Lecture 1:
Natural Language Processing

Instructor: Jackie CK Cheung

COMP-550

Fall 2018

J&M Chapter 1

About Me

Education:

BSc in Computer Science (UBC)	2004-2008
MSc / PhD in Computer Science (Toronto)	2008-2014
Assistant professor at McGill	2015-

Research topics in my lab:

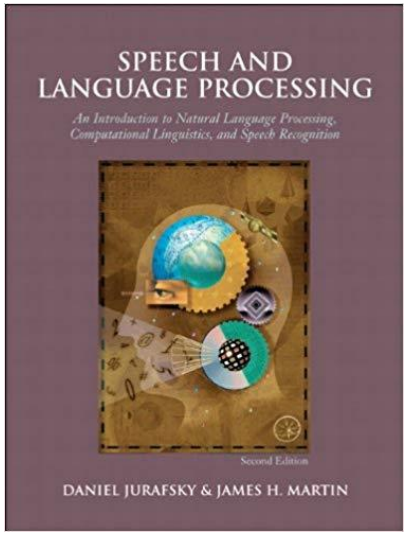
- Natural language generation
- Automatic summarization
- Computational discourse
- Computational semantics
- Commonsense reasoning

Preliminaries

Instructor:	Jackie Chi Kit Cheung
Time and Loc.:	TR 14:35-15:55 in MC 13
Office hours:	T 16:00-18:00 or by appointment in MC108N
TAs:	Malik Altakrori, Sunyam Bagga, Jad Kabbara, Kian Kenyon-Dean
Evaluation:	4 assignments (40%)
	1 midterm (20%)
	1 group project (40%)

Textbook

Jurafsky and Martin. *Speech and Language Processing* (2nd edition)



Hard copy available at bookstore

Draft chapters of 3rd edition available online:

https://web.stanford.edu/~jurafsky/slp3/

The Course Is Full

Current registration: 141/140, with full waiting list.

If you've registered for more courses than you plan to take, please decide soon! Many students are trying to get into this course.

Due to resource and classroom size limits, I cannot extend the class size anymore.

Assignments

Four assignments (10% each)

Involve readings, problem sets and programming component.

- Hand in online through myCourses
- Programming to be done in Python 2.7.
- Also non-programming written components
- Assignment 3 will have two parts with two deadlines (to accommodate midterm date)

General Policies

Lateness policy for assignments:

- < 15 minutes: no penalty
- 15 minutes – 24 hours: 10% absolute penalty
- > 24 hours: not accepted

Plagiarism: just don't do it.

Language policy: In accord with McGill policy, you have the right to write essays and examinations in English or in French.

Course website: http://cs.mcgill.ca/~jcheung/teaching/fall-2018/comp550/index.html

Slides, recordings, other materials and announcements given **in-class** or **on MyCourses.**

Midterm

Worth 20% of your final grade

Wed, Oct 31, 18:05 – 19:25 (80 minutes long)

ADAMS Auditorium

More details as we approach the midterm date.

Final Project

Worth 40%.

Experiment on some language data set

Summarize and review relevant papers

Report on experiments

Must be done in teams of two

Coming up with a project idea:

- Extend a model we see in class
- Work on a relevant topic of interest
- Consult a list of suggested projects, to be posted

