VRIJE
UNIVERSITEIT
BRUSSEL

Proef ingediend met het oog op het behalen van de graad van :
Computerwetenschappen

STATISTISCHE EN COMPUTATIONELE ANALYSES VAN FONOTACTISCHE GEGEVENS UIT TALEN

Stefano Claes

2021-2022

Met supervisie door prof. dr. Bart De Boer en Peter Dekker

wetenschappen en bio-ingenieurswetenschappen

VRIJE
UNIVERSITEIT
BRUSSEL

Thesis submitted in partial fulfilment of the requirements for the degree of sciences and bioengineering sciences: Computerscience

STATISTICAL AND COMPUTATIONAL ANALYSES OF PHONOTACTIC DATA FROM LANGUAGES

Stefano Claes

2021-2022

Under the guidance of prof. dr. Bart De Boer and Peter Dekker

sciences and bioengineering sciences

Dankwoord

Eerst en vooral zou ik graag Peter Dekker bedanken voor de supervisie aan mijn Bachelor thesis. Zonder het wekelijkse overleg over het onderwerp, zouden de bevindingen gedaan in deze thesis zich niet in het stadium bevonden hebben zoals ze nu te zien zijn. Voorts werd ik steeds in de juiste weg geduwd waar nodig en teruggehouden bij het uitspreken van te grote bevindingen in de data. Het was was een eer om een co-supervisor te hebben met deze gelaagde kennis over talen. Ik wil mijn co-promotor alsnog te bedanken voor het te begrijpen van mijn druk schema als voorzitter van Infogroep, voorts wil ik hem ook nog bedanken om me steeds de realisatie te geven dat mijn thesis in een beter stadium zat dan ik zelf wou realiseren.

Met gelijke eer wil ik ook mijn promotor prof. dr. Bart De Boer bedanken. Zijn kennis over de evolutie van talen hielpen me om de uiterlijke veranderingen te maken voor de finale presentatie. Zijn kennis over fonotactische data gaven me weer inzicht over mijn Bachelor thesis en de subtiliteiten die aanwezig zijn in dit wetenschapsdomein.

Naast dit wil ik ook de mensen bedanken die het dichts bij me stonden gedurende mijn studies, zonder hun was er geen sprake van deze bachelor thesis. Ik zou mijn ouders graag willen bedanken om me te laten studeren aan de VUB zonder te twijfelen omtrent de kennis die ik had opgedaan in de vorige studies die ik had gevolgd in het middelbaar. Mijn grootouders zou ik graag willen bedanken om steeds een kalme plek te voorzien waar ik kon studeren. En als laatst zou ik graag nog mijn vriendin bedanken om mijn stres gehalte te verlagen gedurende drukkere periodes en me te motiveren om verder te werken aan deze thesis.

Samenvatting

Talen evolueren doorheen de tijd en vormen hierdoor een soort van familieboom, dit soort bomen heten fylogenetische bomen of een fylogenie in evolutionaire biologie (Felsenstein & Felenstein, 2004). In deze Bachelor thesis wordt de fylogenetische boom van Austronesische talen gebruikt, deze boom zal vergeleken worden met de historische data om vervolgens na te gaan met welke statistische significantie deze boom overeenkomt met de historische data. De historische data die zal gebruikt worden om deze fylogenetische bomen te testen zullen enkelvoudige klanken zijn en dubbele klanken ook gekend als fonen en dubbele fonen. Deze klanken worden gehaald uit woordenlijsten van verschillende talen.

Acknowledgements

First and foremost , I would like to thank Peter Dekker for his supervision on my Bachelor's thesis. Without this weekly talks about the subject, the foundlings which we will see in this thesis wouldn't be realized. Furthermore I was always guided in the right direction when making mistakes and tranquilized when making too euphoric findings in the data. It was a great pleasure having a co-supervisor who had a very layered knowledge in evolution languages. I want to thank my co-promotor for apprehending my busy schedule as a President from Infogroep and giving me the realization that the thesis was in a better stadium than I would realize myself.

With equal gratitude I want to thank my promotor prof. dr. Bart De Boer. His knowledge in evolution of languages helped me making adjustments to what will be the final presentation. His knowledge in phonotactic data gave me more insight into this Bachelor's thesis and the subtleties that lie inside this science-domain.

Next to this I want to thank my close relatives, without them this Bachelor's thesis wouldn't be here. I would like to thank my parents for letting me study at the VUB without doubting on my prior knowledge gained from the studies I did before. I would like to thank my grand-parents for always keeping a calm environment where I could study. And finally I would like to thank my girlfriend to distress me during more busy moments and motivating me to write further for this thesis.

Abstract

Languages evolve trough time and develop some kind of family tree, this tree is called a phylogenetic tree or phylogeny in Evolutionary biology (Felsenstein & Felenstein, 2004). In this Bachelor's thesis the phylogenetic tree of Austronesian languages is used, this tree will be compared with historical data to see too which statistical significance this tree reflects the found historical data. The historical data used to test these phylogenetic trees are single-segment sequences and double-segment sequences also known as phones and biphones. These are extracted from lexicons from different languages.

Contents

1 Introduction **4**
1.1 History of phylogenetic research 4
1.2 Motivation for new method 4
1.2.1 Loanwords 5
1.2.2 Semantic shifts and grammatical changes 5
1.2.3 Phonotatctic data 5
1.3 Phylogenetic signal tests 6
1.3.1 D statistic 6
1.3.2 K statistic 8

2 Method **11**
2.1 Materials 12
2.1.1 Reference phylogeny 12
2.1.2 Phonotactic data 12
2.2 Data pre-processing 13
2.2.1 Character segmentation 13
2.2.2 Feature engineering 15
2.2.3 Data formatting 16
2.3 Phylogenetic signals 18
2.3.1 D statistic 18
2.3.2 K statistic 20

3 Results **20**
3.1 D statistic 21
3.1.1 Austronesian languages 21
3.2 K statistic 28
3.2.1 Austronesian languages 28
3.2.2 Western Lamaholot dialects 35

4 Conclusion **39**

5 Discussion and future work **40**

References **41**

1 Introduction

1.1 History of phylogenetic research

Phylogenetic trees describe a proposed evolution from languages, these tree's are constructed by looking at the linguistics from different languages. By example they look like the figure displayed below, note that this is solely a figure for demonstration and contains no concrete data for this research.

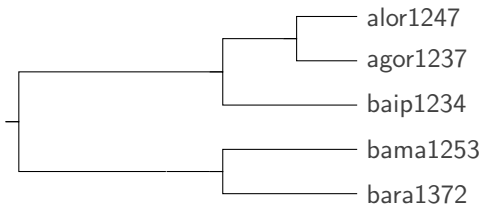


Figure 1: An example of a language phylogeny with the Glottocode of the language at the tip. Glottocodes are unique codes that have been made to identify each language/dialect with a code.

Well documented historical linguistic data creates the opportunity to build these phylogenetic trees (Blust, 2009). The historical data covers different sorts of changes that occur in languages, a summation of this can be made as follows:

- Phonological change of languages: The organisation of sounds in languages and how they have evolved.
- Grammatical change of languages: The way how grammars of languages have evolved throughout time.
- Semantic changes: The way how meaning of words evolve.

Combining these sources make it possible to generate a phylogentic tree using the comparative method, this involves a method that can be summarized as a set of instructions (Durie & Ross, 1996). During this research we will use the phonological data to question the phylogenetic tree's made up for a set of languages. Language research is mostly done by classifying older documents containing information of the language. This classification however is prone to errors since this is a combination of the work of multiple researchers. All of them maybe having their own method of conducting research. This can lead to discrepancies between languages where there are non, or similarities between languages which aren't related. To solve this issue in this Bachelor's thesis phonemes will be used as our main set of data. Next to this the reference phylogeny made by the comparative method will be used. This reference phylogeny will be tested against the phonemic data found of these Austronesian languages.

1.2 Motivation for new method

During this research the method of following paper has been followed (Macklin-Cordes, Bowern, & Round, 2021). Instead of using Pama-Nyungan vocabularies we use Austronesian vocabularies extracted from LexiRumah(Edwards, Kaiping, & Klamer, 2022).[1] The phylogenetic tree is

[1]LexiRumah (from Lexicon and Indonesian Rumah = house)

extracted from Glottolog, this tree is the result from combining multiple sources that conducted classification of these languages (Hammarström, Forkel, Haspelmath, & Bank, 2021). Former studies reflected on testing the significance of phylogenetic trees by comparing them to the historical lexical data using a comparative method (Felsenstein, 1985a). This however is prone to errors due to multiple factors, the first factor we can address are loanwords.

1.2.1 Loanwords

Loanwords [2] make it more difficult to construct these phylogenetic trees since they remain the same in multiple languages and thus make it difficult to find the common ancestor language. The reason to prefer single or double segment sounds to test this phylogenetic tree lies in the fact that these loanwords are still subject to the structural constraints of languages (i.e., The word for computer is the same in Dutch as it is in English although the pronunciation of the word differs greatly) (Kang, 2011). Meaning that in a loanword the single and double segment characters have a great chance of differing from each other which raises the amount of historical data that can be extracted from languages.

1.2.2 Semantic shifts and grammatical changes

The appearance of semantic shifts[3], also generate difficulties in reconstructing a language phylogeny (Blust, 2010). This also creates noise in the data which makes it indistinguishable from the phylogenetic signal needed for constructing the tree. Languages also have grammatical rules (i.e. The way past tense is formed) , these rules however evolve at a faster rate than words in languages (Greenhill et al., 2017). Making it difficult to construct a phylogenetic tree on the basis of this data.

1.2.3 Phonotatctic data

The set of rules for possible sound sequences is called phonotactics, in this research we will use phonetic data retrieved from word lists found in the online database of LexiRumah. Using phonetic data increases the amount of data, this due to the fact that words exist of multiple phonemes. Some languages are less documented as others, making it difficult to compare their vocabularies with other languages. Since the phonetics of the languages form some similarities between the words documented of this language. This similarities can also be shared with other languages, so we can compare languages with more statistical power. Mainly the conclusion can be opposed that there is more phonetic data of languages. This phonetic data is independent from vocabulary or grammar changes in languages and by so makes it a good anchor point to compare this with the proposed language phylogeny. In this Bachelor's thesis we want to compare the possible phoneme sequences as in biphones with the reference phylogeny of a set of languages. With all this in mind the research question can be defined as follows:

Research question

How well does phonotactic data of languages match with the language phylogeny?

Furthermore a big part of the method used in this Bachelor's thesis is based on the method devised in (Macklin-Cordes et al., 2021). In this research however other well documented families

[2] a word adopted from a foreign language with little or no modification.
[3] semantic change is a change in one of the meanings of a word

of languages are researched. Here the phylogenetic tree of Austronesian languages is researched, and a subset of the Austronesian languages. The subset researched in this thesis are the western Lamaholot dialects. Moreover the research isn't limited over double segment phonetic characters, here also single segment phonetic characters are examined.

1.3 Phylogenetic signal tests

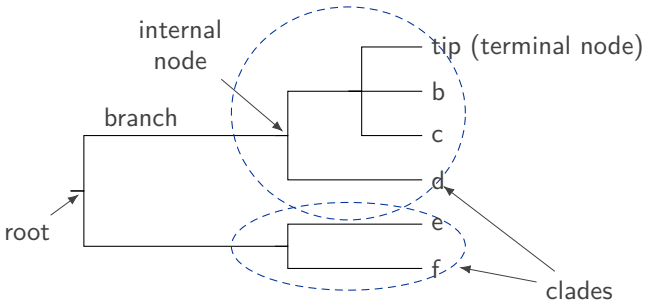
This phonotactic data consist of data extracted from phonetic wordlists, meaning that these wordlists contain the phonemes from the desired language. Phonotactic data is independent from the existence of loanwords or semantic changes, thus increase the amount of usefull data. The data will be used to track phylogenetic signals in the phonotactic tree. A phylogenetic signal is the tendency for related species[4] to resemble each other more than the rest of the phylogenetic tree (Blomberg, Garland, & Ives, 2003). To test these phylogenetic signals the K statistic (Blomberg et al., 2003) and D statistic (Fritz & Purvis, 2010) is used.

1.3.1 D statistic

For the D statistic binary values are used meaning the existence of phonemes in a language will be checked. Phones and Biphones will be used in this statistic, each of these phonemes will have a certain discrete outcome describing if it has a phylogenetic nature throughout the phylogenetic tree. The D statistic is calculated by following formula:

$$D = \frac{\sum d_{obs} - mean(\sum d_b)}{mean(\sum d_r) - mean(\sum d_b)} \tag{1}$$

To explain the D statistic this tree will be used to clarify the subtleties.



Calculating the $\sum d$ is done from the tips of tree up to the root-node each time calculating the sum of differences between the sister nodes. Important to notice is that a feature[5] at a tip node is either 0 or 1 depending on its precedence. This rises to two possibilities all nodes are the same (In which case the, the difference 0) else the difference is 1. The internal node will get the average value from their daughter tips (which, in a fully bifurcating phylogeny, will either be 0, 0.5 or 1). This process is repeated until all nodes of the tree have been reached (Macklin-Cordes et al., 2021). The D value is calculated multiple times each time with different data at the tips coming from a randomly shuffled version from the tree and one generated by a model of Brownian motion, this model of Brownian motion has a certain threshold so it evolves at a realistic rate expected from languages. Both of these models are needed to compare the documented evolution

[4]Species being languages in this Bachelor's thesis
[5]In this thesis a phone or biphone

next to structural duplicates that have respectively random values at the tips[6] and a Brownian threshold model. A phylogenetic signal in a trait is motion of species being languages in our case to resemble each other more than other languages. This is clearly no random motion, so comparison is done with random motions to check the presence of a phylogenetic signal.

