

Final Project Rubric

LING 414: Statistical Methods in Linguistics
Fall 2026

University of Rochester
Department of Linguistics

1 Overview

- ❗ **Final Project Weight:** 40% of total course grade (LING 414)
 - **Pre-proposal Meeting:** 5% of project grade (2% of course grade)
 - **Proposal:** 5% of project grade (2% of course grade)
 - **Presentation:** 20% of project grade (8% of course grade)
 - **Code Submission:** 20% of project grade (8% of course grade)
 - **Paper Submission:** 50% of project grade (20% of course grade)

2 Project Options

Students must choose one of three project types. Each has specific requirements detailed below.

2.1 Option 1: Corpus Analysis

📋 Requirements:

- Use an open source corpus or one from the Linguistic Data Consortium.
 - UR students have free LDC access. See [this page](#).
- Develop data extraction tools.
 - **Good:** I used spaCy to extract all sentences with passive constructions from the NOW corpus, then calculated their frequency by genre, ...
 - **Bad:** I searched COCA for what preceded this one word, then copied and pasted from the table it output into the paper.
- Create processed dataset for analysis using best practices for data representation learned in the course, such as tidy data principles.
- Design analysis pipeline to test a specific linguistic hypothesis about patterns in the corpus.

2.2 Option 2: Judgment/Reading Experiment Design

🔧 Requirements:

- Design complete experimental protocol using experimental design principles from the course.
- Specify participant recruitment strategy and inclusion/exclusion criteria.
- Develop stimuli set for testing specific linguistic hypothesis.
- Define the procedure for data collection.
- Create a data collection instrument using a standard framework (PsychoPy, jsPsych, etc.).
- Generate simulated datasets.
 - Unlikely you will be able to collect data from human subjects within the project timeline.
- Develop a full analysis pipeline, including power analysis based on simulated data.

2.3 Option 3: Mechanistic Interpretation Study

🧠 Requirements:

- Select set of pre-trained deep learning models of interest.
- Choose probing method(s) for mechanistic interpretation.
- Identify linguistic phenomenon(s) to investigate.
- Design probing experiments to:

- Isolate specific model components (layers, attention heads, neurons).
- Measure model behavior on controlled inputs.
- Test hypothesis about internal workings of a deep learning model.
- Use statistical techniques developed in this course.
- Provide mechanistic interpretation of model behavior.

3 Git Repository Structure

All projects must be submitted via a Git repository with the following structure:

project-root/	
- README.md	Project overview and setup instructions
- DESCRIPTION/pyproject.toml	Dependencies and environment setup
- data/	
- raw/	Original data files
- processed/	Cleaned/processed data
- code/	
- preprocessing/	Data-preparation scripts
- analysis/	Statistical-analysis scripts
- visualization/	Plotting and figure generation
- results/	
- figures/	Generated plots and visualizations
- tables/	Statistical-output tables
- paper/	
- paper.pdf	Final paper submission
- paper.tex	LaTeX source (optional)
- docs/	Additional documentation using standard documentation tools (e.g. mkdocs)

4 Paper Structure Requirements

4.1 All Project Types

Section	Content Requirements	Page Range
Abstract	Research question, methods, key findings	~ $\frac{1}{3}$ page
Introduction	Background, motivation, hypothesis	1–2 pages
Lit Review	Relevant prior work, theoretical framework	2–3 pages
Methods	Detailed methodology, statistical approach	2–3 pages
Results	Findings with figures/tables	3–4 pages
Discussion	Interpretation, limitations, implications	2–3 pages
Conclusion	Summary and future work	0.5–1 page
References	Complete citations	1–2 pages
Total		12–18 pages

4.2 Option-Specific Requirements

4.2.1 Option 1: Corpus Analysis

- **Methods Section (2–3 pages):**
 - Corpus selection justification with size, genre, and time period specifications.
 - Data extraction pipeline architecture and tool specifications (e.g., spaCy, NLTK).
 - Linguistic feature operationalization and annotation scheme.
 - Statistical analysis plan including distributional assumptions and corrections for multiple comparisons.
- **Results Section (3–4 pages):**
 - Descriptive statistics of corpus composition and extracted features.
 - Frequency distributions and co-occurrence patterns with confidence intervals.

- Statistical tests of linguistic hypotheses with effect sizes.
- Visualizations: frequency plots, heatmaps, network graphs of linguistic relationships.
- **Required Appendices:**
 - Sample of extracted data (minimum 100 examples) with annotations.
 - Complete extraction scripts with documentation.
 - Corpus preprocessing decisions and edge case handling.
 - Inter-annotator agreement metrics if manual annotation was performed.

