New technologies in language studies

Livia Oushiro
(University of Campinas)

oushiro@iel.unicamp.br
https://oushiro.github.io/

Stockholm, February 20-21, 2018

Objectives

- Learn how to make transcriptions in ELAN (Day1, morning)
- Discuss basic concepts of R and Statistics
- Take the fist steps in a statistical analysis
  - Graphic explorations
  - Raising hypotheses
  - Testing hypotheses
  - Interpretation and follow up analyses

→ To learn how to use any new software, you need to practice!

Why use new technologies?

To optimize your work and spend more time on what really matters: your readings and your analyses!

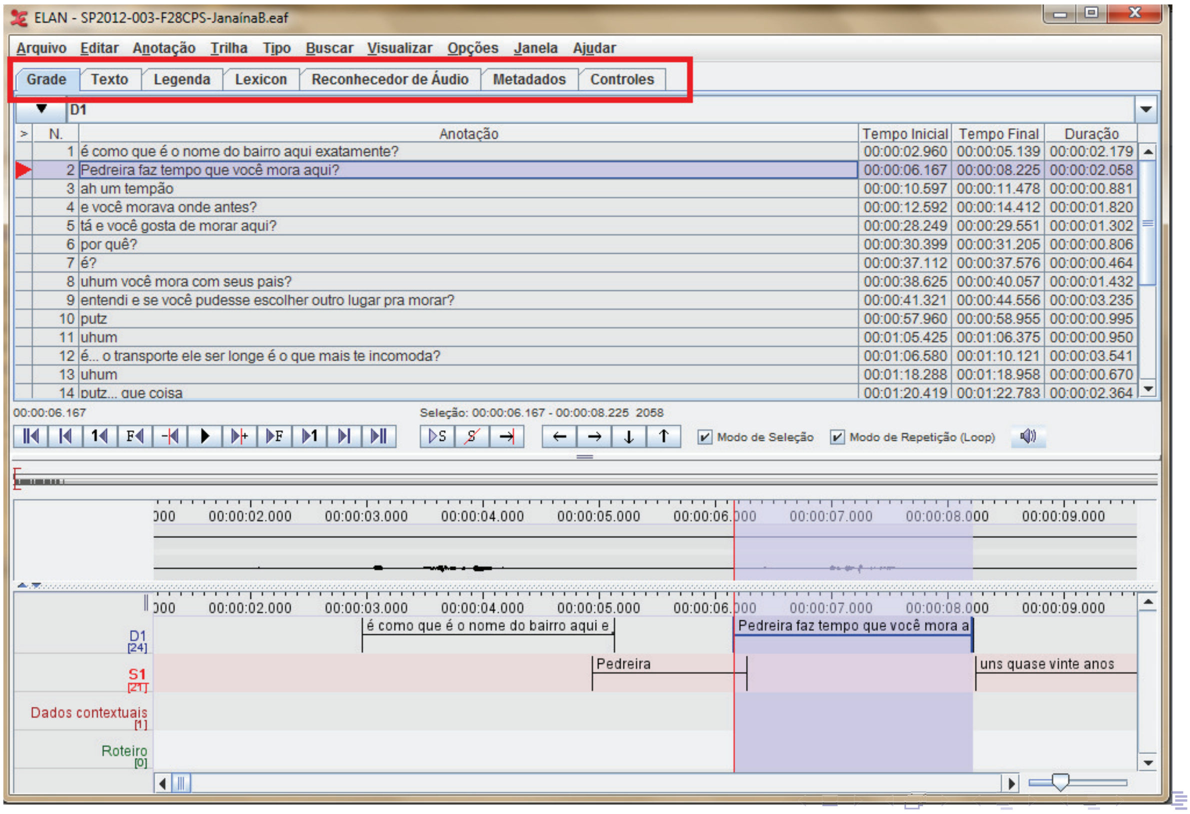
Advantages

Rosenfelder (2011)

- audio and video file transcriptions
- synchronization between media file and transcript/annotation
- possibility to create multiple tiers
- automatic searches in a corpus through regular expressions
- different formats for exporting transcript (.txt, .TextGrid, etc.)
- free and made for linguists!

ELAN's main functions

Tabs


ELAN - SP2012-003-F28CPS-JanainaB.eaf
Arquivo Editar Anotação Trilha Tipo Buscar Visualizar Opções Janela Ajudar
Grade Texto Legenda Lexicon Reconhecedor de Áudio Metadados Controles
D1
Modo de Seleção
Modo de Repetição (Loop)
D1 [24]
S1 [21]
Dados contextuais [1]
Roteiro [0]

N.	Anotação	Tempo Inicial	Tempo Final	Duração
1	é como que é o nome do bairro aqui exatamente?	00:00:02.960	00:00:05.139	00:00:02.179
2	Pedreira faz tempo que você mora aqui?	00:00:06.167	00:00:08.225	00:00:02.058
3	ah um tempão	00:00:10.597	00:00:11.478	00:00:00.881
4	e você morava onde antes?	00:00:12.592	00:00:14.412	00:00:01.820
5	tá e você gosta de morar aqui?	00:00:28.249	00:00:29.551	00:00:01.302
6	por quê?	00:00:30.399	00:00:31.205	00:00:00.806
7	é?	00:00:37.112	00:00:37.576	00:00:00.464
8	uhum você mora com seus pais?	00:00:38.625	00:00:40.057	00:00:01.432
9	entendi e se você pudesse escolher outro lugar pra morar?	00:00:41.321	00:00:44.556	00:00:03.235
10	putz	00:00:57.960	00:00:58.955	00:00:00.995
11	uhum	00:01:05.425	00:01:06.375	00:00:00.950
12	é... o transporte ele ser longe é o que mais te incomoda?	00:01:06.580	00:01:10.121	00:00:03.541
13	uhum	00:01:18.288	00:01:18.958	00:00:00.670
14	putz... que coisa	00:01:20.419	00:01:22.783	00:00:02.364

00:00:06.167

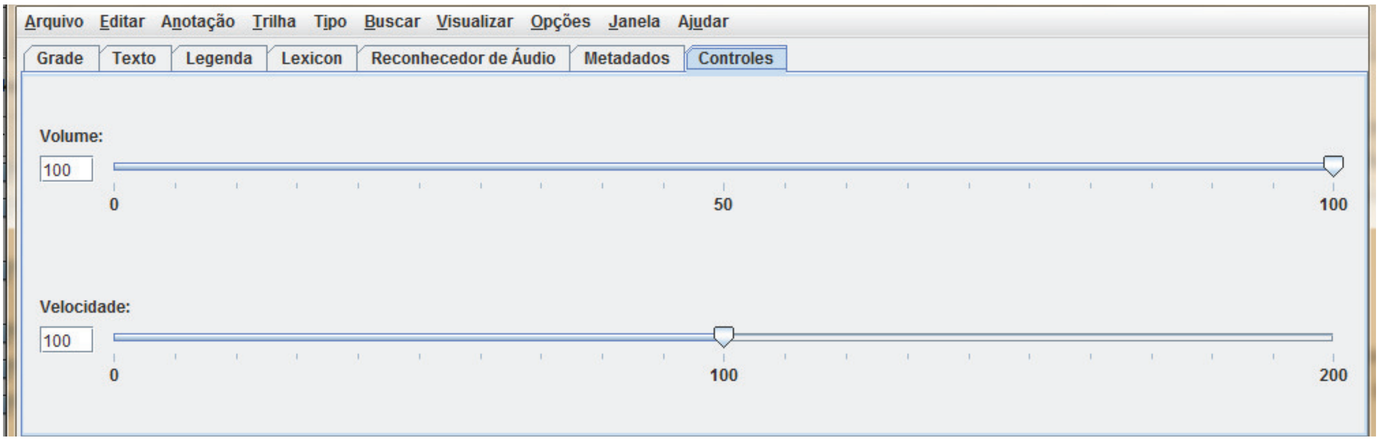
Seleção: 00:00:06.167 - 00:00:08.225 2058

