

What is Natural Language Processing (NLP)?

- Computer analysis of a natural language (wait, what's a natural language?)
 - answers the question “what is property z of language segment x?”
- Computer generation of a natural language
 - given context c (which includes instruction), produce appropriate natural language segment x
- Since this is all AI, we assume that a human can do the analysis and generation tasks we’re asking a computer to do; goal is to be suitably indistinguishable
- Could we extend this to definition to a **superhuman** use/understanding of language? What is fair game and what isn’t?

Let’s talk about “LLMs” (GPT/Gemini/Llama/Claude/...) and the impact they have on this class

What do we want NLP to do? In the age of LLMs are these “solved problems”?

- ...you tell me!
- Chatbots
- Customer service agents
- Question Answering (search)
- Translation
- Grammar Correction
- Summarization/Analysis
- Writing in general, with lots of subcategories
 - article
 - patent
 - report
 - fiction
 - (generally) information delivery
 - (generally) emotional response evoking
- Article/patent/report/fiction
- Opinion/sentiment analysis
- Essay evaluation/plagiarism detection
- Knowledge base construction
- Author identification

What is the purpose of teaching computers to do things that humans can do?

How might NLP/CL research help us learn about language?

What will you learn in the class and how will you learn it?

- You are likely already good at class taking; as a PhD student, what’s most important now is learning how to conduct research in a complicated topic and then to write and speak at an academic level about that research. My job is to introduce you to tools you need to

get started in the topic (here, NLP) and to guide you in the research, writing and speaking.

- Here is the general approach to the class:
 - The first part will cover the basic tools you need to understand the majority of current and slightly older research and some experimentation and data fundamentals. You may have experienced parts of this in other ML/AI classes, but there will generally be an NLP “spin.”
 - Thereafter we’ll discuss the dominant NLP “fields” and “tasks” that are generally the settings for research products. There are various special topics, especially near the end, and there is some fluidity in what we cover so if you don’t see a topic you wanted to learn more about, let me know!
 - Starting roughly in October, each class will begin with two 15 minute presentations students on a paper from this year’s NAACL or ACL conferences and a discussion of the paper.
 - We’ll take a break at about 11 each class, have a snack and some open discussion or just check email (I prefer the open discussion when we can).
 - The class works better when you’re engaged, interrupting me, asking questions, challenging my assertions, bringing up extra points from your own experience, etc.

Who am I?

- Research Associate Professor based at ISI (in Marina del Rey) – means I do more research, less teaching
- areas of interest: machine translation, multilinguality, low-resource issues, dialogue, game playing, automata theory, transfer learning
- I got my PhD here in 2010, I’ve been working here since 2014.

What other professor-types working at USC work in this space (there may be more than I have here so please correct me)?

- Mostly on campus
 - Xiang Ren
 - Shri Narayanan
 - Robin Jia
 - Jesse Thomason
 - Swabha Swayamdipta
 - Dani Yogatama
 - Jieyu Zhao
 - Morteza Dehghani
- Mostly at ISI (MdR)
 - Xuezhe Ma
 - Jay Pujara
 - Ulf Hermjakob
- Mostly at ICT
 - David Traum

- Kallirroi Georgila
- Jonathan Gratch
- Andrew Gordon
- Ron Artstein
- Mostly at ISI (Boston)
 - Elizabeth Boschee
 - Marjorie Freedman
 - Scott Miller
 - Ralph Weischedel

Who are you? In small groups of 2-3 find out about each other, then we'll reconvene and you'll introduce **each other** to the rest of the class. Here are some things to tell us about the person you are introducing:

- What they like to be called
- What their research interests are
- What they find interesting about language, linguistics, NLP, etc
- Anything personal they feel comfortable sharing with the class (default: what your favorite bridge is – can't say golden gate)

More logistics!

