A quantitative analysis of the variational pattern of French loans in Luxembourgish showcasing R (and Quarto)

Peter Gilles

2026-03-02

Table of contents

The research question

Drawing on data from our Schnëssen project, we seek to understand the driving factors behind the choice of a variant of Germanic origin (Luxembourgish or German) or French origin.

Example:

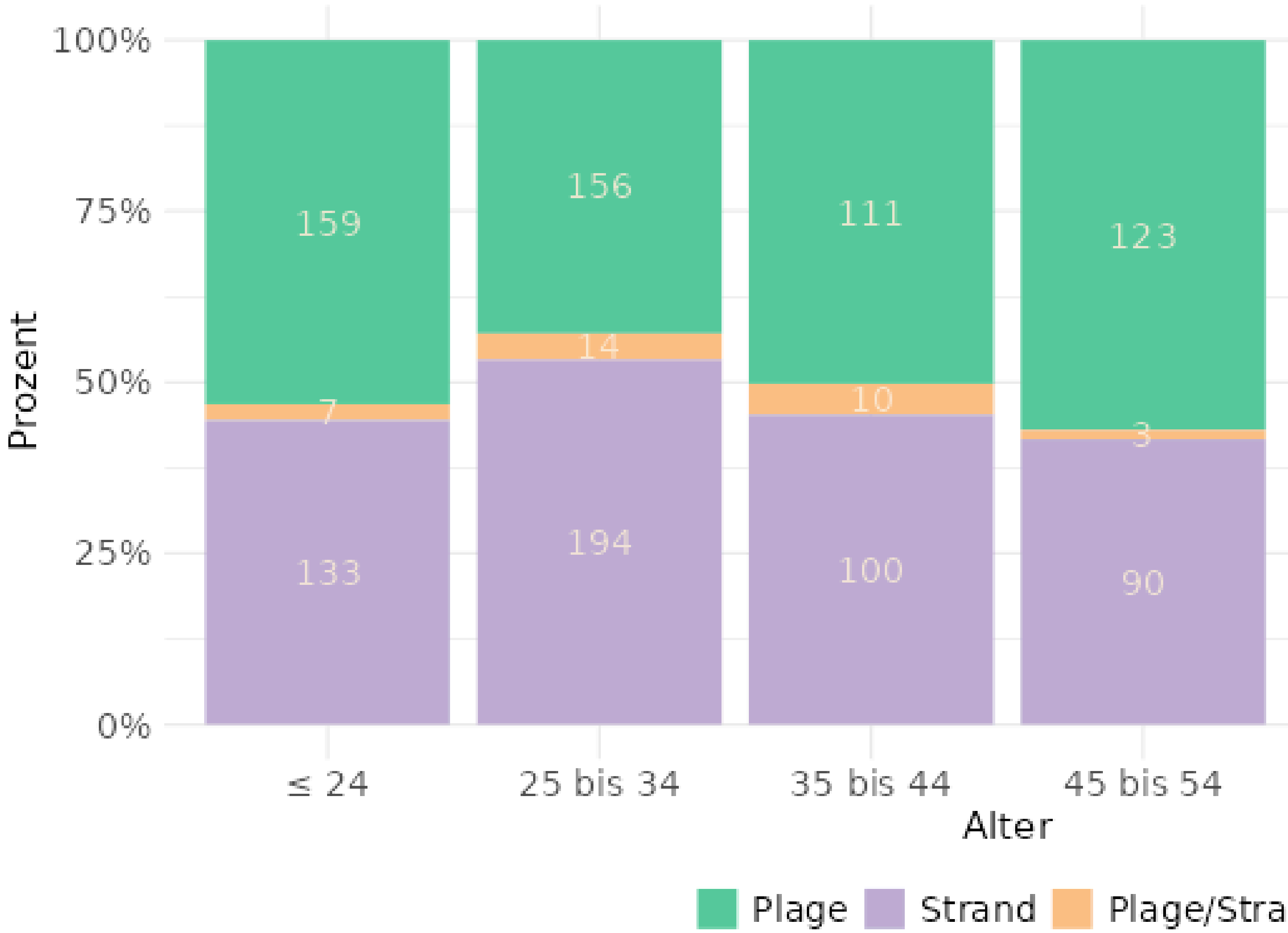
- Germanic origin: *Stréihallem, Strand, Fernbedienung* vs. French origin: *Schallimo, Plage, Telecommande*

This linguistic choice as the **dependent variable** is influenced by social or other linguistic variables as **independent variables:**