Parameters D statistic (Fritz & Purvis, 2010)

- $\sum d_b$: Sums of differences expected under Brownian motion.
- $\sum d_r$: Sums of differences expected from a random phylogenetic signal.
- $\sum d_{obs}$: Sum of differences between two ends of each branch in the phylogeny

These D values depend on the presence of traits in the tips of the tree, figure 2 displays how model distributions between tips compare to the signal gained from the D test.

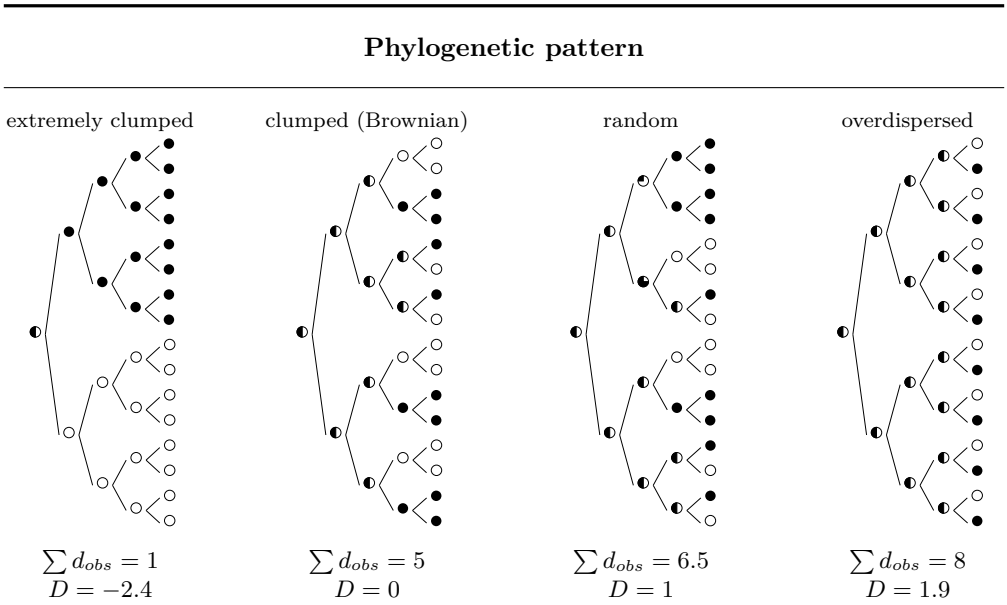


Figure 2: Phylogenetic patterns and their respective D value, figure based on (Fritz & Purvis, 2010)

Extremely clumped

This phylogenetic pattern corresponds to a certain phoneme being present in every tip descending from one of the root branches, the other root branches should in this case show a total absence of this phoneme. This can be the result of a strong tendency to resemble each other more than other cultures in the phylogenetic tree (i.e. Two cultures living on different islands which never have been in contact with each other).

[6] Which destroy the phylogenetic signals

Overdispersed
The exact opposite being overdispersed can be explained with another metaphor, since phylogenetic tree's are also used in biological research (Kawahara et al., 2019) this will be used as metaphor. Imagine subdividing the tips of the phylogeny between male and female species, furthermore take this also as one of the traits/features to be used in the D statistic. This will lead to high D statistic values and show with this metaphor no extra information for detecting phylogenetic signals.

Brownian and Random reference phylogeny's
Subsequently the two other phylogenetic patterns will be used as comparative phylogeny's. The D statistic is standardized so is applicable for different sets of data and different shapes of tree's. The disadvantage from the D statistic is that it requires a big sample size (>50) constraining the amount of languages we could test. These D values gained from the D statistic all have a two ρ values. The first value is for D = 0, corresponding for the significance if a phylogenetic signal is present for a certain phoneme. The second ρ value is for D = 1, this corresponds for the significance that the phoneme shows a random distribution relative to the phylogenetic structure. As said before, $\sum d_r$ and $\sum d_b$ consist from the sums of sister tip/clade differences. These differences can be seen as scores, the scores from $\sum d_r$ and $\sum d_b$ will be compared with each other. We will define the ρ_1 ($H_0(D = 1)$) as the fraction of scores from $\sum d_r$ smaller than $\sum d$, to the contrary we will define ρ_2 ($H_0(D = 0)$) as the proportion of scores from $\sum d_b$ greater than $\sum d$ (Macklin-Cordes et al., 2021).

1.3.2 K statistic

A second statistic used in this Bachelor's thesis is the K statistic, this statistic can be used for continuous values (Blomberg et al., 2003). Prior research with the K statistic has been conducted on comparing the reference phylogeny of animal species using their body/bone measurements as features (Klaczko, Ingram, & Losos, 2015) (Cubo, Ponton, Laurin, De Margerie, & Castanet, 2005). In this bachelor's thesis we will use the phonemes frequencies as features to compare this with the language phylogeny. To calculate the K statistic multiple resources are needed, consisting of:

Resources needed for K statistic (Blomberg et al., 2003)

1. Character data, in our case phonemes.
2. A reference phylogeny which has to be an accurate representation of the phylogenetic history and been generated independent from 1.
3. A Brownian model of evolution.

Important to notice is that this Brownian model will simulate increases or decreases in usage of phonemes trough out the phylogenetic tree. The parameters of this Brownian model need to be fine-tuned to showcase realistic increases or decreases in usage. Thus this is dependable on how fast the phonetics of languages evolve through time.

The K statistic has the desirable property to have with a sample size of at least 20 or more languages a type 1 error rate at a nominal $\alpha = 0.05$, the power[7] of the test equals 0.8. This

[7]Power is the probability of accepting the alternative hypothesis, when the null hypothesis is wrong.

power only rises when increasing the amount of languages with phonotactic data. As in the D statistic each phoneme will contribute a certain value, the K value in this case will be calculated as follows:

Derivation K statistic (Blomberg et al., 2003)

$$MSE_0 = \frac{(X - \hat{a})' \cdot (X - \hat{a})}{n - 1}{}^{a} \tag{2}$$

- $\hat{a}$ is the estimate of the phylogenetically correct mean.(Garland, Midford, & Ives, 2015)
- X is the data vector containing n values, character frequencies in this research.

To put more emphasis on $\hat{a}$, this is calculated by computing the phylogenetically weighted estimate of the mean value for a set of species. This mean is actually the estimated value at the root of the phylogeny. This is calculated as follows in the R package Motmot by the function phyloMean (Puttick, Ingram, Clarke, & Thomas, 2019).

$$MSE_0 = \frac{(U - \hat{a})' \cdot (U - \hat{a})}{n - 1} \tag{3}$$

Where $U = DX$, U is formed by transforming the X vector with the least squares procedure. The matrix D on the other hand satisfies the equation DVD' = I. In this equation the V corresponds to the variance-covariance matrix and I embodies the identity matrix. Combining these two extensions of formulas gives rise to following formula:

$$\frac{MSE_0}{MSE} = (\frac{1}{n-1}) \cdot (tr(V) - \frac{n}{\sum\sum V^{-1}}) \tag{4}$$

- $tr(V)$ also known as the trace of a matrix, is the sum of all diagonal elements
- $\sum\sum V^{-1}$ is the sum of all elements from the inverse matrix of V.

Furthermore this K value is calculated by dividing the observed data to the expected data generated under a Brownian evolution model.

$$K = \frac{Observed(\frac{MSE_0}{MSE})}{Expected(\frac{MSE_0}{MSE})} \tag{5}$$

As a consequence, when $K < 1$ then close relatives in languages have less tendency to mimic each other than expected under the Brownian motion. The opposite $K > 1$ is also possible, then close relatives in languages show more similarities than expected under the Brownian motion. When the Brownian model parameters are correctly chosen, then the optimum K value for a phylogenetic signal is a value near 1.

[a] Note that apostrophe corresponds to the transpose.

Next to this the K statistic will present a randomization procedure to test to which degree a phylogenetic signal in the dataset is statistically significant.

Phylogenetic independent contrasts

The randomization procedure used by the K statistic is based on (Felsenstein, 1985b) phylogenetic independent contrasts (PIC) method. The key insight was that two features for example x (body size) and y (brain size) could not be seen as independent between 2 sister tips. The same can happen between different phonemes if they're for example very frequently used together. This dependence raises a few problems when analyzing some phylogenetic trees. Take for example next example:

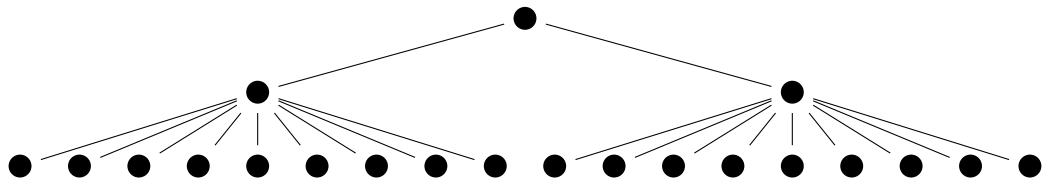


Figure 3: A demonstration of a reference phylogeny which can be difficult to to construct due to possible minor differences between tips. The figure is based on the previous work of (Felsenstein, 1985b).

Please note that the dots may not be confused with their meaning in figure 2, they don't represent the absence or presence of a trait.

This shows the data of 20 species in which 2 pairs of 10 species are closely related, they evolved from a common ancestor but might have diverged feature wise. Collecting the data from this features could lead to confusing plots, which might overshadow a separation between 2 groups of diverged species. A possible plot could look like following:

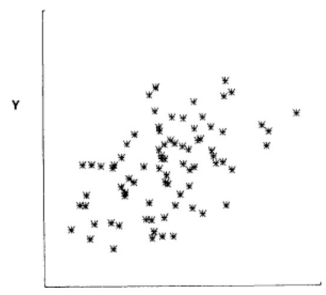


FIG. 6.—A typical data set that might be generated for the phylogeny in fig. 5 using the model of independent Brownian motion (normal increments) in each character.

(a) All data points without separation of closely related species.

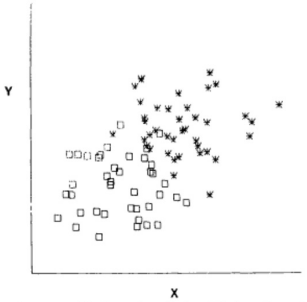


FIG. 7.—The same data set, with the points distinguished to show the members of the 2 monophyletic taxa. It can immediately be seen that the apparently significant relationship of fig. 6 is illusory.

(b) All data points with separation of closely related species.

Figure 4: Images consulted from (Felsenstein, 1985b)

Another insight in (Felsenstein, 1985b) was that the differences of a feature between adjacent tips must be independent. By calculating these differences it is possible to quantify a rate of evolution from this feature. Mathematically these contrasts are calculated by following formula:

$$PIC = C_{ij} = X_i - X_j \tag{6}$$

The rate of evolution of a certain feature can be calculated by dividing this value by the variance.

$$S_{ij} = \frac{X_i - X_j}{v_i + v_j} \tag{7}$$

To calculate the contrasts of internal nodes, following method is used.

Method for calculating contrasts (Felsenstein, 1985b)

1. Find two tips in the phylogeny that are adjacent (say nodes i and j) and have a common ancestor, say node k.
2. Compute the contrast $X_i - X_j$. This has expectation zero and variance proportional to $v_i + v_j$.
3. Remove the two tips from the tree, leaving behind only the ancestor k, which now becomes a tip. Assign it the character value calculated by following formula:

$$X_k = \frac{(\frac{1}{v_i}) \cdot X_i + (\frac{1}{v_j}) \cdot X_j}{\frac{1}{v_i} + \frac{1}{v_j}} \tag{8}$$

4. Lengthen the branch below node k by increasing its length from v_k to $v_k + \frac{v_i \cdot v_j}{v_i + v_j}$.

This lengthening is needed because the weighted average calculated in 3 does not compute the exact value of the feature, it only estimates this. The error of this estimation is statistically indistinguishable from an extra burst of evolution after node k. This will reduce the number of tips in the tree by 1. Step 1 to 4 will be repeated until there is only 1 tip left in the tree.

As mentioned before the tree undergoes a series of random permutations, if the original tree has a variance of PICs lower than 95 percent of these random permutations. Than the null hypothesis of no phylogenetic signal can be rejected at the 95 percent confidence level.

2 Method

The method used in this Bachelor's thesis consists of collecting phonemic word lists and segmenting these word lists into phonemes, as you might recall this method is based on (Macklin-Cordes et al., 2021).Next to this a reference phylogeny is retrieved to compare this to the historical data retrieved from the word lists. These comparisons happen by the D statistic for binary data and the K statistic for continuous data. The results will show if according to the statistic some features (phonemes) show a phylogenetic signal throughout the history of the phylogenetic tree. Part of the method will be by first calculating the D and K statistic for all the mutual languages between the Austronesian language phylogeny from Glottolog and the historical data retrieved from Lexirumah. After this the research will be continued by smaller subsets of the Austronesian languages, in this case only enough data was found for the Western Lamaholot dialects. With a mutual amount of 22 languages only the K statistic is calculated for the Western Lamaholot dialects.

