Ling 8430: Information, Structure, and Language

Course Number: LING-8430 (Fall, 2026)
Time: Monday and Wednesday, 12:30 - 13:45
Location: MacGuire 102

Instructor: Ethan Wilcox
Office: Poulton Hall 248 (OH: Monday, 11:00 am – 12:00 pm)
Email: ethan.wilcox@georgetown.edu
Website: https://wilcoxeg.github.io/

Summary: This seminar brings together multiple 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. Recently, deep-learning based AI models have demonstrated exceptional abilities at producing fluent, humanlike outputs, challenging longstanding beliefs about the computational mechanisms required to learn language. This seminar will focus on the intersection between these structural, informational, and neural-network–based approaches. 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 organization of the lexicon and words, how languages are processed, and how linguistic structures are represented inside neural networks. Students will gain experience reading and presenting research papers in this area. Students should be 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 communicating and presenting these topics to groups

Syllabus: The class will be divided into four modules: An introductory module; a module focusing on information-based modeling approaches; a module based on artificial neural network (ANN) modeling approaches; and a module where each student presents a paper of their choice.

Week	Module	Reading
Aug 31, 2026 (1 session, August 26th)	Introduction	**Session 1:** Welcome and course overview (Ethan)
Aug 31, 2026	Introduction	**Session 1:** The parts of a CL, NLP, and CogSci research paper (Ethan; no reading)

		Session 2: Experimental methods in CogSci, NLP, and CL (Ethan; Reading: *MacKay, 2003, Information Theory, Inference, and Learning Algorithms Chapter 2*)
Sep 7, 2026 (2 sessions; Sept 7th is a Monday schedule)	Information-based Approaches	**Session 1:** Piantadosi et al., 2011, Word lengths are optimized for efficient communication [form] **Session 2:** Kopeling et al., 2017, The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort [form]
Sep 14, 2026	Information-based Approaches	**Session 1:** Mansfield, 2022, The word as a unit of internal predictability [form] **Session 2:** Gibson et al., 2013, Rational integration of noisy evidence and prior semantic expectations in sentence interpretation [processing]
Sep 21, 2026	Information-based Approaches	**Session 1:** Levy, 2008, Expectation-based syntactic comprehension (Sections 1- 4 in detail) [processing] **Session 2:** Futrell et al., 2020, Lossy Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing (Sections 1 - 3 in detail) [processing]
Sep 28, 2026	Information-based Approaches	**Session 1:** Shain et al., 2023 Large-scale evidence for logarithmic effects of word predictability on reading time [processing] **Session 2:** Giulianelli et al., 2026 Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction [processing]
Oct 5, 2026	ANN-based Approaches	**Session 1:** Millière, 2025, Language models as models of language (no paper presentation) **Session 2:** Oh and Schuler, 2023, Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? [processing]
Oct 12, 2026 (1 Session; Oct 14th)	ANN-based Approaches	**Session 1:** Wang et al., 2026, Simulating Human Memory with Language Models [processing]
Oct 19, 2026	ANN-based Approaches	**Session 1:** Botch and Finn, 2026, Sensory context as a universal principle of language in humans and LLMs [processing] **Session 2:** Boguraev et al., (2025), Causal Interventions Reveal Shared Structure Across English Filler–Gap Constructions [representation]

Oct 26, 2026	ANN-based Approaches	**Session 1: TBD** **Session 2:** Padamanabhan et al.,, 2025, On language models' sensitivity to suspicious coincidences [acquisition / bias]
Nov 2, 2026	ANN-based Approaches	**Session 1:** Bencomo, Gupta, et al., 2025, Teasing apart architecture and initial weights as sources of inductive bias in neural networks [acquisition / bias] **Session 2:** Goodale and Mascarenhas, 2026, Fodor and Pylshyn's systematicity challenge still stands [acquisition / bias]
Nov 9, 2026	Student Choice	**Session 1:** **Session 2:**
Nov 16, 2026	Student Choice	**Session 1:** **Session 2:**
Nov 23, 2026 (1 session, Nov. 23rd)	Student Choice	**Session 1:**
Nov 30, 2026	Student Choice	**Session 1:** **Session 2:**
Dec 7, 2026 (1 session, Dec 7th)	Student Choice	**Session 1:**

Course Requirements

- **Reading:** Students should complete the readings for each session. I find that reading is always more fruitful when done in an active manner, i.e., by highlighting, taking notes, and writing comments in the margins.
    a. For the final module, enrolled students will propose and present a paper of their choice. Please send me proposals for your paper by October 20th (although sooner is fine too). Papers should connect with a main theme of the class, but I am open to a wide range of topics and methods 🙂

- **Reading Responses:** For each reading, a reading response should be submitted on Canvas 2 hours before class (i.e., 10:30). Ethan will compile, anonymize, and share excerpts from responses as a basis for class discussion. Responses can be short (i.e., ~150 words or so) and should be divided into three parts:
    a. *Clarification / Understanding Questions:* Indicate any part of the reading that you didn't understand or wasn't sufficiently explained. It's expected that you will not fully

understand a paper the first time you read it, so please don't be shy! My guess is that if you don't understand something, other's don't either, and all will benefit from a discussion about it during class :)

b. *Discussion Questions:* Indicate at least one open-ended or discussion question you have based on the reading.
c. *Opinionated Response:* Please submit a short paragraph (<100 words) that contains both a critique and constructive comment based on a specific section of the paper (e.g., the methods section, the research question, literature review)
    - For the critique: List 2-3 shortcomings or weaknesses of the section. Did the authors make any moves that you believe were not sufficiently justified? What were the limitations of their methods? Models? Data?
    - For the constructive feedback: List 2-3 ways that this (or a different section) could be strengthened. For example: What additional data would make the argument more convincing? What statistical tests could be employed for a more robust analysis?

- **Paper Presentation:** Each student will be responsible for presenting three papers, one in each module of the class. The purpose of the presentation is to structure group discussion, not to provide a comprehensive overview of the entire paper. Presentations should be made in the class' **shared google slides document** and cover the following:

    - Research Question: What is the main scientific question and hypothesis of the paper?
    - Background: What have previous studies/papers concluded about this hypothesis, or the over-all research topic?
    - Methods: What methods did the authors use to answer their hypothesis?
        - Psycholinguistic methods: How was human data collected?
        - Datasets: What dataset did the authors use?
        - Computational Models: What was their computational or mathematical model?
    - Results: What were their results? **Please include a slide copying any major or important figures!** (This is extremely useful for group discussion)
    - Conclusions: How do their results answer the original research question?

- **Attendance and In-Class Discussion:** Students are expected to attend and participate in class discussions.

    The structure of each session will be roughly:
    a. Review clarification questions
    b. Paper presentation
    c. Group discussion based on submitted discussion questions and opinionated responses

Prerequisites:

- **Coding Experience:** Students are expected to have a basic understanding of at least one programming language, which is necessary for understanding computational methods and models.

- **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.

Grading: I will use the following rubric for determining final grades:

Paper Presentations	45% (15% each)
Reading Responses	20% (~1% each)
Participation	20%
Attendance	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, however you are expected to read the assigned papers. If you want to use GenAI to summarize papers, explain concepts, generate ideas, or produce rough drafts – that's all OK. However, all submitted work should be edited, reviewed and checked by you.