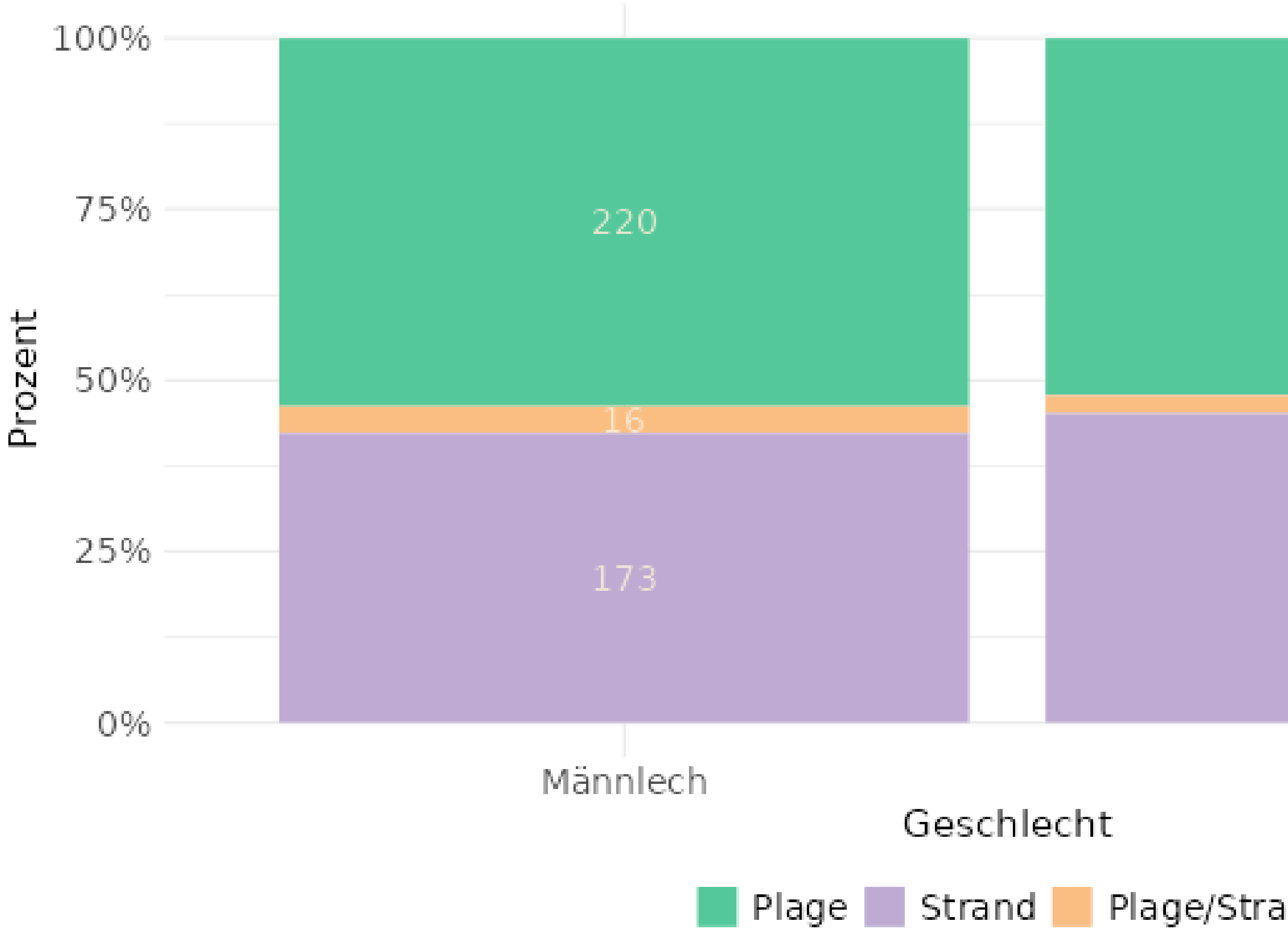
- age

- gender
- education
- language competencies (French, German)