Project Steps

Paper or project proposal

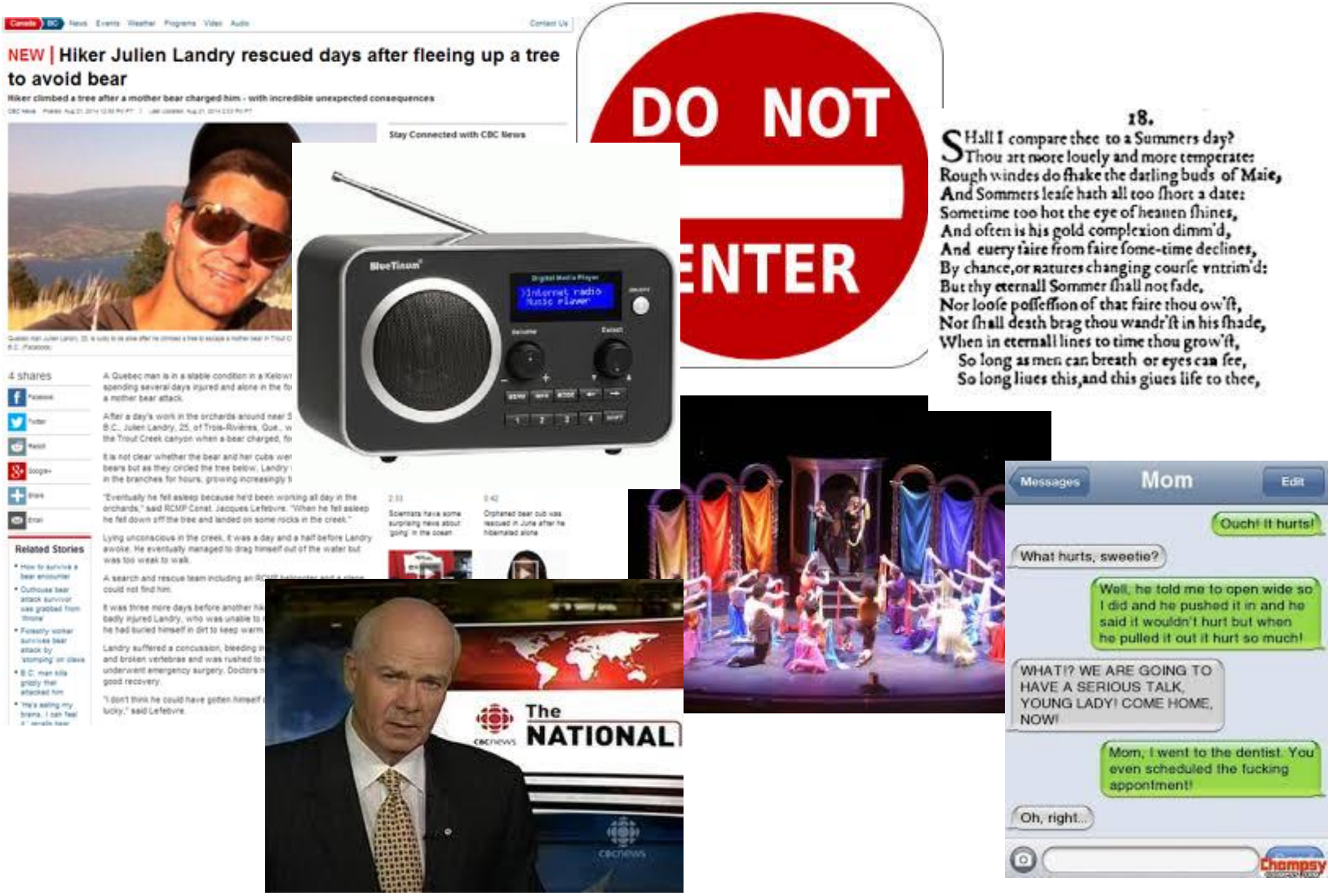
Progress update

Final submission

Due dates to be announced

Computational Linguistics and Natural Language Processing

Language is Everywhere



Languages Are Diverse

6000+ languages in the world

language

langue

ਭਾਸ਼ਾ

語言

idioma

Sprache

lingua

→lingyourlanguage

https://lingyourlanguage.com/ (My high score is 720)

What is Language?

Some properties:

- Form of communication
- Arbitrary pairing between form and meaning
- Primarily vocal (exception: sign languages)
- Highly expressive and productive
- Nearly universal (barring developmental disorders)

How do these compare?

