Quantitatively exploring Australian Languages

Final Paper for the Linguistic Data Analysis using R Course

Benedict Wüthrich – 16-114-456

2024-08-23

Introduction

On the Australian continent there are over 333 reported languages which can be roughly categorized into either Pama-Nyungan languages or Non-Pama-Nyungan languages. It is important to note that the latter does not imply any genealogical link of the included languages, while the former has been shown as a cohesive language family. Non-Pama-Nyungan is more of a collective term for Australian languages that are not thought of as Pama-Nyungan and actually covers at least different language families. Exactly which languages belong to which family is an ongoing debate(Bowern & Koch, 2004).

Pama-Nyungen languages have been spoken for over 5000 years, make up over 306 different identified languages and the speakers cover about 80% of the landmass (Bouckaert 2018, 741). Given this long history, it is likely that the Pama-Nyungan languages have been competing against Non-Pama-Nyungan languages for a long time and ended up limiting the Non-Pama-Nyungan languages to the northernmost part of the continent. Nowadays, both have to compete against the English language.

One of the main actors in describing Australian languages is Robert M. W. Dixon. With "The Languages of Australia", he published a comprehensive description of Australian languages as early as 1980 and another study of Australian languages in 2004 in his monograph "Australian languages. Their Nature and Development". But these works are not without controversy. Dixon vehemently denied the applicability of the comparative method on Australian languages, which lead to discussion surrounding his works. Claire Bowern and Harold Koch wrote "Australian languages. Classification and the comparative method". Dixon's skepticism is described as an erroneous phylogenetic assessment which is "so bizarrely faulted, and such an insult to the eminently successful practitioners of Comparative Method Linguistics in Australia, that it positively demands a decisive riposte."(O'Grady and Hale in Bowern and Koch, 2004: 69), which serves as a perfect summary for what the book tries to achieve.

In more recent publications, such as the paper by Bouckaert et al. (2018), computer-assisted methods have gained ground. This paper specifically deals with the genealogy of Pama-Nyungan languages. Using their data consisting of basic vocabulary of 306 different Pama-Nyungan languages, the authors created a phylogenetic tree that shows the diversification of the Pama-Nyungan language family. Additionally, the same data was used to find a likely homeland of the family, which they postulated to be in the Gulf Plains region.

In this paper I explore the geographical distribution of the Australian languages and how they compare to each other in terms of typological features such as word order and phonemic inventory. My goal is to gain an understanding of how language contact might have affected these features. To make this both simpler to understand as well as simpler to execute, I will be adopting the reductionist point of view of pooling the Non-Pama-Nyungan languages into a single category. The goal is to gain an understanding of how these features are distributed in the two "families" and if any conclusion can be drawn regarding contact of them. My initial hypothesis is that there might be an apparent correlation between geographical proximity and similarity in the features analysed.

Data

The data used for this project originates from the cldf data by WALS. I initially intended to use the phoible cldf data as well, but ended up deciding against it as it did not add an appropriate amount of value to the final product.

The data wrangling for this project was manageable, but took a lot of work to get started. The biggest challenge posed was figuring out how to load cldf dataset into R and how to manipulate it appropriately. I struggled with that part for a while as it also had me refresh my memory on how to manipulate data frames efficiently and how to structure the creation of new data frames from the given sets without creating too much redundancy. I created a separate R script file that entails all the necessary data wrangling and manipulation for the different plots and maps used in this project. I did this, so that I can manipulate the code more easily and keep an overview of the changes, only commiting them to the Rmd file, when I felt the code achieved what I intended. The other scripts rely on this data wrangling script to be run first, so as to load the required data frames into the global environment. This Rmarkdown file compiles them all into one coherent document, though without saving the plots and maps again. The separate scripts can be found in the corresponding GitHub repository, a link to which can be found at the end of this document.

I started out by loading in the cldf using the rcldf package. From this a wide data frame was created using the value table and filtering for Australian languages.

```
##loading in data from phoible and WALS and limiting them to languages spoken in Australia
#Let's start with WALS

wals_cldf <- cldf(here("data/WALS/cldf"))
#this loads the cldf data, but is not yet manipulable. Do to this, we need to get it into a dataframe o
# summary(wals_cldf)
# this last line was commented out to streamline this pdf

#This creates a tibble that will be manipulated further at different points.
wals_value_aus <- as.cldf.wide(wals_cldf, "ValueTable") %>%
  filter(Macroarea == "Australia")
```

```
## Joining Parameter_ID -> ParameterTable -> ID
```

```
## Joining Code_ID -> CodeTable -> ID
```

```
## Joining Language_ID -> LanguageTable -> ID
```

```
## Joining Example_ID -> ExampleTable -> ID
```

Next up was creating separate data frames that specifically only included either Pama-Nyungan or Non-Pama-Nyungan languages and entries covering the word order feature. Additionaly, I renamed the relevant column to be more descriptive and therefore easier to work with, then mutated it into factors, so the different entries could be clearly ordered. In the case of the Non-Pama-Nyungan languages, another step was taken where I added in missing data into the relevant column, as it was entered as *NA*. Another column next to it had the missing data, so I manually added the entries back in.

```
##Word Order Data Wrangling
#creates two separate df, for PN and nPN respectively.

wals_worder_PN <- wals_value_aus %>%
  filter(Chapter_ID == "81" &
```

```
            Family == "Pama-Nyungan") %>% #filtering
  rename("WordOrder" = "Name.CodeTable") %>% #renaming for clarity
  mutate(WordOrder = factor(WordOrder,
                            levels = c("SOV",#reordering for neatness
                                       "SVO",
                                       "VOS",
                                       "OSV",
                                       "No dominant order")))

wals_worder_nPN <- wals_value_aus %>%
  filter( Chapter_ID == "81" &
            Family != "Pama-Nyungan") %>% #filtering
  rename("WordOrder" = "Name.CodeTable") %>% #renaming for clarity
  mutate(WordOrder = factor(WordOrder,
                            levels = c("SOV",#reordering for neatness
                                       "SVO",
                                       "SVO or VOS",
                                       "OSV",
                                       "No dominant order"))) %>%
  mutate(WordOrder = case_when(
    Language_ID.ValueTable == "ung" ~ "OVS",
    Language_ID.ValueTable == "myi" ~ "OVS",
    Language_ID.ValueTable == "grr" ~ "No dominant order",
    TRUE ~ WordOrder))#this bit was added as three languages didn't have entries
#in the new WordOrder Column. Thankfully the necessary information was in the
#Column next to it, so we add it back in
```

In a next step, I created several more data frames with the intention of looking at the phoneme inventory of the language families. The relevant features I picked to accomplish this task were the relative consonant inventory size, relative vowel quality inventory size and consonant to vowel ratio. Where helpful, I changed the relevant feature name to factors and ordered them in a reasonable way, so as to make the plots more easily readable in the end.

```
## Phoneme Inventory
###essentially the same process as above, but for a different Features

###PN
#separating by ConsonantInventories
wals_cons_PN <- wals_value_aus %>%
  filter( Chapter_ID == "1" &
            Family == "Pama-Nyungan") %>%
  rename("ConsonantInventories" = "Name.CodeTable") %>%
  mutate(ConsonantInventories = factor(ConsonantInventories,
                                       levels = c("Large",
                                                  "Moderately large",
                                                  "Average",
                                                  "Moderately small",
                                                  "Small")))

#separating by Vowel Quality Inventories
wals_vowels_PN <- wals_value_aus %>%
  filter( Chapter_ID == "2" &
            Family == "Pama-Nyungan") %>%
```

```
  rename("VowelQualityInventories" = "Name.CodeTable")


#separating for Consonant/Vowel Ratio and ordering the levels
wals_ratio_PN <- wals_value_aus %>%
  filter( Chapter_ID == "3" &
            Family == "Pama-Nyungan") %>%
  rename("ConsonantVowelRatio" = "Name.CodeTable") %>%
  mutate(ConsonantVowelRatio = factor(ConsonantVowelRatio,
                              levels = c("High",
                                         "Moderately high",
                                         "Average",
                                         "Moderately low")))

###nPN
#separating by ConsonantInventories
wals_cons_nPN <- wals_value_aus %>%
  filter( Chapter_ID == "1" &
            Family != "Pama-Nyungan") %>%
  rename("ConsonantInventories" = "Name.CodeTable") %>%
  mutate(ConsonantInventories = factor(ConsonantInventories,
                                       levels = c("Average",
                                                  "Moderately small",
                                                  "Small")))

#separating by Vowel Quality Inventories
wals_vowels_nPN <- wals_value_aus %>%
  filter( Chapter_ID == "2" &
            Family != "Pama-Nyungan") %>%
  rename("VowelQualityInventories" = "Name.CodeTable")


#separating for Consonant/Vowel Ratio and ordering the levels
wals_ratio_nPN <- wals_value_aus %>%
  filter( Chapter_ID == "3" &
            Family != "Pama-Nyungan") %>%
  rename("ConsonantVowelRatio" = "Name.CodeTable") %>%
  mutate(ConsonantVowelRatio = factor(ConsonantVowelRatio,
                                      levels = c("Moderately high",
                                                 "Average",
                                                 "Moderately low")))
```

