Linguistic Data Analysis - Final Project

Benedict Wuethrich

2024-07-10

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

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 categories like word order or phoneme inventory. 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.

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 relatively extensive. 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.

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
```

```r
glimpse(wals_value)
```

```
## Rows: 5,096
## Columns: 40
## $ ID                          <chr> "26A-abn", "30A-abn", "31A-abn", "32A-abn",~
## $ Language_ID.ValueTable      <chr> "abn", "abn", "abn", "abn", "abn", "abn", "~
## $ Parameter_ID.ValueTable     <chr> "26A", "30A", "31A", "32A", "33A", "35A", "~
## $ Value                       <chr> "2", "1", "1", "1", "2", "4", "2", "3", "1"~
## $ Code_ID                     <chr> "26A-2", "30A-1", "31A-1", "32A-1", "33A-2"~
## $ Comment.ValueTable          <chr> NA, NA, NA, NA, NA, NA, NA, "<div class=\"E~
## $ Source.ValueTable           <chr> "Hercus-1994[passim]", "Hercus-1994", "Herc~
## $ Example_ID                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "ig~
## $ Name.ParameterTable         <chr> "Prefixing vs. Suffixing in Inflectional Mo~
## $ Description.ParameterTable <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ ColumnSpec                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Chapter_ID                  <chr> "26", "30", "31", "32", "33", "35", "39", "~
## $ Parameter_ID.CodeTable      <chr> "26A", "30A", "31A", "32A", "33A", "35A", "~
## $ Name.CodeTable              <chr> "Strongly suffixing", "None", "No gender", ~
## $ Description.CodeTable       <chr> "Predominantly suffixing", "None", "No gend~
## $ Number                      <int> 2, 1, 1, 1, 2, 4, 2, 3, 1, 6, 1, 1, 1, 2, 4~
## $ icon                        <chr> "c0000dd", "cffffff", "cffffff", "cffffff",~
## $ Name.LanguageTable          <chr> "Arabana", "Arabana", "Arabana", "Arabana",~
## $ Macroarea                   <chr> "Australia", "Australia", "Australia", "Aus~
## $ Latitude                    <dbl> -28.25, -28.25, -28.25, -28.25, -28.25, -28~
## $ Longitude                   <dbl> 136.25, 136.25, 136.25, 136.25, 136.25, 136~
## $ Glottocode                  <chr> "arab1267", "arab1267", "arab1267", "arab12~
## $ ISO639P3code                <chr> "ard", "ard", "ard", "ard", "ard", "ard", "~
## $ Family                      <chr> "Pama-Nyungan", "Pama-Nyungan", "Pama-Nyung~
## $ Subfamily                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Genus                       <chr> "Central Pama-Nyungan", "Central Pama-Nyung~
## $ GenusIcon                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ ISO_codes                   <chr> "ard", "ard", "ard", "ard", "ard", "ard", "~
## $ Samples_100                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ Samples_200                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ Country_ID                  <chr> "AU", "AU", "AU", "AU", "AU", "AU", "AU", "~
## $ Source.LanguageTable        <chr> "Hercus-1994", "Hercus-1994", "Hercus-1994"~
## $ Parent_ID                   <chr> "genus-centralpamanyungan", "genus-centralp~
## $ Language_ID.ExampleTable    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "ab~
## $ Primary_Text                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "Ma~
## $ Analyzed_Word               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "Ma~
## $ Gloss                       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "Fi~
## $ Translated_Text             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "He~
## $ Meta_Language_ID            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Comment.ExampleTable        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

Next up was creating seperate dataframes that specifically only included either Pama-Nyungan or Non-Pama-Nyungan languages. Additionaly, the