Age

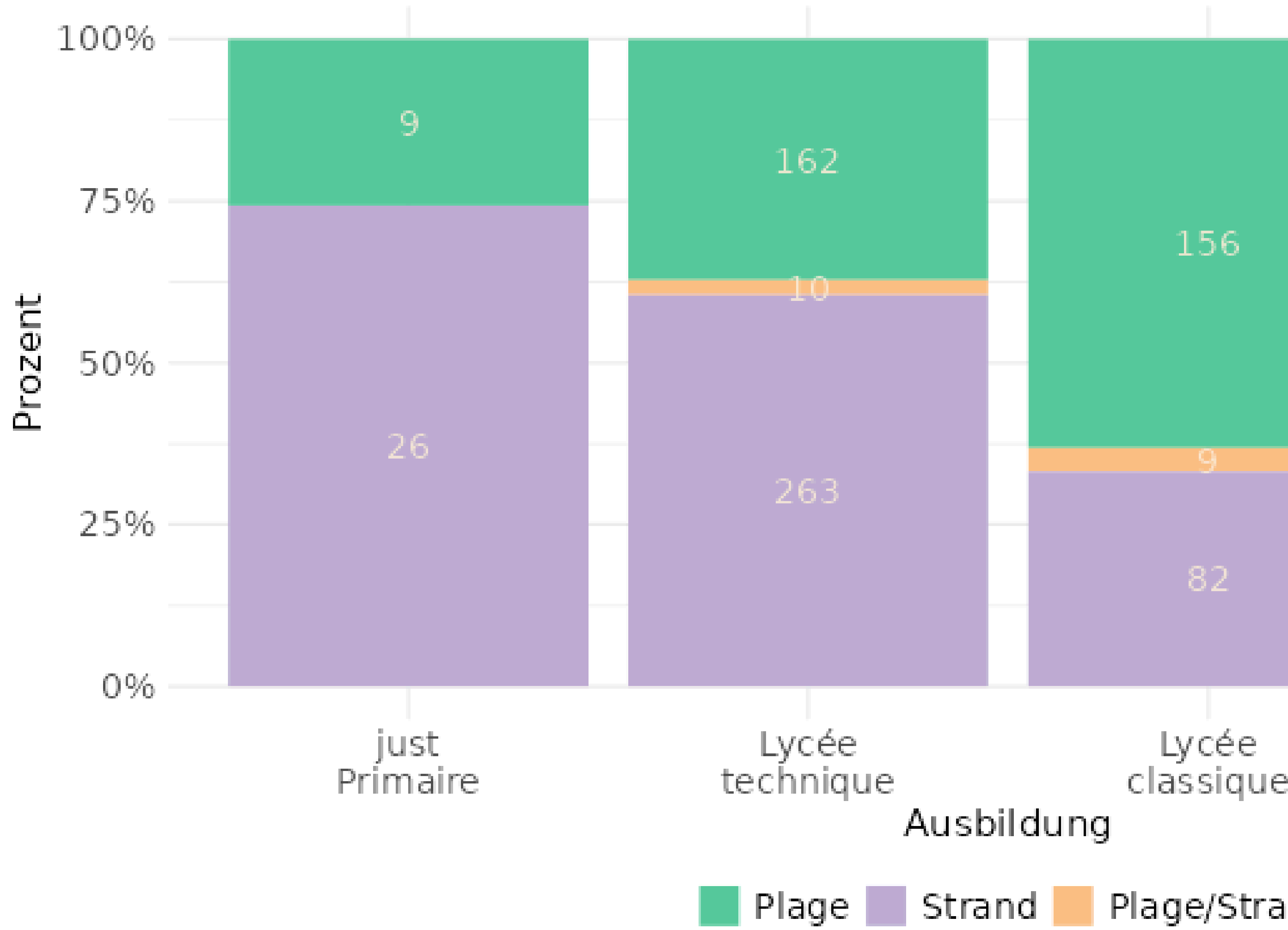


Gender



100%
75%
50%
25%
0%
Prozent
220
16
173
Männlech
Geschlecht
Plage
Strand
Plage/Stra

Education



Corollary aim of this presentation:

- showcase, how these questions will be adressed in a coherent and systematic way using R - for data wrangling and statistics -, Quarto - for writing-up and layout - and GitHub - for publishing and dissemination.

References/Tutorials

Statistics for Language Variation & Change with R

- LADAL - Language Technology and Data Analysis Laboratory
- https://lingmethodshub.github.io/
- [1]

Quarto

- https://quarto.org/

The data

Collected with Schnëssen app

- audio data for +800 linguistic variables inserted in translation tasks from German or French, image descriptions etc.
- per variable: 300 up to 1500 responses
- coded for variable, variant, social data of respondent

Sub-set for this study extracted as data frame/tibble in R.

```r
# load and display the dataset
input_data <- readRDS("input_data.rds")
```

The dataset has 38452 rows.

