB function to value each languages, which is a way to convert text to number. Then we use hierarchial Clustering to model the relationship of different languages and finally build phylogenetic. The main idea of hierarchical clustering was by comparing the distance to divide the space into several classifications. Forming a matrix by putting the object on the column and row and then choosing a distance: such as EuclideanDistance, Standardized Euclidean distance, cosine distance, Mahalanobis distance, Chebyshev Distance etc. Then calculating the distance and selecting the minimum distance and combine them together. Finally, using the combined data as new object. Iterating the above step we could get the language clustering by python. Notice that the languages are all recorded in Latin words in our research. We also simply list the different methods for distance measurement we used.



```
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'
};
```

FIG. 1. A extracted graph of the word sample from Swadesh list for Indo-European.



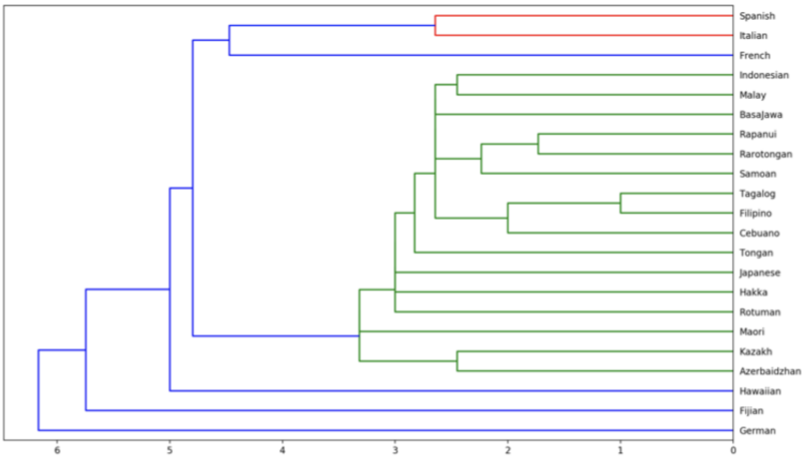
```
English=
{'one','two','three','four','five','six','seven','eight'
,'nine','ten'};
Irish=
{'aon','do','tri','ceathair','cuig','se','seacht','ocht'
,'naoi','deich'};
Greek=
{'hen','duo','treis','tettares','pente','hex','hepta','o
kto','ennea','deka'};
Latin=
{'unus','duo','tres','quattuor','quinque','sex','septem'
,'octo','novem','decem'};
Italian=
{'uno','due','tre','quattro','cinque','sei','sette','ott
o','nove','deici'};
French=
{'un','deux','trois','quatre','cinq','six','sept','huit'
,'neuf','dix'};
```

FIG. 2. A extracted graph of the word sample from Swadesh list for Austronesian.

Different kinds of distances are used for tree construction.

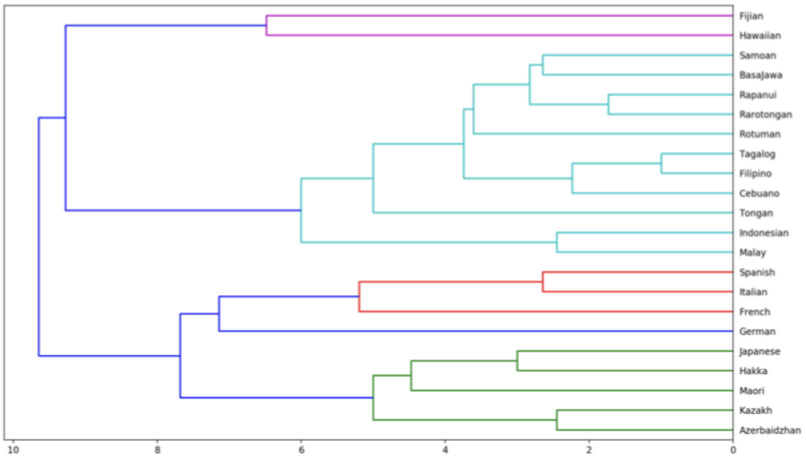
1. Single Linkage/Nearest Neighbor Method

$D(A, B) = min(d(y_i, y_j))$, where $d(y_i, y_j)$ is Euclid distance.



2. Complete Linkage/Farthest Neighbor Method The distance between groups is defined by the distance between the most distant individuals of the two groups.

