**INITIAL OBSERVATIONS PERSON**

## ###PACKAGES USED##

install.packages("ggplot2")

install.packages("esquisse")

install.packages("shiny")

install.packages ("dplyr")

library("esquisse")

library("ggplot2")

library("shiny")

**library("dplyr")**

## STEP #1 GRPH

Create a plot for each number/case combinations.

First, create a table of frequencies based on a +2 occurrence subset.

### #these create the data frames to draw the dot plot

genplu.gph1 <- PerGenPlu.wide %>% filter(tot.occ >= 2) %>%

group\_by(tot.occ, tot.form) %>%

summarise(n = n())

ggplot(data = genplu.gph1, aes(x = tot.occ, y = tot.form, size = n)) +

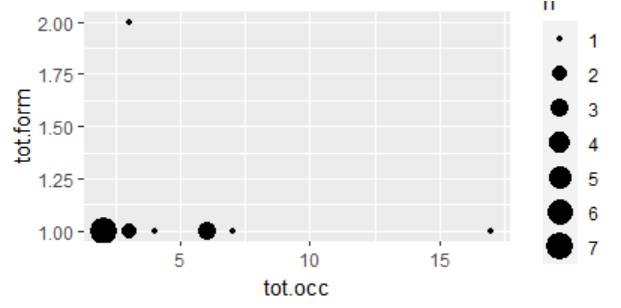
geom\_point()

Afterwards, create a graph of scattered plots where size is adjusted for frequency. We want to answer two questions:  
 a) How many people have used more than 2 occurrences for each parameter?

b) Which percentage of the +2 occurrences (i.e. use) use +1 form?

# NOM.SG

Only one person seems to have used more than one form, so this dataset is not interesting, although it may be because it is not that frequent (it is used less than 5 times by most).



## Code

# Filter the data and assign it to a new dataframe

nomsg.gph1 <- PerNomSin.wide %>%

filter(tot.occ >= 2)

# Group by variables and calculate the count

nomsg.gph2 <- nomsg.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

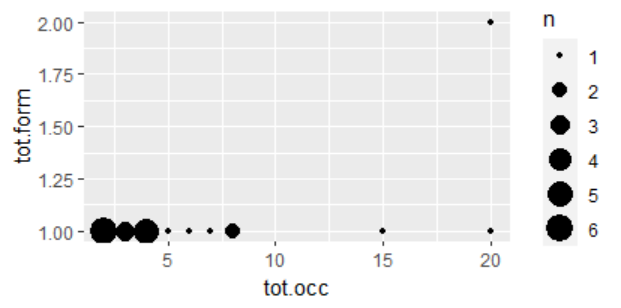
# Plot the data

ggplot(data = nomsg.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# NOM.PL

This form is not very interesting. It is very frequent, but only one person has used more than one form.



## Code

PERnompl.gph1 <- PerNomPlu.wide %>%

filter(tot.occ >= 2)

PERnompl.gph2 <- PERnompl.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = PERnompl.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

## Percentage

It was only one person, but I still wanted to check the percentage of people using more than one form, which, according to R, is 4.55 %

### Code

# Calculate the percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences

perc <- sum(PERnompl.gph2[PERnompl.gph2$tot.form >= 2, "n"]) / sum(PERnompl.gph2$n) \* 100

# Print the percentage

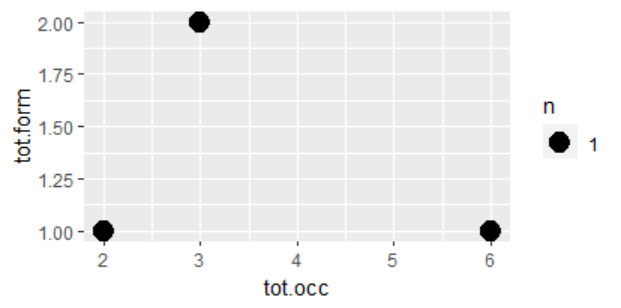
cat("Percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences:", round(perc, 2), "%\n")

# ACC.SG

It only has one form (B4), so we cannot do anything with it.

# ACC.PL

The data for this parameter are also uninteresting. The form does not seem to be very frequent (against our initial predictions), and only 1/3 of the people using it more than once has a second acc.pl form. Note, however, that the other speaker had only used it twice, so chances for variation are also a lot more limited.



