Syllabus: Textual data analysis with R (3 credits)

Linguistics 460 (Moreton)*

2025 August 19 (T)

Time:	TΘ 5:00–6:15
Places:	Fetzer 109 or: Zoom: Meeting ID: 963 8032 6331, passcode: 543792[a] *This is an in-person class. Zoom links are provided in case an emergency forces us to switch to remote instruction.*
Instructor	Elliott Moreton, `moreton@unc.edu`
Office hours:	(Tentatively:), Θ 11–12, F 2:15–3:15, and by appointment
Office:	Smith 101 or: Zoom: Meeting ID 973 4347 1746, passcode 531096[b]
Textbooks:	Freeman and Ross, *Technical Foundations of Informatics*[c] Wickham, Çetinkaya-Rundel, and Grolemund, *R for Data Science*[d] Silge and Robinson, *Text mining with R: a tidy approach*[e] Navarro, *Learning statistics with R: a tutorial for psychology students and other beginners*[f] **These books are available free on line.**

[a] `https://unc.zoom.us/j/96380326331?pwd=MWl6MHo4S2RVT214UlIrOEZFUFVLQT09`
[b] `https://unc.zoom.us/j/97343471746?pwd=cGRMUElraDAyRVY4UmcxT1RFcmxiZz09`
[c] `https://info201.github.io/`
[d] `https://r4ds.hadley.nz/`
[e] `https://www.tidytextmining.com/`
[f] `https://learningstatisticswithr.com/book/`

This is an in-person class. Zoom links are provided in case an emergency forces us to switch to remote instruction.

1 Target audience, course goals, and learning objectives

LING 460 is intended to introduce methods for working with textual data (corpora, databases, etc.) to an audience of linguists and others interested in using linguistic data to test hypotheses. The main topics include

- The R programming language[1]
- Processing text data (words and documents)
- Statistical hypothesis testing

These include both conceptual knowledge and practical skills. By the end of the course, students will have learned:

[1] `https://cran.r-project.org/index.html`

- How to organize, structure, and clean textual data
- How to use data-visualization tools to find patterns in data
- How to use regular expressions to search for patterns in text
- How to do some basic text-mining tasks (e.g., sentiment analysis, n-gram analysis, topic modelling)
- How to use machine-learning classification models to make predictions based on patterns in data
- The logic behind some of the most commonly-used statistical tests, and when and how to use them

The course will culminate in a student-designed text-based data project.

2 Prerequisites and requirements

There are no prerequisites. This class is designed for people with little or no experience in programming, statistics, or linguistics. Those with extensive background in statistics or computer science are encouraged to take a more advanced course.

LING 460 fulfills the following requirements:

- Research and Discovery requirement in IDEAS In Action
- Data Science minor core requirement (similar to STOR 120)
- Cognitive Science minor requirement
- Elective credit for Linguistics major and minor

3 Approximate semester schedule

Unit 1 (Weeks 1–4): **R and data-science basics**	R and RStudio. Data types, "tidy" data, and data visualization using `ggplot2`. Strings. "Tidy" text. Sentiment analysis.
Unit 2 (Weeks 5–8): **Machine learning for document classification.**	Text-mining concepts: document-term matrices, term frequency, tf-idf. Topic modelling. Statistical models as learning models for document classification: Naïve Bayes and LASSO regression.
MIDTERM	
Unit 3 (Weeks 8–9): **Experiment design.**	Independent and dependent variables. Correlational vs. manipulational studies. Confounds and controls. Final projects.
Unit 4 (Weeks 10–13): **Inferential statistics.**	Probability distributions, sampling theory, confidence intervals. Laws of Large Numbers. Central Limit Theorem. Logic of frequentist hypothesis testing. Statistical models and model fitting for hypothesis testing. Wilkes's Theorem and the likelihood-ratio test. Work on final projects
Project presentations (Weeks 14–16)	In-class presentations of projects. Because this is such a large class, THERE WILL BE PROJECT PRESENTATIONS THE TUESDAY OF THANKSGIVING WEEK; please plan to be there on that day. Missing that day in person will not be an excused absence.
FINAL EXAM	

4 Normal weekly schedule

Here's what will happen in a normal week:

	Tuesday	Thursday
⋮	⋮	⋮
Week n	Lab n due Reading n assigned	Start Lab $n+1$ Quiz n due
Week $n+1$	Lab $n+1$ due Reading $n+1$ assigned	Start Lab $n+2$ Quiz $n+1$ due
⋮	⋮	⋮

5 Class in the time of coronavirus

As I write this, the University's plan is to remain open for a full 15-week semester, and to hold classes in person. If the virus situation deteriorates, we may go to all-Zoom classes, at the Zoom link on p. 1 of this syllabus. Otherwise, we will meet in person and the class will not be streamed or recorded.