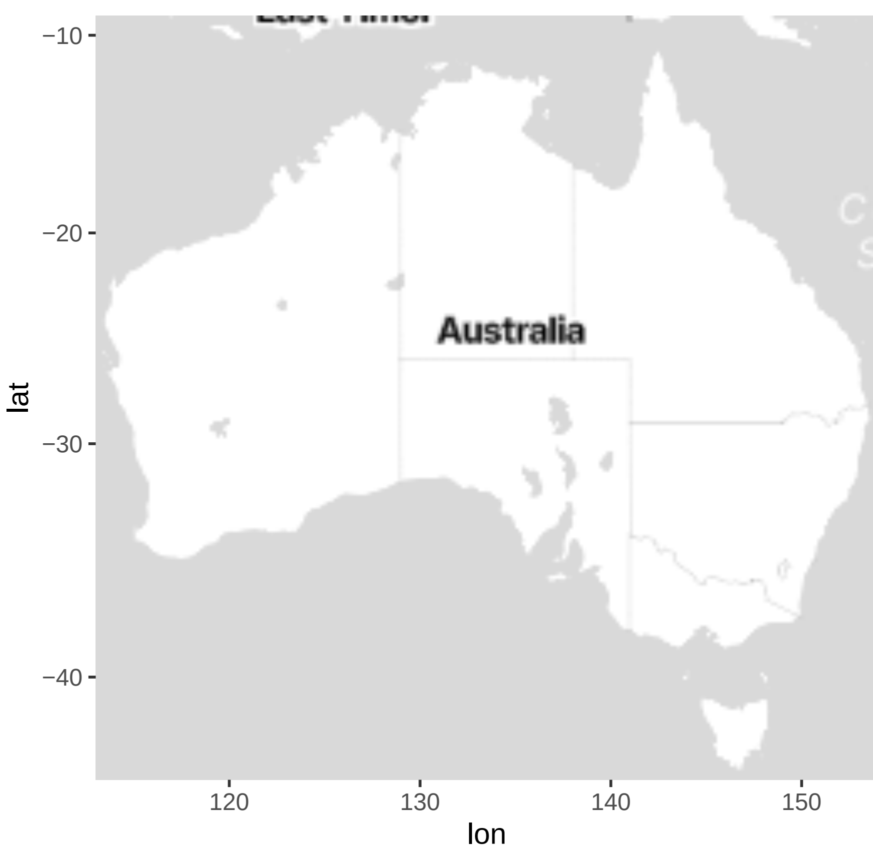
```r
##Word Order Data Wrangling
#creates two seperate df, for PN and nPN respectively. Also renames relevant columns to be more descrip

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

```
            Family == "Pama-Nyungan") %>%
  rename("WordOrder" = "Name.CodeTable")

wals_worder_nPN <- wals_value %>%
  filter( Chapter_ID == "81" &
            Family != "Pama-Nyungan") %>%
  rename("WordOrder" = "Name.CodeTable")
```

```
map_AUS <- get_stadiamap(bbox = c(left = 113,
                                  bottom = -44,
                                  right = 154,
                                  top = -9), #here the map boundaries given by ChatGPT earlier came in
                         zoom = 3,
                         maptype = "stamen_toner_lite",
                         color = "color")
```

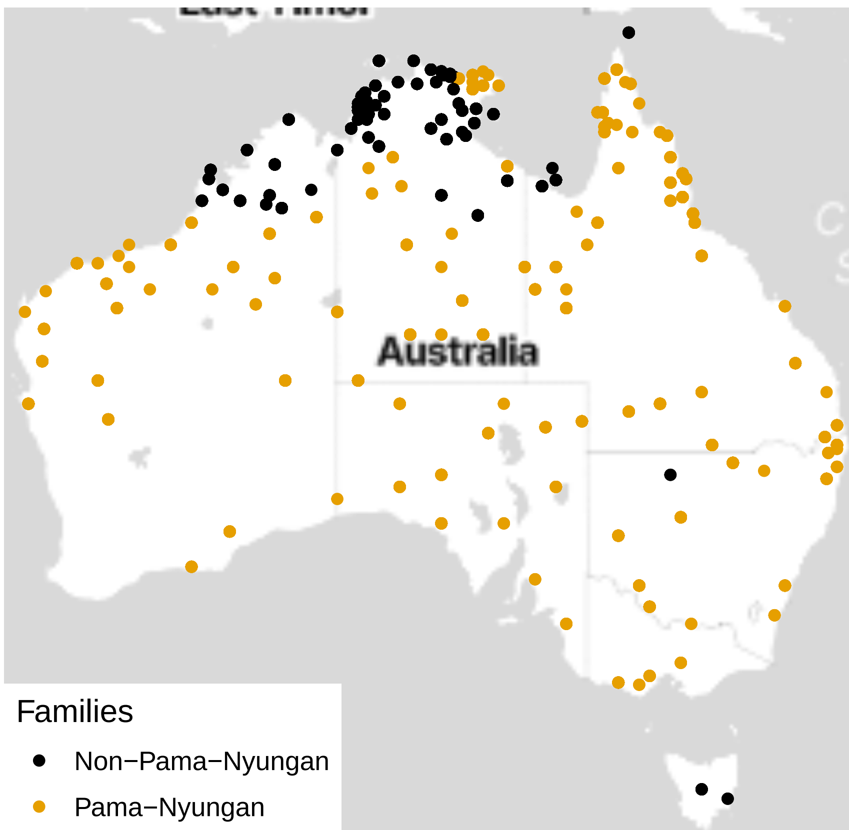
```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```
ggmap(map_AUS)
```



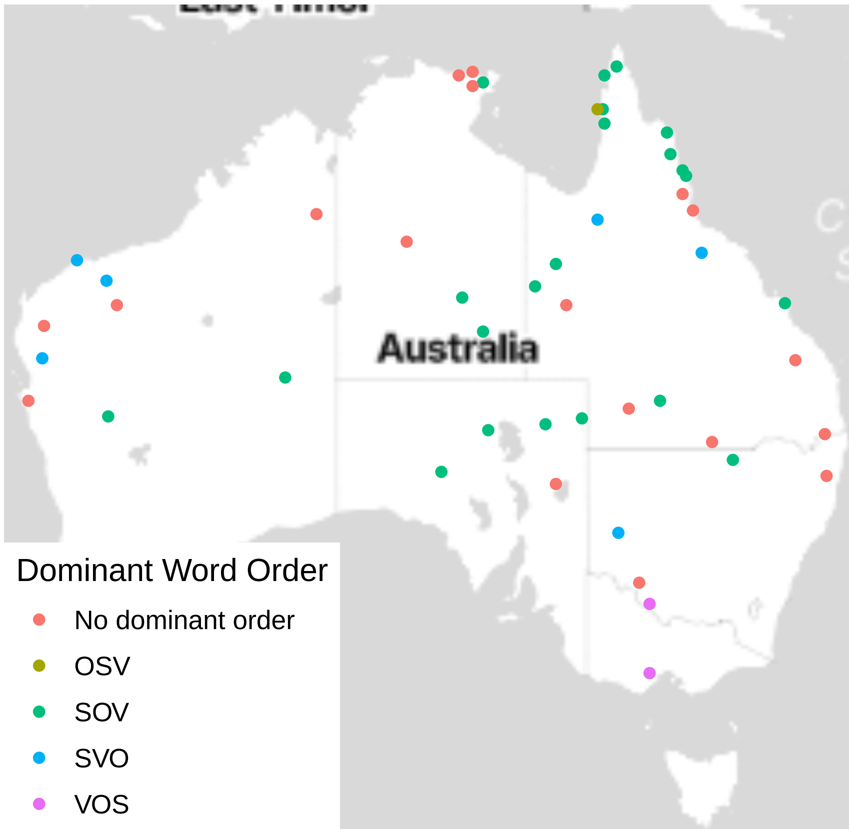