## Code

PERAccPlu.gph1 <- PerAccPlu.wide %>%

filter(tot.occ >= 2)

PERAccPlu.gph2 <- PERAccPlu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = PERAccPlu.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

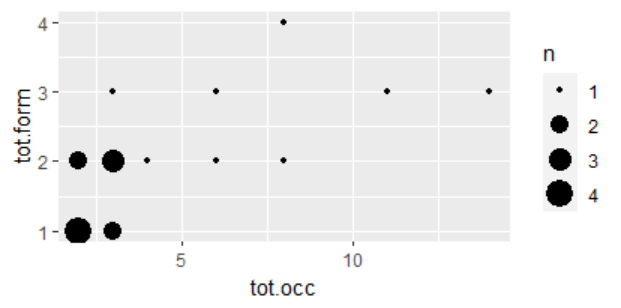
geom\_point()

# GEN.SG

It only has one form (B6), so we cannot do anything with it.

# GEN.PL

This parameter is the most colourful. Only one speaker seems to have used 4 different forms, but in general, there seems to be a regular correlation between the frequency of occurrences and the number of forms; i.e. the more utterances, the more forms people produce.



## Code

PERGenPlu.gph1 <- PerGenPlu.wide %>%

filter(tot.occ >= 2)

PERGenPlu.gph2 <- PERGenPlu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = PERGenPlu.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

## Percentage

I have tried to calculate the percentage of speakers using more than one form,out of those uttering more than one gen.pl. According to R, it is 68.42%.

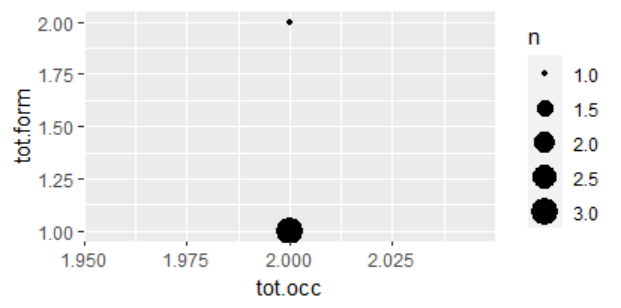
### Code

perc <- sum(PERGenPlu.gph2[PERGenPlu.gph2$tot.form >= 2, "n"]) / sum(PERGenPlu.gph2$n) \* 100

cat("Percentage of individuals with 2 or more forms out of all individuals with 2 or more occurrences:", round(perc, 2), "%\n")

# ADNM

Not surprisingly, this form is extremely rare. Only two speakers have used it. One of them has used two forms, whilst the other has been coherent within their 3 utterances. Given the scale of the data, we cannot do much further with it.



## Code

PERadnm.gph1 <- PerADMN.wide %>%

filter(tot.occ >= 2)

PERadnm.gph2 <- PERadnm.gph1 %>%

group\_by(tot.occ, tot.form) %>%

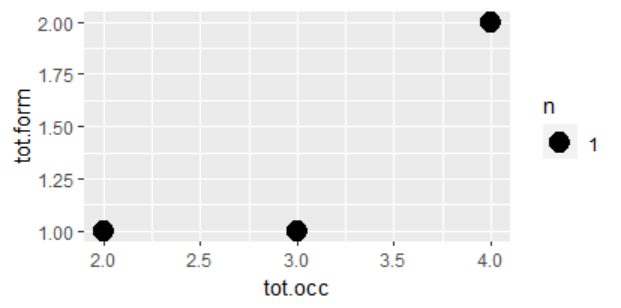
tally(name = "n")

ggplot(data = PERadnm.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# DAT.SG

This form is not very frequent and only one speaker (out of 3) has used more than one form.



## Code

PERDatSin.gph1 <- PerDatSin.wide %>%

filter(tot.occ >= 2)

PERDatSin.gph2 <- PERDatSin.gph1 %>%

group\_by(tot.occ, tot.form) %>%

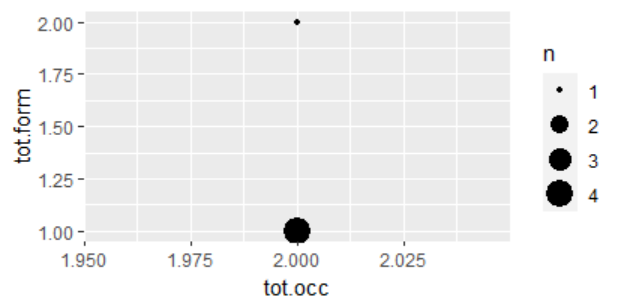
tally(name = "n")

ggplot(data = PERDatSin.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# DAT.PL

This parameter is also not interesting. Only 5 people have more than dat.pl of it, but never more than 2, and out of them, only one has used more than 1 form.



