Linguistic Data Analysis - Final Project

Benedict Wuethrich

2024-07-11

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 aren't thought of as Pama-Nyungan and actually contain 27 different language families. Exactly which langauges belong to which category is an ongoing debate.

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 are having 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 has 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 were not without their controversy. Dixon vehemently denied the applicability of the comparative method on Australian Languages, which lead to a rather large amount of 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, like 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, they 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 will be looking at the geographical distribution of the Australian languages and how they compare to each other in terms of categorical features such as word order, use of reduplication and more. 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. Furthermore, for certain features, I will focus solely on the Pama-Nyungan languages.

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 dataframes eficiently and how to structure the creation of new dataframes from the given sets without creating too much redundancy. I created a seperate Rscript file that entails all the necessary data wrangling and manipulation for the different plots and maps used in this project. The other scripts rely on this Data Wrangling script to be run first, so as to load the required dataframes. This Rmarkdown file compiles them all into one coherent document, though without saving the plots and maps again. The seperate 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 dataframe was created using the Value Table and filtering for Australian languages only.

```
##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)
```

```
## A Cross-Linguistic Data Format (CLDF) dataset:
## Name: The World Atlas of Language Structures Online
## Path: C:/Users/Benedict/Documents/GitHub/LDAR-Project_BW/LDAR-FP-BW/data/WALS/cldf
## Type: http://cldf.clld.org/v1.0/terms.rdf#StructureDataset
## Tables:
##   1/12: areas.csv (3 columns, 11 rows)
##   2/12: CodeTable (6 columns, 1143 rows)
##   3/12: ContributionTable (11 columns, 152 rows)
##   4/12: contributors.csv (4 columns, 55 rows)
##   5/12: countries.csv (2 columns, 192 rows)
##   6/12: ExampleTable (8 columns, 3907 rows)
##   7/12: language_names.csv (4 columns, 7377 rows)
##   8/12: LanguageTable (17 columns, 3573 rows)
##   9/12: MediaTable (8 columns, 153 rows)
##   10/12: ParameterTable (5 columns, 192 rows)
##   11/12: TreeTable (8 columns, 254 rows)
##   12/12: ValueTable (8 columns, 76475 rows)
## Sources: 7373
```

```
#This creates a tibble that will be manipulated further at different points.
wals_value <- 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 seperate dataframes that specifically only included either Pama-Nyungan or Non-Pama-Nyungan languages. Additionaly, the relevent column was renamed to be more descriptive and therefore

easier to work with, then mutated into factors, so the different entries could be clearly ordered. In the case of the Non-Pama-Nyungan languages, another step was taken where we add 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 seperate df, for PN and nPN respectively.

wals_worder_PN <- wals_value %>%
  filter( Chapter_ID == "81" &
            Family == "Pama-Nyungan") %>%
  rename("WordOrder" = "Name.CodeTable") %>%
  mutate(WordOrder = factor(WordOrder,
                            levels = c("SOV",
                                       "SVO",
                                       "VOS",
                                       "OSV",
                                       "No dominant order")))

wals_worder_nPN <- wals_value %>%
  filter( Chapter_ID == "81" &
            Family != "Pama-Nyungan") %>%
  rename("WordOrder" = "Name.CodeTable") %>%
  mutate(WordOrder = factor(WordOrder,
                            levels = c("SOV",
                                       "SVO",
                                       "SVO or VOS",
                                       "OSV",
                                       "No dominant order"))) %>%
  mutate(WordOrder = case_when( #this bit was added as three languages didn't have entries in the new W
    Language_ID.ValueTable == "ung" ~ "OVS",
    Language_ID.ValueTable == "myi" ~ "OVS",
    Language_ID.ValueTable == "grr" ~ "No dominant order",
    TRUE ~ WordOrder))
```

