Land Acknowledgement

McGill University is on land which has long served as a site of meeting and exchange amongst Indigenous peoples, including the Haudenosaunee and Anishinabeg nations. We acknowledge and thank the diverse Indigenous peoples whose presence marks this territory on which peoples of the world now gather.

COMP/LING 445: Computational Linguistics

Course Information

Instructor:	Roger Yu-Hsiang Lo (`roger.lo@xxx.yyy`)
Credits:	3
Time:	Tues/Thur 8:35 AM–9:55 AM
Location:	TBD
Online discussion forum:	TBD
Teaching assistant:	John Doe (`john.doe@xxx.yyy`)
Instructor office hours:	Wed 1:00 PM–2:00 PM @TBD
TA office hours:	Fri 11:30 AM–12:30 PM @TBD

Course Overview

Computational linguistics is a scientific discipline focused on understanding written and spoken language through computational methods. In the first half of this course, we will explore foundational concepts and methods in computational linguistics. In the second half, we will examine recent developments and common applications of this field.

We begin with an overview of the key issues before delving into regular expressions and finite state machines, which are not only useful for matching words but also for describing morphological patterns across languages. Next, we will study language modelling, a cornerstone of computational linguistics. We will then explore classification methods and basic model evaluation techniques in machine learning, as many language-related tasks can be framed as classification problems. Subsequently, the course will cover part-of-speech tagging, computational semantics—vector semantics in particular—which have broad applications in natural language processing (NLP). Later, we will focus on neural network-based techniques, the driving force behind most modern language technologies. The course concludes with an exploration of real-world applications of computational linguistics.

Learning Objectives

Upon completion of this course, you will be able to:

- Explain key concepts underlying current computational linguistic research;
- Implement algorithms or use existing libraries for common NLP tasks;
- Empirically evaluate the performance of computational models, including their errors.

Prerequisites & Restrictions

COMP 250 *Introduction to Computer Science* and MATH 240 *Discrete Structures* or permission from the instructor. Some background in linguistics at the level of LING 201 *Introduction to Linguistics* would be useful, though not critical.

Note that this course is not open to students that have taken or are taking COMP 445 *Computational Linguistics*.

Course Materials

Required

- Jurafsky, Daniel, and James H. Martin. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models*. 3rd edition. (Book draft)
- Lewis, Patrick et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474. (Full text)
- Xi, Zhiheng et al. 2023. The rise and potential of large language model based agents: A survey. (Full text)

Optional

- Goldberg, Yoav. 2017. *Neural network methods for natural language processing*. Morgan & Claypool. (Full text)
- Eisenstein, Jacob. 2018. *Natural language processing*. (Full text)

Course Format

Most class meetings will be lecture-based, with some lectures incorporating live-coding demonstrations. While you will be able to follow the lectures without completing the assigned readings and videos, I strongly encourage you to review them in advance to deepen your understanding of the topics. Lecture notes will be posted prior to class.

Assessment

- **Homework assignments** (60%; 10% per assignment): There will be six mandatory homework assignments designed to help you understand course concepts and techniques and apply them to practical problems. Assignments are **due on Friday at midnight** (see the Tentative Schedule & Topical Outline below for specific dates) and should be submitted electronically. Except for HW 1, you may work in groups of up to **three** people for each assignment. Late submissions will incur a 10% deduction for every 24 hours past the deadline. However, you have a five-day "grace period" for **ONE** assignment, allowing for a late submission without any point deduction (Life happens sometimes!). If you choose to use this option, please

clearly indicate it on your assignment. If you have an emergency, please reach out to the instructor as soon as possible.

- **Final project** (30%): The final project offers an opportunity for you to explore a topic in computational linguistics in depth. You may choose from the following project types:
  - **Independent research**: Extend techniques from a prior study or apply them to your own research;
  - **Error analysis**: Conduct a detailed linguistic analysis of errors produced by an NLP application, focusing on what they reveal about the underlying model;
  - **Literature review**: Survey and synthesize research on a specific theme in the field.

  If you are uncertain about the scope of your project, feel free to consult me at any time. Midway through the course, I will check in to discuss your plans.

  Projects can be completed individually or in groups of up to **three** people. The effort for group projects is expected to scale linearly with the number of group members. All group submissions must include a paragraph detailing each member's contributions.

  We will also offer a "default project" (on topic modelling) in lieu of a final project if you are unable to identify a topic of interest. While this alternative exists, I encourage you to take advantage of this opportunity to pursue a self-directed project. Computational linguistics is a booming and exciting field, and you are likely to find a topic that piques your interest!

- **Class attendance and participation** (10%): Attendance and participation will be assessed through in-class quizzes during lectures.