- Programming language (e.g., C, Python, Java)
- Vocalizations by your favourite animal
- Written English

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Domains of natural language

- Acoustic signals, phonemes, words, syntax, semantics, …
- Speech vs. text
- **Natural language understanding** (or **comprehension**) vs. **natural language generation** (or **production**)

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Goals

- Language technology applications
- Scientific understanding of how language works

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Methodology and techniques

- Gathering data: language resources
- Evaluation
- Statistical methods and machine learning
- Rule-based methods

Natural Language Processing

Computational linguistics and **natural language processing (NLP)** are sometimes used interchangeably.

Slight difference in emphasis:

NLP	**CL**
Goal: practical technologies	Goal: how language actually works
Engineering	Science

Understanding and Generation

Natural language understanding (NLU)

- Language to form usable by machines or humans

Natural language generation (NLG)

- Traditionally, semantic formalism to text
- More recently, also text to text

Most work in NLP is in NLU

- c.f. linguistics, where most theories deal primarily with production

Personal Assistant App

Understanding

Call a taxi to take me to the airport in 30 minutes.

What is the weather forecast for tomorrow?

Generation

Machine Translation

I like natural language processing.

Automatische Sprachverarbeitung gefällt mir.

Understanding

Generation

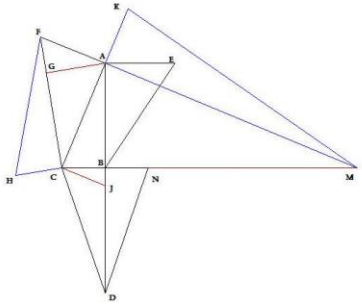
Computational Linguistics

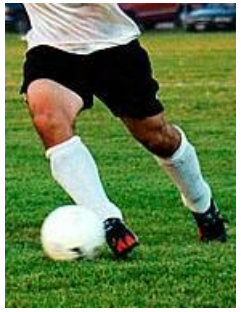
Besides new language technologies, there are other reasons to study CL and NLP as well.

The Nature of Language

First language acquisition

- Chomsky proposed a **universal grammar**
- Is language an “instinct”?





What innate knowledge must children already have in order to learn their mother tongue, given their exposure to linguistic inputs?

- Train a model to find out!

The Nature of Language

Language processing

Some sentences are supposed to be grammatically correct, but are difficult to process.

Formal mathematical models to account for this.

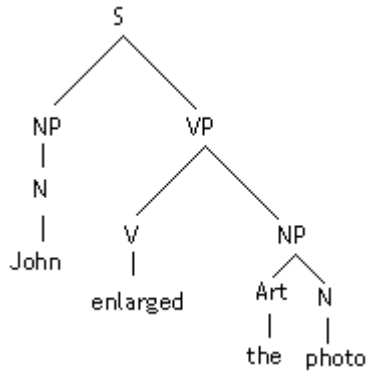
The rat escaped.

The rat the cat caught escaped.

?? *The rat the cat **the dog chased** caught escaped.*

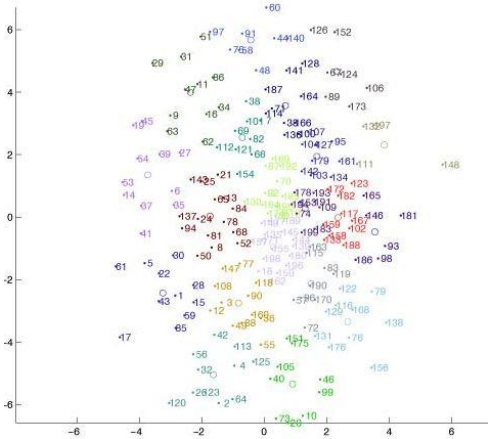
Mathematical Foundations of CL

We describe language with various formal systems.