If you are too ill (with Covid or otherwise) to attend class safely, please contact me in advance to arrange to attend by Zoom. Others will not be admitted into the Zoom room.

6 Where to find course components

The main tools we will be using to communicate in this course are the following:

1. The class log[2], on the World Wide Web, updated after each class. Here you will find
   (a) A brief outline of what was covered each day
   (b) A list of any assignments made that day
2. The Canvas site[3]. Everyone who is enrolled in the class should already have access to it. Our class's ID, if you need it, is `LING460.001.FA25`. The main things we will need there are
   (a) Course materials like slides, handouts, and readings (under `Modules`)
   (b) The place to pick up lab assignments (under `Assignments`).
   (c) The place for on-line quizzes (under `Quizzes`)
   (d) A discussion forum for asynchronous collaboration (under `Discussions`)
   (e) The gradebook (under `Grading`)
3. The Zoom meeting links (see p. 1 of the syllabus). If Zoom is not already installed on your computer, please go to `zoom.unc.edu`[4] to get it. *This is an in-person class. The Zoom links are provided in case an emergency forces us to meet on-line.*

[2] `http://users.castle.unc.edu/~moreton/Ling460/460log.html`
[3] `http://canvas.unc.edu`
[4] `http://zoom.unc.edu`

7 Course components and assessment

The assessed components of this course include:

Reading quizzes (10%) These are meant to track progress in doing and understanding the readings and attending class. They will be done through the `Quizzes` section of the Canvas site, i.e., on-line, with two attempts and unlimited time. Normally, these will be due on Thursday before class.

Lab homeworks (20%) There will be about 8 of them, spread across the semester. They consist of small programming (coding) projects. They will normally be assigned on Thursday, when we will start working on them *in class*, and due Tuesday before class.

Each lab homework will be graded pass/fail. "Pass" means "made a serious attempt to answer each question correctly, even if the answer turned out to be wrong". *Simply putting superficial and obviously wrong answers will not count.*

You may drop or miss *one* of the lab homeworks.

Attendance and participation (10%) You are expected to come to class in person, having done the reading and thought about it until either (a) it makes sense, or (b) you can express precisely what about it doesn't make sense; either way, you'll have something to talk about in class.

If I start getting the impression that people aren't doing the readings, I'm going to institute pop quizzes. These are annoying because they waste class time, but coming to class without having done the reading wastes even more class time.

Missing classes will make it hard to keep up. It will also lower your participation grade. If you miss a class, it is your responsibility to get missed materials from me or other students. Always check the website if you have been absent.

To encourage participation, I will randomly call on students in class. You may be asked to explain a code snippet, propose a solution to a problem, answer another student's question, or otherwise contribute to class. Your responses will *not* be graded, and neither will your spontaneous contributions. (However, absences detected in this way count as absences!)

Exams (30%) There will be two, a midterm and a final, weighted equally and both cumulative from the beginning of the course. They will be done during the scheduled class time, and will be on paper.

Final project (30%) In the last half of the semester, students, working in groups of three, will choose a research question, then design, execute, and analyze an experiment to answer it, and finally present the question and the results to the class. This will take place in several steps, and I'll be giving details as each one comes up. All members of the group must contribute to coding and analysis. The grade will have a group component and an individual component.

Numeric grades will be converted to UNC's letter-grade system by mapping the numeric range from 60 to 100 onto the 10 passing letter grades from D to A, with four numeric points per step (except that A has 5 points, 96 to 100).

