INITIAL OBSERVATIONS OF THE ‘person’ DATA

# Introductory remarks

I have started by gen pl, because it’s the most prominent form and where there is room for more variation. The rest will follow a more “traditional logic”.

Some remarks:

-Only one form was recorded for the acc sg, the ins sg and loc pl, and none for loc sg. Therefore, they have not been included in these observations.

-Unfortunately, there was no file for ins pl, so I haven’t analysed anything.

-Before running other tests, I wanted to check whether there is any ratio of “total forms” to “total occurrences”, in relation to the length of the recordings. That is to say, whether having longer recordings and the frequency of using this form would give more variation. I have run this test with all forms, with predictable results.

-I have primarily used the ggplot2 package in R and esquisse (as a user-friendly interface) for creating ggplots. I am attaching a full list of the packages used below.

-Given that in some datasets the names of the variables had slightly different names the codes for each form. In some cases, I have condensed the code a bit using […] for minor issues concerning font size or subtitles.

-I have also attached the images, which are available on the shared folder as well (>FaF/Stats/Person Plots).

## ###PACKAGES USED##

install.packages("tidyverse")

install.packages("ggplot2")

install.packages("esquisse")

install.packages("shiny")

install.packages("broom")

install.packages("cowplot")

library(“tidiverse”)

library("esquisse")

library("broom")

library("cowplot")

library("tidyverse")

library("ggplot2")

library("shiny"

# GENITIVE PL

First, I wanted to check whether there is any ratio of “total forms” to “total occurrences”, in relation to the length of the recordings. That is to say, whether having longer recordings and the frequency of using this form would give more variation, which provides predicable results.

Chart, scatter chart

Description automatically generated

There is no speaker who has spoken for more than an hour (length 4) who has not produced at least two different forms.

## ###CODE GEN PL Length-frequency-overabundance###

ggplot(PerGenPlu.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

theme\_minimal()

I wanted to see which of the three main variables “age”, “gender”, “village” would trigger more overabundance (i.e. a higher tot.form), using heat maps.

Chart, box and whisker chart

Description automatically generated

Bahdanawka has the most atypical distribution of all for age-overabundance, whilst for most villages, there seems to be a clearer correlation. The distinction between male-female doesn’t seem to be very relevant, and village wise; Pare is the “most boring” (but because we have very little data), whereas Tatar’ja and Žydče present more variation.

## ###CODE GEN PL Overabundance across villages, age and gender ###

ggplot(PerGenPlu.wide) +

aes(

x = Gender,

y = Age.group,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_viridis\_c(option = "viridis", direction = 1) +

scale\_color\_viridis\_c(option = "viridis", direction = 1) +

labs(

title = "Overabundance across villages,

age and gender",

subtitle = "GEN PL",

caption = "Overabundance across villages,

age and gender (GEN PL)"

) +

theme\_minimal() +

facet\_wrap(vars(Village))

# NOMINATIVE SG

The form does not appear on the corpus of speakers who spoke for less than 30 min, which suggests it may not be a frequent form. Interestingly, most of the variation happens on the corpus of people who used the nom sg less than 7 times.

Chart, scatter chart

Description automatically generated

## ###CODE NOM PL Length-frequency-overabundance###

ggplot(PerNomSin.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Lentgh-frequency-overabundance",

subtitle = "NOM SG",

caption = "NOM SG Lentgh-frequency-overabundance"

) +

theme\_minimal() +

theme(

axis.title.y = element\_text(face = "bold"),

axis.title.x = element\_text(face = "bold")

)

Looking at the variables we are mostly interested in, again Pare doesn’t tell us much because of the lack of data, but surprisingly, Podlasie (Poland) for which there is significantly more data, the results are pretty homogeneous. Tatar’ja and Žydče, once again show an interesting picture. This time it seems as if middle-aged people would be very conservative, whilst young and older speakers more innovative/heterogeneous.

Chart

Description automatically generated

## ###CODE NOM.SG \_ Overabundance across villages, age and gender###

ggplot(PerNomSin.wide) +

aes(

x = Gender,

y = `Age group`,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_viridis\_c(option = "viridis", direction = 1) +

scale\_color\_viridis\_c(option = "viridis", direction = 1) +

labs(

title = "Overabundance across villages,

age and gender", […]

) +

theme\_minimal() +

theme(

axis.title.y = element\_text(size = 12L,

face = "bold"),

axis.title.x = element\_text(size = 12L,

face = "bold")

) +

facet\_wrap(vars(Village))

# NOMINATIVE PL

As for the length-total occurrences & overabundance ratio, there is more or less a regular correlation between frequency and overabundance; however, there is a big gap and overabundance suddenly increases on super frequent forms, but not on the longest texts.