Analysis

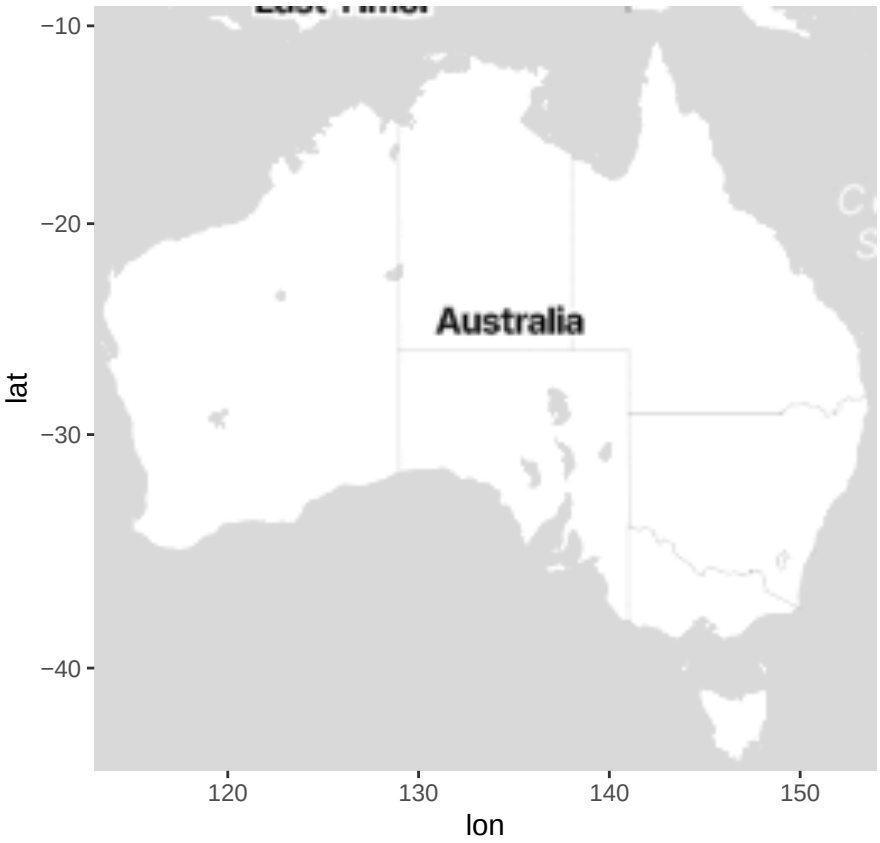
First off, I explore the distribution of Pama-Nyungan languages compared to Non-Pama-Nyungan languages. To do this, I first load a map of Australia through stadiamap. Next, I modify an existing data frame in order to make the legend of the final map more easily readable. Using this new data frame, I plot the locations of the languages onto the map.

```
map_AUS <- get_stadiamap(bbox = c(left = 113,
                                  bottom = -44,
                                  right = 154,
```

```
                                top = -9),
                    zoom = 3,
                    maptype = "stamen_toner_lite",
                    color = "color")
```

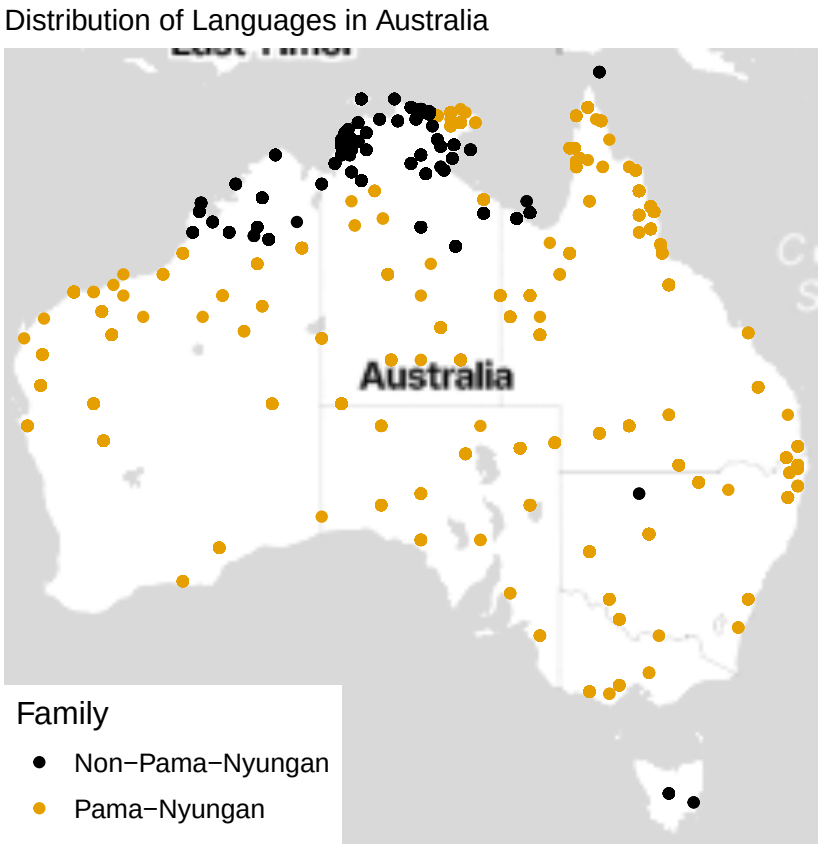
```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```
ggmap(map_AUS)
```




```
#This map will show the distribution of PN and nPN languages.
#While reductionist, I want to make it easier to read,
#so I mutate all nPN languages to simply "Non-Pama-Nyungan"


wals_nPN <- wals_value_aus %>%
  mutate(Family = if_else(Family == "Pama-Nyungan", "Pama-Nyungan", "Non-Pama-Nyungan"))

map_AUS_family <- ggmap(map_AUS) +
  geom_point(data = wals_nPN,
             aes(x = Longitude,
                 y = Latitude,
                 color = Family),
             show.legend = T) +
  scale_color_colorblind() + #better color
```

```
  theme_map() + #removes axes labels and puts the legend in bottom left corner of map
  theme(legend.text = element_text(size = 10),
        legend.title = element_text(size = 12)) +
  labs(color = "Family")+
  ggtitle("Distribution of Languages in Australia")
map_AUS_family
```



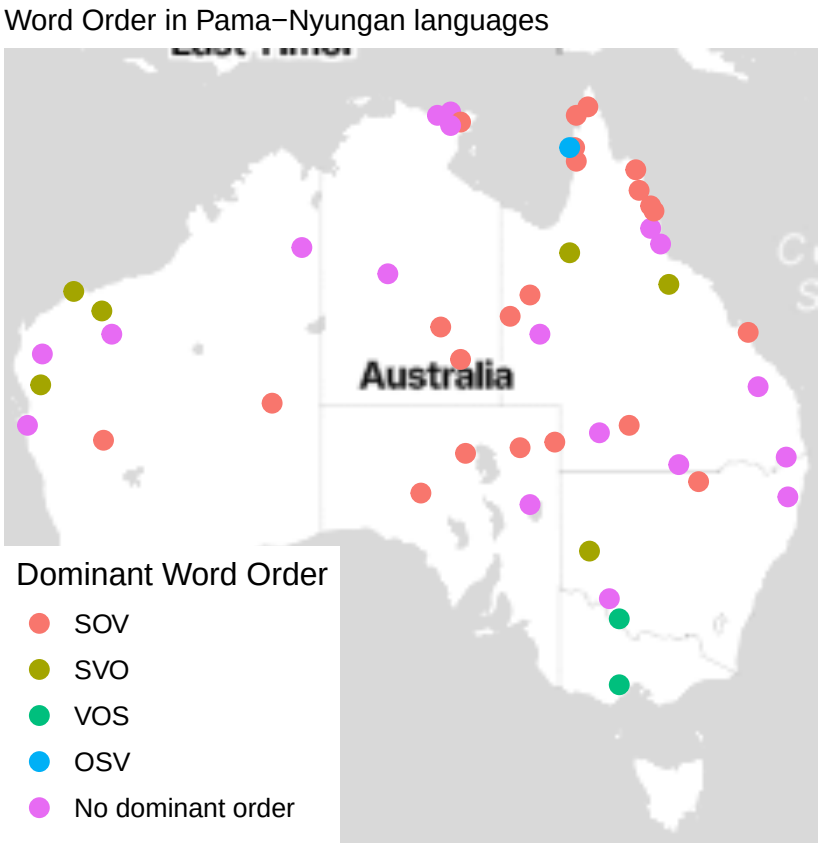
This map illustrates the distribution of Pama-Nyungan languages compared to Non-Pama-Nyungan languages. The latter are, with a few exceptions, visibly confined to the northernmost part of Australia, whereas the Pama-Nyungan languages are spread all throughout the continent.