## Code

PERDatPlu.gph1 <- PerDatPlu.wide %>%

filter(tot.occ >= 2)

PERDatPlu.gph2 <- PERDatPlu.gph1 %>%

group\_by(tot.occ, tot.form) %>%

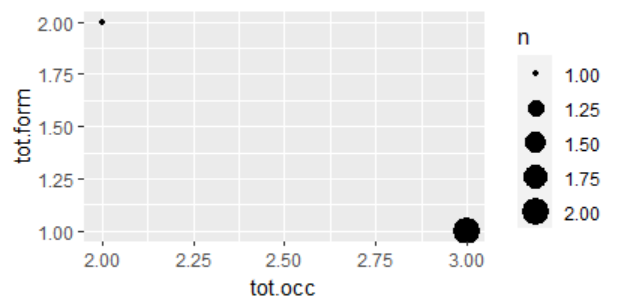
tally(name = "n")

ggplot(data = PERDatPlu.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# INS.PL

Only three speakers have produced more than one ins.pl, and out of these, only one speaker has used two forms (interestingly, being the one with less occurrences). Anyway, this data cannot tell much about this, because the ins.pl of this noun seems very infrequent.



## Code

PERInsPl.gph1 <- PerInsPlu.wide %>%

filter(tot.occ >= 2)

PERInsPl.gph2 <- PERInsPl.gph1 %>%

group\_by(tot.occ, tot.form) %>%

tally(name = "n")

ggplot(data = PERInsPl.gph2, aes(x = tot.occ, y = tot.form, size = n)) +

geom\_point()

# STEP #2 MULTILEVEL LOGISTIC REGRESSIONS

I have only focused on the main (juicy) three forms: nom.pl, acc.pl, gen.pl

## NOM.PL

I start by creating a dataframe that excludes the “FALSE” instances from the long dataset.

PerNomPlu.long\_filtered <- PerNomPlu.long %>%

filter(pres == TRUE)

I have tested this against gender, age & village

Before running more tests, I wanted to have an overview of the frequency of forms. B3 is definitely the main form, the other two are rather marginal.

A picture containing screenshot, diagram, line

Description automatically generated

#### Code

ggplot(PerNomPlu.long\_filtered) +

aes(x = form, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

### Gender& Age

Before saying that C3 and D3 are female only forms, I suspect that this has to do more with the distribution of the participants in the villages where I have worked.

A screenshot of a computer game

Description automatically generated with low confidence

#### Code

ggplot(PerNomPlu.long\_filtered) +

aes(x = Gender.age, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

### Area

In this case, B3 is a universal form, whereas C3 and D3 have a regional marking. In any case, their use is so marginal, as we have seen with the first test, that, I would be very careful about any conclusion.

A picture containing screenshot, diagram, square

Description automatically generated

#### Code

ggplot(PerNomPlu.long\_filtered) +

aes(x = Area, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

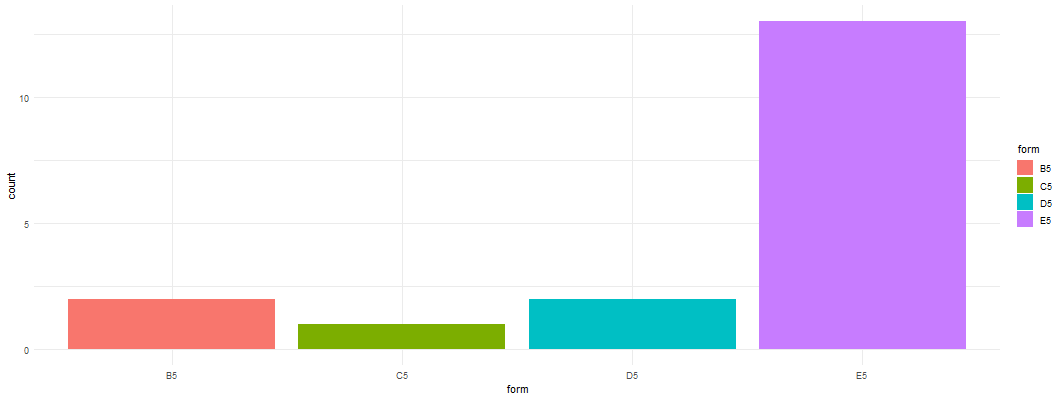
theme\_minimal()

## ACC.PL

This is the code for the dataframe that filters out the false presences.