8 Policies

Attendance. If you must miss class because of a medical or family emergency, you should let me know *beforehand* by emailing or buttonholing me in person. If you miss a class, it is your responsibility to get missed materials from me or other students. Always check the class log and the Canvas website if you have been absent.

Late assignments. Homework solutions will normally be discussed in class the day the assignment is due. Therefore, as a general rule, NO LATE ASSIGNMENTS WILL BE ACCEPTED FOR CREDIT. To accommodate for emergencies, one lab-homework grade will be dropped (as noted above in the section about lab homeworks). For anything more than that, please see the University Approved Absences Office[5] and let me know by email as soon as you are able to arrange alternatives.

Collaboration and citation. It is a really good idea to discuss assignments with other people in the class and solve the problems together. However, each person (or each partnership) should write up their solution alone. If you work with other people, or look up information in sources that aren't officially part of this course, you are required to acknowledge them in the writeup. There is no shame in collaborating with other people, or in digging out information independently, but you need to give credit where it is due.

Generative artificial intelligence. "People" in the preceding paragraph means "humans". Use of generative AI for this class is governed by the guidelines developed by UNC-CH's Generative AI Committee, which can be found at this link[6] starting at the heading "Syllabus guidelines for generative AI" and continuing through the end of the page. The instructor (and the Honor Court, if it comes to that) will expect you to have read these guidelines. They apply to every aspect of the course, as long as they are not superseded by explicit written instructions from the instructor in an assignment.

Recording. Permission to make audio or video recordings of class will be given only in special circumstances (e.g., to students with hearing impairments).

Dates are still tentative at this point. I'll give at least two weeks' notice of the midterm, and will provide an exam syllabus (a study guide) one week before each exam.

The Carolina Honor Code[a] is in effect in this class, and I will treat violations seriously. You should review it. If you have questions about interpretation, you should bring them to me.

[a] `http://instrument.unc.edu`

9 General UNC-CH course policies and resources

Accessibility Resources The University of North Carolina at Chapel Hill facilitates the implementation of reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability or pregnancy complications resulting in difficulties with accessing learning opportunities.

All accommodations are coordinated through the Accessibility Resources and Service Office. See the ARS Website for contact information: `https://ars.unc.edu` or email `ars@unc.edu`.

[5] `https://uaao.unc.edu/`

[6] `https://web.archive.org/web/20240624194836/https://provost.unc.edu/student-generative-ai-usage-guidance/`

AI Use Policy Use of generative AI in this class is governed by the guidelines developed by UNC-CH's Generative AI Committee, which can be found at this link[7] starting at the heading "Syllabus guidelines for generative AI" and continuing through the end of the page. The instructor (and the Honor Court, if it comes to that) will expect you to have read these guidelines.

Attendance Policy No right or privilege exists that permits a student to be absent from any class meetings, except for these University Approved Absences:

1. Authorized University activities
2. Disability/religious observance/pregnancy, as required by law and approved by Accessibility Resources and Serviceand/or the Equal Opportunity and Compliance Office(EOC)
3. Significant health condition and/or personal/family emergency as approved by the Office of the Dean of Students, Gender Violence Service Coordinators, and/or the Equal Opportunity and Compliance Office (EOC).

Please communicate with me early about potential absences. Please be aware that you are bound by the Honor Code when making a request for a University approved absence.

Code of Conduct All students are expected to adhere to University policy and follow the guidelines of the UNC Code of Conduct. Additional information can be found at `https://studentconduct.unc.edu/`.

Counseling and Psychological Services UNC-Chapel Hill is strongly committed to addressing the mental health needs of a diverse student body. The Heels Care Network website[8] is a place to access the many mental health resources at Carolina. CAPS is the primary mental health provider for students, offering timely access to consultation and connection to clinically appropriate services. Go to the CAPS website[9] or visit their facilities on the third floor of the Campus Health building for an initial evaluation to learn more. Students can also call CAPS 24/7 at 919-966-3658 for immediate assistance.

Diversity Statement I value the perspectives of individuals from all backgrounds reflecting the diversity of our students. I broadly define diversity to include race, gender identity, national origin, ethnicity, religion, social class, age, sexual orientation, political background, and physical and learning ability. I strive to make this classroom an inclusive space for all students. Please let me know if there is anything I can do to improve, I appreciate suggestions.