Chart, scatter chart

Description automatically generated

## ###CODE NOM PL Length-frequency-overabundance###

ggplot(PerNomPlu.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Length-frequency-overabundance",[…]

) +

theme\_minimal() +

theme(

axis.title.y = element\_text(size = 16L,

face = "bold"),

axis.title.x = element\_text(size = 16L,

face = "bold")

)

Concerning the village, age & gender variables, the data seem insufficient, as the max difference is two form per speaker. Podlasian men of all ages have two forms, but the distribution is less clear for other villages. Again, Tatar’ja has an odd discontinuity.

Timeline

Description automatically generated

## ###CODE NOM PL Overabundance across villages, age and gender###

ggplot(PerNomPlu.wide) +

aes(

x = Gender,

y = `Age group`,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_gradient() +

scale\_color\_gradient() +

labs(

title = "Overabundance across villages,

age and gender",

subtitle = "NOM PL",

caption = "NOM PL Overabundance across villages,

age and gender"

) +

theme\_minimal() +

theme(

axis.title.y = element\_text(size = 16L,

face = "bold"),

axis.title.x = element\_text(size = 16L,

face = "bold")

) +

facet\_wrap(vars(Village))

# ACCUSATIVE PL

There are no many occurrences of this form, most people have only used it 3 times or less, so there is not much to observe concerning the length-frequency and overabundance correlation. It only appears on the speeches of people who have spoken for a long time, which indicates it’s very infrequent.

Chart, scatter chart

Description automatically generated

## ###CODE ACC PL Lenthg-frequency-overabundance###

ggplot(PerAccPlu.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Lenthg-frequency-overabundance",

subtitle = "ACC PL",

caption = "ACC PL Lenthg-frequency-overabundance"

) +

theme\_minimal() +

theme […]

)

Concerning the variables we are most interested in, we have little data. The distinction of two forms only seems relevant in Bahdanawka and has no bearing at all in Podlasie, Žydče or Pare (for the former two, there are many speakers, though).

Chart

Description automatically generated with medium confidence

## ###CODE ACC PL Overabundance across villages, age and gender###

ggplot(PerAccPlu.wide) +

aes(

x = Gender,

y = Age.group,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_gradient() +

scale\_color\_gradient() +

labs(

title = "Overabundance across villages, age and gender", […]"

) +

theme\_minimal() +

theme() +

facet\_wrap(vars(Village))

# GENITIVE SG

The first test on length-frequency-overabundance shows us that there is nothing else to study here, as speakers stick to always the same form (in spite of having several options.

Chart, scatter chart

Description automatically generated

## ###CODE GEN SG length-frequency-overabundance###

ggplot(PerGenSin.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Length-frequency-overabundance",

subtitle = "GEN SG",

caption = "GEN SG Length-frequency-overabundance"

) +

theme\_minimal() +

theme(

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

plot.subtitle = element\_text(size = 14L),

plot.caption = element\_text(size = 11L),

axis.title.y = element\_text(size = 16L,

face = "bold"),

axis.title.x = element\_text(size = 16L,

face = "bold")

)

# ADNUMERATIVE

The adnumerative is also quite rare in the corpus. For most users it only has one form. However, I will proceed to the full analysis admitting the stats will be quite flawed (probably only based on one speaker using a second form once).

Chart, scatter chart

Description automatically generated

## ###CODE Length-frequency-overabundance###

ggplot(PerADMN.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Length-frequency-overabundance"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"),

plot.caption = element\_text(size = 13L),

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

face = "bold")

)

The second analysis doesn’t give any statistically significant results and reinforces my previous hypothesis (it’s probably only one speaker using a different form).

Timeline

Description automatically generated

## ###CODE ADNM Overabundance across villages, age and gender###

ggplot(PerADMN.wide) +

aes(

x = Gender,

y = `Age group`,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_gradient() +

scale\_color\_gradient() +

labs(

title = "Overabundance across villages,

age and gender",

subtitle = "ADNM",

caption = "ADNM Overabundance across villages, age and gender"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"), […]

plot.caption = element\_text(size = 13L),

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

face = "bold")

) +

facet\_wrap(vars(Village))

# DATIVE SG

After running the first test, we see that there is probably only one speaker (it refers to people who has spoken for more than an hour) who has used two different forms. Even other speakers who have spoken for an hour only have 2-3 tokens of the dat sg and they have been consistent with the form. Therefore, I don’t think data on the next step will provide any useful results.