PerAccPlu.long\_filtered <- PerAccPlu.long %>%

filter(pres == TRUE)

I have tested this against gender & age and area. But before engaging into more details, let us have a general overview of the distribution of the forms. E5 is definitely the winner, the rest are rather marginal.

#### Code

ggplot(PerAccPlu.long\_filtered) +

aes(x = form, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

### Gender & age

The dataset is very small, but is curious to see that, in this case, the use of different forms seems segregated by gender, or age. With the old data, we saw that younger speakers (who are less) seem to be very coherent with the forms, whereas the oldest speakers seem more open to variation. With this new arrangement, it looks completely different. There seems to be a preference based on gender for the non-E5 forms, but this also transcends age.

A screenshot of a computer game

Description automatically generated with low confidence

#### Code

ggplot(PerAccPlu.long\_filtered) +

aes(x = Gender.age, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

### Area

There seem to be regional preferences. Definitely D5 is a form marked in Podlasie, and C5 only appears in Torokan (West). However Tolkovo (West) and Bahdanawka (East) have the same forms, so the picture is a bit odd.

A picture containing screenshot, colorfulness, diagram

Description automatically generated

#### Code

ggplot(PerAccPlu.long\_filtered) +

aes(x = Area, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

## GEN.PL

This is the code for the dataframe that filters out the false presences.

PerGenPlu.long\_filtered <- PerGenPlu.long %>%

filter(pres == TRUE)

Given that there are many forms (8), I have decided to have a look at the distribution in the frequency of all of them, just to have a clearer picture before testing for the dependent variables. The clear winners are E7 and B7, respectively. The rest of the forms are infrequent.

A picture containing screenshot, colorfulness, square, line

Description automatically generated

#### Code

ggplot(PerGenPlu.long\_filtered) +

aes(x = form, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

labs(title = "GEN.PL distribution") +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L, face = "bold"),

axis.title.y = element\_text(size = 20L),

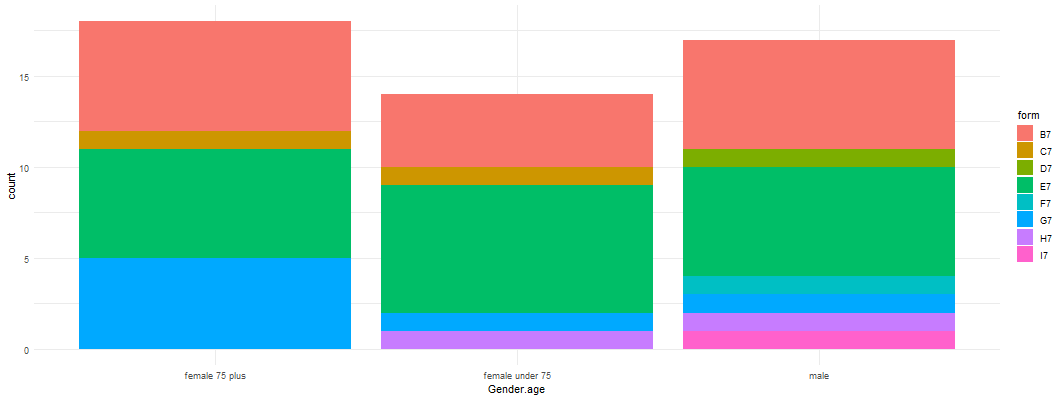
axis.title.x = element\_text(size = 20L),

legend.text = element\_text(size = 20L) # Adding this line for legend text size

)

### Gender & Age

It seems that male speakers are slightly more heterogenous; but also younger Female speakers. There is a relative homogeneous distribution of B7 and E7 across all. G7 seems preferred by older women and forms such as I7 or D7 have only been used by men, whilst C7 only by women.



#### Code

ggplot(PerGenPlu.long\_filtered) +

aes(x = Gender.age, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()

### Area

Podlasie comes up as the most diverse region, it seems that all forms are present in that variety, whereas in others the choice is more limited. Interestingly, B7 has also limited use in Podlasie.

A picture containing screenshot, colorfulness, diagram, square

Description automatically generated

#### Code

ggplot(PerGenPlu.long\_filtered) +

aes(x = Area, fill = form) +

geom\_bar() +

scale\_fill\_hue(direction = 1) +

theme\_minimal()