4.2.2 Option 2: Experiment Design

- **Methods Section (2–3 pages):**
 - Experimental design specification (within/between subjects, factorial structure, counterbalancing).
 - Participant demographics, recruitment strategy, and inclusion/exclusion criteria.
 - Stimuli construction methodology with linguistic controls and norming procedures.
 - Power analysis with assumed effect sizes based on prior literature.
 - Data collection procedure timeline and experimental interface details.
- **Results Section (3–4 pages):**
 - Simulated data generation process with parameter justification.
 - Analysis of simulated datasets showing expected patterns.
 - Power curves and sample size recommendations.
 - Mixed-effects model specifications and random effects structure.
 - Sensitivity analyses for different effect size scenarios.
- **Required Appendices:**
 - Complete stimuli list with item characteristics and counterbalancing scheme.
 - Experimental interface screenshots and participant instructions.
 - Simulation description and parameters for reproducibility.

4.2.3 Option 3: Mechanistic Interpretation

- **Methods Section (2–3 pages):**
 - Model selection justification with architecture specifications.
 - Probing methodology details (linear probes, attention analysis, causal interventions).
 - Dataset construction for probing experiments with linguistic controls.
 - Statistical framework for interpreting probe accuracy and significance.
 - Comparison methodology across layers, attention heads, or model variants.
- **Results Section (3–4 pages):**
 - Layer-wise analysis of linguistic feature encoding.
 - Attention pattern visualizations and interpretations.
 - Statistical validation of mechanistic hypotheses.
 - Comparison of representations across model architectures.
 - Ablation study results showing component importance.
- **Required Appendices:**
 - Model specifications, hyperparameters, and training details.
 - Complete probing datasets with linguistic annotations.
 - Visualization gallery of attention patterns and activations.
 - Reproducibility checklist with random seeds and compute requirements.
 - Supplementary statistical analyses and robustness checks.

5 Grading Criteria

5.1 Presentation Assessment (20% of Project Grade)

5.1.1 Required Slide Structure

Academic Presentation Format (15–20 slides total):

1. **Title Slide:** Title, authors, affiliations, date.
2. **Outline/Roadmap:** Overview of talk structure.
3. **Introduction/Motivation:** Why this research matters (2–3 slides).
4. **Background/Literature Review:** Previous work and context (3–4 slides).
5. **Research Questions:** Clear statement of hypotheses (1 slide).
6. **Methods:** Data, tools, analysis approach (3–4 slides).
7. **Results:** Key findings with visualizations (4–5 slides).
8. **Discussion:** Interpretation and implications (2–3 slides).
9. **Conclusions:** Summary and future work (1–2 slides).
10. **Acknowledgments:** Collaborators, resources (1–2 slides).
11. **References:** Key citations.
12. **Appendix Slides:** Additional details for Q&A.

5.1.2 Slide Design Requirements

Component	Criteria	Weight
Visual Design	Clean layout, readable fonts (24pt+), appropriate color scheme	15%
Content Quality	Clear, concise text; effective use of figures/tables	25%
Structure	Logical flow, proper transitions, timing	20%
Delivery	Clear speech, eye contact, engagement with audience	20%
Q&A Handling	Thoughtful responses, depth of understanding	20%

5.1.3 Submission Requirements

-  **Presentation Materials to Submit:** PDF version of slides
Due: 24 hours before presentation date

5.2 Code Assessment (20% of Project Grade)

Oral Assessment Format:

- 10–20 minute interview during finals period.
- Code will be shown without comments.
- Students must explain functionality line-by-line.
- Questions will test understanding of implementation choices.

Component	Criteria	Weight
Functionality	Code runs without errors, produces correct output	40%
Understanding	Can explain all code decisions during oral assessment	30%
Organization	Clear structure, appropriate modularization	15%
Documentation	README, inline documentation	15%

5.3 Paper Assessment (50% of Project Grade)

Component	Criteria	Weight
Research Question	Clear, testable hypothesis; appropriate scope	15%
Methodology	Appropriate statistical methods; sound design	20%
Analysis	Correct implementation; thorough exploration	20%
Results	Clear presentation; appropriate visualizations	15%
Discussion	Insightful interpretation; acknowledges limitations	15%
Writing Quality	Clear, concise, well-organized	10%
Citations	Complete, properly formatted	5%

6 Detailed Grading Rubric

6.1 Exemplary (A: 93–100%)

- **Code:** Elegant, efficient implementation with excellent documentation.
- **Analysis:** Sophisticated multifaceted statistical approach that uses advanced techniques appropriately.
- **Paper:** Publication-quality writing with novel insights.
- **Oral:** Demonstrates deep understanding, can discuss alternatives.

