Linguistic Data Analysis using R -- Syllabus FS2026

Instructor: Sandra Auderset (Office B163) sandra.auderset@unibe.ch
Time and Place: Tuesdays 16:15-18:00 in F-103

Course description

This course introduces students to data analysis using R, with a focus on linguistic data and research questions. Students will learn to wrangle, visualize, and analyze data, while developing reproducible research practices. No prior programming experience required.

By the end of this course, you will be able to:

- read, clean, and transform linguistic datasets
- perform exploratory data analysis and basic statistical operations
- create publication-quality visualizations
- apply clustering and correlation techniques to linguistic data
- use version control (GitHub) for reproducible research
- critically evaluate data sources, analysis methods, and working with generative AI

Course policies

Attendance:

- attendance is not mandatory, but strongly recommended (hands-on course)
- if you miss class, it is your responsibility to catch up (materials on ILIAS, notes from colleagues)

Homework:

- complete assigned readings before the class session
- submit homework via ILIAS by due date
- you can discuss approaches with peers (I would recommend this), but write your own code
- in general, I don't accept late work unless you arrange for an extension before the due date; extensions are only granted in exeptional circumstances

Collaboration and AI use:

- projects should reflect your own, individual work
- LLMs (ChatGPT, Claude, etc.) permitted for debugging or if you get stuck, but:
  - you must understand every line of code you submit and be able to explain it when asked
  - cite AI assistance with the prompt in code comments
  - direct copying without understanding is academic misconduct

Evaluation

- homework: 30%
- project proposal: 10%
- project plan and presentation: 10%
- final paper with code: 50%

Schedule -- subject to change!

Week		Content of class	Readings	Homework
1	17.02.	Why data analysis for linguistics?; What is a database?; What is R and the tidyverse?; Course overview	Winter 2020: Sections 0.2 + 0.3 R4DS 2023: Introduction OHLDM 2021: Chapter 1.1 (1.2 optional)	install R and RStudio, create GitHub account, install GitHub Desktop
2	24.02.	Getting started with R data formats & types; File versioning and why it matters; Intro GitHub and GitHub Desktop	Winter 2020: Sections 1.1-1.11 R4DS 2023: 2 Workflow: basics; 7 Data import Intro to Github Desktop	Homework 1: Reflecting on open data/code in linguistics -> due on 02.03.
3	03.03.	Homework discussion; Reading in data; Getting started with the tidyverse: Intro to pipes	Winter 2020: Sections 1.12-1.15; 2.1-2.3, 2.5 R4DS 2023: 3 Data transformation	--
4	10.03.	Basics of data wrangling; How to get help; Final project overview	R4DS 2023: 5 Data tidying, 12 Logical vectors, 13 Numbers, 14 Strings, 16 Factors, 18 Missing values	Homework 2: R basics and GitHub -> due on 17.03.
5	17.03.	Intro to RMarkdown; Tidy data and transformation: understanding 'tidy' and 'messy' data; cleaning and transforming data	RMarkdown Guides Winter 2020: Chapter 3 R4DS 2023: 19 Joins Tidyverse Style Guide: 1, 2, 4	--
6	24.03.	Data wrangling: summarizing and aggregating data; joining data sets; Intro to descriptive statistics	Winter 2020: Chapter 3 Exercises R4DS 2023: 9 Layers, 10 Exploratory data analysis	Homework 3: Data wrangling -> due on 10.04.
7	31.03.	Data visualization basics: getting started with ggplot2; exploratory data analysis	R4DS 2023: 11 Communication GGPLOT2: 3 Individual geoms, 8 Annotations, 11 Colour scales and legends	--
	Easter break			
8	14.04.	Advanced data visualization: labels, palettes, and other customization	GGPLOT2 book: 13 Build a plot layer by layer ggmap tutorial	Homework 4: Data visualization -> due 28.04
9	21.04.	Maps: making maps with ggmaps and maps	--	--
10	28.04.	No class: Project work session with peer review	DataNovia: Cluster analysis datarundown: Types of clustering	**project proposal due (on GitHub)**

11	05.05.	Intro to cluster analysis	--	
12	12.05.	More cluster analysis	Bender et al. 2025; Quint & Allassonnière-Tang 2023; optional: Bender et al. 2021	--
13	19.05.	Working with AI; Use case: Random Forests for case paradigms	OHLDM 2021: Chapter 4 and 5	prepare presentation
14	26.05.	Project presentations; Class reflection (ethics of open science, ethics of AI, etc.)	--	**project presentations and plan due (before class)**
	15.07.	**Final project due**		

Materials and readings

Software (all free and open-source):

- R (version 4.0 or higher): https://www.r-project.org/
- RStudio Desktop: https://posit.co/download/rstudio-desktop/
- GitHub Desktop: https://desktop.github.com/
- GitHub account: https://github.com/

Textbooks (most freely available online):

- R4DS = Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. R for Data Science. 2nd edition. O'Reilly. https://r4ds.hadley.nz/
- GGPLOT2 = Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen. ggplot2: Elegant graphics for data science. 3rd edition. https://ggplot2-book.org/
- Winter 2020 = Winter, Bodo. 2020. Statistics for Linguists – An Introduction Using R. Routledge. [Available via university library]
- OHLDM = Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller, Lauren B. Collister (eds.). 2022. The Open Handbook of Linguistic Data Management. MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001
- Additional resources:
    - Tidyverse Style Guide: https://style.tidyverse.org/
    - RStudio Cheat Sheets: https://posit.co/resources/cheatsheets/
    - RMarkdown Guide: https://www.dataquest.io/blog/r-markdown-guide-cheatsheet/
    - RMarkdown Guide: https://rmarkdown.rstudio.com/

Additional readings:

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Bender, Emily M., Eamon Costello, Kyungmee Lee, Robert Farrow, and Giselle Ferreira. 2025. "Unsafe AI for Education: A Conversation on Stochastic Parrots and Other Learning Metaphors ⚠️". Journal of Interactive Media in Education 2025 (1): 10. https://doi.org/10.5334/jime.1079.

- Quint, Nicolas and Allassonnière-Tang, Marc. 2023. Inferring case paradigms in Koalib with computational classifiers. *Corpus Linguistics and Linguistic Theory* 19(2), pp. 237-269. https://doi.org/10.1515/cllt-2021-0028