Word Order

Matthew Dryer describes the word order of different languages as "perhaps the single most frequently cited typological feature of languages" (Dryer, 2013). I do not intend on making an exception in this matter either. To gain insight on the matter for the context of the Australian languages, I will be creating separate maps for both Pama-Nyungan and Non-Pama-Nyungan, as well as separate bar plots to explore the frequency of each within the respective categories.

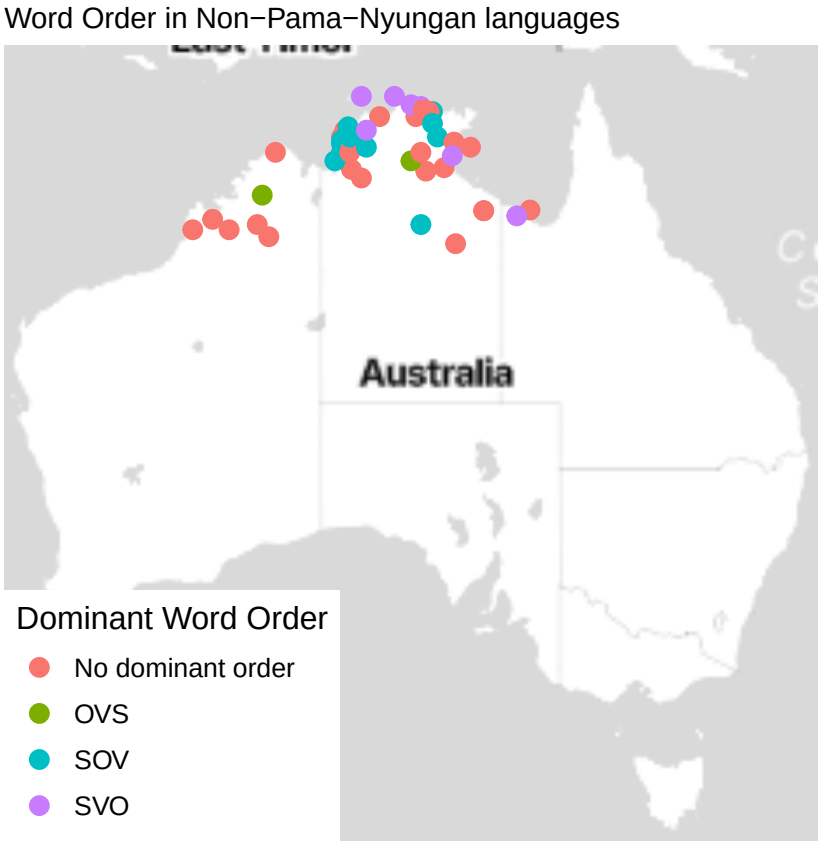
```
#a different map that shows the dominant word order of Australian languages
#first we need to filter for data that entails the Word order.
#This is chapter 81 of WALS.
#we further separate them by PN and nPN

##PN
```

```
map_PN_worder <- ggmap(map_AUS) +
  geom_point(data = wals_worder_PN,
             aes( x = Longitude,
                  y = Latitude,
                  color = WordOrder),
             show.legend = T,
             size = 3) +
  theme_map() +
  theme(legend.text = element_text(size = 10),
        legend.title = element_text(size = 12)) +
  labs(color = "Dominant Word Order") +
  ggtitle("Word Order in Pama-Nyungan languages")
map_PN_worder
```




```
##nPN
map_nPN_worder <- ggmap(map_AUS) +
  geom_point(data = wals_worder_nPN,
             aes( x = Longitude,
                  y = Latitude,
                  color = WordOrder),
             show.legend = T,
             size = 3) +
  theme_map() +
  theme(legend.text = element_text(size = 10),
        legend.title = element_text(size = 12)) +
```

```
  labs(color = "Dominant Word Order") +
  ggtitle("Word Order in Non-Pama-Nyungan languages")
map_nPN_worder
```




```
#Note: the size of the dots was reduced compared to the code in Maps.R,
#so as to make the individual dots more clearly distinguishable after the pdf is knit.
```

The maps paint an unclear picture. There is little insight to be gained from these maps alone. Maybe one could argue that there's a slight concentration of Pama-Nyungan languages that prefer SOV word order within the outback of Australia. Though the northernmost part of Queensland also shows such a trend and it is hard to say if either is truly meaningful. Looking at the map regarding the Non-Pama-Nyungan languages, not much more can be said. There is an apparent general trend to not having a clear dominant order though. Let's plot the frequencies of the various word orders for the two language categories.