cat + z > cats

cat + z	*SS	Agree	Max	Dep	Ident
catiz				*!	
catis				*!	*
catz		*!			
cat			*!		
☞ cats					*



$$\begin{bmatrix} \text{phon } \textit{made} \\ \text{synsem} \mid \text{loc} \mid \text{cat} \begin{bmatrix} \text{head} \begin{bmatrix} \text{maj } v \\ \text{vform } \textit{finite} \end{bmatrix} \\ \text{valence} \begin{bmatrix} \text{subj} \langle \text{NP} \rangle \\ \text{compls} \langle \text{VP}[\textit{base}], \text{NP} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{phon } \textit{force} \\ \text{synsem} \mid \text{loc} \mid \text{cat} \begin{bmatrix} \text{head} \begin{bmatrix} \text{maj } v \\ \text{vform finite} \end{bmatrix} \\ \text{valence} \begin{bmatrix} \text{subj} \langle \text{NP} \rangle \\ \text{compls} \langle \text{VP}[\textit{inf}], \text{NP} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Mathematical Foundations of CL

Mathematical properties of formal systems and algorithms

- Can they be efficiently learned from data?
- Efficiently recovered from a sentence?
- Complexity analysis

Implications for algorithm design

Types of Language

Text

- Much of traditional NLP work has been on news text.
  - Clean, formal, standard English, but very limited!
- More recent work on diversifying into multiple domains
  - Political texts, text messages, Twitter

Speech

- Messier: disfluencies, non-standard language
- Automatic speech recognition (ASR)
- Text-to-speech generation

Domains of Language

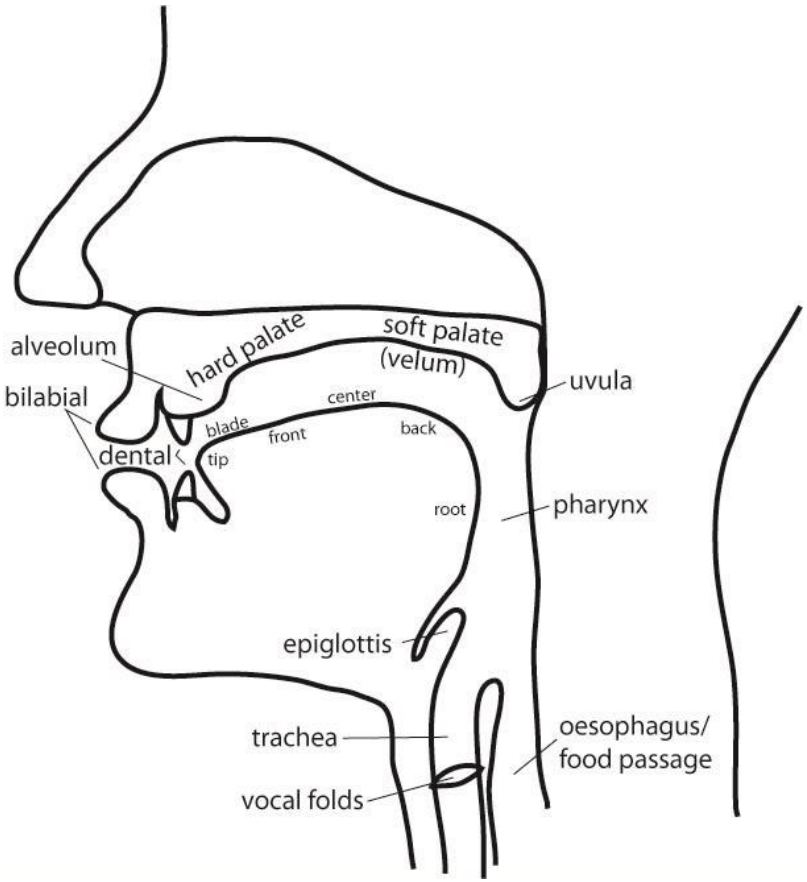
The grammar of a language has traditionally been divided into multiple levels.