- If you stay actively engaged (ask questions and contribute to class and on piazza/slack, get your work in on time, apply recommendations to how to improve in HWs/project) it's an easy A.
 - There is no punishment curve and there is always a reward curve (I have no problem with everyone getting an A if everyone deserves it)
 - If you are concerned about your grade, come and see us at office hours and we can assess how we think you're doing
 - Note the drop dates and how they mesh with assignment due dates; you can do an assignment early and have us look at it if you're concerned and a deadline is approaching. See the website (<https://jonmay.github.io/USC-CS662/>) but do point out any errors
- Reading
 - Partly from the two (freely available) textbooks, maybe also some papers relevant to lecture
 - NLP moves fast – probably you don't want to buy any books
 - Eisenstein is in print but it was printed in 2019 and released in 2018; there's barely any transformer in it!
 - Jurafsky and Martin 3rd edition has been written and rewritten continuously; they no longer have an anticipated publication date!
 - Papers that are presented in class
 - Expected to be read before the class
 - I may ask questions, socratically (i.e. I call people out by name) – be prepared!
- Evaluation
 - Homework (30%)

- Three of them, each 3 weeks long, each 10% of grade
 - In each, you will write the code for, train, and evaluate one or more NLP models. You will try to make the models better using ideas from literature
 - Write the code in python
 - Write the report in LaTeX
 - **Communication is most important** – reports should resemble a paper, addressing the research questions, how they were addressed, what the finding was. Tell a story of the research (not a log of the tasks you did!)
- Project (40%)
 - Proposal (5%), Report v1 (5%), presentation (10%), final report (20%)
 - Reproduction of a 2023 EMNLP paper (see project handout for details)
 - Must form groups – scoring is done for the whole group. Group size and make-up dictated by the professor (the general principle is that you must team with someone who has a different program/focus area background than you)
- Class participation (30%)
 - Coming to class, asking questions beyond “assigned” questions (see below), engaging on piazza in useful ways (10%)
 - Assigned questions before in-class selected paper presentation, assigned questions during final project presentations (5% each)
 - Your in-class paper presentation (10%)
 - Presentation of a 2024 NAACL/ACL paper in 15 minute presentations (details to come)
- Deadlines
 - Code and written work for homeworks, project proposal and report v1 due 23:59:59 AoE on announced due date
 - 20% of the **possible** total off for every calendar day (AoE) thereafter; don’t bother if it’s ≥ 5 days late
 - For these items **only** you have four cumulative extension days to be used as you like
 - If a project milestone is late, all group members are assessed extension days (or lose score)
 - Final report is due at the announced final exam period time (check website) – no late days allowed
 - In-person elements must be done in-person
- Extra credit vs extra mile
 - There is no extra credit in this class
 - You cannot get full credit on homeworks or the project without “going the extra mile.” This means trying something novel, in addition to the requirements of the assignment (and along the same lines; examples will be suggested). It’s not a box to tick – we will judge how much of an extra mile your ‘extra mile’ contributions are.
- Ethical participation in class
 - Read the [Viterbi Code of Integrity](#); it applies, and I will abide by it.

- Respect others' opinions and backgrounds; honor requests and understand that you may cause offense without intending to – you still caused offense and you need to own and understand that. I have my own biases and may offend. It's up to me to understand when I've done that and be up front about it but if you want to call me out on that I very much appreciate the assistance!
- All code and all writing must be your own and not copied from the web or any other person. Attempts to modify others' code or writing by changing small elements are easier to detect than you may think
- But what about writing tools? Here is our current policy:
 - Spell check, grammar check (e.g. grammarly), template-based code generation and completion and other low level tools are fine to use and need not be mentioned.
 - **Generative language, code, and vision models** (e.g. ChatGPT, Llama, Midjourney, Copilot, etc.; if you are unsure, ask and don't assume!!) **can be used with the following caveats:**
 - You must declare your use of the tools in your submitted artifact. **If you don't declare the tool usage but you did use these tools, we will consider that plagiarism**
 - For code and image generation, you must indicate the prompt used and output generated
 - **Please don't use generated images in slides just for "style"** it is extremely distracting and instantly recognizable. Obviously if the work is about generation or even if you are using generation to make icons/tools you are using it's ok but as a space-filler, better to find clip art (or even have a wall of text)
 - For text generation you must provide either a link to the chat session you used to help write the content or an equivalent readout of the inputs you provided and outputs received from the system.
 - **You will lose credit if "the AI" is doing the work rather than you. Here are some examples:**
 - bad prompt (plagiarism): "Here is a paper <contents of paper>. Here are the report assignment instructions <contents of assignment>. Write a report about the paper, following the instructions.
 - better prompt (not plagiarism, but possibly not full marks): "Here are the instructions <instructions>. Here are my rough ideas <bulleted list>. Write a report from the ideas that follows the instructions.
 - good prompt (full marks): "Here are the instructions <instructions>. Here is a draft of the report <report v1>. Improve spelling and grammar and writing so that it is compliant with the instructions and is interesting to read."