D1: é como que é o nome do bairro aqui e | Pedreira faz tempo que você mora a

S1: Pedreira | uns quase vinte anos

ELAN's main functions

Controls

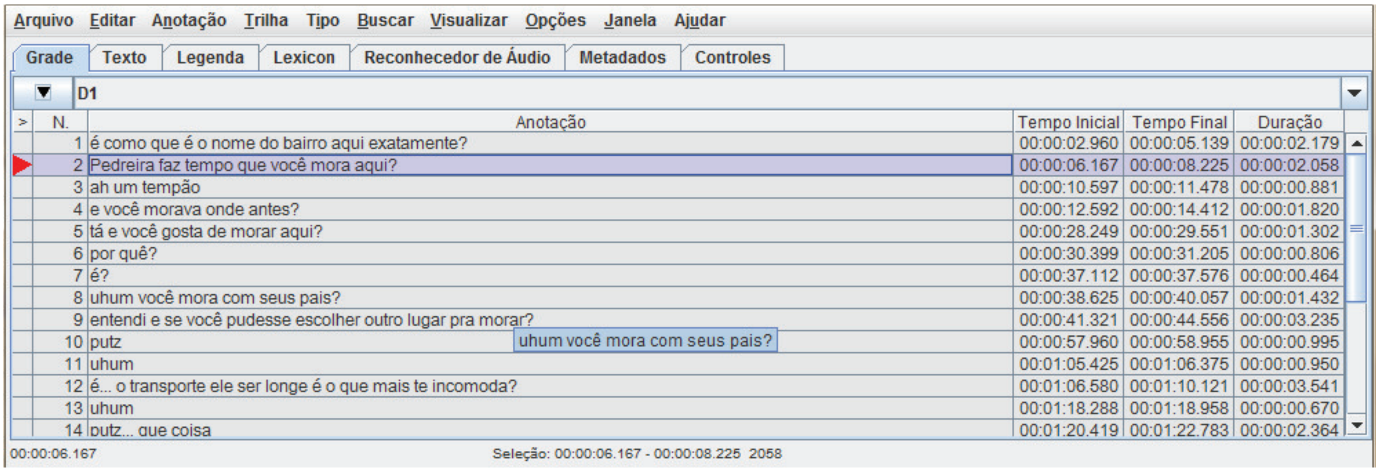

Arquivo Editar Anotação Trilha Tipo Buscar Visualizar Opções Janela Ajudar
Grade Texto Legenda Lexicon Reconhecedor de Áudio Metadados Controles
0
50
100
0
100
200

Volume: 100

Velocidade: 100

ELAN's main functions

Grid



>	N.	Anotação	Tempo Inicial	Tempo Final	Duração
	1	é como que é o nome do bairro aqui exatamente?	00:00:02.960	00:00:05.139	00:00:02.179
▶	2	Pedreira faz tempo que você mora aqui?	00:00:06.167	00:00:08.225	00:00:02.058
	3	ah um tempão	00:00:10.597	00:00:11.478	00:00:00.881
	4	e você morava onde antes?	00:00:12.592	00:00:14.412	00:00:01.820
	5	tá e você gosta de morar aqui?	00:00:28.249	00:00:29.551	00:00:01.302
	6	por quê?	00:00:30.399	00:00:31.205	00:00:00.806
	7	é?	00:00:37.112	00:00:37.576	00:00:00.464
	8	uhum você mora com seus pais?	00:00:38.625	00:00:40.057	00:00:01.432
	9	entendi e se você pudesse escolher outro lugar pra morar?	00:00:41.321	00:00:44.556	00:00:03.235
	10	putz	00:00:57.960	00:00:58.955	00:00:00.995
	11	uhum	00:01:05.425	00:01:06.375	00:00:00.950
	12	é... o transporte ele ser longe é o que mais te incomoda?	00:01:06.580	00:01:10.121	00:00:03.541
	13	uhum	00:01:18.288	00:01:18.958	00:00:00.670
	14	putz... que coisa	00:01:20.419	00:01:22.783	00:00:02.364

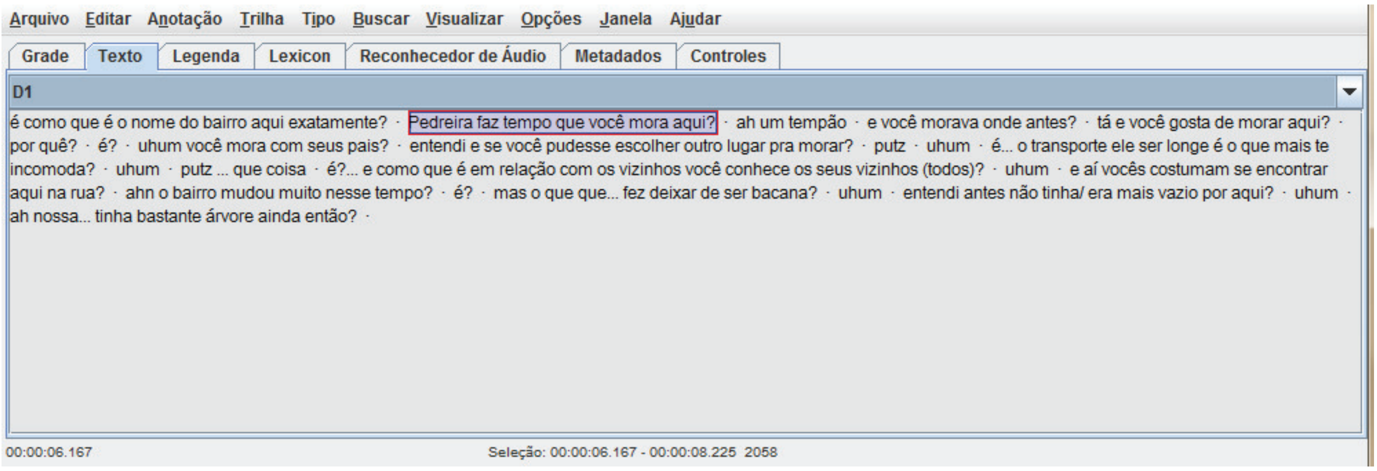
uhum você mora com seus pais?

00:00:06.167

Seleção: 00:00:06.167 - 00:00:08.225 2058

ELAN's main functions

Text



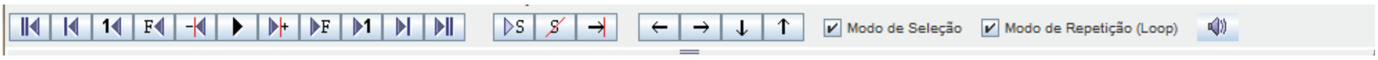
é como que é o nome do bairro aqui exatamente? · Pedreira faz tempo que você mora aqui? · ah um tempão · e você morava onde antes? · tá e você gosta de morar aqui? · por quê? · é? · uhum você mora com seus pais? · entendi e se você pudesse escolher outro lugar pra morar? · putz · uhum · é... o transporte ele ser longe é o que mais te incomoda? · uhum · putz ... que coisa · é?... e como que é em relação com os vizinhos você conhece os seus vizinhos (todos)? · uhum · e aí vocês costumam se encontrar aqui na rua? · ahn o bairro mudou muito nesse tempo? · é? · mas o que que... fez deixar de ser bacana? · uhum · entendi antes não tinha/ era mais vazio por aqui? · uhum · ah nossa... tinha bastante árvore ainda então? ·

00:00:06.167

Seleção: 00:00:06.167 - 00:00:08.225 2058

ELAN's main functions

Control buttons

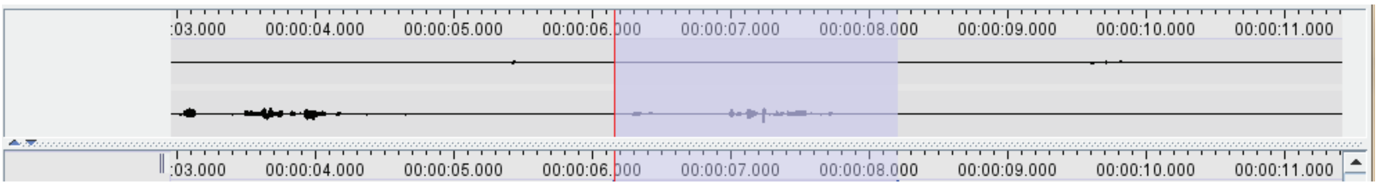


☑ Modo de Seleção ☑ Modo de Repetição (Loop)

(but you'll be using keyboard shorcuts!)

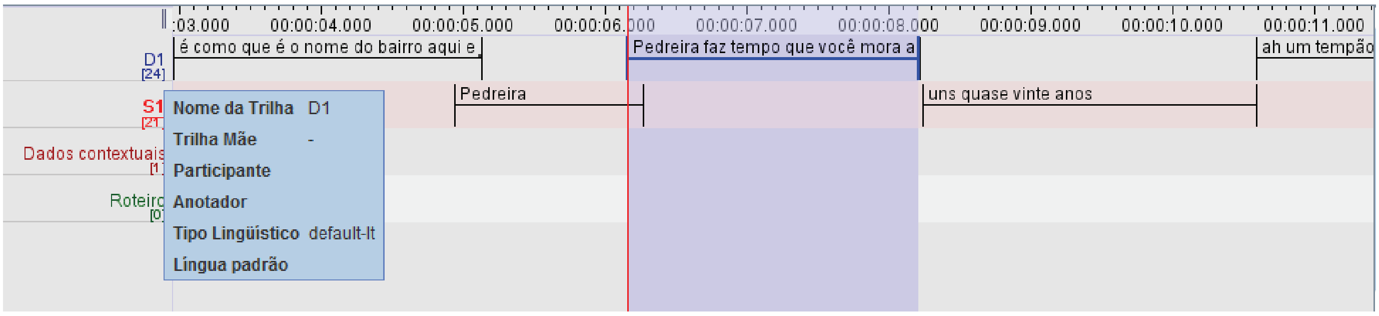
ELAN's main functions

Waveform (only for .wav files)