Give an overview as table:

```r
DT::datatable(input_data)
```

```
Warning in instance$preRenderHook(instance): It seems your data is too big for
client-side DataTables. You may consider server-side processing:
https://rstudio.github.io/DT/server.html
```

id	variant	variable	French_origin	domain	urbanity	socio_index

The tibble `input_data` has the following structure.

```
str(input_data)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   38452 obs. of  16 variables:
 $ id                 : Factor w/ 3151 levels "1000","1004",..: 393 452 508 555
726 783 811 821 828 849 ...
 $ variant           : Factor w/ 116 levels "Zwiwwel","Ënn",..: 1 1 2 2 1 2 2 2
2 2 ...
 $ variable          : Factor w/ 50 levels "Variant_Ënn",..: 1 1 1 1 1 1 1 1 1
1 ...
 $ French_origin    : chr  "non-French" "non-French" "French" "French" ...
 $ domain           : chr  "lexicon" "lexicon" "lexicon" "lexicon" ...
 $ urbanity         : Factor w/ 3 levels "smaller towns",..: 1 1 2 1 1 1 1 2 2
1 ...
 $ socio_index      : Factor w/ 15 levels "(0.122,0.278]",..: 5 1 5 5 4 2 2 2 5
1 ...
 $ socio_index_raw  : num  0.89 0.243 0.821 0.856 0.594 ...
 $ region           : Factor w/ 5 levels "Norden","Osten",..: 3 4 2 4 3 4 3 4 2
```

```
4 ...
 $ first_language   : Factor w/ 1 level "Jo": 1 1 1 1 1 1 1 1 1 1 ...
 $ age6             : Ord.factor w/ 6 levels "≤ 24"<"25 bis 34"<..: 2 2 2 2 2 2
3 2 5 2 ...
 $ age              : Factor w/ 3 levels "young","middle-aged",..: 1 1 1 1 1 1
2 1 3 1 ...
 $ gender           : Factor w/ 2 levels "male","female": 2 2 1 2 2 2 2 2 2 2 ...
 $ education        : Ord.factor w/ 4 levels "other"<"Technical\nschool"<..: 4
4 4 4 4 4 4 4 3 4 ...
 $ competence_french: Factor w/ 3 levels "French\nlow",..: 3 2 2 2 2 3 3 2 3
3 ...
 $ competence_german: Factor w/ 3 levels "German low","German average",..: 3 3
3 3 3 3 3 3 3 3 ...
```

```
summary(input_data)
```

```
       id                   variant                          variable
 1826   :   48   Kannapee       : 1188   Variant_Déierendoktesch: 1704
 6398   :   48   Toilettëpabeier: 1126   Variant_Kannapee       : 1626
 6410   :   48   Ënn            : 1120   Variant_Ënn            : 1589
 1700   :   47   Kamäin         : 1087   Variant_Schaarschtech  : 1378
 3818   :   47   Déierendokter  : 1045   Variant_Schallimo      : 1278
 203    :   46   Kürbis         : 1034   Variant_Kürbis         : 1224
 (Other):38168   (Other)        :31852   (Other)                :29653
 French_origin         domain                        urbanity
 Length:38452       Length:38452       smaller towns     :17007
 Class :character   Class :character   rural areas       :16651
 Mode  :character   Mode  :character   1\nStad Lëtzebuerg: 4794




        socio_index    socio_index_raw      region      first_language
 (0.278,0.434]:11792   Min.   :0.1230   Norden : 4060   Jo:38452
 (0.589,0.744]: 7396   1st Qu.:0.3434   Osten  : 6874
 (0.434,0.589]: 7380   Median :0.4746   Süden  :10643
 (0.122,0.278]: 5969   Mean   :0.4998   Zentrum:16875
 (0.744,0.9]  : 5619   3rd Qu.:0.6373   #ERROR!:    0
 (0.313,0.46] :   52   Max.   :0.8997
 (Other)      :  244
        age6                  age           gender
 ≤ 24     : 7451   young      :17977   male  :11634
 25 bis 34:10526   middle-aged:13313   female:26818
 35 bis 44: 6799   old        : 7162
 45 bis 54: 6514
 55 bis 64: 5106
```

```
65+       : 2056

                       education          competence_french
other                       :    0   French\nlow     : 2675
Technical\nschool           :13431   French\naverage:18193
Classical\nsecondary\nschool: 6881   French\nhigh    :17584
University                  :18140



     competence_german
German low    :  119
German average: 4377
German high   :33956
```

What's inside the tibble?

