LING 334 / CS 396 - Introduction to Computational Linguistics

Week 1

What is computational linguistics?
Course overview and policies, Quest login

Who are we? - Instructor

Rob Voigt

robvoigt@northwestern.edu

Assistant Professor
of Linguistics

Who are we? - Teaching Assistants

Grace LeFevre

gracelefevre2026@u.northwestern.edu

PhD Student
in Linguistics

Chris Coleman

ccol@u.northwestern.edu

PhD Student
in Computer Science

Who are we? - Undergraduate Peer Mentor

Michelle Zhang

michellezhang2023@u.northwestern.edu

Senior CS Major, Ling/Math minors

Who is this class for?

- Linguists, social scientists, computer scientists
- Basic programming experience required!
- Basic understanding of probability necessary (or catch up)
- Applications:
  - research in linguistics or social science
  - research in CS / machine learning / neural nets
  - industry jobs and data science

Who is this class for? - other relevant classes

- CS 337 - NLP
- CS 348 - AI
- CS 349 - ML
- CS 397 - Seminar in Statistical Language Modeling
- CS 449 - Deep Learning
- CS 497 - Deep Learning for NLP

Who is this class for? - other relevant classes

- LING 330 - Stat methods
- Seminars and grad courses I teach down the road, e.g.

  Fall ‘23: Advanced Analysis for Computational Linguistics

When and **where** will we see each other?

Here in Pancoe Auditorium

recorded if you can't make it and for reference

Panopto on Canvas

Tuesdays	usually lecture, discussion
Thursdays	more mixed; can be lecture, group work, peer evaluation, extra OH

When and **where** will we see each other?

Office hours	*Rob*	Tuesdays 10-11am and by appt
	Grace	Thursdays 12:30-2:30pm
	Chris	Fridays 10am-noon
	Michelle	Mondays and Wednesdays 2-4pm

Ed discussion board for questions

Be involved, be interested, help each other out!

(quick demo)

What will we learn?

- Basic problems and algorithms for CL/NLP
- Programming and math (nothing too crazy)
- Historical trajectory, foundations for neural
- Focus on linguistic understanding and applications

What is computational linguistics?

The use of computational methodologies to understand language and language use at a large scale

The development of algorithms and statistical tools to allow computers to process human languages

Computational Linguistics (CL)

Natural Language Processing (NLP)

Engineering, Social Science, and beyond

Deeply interdisciplinary field - questions like:

How can we build a robot that can talk to us?

How do we organize and access all the information (written in human language) in books and on the internet?

How can computational models represent linguistic ones?

How does language use at a large scale demonstrate aspects of human psychology?

Historically Intertwined → Now Less Clear

Originated with Machine Translation in the 50s

Political/military purposes - translate Russian into English
(and vice versa)

"We'll have this solved in a decade." ... a few times.
Turned out to be a hard problem!

ALPAC Report, 1966 - heavily critical of the progress thus far

Historically Intertwined → Now Less Clear

MT introduced numerous layers of difficulty:

lexical, syntactic, pragmatic, contextual

Early work connected with theoretical linguistics more directly:

What is the computational complexity of natural languages?

Computational semantics -
defining and reasoning with
formal representations of meaning

(2a) *English:*
Ron ate a cookie in the Oval Office.

(2b) *NFLT Display Syntax:*

```
(SOME X5
  (COOKIE inst:X5)
  (EAT agt:RON ptnt:X5
                loc:OVAL-OFFICE))
```

(Creary and Pollard 1985)

Historically Intertwined → Now Less Clear

As CL / NLP got more applied (and useful), the link loosened:

1990s - beginning of statistical revolution in NLP
rise of machine learning, feature engineering

(field as a whole became much more empirical)

2010s - beginning of neural revolution in NLP
rise of vector representations of meaning

(core characteristic is uninterpretable features)

Sidenote of Social Importance

CL/NLP is probably among the most gender-diverse CS subfields

Many important women historical (and contemporary) figures!

Margaret Masterman

Early MT, notion of "interlingua"

Karen Spärck Jones

Early Information Retrieval, tf-idf

Unique Properties of Language (that make it hard)

Polysemy

7 always means 7;

‘Waldorf’ can be a hotel, or a school, or a salad, or a muppet

Ambiguity and Vagueness

We saw her duck … pet? dodgeball?

The western part of North America …Wyoming? Manitoba?

Unique Properties of Language (that make it hard)

Sparsity

Many ways to say “the same thing”

Q: Where is he?

He went to the store

Oh, Johnny left to get groceries

Out to grab the essentials

Nested / Recursive / Infinite

Stacey ate the candy that Naveen had found next to the adorable cat statue on the table that had been left out from last week’s party where Bill had unexpectedly proclaimed his love for Maurice and …

Unique Properties of Language (that make it hard)

These properties (among others)

are what make language

useful for humans

and **interesting to study**!

So in This Class? - **Basic Algorithms**

Algorithm?

Wikipedia - “finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation”

Methods for solving problems!

Tend to address some of these common challenges

The building blocks of the CL/NLP “thought process”

So in This Class? - **Programming Skills**

Translating conceptual understanding, pseudocode, and math into actual, working code that we can run

More on-your-own than LING300/CS110 etc

Still some scaffolding

So in This Class? - **Some Math**

Much of contemporary NLP relies on probability

If you have high school math, you can learn it

Probability primer by Sharon Goldwater on course webpage!
(very important if your probability is rusty)

So in This Class? - **Concepts and Applications**

Why are we doing what we're doing?