ELAN's main functions

Tiers



é como que é o nome do bairro aqui e

Pedreira faz tempo que você mora a

ah um tempão

Pedreira

uns quase vinte anos

Nome da Trilha D1
Trilha Mãe -
Participante
Anotador
Tipo Lingüístico default-lt
Língua padrão

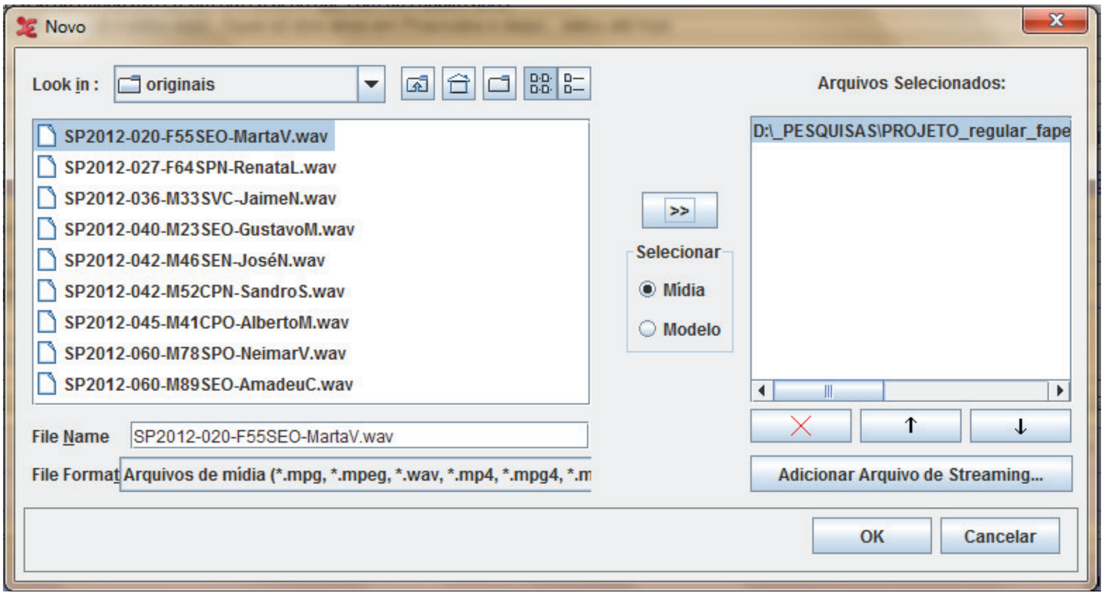
Shortcuts

Edit > Preferences > Edit shortcuts...

	Default	Suggestion
Annotation mode – Annotation editing		
Delete annotation	[Alt] + [D]	[Ctrl] + [Delete]
Modify active annotation value	[Alt] + [M]	[Ctrl] + [M]
New annotation here	[Alt] + [N]	[Shift] + [Enter]
Remove annotation value	[Alt] + [Delete]	[Shift] + [Delete]
Segmentation Mode – Media navigation		
Pause/play the media	[Ctrl] + [Space]	[Shift] + [Space]
Play selection	[Ctrl] + [Shift] + [Space]	[Ctrl] + [Space]
Set time one second back	[Shift] + [←]	[Ctrl] + [←]
Set time one second ahead	[Shift] + [→]	[Ctrl] + [→]
Go to previous pixel	[Ctrl] + [Shift] + [←]	[Shift] + [←]
Go to next pixel	[Ctrl] +[Shift] + [→]	[Shift] + [→]
Segmentation Mode – Selection		
Clear selection	[Alt] + [Shift] + [C]	[Esc]

New transcription

File > New...



Novo

Look in : originais

- SP2012-020-F55SEO-MartaV.wav
- SP2012-027-F64SPN-RenataL.wav
- SP2012-036-M33SVC-JaimeN.wav
- SP2012-040-M23SEO-GustavoM.wav
- SP2012-042-M46SEN-JoséN.wav
- SP2012-042-M52CPN-SandroS.wav
- SP2012-045-M41CPO-AlbertoM.wav
- SP2012-060-M78SPO-NeimarV.wav
- SP2012-060-M89SEO-AmadeuC.wav

File Name: SP2012-020-F55SEO-MartaV.wav

File Format: Arquivos de mídia (*.mpg, *.mpeg, *.wav, *.mp4, *.mpg4, *.m

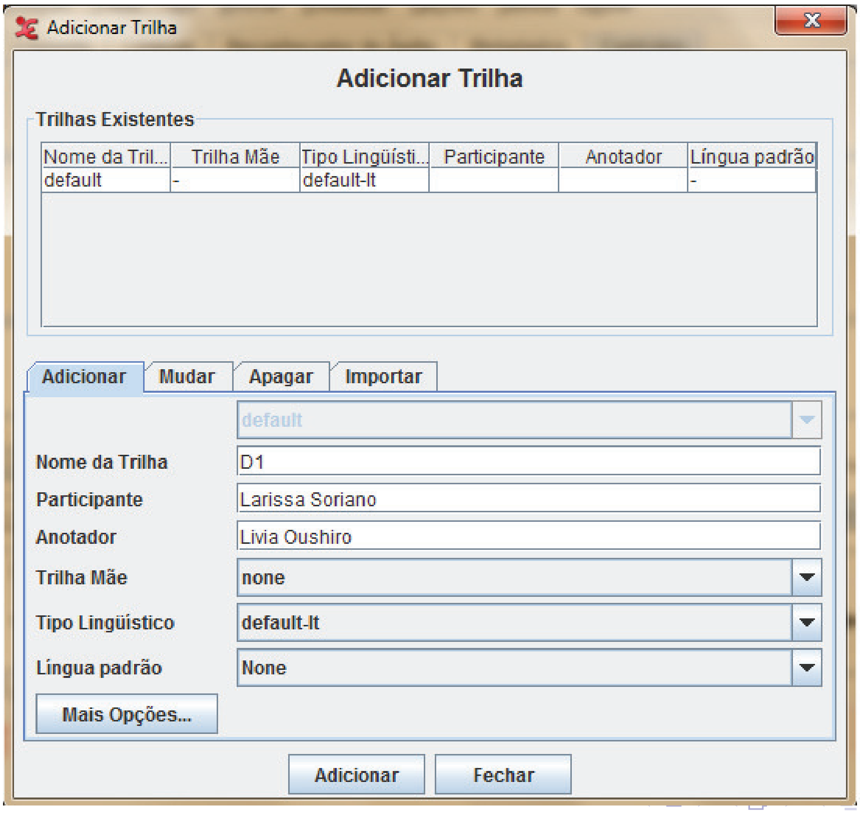
Selecionar: ● Mídia ○ Modelo

Arquivos Selecionados: D:_PESQUISAS\PROJETO_regular_fape

- ELAN: extension .eaf
- N.B.: The .pfsx file links the transcription and media files

Creating new tiers

Tier > Add new tier... (or [Ctrl] + [T])



Adicionar Trilha

Adicionar Trilha

Trilhas Existentes

Nome da Tril...	Trilha Mãe	Tipo Lingüísti...	Participante	Anotador	Língua padrão
default	-	default-lt			-

default

Nome da Trilha: D1

Participante: Larissa Soriano

Anotador: Livia Oushiro

Trilha Mãe: none

Tipo Lingüístico: default-lt

Língua padrão: None

Workflow

See Rosenfelder (2011:17–18)

Statistical analyses

- Three main objectives: synthesize, explain, predict
  - Descriptive Statistics: tables, plots
  - Inferential Statistics: tests that seek to generalize the observations on a sample to the population in general