- Everything you do in this class must be newly done for the class; you can't submit work you submitted elsewhere here. That's self-plagiarism (which is a form of plagiarism).
- If we have determined you have violated the honor code we will invoke punishments as deemed necessary; this can mean a zero on an assignment, a reduced letter grade in the class, and/or even a failing grade. Punishments can occur at any time after violations (usually they will occur at the worst possible time). I hate doing this but I will if necessary (ask around) – the consequences can be very severe!
- Office hours
 - Check website for the latest. Jon generally tries to have one hour on class days, usually right before or after class.

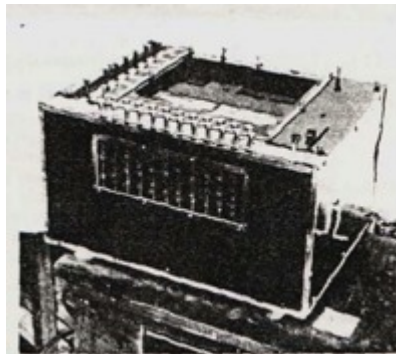
Connections to other areas

- Machine Learning -- the biggie, now. NLP: ML is a tool we use. ML: NLP is a data set we use.
- Linguistics -- ML folks might think of this as 'guided search' both within a problem and when considering what the problems are. But also consider we are trying to understand language and can use NLP/CL techniques to do so.
- Cognitive Science / Psychology -- see also Linguistics, but a level up. We are fascinated by humans' ability to learn language that is far better than computers' and we're not quite sure why this is. How do theories of learning in humans correlate with theories of humans in machines? If an adequate model can be built, that avoids the trouble that comes with doing HSR (especially when studying learning, which is often on young children) but no model is ever accurate and all are biased.
- Philosophy, especially as it relates to psychology/linguistics: What does it mean for an entity to "understand" language? This and related topics are a raging debate
- information theory -- language is a means by which humans communicate and the communication capacity/compression/confusability of communications is baked in to our studies, particularly when discussing (cross-)entropy and mutual information. How much redundancy is naturally in our language stream? How much needs to be there? Does it matter when we shift modality?
- theory of computation -- very important for search over complex spaces (e.g. for a syntactic tree or semantic graph) and for recognizing what transformations are and are not possible. Strongly connected via chomsky hierarchy, of which we'll only talk a little
- data science, political science, etc. -- NLP: these are good subdomains to try our stuff. X science: NLP is a good tool to demonstrate my theories. Currently looking at using NLP to recognize and combat (e.g. state-sponsored) influence campaigns, en masse analysis of traits of different populations online, detection in bots.
- ethics
 - what threats exist?
 - who benefits and who loses?
 - what safeguards should be put in place?

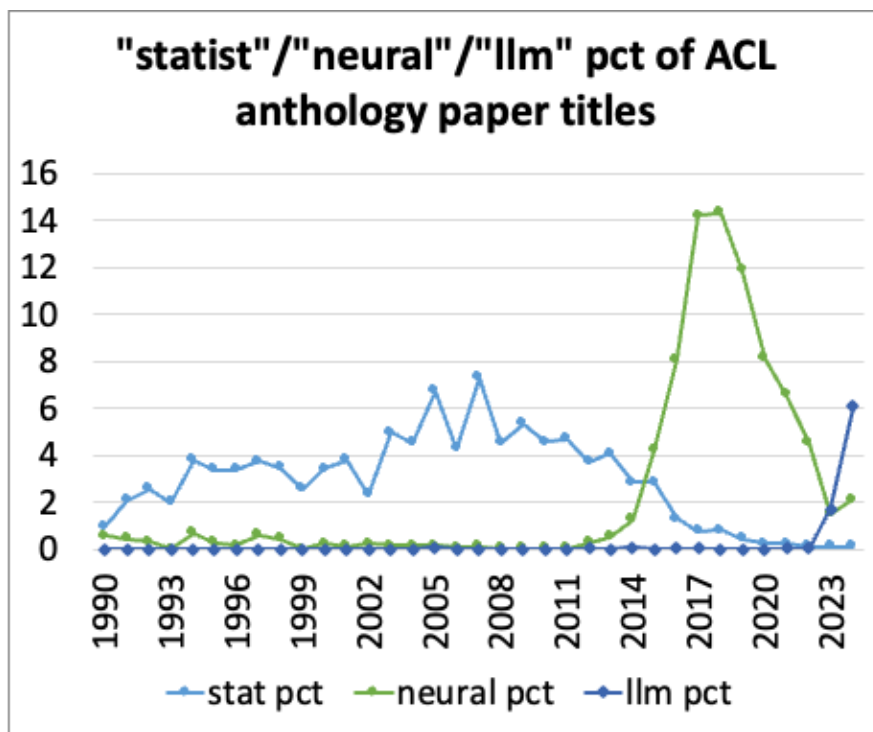
- what freedoms should be preserved?
- Other areas you're interested in not covered here?