2.1 Materials

As of materials two resources have been used, the phonemic data is retrieved from Lexirumah (Edwards et al., 2022). The Reference phylogeny is retrieved from Glottolog (Hammarström et al., 2021). Both of these resources contain different languages, only the intersection of languages can be used.

2.1.1 Reference phylogeny

During the making of this research, a Phylogenetic tree of the Austronesian languages was used. This tree consisted of 1270 tips or languages, the research for this classification was heavily based on the work of (Blust, 2009). Since the phylogenetic tree consists of 1270 tips, the phylogenetic tree is difficult to display and won't be shown. Noteworthy is that the root of this language tree contains and should represent the Proto-language of this tree, although uncertain as exact figure it is believed that Proto-Austronesian dates from 6000BP[8] (Blust, 1995).

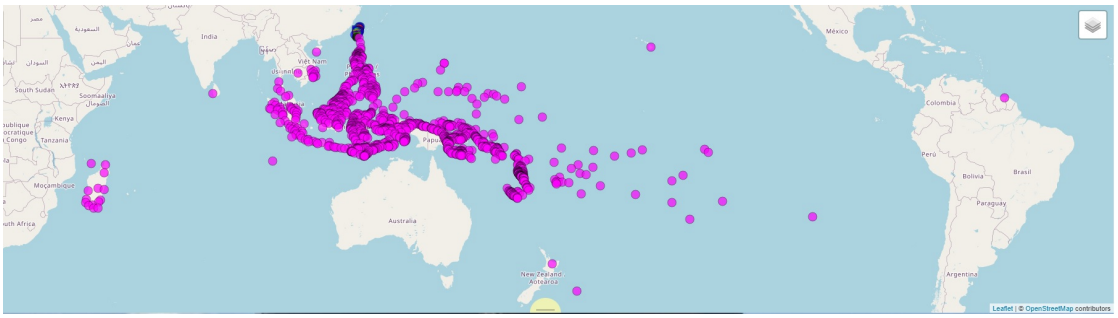


Figure 5: The geographical spread of Austronesian languages , image consulted from (Hammarström et al., 2021)

2.1.2 Phonotactic data

Next to the reference phylogeny, word lists in form of the International Phonetic Alphabet (IPA) is used. The dataset of lexirumah consists of 173 languages which makes it substantially smaller than the dataset of Glottolog.

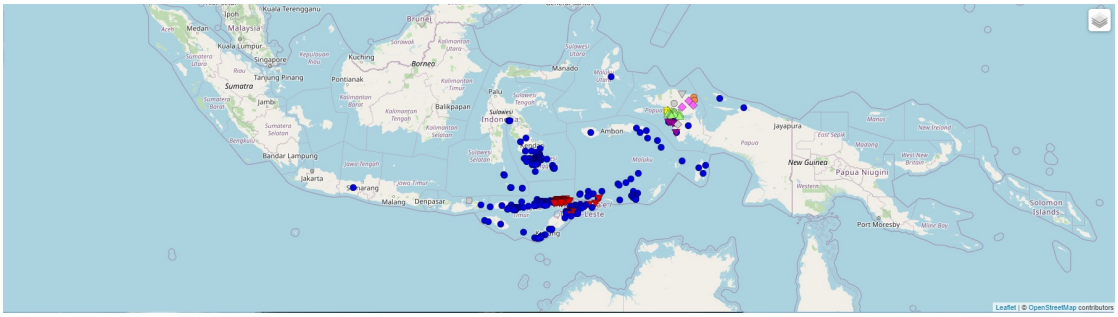


Figure 6: The geographical spread of Austronesian languages (Austronesian languages are represented by blue dots), image consulted from (Edwards et al., 2022)

[8]Before present

The word lists for each language differ in amount of words, in (Macklin-Cordes et al., 2021) they put a threshold of 250 for a minimum amounts of words. In this bachelor's thesis however we put no threshold.

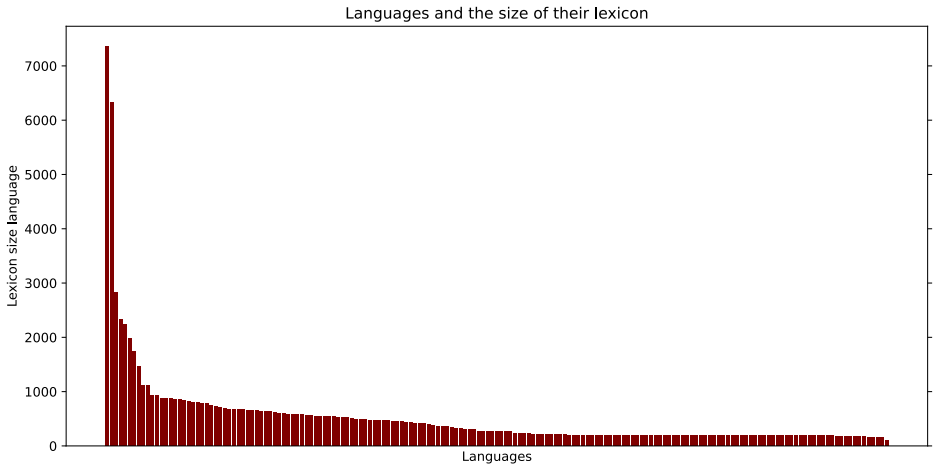


Figure 7: The sizes of the vocabularies of each Austronesian language retrieved from the Lexirumah database

The biggest word-list contains 7362 words while the smallest word-list contains 108 words.

2.2 Data pre-processing

Since the data exists of phonetic word-lists, there needs to be some prepossessing. This is because a major part of the research is conducted on phones and biphones. The first step in this multilayered pipeline consists of extracting all possible phonemic characters that occur in the phonetic word-list. This phonemic characters will be regarded as the features of the languages which will be tested for phylogenetic signals.

2.2.1 Character segmentation

Character segmentation will in the case of phones take every phonemic character found in the language and add it to the data. The case of biphones is a little more subtle, every combination of two phonemic characters behind each other are taken, but a word starting or ending with a certain phonemic character is also judged as a biphone[9]. This eventually leads to some data which give a less good representation of a biphone. This is due to the fact that a biphone is defined as follows: "Biphones are understood to be just left or right context dependent (mono)phones"(Dobrisek, Mihelic, & Pavesic, 1999). To clarify a biphone is a sound retrieved as a combination of two phones put sequentially after each other. These phones can't be altered in pronunciation when defining a biphone. When a phone gets altered during their combination a diphone is formed. Diphones are essentially the sounds from transitions between two phones. Since it is difficult

[9]The character to denote this in the data is #

to distinguish both when extracting from raw word lists we will use the terms interchangeably. Next to this not enough papers where found to make the effects of diphones on our research to be neglected. Furthermore multiple aproaches have been tried to do this segmentation, in linguistics multiple phonetic alphabets exist. The most general alphabet is preferred here because all the data is based on many independent researches of languages, these can be found at `https://lexirumah.model-ling.eu/sources`. Two approaches can be considered for the phonemic data one of them being the IPA (International phonetic alphabet) the other one being the ASJP (Automated Similarity Judgment Program).

International Phonetic Alphabet
The International Phonetic Alphabet is an alfabet consisting of 107 segmental letters, 44 diacritics and four extra-lexical prosodic[10] marks[11]. The data from LexiRumah is made up in the IPA format, although this is an international standard the data at Lexirumah is made up on the findings of multiple researchers. These researchers may have made up their own methodology on documenting these languages, as a consequence multiple researchers may document the same phoneme differently because of their own interpretations. Subsequent research on these phonemes can lead to outliers due to the usage of diacritics. Diacritics can make very unique characters which make it very difficult to cluster the data [12]. To have less noise from this erroneous data the Automated Similarity Judgment Program is considered.

Automated Similarity Judgment Program
The Automated Similarity Judgment Program (ASJP)(Brown, Holman, Wichmann, & Velupillai, 2008) is a database that consists following amounts of data(as of June 29, 2022):

- Word-lists: 9788
- Synsets: 400,537
- Words: 445,359
- Distinct Glottolog families: 395

The database was originally intended to compare the similarity of words with the same meaning from different languages. Digging further in this the goal was to classify these languages. Next to this a code has been created called ASJPcode(Brown, Holman, & Wichmann, 2013a), this code is multi to one mapping from IPA. ASJPcode will map all characters retrieved from IPA strings to 41 different phonemes. The clustering[13] of these characters happen via a two step algorithm.

[10] properties of sounds based on in which situation they're used, for example for marking a difference in sound when shouting a word.
[11] `https://www.internationalphoneticassociation.org/content/full-ipa-chart`
[12] A research where this deviated from.
[13] Noteworthy is that the IPA encoding is in fact also a clustering of characters, some Austronesian languages are documented in phonemic encodings that are non existent in IPA(Blust, 2009).

ASJPcode clustering for IPA (Brown et al., 2013a)

The algorithm only compares words that are retrieved from the same language phylogeny, using only this words the algorithm will start it's 2 step verification.

1. **Potential correspondence**

First the algorithm searches for a single word from the standardized 40 item list, here two languages will be compared if a word is found that is written phonetically the same way except one character. If this happens the character is marked as a potential correspondence.

2. **Actual correspondence**

An actual correspondence is when at least two words are found that only differ with the same phonetic character.

The algorithm also deals with some special cases, for example a symbol can also correspond to the abscence of symbol in another word(Denoted by the $\emptyset$ symbol). An overview of this mapping can be found in the appendix, this is copied from next paper with supplementary materials (Brown, Holman, & Wichmann, 2013b).

2.2.2 Feature engineering

The data retrieved from the character segmentation needs to be counted, to have the overal distribution of these phonemic characters. As mentioned in the research question, D-test features will be binary and show the presence or absence of a character. K-features however will show the frequency.

D statistic features
Two different datasets are created for the D-test. One is for the single segment phonemic sounds, the other one for the double segment sounds. As mentioned in 1.3.1 there will be data describing the presence of a phoneme for each language. In short 1 will mark the presence of a phone, 0 will mark the opposite. When doing so for the phones we find that the data consists of about 41 characters, ranging over 133 Austronesian languages. A possible research included subdividing the Austronesian languages in regions that had met in history and possibly started resembling each other more and more. And thus so showing a phylogenetic signal. However this isn't possible due to the minimum of 50 entries to state something with enough power with the D-test. Biphones of the Austronesian languages are researched too, since this includes all combinations of 2 characters more features are available. This increases the data to 41^2 or 1681 features. When evaluating the data with the D statistic, a feature can't have the same value in each row. When this happens the phoneme needs to be deleted since it wouldn't give us more information about the evolution of a language phylogeny.

K statistic features
K statistic features differ in the fact that they represent frequencies, furthermore they represent conditional probabilities. Two types are defined, by definition they are called the Markov chain forward transition probability and the Markov chain backward transition probability.

Markov chain transition probabilities (definition)

Markov chain forward transition probability
Consider a biphone "xy", if defining the forward probability following formula is used:

$$\frac{P(xy)}{p(x)} \tag{9}$$

To clarify the amount of occurences of "xy" will be devided by the amount of occurences by the single segment character "x".
Markov chain backward transition probability
It's obvious that the backward transition probability will define the amount of occurences of "xy" devided by the character "y". As follows a justification with the mathimatical formula:

$$\frac{P(xy)}{p(y)} \tag{10}$$

The forward transition probability is used because strings aren't constructed in most cases as independent segments. In most cases a certain phone say "y" in the phone "xy" is dependent on the phone that came before it ("x" in this case), to minimize this effect forward transition probabilities are used. The opposite is true as well so for this case the backward probability are utilized. These frequencies are based on their frequency in word lists, note that the frequency of a word occurring in their regular language has no effect on the result. By so every word is only counted once and the result isn't effected by word frequency effects. However this being desirable during this research, coined words show some effect on the phoneme frequencies since speakers of that language may tend to use that word more frequently (Zuraw, 2000; Coleman & Pierrehumbert, 2003; Ernestus & Baayen, 2003; Albright & Hayes, 2003; Eddington, 2004; Hayes & Londe, 2006; Gordon, 2016). As mentioned in (Macklin-Cordes et al., 2021), this raises the possibility to extending this research with word frequencies included.

2.2.3 Data formatting

All the data can't be used for processing since the D statistic and K statistic can't process some types of column data. The D statistic can't handle data where all rows have the same value for a certain column. The K statistic on the other hand can't handle data that solely exist of "nan" values. This decreases the amount of data since a lot of the proposed biphones don't exist in the range of Austronesian languages. The reductions lead to following amounts of data:

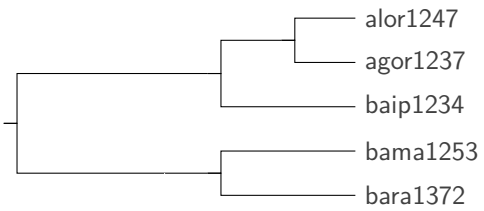
Data amounts

Note that more tests will be conducted and some cover a subset of the Austronesian languages. The subset that is researched for phylogenetic signals in this Bachelor's thesis is the Western Lamaholot dialects.

