**CSCI 384 Computational Linguistics**
Fall 2025

Introduction and warm-up unit:

- General introduction to the course (**today**)
- Lab: Python and relevant libraries (Friday)

Today:

- Computational linguistics in the context of related fields
- Syllabus and course details
- Basic vocabulary of computational linguistics

Related fields

- Computational linguistics
- Natural language processing
  - Text processing
  - Speech processing
- Language technologies
  - Spelling and grammar correction
  - Machine translation
  - Speech to text and text to speech
  - Optical character recognition
  - Assisted writing
  - Information retrieval
  - Stylometry
  - Text classification
  - Conversational agents
  - Language learning tools
- Statistical language modeling
  - Large language models
- Machine learning

I. Traditional methods
   A. Introduction (2 days)
      - NLP terminology
      - Python
      - NLTK and other libraries
   B. Regular expressions (2 days)
      - Definition, rules, and limitations
      - Regex-based chatbots
   C. Edit distance (1 day)
   D. Information theory (2 days)
      - Concepts and terminology
      - Noisy channel model
      - Compression
      - Character-based language models
   E. Statistical language models (4 days)
      - *n*-grams and other language statistics
      - Lexical language models
      - Smoothing and interpolation
      - Applications and evaluation
   F. Part-of-speech tagging and hidden Markov models (4 days)
      - Lexical categories
      - Hidden Markov model definition
      - Viterbi algorithm
   G. Parsing (3 days)
      - Context-free grammars
      - CKY parsing algorithm

II. Machine-learning methods
   A. Basic training in machine learning (2 days)
      - Concepts and tasks
      - Bag-of-words model
   B. Naive Bayes classification and sentiment analysis (3 days)
      - Bayes's theorem
      - Sentiment analysis
   C. Stylometry and authorship attribution (3 days)
      - Concepts and problem statement
      - Techniques and practice
   D. Neural nets and related models (2 days)
      - Neural net basics
      - Recurrent neural nets
   E. Vector semantics and word embeddings (3 days)
      - The idea of word vectors
      - Word2vec and other approaches
   F. RNNS and Machine translation (4 days)
      - RNNs and LSTMs
      - Transformers
      - Machine translation
   G. Large language models (3 days)

**Academic Integrity.**

*For programming assignments,* collaboration among students enrolled in the course is permitted. Getting code solution to programming problems from the Internet or getting help on programming problems from anyone not currently enrolled in the course is strictly prohibited. For help on the programming language and libraries, students are encouraged to use the official documentation for those tools. For each programming assignment, students must keep track of any use of StackOverflow and similar sites, AI overviews of searched questions, and Copilot and similar tools. When they submit their solution to the assignment, they must include a README file reporting this use and reflecting on how these tools aided or hindered their learning on this assignment.

*For quizzes* that are administered through Canvas, students are *permitted but not encouraged* to reference their notes and official course materials; but students should **not confer with classmates or others** on quizzes.

*For reading responses,* students must submit original writing without using any form of generative AI. If students use AI tools for other purposes, such as reading an AI summary *in addition* to the assigned reading, they should report this in their response.

| | |
|---|---|
| **rhetorical** | larger-context meaning, persuasion, discourse |
| **semantic** | literal or immediate-context meaning |
| **syntactic** | structure, grammar |
| **morphological** | word forms and changes, prefixes and suffixes |
| **lexcial** | words |
| **phonetic/orthographical** | sounds / letters |

# Tokens, types, and lemmas

- Token
- Type
- Lexeme
- Lemma
- Wordform
- Corpus
- Vocabulary

# Tokens, types, and lemmas

*I rose to saw off the still rose that I saw still grew near the still.*

16 tokens.

Vocabulary: { *grew, I, near, off, rose, saw, still, that, the, to* }

Types occurring once: to, off, that, grew, near.

Types occurring twice: I, rose, saw, the.

Type occurring three times: still

Distinct lexemes/lemmas: I, rise, to, saw (verb-cut), off, the, still (adjective), rose (noun), that, see, still (adverb), grow, near, still (noun).

Coming up:

- Learn Python
- Take Python quiz (Thurs, Aug 28)
- Read J&M, Sections 2.(0–4) (Tues, Sept 2)
- Regular expressions quiz (Tues, Sept 2)
- Do Python warm-up assignment (Wed, Sept 3)

Next time: Introduction to Python and NLTK *in the lab*.