Equal Opportunity and Compliance — Accommodation Equal Opportunity and Compliance Accommodations Team (Accommodations — UNC Equal Opportunity and Compliance[10]) receives requests for accommodations for disability, pregnancy and related conditions, and sincerely held religious beliefs and practices through the University's Policy on Accommodations. EOC Accommodations team determines eligibility and reasonable accommodations consistent with state and federal laws.

Grade Appeal Process If you have any concerns with grading and/or feel you have been awarded an incorrect grade, please discuss it with me as soon as possible. If we cannot resolve the issue, you may talk to our director of undergraduate studies or department chair.

[7] `https://provost.unc.edu/student-generative-ai-usage-guidance/`
[8] `http://care.unc.edu/`
[9] `https://caps.unc.edu/`
[10] `https://eoc.unc.edu/accommodations/`

Honor Code Statements

1. All students are expected to follow the guidelines of the UNC honor code. In particular, students are expected to refrain from "lying, cheating, or stealing" in the academic context. If you are unsure about which actions violate that honor code, please see me or consult honor.unc.edu. (source: Department of Asian Studies)
2. Students are bound by the Honor Code in taking exams and in written work. The Honor Code of the University is in effect at all times, and the submission of work signifies understanding and acceptance of those requirements. Plagiarism will not be tolerated. Please consult with me if you have any questions about the Honor Code. (source: syllabus from section of HIST 486 offered in 2015)
3. The University of North Carolina at Chapel Hill has had a student-administered honor system and judicial system for over 100 years. The system is the responsibility of students and is regulated and governed by them, but faculty share the responsibility. If you have questions about your responsibility under the honor code, please bring them to your instructor or consult with the office of the Dean of Students or the Instrument of Student Judicial Governance. This document, adopted by the Chancellor, the Faculty Council, and the Student Congress, contains all policies and procedures pertaining to the student honor system. Your full participation and observance of the honor code is expected (honor.unc.edu). (source: syllabus from section of GEOG 67 offered in 2015)

Syllabus changes The instructor reserves the right to make changes to the syllabus including project due dates and test dates. These changes will be announced as early as possible.

Title IX Resources Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Reports can be made online to the EOC[11] or by contacting the University's Title IX Coordinator, Elizabeth Hall (titleixcoordinator@unc.edu), or the Report and Response Managers in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu). Please note that I am designated as a Responsible Employee, which means I must report to the EOC any information I receive about the forms of misconduct listed in this paragraph. If you'd like to speak with a confidential resource, those include Counseling and Psychological Services, the University's Ombuds Office, and the Gender Violence Services Coordinators (gvsc@unc.edu). Additional resources are available at `safe.unc.edu`.

Undergraduate Testing Center The College of Arts and Sciences provides a secure, proctored environment in which exams can be taken. The center works with instructors to proctor exams for their undergraduate students who are not registered with ARS and who do not need testing accommodations as provided by ARS. In other words, the Center provides a proctored testing environment for students who are unable to take an exam at the normally scheduled time (with pre-arrangement by your instructor). For more information, visit `http://testingcenter.web.unc.edu/`.

Additional student resources 1. The Learning Center: The UNC Learning Center is a resource both for students who are struggling in their courses and for those who want to be proactive and develop sound study practices to prevent falling behind. They offer individual consultations, peer tutoring, academic coaching, test prep programming, study

[11] `https://eoc.unc.edu/report-an-incident/`

skills workshops, and peer study groups. If you think you might benefit from their services, please visit them in SASB North or visit their website to set up an appointment: `http://learningcenter.unc.edu`.

2. The Writing Center: The Writing Center is located in the Student and Academic Services Building and offers personalized writing consultations as well as a variety of other resources. This could be a wonderful resource to help with your writing assignments in this course (and any assignments in your other courses). You do not need a complete draft of your assignment to visit; they can help you at any stage! You can chat with someone in the writing center or set up as appointment on their website: `http://writingcenter.unc.edu`.
3. Resources for Success in Writing: UNC has a Writing Center that provides one-on-one assistance to students free of charge. To make an appointment, browse the Writing Center's online resources, or submit a draft online. They have additional useful information, such as handouts on how to cite online.