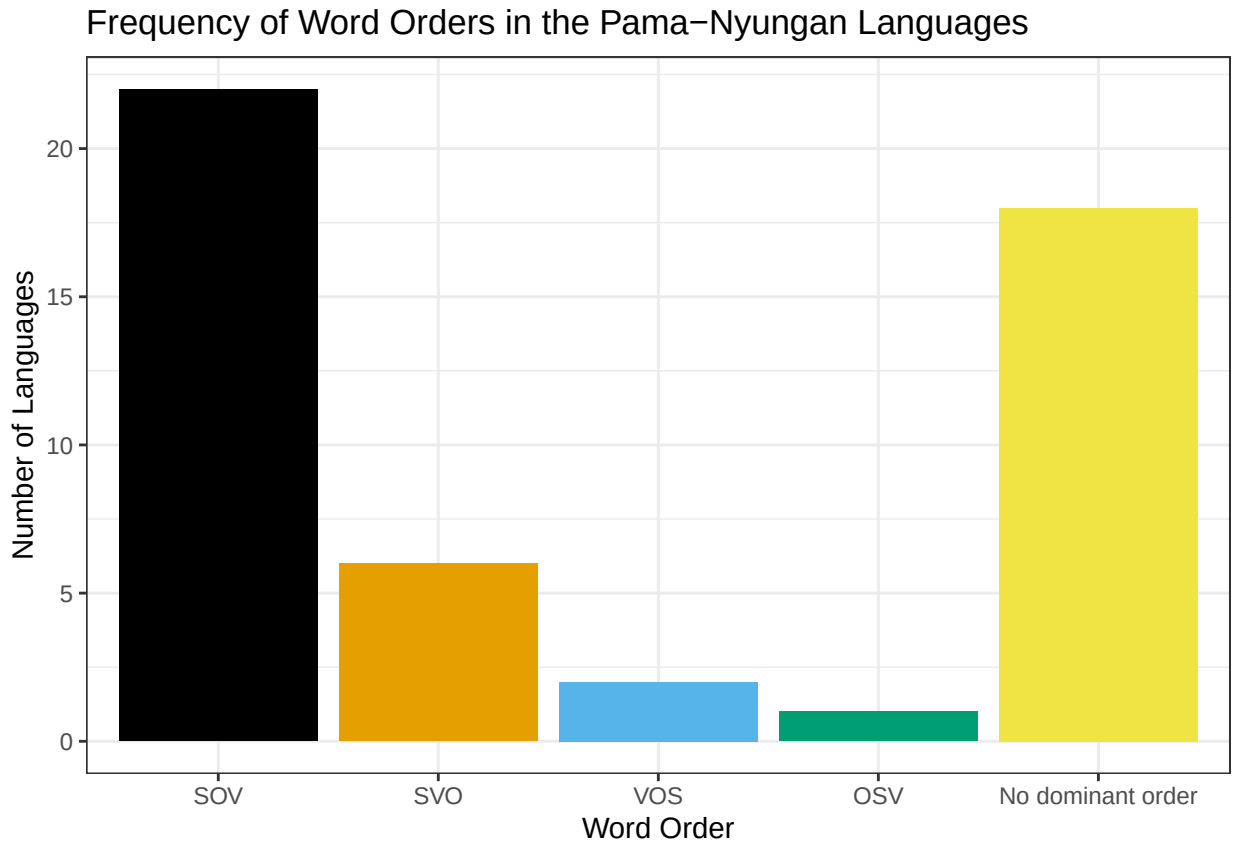
```
##Word Order Plots

###Word Order PN
histo_PN_worder <- ggplot(data = wals_worder_PN,
                          aes(x = WordOrder,
                              fill = WordOrder)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Word Order") + # renaming the axes
  theme_bw() + # adjusting the theme
```

```
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Frequency of Word Orders in the Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histo_PN_worder #getting a preview
```

```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

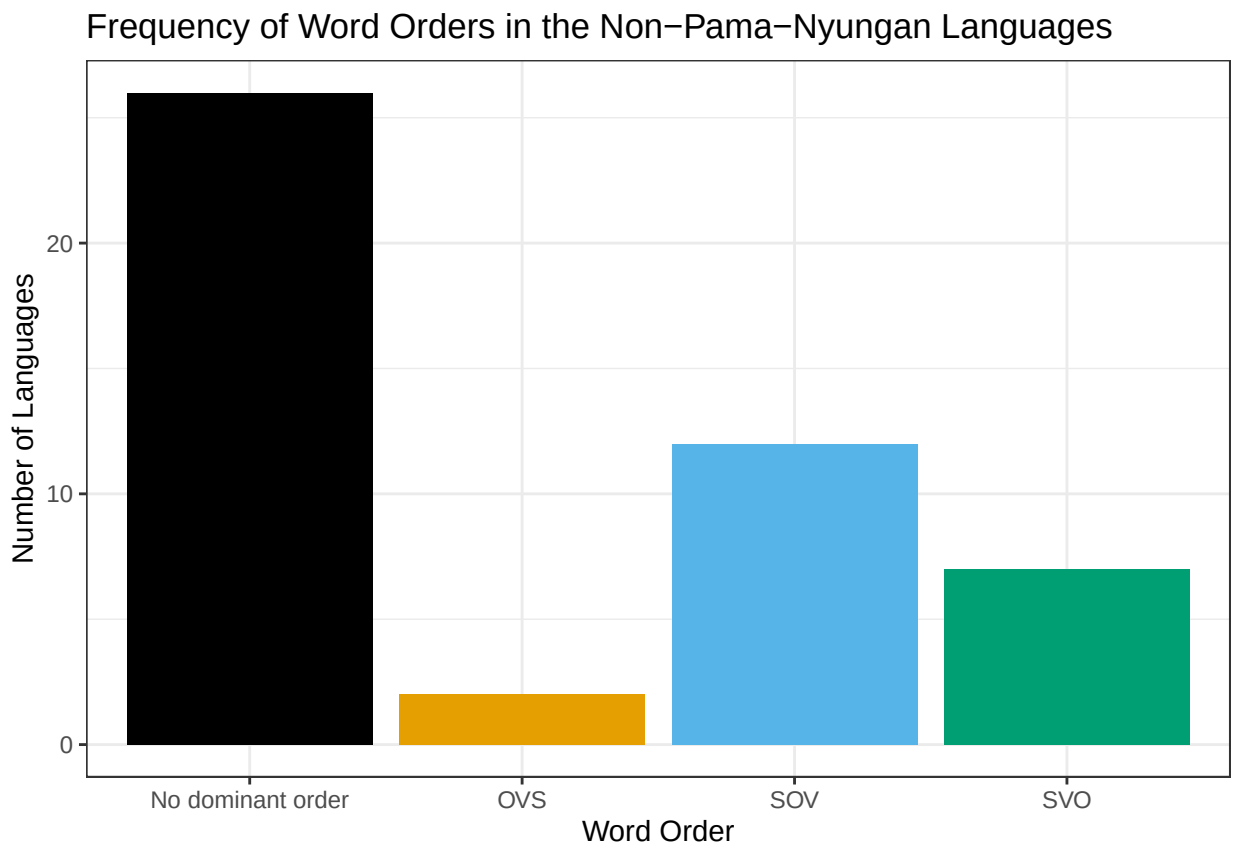



```
###Word Order nPN

histo_nPN_worder <- ggplot(data = wals_worder_nPN,
                           aes(x = WordOrder,
                               fill = WordOrder)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Word Order") + # renaming the axes
  theme_bw() + # adjusting the theme
```

```
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Frequency of Word Orders in the Non-Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histo_nPN_worder #getting a preview
```



With this additional information, we can now more clearly see that the Pama-Nyungan languages have a very clear tendency of having a SOV word order, as almost half of all languages in the dataset have this as their dominant word order. The trend of starting off a sentence with the subject continues, with six more languages having a SVO word order. Having no dominant word order is also quite common in this dataset, with a total of 18 languages fitting that profile. Only a small minority, a total of three languages, have a word order that does not fit those three major categories.
The Non-Pama-Nyungan languages show a similar pattern. Although here the clear majority does not have a dominant word order, the majority of those who do have one, will have one that starts their word order with the subject of the sentence.

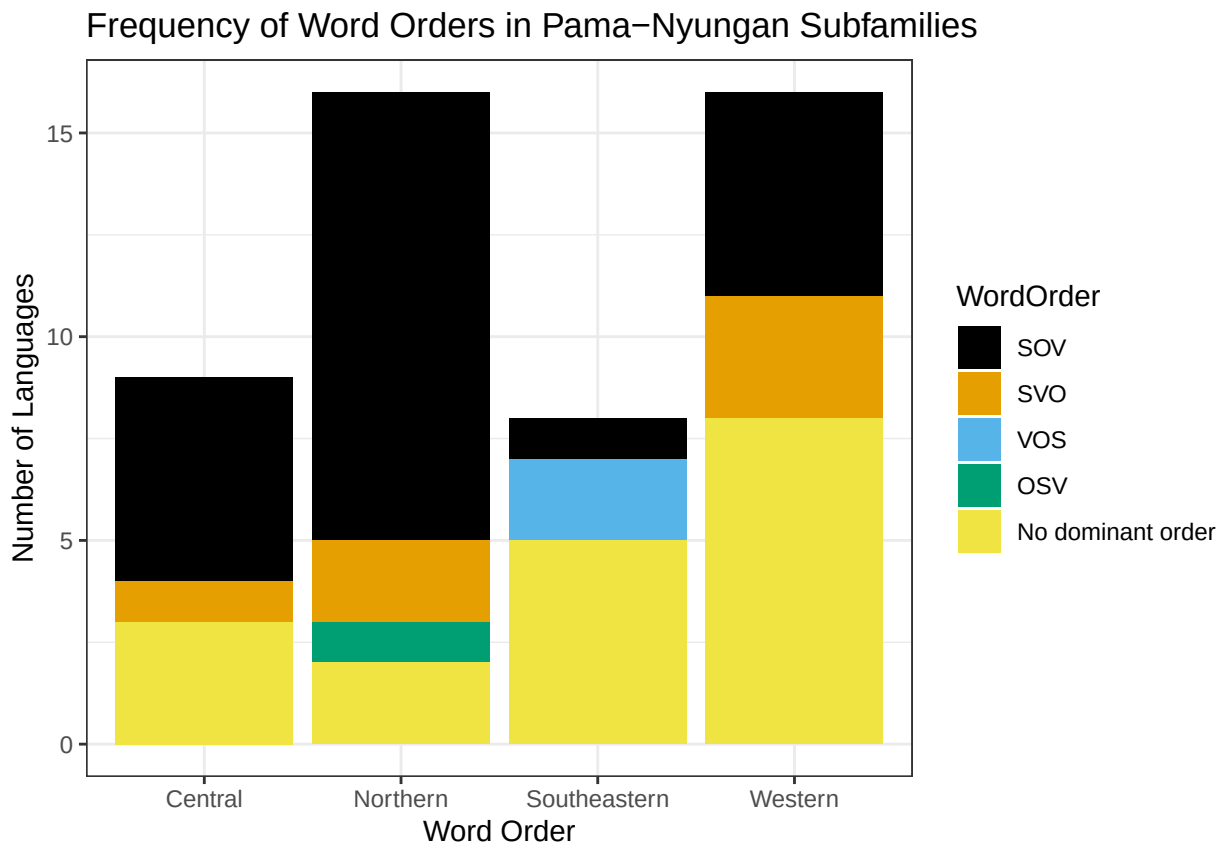
Circling back to the Pama-Nyungan languages, let's look at the Word Order distribution if we separate them by genus. This might give us additional insight into the regional distribution of word order in Australia.