- D statistic for phones of the Austronesian languages:
  - Amount of languages: 133
  - Amount of single segment phonemes: 41
- D statistic for biphones of the Austronesian languages:
  - Amount of languages: 133
  - Amount of double segment phonemes: 151
- K statistic for biphones of the Austronesian languages:
  - Amount of languages: 133
  - Amount of double segment phonemes: 314
- K statistic for biphones of the Western Lamaholot dialects:
  - Amount of languages: 22
  - Amount of double segment phonemes: 98

As menstioned before, conducting this research on lesser documented dialects is not possible since of the minimum amount of data required to state phylogenetic signals by a certain statistical power.

After all this the data is ready to be combined and used by the R package functions. At figure 8 this is visualized with a small example. Note that the phylogeny doesn't represent the phylogeny used during this research, it's just to give an idea of the different data used in this thesis.



(a) An example of a language phylogeny with the Glottocode of the language at the tip.

Name	a_e	c_e	d_g	b_p
alor1247	1	0	1	0
agor1237	1	1	1	0
baip1234	0	1	1	0
baip1234	1	0	1	1
bama1253	0	1	1	1
bara1372	1	1	1	1

(b) An example of the phonemic data, Glottocodes as row names and biphones as features. The column data itself is an example for D statistic data since it is binary.

Figure 8: Left the language phylogeny, right the historical data.

2.3 Phylogenetic signals

The main objective of this Bachelor's thesis is to find out which phonemes show phylogenetic signals. For the D statistic a modded function of "phylo.d" is used, this function originates of the caper(Orme, 2013)[14] package in R. For the K statistic the function "multiPhylosignal" is used. This time the Picante package (Kembel et al., 2010) is used in combination with some additional analysis. Both statistics will eventually give us some distribution of values, these values will be plotted and will have some statistical properties that will be researched. The points of interest here are:

- Mean
- Standard deviation
- Median
- Skewness
- Kurtosis

2.3.1 D statistic

The D statistic calculates more than just the D value per phoneme, as mentioned before 2 null hypotheses are calculated each of them responsible for the significance of a certain phylogenetic tendency. The tendencies can be as follows:

- Phylogenetic(significantly below 1 and not significantly different from 0): The phoneme shows a phylogenetic signal troughout the phylogenetic tree. There is a clear track of some phonemes being present in closely related branches, they have the tendency to mimic each other more than more distant branches.
- Random(significantly above 0 and not significantly different from 1): The phoneme has a random appearance in the phylogenetic tree, some very distant branches may have the phoneme while close sister languages may have this phonemic sound not in common.
- More dispersed(significantly above 1): The phoneme is scattered in an unusual way, closely related languages may not have the sound in their data while more distant languages do have the sound. This needs to happen in some structural way since otherwise a random structure is considered.
- More clumped (significantly below 0): There is no tendency of languages mimicking each other, the right branch of the root may all have the sound the other branch shows a complete absence.
- Indeterminate (neither H0 rejected, not significantly distinct from 0 nor 1): Due to contradiction or lack of data, no conclusion about the tendency can be stated.
- $0 < D < 1$ (both H0s rejected, significantly above 0 and below 1): The historical data rejects the null hypothesis of a random distribution as the null hypothesis of a phylogenetic distribution.
- N/A: The sound isn't present in the historical data (This is no return value of the function but is used in some plotted data)

[14] Comparative Analyses of Phylogenetics and Evolution in R

Phylo.d function (Orme, 2013)

Arguments

- Data: A data object transformed to make it comparative, or in short "comparative.data" and "data.frame" object.
- Phy: When the data isn't transformed to a "comparative data" object, an object of class "phylo" is needed.
- Names.col: The name of the first row in the in the dataframe, with the assumption that this column contains all the tips of the phylogeny. In our case this is the column with all the Glottocodes.
- Binvar: A certain phoneme that will be the binary variable of interest.
- Permut: The amount of permutations to be checked in the randomisation test.

Note that the Phylo.d function contains more arguments, however these are never used during the process.

Return values

- DEstimate: The estimated D value for a certain phoneme.
- Pval1: A p value that indicates whether the DEstimate is significantly different from 1. So to clarify this is the significance that a phoneme shows a random distribution relative to a phylogenetic structure.
- Pval0: A p value that indicates whether the DEstimate is significantly different form 0. Conversely this is the significance of a phylogenetic signal being present for a certain phoneme.

Note that yet again more return values are present, however these show lesser relevance in this research and thus aren't mentioned.

Calculating the D statistic will be done for single and double segmented characters, biphones however will be constructed by two phones or a word boundary which is denoted by the character "#". Moreover the biphone data consists of 3 values which are transformed to a binary format to comply with the R package functions. To sum up:

- 1: The biphone is present in the language
- 0: The biphone (take for example "xy") isn't present in the language but both phones "x" and "y" appear in the language.
- nan: The biphone "xy" isn't present nor are one or both of the phones "x" "y".

It's important to notice that all nan values will be projected to 0. Furthermore the values $\sum d_r$ and $\sum d_b$ are calculated via the procedure explained in 1.3.1. The amount of permutations used in this calculation are 10000 per biphone or phone. As of the statistical significance we use a standard value of $\rho = 0.05$. For a quick overview the D test can be summarized as demonstrated in the frame below.

D statistic test method

For each biphone or phone, only calculate the D value if there are at least 50 non "nan" values and these values aren't all the same. For each phone or biphone calculated the next three values:

- The D value
- ρ_0 for the first null hypothesis $H_0(D = 0)$. (The significance of a phylogenetic signal being present for this character.)
- ρ_1 for the first null hypothesis $H_0(D = 1)$. (The significance of a phoneme showing a random distribution relative to the phylogenetic signal.)

A result will be constructed by combining the results from the D value and the two ρ values.

2.3.2 K statistic

During the calculation of the K statistic the function "multiPhylosignal" is used, since extra analysis is preferred an extra dataframe is created. This dataframe includes the PIC variance or phylogenetic independent contrast variance is calculated. This extra analysis will give us some idea what the rate of evolution per character is. To sum up following data will be collected:

- K: The K value for a specific phonemic sound.
- PIC.variance: The rate of evolution for a certain phonemic character.
- significant: Testing whether a result is significant.

K statistic test method

For each biphone, only calculate the K value if there are at least 20 non "nan" values and these values aren't all the same. For each biphone the next two values are calculated:

- K value for the specific biphone
- ρ value based on the randomization precedure for $H_0(k = 0)$.

As explained in 1.3.2 a random procedure is needed, for this procedure 10000 random permutations per phonemic character are made. This way the calculation for a specific K value converges to a certain mean.

3 Results

The results are devided between phones and biphones, for the K statistic an additional subset of the Austronesian languages is used this being the Western Lamaholot dialects. Other regions and families of languages where considered but none had enough data for the required minimum languages of 20(Or 50 in the D statistic).

3.1 D statistic

For the D statistic only the Austronesian languages will be researched since of the minimum of 50 languages.

3.1.1 Austronesian languages

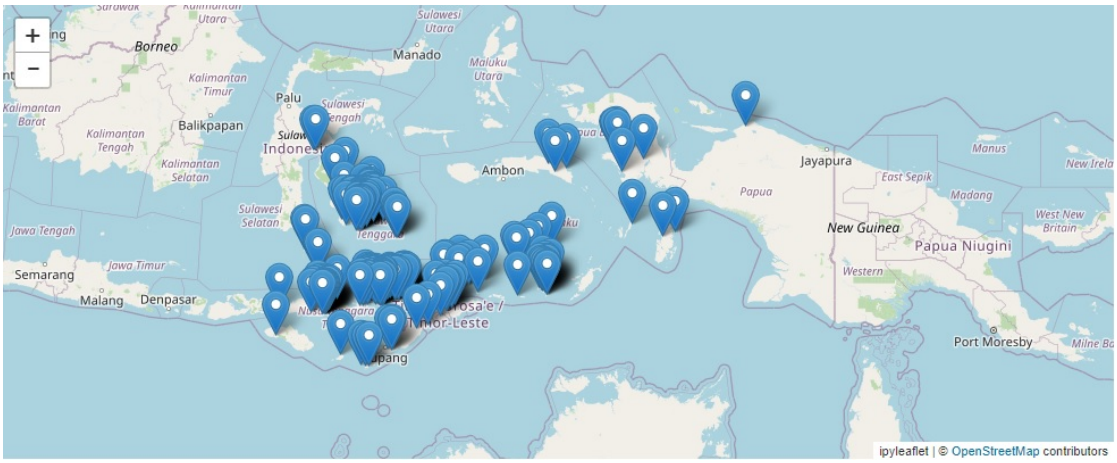
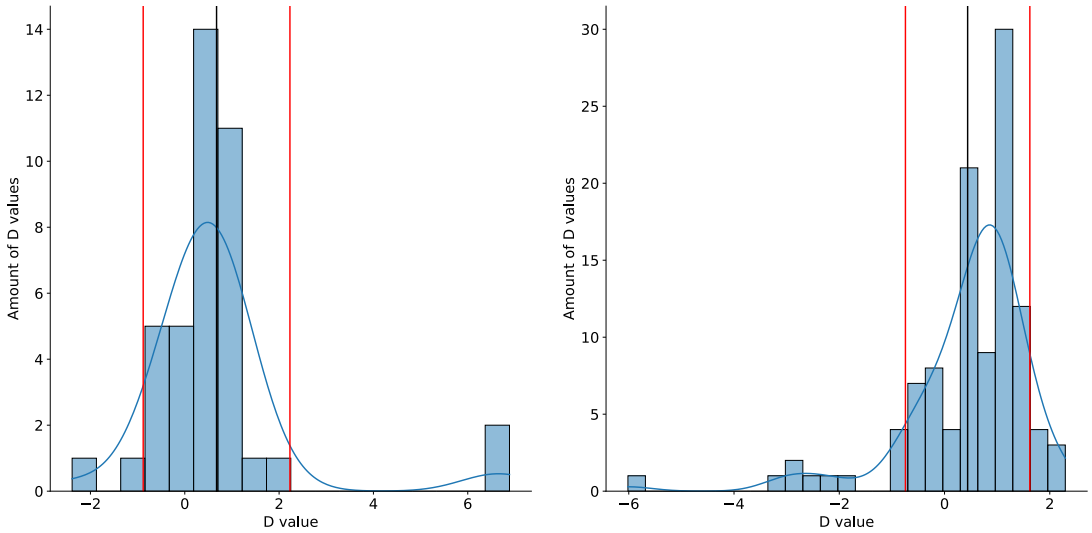


Figure 9: Geographical spread of the languages that Lexirumah and Glottolog have in common, i.e., the languages that will be analyzed.

The map seen on figure 9 is generated by displaying all geographical positions where languages are proposed to be existing, note that some languages have multiple markers due to the fact they are spoken in more than 1 location. Moreover phylogenetic signals can be expected from languages residing on the same island, although this is plausible but caution need to be taken with such statements.The take-away message here should be that the geographical spread of languages is solely an indication of languages copying each other they still can differ by a substantial amount, the more important aspect here is the amount of contact that languages have undergone during their history. Multiple elements take part in the way languages can start mimicking each other. Languages can change grammatically, for example studies from (Klamer, 2020) and (Moro, 2019) have shown that languages can change in their morphology[15] due to their contact with other languages. These changes in morphology may also induce changes in the phonetics of the language and by so show phylogenetic signals.

[15]The study of words on how they're formed and their relation to other words in a specific language. More specific it studies prefixes and suffixes of words.

Single segment characters (phones)



(a) The distribution of D values for phones calculated with ASJPcode as phonemic character base.

(b) The distribution of D values for phones calculated with IPA as phonemic character base.

Figure 10: A comparison between the D statistic distributions of 2 phonemic alphabets using phones as phonemic sounds.(Black line denotes the mean, red lines denote the left and right standard deviations)

Character encoding comparisons

At figure 10 two distributions are shown, as mentioned before ASJPcode is a many-to-one mapping of the IPA alphabet. This due to the fact that the IPA alphabet is extensive and has many characters. This can lead to discrepancies between different researchers who are investigating two languages that are nearly the same. Due to their own interpretation of some phonemic sounds different characters from the IPA alphabet could be used for the same sound. In data this discrepancy can be observed as outliers which make it more difficult to analyze the phonemic sounds. Although 10a 10b both show outliers, during the continuation of this Bachelor's thesis ASJPcode is used. This decision is taken due to the fact of data limitations, and due to the probability of discrepancies in previous research. To compare both phonemic alphabets, for the first D statistic plot both calculations will be shown.

ASJPcode

- Amount of phones: 41
- Mean: 0.67
- Standard deviation: 1.55
- Median: 0.42
- Skewness: 2.72
- Kurtosis: 10.07 (leptokurtic)

IPA

- Amount of phones: 109
- Mean: 0.43
- Standard deviation: 1.18
- Median: 0.69
- Skewness: -2.38
- Kurtosis: 8.71 (leptokurtic)

Both phonemic encodings show a leptokurtic distribution, this is due to the fact the distribution contains more outliers relative to a normal distribution. This is visible in figure 10, both distributions show tall narrow peaks and long tails. The distributions however differ in respect to the skewness, the ASJPcode has a positive skew meaning that phonemes have the tendency to show an over-dispersed distribution at the tips of the tree. This can be due to the lack of data (outliers responsible for the big skew) or a bad mapping of characters. Since the data only exists of 41 characters it is difficult to analyze why this distribution has this positive skewness. IPA code on the other hand has as negative skew, meaning many values show a clumped nature. Again this can be due to to the lack of data and thus outliers having a big impact on the skew, another explanation is that certain phonemic sounds where developed in a very early stage of the language tree. One subset of all closely related Austronesian languages may contain this sound while another subset of languages which diverged from a very early stage in the tree may not contain this phonemic sound. Visually this can be interpreted as follows for the presence of a certain phone:

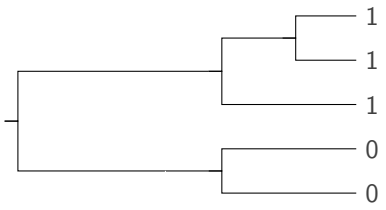


Figure 11: Visual representation of a phylogenetic language tree where a certain phonemic sound is only used in one root branch of the tree. Meaning a certain sound was developed at an early stage and diverged after.