Grading Scale

Percentage grades will be assigned for all assessments and converted to final letter grades based on the scale published by the university:

Letter grade	% grade	Definition
A	85–100	Excellent performance
A−	80–84	
B+	75–79	Good performance
B	70–74	
B−	65–69	
C+	60–64	Satisfactory performance
C	55–59	
D	50–54	
F	0–49	Unsatisfactory performance (fail)

Communication

For course-related questions, please follow these steps for the quickest response:

1. Consult this syllabus.

2. Post your question on the online discussion forum or ask classmates.
3. Meet with me during office hours.

For personal questions, feel free to email me directly. I aim to respond within 48 hours.

Accessibility

- **Accommodation for students with disabilities**: Students requiring academic accommodations due to a disability or medical condition should reach out to Student Accessibility & Achievement. More information is available on this page.
- **Well-being**: Being a student at any level can be challenging. You should always prioritize your well-being if you experience physical or psychological difficulties. Please refer to Student Wellness Hub for resources provided by the university.

McGill Policy Statements

Academic integrity

McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures. (See McGill's guide to academic honesty for more information.)

L'université McGill attache une haute importance à l'honnêteté académique. Il incombe par conséquent à tous les étudiants de comprendre ce que l'on entend par tricherie, plagiat et autres infractions académiques, ainsi que les conséquences que peuvent avoir de telles actions, selon le Code de conduite de l'étudiant et procédures disciplinaires. (Pour de plus amples renseignements, veuillez consulter le guide pour l'honnêteté académique de McGill.)

Language of submission

In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French written work that is to be graded. This does not apply to courses in which acquiring proficiency in a language is one of the objectives.

Conformément à la Charte des droits de l'étudiant de l'Université McGill, chaque étudiant a le droit de soumettre en français ou en anglais tout travail écrit devant être noté, sauf dans le cas des cours dont l'un des objets est la maîtrise d'une langue.

Copyright

Use of generative artificial intelligence (GenAI) tools

You may choose to use GenAI tools as you work through the assignments in this course. However, you should be aware that the code/text generated by GenAI may by inaccurate, biased, or incomplete. You are ultimately accountable for the work you submit, and any content generated or supported by an artificial intelligence tool must be documented appropriately. The documentation should include what tool(s) were used, how they were used, and how the results from the GenAI were incorporated into the submitted work.

Extraordinary circumstances

In the event of extraordinary circumstances beyond the University's control, the content and/or assessment tasks in this course are subject to change and students will be advised of the change.

Tentative Schedule & Topical Outline

Wk #	Date	Topics	Readings	Due
1	08/29 (Thur)	**Introduction** - Course overview - Software set-up	- Install `Jupyter Notebook`	
2	09/03 (Tues) 09/05 (Thur)	**Computational morphology** - Regular expressions - Finite state machine - Edit distance	- SLP ch. 2	HW 1 due on 09/06
3	09/10 (Tues) 09/12 (Thur)	**Language modelling** - N-grams - Perplexity - Smoothing	- SLP ch. 3	
	Tuesday, September 10, is the add/drop deadline			
4	09/17 (Tues) 09/19 (Thur)	**Classification** - Naive Bayes - Logistic regression - Evaluation metrics - Data split	- SLP ch. 4, 5 - Video: Bayes theorem	HW 2 due on 09/20
5	09/24 (Tues) 09/26 (Thur)	**POS tagging** - Hidden Markov Model - CRFs	- SLP ch. 17	
6	10/01 (Tues) 10/03 (Thur)	**Computational semantics** - Vector semantics - Embedding - Similarity metrics	- SLP ch. 6 - Video: vectors	HW 3 due on 10/04

[continued on the next page]

Wk #	Date	Topics	Readings	Due
7	10/08 (Tues) 10/10 (Thur)	**Neural network** - Feed-forward networks - Backpropagation - Neural language models	- SLP ch. 7 - Video: NN part 1, 2, 3, 4	HW 4 due on 10/18
	~~10/15 (Tues)~~ ~~10/17 (Thur)~~	Fall reading break (no class)		
8	10/22 (Tues) 10/24 (Thur)	**RNNs & Transformers** - Vanishing gradients - LSTM - Attention - Layer normalization	- SLP ch. 8, 9 -Video: transformers, attention	
9	10/29 (Tues) 10/31 (Thur)	**Large language models** - Pre-training & fine-tuning - RLHF - Prompt engineering - Jailbreaking	- SLP ch. 10 - Video: LLMs	HW 5 due on 11/01
10	11/05 (Tues) 11/07 (Thur)	**Sentiment analysis** **Information extraction**	- SLP ch. 20, 22	
11	11/12 (Tues) 11/14 (Thur)	**RAG** **LLM agent**	- Lewis et al. (2020) - Xi et al. (2023)	HW 6 due on 11/15
12	11/19 (Tues) 11/21 (Thur)	**Automatic speech recognition** - CTC - Text-to-speech	- SLP ch. 16	
13	11/26 (Tues) 11/28 (Thur)	**Vision-language model**	- Blog: Intro to VLM	
14	12/03 (Tues)	**What's next**		Final project due on 12/20