Super Brief and Selective History of NLP:

- NLP was indistinguishable from MT for a long time for some weird reason
- Athanasius Kircher -- 'polygraphy' (1630) for translation without understanding
 - side note: kircher is really weird guy! he
 - incorrectly deciphered hieroglyphics
 - Went into Vesuvius' erupting crater for his geology research
 - invented a cat piano
 - check him out at the Museum of Jurassic Technology just outside Culver City (in 2025 due to other exhibits)
- Becher mechanical meta-language for language-to-meaning (in the service of NLP): 1666
- Georges Artsruni mechanical brain (1930)
- Computers proposed for translation (1949...3 years after computer invented)
- ACL founded (as the ``Association for Machine Translation") in 1962; produced ACL conference (since 1962 but records at <https://aclanthology.org/> go back to 1979) and CL journal (since 1975). EACL started in 1983, NAACL in 2000 but merged with HLT which started in 1986. EMNLP started in 1996; there are others too!



- The US defense department has long had a desire to gain situational awareness, know what the Russians were saying, etc. So DARPA/ARPA/IARPA has always been part of the underlying funding and support of this work.
- Government programs tried (and still try) to get researchers to solve very complicated end-to-end tasks such as “extract all useful knowledge from ship-to-ship communications.” We have been breaking them down to their component parts, this leads to the tasks I’ll describe next
- Through the mid-1980s a CL/NLP paper would be a description of a system or a grammar or formalism along with, perhaps, some examples of what it could do/represent. Rule-based, by thinking carefully about the various ways, e.g. “a psychiatrist would respond to a paranoid patient” or “interpret this word problem about physics”
- Starting in the mid-1980s more attention was paid to corpus-based or statistical approaches that used data, learned probabilistic models, and actually started to work well on *unseen* data.
- Starting in the mid-2010s we were able to use GPUs to seriously train neural networks, getting often really nice results, and attracting more attention from big business. As data and compute availability have grown we can increasingly complete larger end-to-end tasks.
- In November 2022, ChatGPT was publicly released, drastically increasing public awareness of the field, which as of this writing (updated fall 2024) has implications in labor rights movements, political campaigns, and seemingly every other aspect of modern life. [It is very weird to see nonstop ads for your niche research on the olympics]



So can't we just [use](#) ChatGPT to do everything we want to do?

🔮 GPT-4



Create a list of three reasons why I want to study natural language processing. Use no more than 20 words per reason.



1. Unlock insights from vast textual data; drive informed decision-making.
2. Improve human-computer interaction; create intuitive systems.
3. Explore linguistic patterns; advance artificial intelligence research.



2022, with GPT 3:

Playground

If x is twice the square of 20 percent of y, and y is 45, what is x?

x is the square of 40.

[2023](#), with ChatGPT (uses Chain of Thought)

JO If x is twice the square of 20 percent of y, and y is 45, what is x?



Let's break down the problem step by step.

1. First, let's find 20 percent of y:
 $20\% \text{ of } y = 0.20 * y = 0.20 * 45 = 9.$
2. Now, let's square this value:
 $(20\% \text{ of } y)^2 = 9^2 = 81.$
3. Finally, x is twice this squared value:
 $x = 2 * (20\% \text{ of } y)^2 = 2 * 81 = 162.$

So, when y is 45, the value of x would be 162.

2023 (trust me, it's wrong, try adding 23098757832900183773 and 198796398893701)

JO What is 23098757832900183773 + 198796398893701?



The sum of 23098757832900183773 and 198796398893701 is:

$$23098757832900183773 + 198796398893701 = 23297554231793987474.$$

2024 (uses tools):

what is 23098757832900183773 + 198796398893701?