Using crosstables - Responses by age

```r
table(input_data$age)
```

```

     young middle-aged         old
     17977       13313        7162
```

- Responses by age and French_origin

```r
table(input_data$age, input_data$French_origin)
```

```

              French non-French
 young          6768      11209
 middle-aged    5229       8084
 old            3037       4125
```

- Responses by competence_french and French_origin

```r
table(input_data$competence_french, input_data$French_origin)
```

```

                  French non-French
```

```
  French\nlow        879       1796
  French\naverage   6776      11417
  French\nhigh      7379      10205
```

- Can we have this in percentages?

```
prop.table(table(input_data$competence_french, input_data$French_origin))
```

```


                    French non-French
  French\nlow     0.02285967 0.04670758
  French\naverage 0.17621970 0.29691564
  French\nhigh    0.19190159 0.26539582
```

- Mean number of responses by id

```
mean(table(input_data$id))
```

```
[1] 12.20311
```

- How many different `ids` (= speakers) are in the dataset?

```
length(unique(input_data$id))
```

```
[1] 3151
```

The analysis

1st analysis: regression analysis

Hypothesis: The choice of a French variant (dependent variable/response variable) is influenced by social factors (independent variables/predictors).

Choice of regression analysis dependent on the nature of the dependent variable:

- binary: logistic regression
- count: Poisson regression
- continuous: linear regression

Our dependent variable *French_origin* is binary, thus we will use logistic regression.

Logistic regression

Prepare data

```

```

```r
# see: https://slcladal.github.io/regression.html#Random_Effects
library(tidyverse)
```

```
Warning: Paket 'tidyr' wurde unter R Version 4.3.2 erstellt
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
library(lme4)
```

```
Lade nötiges Paket: Matrix

Attache Paket: 'Matrix'

Die folgenden Objekte sind maskiert von 'package:tidyr':

    expand, pack, unpack
```

```r
library(sjPlot)

data <- input_data %>%
  # filter for domain (lexicon or phonology)
  filter(domain == "lexicon") %>%
  # convert binary variable to values 0 and 1
  mutate(across(French_origin, str_replace, "non-French", "0")) %>%
  mutate(across(French_origin, str_replace, "French", "1")) %>%
  mutate(French_origin = as.integer(French_origin)) %>%
  mutate(age = factor(age, ordered = FALSE)) %>%
  mutate(age6 = factor(age6, ordered = TRUE)) %>%
  mutate(education = factor(education, ordered = FALSE)) %>%
  mutate(competence_german = factor(competence_german, ordered = FALSE)) %>%
  mutate(competence_french = factor(competence_french, ordered = FALSE)) %>%
  mutate(urbanity = factor(urbanity, ordered = FALSE))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `across(French_origin, str_replace, "non-French", "0")`.
Caused by warning:
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
Supply arguments directly to `.fns` through an anonymous function instead.

  # Previously
  across(a:b, mean, na.rm = TRUE)

  # Now
  across(a:b, \(x) mean(x, na.rm = TRUE))
```

The regression model will be fit step by step ('stepwise regression') by starting out with a base model without predictors and then adding one predictor after the other. After each step the model will be compared statistically with the previous one. If better, the predictor is retained, if the model is worse or the same, the predictor is eliminated. Step by step then the explanatory predictors are detected.

Logistic Regression

Fit base model. Use *id* and *variable* as random effects.

```
m0.lmer <- lmer(formula=French_origin ~ 1 + (1|id) + (1|variable), REML = T,
data = data)
```

Add age as first predictor.

```
m1.lmer <- lmer(formula=French_origin ~ age + (1|id) + (1|variable), REML = T,
data = data)
tab_model(m1.lmer)
```

Compare which model is performing better.

```
anova(m1.lmer, m0.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m0.lmer: French_origin ~ 1 + (1 | id) + (1 | variable)
m1.lmer: French_origin ~ age + (1 | id) + (1 | variable)
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m0.lmer    4 38443 38477 -19218    38435
m1.lmer    6 38411 38462 -19199    38399 36.505  2  1.183e-08 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Age is a significant contribution to the model.   The *age* model is significantly better than the previous one.

Adding further predictors, starting with *gender*. Instead of refitting the model, we can update the model.

```r
m2.lmer <- update(m1.lmer, .~.+ gender)
tab_model(m2.lmer)
```

Compare again, which one is better.