Analysis

First off, let's see 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 dataframe in order to make the legend of the final map more easily readable. Using this new dataframe, 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")
```

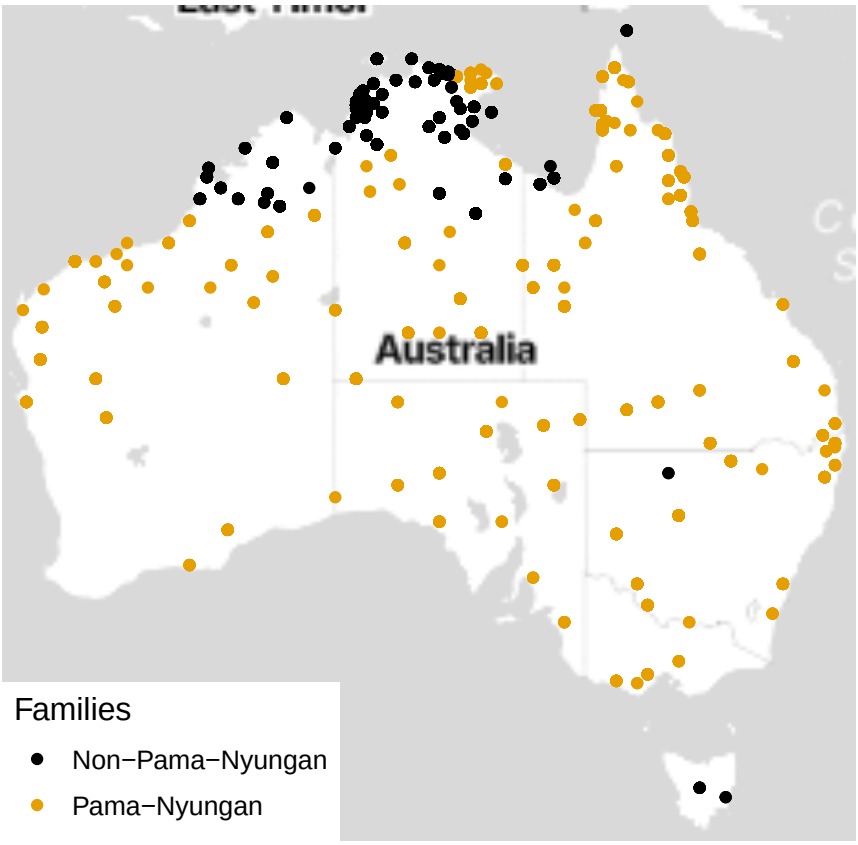
```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```
ggmap(map_AUS)
```


```
#This map will show the distribution of PN and nPN languages. While reductionist, I want to make it eas


wals_nPN <- wals_value %>%
  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 = "Families")
map_AUS_family
```



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

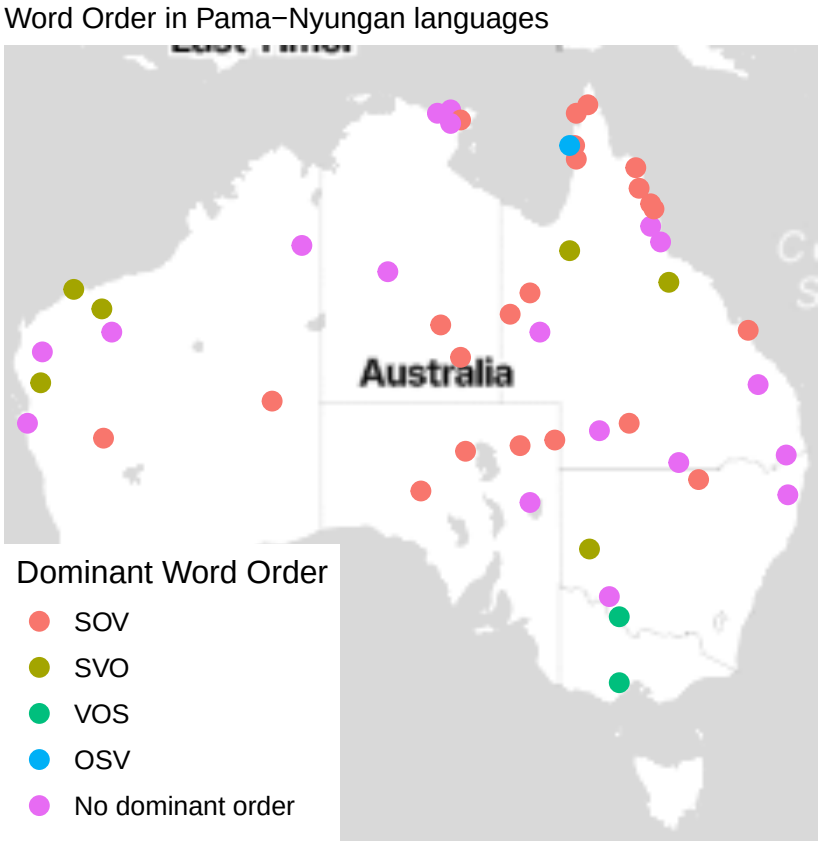
Word Order

The word order of different languages is described as "perhaps the single most frequently cited typological feature of languages" by Dryer and Haspelmath on the website concerning the Chapter about word order. 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 seperate maps for both Pama-Nyungan and Non-Pama-Nyungan, as well as seperate bar plots to explore the frequency of each within the respective categories.

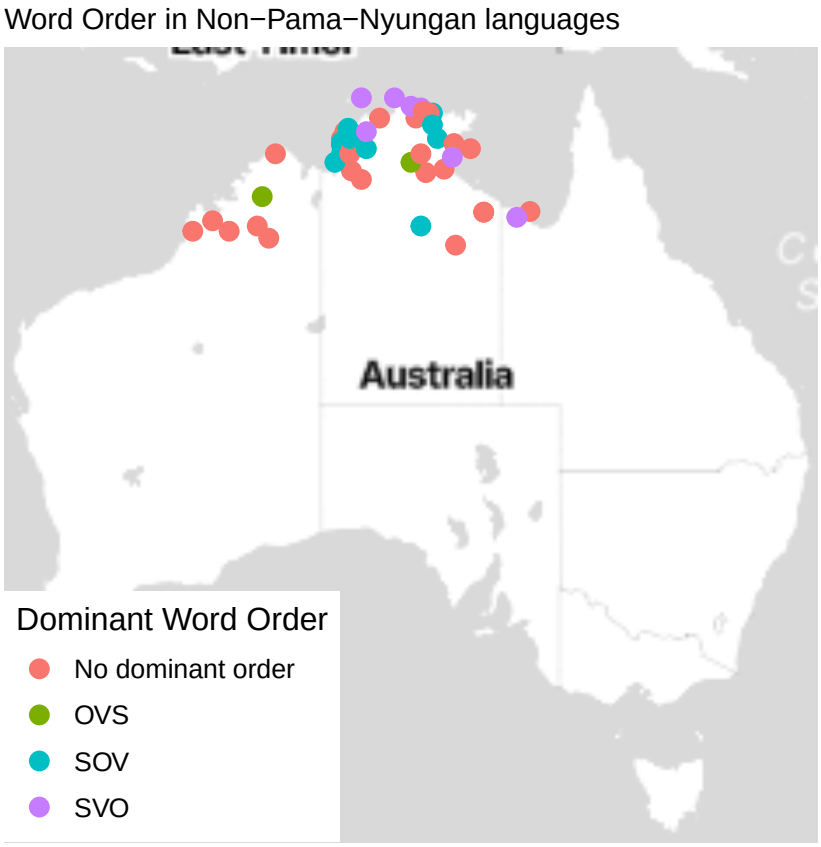
```
#a different map that shows the dominant word order of Australian languages
#first we need to filter for data that entails the Word order. This is chapter 81 of WALS. we further s