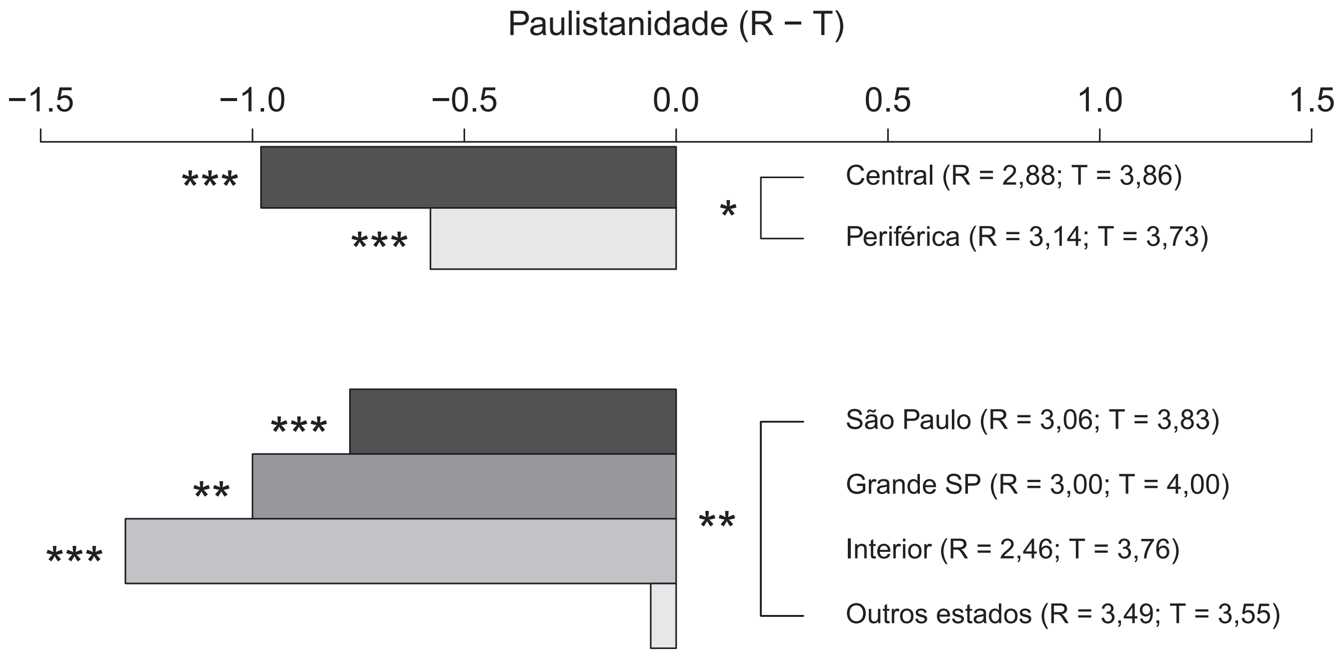
R: what it is and what it does

- Free software (available for Windows, Linux, MacOS)
- Programming language for statistical and graphical computations
- With R you can...
  - perform statistical analyses
  - make graphics
  - compile corpora
  - annotate corpora
  - make concordances
  - make frequency lists
  - ...

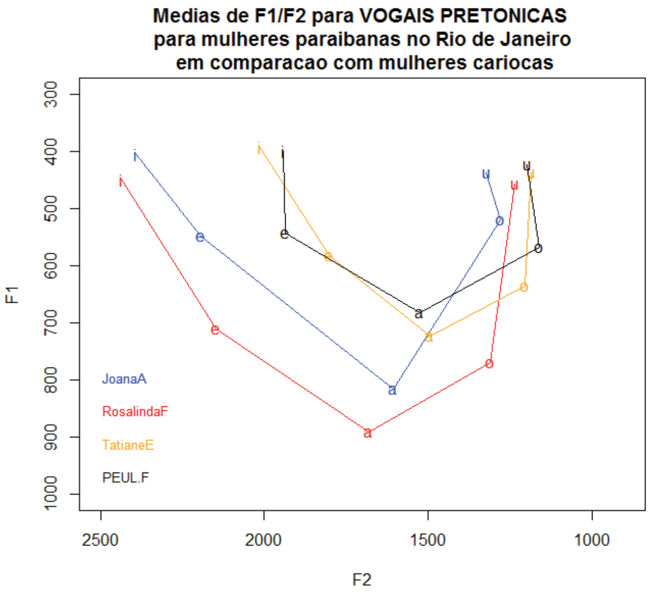
Advantages

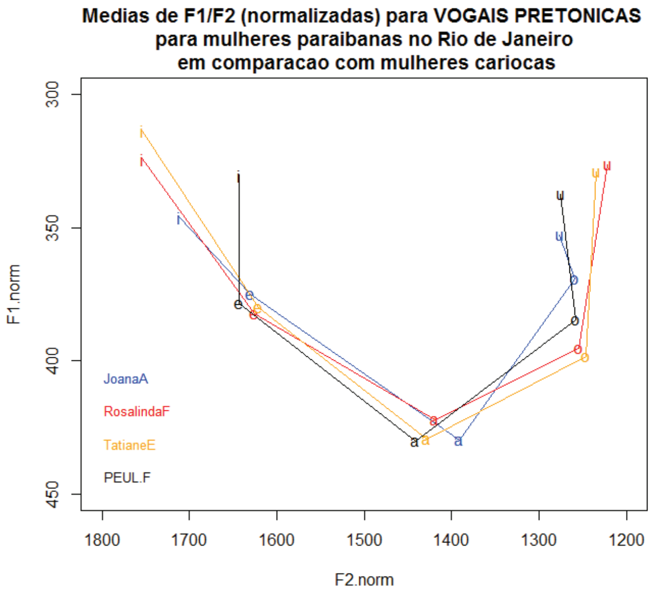
- Free, open source
- Flexibility in data manipulation (both textual/linguistic and numeric)
- Analyses of different types of variables (cf. GoldVarb)
- Analyses of interaction between predictors
- Mixed effects models
- Figures and graphics