What linguistic phenomena motivate us to do it this way?

Applying methods to real-world datasets!

Error analysis! Ethical concerns!

What will we **not** explicitly cover?

(but you can learn if you're motivated!)

Sequence models

HMMs, CRFs, seq2seq, etc

Very important, needs its own class and more math

Implementation and training of neural models

We will cover foundations (vector semantics, classification) and have an applied assignment using these models

How will we learn it?

Course structure
and policies:

- Schedule
- Assignments
- Grading and Evaluation
- Agreements

Syllabus and Schedule

Syllabus and policies on course website:

https://faculty.wcas.northwestern.edu/robvoigt/courses/2023_spring/ling334/

Learning Structure - **Materials**

All available for free on course website

Starting point using Jurafsky and Martin's *Speech and Language Processing*

Learning Structure - **Assignments**

Out on Tuesdays, due the following Wednesday night

Largely programming, some qualitative aspects

Generally in-class peer review / support / evaluation at the beginning of class Thursdays

Learning Structure - **Assignments**

Structured flexibly to accommodate range of backgrounds:

- Core assignment - basics we want everyone to do
  - has corresponding autograder and your first goal is to pass all its tests.

- Extensions - various proposed ways to go beyond the basics and learn more
  - can include additional programming, qualitative analysis, reading papers in the field, or whatever you can dream up!

Learning Structure - **Assignments**

You can work on them anyway, anyhow

Your own machine, or on Quest, or on Quest Analytics, etc

But your assignments **must** run on Quest

Assignment 0 (very quick) due before class Tuesday - demo

Learning Structure - **In Class**

A few group work projects in class

Regexes Tuesday!

Regular peer evaluations (focus on understanding and style)

Questions are **always** welcome - please stop me!

... just know with ~100 people I may sometimes have to move on, follow up on Ed and in OH

Learning Structure - **Final Project / Assignment**

Very open to possibilities!

Do something self-directed, fun, exciting!

There will also be an alternative final assignment.

Will discuss in detail again about halfway through - but talk with me and the teaching team at any point about ideas.

Grading and **Evaluation**

Heavy emphasis on qualitative feedback where possible, peer support, self-directed learning

We will record completion, passing base auto-grader;
then provide qualitative feedback on extensions

Letter grades at the end based on effortful completion,
midterm and final self-evaluations (quick demo)

The point of this whole thing is for you to learn, period!

Flexibility

Class attendance is not strictly required

Deadlines are not strictly enforced

- however, if submitted greater than a week past the deadline, the teaching team will not review your work

Grading is (mostly) on your own scale relative to your own goals

This is a blessing and a curse! Prioritize appropriately.

What constitutes **strong performance**?

Effort and Engagement with learning.

Performance relative to *you*, not absolute performance.

Challenge yourself.

We have a very broad range of backgrounds and skill levels!

You are smart, you are adults -
We provide a structure, but it's ultimately on you.

What constitutes **strong performance**?

There is a lower bound:

> Do basic reading, complete core assignment (make it work and pass auto-grader)
>
> Be present, be engaged, be communicative

What constitutes **strong performance**?

There is no upper bound:

Each week will have extra “relevant readings”

Each assignment will have a number of possible extensions

You can start working early on your final projects

Plus whatever you can dream up

What constitutes **strong performance**?

There are no-nos:

Doing the “bare minimum” (relative to you)

Lack of communication

Academic integrity violations

Academic Integrity

I will attempt for this to be the only time I talk about this.

Here's the line:

Talking with fellow students about questions so you understand better and work through problems
❤️🎇👍 **¡AMAZING!**👍🎇❤️

Looking directly at others' work in order to copy parts of their code, taking credit for things you didn't do
💔💥👎 **HORRIBLE. HORRENDOUS. PLEASE NO.** 👎💥💔

Academic Integrity

Flexibility in structure does not mean flexibility on integrity.

Know that it can be very hard
to lie on a self-evaluation.

Cheating and dishonesty are so
unnecessary, and just so incredibly *lame*.

Don’t Waste Our Time,
Don’t Disappoint Dwight.

Agreements

I see this class as entering into a set of mutual agreements,
on top of the basic agreements of the university
(like academic integrity and so on)

We're building a community of learners interested in this topic!
(I'm a learner too.)

By registering, you agree to certain things -
By being the instructor, I agree to certain things.

You agree to:

Invest substantial time and effort in this course this quarter

Hold yourself accountable for your own progress

Be honest in assignments, self-evaluations

Stay on top of your work, and ask for help when needed

Be open to constructive feedback

Challenge yourself

Communicate with me when any of the above falls through

I agree to:

Invest substantial time and effort in your process of learning

Prepare well for class, construct meaningful assignments

Make myself available to help

Be open to criticism and commentary

Provide structures for learning

Communicate with you when any of the a
(it totally could, I have small kids at h

The key idea for me is **mutual respect.**

I respect your time, intelligence, integrity, and effort.

I ask you to respect our time, intelligence, integrity, and effort.

WE ARE HERE TO HELP

No such thing as a dumb question here.

We’re on the same team, this is not an adversarial relationship!

For next time!

Work through the setup in Assignment 0.

Read and get started on regular expressions.

Hooray! Very glad to have you in class.

Have a good weekend.