The sum of 23098757832900183773 and 198796398893701 is 23098956629299077474. [~]

Every year since ~2017 I have given some form of this “here are things we can’t do” and it keeps getting harder and harder! Here is a failure case in 2024:

how many "r"s are there in strawberry?



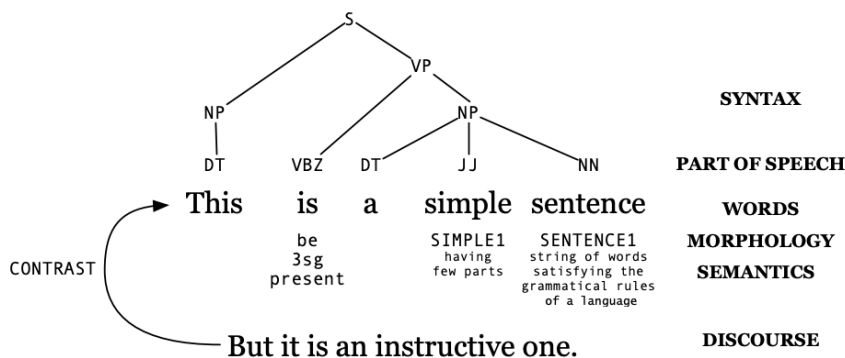
The word "strawberry" contains **two** "r"s.



(upon further prompting it can be made to realize its error)

Why is NLP hard and interesting? Ambiguity. It exists at all levels, so let's discuss those levels.

Breakdown of parts:

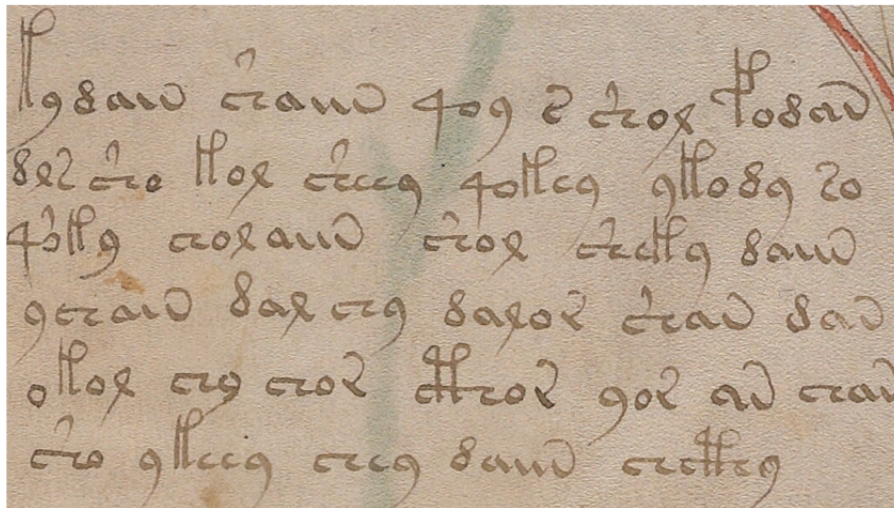
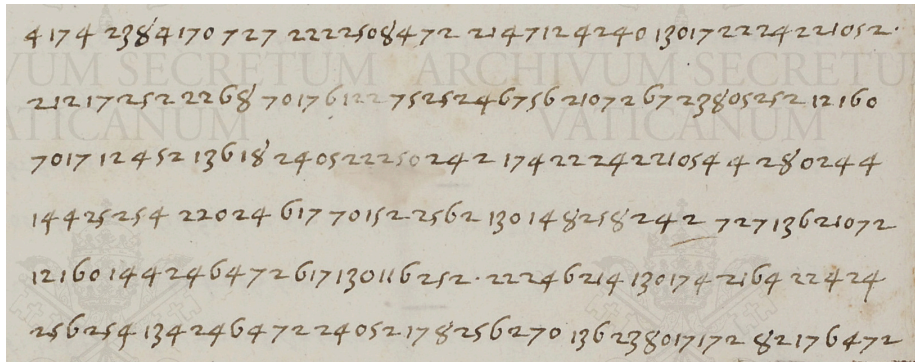


[fig from Nathan

Schneider. could also add dependency]