6.2 Proficient (B: 83–92%)

- **Code:** Correct implementation with good organization.
- **Analysis:** Appropriate methods correctly applied.
- **Paper:** Clear writing with sound conclusions.
- **Oral:** Can explain all code with minor prompting.

6.3 Developing (C: 73–82%)

- **Code:** Mostly functional with some errors or inefficiencies.
- **Analysis:** Basic methods with some misunderstandings.
- **Paper:** Adequate writing with unclear sections.
- **Oral:** Struggles to explain some code sections.

6.4 Inadequate (D–F: < 73%)

- **Code:** Major errors or incomplete implementation.
- **Analysis:** Incorrect methods or misinterpretation.
- **Paper:** Poor organization or missing components.
- **Oral:** Cannot explain core functionality.

7 Submission Requirements

📅 Timeline:

- **Pre-proposal Meeting:** Scheduled on or before Oct. 16.
- **Project Proposal:** Due Oct. 28 (Week 9).
- **Project Presentations:** Dec. 7, 9, or 14 (Weeks 15–16).
- **Paper & Code Submission:** Due Dec. 14.
- **Oral Assessments:** Scheduled during the Dec. 18–23 finals period.

7.1 Proposal Requirements (1–2 pages)

1. Research question and hypothesis.
2. Chosen project option and justification.
3. Data source or experimental-design overview.
4. Proposed statistical methods.
5. Timeline for completion.

7.2 Presentation Requirements (15 minutes)

1. Research question and motivation.
2. Methodology overview.
3. Preliminary results (if available).
4. Challenges and next steps.
5. 5 minutes for questions.

8 Academic Integrity

⚠ Important Policies:

- All code must include citations for external sources.
- LLM-generated code must be disclosed and you must provide prompts used.
- You may not use LLMs for writing the paper, except to generate tables, figures, etc.

9 Resources and Support

9.1 Corpus Resources

9.1.1 Institutional & Licensed Corpora

- **LDC:** Linguistic Data Consortium (ldc.upenn.edu)
- **BNC:** British National Corpus - 100M words balanced (english-corpora.org/bnc)

9.1.2 Large Web-Scale Corpora

- **Common Crawl:** 50+ billion web pages, petabytes of data (commoncrawl.org)
- **NOW Corpus:** 15.6+ billion words, updated monthly (english-corpora.org/now)
- **GloWbE:** 1.9 billion words web-based English (english-corpora.org/glowbe)
- **COCA:** Corpus of Contemporary American English (english-corpora.org/coca)

9.1.3 Multilingual & Parallel Corpora

- **OPUS:** Largest parallel corpus, 100+ languages (opus.nlpl.eu)
- **Universal Dependencies:** Treebanks for 100+ languages (universaldependencies.org)
- **Tatoeba:** Community sentences in 400+ languages (tatoeba.org)
- **OpenSubtitles:** Movie/TV subtitles, 60+ languages (opensubtitles.org)

9.1.4 Specialized & Literary Corpora

- **Project Gutenberg:** 60,000+ free books (gutenberg.org)
- **Wikipedia Dumps:** Full Wikipedia text (dumps.wikimedia.org)
- **Reddit Corpus:** Social media discussions (pushshift.io/reddit)
- **ArXiv:** Scientific papers, 270GB fulltext (arxiv.org/bulk_data)

9.1.5 Curated Dataset Collections

- **GitHub NLP Datasets:** Comprehensive list (github.com/niderhoff/nlp-datasets)
- **Hugging Face Datasets:** 70,000+ datasets (huggingface.co/datasets)

9.2 NLP Tools

- **spaCy:** Industrial-strength NLP (spacy.io)
- **Stanford CoreNLP:** Java-based toolkit (stanfordnlp.github.io/CoreNLP)
- **NLTK:** Natural Language Toolkit for Python (nltk.org)
- **Stanza:** Neural NLP pipeline (stanfordnlp.github.io/stanza)

This rubric is subject to clarification. Updates will be announced on Zulip.
Last updated: August 25, 2026