##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 do
```

The maps paint an unlcear 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.

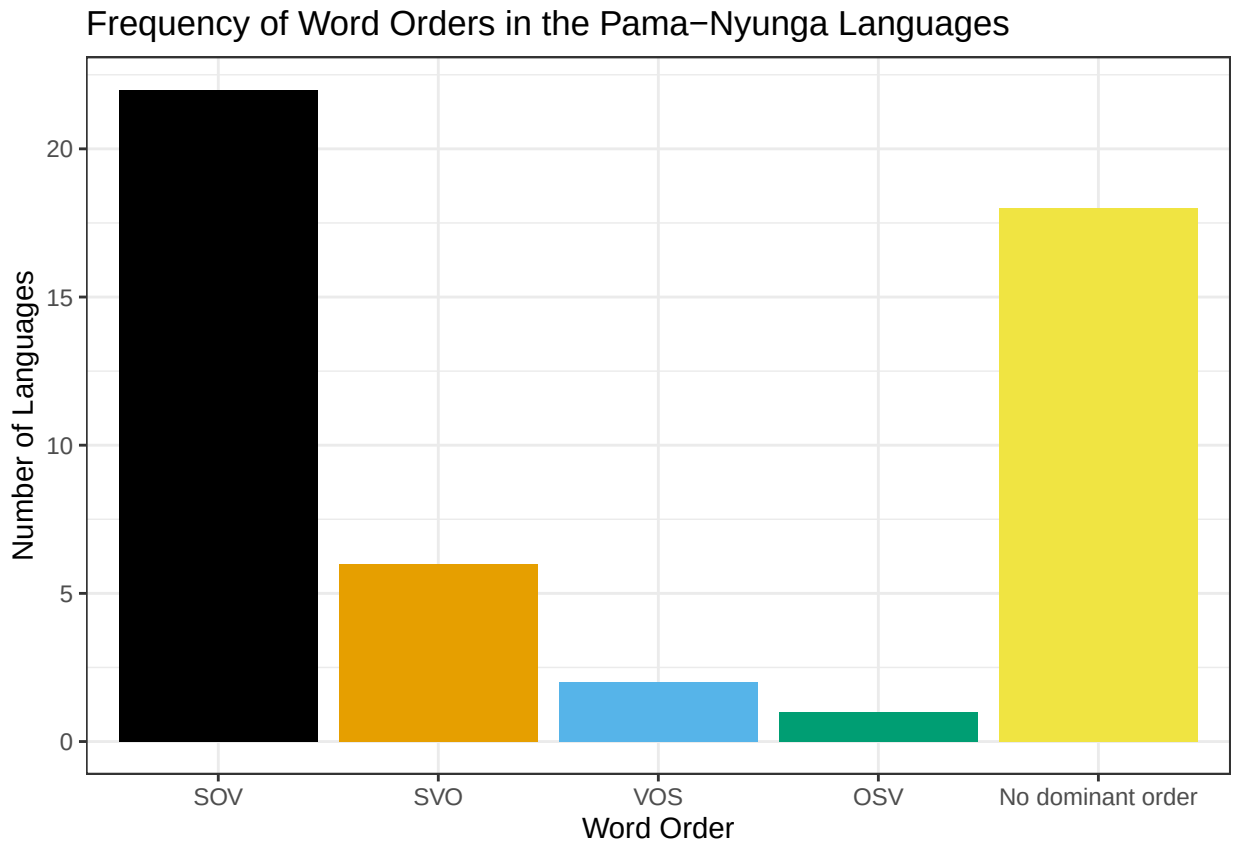
```
##Word Order Plots