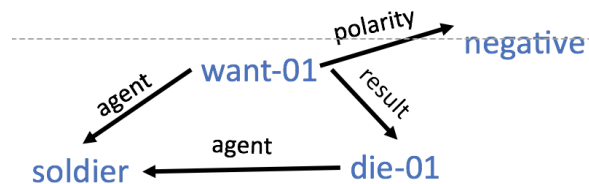
- pre-text (speech channel)
 - phonetics -- what mouth sound has been produced?
 - [l] = alveolar lateral approximant (lace)
 - [R] = alveolar tap (race)
 - [r] = alveolar trill (rey (Spanish))
 - phonology -- what are the meaningfully distinct sounds (governed by each language)?
 - English: [R] vs. [r] conflated
 - Japanese: [R] vs. [r] vs. [l] conflated
 - Hindi: [d] (dental) vs. [dʰ] (dental, glottal) distinct
 - https://en.wikipedia.org/wiki/Hindustani_phonology examples (retroflex are good ones to confuse my american ears)
 - example: ɒ̌ɪs ɪz ə 'sɪmpl 'sentəns
- pre-text (vision channel)

- orthography -- what makes a character? Particularly difficult when dealing with unknown writing set, especially handwritten:



- morphology -- what are the minimal meaning-laden parts of a word that are useful to distinguish? (Why distinguish? For handling novelty (wug test), collapsing statistics...)
 - English is pretty weak here (only Chinese is weaker):
 - inflection: 'talks' = 'talk (verb) + s (present 3rd singular)' or 'talk (noun) + s (plural).
 - Turkish agglutination: uyarlaştıramadıklarımızdanmışınızcasına = '(behaving) as if you are among those whom we could not civilize'
- words (lexemes)
 - Fundamental atom of data processing. We can describe it as we like. Could be an intuitive definition...
 - Is a word 'a single unit of meaning?' 'text separated by whitespace?'
 - What about Chinese? Or Thai? Or long compounds/agglutinations in Turkish, German, Finnish?
 - What about whitespace-separated units that function noncompositionally ('New York', 'take out')?
 - What about hyphenated and punctuated text? (Tricky example: 'New York-New Haven Railroad')

- In practice we more or less use a half solution: common punctuation is separated, some specific decisions are made about contractions, and otherwise it's just whitespace separated. This is less useful for morphologically rich and whitespace-free languages, where other strategies are used
- syntax -- how to properly put words together to form a sentence
 - part-of-speech tags:
 - the/DT (determiner) blue/JJ (adjective) boat/NN (noun)
 - *boat/NN blue/JJ the/DT
 - Constituencies:
 - S = NP[the blue boat] VP[sailed home]
 - * VP[sailed home] NP[the blue boat]
- semantics -- what does a word in a sentence mean, and how do the words meaningfully relate to each other?
 - Consider the word `want`: What are its distinct definitions?
 - Now consider the sentence `The soldier did not want to die.`: What definition of `want` is being used?
 - Who is doing the wanting? Who is doing the dying? What is (not) wanted?



- pragmatics -- what does the *speaker* (as opposed to the sentence) mean (often in conversational context)?
 - Human: Can you get me a table at N/Naka tomorrow at 7?
 - *AI: Yes, I have that ability.
 - AI: OK, your reservation is made.
 - Also seen in “dogwhistle” or coded language
 - “I don’t like my opponent’s **globalist** agenda” (antisemitic)
 - “agencies tied to **big pharma** and the **chemical industry** are problematic” (anti-vax)
- discourse – how are sentences/clauses/etc related to each other or to the whole text?
 - In a news article, main idea vs supporting argument, background, contrasting opinion
 - I bought a car. Therefore [Result] I had less money.
 - Darwin is a geologist. He tends to be viewed now as a biologist [concession] but on the Beagle his main work was geology [evidence].

Ambiguity

- “Add pineapple sage, pineapple, sage, good and plentys, and mike and ikes to the shopping list” (punctuation/segmentation recognition)

- Teacher strikes idle kids (greedy pos tagging – NN VB JJ NN vs JJ NN VB NN)
- The horse raced past the barn fell (greedy segmentation – (the horse) raced past (the barn) ... vs. (the horse raced past the barn) fell)
- Kids make nutritious snacks (semantics of “make”)
- Stolen painting found by tree (role labeling of “tree” – agent vs. location)
- “Is that window still open?” vs “Please close the window” (pragmatics)
- Dinosaurs didn't read. Now they are extinct. (discourse label. Causal?)