This hypothesis however can't be confirmed due to the lack of data, as mentioned in 2.1.1 the phylogenetic tree consists of 1270 tips. The "phylo.d" function however was only calculated on 133 tips of this phylogenetic tree. Diverging from exploring both phonemic alphabets, the results will now only be focused on the ASJPcode. Exploring the results leads to the quantitative summarization below[16].

- **Amount of phylogenetic results: 17 (41 % of the phonemic characters)**
- Amount of Indeterminate (neither H0 rejected) results: 12 (29 % of the phonemic characters)
- Amount of random results: 1 (2 % of the phonemic characters)
- Amount of More dispersed results: 0 (0 % of the phonemic characters)
- Amount of More clumped results: 10 (24 % of the phonemic characters)
- Amount of $0 < D < 1$ (both H0s rejected) results: 1 (2 % of the phonemic characters)

Note that these results are expected to have 5 % of false positive discoveries, taking this into account approximately 2 of the 40 rejected null hypothesis show falsive information. It's clear that when using the D statistic method, phylogenetic signals are present. Furthermore the results suggest that using a many-to-one mapping from IPA to ASJPcode induces more statistically significant results.

Figure 12: Heatmap showing the kind of historical signal per phonemic character of the ASJPcode, grid colour-codes are used for each single segment character to show their result of its respective significance test.

To analyze figure 12 to greater extend, the ASJPcode Table needs to be consulted. This can be found on `https://web.archive.org/web/20220507112302/https://en.wikipedia.org/wiki/Automated_Similarity_Judgment_Program`. These rusults are in sharp contrast with previous studies, since these where mostly done in IPA encoding (Blust, 2009). Given a sufficient amount of languages, in our case more than 50. The D values will present a phylogenetic signal independent from the tree size (the amount of languages contained in the tree) and the branching pattern of the tree. To check the robustness of the results, a plot is made where the y-axis defines the amount of languages where a certain phone is present. The x-axis will show the corresponding D value for the character.

[16]These calculations where also performed on the IPA encoding, 46% of the results where Indeterminate so no continuation with this encoding was considered.

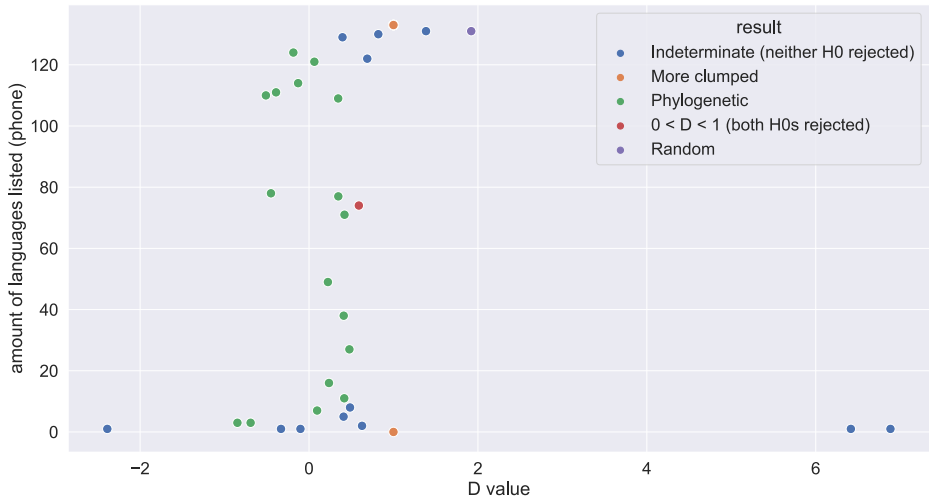


Figure 13: Phones plotted against the number of languages where they appear in. Colors represent the kind of statistical result retrieved from the D statistic for a certain phone.

Figure 13 displays that phones that aren't phylogenetic either are present in 2 extremes. The first extreme is a phonemic sound that is present in most languages, this is logical when looking to figure 2. Phylogenetic signals are the tendencies of closely related species mimicking each other, when all languages share the same value no phylogenetic signal is present. The expectation is "Indeterminate" or "more clumped" in this case. However when there is a lack of data, i.e. the language phylogeny contains more languages than the historical phonemical data it is unsure to classify a phonemical sound "more clumped" when found to be in all entries of the phonemical data. The other extreme when a phoneme is only found in a few languages is prone to be Indeterminate. Indeed if this is the case it is mostly regarded as an outlier, a phoneme which may be very unique or wrongly documented.

Double segment characters (biphones)
As of the single segment characters, double segment characters undergo the same statistical test. Again 10000 permutations are used in combination with a ρ value of 0.05.

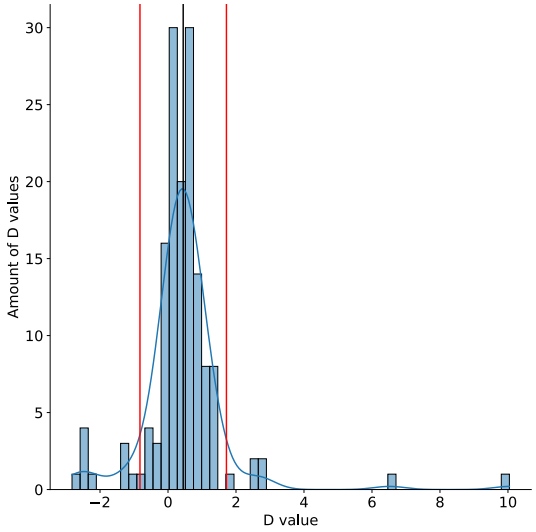


Figure 14: The distribution of D values for biphones calculated with ASJPcode as phonemic character base. Black line denotes the mean, red lines denote the left and right standard deviations

- Amount of biphones: 151
- Mean: 0.45
- Standard deviation: 1.27
- Median: 0.41
- Skewness: 3.21
- Kurtosis: 24.66 (leptokurtic)

Again a leptokurtic distribution is present, meaning there is a narrow peak with a long tail. Skewness has a positive value, although this looks to be the case of some outliers. These outliers show some over dispersed behaviour. With this in mind further research leads to following results:

- **Amount of phylogenetic results: 68 (45 % of the phonemic characters)**
- Amount of Indeterminate (neither H0 rejected) results: 78 (52 % of the phonemic characters)
- Amount of random results: 4 (3 % of the phonemic characters)
- Amount of More dispersed results: 0 (0 % of the phonemic characters)
- Amount of More clumped results: 1 (1 % of the phonemic characters)

- Amount of $0 < D < 1$ (both H0s rejected) results: 0 (0 % of the phonemic characters)

Noteworthy here is that biphones tend to have significantly more "indeterminate" results than phones tend to have in the D statistic. 52% of the characters have an indeterminate result, suggesting that for biphones the D statistic is a low information yielding method for discovering phylogenetic signals. Although since of the high percentage for phylogenetic results a phylogenetic signal seems to be present.

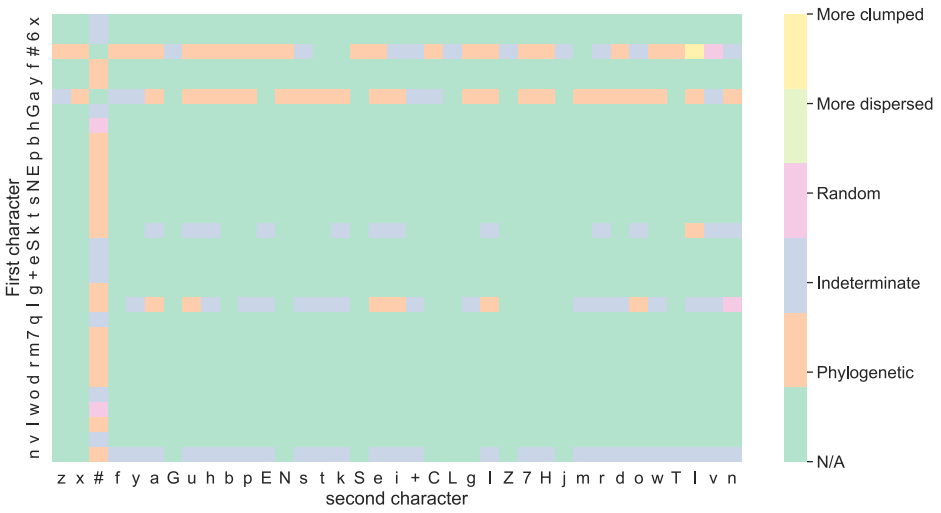


Figure 15: Heatmap showing the kind of historical signal per phonemic character of the ASJP-code, the y-axis represents the first phonemic character of the specific biphone the second character is displayed on the x-axis. Grid colour-codes are used for each single segment character to show their result of its respective significance test.

Note that in figure 15 many combinations are displayed as darker green, these biphones are non existent in any of the languages. This is due to the sparsity of combinations possible between phones. Another factor is also responsible for the sparsity in this matrix, these are the symbols "#". This represents on the y-axis the start of a word, while the character on the x-axis will represent the first phone appearing in the word. When using this the "#" character on the x-axis this denotes that this is the ending of a specific word, the character on the y-axis displays in this case the last phoneme heard in the word. Multiple patterns stand out in this heatmap, biphones starting with the character "a" a "low central vowel, unrounded" show phylogenetic signals in combination with other phones. Another character showing phylogenetic signals is the character "l" a "voiced alveolar lateral approximate" followed with some phones. The last phylogenetic signal found is the biphone "kl" with k being a "voiceless velar stop", although found to show a phylogenetic signal no cognate sets where found that confirm it's existence.Interesting here is the description of the ASJPcode this opens bridges to research this phylogenetic signals with higher accuracy. For example researching the IPA characters residing under the specific "k" , "a" and "l" characters. 2 other signals stand out, these are all sounds at the beginning or ending of a word, this suggests that closely related languages mimick each other in terms of word prefixes

and word suffixes.
Words have an underlying morphology which make some characters appear more frequently, examples of this are the prefix “an” and suffix “al” in Austronesian languages (Blust, 2003).

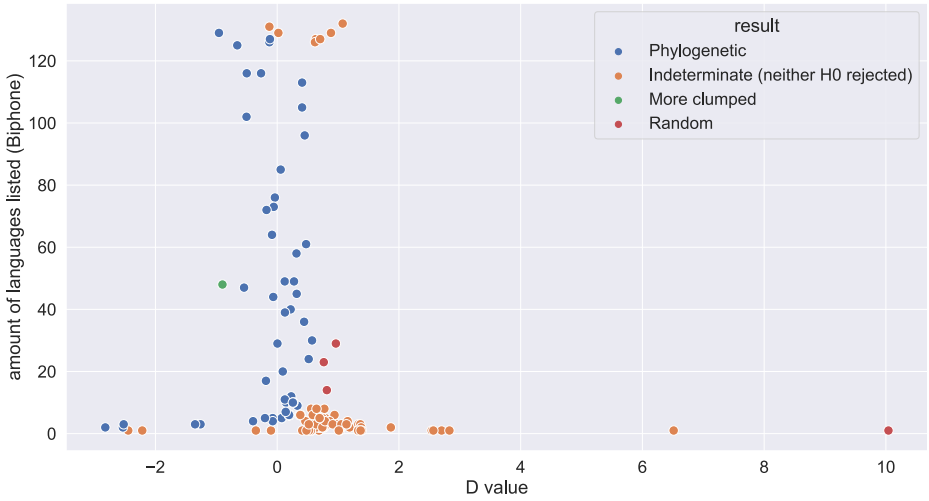


Figure 16: Bihones plotted against the number of languages where they appear in. Colors represent the kind of statistical result retrieved from the D statistic for a certain phone.

Again a robustness check is done to determine how trustworthy the results are, the results for biphones show the same structure but with more data. Characters that only appear in a few languages show indeterminate results as well as characters that appear in almost all languages. Although some phylogenetic signals are found with the D statistic, some finer data extraction will be done to find more characters with phylogenetic signals.

3.2 K statistic

Since the requirements of the K statistic are more easily attainable, a subset of Austronesian languages is researched as well. This subset is the Western Lamaholot dialect group, it might be revealing if from this subset different phylogenetic signals are detected

3.2.1 Austronesian languages

The first graph plotted is the density curve for all the K values, noteworthy before observing the graph is that the optimum K value for a phylogenetic signal should be equal to 1(Blomberg et al., 2003). This research is comparable to the research of biological traits, in (Galván, Rodríguez-Martínez, & Carrascal, 2018) for example the traits represent the pigmentation of birds on several spots of their body. Instead of using biological traits, character frequencies are used.