```r
anova(m2.lmer, m1.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m1.lmer: French_origin ~ age + (1 | id) + (1 | variable)
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + gender
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m1.lmer    6 38411 38462 -19199    38399
m2.lmer    7 38411 38471 -19199    38397 1.1189  1     0.2902
```

Gender is not significant. Remove it? Let's try for the interaction with age.

```r
m2.lmer <- update(m1.lmer, .~.+ gender*age)
tab_model(m2.lmer)
```

```r
anova(m2.lmer, m1.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m1.lmer: French_origin ~ age + (1 | id) + (1 | variable)
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + gender + age:gender
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
```

```
m1.lmer    6 38411 38462 -19199    38399
m2.lmer    9 38415 38492 -19199    38397 1.4425  3     0.6956
```

None of interactions is significant and the model is not performing better than the previous one. Gender will be removed . Interestingly, and contrary to general sociolinguistic assumptions, gender seems to play no role in the choice of a French variant.

Add predictor *competence_french*.

```
m2.lmer <- update(m1.lmer, .~.+ competence_french)
tab_model(m2.lmer)
```

```
anova(m2.lmer, m1.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m1.lmer: French_origin ~ age + (1 | id) + (1 | variable)
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m1.lmer    6 38411 38462 -19199    38399
m2.lmer    8 38324 38393 -19154    38308 90.154  2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, this is one is performing well.

Maybe *competence_french* is interacting with *age*?

```
m3.lmer <- update(m1.lmer, .~.+ competence_french*age)
tab_model(m3.lmer)
```

```
anova(m3.lmer, m2.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
```

```
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french
m3.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
age:competence_french
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m2.lmer    8 38324 38393 -19154    38308
m3.lmer   12 38330 38432 -19153    38306 2.7069  4      0.608
```

No interaction! Removed from the model.

Add predictor *competence_german*.

```
m3.lmer <- update(m2.lmer, .~.+ competence_german)
tab_model(m3.lmer)
```

```
anova(m3.lmer, m2.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french
m3.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
competence_german
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m2.lmer    8 38324 38393 -19154    38308
m3.lmer   10 38328 38414 -19154    38308 0.2494  2     0.8828
```

Not significant! Removed from the model.

Add predictor *education*.

```
m3.lmer <- update(m2.lmer, .~.+ education)
tab_model(m3.lmer)
```

```
anova(m3.lmer, m2.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
```

```
m2.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french
m3.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
education
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m2.lmer    8 38324 38393 -19154    38308
m3.lmer   10 38297 38382 -19138    38277 31.709  2  1.302e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Education is a significant contribution to the model. Speakers coming from a classical school or university, use significantly more French variants.

Now adding predictors relating to the location of the speaker, starting with the degree of *urbanity*.

```r
m4.lmer <- update(m3.lmer, .~.+ urbanity)
tab_model(m4.lmer)
```

```r
anova(m4.lmer, m3.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m3.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
education
m4.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
education + urbanity
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m3.lmer   10 38297 38382 -19138    38277
m4.lmer   12 38240 38343 -19108    38216 60.343  2  7.881e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Urbanity is a ignificant contribution to the model! Check the algebraic sign for the estimates: In rural areas significant less French variants are used. In the capital significant more French variants are used.

Add predictor *socio-economic index*; ranges from 0 to 1 and is based on the share of single parents, mean salary, share of persons with RMG, level of unemployment *per commune*. 0 = favorable commune, 1 = defavorable commune (see STATEC).

```
m5.lmer <- update(m4.lmer, .~.+ `socio_index_raw`)
tab_model(m5.lmer)
```

```
anova(m5.lmer, m4.lmer, test = "Chi")
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
m4.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
education + urbanity
m5.lmer: French_origin ~ age + (1 | id) + (1 | variable) + competence_french +
education + urbanity + socio_index_raw
        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
m4.lmer   12 38240 38343 -19108    38216
m5.lmer   13 38206 38317 -19090    38180 36.432  1  1.581e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *socio-economic index* is a significant contribution to the model. The higher the index, i.e. the less favorable the commune, the less French variants are used.

This is our final model. Speakers **favoring French variants** show the following social characteristics:

- rather old
- average to high competence in French
- education in classical school or university
- rather living in the capital, then in rural areas
- low socio-economic index

2nd analysis: correspondence regression analysis

See; [2]

Run a regression with two predictors for all **variants, instead for the variables.** Group similar variants together in a two-dimensional space.

Run the correspondence regression analysis using the R package *corregp*. We are using two predictors, *competence_french* and *age* against the response variable *variant*.