```
#This map will show the distribution of PN and nPN languages. While reductionist, I want to make it eas


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
```

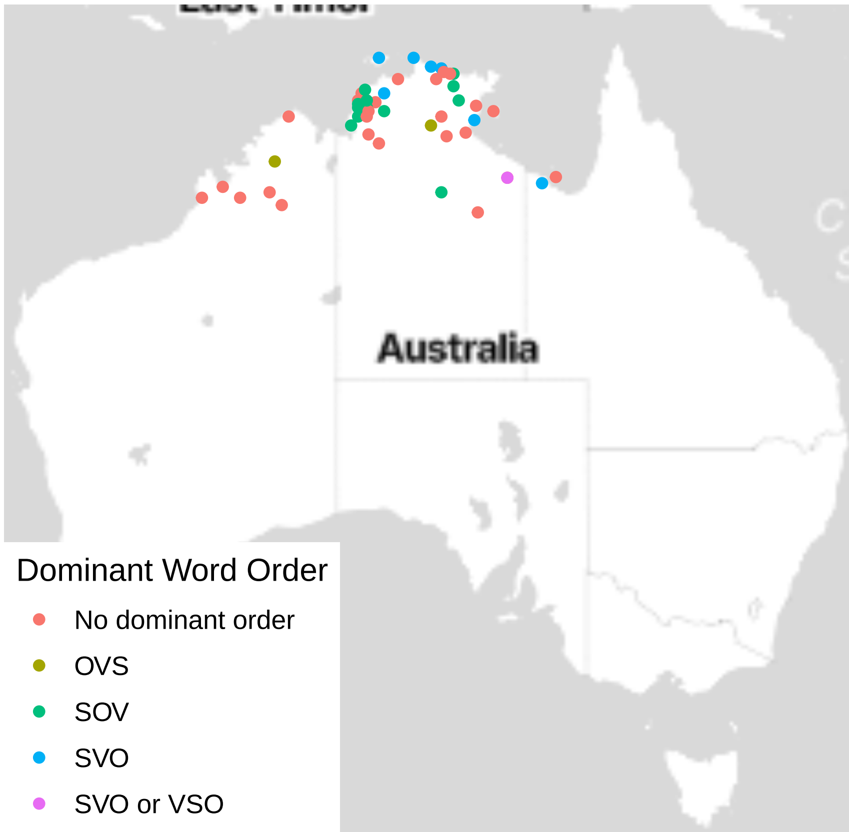