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Phonetics

Study of the speech sounds that make up language

Articulation, transmission, perception



peach [phi:tsh]

Involves closing of the lips, building up of pressure in the oral cavity, release with aspiration, …

Vowel can be described by its formants, …

Phonology

Study of the rules that govern sound patterns and how they are organized

peach	[phi:tsh]	/pi:t͡ʃ/
speech	[spi:tsh]	/spi:t͡ʃ/
beach	[bi:tsh]	/bi:t͡ʃ/

The p in peach and speech are the same phoneme, but they actually are phonetically distinct!

Morphology

Word formation and meaning

antidisestablishmentarianism

anti- dis- establish -ment -arian -ism

establish

*establish**ment***

*establishment**arian***

*establishmentarian**ism***

***dis**establishmentarianism*

***anti**disestablishmentarianism*

Syntax

Study of the structure of language

**I a woman saw park in the.*

I saw a woman in the park.

The first sentence is not well formed (it is **ungrammatical**), while the second one is.

- Words must be arranged in a certain order in a certain way to be a valid English sentence!

Syntax

http://explosm.net/comics/1682/

There are two meanings for the first sentence in the comic! What are they? This is called **ambiguity**.

Semantics

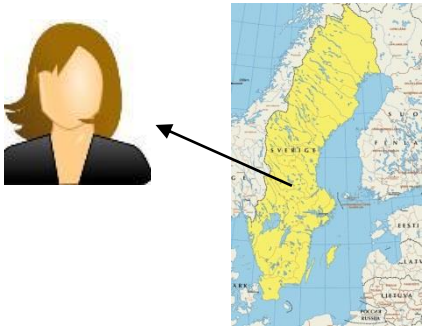
Study of the meaning of language

bank

Ambiguity in the **sense** of the word

Semantics

Ross wants to marry a Swedish woman.