```
library(corregp)
```

```
Lade nötiges Paket: diagram
```

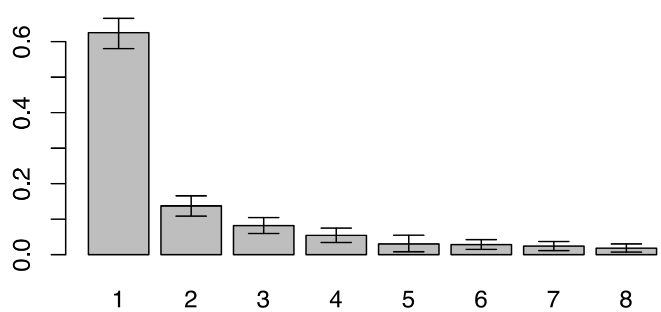
```
Lade nötiges Paket: shape
```

```
Lade nötiges Paket: rgl
```

```
corr.crg <- corregp(variant ~ competence_french * age, data=input_data,
                    part="variable", b=3000)
```

A Screeplot shows the amount of variation explained per so-called 'latent variable'.

```
screeplot(corr.crg, add_ci=TRUE, type="%")
```



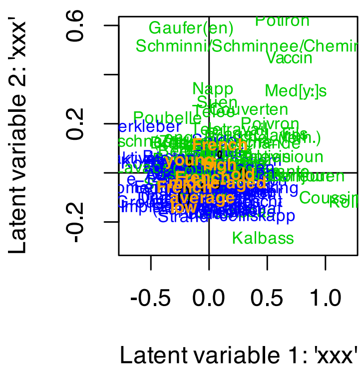
In this case, 65 % of variation is explained by the first 'latent variable' and 13 % by the second latent variable. A considerable amount of variation is thus explained through these two predictors, *competence_french* and *age*.

Now plot this in a two-dimensional space. Use green for variants with French origin and blue for variants with Germanic origin. In addition, plot also the values for *age* (young - middle-aged - old) and *competence_french* (low - average - high).

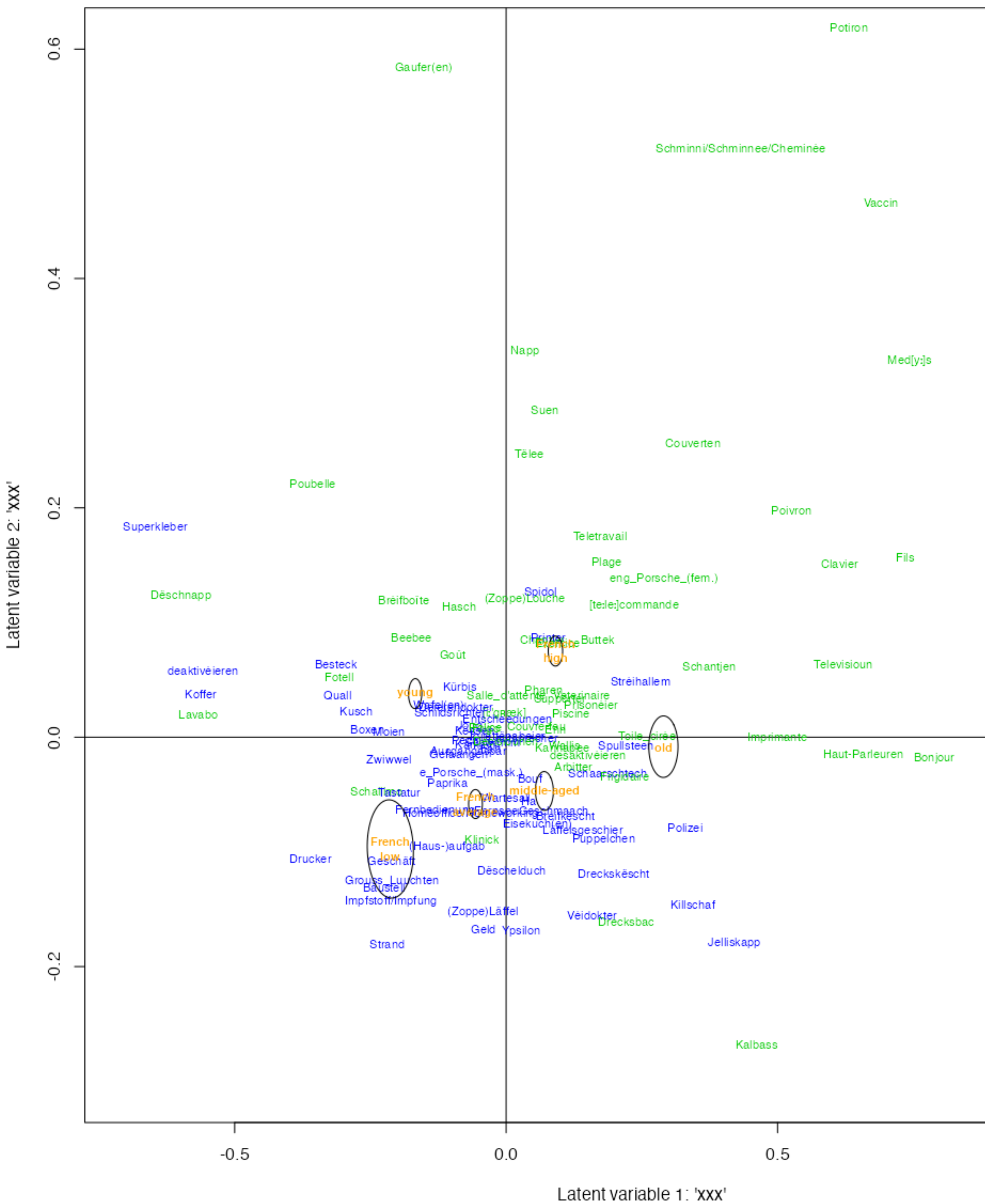
```
# Colors for plotting:
corr.col <- ifelse(xtabs(~variant + French_origin, data=input_data)[,"French"]
> 0,
                       "green3", "blue")

