Computational Linguistics I

LING 4424 / CS 4744 / COGST 4240
Spring 2023

TR 11:25AM - 12:40PM
Goldwin Smith Hall 142

Mats Rooth
Office: Morrill 203A (entry through the Linguistics Department office)
Office hour: Wednesday 4:45PM-5:45PM, meetings also by appointment
Meetings in person

John Starr
Office: Computational Linguistics Lab, Morrill B07
Office hour: Friday 10AM-11AM, meetings also by appointment
Meetings in person

This course introduces computational methods and models in varied subfields of linguistics, including syntax, semantics, phonology, and phonetics. It covers formalisms and computational models that are relevant to theoretical linguistics, and nuts-and-bolts computational methods. The emphasis is on symbolic models. Neural models are covered in Natural Language Processing (CS 4740/LING 4474/COGST 4740) and Computational Linguistics II (LING 4434/CS 4745).

Prerequisites: Elementary Python (CS 1133 suffices) and
LING 1101 or CS 2800 or PHIL 2310

Topics

1. Tree syntax, Context free grammar, CF Parsing
2. Feature constraint grammar
3. Logical semantics
4. Generative phonology and Finite state transducers
5. Minimalist grammar and Multiple context free grammar
6. OT phonology (optimality theory)
7. HMM-GMM speech recognition (Hidden Markov models)
8. End-to-end system – truth values of spoken sentences

Requirements

Circa six problem sets 55%
Midterm prelim (covering 1-3 and the FST part of 4) 20%
Second prelim (covering 4-7) 20%
Participation in class and on forum 5%

Class attendance is obligatory. You may miss two classes without penalty, after that the penalty is 2 points per missed class. Email the instructor and the TA if you plan to miss a class, or otherwise miss one. Sign in at the start of class.

Please submit problem sets on time. There is a penalty of 5% per day for problem sets submitted late.

Letter grades are assigned on a curve, with the distribution typical in Linguistics classes. Scores on problem sets with unusually low means are scaled up linearly before scores are distributed.

Computational environment

In Computational Linguistics and NLP, Python is the language of choice for research and development, communicating ideas, and exchanging functionality. It is a basic teaching language at Cornell. You should have familiarity with elementary Python coming into the class.

Many lectures and problem sets use Jupyter notebooks. It will be necessary to install various packages, some requiring different versions of Python 3. Virtual environments (virtualenv) are a good way of dealing with this. Installation information will be distributed as we go. While not everything has been tried out in advance, we believe things will work on your laptops using the OSX and Linux operating systems, and under the Windows operating system with Windows Subsystem for Linux (WSL). As a backup, there is an Ubuntu Linux server `kuno.compling.cornell.edu` where you can get an account and do your coding. John Starr is the contact for setup on Windows and Linux, and Mats Rooth for OSX and Linux.

These are some of the toolkits we will use.

NLTK (Natural Language Toolkit) --- Python platform for working with natural language models and data. It includes nice graphics, e.g. for drawing trees, and works in Jupyter notebooks.

HFST (Helsinki Finite State Transducer Technology) --- Toolkit for the Finite State Calculus, which is a language of extended regular expressions including operations of intersection and complement (on top of the usual operations for regular expressions), and operations on relations in addition to sets. It works in notebooks. We use it for Phonology.

Parsers in Python for context free grammar, minimalist grammar, and multiple context free grammar from Edward Stabler and Peter Ljunglöf**.**

Kaldi and PyKaldi --- Toolkit for Hidden Markov modeling of speech signals.

Readings

Natural Language Processing with Python, by Steven Bird, Ewan Klein, and Edward Loper. Online.

Online documentation for NLTK.

Finite State Morphology, by Kenneth R. Beesley and Lauri Karttunen ($40 or less at Amazon or Abebooks). Used for computational phonology.

Lecture notes from Edward Stabler*, Computational Linguistics* (2013 version).

Speech and Language Processing, by Daniel Jurafsky and James H. Martin. Chapters from the 3rd edition are free online. A couple of chapters will be used.

Lecture notes prepared for the class.

Course mechanics and interaction

Basic mechanics such as announcements and homework submission are via Canvas and/or CMSX. We will use Ed to answer questions about the lectures, homework, and the project and other discussion. It is linked through Canvas. Please post your questions there, and answer questions when you can.

Academic Integrity

Your conduct in the course is governed by Cornell policy on academic integrity. A key idea is to attribute any sources or assistance you use. Assignments will state whether group work is permissible. When group work is not permitted, you *may not* look at solution code or solution text written by another student. It's permissible to discuss matters of interpretation or general strategy. This should preferably be done on the forum. It's permissible to solve a problem by finding the solution with internet searches, in a textbook, or in a technical article. If you do that, cite the source in your submission, but don't post it on the forum. In coding problems, it is permissible to use an AI coding system or AI-assisted IDE. If an AI made a substantial contribution to your solution, cite it in your submission. In parts calling for you to write sentences and paragraphs in English, it is not permissible to use text generation systems such as ChatGPT.

Lectures and course materials are copyrighted, you may not record them or convey them to note-taking services or the like.

Special Accommodations

Please give the instructor any Student Disability Services (SDS) accommodation letter as early as possible so that needed academic accommodations can be arranged. If you need an immediate accommodation, please speak with the instructor after class or email mr249@cornell.edu and/or SDS at sds_cu@cornell.edu. SDS is located on level 5 of Cornell Health, 110 Ho Plaza, 607-254-4545, https://sds.cornell.edu/.