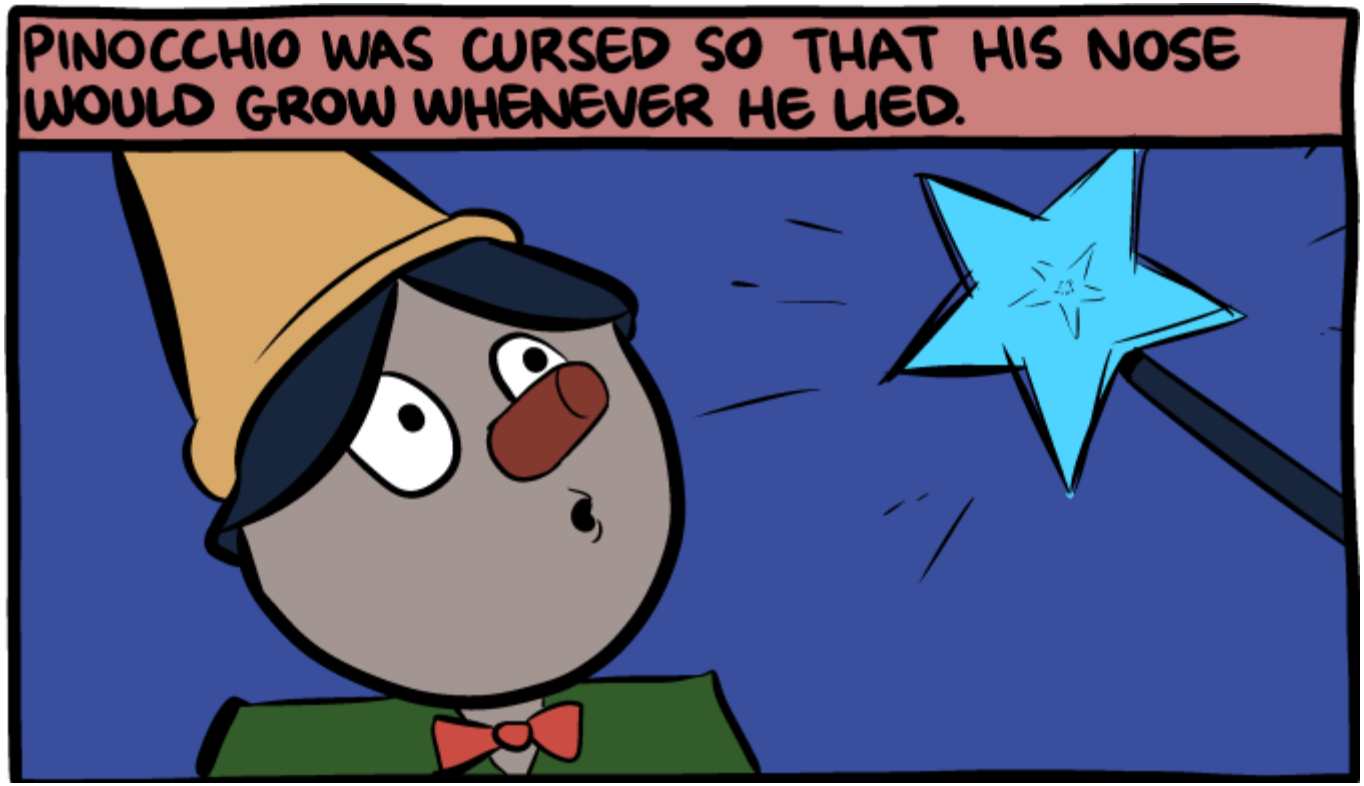


Pragmatics

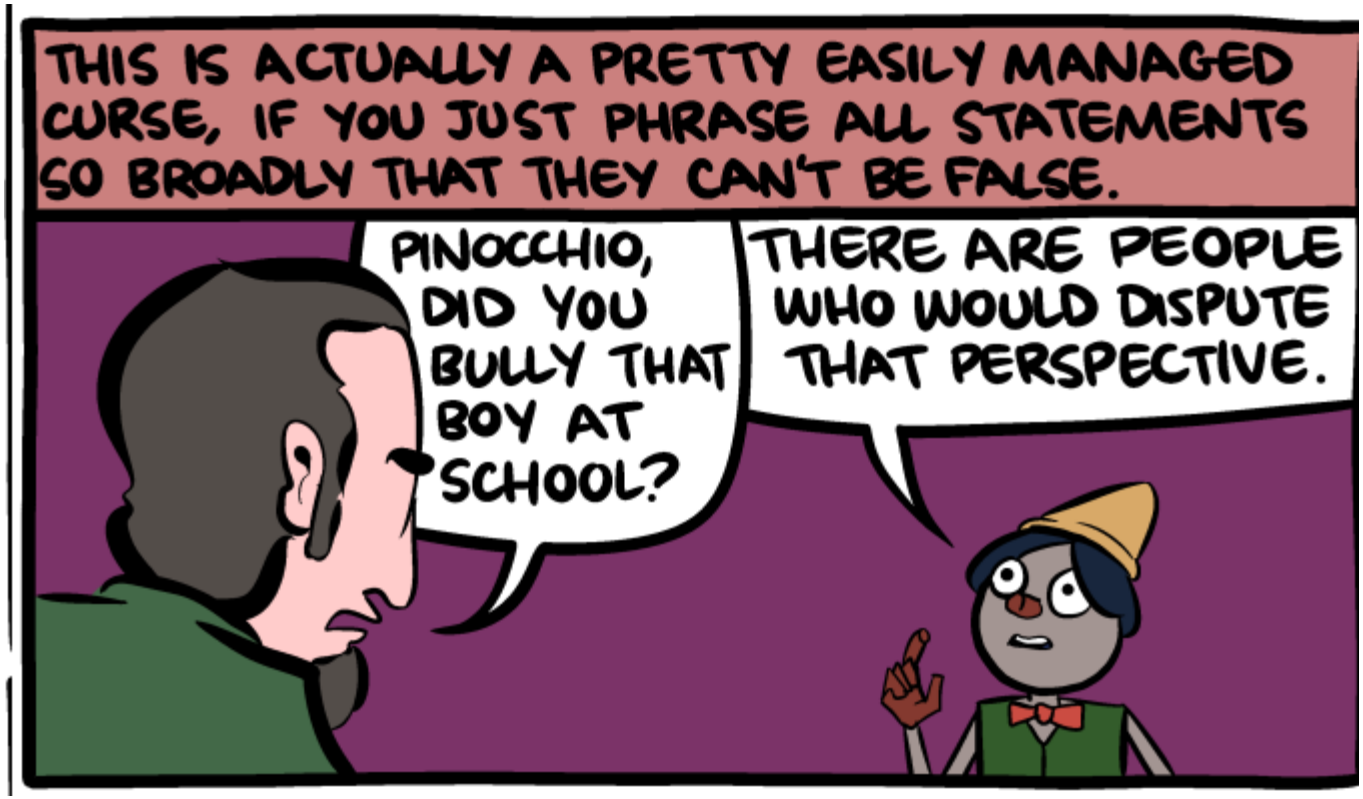
Study of the meaning of language in context.