Figures in R

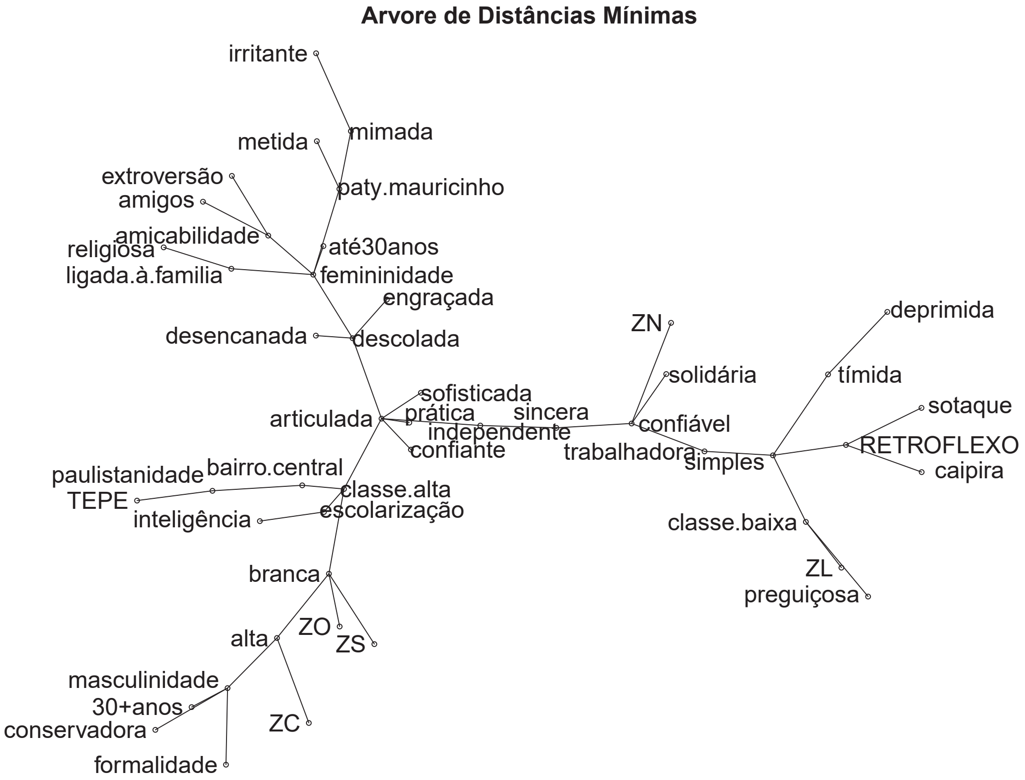


Figures in R

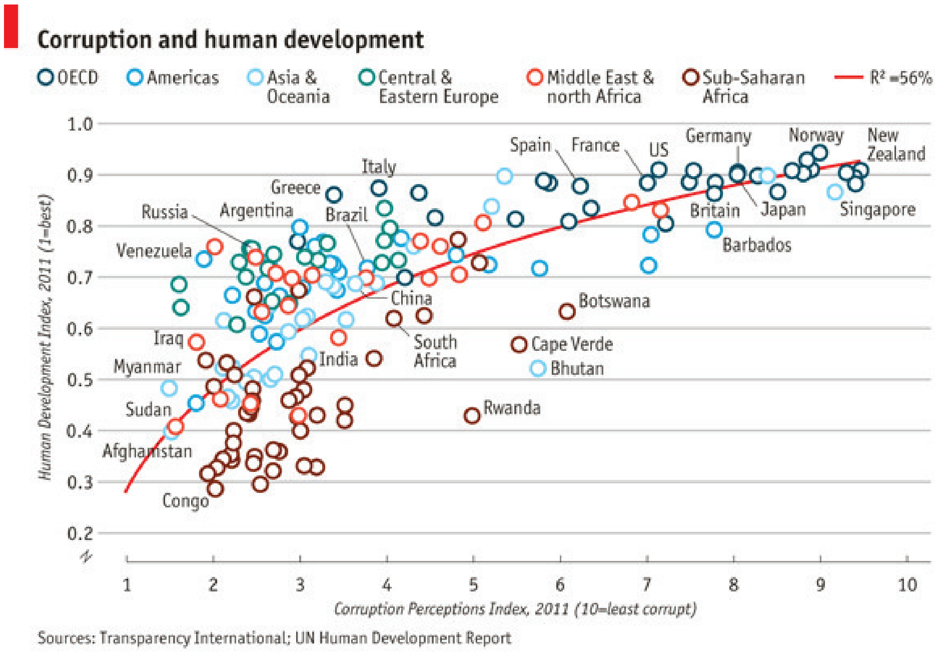




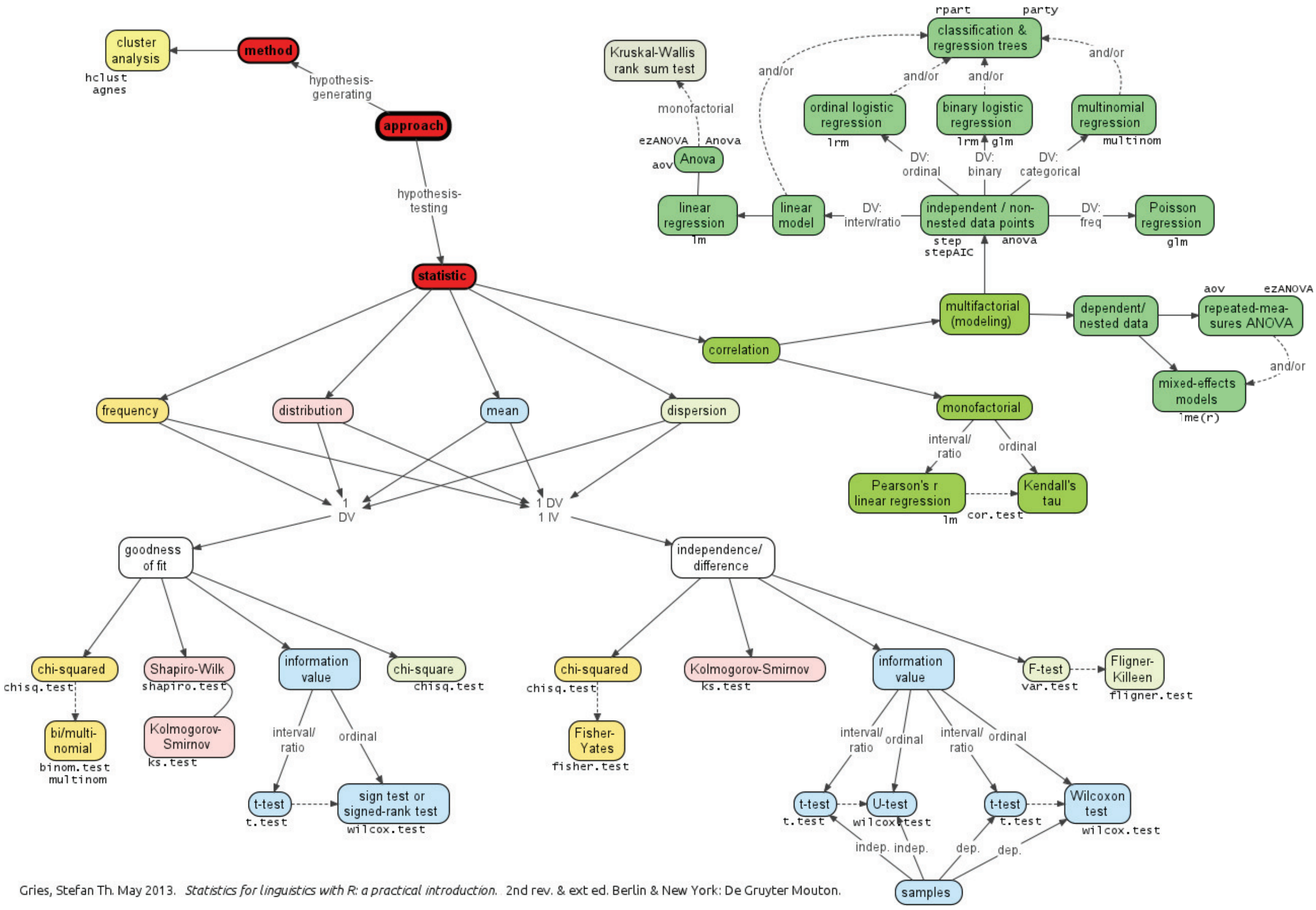
Figures in R



Figures in R



http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html



Gries, Stefan Th. May 2013. *Statistics for linguistics with R: a practical introduction.* 2nd rev. & ext ed. Berlin & New York: De Gruyter Mouton.

Your data and your research questions should determine what statistical tests to be run, and not the other way round

Using R means having **a new stance towards your data**

Installation

- R
  - Go to http://cran.r-project.org/ and download the latest version to your operational system (Linux, Mac, Windows)
  - Download and install the program
- RStudio: "friendlier" UI
  - Go to http://www.rstudio.com/ide/download/ and download the latest version to your operational system
  - Install and start the program
  - N.B.: it's necessary to have R installed to run RStudio

First contact

- Because it is a programming language, the user must instruct the program what is to be done through command lines
  - Disadvantage: for most commands, there are no pre-programmed buttons
  - Advantage: Because it doesn't have limited command buttons, the number of options the program offers is much wider than others (like GoldVarb X, SPSS, Excel, Calc)
  - Advantage: The user can save a sequence of commands in scripts, which can be reutilized and adapted later
  - Advantage: If you don't know which analysis is more adequate to your data, operating a series of pre-programmed buttons can be more harmful than beneficial...

RStudio's interface

- Source: script files
- Environment/History: objects in R's current session memory / history of command lines
- Console: where the command lines are executed
- Files, Plots, Packages, Help, Viewer

Intro to R swirl course

- swirl: Interactive interface for learning R in R
- Day 1, afternoon

`http://swirlstats.com/students.html`

- '...' means you should hit ENTER to continue
- `skip()` "skips" a question
- `play()` allows you to temporarily leave the swirl environment
- `nxt()` goes back to the tutorial

Types of variables

Types	How R reads them
categorical/nominal	factor
ordinal	factor/integer
numeric/continuous	integer/numeric

- All numeric variables are also ordinal
- All ordinal variables can be turned into nominal variables
- Therefore: given the chance to code a variable as numeric, do it!

How to organize your data file

The case-by-variable format (Gries 2013:15-26)

- The first row contains the names of the variables
- Each row represents one and only one case (one observation of the response/dependent variable)
- Each of the following columns represents one and only one variable
- Missing data are entered as "NA", and are not represented by empty cells
- Suggestions
  - code nominal variables as characters, not numbers (e.g., "1st", instead of "1")
  - don't use characters such as space, comma, tab, #, quotation marks, diacritics etc. for the variables or the variants
  - employ maximally simple but also maximally informative names for variables and variants

Tables

templateAnalyses.R, lines 30–53

- Frequency distribution of one variable: `table(); addmargins()`

```
> aa <- with(data, table(variable)) ¶
> addmargins(aa) ¶
```

- Frequency distribution between two variables: `table(); addmargins()`

```
> aa <- with(data, table(IV, DV)) ¶
> addmargins(aa) ¶
```

- Proportion table: `prop.table()`

```
> aa <- with(data, table(IV, DV)) ¶
> prop.table(aa, 1) ¶ #proportion by line
> prop.table(aa, 2) ¶ #proportion by column
> prop.table(aa) ¶ #general proportion
```

Barplot

templateAnalyses.R, lines 56–96

- `barplot()`

```
barplot(  x,                      #Table to be plotted
          beside=T,               #Bars side by side?
          horiz=F,                #Horizontal bars?
          main="",                #Plot title
          xlab="",                #Name of the variable on the x-axis
          ylab="",                #Name of the variable on the y-axis
          names.arg=c(...),       #Name of the variants on the x-axis
          legend.text=T,          #Plot legend?
          xlim=c(0,10),           #Limit values on the x-axis
          ylim=c(0,100),          #Limite values on the y-axis
          cex.axis=1,             #Proportion of the font size of numerical valu
          cex.names=1,            #Proportion of font size for axes labels
          col=NULL)               #Column colors
```

Valid for all types of plots

- See function `legend()` in templateAnalyses.R, lines 82–86
- For a list of colors, see `http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf`
- To save plots: Plots > Export > Save as Image/PDF...