Chart, scatter chart

Description automatically generated

## ###CODE DAT SG Length-frequency-overabundance###

ggplot(PerDatSin.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Length-frequency-overabundance",

subtitle = "DAT SG",

caption = "DAT SG Length-frequency-overabundance"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"),

plot.subtitle = element\_text(size = 14L,

face = "bold"),

plot.caption = element\_text(size = 13L),

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

face = "bold")

)

The results have been a bit different from what I expected. Still, there is a light correlation between speakers having more time and using more than one form, but this only applies to a handful of speakers, all in Belarus. It is impossible to associate it to a specific age group, because appears everywhere (though most prominently amongst the old speakers in Tatar’ja) and slightly more marked in women, in part because we have more women and on average, they spoke for much longer than men.

Timeline

Description automatically generated

## ###CODE DAT SG Overabundance across villages, age and gender###

ggplot(PerDatSin.wide) +

aes(

x = Gender,

y = `Age group`,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_gradient() +

scale\_color\_gradient() +

labs(

title = "Overabundance across villages,

age and gender",

subtitle = "DAT SG",

caption = "DAT SG Overabundance across villages,

age and gender"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"),

plot.subtitle = element\_text(size = 14L,

face = "bold"),

plot.caption = element\_text(size = 13L),

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

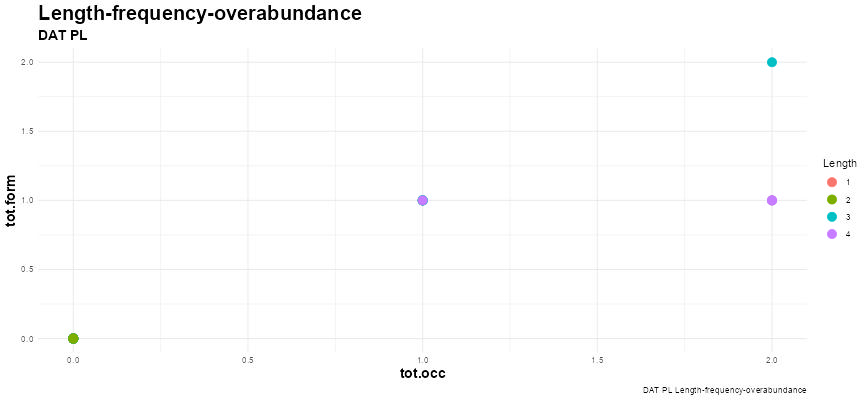
face = "bold")

) +

facet\_wrap(vars(Village))

# DATIVE PL

The variation between forms only seems to happen in speakers with a very large input. And even though, nobody has used the dat pl more than three times in the corpus, therefore, the results are very limited.



## ###CODE DAT PL Length-frequency-overabundance###

ggplot(PerDatPlu.wide) +

aes(

x = tot.occ,

y = tot.form,

fill = Length,

colour = Length

) +

geom\_point(shape = "circle", size = 4L) +

scale\_fill\_hue(direction = 1) +

scale\_color\_hue(direction = 1) +

labs(

title = "Length-frequency-overabundance",

subtitle = "DAT PL",

caption = "DAT PL Length-frequency-overabundance"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"),[…]

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

face = "bold")

)

Despite predicting very bad results, I have decided to take a second test of the distribution of these very rare forms across speakers of different ages, villages and genders, where it seems the variation between forms is only an issue in Belarusian villages, especially in Žydče (East) and Tatar'ja (West) even though this does not reflect in other neighbouring varieties geographically (and structurally) much closer to each other. In the case of Pare (East), it may be also related to the fact that the corpus is very small, conversely not for Bahdanawka (also East).

Timeline

Description automatically generated with low confidence

## ###CODE DAT PL Overabundance across villages,age and gender###

ggplot(PerDatPlu.wide) +

aes(

x = Gender,

y = `Age group`,

fill = tot.form,

colour = tot.form

) +

geom\_tile() +

scale\_fill\_gradient() +

scale\_color\_gradient() +

labs(

title = "Overabundance across villages,

age and gender",

subtitle = "DAT PL",

caption = "DAT PL Overabundance across villages,

age and gender"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 20L,

face = "bold"),

plot.subtitle = element\_text(size = 14L,

face = "bold"),

plot.caption = element\_text(size = 13L),

axis.title.y = element\_text(size = 14L,

face = "bold"),

axis.title.x = element\_text(size = 14L,

face = "bold")

) +

facet\_wrap(vars(Village))

# Summary

Not surprisingly some forms are more frequent than others and we can only really extract meaningful results from the very frequent ones. The division between gender does not seem as clear cut as initially expected and age, which only seems relevant for some parameters, is very strongly attached to the former. Village ~location doesn’t seem to be as relevant (East-West) as initially thought concerning variation (of course, some variants only appear in some areas, for which Podlasie/Poland stands out from the rest). Pare has a very small corpus, and as previously said, many of the variation issues seem to be related to frequency; i.e. the longer the speakers have spoken for, the higher the chances of using multiple forms for the same slot.