###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-Nyunga 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.
```

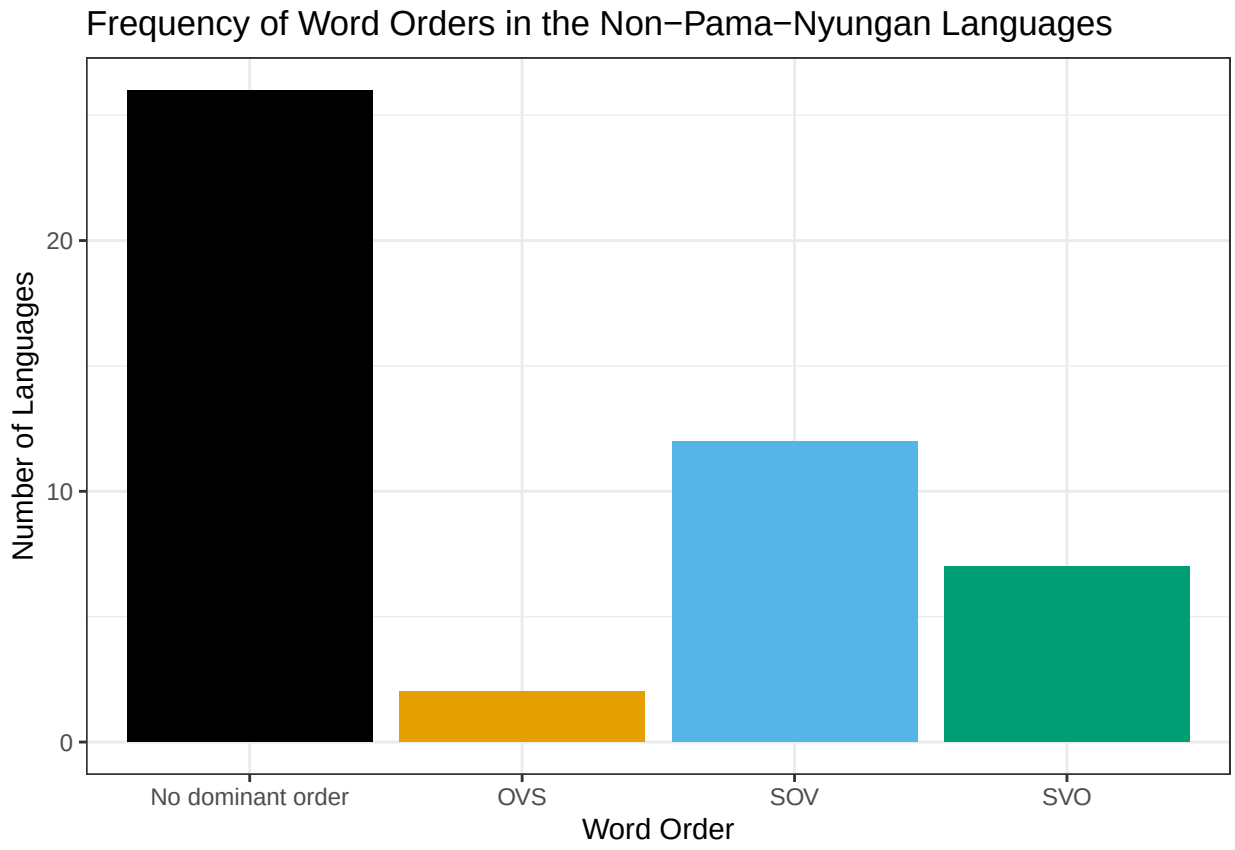



```
###Word Order nPN

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 tital 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 majoriy of those who do have one, will have one that starts their word order with the subject of the sentence.