Line charts

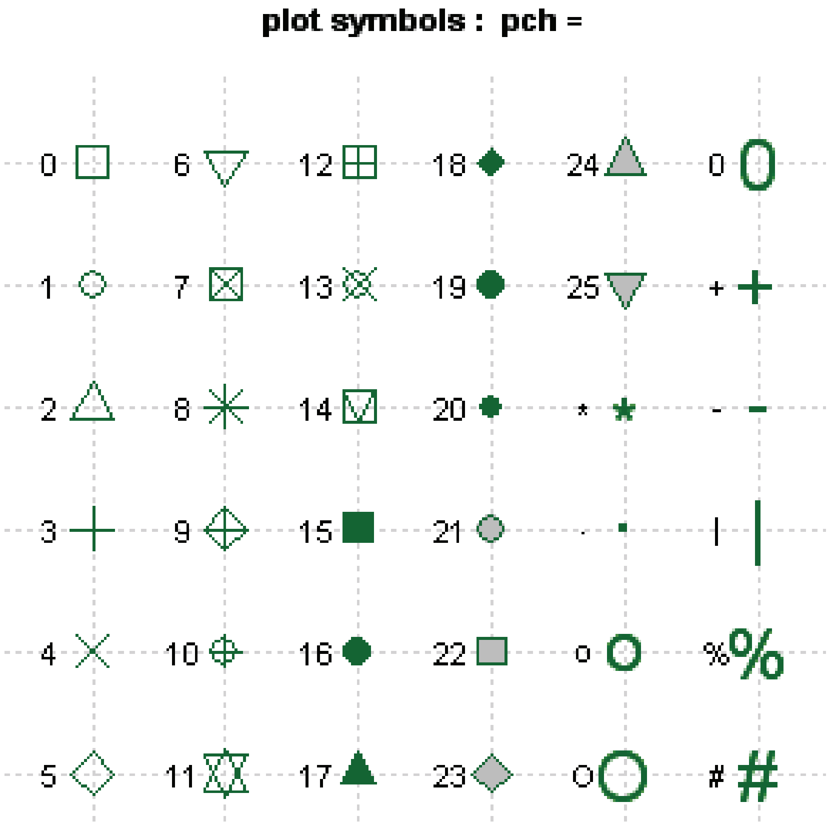
templateAnalyses.R, lines 97–124

- `plot()`

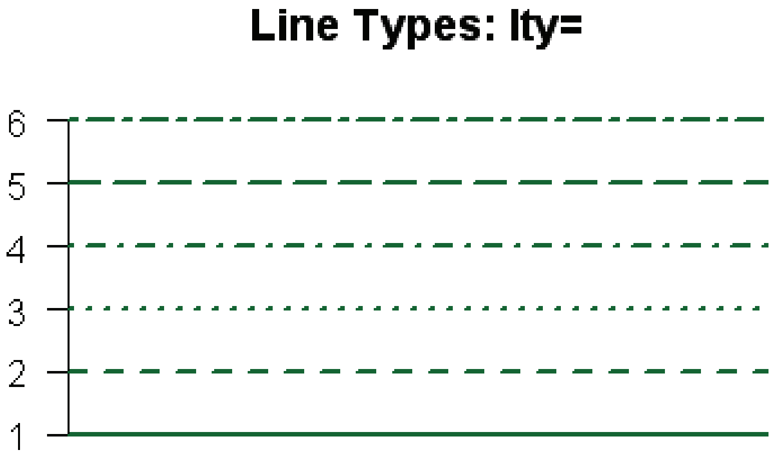
```
plot(   x,                    #Table to be plotted
        type="o",             #See ?plot
        pch=19,               #plot symbols (see next slides)
        lty=1,                #line type (see next slides)
        col="black",          #symbol and line color
        axes=F,               #plot axes?
        ylim=c(0,100),        #Limit values on the y-axis
        xlab="",              #Name of the variable on the x-axis
        ylab="")              #Name of the variable on the y-axis
```

- See also `axis()`, `box()` e `title()` in templateAnalyses.R, lines 109–119

Plot symbols for “pch”



Line types for “lty”



Proportion and Chisquare test

templateAnalyses.R, lines 175–233

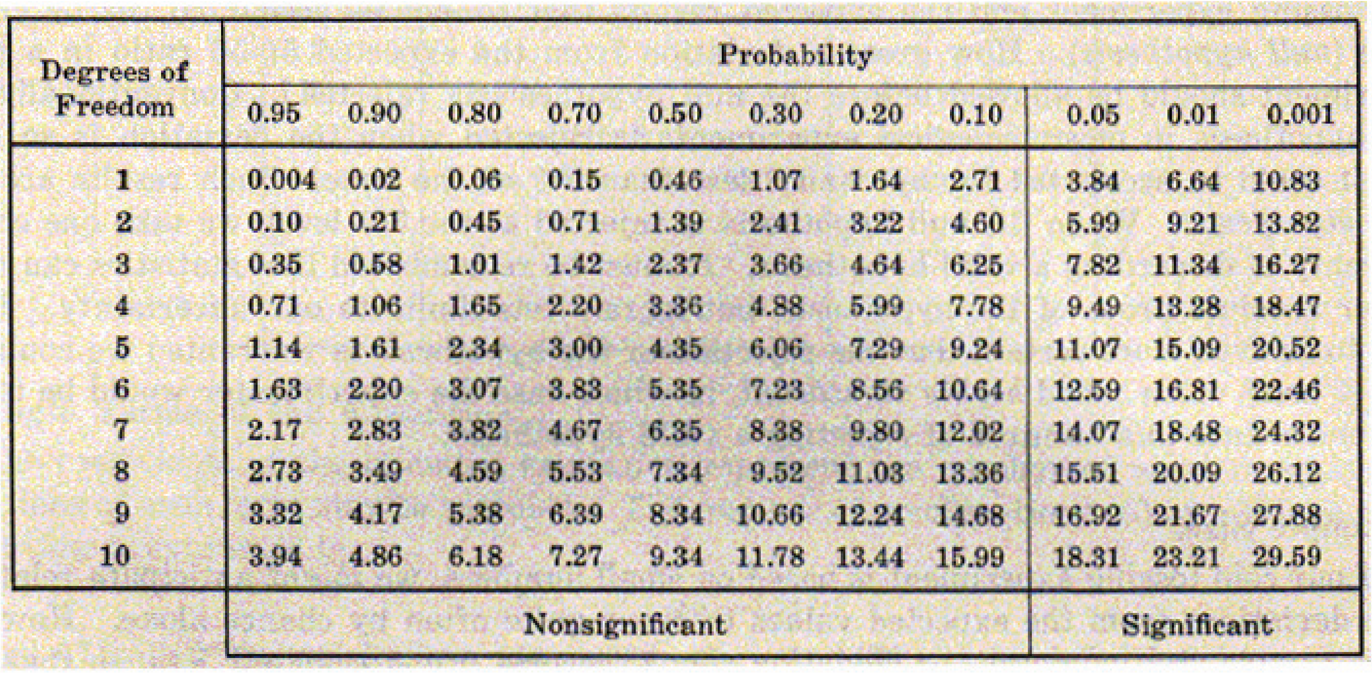
- For nominal variables
- Tests if there's difference between proportions
- Functions `prop.test()` and `chisq.test()`

`chisq.test(x)` #x is a frequency table

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Degrees of freedom: (n-rows - 1) x (n-columns - 1)

Table of probabilities of the chi-square distribution



Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

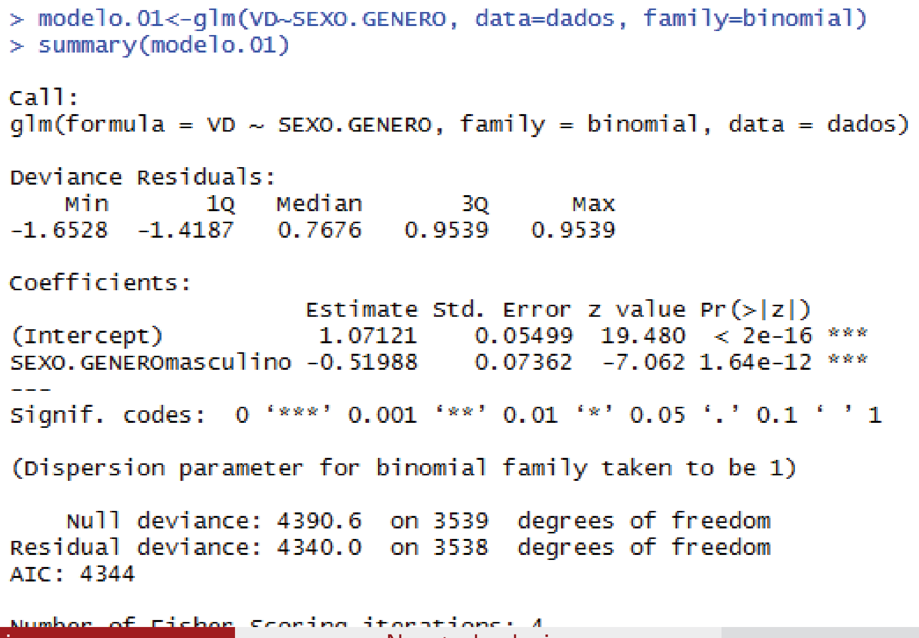
Useful functions

- Função `factor()`: turns vectors into factors
- Função `levels()`: assigns values to the levels of a factorial variable
  - reorganize the order of factors of a nominal variable
  - amalgamate factors of a nominal variable

