

# Syllabus tl;dr

## GNED 495: Computational linguistics

*Spring 2019*

Location: Ruffner 350, TR 2pm  
Website: <http://cs.longwood.edu/courses/cm318>  
Professor: Don Blaheta  
Office: Ruffner 334  
Phone: x2191  
Email: [blahetadp@blahedo.org](mailto:blahetadp@blahedo.org) or [blahetadp@longwood.edu](mailto:blahetadp@longwood.edu)  
Office hours: Mon/Wed 1:30-2:30pm; Tue 11-noon; Fri 2-3pm

### Textbook and resources

The textbook is *Language and computers*, by Dickinson, Brew, and Meurers (ISBN 978-1-4051-8305-5); also read Language Log at <http://languagelog ldc.upenn.edu/n11/> regularly.

Programs can be developed using Python 3 or any other language.

### Graded work

- Preparation and participation 5%
- “Lab” programming 10%
- Homework 10%
- Projects (see below) 15% each, total 45%
- Exams 15% each, total 30%

**Exam 1, Thu 28 February in class**

**Exam 2, Thu 2 May at 11:30am**

### Projects

In each project, you will identify a linguistic question, explain why it is sociolinguistically relevant or interesting, and then try to answer it, using computational tools that you access and build:

- First, you will formulate or reformulate the question into an explicitly quantifiable form;
- then scrape or acquire the data from an appropriate source, and write and document a program to process the data and answer the quantified question;
- and finally analyse and interpret that computational result to answer the original underlying question of interest.
- Throughout the process, you will construct small, exact test cases to test your program, as well as using your understanding of English and other languages to evaluate the larger results for correctness and reasonableness.

At the end of the project, you will submit the program itself and your data (or a link to it) but the primary artifact you will produce is a written report on your work. It should explain the full quantitative process in a way that would make the experiment reproducible; but it should also explain the conclusion in a way that would be understandable and persuasive to an interested audience that does not necessarily know or understand the intervening quantitative details.

## Specifics

Project 1: Stake a claim of the general form “X talks a lot/not much about Y.” X could be a person (such as Donald Trump or JK Rowling) or a group (such as Democrats or the Roman Catholic Church). Y is a topic. How will you defend and prove such a claim empirically and quantitatively?

Project 2: Answer a question of the general form “How real is prescriptive grammar rule X?” X should be an oft-cited rule (such as: no singular they, no split infinitive, which/that, no sentence-ending preposition, anything from Strunk and White—but in any case, give a citation for the rule!). You will use data from real and acknowledged-good writers (such as Jane Austen, Shakespeare, Dickens, Hemingway, or JK Rowling) to defend or debunk the rule. What will you count, and how will you make your argument?

Project 3: Explore data and draw a conclusion of the general form “Language learners from L1 X to L2 Y are particularly prone to make error Z.” X and Y are languages, and for practical reasons one of them is probably English. Z can be any linguistic error. What sorts of statistics would be used to draw that kind of conclusion? What sorts of data would you need to gather to support it—how would you find it, clean it, and determine its usability?