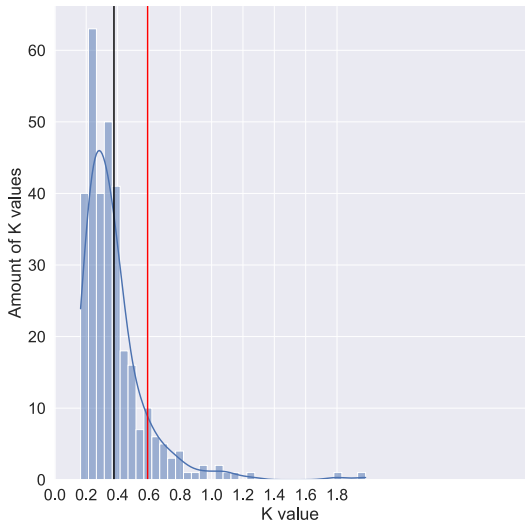


Figure 17: The distribution of K values for biphones of the Austronesian dataset calculated with ASJPcode as phonemic character base. Black line denotes the mean, red line denotes the right standard deviation

When calculating the K statistic with 10000 permutations, 314 biphones and a ρ value of 0.05 we find following results:

- **Mean: 0.377 (Weak phylogenetic signal among biphones, desired value for a phylogenetic signal = 1)**
- Standard deviation: 0.215
- Kurtosis: 16.78 (leptokurtic)
- Amount of statistical significant values: 118 (38% of the total amount of biphones)

Noticeable is that most characters show a weak phylogenetic signal (mean = 0.377) furthermore figure 17 displays a very high narrow peak confirmed by the high value for the kurtosis. What makes the result difficult to interpret is that most characters have an insignificant K value, to challenge this a second heatmap is made with only significant characters.

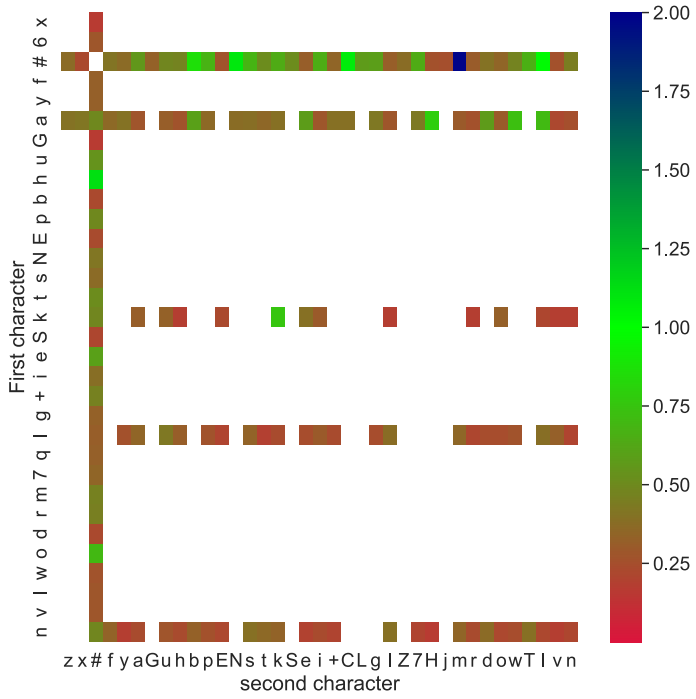


Figure 18: Heatmap containing all values gained for the forward transition probabilities extracted from the Austronesian languages, bright green values denote a strong phylogenetic signal. Red and blue define both weak phylogenetic signals. The y-axis denotes the first character of the biphone while the x-axis displays the second character.

The same suggestion can be made here that the prefix and suffix of closely related languages tend to mimick each other, this suggestion is made because of the fully colored row and column for the character #[17]. An extra observation can be made in this regards, the phylogenetic signal seems to be stronger for suffixes of languages since the horizontal row of # appears to have brighter green. Other characters besides the phonemic character “a” show weak phylogenetic signals.

[17]Which denotes the end and start of a word.

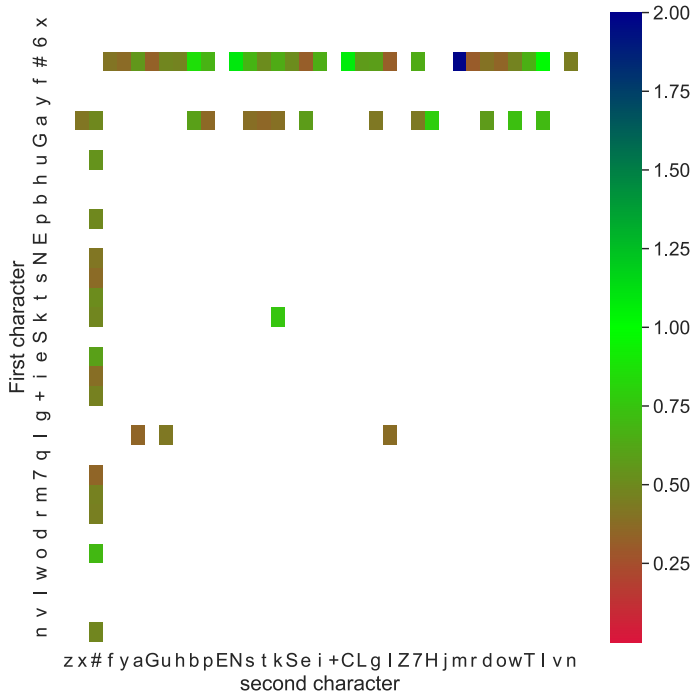


Figure 19: Heatmap containing all significant values gained for the forward transition probabilities extracted from the Austronesian languages, bright green values denote a strong phylogenic signal. Red and blue define both weak phylogenetic signals. The y-axis denotes the first character of the biphone while the x-axis displays the second character.

When elimanting all insignificant characters, a simular result as in 15 is conceived. With the K statistic it is however possible to see that phylogenetic signals tend to be stronger for phones that are the absolute suffix of a phonemic word. The suggestion here is that languages tend to have a form of conjugation, this conjugation happens for example in Sulawesi (Berg, 1996) at the end of words. Next to the similarities to the D statistic, the biphone “kl” isn’t significant here. A biphone that is significant in the K statistic but not in the D statistic is the biphone “kk”. The biphone “k” is frequently used in the Toba Batak language (Nababan et al., 1981). Figure 20 demonstrates the geographical location of the Toba Batak language, note that it is located in northern Sumatra.

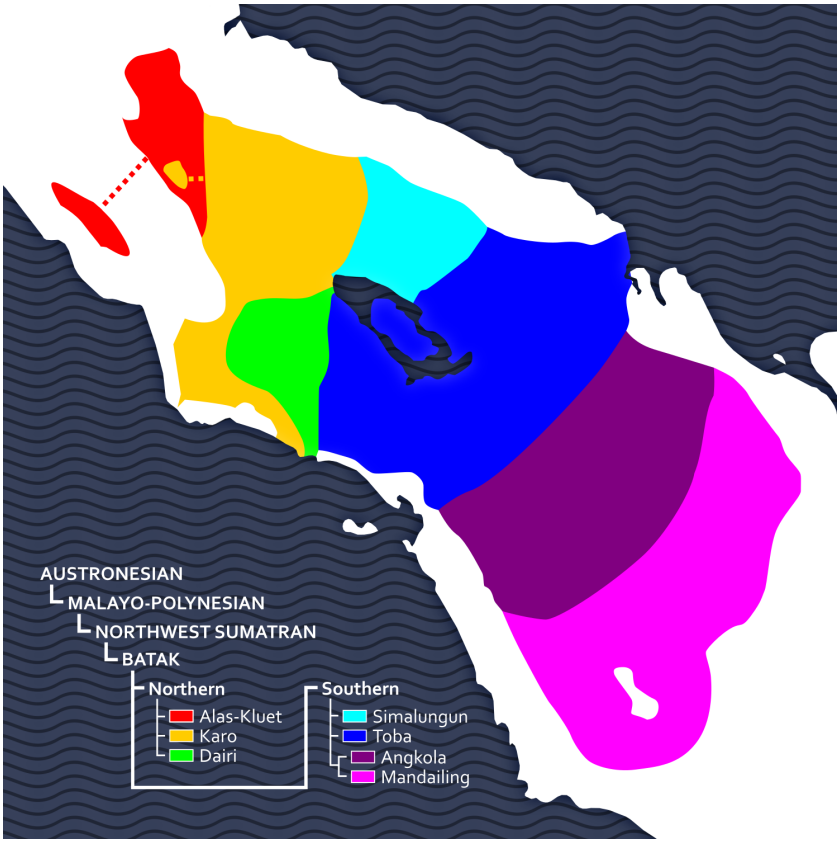


Figure 20: Figure showing the geographical location on where Toba Batak languages are spoken.
Title: "The distribution of Batak languages in northern Sumatra. Toba Batak is the majority language in the blue-colored areas labeled with its ISO 639-3 code "bbc"."
author: Masjawad99
Source:`https://web.archive.org/web/20220529023816/https://commons.wikimedia.org/wiki/File:Batak_languages.png`
Licence:CC-BY-SA-4.0

To check whether the results are robust, a check is done by plotting the phonemic characters by their K score against the amount of languages they are represent in. In this scatterplot the results are marked by there significance. Hereby no distinction is made for the forward and backward transition probabilities, both are plotted to maximize the amount information to be extracted.

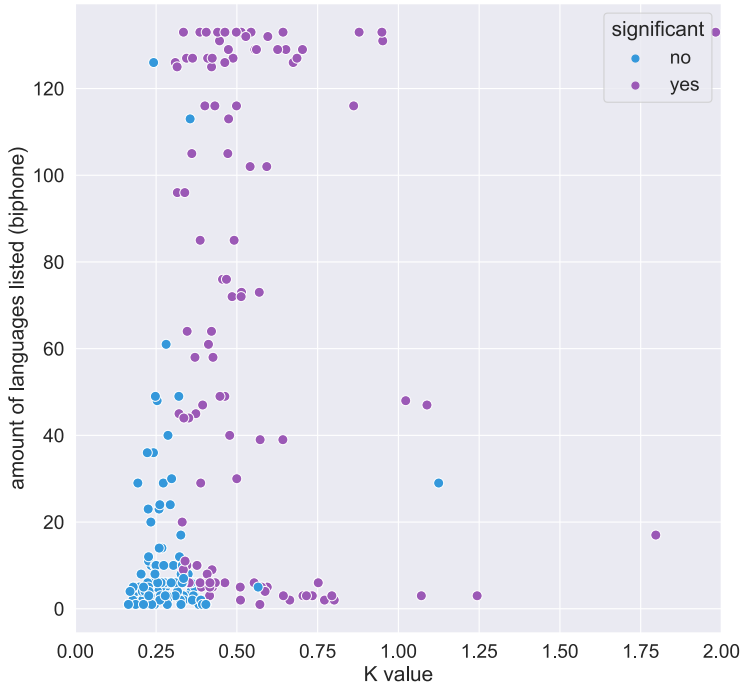


Figure 21: K value estimated values for forward and backward transition frequencies, plotted against the amount of Austronesian languages they appear in.

Figure 21 shows a narrow clustering around the mean value of 0.377. The variance however shows a very imposing characteristic, this can be summarized as follows:

- Unique or less documented biphones (0 to 40 languages): Show a very high variability.
- More frequently found biphones in languages (40 to 120 languages): Show a very low variability.
- Biphones appearing in every language of the historical data retrieved from Lexirumah, show a high variability.

Furthermore as expected the significance of results increases in frequency by the amount of languages where the biphone is present in.

At last we check whether there is a difference between the Forward and backward transition probabilities.

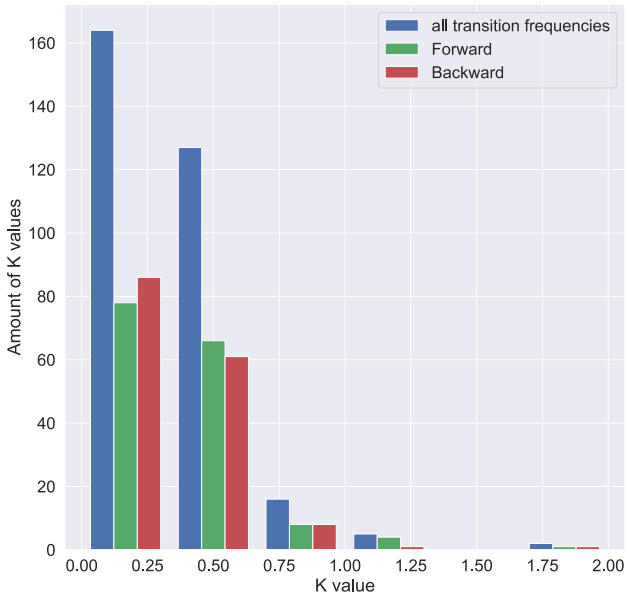


Figure 22: A histogram showing the difference between the forward and backward transition frequencies using the data of Austronesian languages.

Looking to figure 22 suggests that there is no difference between forward and backward transition frequencies of biphones. However there means differ by a substantial amount, a statistical analyses can be found in summation below:

- Mean K values for forward transition frequencies: 0.39
- Mean K values for backward transition frequencies: 0.36
- t-test result from independent samples: 1.01
- sig: 0.31

Although a difference was calculated in terms of the means of the two transition frequencies, according to the t-test they can be treated as equal distributions when testing with a significance ρ of 0.05.

3.2.2 Western Lamaholot dialects

The western Lamaholot dialect ranges over a few islands, the documented dialects have a geographical spread as seen below.