plot(corr.crg,      x_ell=TRUE,      xsub=c("competence_french",      "age"),
col_btm=corr.col, col_top="orange",
```

```
        cex_btm=0.7,  cex_top=0.7,  font_btm=  1,  font_top=2,  hlim=c(-0.7,1.2),
vlim=c(-0.3,0.6),
       hlab  =  "Latent  variable  1:  'xxx'",  vlab  =  "Latent  variable  2:  'xxx'",
add_ori=TRUE)
```



Larger image of correspondence analysis



Latent variable 2: 'xxx'
Latent variable 1: 'xxx'
0.6
0.4
0.2
0.0
-0.2
-0.5
0.0
0.5
Potiron
Gaufer(en)
Schminni/Schminnee/Cheminée
Vaccin
Napp
Med[y:]s
Suen
Couverten
Tëlee
Poubelle
Poivron
Superkleber
Teletravail
Plage
Clavier
Fils
eng_Porsche_(fem.)
Dëschnapp
Breifboîte
Hasch
(Zoppe)Louche
Spidol
[te:le:]commande
Beebee
Buttek
Goût
high
Schantjen
Televisioun
Besteck
deaktivéieren
Fotell
young
Kürbis
Streihallem
Koffer
Quall
Kusch
Lavabo
Piscine
Boxen
Moien
Spullsteen
old
Imprimante
Haut-Parleuren
Bonjour
Zwiwwel
desaktivéieren
Arbitter
e_Porsche_(mask.)
Paprika
middle-aged
Tastatur
French
low
(Haus-)aufgab
Klinick
Drucker
Polizei
Puppelchen
Dëschelduch
Dreckskëscht
Grouss_Luuchten
Baustell
Impfstoff/Impfung
Killschaf
(Zoppe)Läffel
Véidokter
Drecksbac
Geld
Ypsilon
Jelliskapp
Strand
Kalbass

French variants tend to concentrate in the NE quadrant, and Germanic variants more in the lower half. Interestingly, green and blue form - cum grano salis - two clouds. Distances become visible, e.g. *Buttek* associated with 'French high' and *Geschäft* with 'French low'.

The end

- R will help you to understand your data and find the best statistical analysis
- Quarto will help you to write an analysis, a report or even an entire fully-fledged article in a journal layout - all in one place.

References

Bibliography

[1] B. Winter, *Statistics for linguists: an introduction using R*. Routledge, 2020. [Online]. Available: https://www.taylorfrancis.com/books/9781315165547

[2] P. Gilles, "Regional variation, internal change and language contact in Luxembourgish: results from an app-based language survey1", *Taal en Tongval*, vol. 75, no. 1, p. 29, 2023, doi: 10.5117/TET2023.1.003.GILL.