$D(A, B) = max((d(y_i, y_j)))$



3. Average Linkage

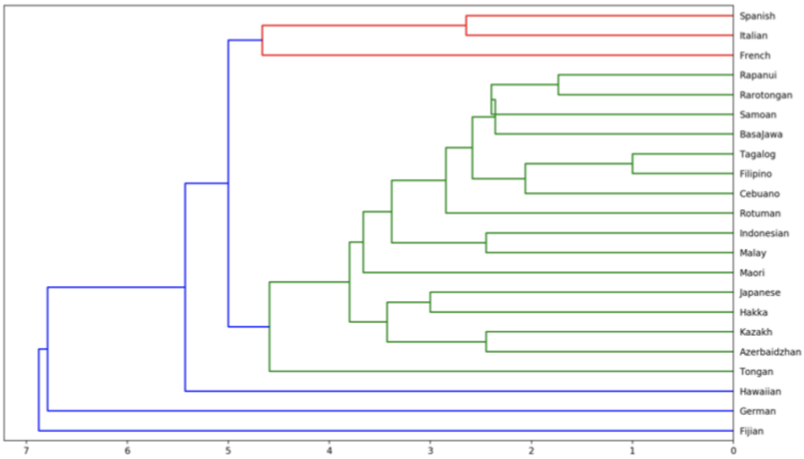
$D(A, B) = \frac{1}{n_A n_B} \sum \sum d(y_i, y_j)$

4. Centroid Method The distance between the two groups is defined as the centroid of mass of each group

$D(A, B) = d(\hat{y_A}, \hat{y_B} = \frac{1}{n_A} \sum y_i, \hat{y_B} = \frac{1}{n_B} \sum y_j$

$y_{AB} = \frac{n_A \hat{y_A} + n_B \hat{y_B}}{n_A + n_B}$

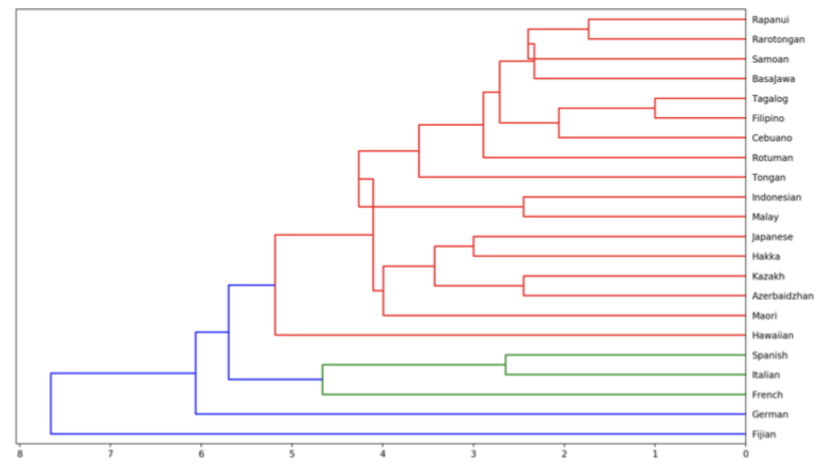
If the combination of the two groups results in a Reversal of the minimum distance (its center of mass changes), it is called Inversion, which is represented in the system tree pattern as Crossover.



5. The Centroid Method Based on the Midpoint

In order to reduce the error casued by large sample based on centroid, midpoint could be used as modification.

$m_{AB} = \frac{1}{2}(\hat{y_A} + \hat{y_B})$



6. Ward Method/Incremental sum of squares

$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$

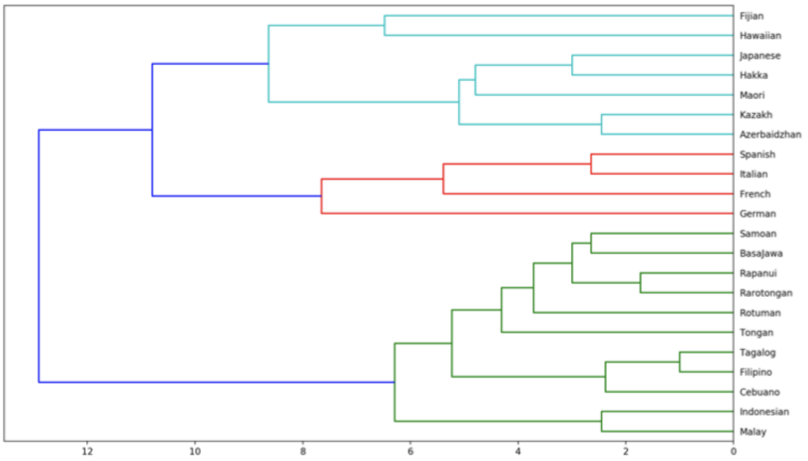
$SSE_A = \sum_{i=1}^{n_A}(y_i - \overline{y_A})'(y_i - \overline{y_A})$ for $y_i \in A$

$SSE_B = \sum_{i=1}^{n_B}(y_i - \overline{y_B})'(y_i - \overline{y_B})$ for $y_i \in B$