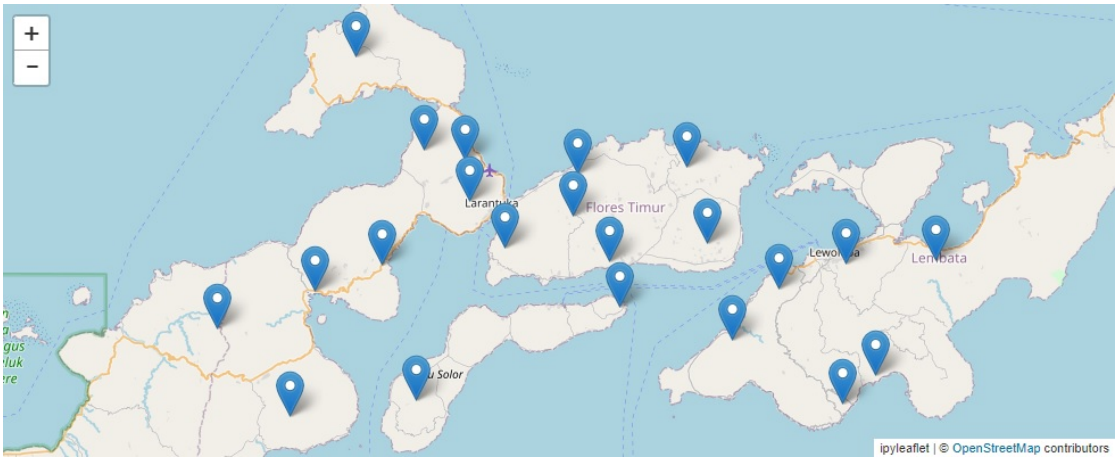


Figure 23: Geographical spread of the western Lamaholot dialects that Lexirumah and Glottolog have in common, i.e., the languages that will be analyzed. The languages are spreaded over the eastern tip of Flores and the islands Solor, Adonara and Lembeta

Since the geographical spread is more contained now, it is interesting to test how this affects the phylogenetic signals that are detected. The first plot to research this difference is the distribution of K values.

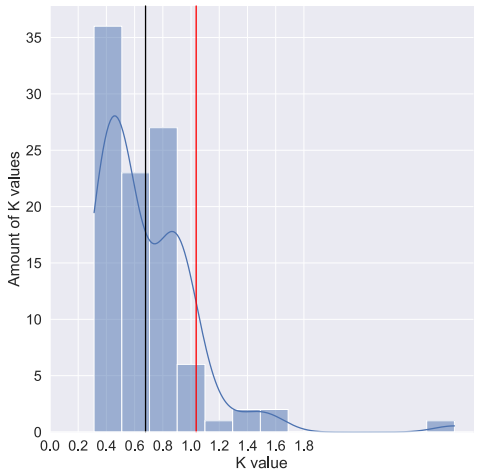


Figure 24: The distribution of K values for biphones of the western Lamaholot subset calculated with ASJPcode as phonemic character base. Black line denotes the mean, red line denotes the right standard deviation

Again the K statistic is calculated with 10000 permutations and a ρ value of 0.05. The amount of biphones is lower this time, capping at 98 different biphones. The summation of the results can be viewed below.

- **Mean: 0.678 (Strong phylogenetic signal among biphones, desired value for a phylogenetic signal = 1)**
- Standard deviation: 0.359
- Kurtosis: 13.47 (leptokurtic)
- Amount of statistical significant values: 37 (38% of the total amount of biphones)

The results are comparable with the findings made in 3.2.1, except this time the K value mean is higher(0.678:western Lamaholot, 0.377: Austronesian) suggesting there is a stronger phylogenetic signal among biphones. Due to the low amount of significant values again 2 heatmaps are plotted to research the phylogenetic signals with greater detail.

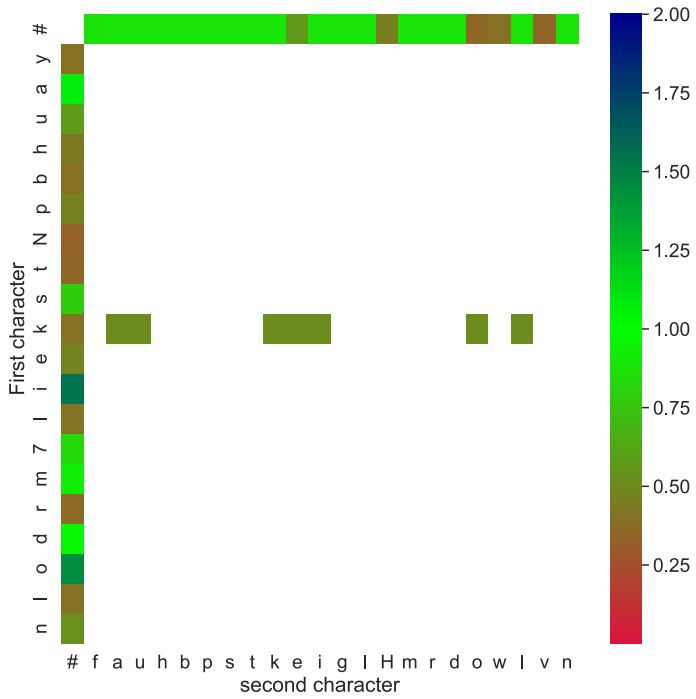


Figure 25: Heatmap containing all values gained for the forward transition probabilities extracted from the western Lamaholot dialects, bright green values denote a strong phylogenetic signal. Red and blue define both weak phylogenetic signals. The y-axis denotes the first character of the biphone while the x-axis displays the second character.

The reoccurring tendency of phylogenetic signals being present in the phonemic prefix and suffix of words continues here, again the phonemic suffix displays a stronger phylogenetic signal. To go more in detail the heatmap with only significant results will be consulted.

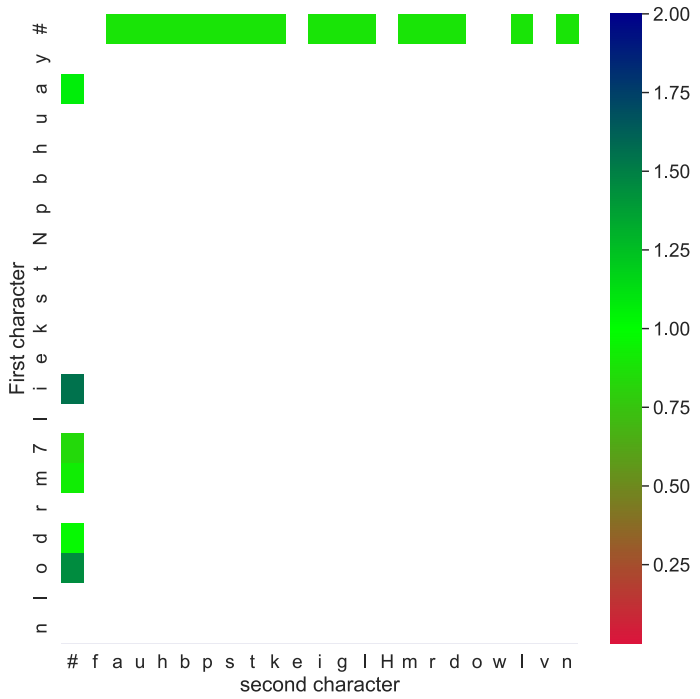


Figure 26: Heatmap containing all significant values gained for the forward transition probabilities extracted from the western Lamaholot dialects, bright green values denote a strong phylogenetic signal. Red and blue define both weak phylogenetic signals. The y-axis denotes the first character of the biphone while the x-axis displays the second character.

No special biphones consisting of 2 phones are discovered here, although the results may suggest that yet again the suffixes of words contain very strong phylogenetic signals. Since of the little amount of data, 22 languages against the opposing minimum of 20 languages a robustness check is done as calculated in the research of all Austronesian languages available in the data.

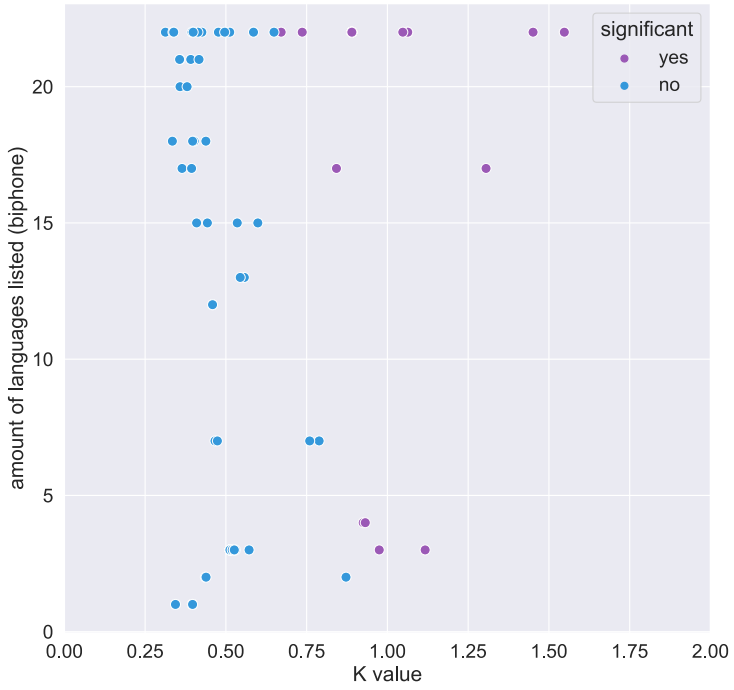


Figure 27: K value estimated values for forward and backward transition frequencies, plotted against the amount of western Lamaholot dialects they appear in.

Figure 27 shows the effects of a lack of data, many results are insignificant even when listed in all languages. This makes the results found by the the first heatmap containing all values unreliable, although this is true this can be the result of a lack of data. For example combinations with biphones starting with the phonemic character “k” can be consulted in (Blust, 2009), if this combination exists for a vast subset of western Lamaholot dialects a phylogenetic signal can be deduced.

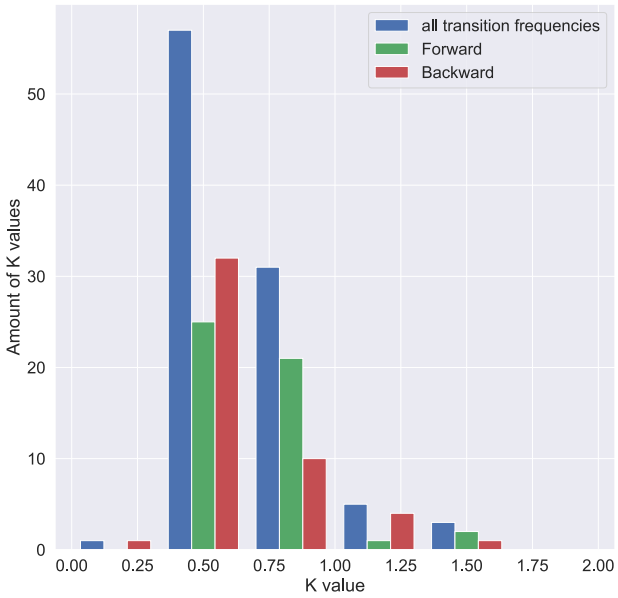


Figure 28: A histogram showing the difference between the forward and backward transition frequencies using the data of western Lamaholot dialects.

When looking at 28 suggestion rises that the forward and backward transition frequencies for biphones aren't equal. While this can be due to the lack of data, a student t-test is conducted to test this significance.

- Mean K values for forward transition frequencies: 0.68
- Mean K values for backward transition frequencies: 0.69
- t-test result from independent samples: 0.42
- sig: 0.67

Following these results we accept both distributions being the same according to the t-test with a ρ value of 0.05.

4 Conclusion

The research method was based on a procedure defined in (Macklin-Cordes et al., 2021). The method was in this research conducted on another language family, with this being the phylogeny of Austronesian languages. Furthermore the method was extended by looking how the phylogenetic signal tests such as the D and K statistic performed on subsets of the phylogeny such as in this research the western Lamaholot dialects. It was found that for single segment phonetic characters (phones) their existence in different languages show phylogenetic signals

with the D statistic. Biphones in regard to the D statistic gave a strong indication of languages having phylogenetic signals in the prefixes and suffixes of languages, next to this combinations with the character “a” and “l” as first phoneme in a biphone raised a substantial amount of phylogenetic signals. The K statistic confirmed the previous research for biphones with the D statistic and suggested that suffixes of languages contain stronger phylogenetic signals than the prefixes. Evaluating the K statistic on a smaller dataset, this being the western Lamaholot dialects introduced stronger phylogenetic signals for biphones. As last it was found that using the ASJPcode clustering gave rise to less outliers and gave little implication on reverting back to the IPA alphabet which is widely used in Austronesian languages research (Blust, 2009).

5 Discussion and future work

In general the research that has been done gives a clear image of phylogenetic signals being present in the reference phylogeny in combination with the historical data, it does not show which languages lie on the basis of these phylogenetic signals. An addition in this regard would be researching the phones and biphones that show a phylogenetic signal.