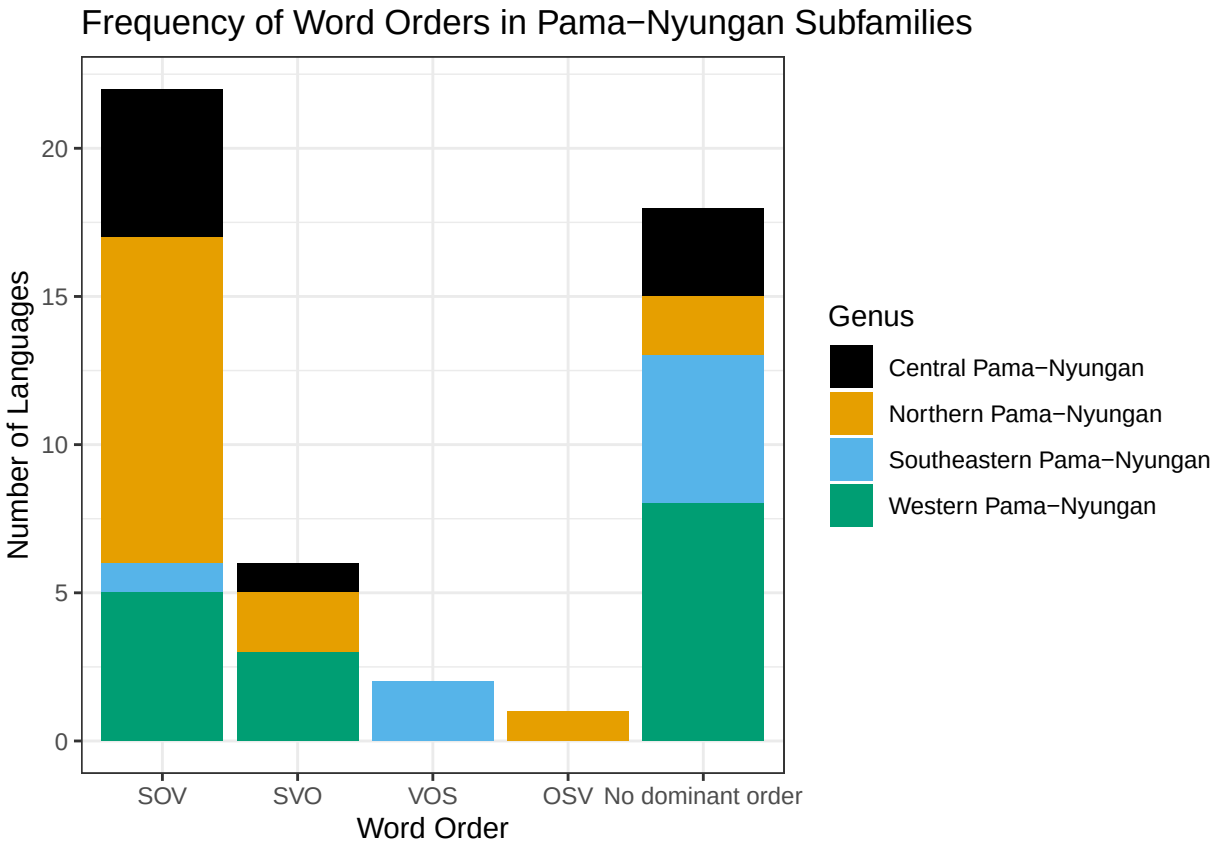
Circling back to the Pama-Nyungan languages, let's look at the Word Order distribution if we separate them by Genus. This might give us additional insight into the regional distribution of word order in Australia.

```
#Worder Plot for PM, fill = genus
histogram_PNGenus_worder <- ggplot(data = filter(wals_worder_PN),
                                   aes(x = WordOrder,
                                       fill = Genus)) +
  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
```



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 geograhical 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, but this plot would have us infer the opposite. The Central Pama-Nyungan languages have a very similar distribution profile as the Northern ones, albeit not as extreme.

Discussion

Accuracy of genus data. simply regional or based? Pooling nPN never a good idea Data represents only around 15% of PN languages

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
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.