Ling 8430: Information, Structure, and Language

Course Number: LING-8430 (Fall, 2024)
Time: Tuesday and Friday, 12:30 - 13:45
Location: Saint Mary's 120

Instructor: Ethan Wilcox
Office: Poulton Hall 248 (OH: Tuesday, 2:00 - 3:00 pm)
Email: ethan.wilcox@georgetown.edu
Website: https://wilcoxeg.github.io/

Summary: This seminar brings together two divergent perspectives on human language. On one hand, linguistics research seeks to describe the structures that underlie human communication systems, often using formal tools such as grammars and logics. On the other hand, research in computer science, in particular information theory, seeks to discover the optimal way to package and transmit information over a channel. This seminar will focus on the intersection between these two programs: To what extent are human languages optimized for efficient communication? Can structural features of human language, or human linguistic behaviors be analyzed using the toolkit developed for efficient information exchange? Topics covered will include the structure of the lexicon, the relationship between syntactic and statistical dependencies, pragmatic inferences, as well as various language processing phenomena. Students will gain experience reading and presenting research papers in this area, and implementing concepts from information theory in code. Students should be proficient in at least one programming language (Python or R), and familiar with basic concepts of probability theory and/or machine learning.

Learning Outcomes:

- Students will become familiar with basic concepts from information theory and statistics and how they can be applied to model linguistic phenomena, including syntax, morphology, semantics, pragmatics and language processing
- Students will gain practice reading and critiquing contemporary research papers on these topics
- Students will learn skills for writing, communicating and presenting these topics to groups
- Students will gain experience formulating novel research questions in this area

Syllabus:

Week	Module	Reading
Aug 26, 2024 (1 session, Aug 29th)	Introduction	
Sep 2, 2024 (1 session, Sept 5)	Introduction	MacKay, 2003, Information Theory, Inference, and Learning Algorithms Chapter 2
Sep 9, 2024	Lexicon, Morphology & Syntax	Piantadosi et al., 2011, Word lengths are optimized for efficient communication

Sep 16, 2024	Lexicon, Morphology & Syntax	Futrell et al., 2019, Syntactic dependencies correspond to word pairs with high mutual information Bonus: Hoover et al., 2021, Linguistic Dependencies and Statistical Dependence
Sep 23, 2024	Lexicon, Morphology & Syntax	Kopeling et al., 2017, The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort
Sep 30, 2024 Friday, Oct 4th will be remote	Language Processing	Gibson et al., 2013, Rational integration of noisy evidence and prior semantic expectations in sentence interpretation Bonus: Levy, 2008, A noisy-channel model of rational human sentence comprehension under uncertain input
Oct 7, 2024 Friday, Oct 11th will be remote	Language Processing	Levy, 2008, Expectation-based syntactic comprehension (Sections 1-5)
Oct 14, 2024	Language Processing	Smith & Levy, 2013, The effect of word predictability on reading time is logarithmic
Oct 21, 2024	Language Processing	Meister et al., 2021, Revisiting the Uniform Information Density Hypothesis
Oct 28, 2024	Language Processing	Futrell et al., 2020, Lossy Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing
Nov 4, 2024	Semantics & Pragmatics	Zaslavsky et al., 2018, Efficient Compression in Color Naming and its Evolution
Nov 11, 2024	Semantics & Pragmatics	Session 1: Degen, 2023, The Rational Speech Act Framework Session 2: Scontras et al., 2018, Probabilistic Language Understanding (Chapter 1)
Nov 18, 2024	Semantics & Pragmatics	Session 1: Scontras et al., 2018, Probabilistic Language Understanding (Chapter 2) Session 2: Scontras et al., 2018, Probabilistic Language Understanding (Chapter 3)
Nov 25, 2024 (1 session; Nov 26th)	Abstract Workshop	
Dec 2, 2024	Conclusion & Project Presentations (Friday)	Futrell and Hahn, 2022, Information Theory as a Bridge between Language Form and Language Function (no presentation, roundtable discussion)

		Bonus: Gibson et al., 2019, How Efficiency Shapes Human Language
Dec 9, 2024 (1 session; Dec 10)	Project Presentations	

Weekly Structure: Each week of class will be organized around a paper, and divided into two sessions

- **Session 1 (typically Tuesday):** ***Workshop.*** This session will introduce concepts from information theory and linguistics that are important for the week's reading. The workshop will be divided into two parts:
    - **Technical Presentation (~45 minutes):** A student will introduce a technical concept from information theory or machine learning that is relevant to the weekly reading, and present a piece of code that illustrates the concept.
    - **Linguistics Concept Presentation (~30 minutes):** A student will introduce a linguistics or psycholinguistics concept that is important to this week's paper, and present a worked example (or examples) that illustrates the concept.

  Presentations are expected to scale with the student's previous experience in linguistics and computer science. I will provide a list of technical and linguistic topics for each class. If I think that a topic is essential for understanding the reading, I will underline it, in which case it must be chosen that week. Otherwise, you are free to choose whichever you like or to propose your own topic, with enough time in advance.