Name	a_e	c_e	d_g	e_a
alor1247	1	0	1	0
agor1237	1	1	1	0
baip1234	0	1	1	0
baip1234	1	0	1	1
bama1253	0	1	1	1
bara1372	1	1	1	1

Table 1: Green: Biphone with phylogenetic signal present, Red: Biphone without phylogenetic signal

It would be possible to filter the original data by phylogenetic signal, and use a PCA or TSNE plot to cluster the languages by their subfamily as defined in the Glottolog database. A more rigorous approach in this regard is possible to, since most phylogenetic signals contain a prefix or suffix phone it is possible to compare this with accepted rules about this languages, for example verbs[18] and their phonological rules. Another subject of future work would be going further on the findings of smaller datasets. It was found for example that western Lamaholot dialects showed stronger phylogenetic signals than the superset of Austronesian languages. It could be researched that for other large enough subsets(> 20) of closely related languages (in respect to the reference phylogeny) this tendency continues or that in fact the western Lamaholot dialects have strong phylogenetic signals due to the suggestive idea of their geographical spread over 4 islands.

[18] Assuming the grammatical conjugation of a language happens at the end of words.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/-experimental study. *Cognition*, *90*(2), 119–161.

Berg, R. v. d. (1996). The demise of focus and the spread of conjugated verbs in Sulawesi. *Pacific Linguistics. Series A. Occasional Papers*, *84*, 89–114.

Bezoui, M. (2019, 03). Speech recognition of Moroccan dialect using hidden markov models. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *8*, 7. doi: 10.11591/ijai.v8.i1.pp7-13

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. In *Evolution; international journal of organic evolution.*

Blomberg, S. P., Rathnayake, S. I., & Moreau, C. M. (2020). Beyond Brownian motion and the ornstein-uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *The American Naturalist*, *195*(2), 145-165. Retrieved from `https://doi.org/10.1086/706339` (PMID: 32017624) doi: 10.1086/706339

Blust, R. (1995). The prehistory of the Austronesian-speaking peoples: A view from language. *Journal of World Prehistory*, *9*(4), 453–510. Retrieved 2022-06-02, from `http://www.jstor.org/stable/25801085`

Blust, R. (1999). Subgrouping, circularity and extinction: some issues in austronesian comparative linguistics. In E. Zeitoun & P. J.-K. Li (Eds.), *Selected papers from eighth international conference on austronesian linguistics* (p. 31-94). Taipei: Academica Sinica.

Blust, R. (2003). Three notes on early Austronesian morphology. *Oceanic Linguistics*, *42*(2), 438–478. Retrieved 2022-05-28, from `http://www.jstor.org/stable/3623246`

Blust, R. (2009). *The Austronesian languages*. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University. Retrieved from `https://books.google.be/books?id=uYspAQAAIAAJ`

Blust, R. (2010). 36 five patterns of semantic change in austronesian languages. *A journey through Austronesian and Papuan linguistic and cultural space*, 525.

Blust, R. (2015, 12). The case-markers of Proto-Austronesian. *Oceanic Linguistics*, *54*, 436-491. doi: 10.1353/ol.2015.0024

Brown, C., Holman, E., & Wichmann, S. (2013a, 03). Sound correspondences in the world's languages. *Language*, *89*, 4-29. doi: 10.2307/23357720

Brown, C., Holman, E., & Wichmann, S. (2013b, 01). Sound correspondences in the world's languages: Online supplementary materials. *Language*, *89*, s1-s76. doi: 10.1353/lan.2013.0012

Brown, C., Holman, E., Wichmann, S., & Velupillai, V. (2008, 11). Automated classification of the world's languages: A description of the method and preliminary results. *STUF - Language Typology and Universals, v.61, 285-308 (2008)*, *61*. doi: 10.1524/stuf.2008.0026

Campbell, L. (2013). *Historical linguistics: An introduction* (NED - New edition, 3 ed.). Edinburgh University Press. Retrieved 2022-05-02, from `http://www.jstor.org/stable/10.3366/j.ctt1g0b5gq`

Chazot, N., Willmott, K. R., Santacruz Endara, P. G., Toporov, A., Hill, R. I., Jiggins, C. D., & Elias, M. (2014). Mutualistic mimicry and filtering by altitude shape the structure of andean butterfly communities. *The American Naturalist*, *183*(1), 26-39. Retrieved from `https://doi.org/10.1086/674100` (PMID: 24334733) doi: 10.1086/674100

Coleman, J., & Pierrehumbert, J. (2003, 06). Stochastic phonological grammars and acceptability.

Cubo, J., Ponton, F., Laurin, M., De Margerie, E., & Castanet, J. (2005, 08). Phylogenetic Signal in Bone Microstructure of Sauropsids. *Systematic Biology*, *54*(4), 562-574. Retrieved from https://doi.org/10.1080/10635150591003461 doi: 10.1080/10635150591003461

Dobrisek, S., Mihelic, F., & Pavesic, N. (1999, 01). Acoustical modelling of phone transitions: biphones and diphones - what are the differences?.

Durie, M., & Ross, M. (1996). *The comparative method reviewed: Regularity and irregularity in language change*. Oxford University Press.

Eddington, D. (2004). *Spanish phonology and morphology: Experimental and quantitative perspectives* (Vol. 53). John Benjamins Publishing.

Edwards, O., Kaiping, G. A., & Klamer, M. (2022, Feb). *lessersunda/lexirumah-data: v3.0.1.* Zenodo. (Funding: Reconstructing the past through languages of the present: the Lesser Sunda Islands Netherlands Organisation for Scientific Research (The Hague) GRANT NUMBER: 277-70-012 URL: https://www.nwo.nl/en/research-and-results/research-projects/i/91/22391.html) doi: 10.5281/zenodo.5951917

Elias, M., Gompert, Z., Jiggins, C., & Willmott, K. (2009, 01). Mutualistic interactions drive ecological niche convergence in a diverse butterfly community. *PLoS biology*, *6*, 2642-9. doi: 10.1371/journal.pbio.0060300

Ernestus, M. T. C., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language*, *79*(1), 5–38.

Felice, R., Randau, M., & Goswami, A. (2018, 09). A fly in a tube: Macroevolutionary expectations for integrated phenotypes. *Evolution*, *72*. doi: 10.1111/evo.13608

Felsenstein, J. (1985a). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1-15. Retrieved from https://doi.org/10.1086/284325 doi: 10.1086/284325

Felsenstein, J. (1985b). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. Retrieved 2022-05-09, from http://www.jstor.org/stable/2461605

Felsenstein, J., & Felenstein, J. (2004). *Inferring phylogenies* (Vol. 2). Sinauer associates Sunderland, MA.

Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *24*(4), 1042-1051. Retrieved from https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2010.01455.x doi: https://doi.org/10.1111/j.1523-1739.2010.01455.x

Galván, I., Rodríguez-Martínez, S., & Carrascal, L. M. (2018). Dark pigmentation limits thermal niche position in birds. *Functional Ecology*, *32*(6), 1531-1540. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.13094 doi: https://doi.org/10.1111/1365-2435.13094

Garland, J., Theodore, Midford, P. E., & Ives, A. R. (2015, 08). An Introduction to Phylogenetically Based Statistical Methods, with a New Method for Confidence Intervals on Ancestral Values1. *American Zoologist*, *39*(2), 374-388. Retrieved from https://doi.org/10.1093/icb/39.2.374 doi: 10.1093/icb/39.2.374

Gordon, M. K. (2016). *Phonological typology* (Vol. 1). Oxford University Press.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, *114*(42), E8822-E8829. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1700388114 doi: 10.1073/pnas.1700388114

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021, Dec). *glottolog/glottolog: Glottolog database 4.5.* Zenodo. doi: 10.5281/zenodo.5772642

Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of hungarian vowel harmony. *Phonology*, *23*(1), 59–104.

Holton, G., & Robinson, L. C. (2014). The linguistic position of the Timor-Alor-Pantar languages. In M. Klamer (Ed.), *Alor-pantar languages: History and typology* (p. 155-198). Berlin: Language Science Press.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Kang, Y. (2011). Loanword phonology. In *The blackwell companion to phonology* (p. 1-25). John Wiley Sons, Ltd. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444335262.wbctp0095` doi: https://doi.org/10.1002/9781444335262.wbctp0095

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., … Breinholt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences*, *116*(45), 22657-22663. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.1907847116` doi: 10.1073/pnas.1907847116

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., … Webb, C. O. (2010). *Picante: R tools for integrating phylogenies and ecology* (Vol. 26).

Klaczko, J., Ingram, T., & Losos, J. (2015). Genitals evolve faster than other traits in anolis lizards. *Journal of Zoology*, *295*(1), 44-48. Retrieved from `https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1111/jzo.12178` doi: https://doi.org/10.1111/jzo.12178

Klamer, M. (2019). The dispersal of Austronesian languages in island south east asia: Current findings and debates. *Language and Linguistics Compass*, *13*(4), e12325. Retrieved from `https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12325` (e12325 LNCO-0767.R1) doi: https://doi.org/10.1111/lnc3.12325

Klamer, M. (2020, 10). From Lamaholot to Alorese morphological loss in adult language contact. In (p. 339-367). doi: 10.1075/tsl.129.07kla

Macklin-Cordes, J. L., Bowern, C., & Round, E. R. (2021). Phylogenetic signal in phonotactics [Journal Article]. *Diachronica*, *38*(2), 210-258. Retrieved from `https://www.jbe-platform.com/content/journals/10.1075/dia.20004.mac` doi: https://doi.org/10.1075/dia.20004.mac

Moro, F. R. (2019). Loss of morphology in Alorese (Austronesian): Simplification in adult language contact. *Journal of Language Contact*, *12*(2), 378 - 403. Retrieved from `https://brill.com/view/journals/jlc/12/2/article-p378_378.xml` doi: https://doi.org/10.1163/19552629-01202005

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, *3*(4), 743-756. Retrieved from `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2012.00196.x` doi: https://doi.org/10.1111/j.2041-210X.2012.00196.x

Nababan, P., et al. (1981). *A grammar of Toba-Batak*. Dept. of Linguistics, Research School of Pacific Studies, The Australian ….

Orme, D. (2013, 01). The caper package: Comparative analysis of phylogenetics and evolution in r.

Petroni, F., & Serva, M. (2008, 08). Language distance and tree reconstruction. *Journal of Statistical Mechanics: theory and experiment*, *2008*, P08012. doi: 10.1088/1742-5468/2008/08/P08012

Puttick, M., Ingram, T., Clarke, M., & Thomas, G. (2019, 12). Motmot: Models of trait macroevolution on trees (an update). *Methods in Ecology and Evolution*, *11*. doi: 10.1111/2041-210X.13343

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Walker, A., & Zorc, R. (2011, 01). Austronesian loanwords in Yolngu-Matha of northeast arnhem land. *Aboriginal History Journal*, *5*. doi: 10.22459/AH.05.2011.07

Wichmann, E. W. H., Søren, & (eds.), C. H. B. (n.d.). *Automated similarity judgment program.* https://web.archive.org/web/20220331213605/https://asjp.clld.org/. (Accessed: 2022-4-1)

Wu, Q. (2020). eefolium: A python package for interactive mapping with google earth engine. *Journal of Open Source Software*, *5*(51), 2305.

Zuraw, K. R. (2000). *Patterned exceptions in phonology.* University of California, Los Angeles.

A ASJPcode Table

ASJPcode	Description	IPA
i	high front vowel, rounded and unrounded	i, ɪ, y, ʏ
e	mid front vowel, rounded and unrounded	e, ø
E	low front vowel, rounded and unrounded	a, æ, ɛ, œ, ɶ
3	high and mid central vowel, rounded and unrounded	ɨ, ɵ, ə, ɜ, ʉ, ɵ, ɞ
a	low central vowel, unrounded	ɐ
u	high back vowel, rounded and unrounded	ɯ, u
o	mid and low back vowel, rounded and unrounded	ɤ, ʌ, ɑ, o, ɔ, ɒ
p	voiceless bilabial stop and fricative	p, ɸ
b	voiced bilabial stop and fricative	b, β
m	bilabial nasal	m
f	voiceless labiodental fricative	f
v	voiced labiodental fricative	v
8	voiceless and voiced dental fricative	θ, ð
4	dental nasal	n̪
t	voiceless alveolar stop	t
d	voiced alveolar stop	d
s	voiceless alveolar fricative	s
z	voiced alveolar fricative	z
c	voiceless and voiced alveolar affricate	ts, dz
n	voiceless and voiced alveolar nasal	n
S	voiceless postalveolar fricative	ʃ
Z	voiced postalveolar fricative	ʒ
C	voiceless palato-alveolar affricate	tʃ
j	voiced palato-alveolar affricate	dʒ
T	voiceless and voiced palatal stop	c, ɟ
5	palatal nasal	ɲ
k	voiceless velar stop	k
g	voiced velar stop	ɡ
x	voiceless and voiced velar fricative	x, ɣ
N	velar nasal	ŋ
q	voiceless uvular stop	q
G	voiced uvular stop	ɢ
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative	χ, ʁ, ħ, ʕ
7	voiceless glottal stop	ʔ
h	voiceless and voiced glottal fricative	h, ɦ
l	voiced alveolar lateral approximate	l
L	all other laterals	ʟ, ɭ, ʎ
w	voiced bilabial-velar approximant	w
y	palatal approximant	j
r	voiced apico-alveolar trill and all varieties of “r-sounds”	r, ʀ, etc.
!	all varieties of “click-sounds”	!, ǀ, ǁ, ǂ

Table 1: Table containing all ASJPcode characters and their respective IPA translations (Brown et al.,2008).