```
#a different map that shows the dominant word order of Australian languages
#first we need to filter for data that entails the Word order. This is chapter 81 of WALS. we further s

##PN
map_PN_worder <- ggmap(map_AUS) +
  geom_point(data = wals_worder_PN,
             aes( x = Longitude,
                  y = Latitude,
                  color = WordOrder),
             show.legend = T) +
```

```r
  theme_map() +
  theme(legend.text = element_text(size = 10),
        legend.title = element_text(size = 12)) +
  labs(color = "Dominant Word Order")
map_PN_worder
```




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



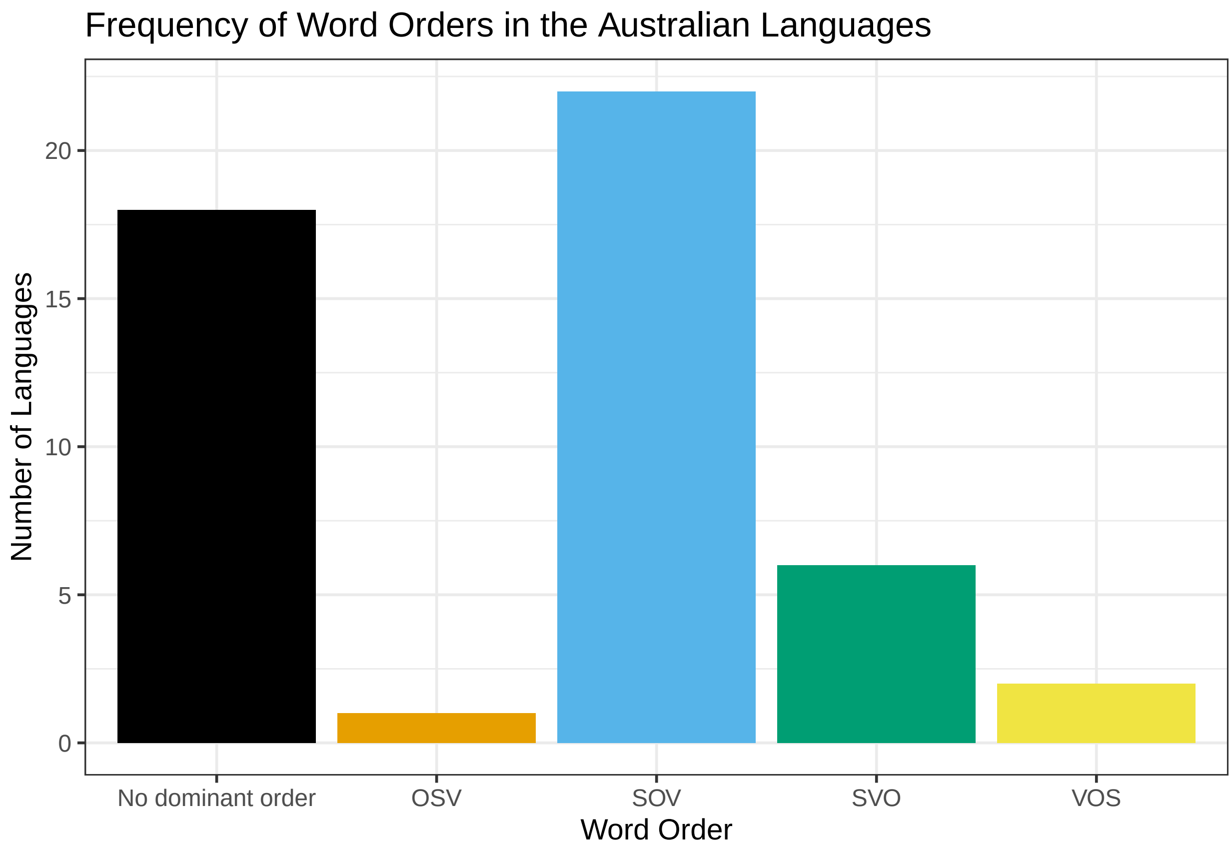

```
#PLOTS


##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 Australian 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 Australian Languages") # adding a descriptive title
```

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

```
histo_nPN_worder #getting a preview
```

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 (Hrsg.) (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.

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.