```
#adapting the df to make the plot more readable in the end

wals_worder_PN_readable <- wals_worder_PN %>%
  mutate(Genus = case_when(
    Genus == "Central Pama-Nyungan" ~ "Central",
```

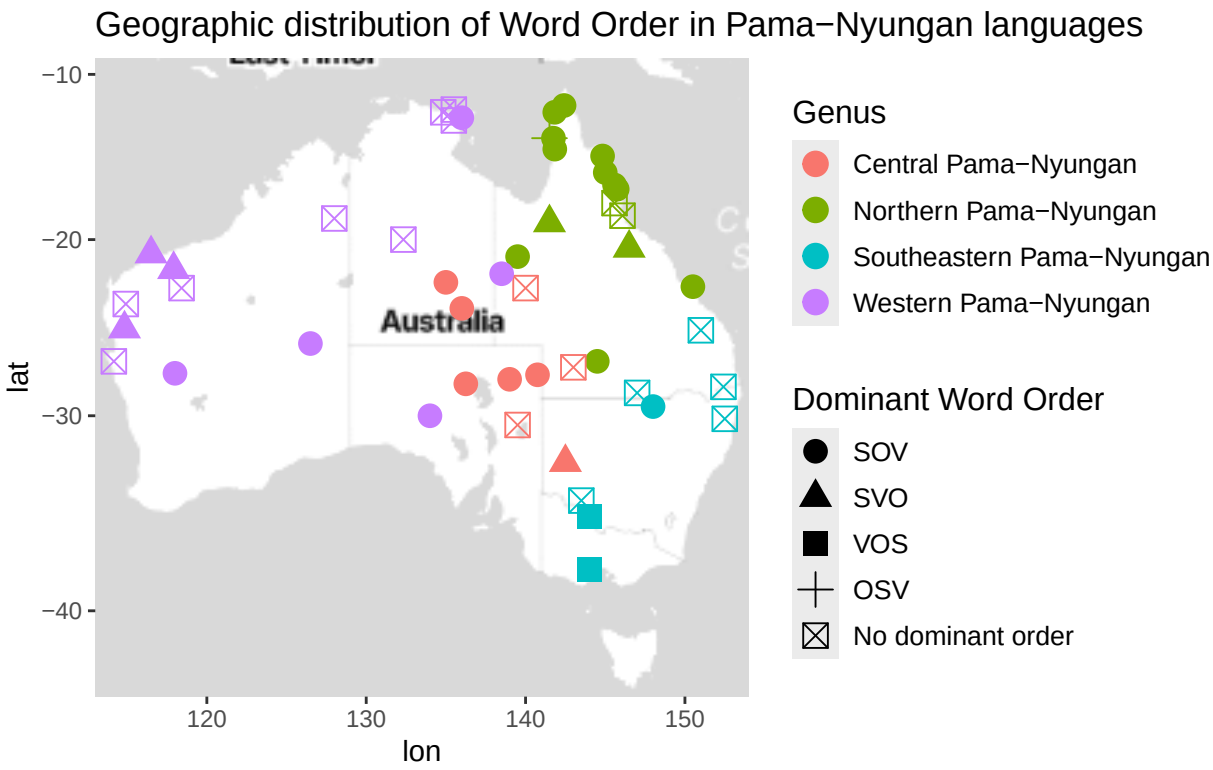
```
    Genus == "Northern Pama-Nyungan" ~ "Northern",
    Genus == "Southeastern Pama-Nyungan" ~ "Southeastern",
    Genus == "Western Pama-Nyungan" ~ "Western",
    TRUE ~ Genus # Keep other values unchanged
  ))


#Worder Plot for PM, x = genus
histogram_PNGenus_worder <- ggplot(data = filter(wals_worder_PN_readable),
                                   aes(x = Genus,
                                       fill = WordOrder)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x = "Word Order") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind() +
  ggtitle("Frequency of Word Orders in Pama-Nyungan Subfamilies") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histogram_PNGenus_worder #getting a preview
```



```
map_PN_worder_genus <- ggmap(map_AUS) +
  geom_point(data = wals_worder_PN,
             aes( x = Longitude,
                  y = Latitude,
                  color = Genus,
                  shape = WordOrder),
             show.legend = T,
             size = 4) + #might be a bit too small for the size I save it as, but overlapping dots are
  #theme_map() +
  theme(legend.text = element_text(size = 10),
        legend.title = element_text(size = 12)) +
  labs(color = "Genus",
       shape = "Dominant Word Order") +
  ggtitle("Geographic distribution of Word Order in Pama-Nyungan languages")
map_PN_worder_genus #getting a preview
```



One thing that becomes apparent immediately when considering this plot is the fact that the Northern Pama-Nyungan languages have a disproportionate amount of SOV word order languages when compared to the other families. It is possible that their geographical proximity to the Non-Pama-Nyungan languages might be an explanation for this phenomenon, though in that case we might also expect the Northern Pama-Nyungan languages to show a tendency towards a lack of dominant word order, though the bar plot would have us infer the opposite. Taking the map into consideration, this theory seems somewhat more plausible, as it is indeed the languages spoken more closely to the Non-Pama-Nyungan languages that do not have a dominant word order. The Central Pama-Nyungan languages have a very similar distribution profile as the Northern ones, albeit not as extreme. The Southeastern Pama-Nyungan languages seem to be the most different compared to their counterparts. While at first glance appearing to have a general

lack of dominant word order like many of the others, they also have a distinctly small number of languages with a SOV word order system, with only one language in the dataset fitting that description. Lastly, the Western Pama-Nyungan languages appear almost prototypical for the language family in general, as they are predominantly without dominant word order, closely followed by the SOV and SVO word orders. Their complete lack of other word orders also correlates with the general rarity of those in the overall dataset.

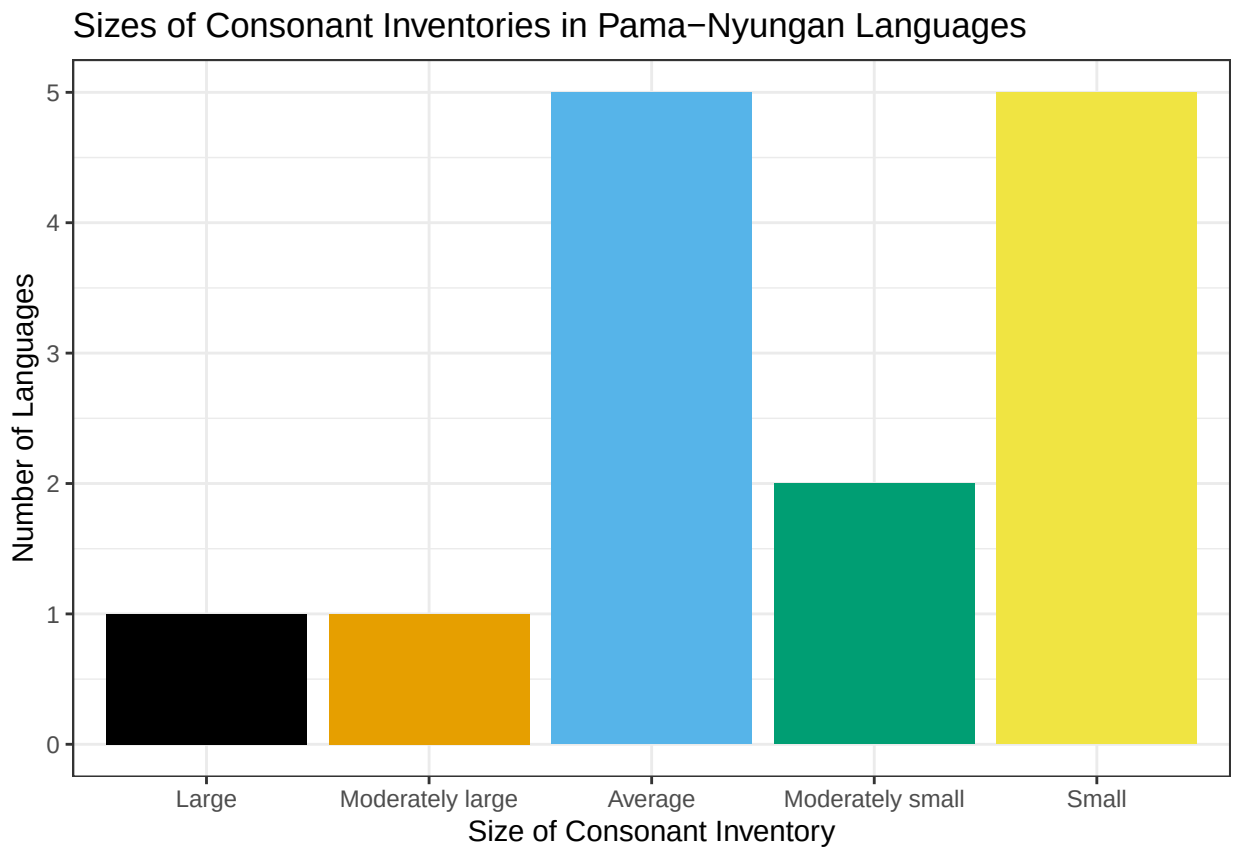
Phoneme Inventory

Next, I will be looking at the phoneme inventories of the two language families. As the data by WALS does not give explicit data on that but rather simply ranges of inventory sizes, I will try my best to gain some meaningful insight from this approach.
Let's start off by comparing the consonant inventories of Pama-Nyungan and Non-Pama-Nyungan.