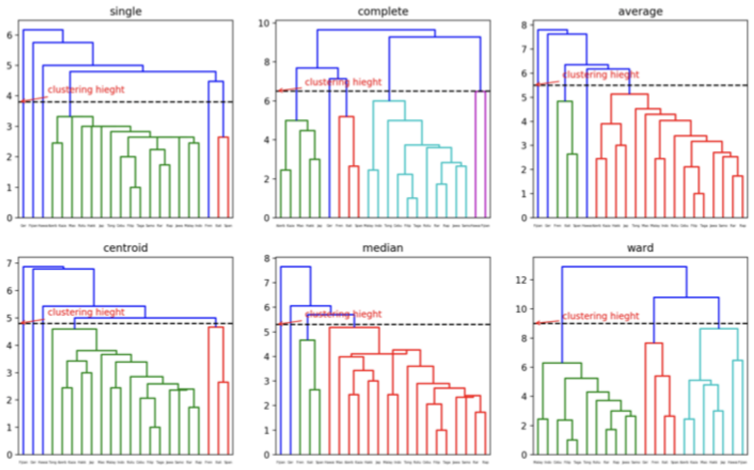
$SSE_{AB} = \sum_{i=1}^{n_AB}(y_i - \overline{y_AB})'(y_i - \overline{y_AB})$ for $y_i \in AB$

$\overline{y_{AB}} = \frac{n_A\overline{y_A}+n_B\overline{y_B}}{n_A+n_B}$

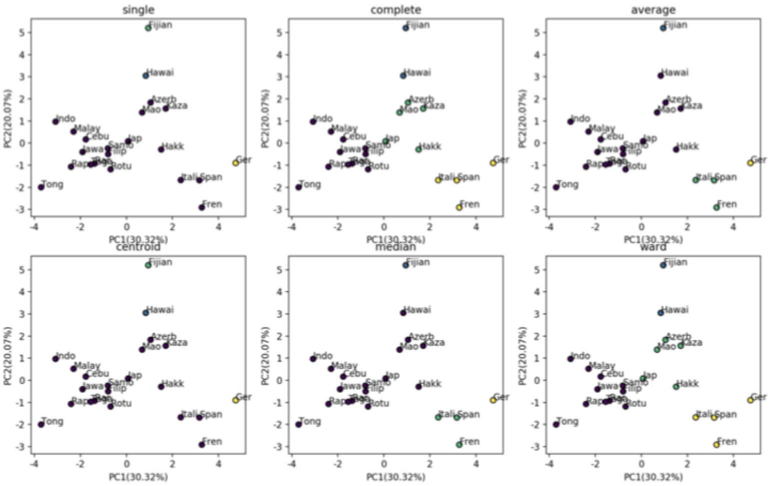
Ward method tends to combine groups with small sample size.



Totally, clustering height was texted on the graph so it was clear to see the classification through different color.



In order to visualize the clustering results, we need to reduce the dimension, in mechine learning PCA method was a nice way to do this.



Based on the six method, the result was very clear. Since the result of this experiment was affected by so many factors, such as sampling, cognates, geographical influence and so on, the algorithm could be modified continually.

4 Discussion

In 1899, Wilhelm Schmidt proposed that the Austronesian family was of the same origin as the Austronesian [3]. In 1970, American scholar argued that the Austronesian family shared the same origin with the Indo-European family. Furthermore, L. Sagart (1994) stated 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 [5]. 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 immigration 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. 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 languages 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, 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.

5 Conclusion

The basic method of study is the comparison of linguistics and biology. The sample words are filtered by the Swadesh list and language distance is measured by Levenshtein distance. Our method gets the language cluster. The phylogenetic tree is constructed by hierarchical clustering. Difference distances measurements are compared and for the sake of classification, PCA analysis is also used. When certain historical and political factors are taken into account, it is reasonable to conclude that the hypothesis that Austronesian originates in southern China is reasonable in terms of the similarity of the data. Finally, we stressed that the accuracy of this method highly relies on the sample choosing and language word translating.

6 Acknowledgement

I would like to say special thank you to my supervisor Dr.Tengwen Long, who provided me lots of guidance and useful database. I am so grateful that I can get these support in my first-try of scientific research.

References

[1] Quentin Douglas Atkinson. *From species to languages: a phylogenetic approach to human prehistory*. PhD thesis, Univ. Auckland, 2006.

[2] Peter Bellwood. A hypothesis for austronesian origins. *Asian Perspectives*, 26(1):107–117, 1984.

[3] Helmut Lukas. Pioneering studies on languages and cultures of southeast asia (w. schmidt, r. heine-geldern, p. schebesta). *Kunsthistorische Museum, Wiena*, 2006.

[4] Laurent Prévot, Chu-Ren Huang, and I-Li Su. Using the swadesh list for creating a simple common taxonomy. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 324–331, 2006.

[5] Laurent Sagart. Proto-austronesian and old chinese evidence for sino-austronesian. *Oceanic Linguistics*, pages 271–308, 1994.