Logistic regressions

templateAnalyses.R, lines 234–239

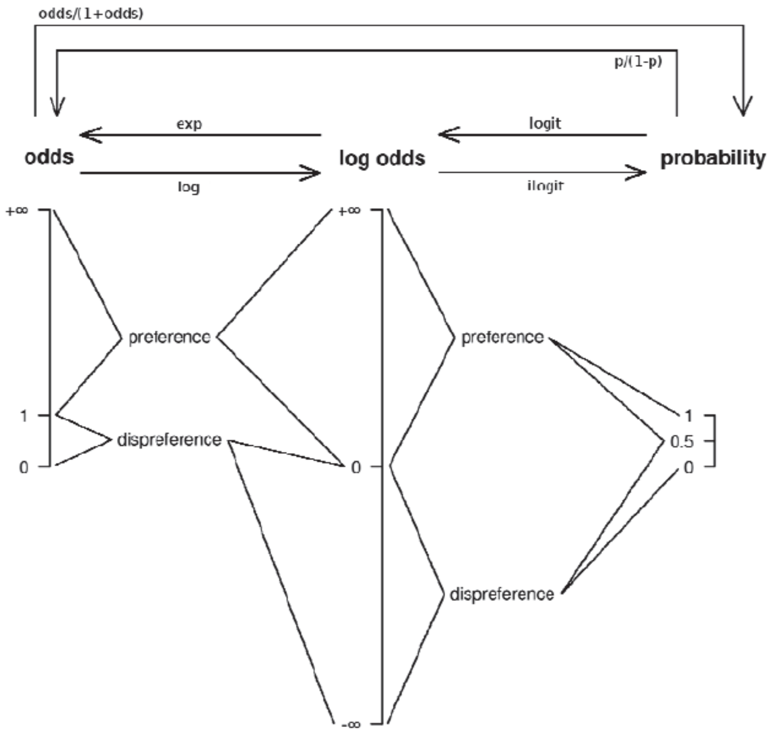
- Função glm()
  ex.: `modelo00 <- glm(DV ~ IV, data = data, family = "binomial")`



```
> modelo.01<-glm(VD~SEXO.GENERO, data=dados, family=binomial)
> summary(modelo.01)

Call:
glm(formula = VD ~ SEXO.GENERO, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6528  -1.4187   0.7676   0.9539   0.9539

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.07121    0.05499  19.480  < 2e-16 ***
SEXO.GENEROmasculino -0.51988    0.07362  -7.062 1.64e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4390.6  on 3539  degrees of freedom
Residual deviance: 4340.0  on 3538  degrees of freedom
AIC: 4344

Number of Fisher Scoring iterations: 4
```

Odds, Logodds, Probabilities

Gries (2013:300)



Interpretation

- Null deviance: how much variability there is if no predictor is included in the model
- Residual deviance: how much variability there is after including predictors
- Therefore: Null deviance - residual deviance: how much variability the included predictors can account for
- Fisher Scoring iterations: if number is too big (say, more than 20), the model is too complex to be run on the data you have and it doesn't converge -> incluide fewer predictors

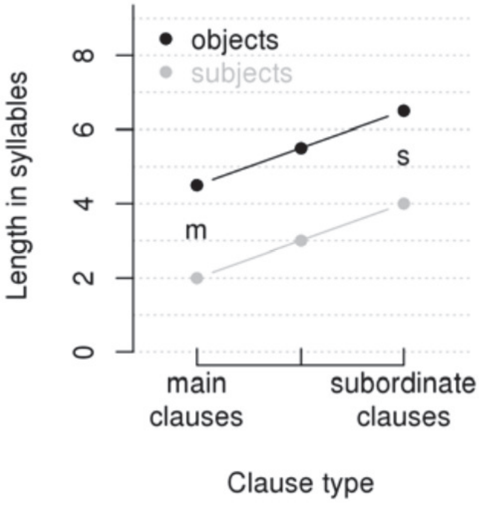
Interpretation *(cont.)*

- Coefficients
  - Intercept: logodds estimate for the *second* level of the DV
  - (Coefficients): difference between the estimate logodds in relation to the intercept, for the *second* level of the DV, when the predictor corresponds to that level

Interaction

Gries 2013:249–253 Example 1

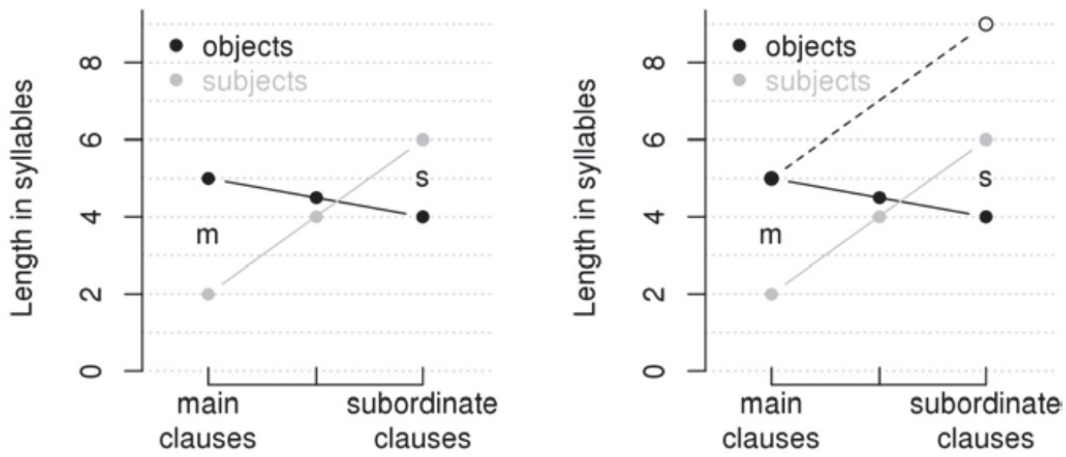
- In multivariate analyses, it's necessary to watch out for possible interactions between predictors
- Independence: addictive effect



Interaction

Gries 2013:249–253 Example 2

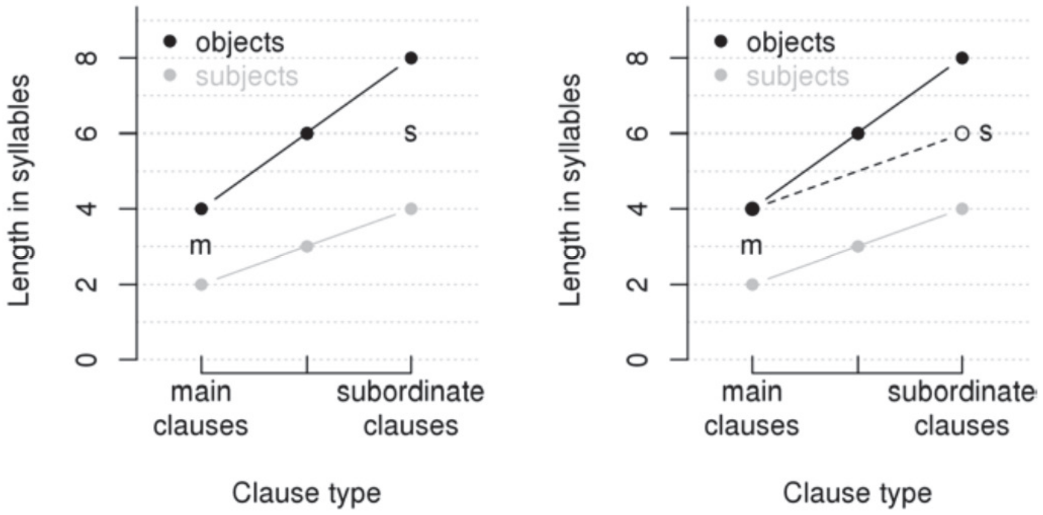
- In multivariate analyses, it's necessary to watch out for possible interactions between predictors
- Interaction: the effect of a predictor cannot be predicted without taking the effect of another predictor on the same response/dependent variable.