```
#Consonant Inventory
histo_PN_cons <- ggplot(data = wals_cons_PN,
                        aes(x = ConsonantInventories,
                            fill = ConsonantInventories)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Size of Consonant Inventory") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Sizes of Consonant Inventories in Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

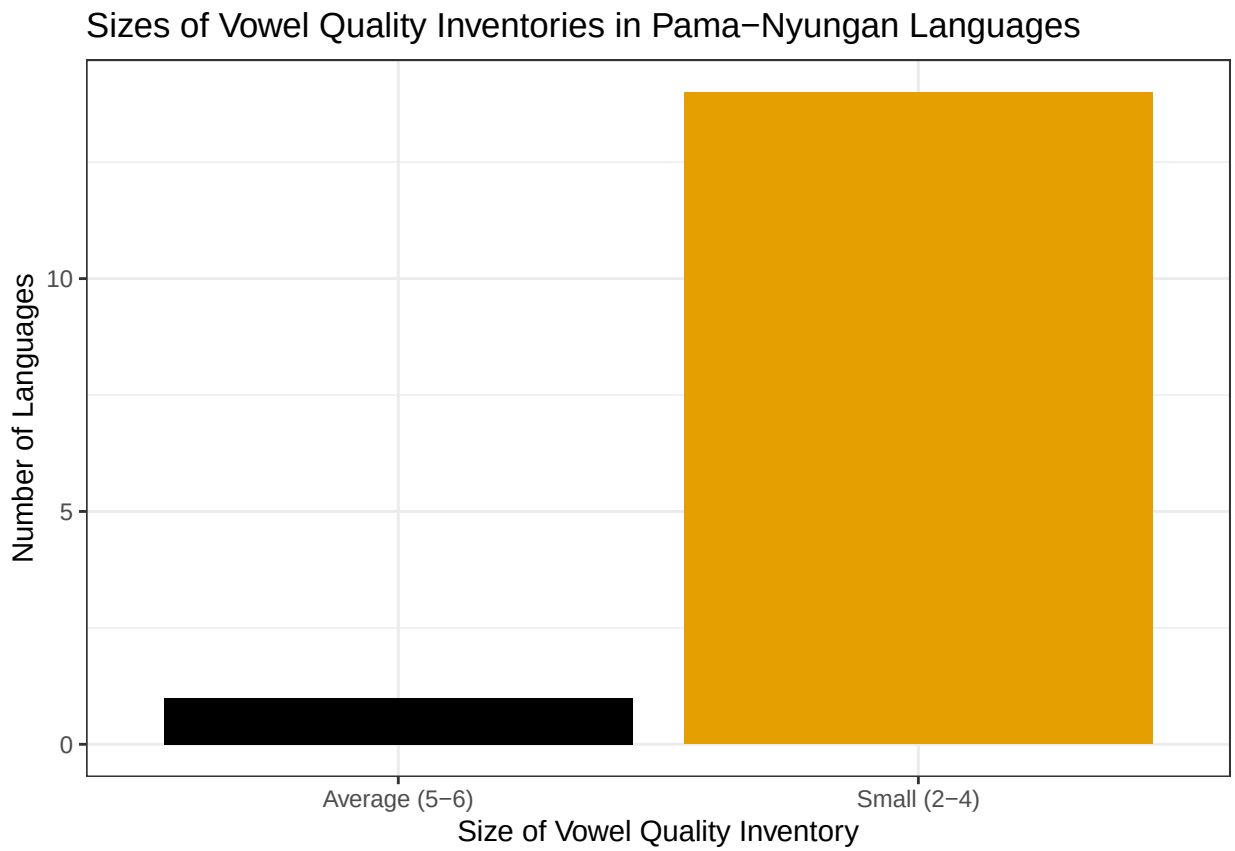
```
histo_PN_cons #getting a preview
```




```
#Vowel Quality Inventory
histo_PN_vowels <- ggplot(data = wals_vowels_PN,
                          aes(x = VowelQualityInventories,
                              fill = VowelQualityInventories)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Size of Vowel Quality Inventory") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Sizes of Vowel Quality Inventories in Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histo_PN_vowels #getting a preview
```