- **Session 2 (typically Friday):** ***Paper Presentation*.** A student will present the paper (~1 ¼ hour). Presentations are expected to walk through the major scientific components of a paper (scientific theory, research question, methods, results, impact), giving an overview of each component, and leaving lots of time for group discussion. Presenters are encouraged to include open-ended questions that can stimulate group discussion. See below for paper presentation guidelines.

- **Sign-up Sheet Link**

N.B. that for the weeks spent on the Rational Speech Act (2nd and 3rd weeks of November), we will not follow this schedule and instead work through the first three chapters of "Probabilistic Language Understanding" by Scontras, Tessler and Goodman (2018). Presentations for these sessions can count as either a main paper presentation or a combined technical / linguistics presentation (as presenters will use the full class time during these presentations).

Prerequisites:

- **Coding Experience:** Students are expected to have a basic understanding of at least one programming language (ideally Python or R).
- **Probability Theory:** This class will involve reading papers that involve lots of probability theory. Ideally, students should have prior experience with random variables, joint and conditional probabilities, expectations of random variables, and Bayes' rule. I will be providing a quick recap

of these topics during the first few sessions. If you haven't encountered these concepts before but want to take the class, talk to me, and I can recommend some additional readings that will help you get up to speed.
- **Machine Learning:** Students should have a basic conceptual understanding of machine learning models (i.e., understand how to interpret their inputs and outputs) but are not required to implement them in this class.

Course Requirements:

- **Attendance:** Students are expected to attend every session. If you can't make a session due to medical or family reasons, please contact me in advance (ideally, at least one week in advance). In class, students are expected to contribute actively to group discussions.
- **Readings**: Students will be expected to complete the weekly readings. It's expected that you have a look at the paper prior to the workshop session (typically Tuesday), and have an in-depth read of the paper prior to the discussion session (Friday).
- **Paper Presentation:** Students will be required to present one paper in class. Presentations should be about 70 minutes long, including group discussion. (So bring discussion questions!) Presentations should cover the following:

    - What is the main scientific hypothesis of the paper?
    - What have previous studies/papers concluded about this hypothesis, or the over-all research topic?
    - What methods did the authors use to answer their hypothesis?
    - What dataset did the authors use?
    - What was their computational or mathematical model?
    - What were their results? What were their conclusions?

  Students should include at least one slide (or section) of their presentation that involves a critical analysis of the work: Did the author's conclusions follow from their results? What are the limitations of the paper? How does the paper change (or not change!) our understanding of human language?

- **Technical Workshop Presentation:** Students will be required to present one technical concept from information theory, and produce a novel piece of code that illustrates this concept. The code should be easily runable in a notebook environment (like Jupyter, Colab, or R markdown). For example, a student could give a mathematical introduction to surprisal and a notebook that plots the surprisal "curves" of words in a sentence. As another example, a student could give a mathematical introduction to Entropy and provide a notebook that visualizes the Entropy of different random variables. The presentation should last about 30 minutes, after which we will all download the code and play around with it in a sandbox environment for about 15 minutes.
    - In order to view R Markdown code, please install R Studio and Jupyter in advance of the second session, if you do not have these already installed on your computer. Please bring laptops to class!

- **Linguistics Workshop Presentation:** Students will be required to present one concept from linguistics or psycholinguistics relevant to the week's reading, along with a worked example that illustrates this concept. Examples include: What is a pronomial system and how do they change cross-linguistically? How do we measure reading times experimentally? The presentation should be about 30 minutes.
- **Research Paper** (Due: Dec 16th AOE) Students will be expected to write a final paper. Although not required, the research paper can reproduce and extend one of the papers discussed during class. Group work is encouraged, although the scope of the research paper should scale with the size of the group. As part of the paper process, the course will include the following:

    - **Paper proposal** (Due: Nov 12th before class): Students must send me a brief (<1 page) proposal for their research topic. Those working in groups can send me one single proposal for the group project.
    - **Abstract workshop** (Due: Nov 26th in class): Students will bring a short, 300-word abstract to class. In class, we'll read each others' abstracts and provide feedback. Each person must submit an abstract that they have authored, meaning that groups will submit multiple abstracts per group. This is to give each person in the class practice writing abstracts and to make sure that each person feels comfortable articulating the main goals, methods, and results of their project.
    - **Research project presentations** (Due: Dec, 5th or 10th, depending on your slot): Students will share their research project and findings with the class. Projects can still be in progress at this point; this is an opportunity for feedback. However, during the presentations groups must clearly articulate the following: What are your research questions? What are your hypotheses? What methods will you use to answer your question? What datasets and/or models will you use (if relevant)? In the presentation, students will be expected to comment on how their research topic relates to the material covered in the course.

I will use the following rubric for determining final grades:

Research Paper	30%
Paper proposal	5%
Abstract workshop	5%
Project Presentation	10%
Paper Presentation	20%
Linguistics Presentation	15%
Technical Presentation	15%

Accommodations: Please contact me if you have a recognized accommodation through the ARC.

AI Policy: In general, the use of AI assistants and coding assistants is allowed in this class. If you want to use Generative AI assistants to summarize papers, explain concepts, generate ideas, produce rough drafts of slides, or proofread material., that's OK. The two exceptions are:

1. You are expected to read the assigned papers
2. You are expected to write abstracts and final papers