Analyzing Language Data with R

LING-469 (Spring 2020)
Mon+Wed 2-3:15
Room: Maguire 103

NOTE: We will also have class on **Tuesday, February 18,** but not **Wednesday, April 8**

Instructor:
Amir Zeldes
E-Mail: amir.zeldes@georgetown.edu
Website: http://corpling.uis.georgetown.edu/amir
Office: Poulton Hall 243 (office hours Wednesdays, 3:30-5:00)

Teaching Assistant:
Logan Peng
E-Mail: sp1184@georgetown.edu

Summary:
This course will teach statistical analysis of language data with a focus on corpus materials, using the freely available statistics software 'R'. The course will begin with foundational notions and methods for statistical evaluation, hypothesis testing and visualization of linguistic data which are necessary for both the practice and the understanding of current quantitative research. As we progress we will learn exploratory methods to chart out meaningful structures in language data, such as agglomerative clustering, principal component analysis and multifactorial regression analysis. The course assumes basic mathematical skills and familiarity with linguistic methodology, but does not require a background in statistics or R.

Course requirements:

Attendance	
Homework assignments	60%
Final exam	25%
Participation	15%

Participation, assignments and final exam:
This course is conceived as an intensive introduction to natural language data analysis, making continuous attendance, submission and correction of homework essentials. Classwork will also employ the statistics software R, so that use of laptops in class is required for every session. If you do not have access to a laptop, please let me know immediately so that we can find a solution as soon as possible.

Homework assignments will be given regularly as either a smaller assignment for completion within a week or larger assignments for the following week. Assignments should be submitted as executable R scripts to facilitate correction and comparison with suggested solutions in class. Answers to prose questions as part of the exercises should be included as comments to the code in the R script.

The final exam will be without the use of a computer, and will not require you to remember complex R code or formulas. Instead, the exam will concentrate on definitions, their application to language data, analysis of example studies and visualizations of data. All types of exam questions will be covered in examples as part of the homework assignments and a mock final exam will be provided for studying.

Absences and timely assignment submission:
Students are expected to attend all classes and to complete all assignments on time. Absences may have an adverse effect on grades in the course, up to and including failure. That said, students may excuse themselves via e-mail from up to three meetings at their discretion, provided that they make up for lost course work and submit the assignments. Any additional absences for special reasons (religious observances, athletic travel, prolonged illness etc.) may be coordinated on a case by case basis with the instructor (documentation may be required as applicable). For this course in particular it is essential that any material missed by students is reviewed in depth until all points are understood – it is very easy to lose touch by missing some of the material, so please let me know about any topics that remain unclear so that we can discuss more in or out of class.

Reference works
This course does not use a text book, however the following books are recommended as reference works:

- Gries , Stefan Th. (2009) *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.
- Gries, Stefan Th. (2016) *Statistics for Linguistics With R: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.
- Oakes, Michael P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Dalgaard, Peter (2008) *Introductory Statistics with R*. New York: Springer.
- Baayen, R. Harald (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.
- Rietveld, Toni & van Hout, Roeland (2005) *Statistics in Language Research: Analysis of Variance*. Berlin/New York: Mouton de Gruyter.

Notice regarding sexual misconduct:
Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. University policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director of Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley, Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.