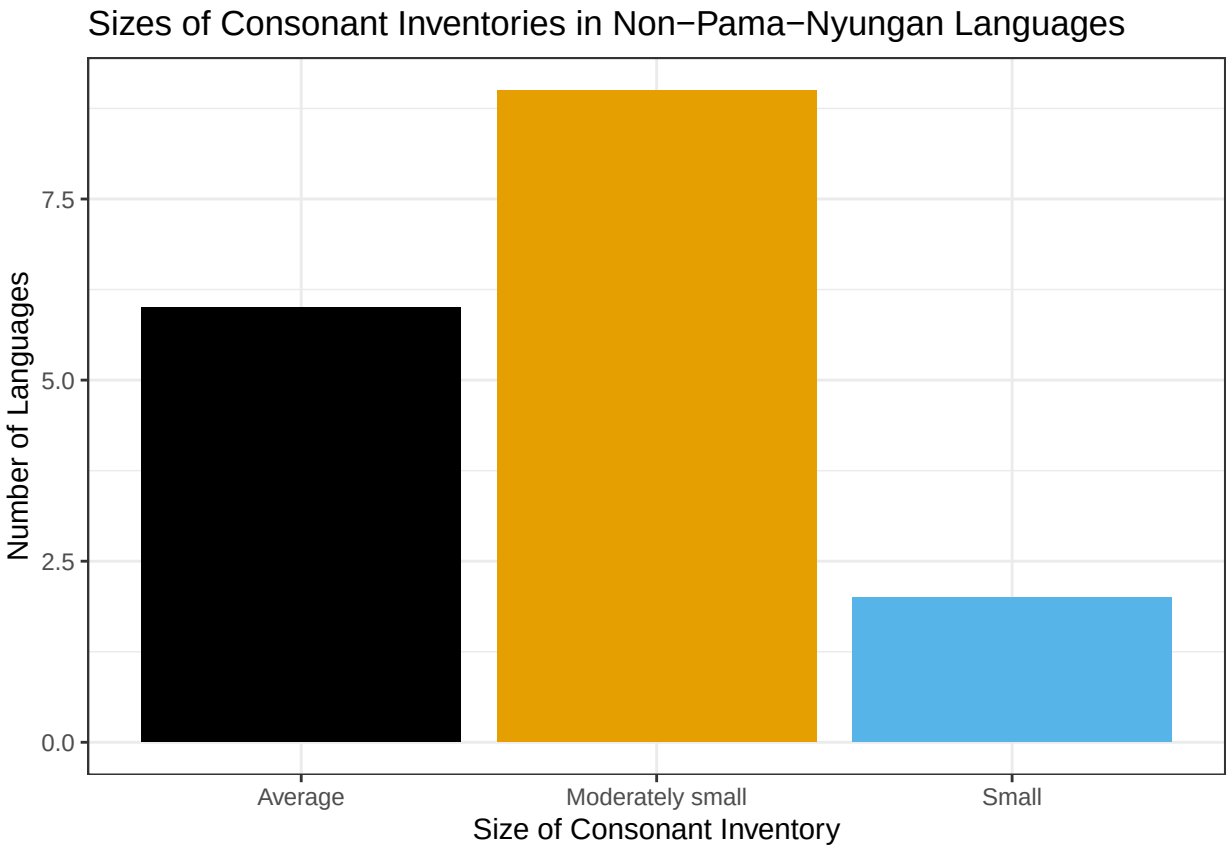


These bar plots clearly show that Pama-Nyungan languages have a strong tendency towards both small consonant as well as small vowel inventories. Let's plot the same for Non-Pama-Nyungan and see if the trend continues.

```
##nPN
#Consonant Inventory
histo_nPN_cons <- ggplot(data = wals_cons_nPN,
                         aes(x = ConsonantInventories,
                             fill = ConsonantInventories)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Size of Consonant Inventory") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Sizes of Consonant Inventories in Non-Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

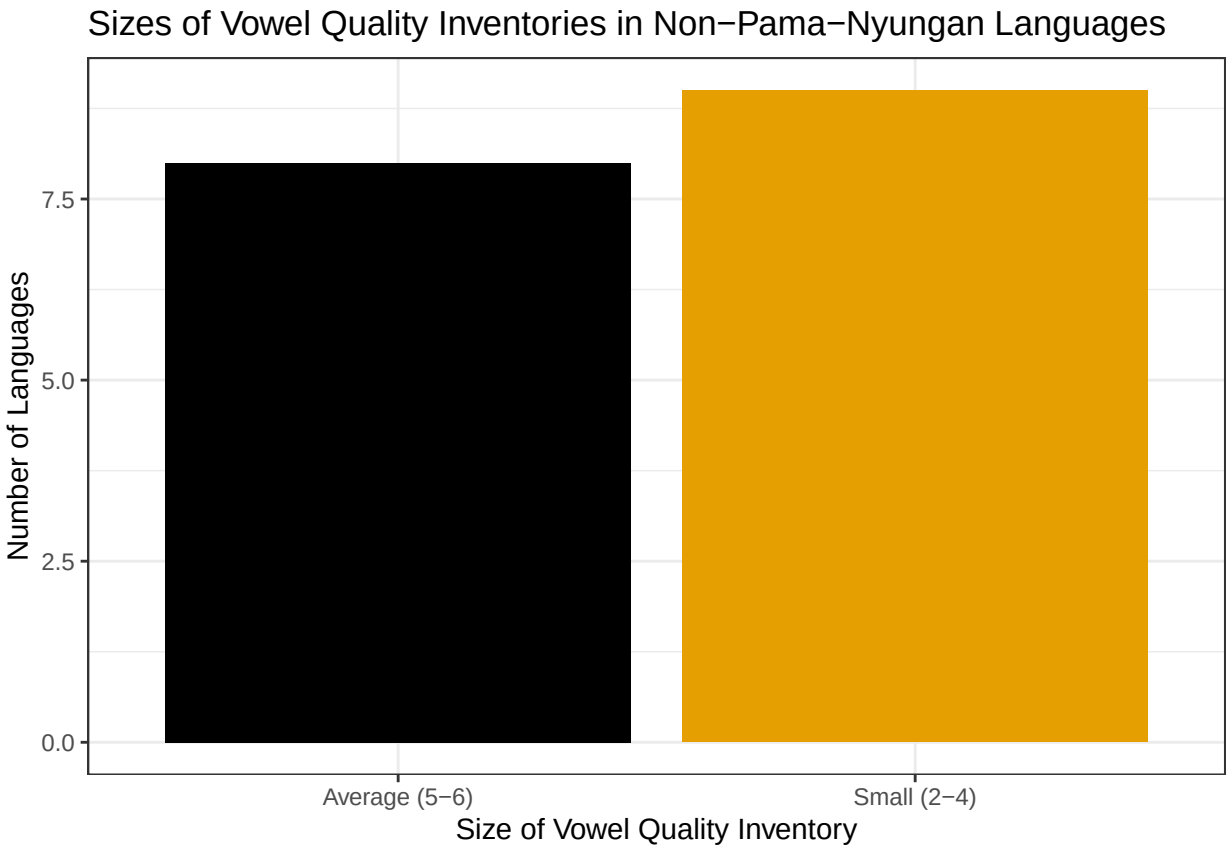
```
histo_nPN_cons #getting a preview
```




```
#Vowel Quality Inventory
histo_nPN_vowels <- ggplot(data = wals_vowels_nPN,
                           aes(x = VowelQualityInventories,
                               fill = VowelQualityInventories)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Size of Vowel Quality Inventory") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Sizes of Vowel Quality Inventories in Non-Pama-Nyungan Languages") # adding a descriptive ti
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histo_nPN_vowels #getting a preview
```

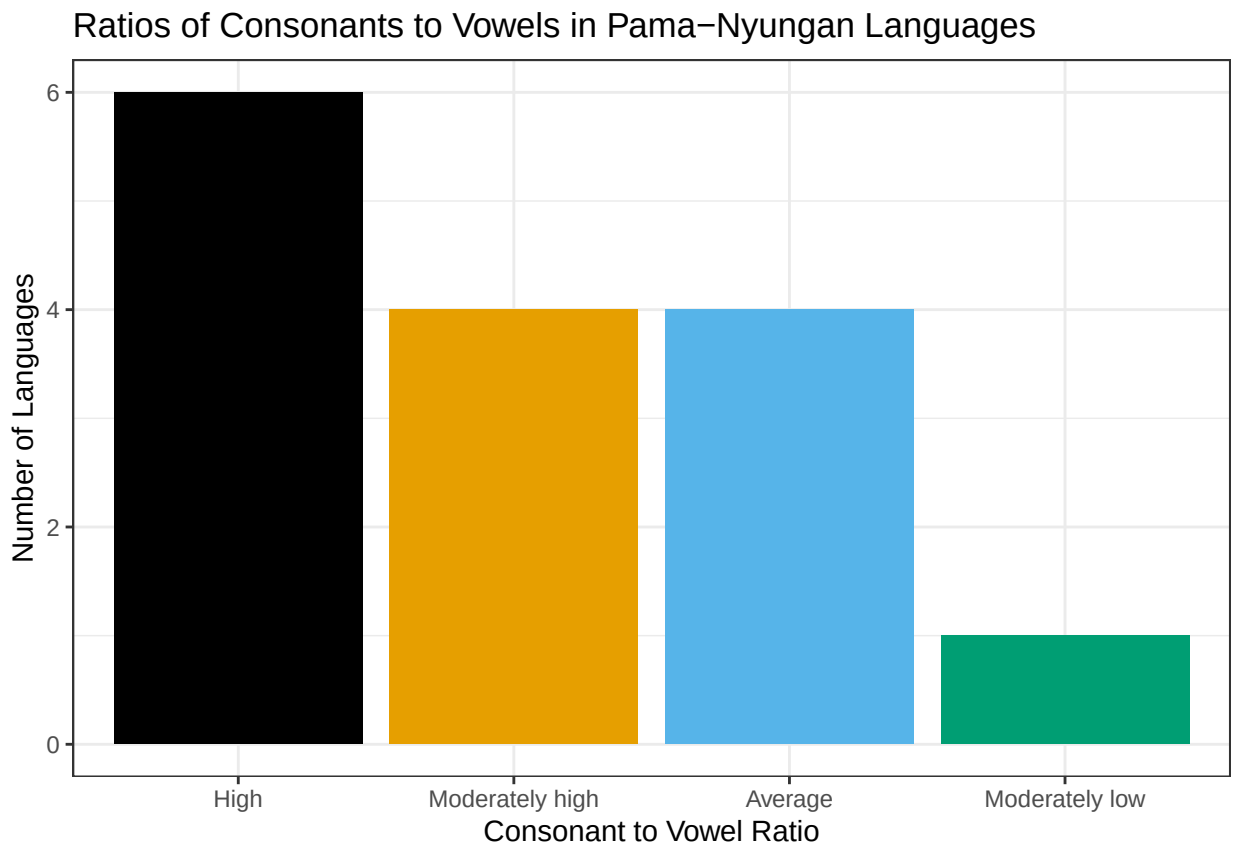


It is apparent that the Non-Pama-Nyungan languages also tend towards rather small consonant inventories, even more so than the Pama-Nyungan languages, as here we do not find any languages with a large or moderately large consonant inventory at all. But the Non-Pama-Nyungan languages are starkly contrasting their Pama-Nyungan counterparts when it comes to vowel inventories. There are a lot more Non-Pama-Nyungan languages that have a vowel quality inventory larger than 4, though the majority is still slightly tending towards a smaller inventory of 4 or less.
Lastly, let's compare the Consonant to Vowel Ratio of Pama-Nyungan and Non-Pama-Nyungan languages.