Interaction

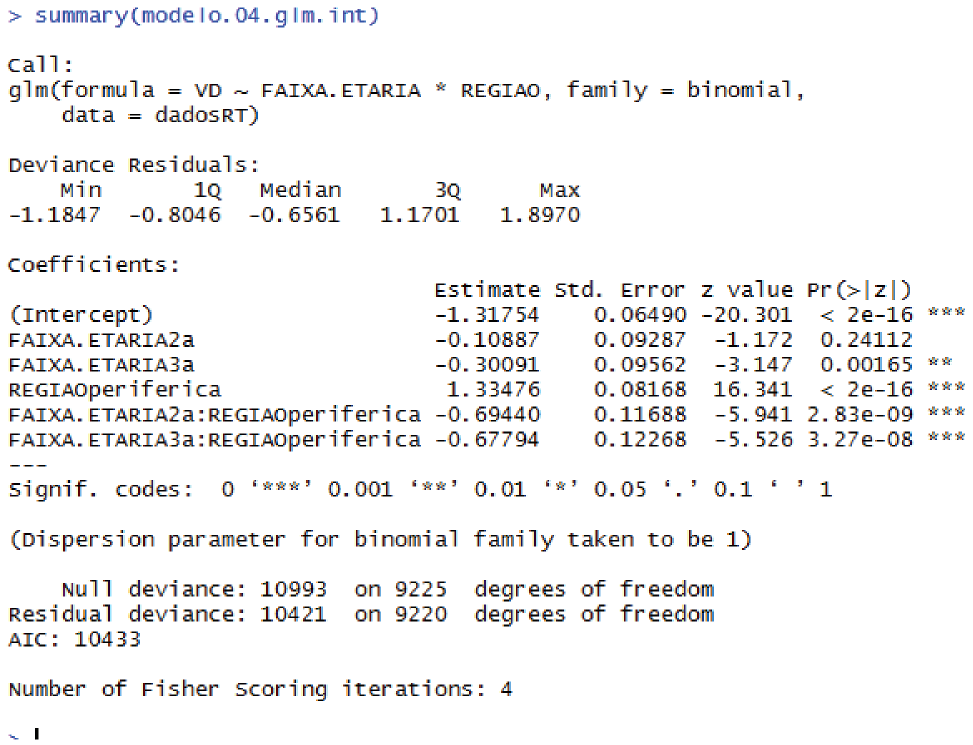
Gries 2013:249–253 Example 3

- In multivariate analyses, it's necessary to watch out for possible interactions between predictors
- Interaction: the effect of a predictor cannot be predicted without taking the effect of another predictor on the same response/dependent variable.



Interactions in regression models

- `model.glm<-glm(DV ~ VI * VI, data = data, family = binomial)`

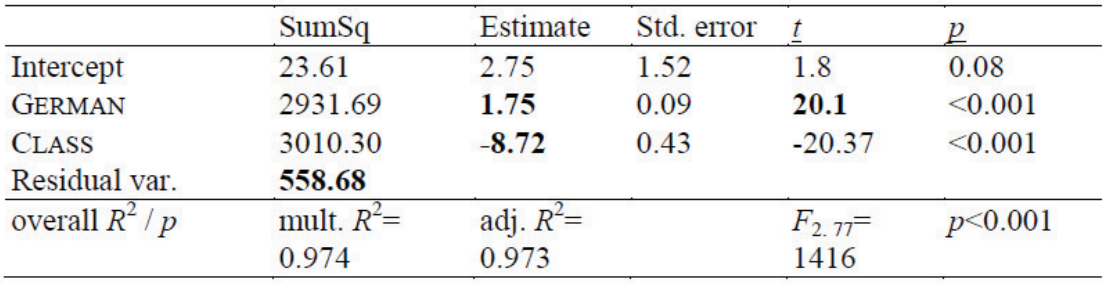


```
> summary(modelo.04.glm.int)

Call:
glm(formula = VD ~ FAIXA.ETARIA * REGIAO, family = binomial,
    data = dadosRT)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1847  -0.8046  -0.6561   1.1701   1.8970

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.31754    0.06490 -20.301  < 2e-16 ***
FAIXA.ETARIA2a                 -0.10887    0.09287  -1.172  0.24112
FAIXA.ETARIA3a                 -0.30091    0.09562  -3.147  0.00165 **
REGIAOperiferica                1.33476    0.08168  16.341  < 2e-16 ***
FAIXA.ETARIA2a:REGIAOperiferica -0.69440    0.11688  -5.941 2.83e-09 ***
FAIXA.ETARIA3a:REGIAOperiferica -0.67794    0.12268  -5.526 3.27e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10993  on 9225  degrees of freedom
Residual deviance: 10421  on 9220  degrees of freedom
AIC: 10433

Number of Fisher Scoring iterations: 4

>
```

How to report results

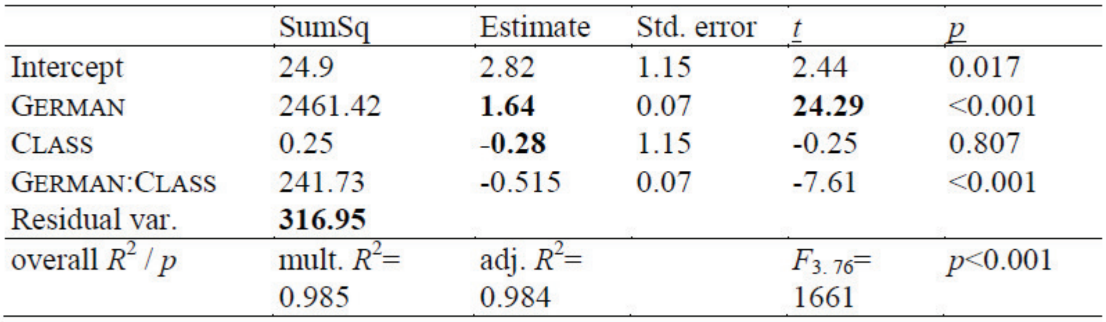
Gries (2013:257)

Table 44. The results of the linear model in (57)



	SumSq	Estimate	Std. error	*t*	*p*
Intercept	23.61	2.75	1.52	1.8	0.08
GERMAN	2931.69	**1.75**	0.09	**20.1**	<0.001
CLASS	3010.30	**-8.72**	0.43	-20.37	<0.001
Residual var.	**558.68**				
overall R^2 / *p*	mult. R^2= 0.974	adj. R^2= 0.973		$F_{2,77}$= 1416	p<0.001

Table 45. The results of the linear model in (58)



	SumSq	Estimate	Std. error	*t*	*p*
Intercept	24.9	2.82	1.15	2.44	0.017
GERMAN	2461.42	**1.64**	0.07	**24.29**	<0.001
CLASS	0.25	**-0.28**	1.15	-0.25	0.807
GERMAN:CLASS	241.73	-0.515	0.07	-7.61	<0.001
Residual var.	**316.95**				
overall R^2 / *p*	mult. R^2= 0.985	adj. R^2= 0.984		$F_{3,76}$= 1661	p<0.001

How to report results

Walker et al. (2014:179)

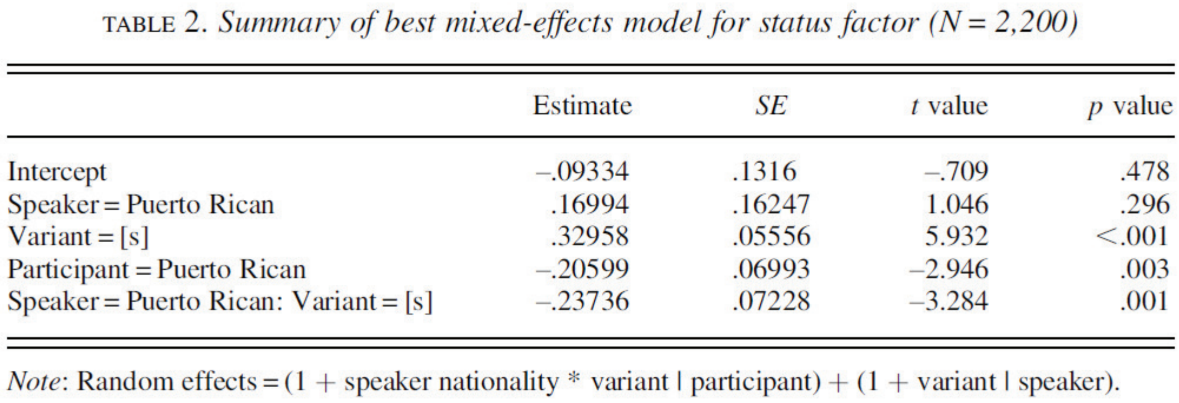


TABLE 2. *Summary of best mixed-effects model for status factor (N = 2,200)*

	Estimate	*SE*	*t* value	*p* value
Intercept	–.09334	.1316	–.709	.478
Speaker = Puerto Rican	.16994	.16247	1.046	.296
Variant = [s]	.32958	.05556	5.932	<.001
Participant = Puerto Rican	–.20599	.06993	–2.946	.003
Speaker = Puerto Rican: Variant = [s]	–.23736	.07228	–3.284	.001

Note: Random effects = (1 + speaker nationality * variant | participant) + (1 + variant | speaker).

How to report results

... and many figures

To learn more

To learn more about a function, type `?nameoffunction` on the Console. E.g.: `?scan`

R manuals: <http://cran.r-project.org/manuals.html>

Baayen, R. H. (2008) Analyzing Linguistic Data. A practical introduction to statistics using R. São Paulo: Cambridge University Press.

Dalgaard, P. (2008) Introductory statistics with R. New York: Springer.

Gries, S. Th. (2009) Quantitative Corpus Linguistics with R. A practical introduction. New York/London: Routledge.

Gries, S. Th. (2013) Statistics for Linguistics with R. Berlin/New York: Mouton de Gruyter.

Levshina, N. (2015) How to do Linguistics with R. Amsterdam: John Benjamins.