→ Literal meaning (semantics) vs. meaning in context:

http://www.smbc-comics.com/index.php?id=3730



Pragmatics


THIS IS ACTUALLY A PRETTY EASILY MANAGED CURSE, IF YOU JUST PHRASE ALL STATEMENTS SO BROADLY THAT THEY CAN'T BE FALSE.
PINOCCHIO, DID YOU BULLY THAT BOY AT SCHOOL?
THERE ARE PEOPLE WHO WOULD DISPUTE THAT PERSPECTIVE.

Pragmatics


YOU CAN ALSO ANSWER TOUGH QUESTIONS BY JUST DESCRIBING THE CURRENT SITUATION.
PINOCCHIO! DID YOU EGG MY DOOR?
IF YOU'RE ACCUSING ME OF SUCH A DEED, I HAVE NOTHING MORE TO SAY.

Pragmatics


CURSED TO NEVER LIE, PINOCCHIO CEASED TO INTERACT WITH TRUTH, EVEN IN SAFE SITUATIONS.
SIR, WOULD YOU LIKE ANY DESSERT?
DESSERT WOULD BE DELICIOUS.

Pragmatics – Deixis

Interpretation of expressions can depend on **extralinguistic** context

e.g., pronouns

I think cilantro tastes great!

The entity referred to (the **antecedent**) by *I* depends on who is saying this sentence.

Discourse

Study of the structure of larger spans of language (i.e., beyond individual clauses or sentences)

I am angry at her.

She lost my cell phone.

I am angry at her.

The rabbit jumped and ate two carrots.

Questions

1. What is the difference between phonetics and phonology?

2. What are two possible readings of this phrase? What level does the ambiguity act at? (i.e., lexical, syntactic, semantic, discourse)
    - *old men and women*

Topics in COMP-550

Progress through the subfields, roughly organized by the level of linguistic analysis

Morphology -> Syntax -> Semantics -> Discourse

NLP problems:

- Language modelling, part-of-speech tagging, parsing, word sense disambiguation, semantic parsing, coreference resolution, discourse coherence modelling

Focus on:

Basic linguistics needed to understand NLP issues

Algorithms and problem setups

Machine Learning in COMP-550

Interspersed throughout the course, and introduced as necessary

Machine learning topics we will cover:

- Feature extraction
- Sequence and structure prediction algorithms
- Probabilistic graphical models
- Linear discriminative models
- Neural networks and deep learning

Optional Tutorials

To help balance/even out different backgrounds:

- Python programming
- Machine learning
- Neural networks

Stay tuned for details!

Applications in COMP-550

Last three weeks of the course focus on language technology applications and advanced topics:

- Automatic summarization
- Machine translation
- Evaluation issues in NLP

Course Objectives

Understand the broad topics, applications and common terminology in the field

Prepare you for research or employment in CL/NLP

- Learn some basic linguistics
- Learn the basic algorithms
- Be able to read an NLP paper

Understand the challenges in CL/NLP

- Answer questions like “Is it easy or hard to…”