```
#Consonant/Vowel L + Ratio
#PN
histo_PN_ratio <- ggplot(data = wals_ratio_PN,
                         aes(x = ConsonantVowelRatio,
                             fill = ConsonantVowelRatio)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Consonant to Vowel Ratio") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Ratios of Consonants to Vowels in Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

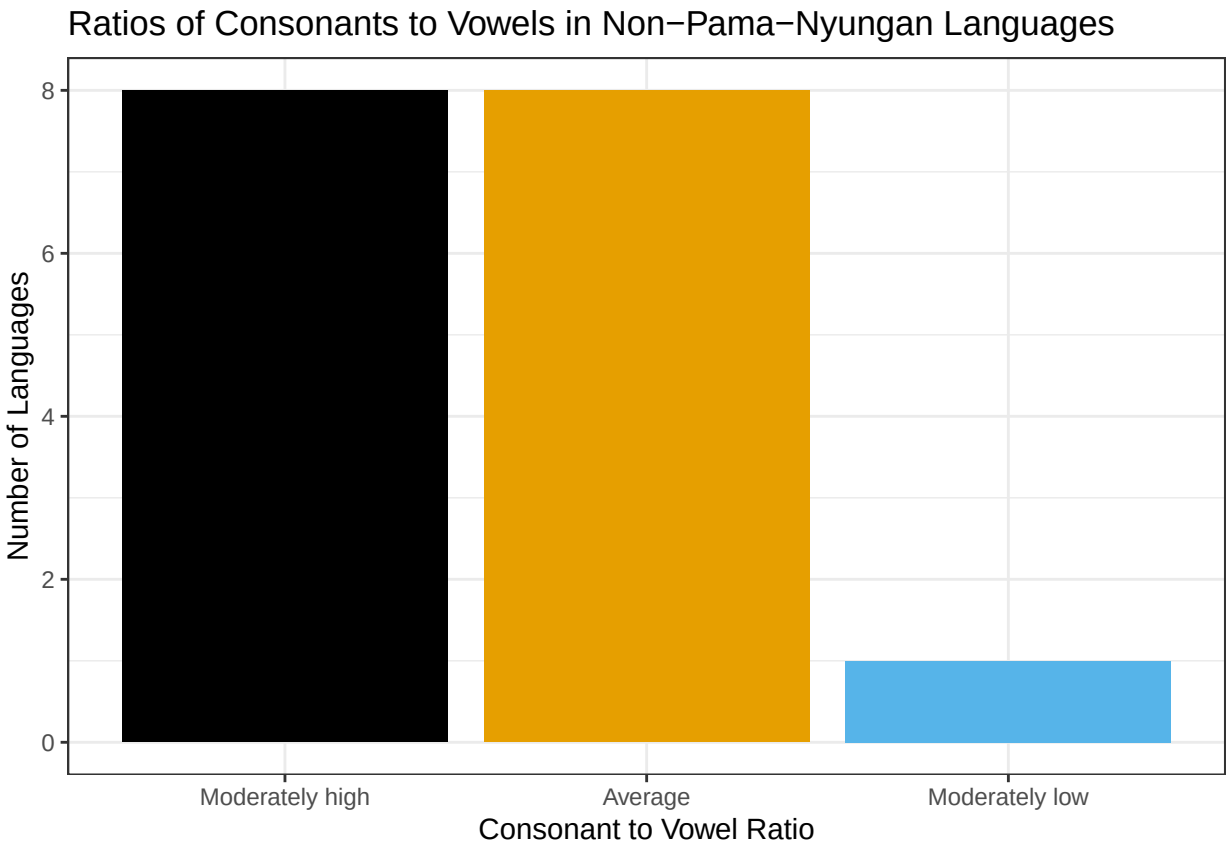
```
histo_PN_ratio #getting a preview
```




```
#nPN

histo_nPN_ratio <- ggplot(data = wals_ratio_nPN,
                          aes(x = ConsonantVowelRatio,
                              fill = ConsonantVowelRatio)) +
  geom_histogram(stat = "count") + # defining the type of plot
  labs(y = "Number of Languages",
       x= "Consonant to Vowel Ratio") + # renaming the axes
  theme_bw() + # adjusting the theme
  scale_fill_colorblind(guide = FALSE) +
  ggtitle("Ratios of Consonants to Vowels in Non-Pama-Nyungan Languages") # adding a descriptive title
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```

```
histo_nPN_ratio #getting a preview
```



Unsurprisingly, the Pama-Nyungan languages show a generally higher ratio of consonants to vowels. This is driven both by their generally slightly bigger consonant inventories and consistently smaller vowel quality inventories respectively. It is worth noting that there is still one Pama-Nyungan language that has a moderately low ratio, which is unexpected given the data. This language, Wik Munkan, does not have any apparent differences that make it stick out in other regards. It is spoken in Northern Queensland, like many other Northern Pama-Nyungan languages. The Non-Pama-Nyungan languages generally tend towards having a moderately high or average consonant to vowel ratio. Here we again have a single outlier in the Western Daly language Maranungku that shows a moderately low ratio,.
It is worth noting that the data points for all of these sets are significantly smaller than the ones regarding word order.

Discussion

The analysis of the typological features discussed above give some limited insight as to the influence of language contact. The word orders are relatively well distributed on the continent, though SOV in Pama-Nyungan is very concentrated in Northern Queensland, whereas in the Non-Pama-Nyungan languages it is found more often in the Northwest of the Northern Territories. For both "families", the languages with no apparent word order are well distributed and do not seem to follow an apparent trend. The analysis of Phoneme Inventories has also not turned up any significant or apparent differences that could be attributed to language contact easily. Admittedly, in retrospect these typological features do not seem to be the best fit for this kind of task. Given this, I have to rescind my initial hypothesis of there being a clear effect of geographical proximity on the typological features discussed. It seems that while that might have some minimal effect, it is not at all conclusive given my limited analysis.

A major point to consider when evaluating all of this data is the fact that it does not cover a representative part of the language families. As established in the introduction, there are over 300 different Pama-Nyungan

languages, yet the dataset for word order contained only 49 data points, so only around 15%. Looking at the Phoneme Inventory is even worse, with only around 15 data points each, so less than 5% of total Pama-Nyungan languages. With a total of 139 Australian languages on glottolog not considered part of the Pama-Nyungan family(Hammarström et al, 2024), the discrepancies are similar, with only 29% of them considered in the data regarding word order and only around 11% in the data points regarding the Phoneme Inventory.

Another important aspect to be aware of is the hand waiving of pooling all the Non-Pama-Nyungan languages into one group. This category actually consists of many different language families altogether that are not nearly as closely related, if at all, as all the Pama-Nyungan languages are, so every conclusion drawn about Non-Pama-Nyungan languages in general should be taken with that in mind.

Finally, I wonder about the accuracy of the genus data that I used in the bar plot where I visually separated the Word Order occurrences by genera of Pama-Nyungan languages. It strikes me as odd that these are simply named by their geographical grouping and I therefore question if they are actually representative of genealogical connectedness. When considering the more recent data on Glottolog, it quickly became apparent, that this is indeed an outdated way of grouping the languages.

With all these shortcomings of mine and the data in mind, I would argue that the data ended up being barely sufficient for what I set out to accomplish, with it ending up being a lot less clear of a result than I initially anticipated, given my initial reading of the literature. Again, I would most likely attribute this to the data only covering a rather small part of the Australian languages and the fact that I only considered this one set of data.

Bibliography

Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson (2018). The origin and expansion of Pama–Nyungan languages across Australia. Nature Ecology & Evolution (2), 741–749.
Bowern, Claire / Koch, Harold (Publ.) (2004): Australian Languages. Classification and the comparative method. Amsterdam / Philadelphia: John Benjamins.
Dixon, R.M.W. (1980): The Languages of Australia. Cambridge: Cambridge University Press.
Dixon, R.M.W. (2004): Australian Languages. Their Nature and Development. Cambridge: Cambridge University Press.
Greenhill, Simon (2024). rcldf: rcldf - Read Linguistic Data In The Cross Linguistic Data Format (CLDF)_. R package version 1.2.0, commit 3979a89dbe4db653873caca62212fc07ebb966e9, https://github.com/SimonGreenhill/rcldf
Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2024. Glottolog 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.10804357 (Available online at http://glottolog.org, Accessed on 2024-08-22.)
Kahle, D., Wickham, H. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
Matthew S. Dryer. (2013) Order of Subject, Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/81, Accessed on 2024-04-16.)
Moran, Steven & McCloy, Daniel (eds.) 2019. PHOIBLE. Jena: Max Planck Institute for the Science of Human History. (Available online at https://phoible.org)
O'Grady, G. N. (1998). Toward a Proto-Pama-Nyungan Stem List, Part I: Sets J1-J25. Oceanic Linguistics, 37(2), 209–233.
Schmidt, W. (1919). Die Gliederung der australischen Sprachen: geographische, bibliographische, linguistische Grundzüge der Erforschung der australischen Sprachen. Mechitharisten-Buchdruckerei.
Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer (2024). viridis(Lite) - Colorblind-Friendly Color Maps for R. viridis package version 0.6.5.
Wickham, H et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4 (43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686
Zuckermann, G. et al. (2021) LARA in the Service of Revivalistics and Documentary Linguistics: Community Engagement and Endangered Languages. Proceedings of the Workshop on Computational Methods for Endangered Languages (1), 13-23.

Everything related to this project can be found in this GitHub repository.