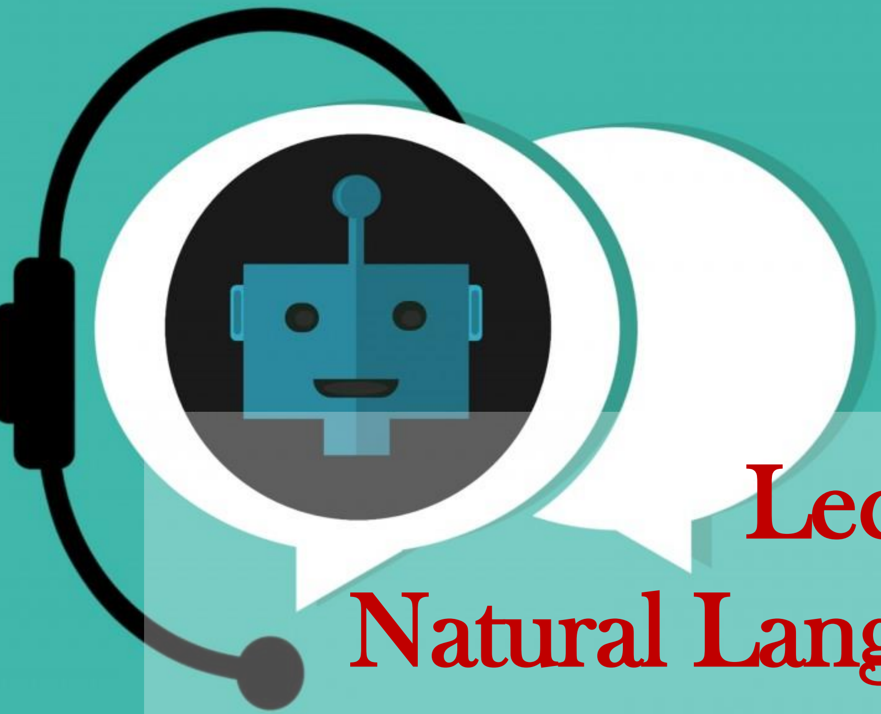


Lecture 1:
Natural Language Processing

Instructor: Jackie CK Cheung

COMP-550

Fall 2020

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 pragmatics and discourse

Computational semantics

Common sense reasoning in text

General Information

Instructor:	Jackie Chi Kit Cheung
Time and Loc.:	Posted online Mon and Wed
Office hours:	Start next week. TBD
TAs:	Ali Emami
	Zhi Wen
	Dora Jambor
	David Venuto
	Ashita Diwan

Lectures and Engagement

Pre-recorded and posted Mondays and Wednesdays

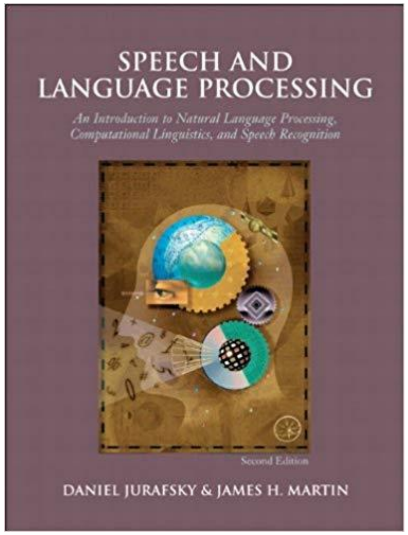
- There are 202 students registered!

We will try to create a supportive and engaging environment:

- Office hours at different times of the day
- Discussion forum will be monitored
- Active participation will be expected

Course Materials

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


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Hard copy available at bookstore

Draft chapters of 3rd edition available online:

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

Other useful resources linked to in course outline

Evaluation

Group final project	40%
Programming assignments	20%
Reading assignments	20%
Study groups	10%
Online quizzes	10%

Final Project

Group project in **a team of three** (40%)

Experiment on some language data set

Summarize and review relevant papers

Report on experiments

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

There will be a proposal and a final deadline.

Assignments

Programming assignments (2 x 10% each)

To be done in Python 3

Reading assignments (4 x 5% each)

Classic papers and recent papers on topics in NLP

Summarize and discuss in a short report

Late policy for assignments

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

Study Groups (10%)

We will ask you to form small study groups (size: 3 – 6) that meet regularly to discuss course topics

- Guided discussions; questions and topics to be posted
- Short write-up to summarize discussions
- More details to be posted
- Could even watch the lecture recordings together!

We will help with forming groups, respecting everybody's constraints (time zone, computer equipment, connectivity, other issues).

Online Quizzes (10%)

To be posted most weeks on myCourses

Goal is to check your understanding of the course materials from that week

- Short answer and multiple choice
- Computation questions and knowledge tests
- Multiple attempts allowed

General Policies

Plagiarism

- Just don’t do it—I regularly catch and submit cases
- See course outline for full policy and link to McGill's academic integrity policy

Language policy

You have the right to submit written assignments and reports in English or in French.

Slides, recordings, other materials and announcements given on myCourses.

Questions?

For general questions about course organization, go to myCourses:

Discussions > Course Organization

What is Language? What is NLP?

What is Language?

Some properties:

- Form of communication
- **Arbitrary** pairing between form and meaning
- Highly expressive and productive
- Nearly universal (barring developmental disorders)
- Uniquely human

How do these compare?

- Vocalizations by your favourite animal (e.g., meowing, barking, whalesong)
- Traffic signs and symbols
- Programming language (e.g., C, Python, Java)

Languages Are Diverse

6000+ languages in the world

Arbitrary pairing of form and meaning:

language

langue

ਭਾਸ਼ਾ

語言

idioma

Sprache

lingua

lugha

→lingyourlanguage

https://lingyourlanguage.com/ (My high score is 513 on Omniglot)

Language and Computers

We can now communicate with computers in rudimentary ways.

How is this possible? What are the limitations? Are we close to having the AI on *Her*?

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

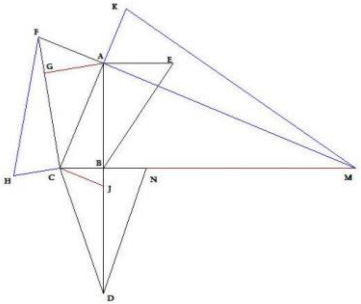
Computational Linguistics

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.

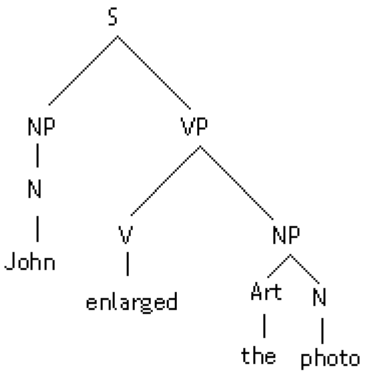
The rat escaped.

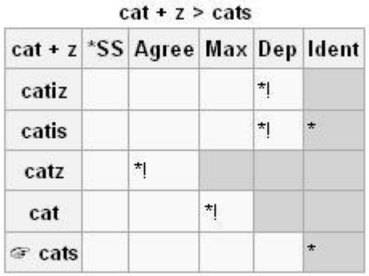
The rat the cat caught escaped.

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

Mathematical Foundations of CL

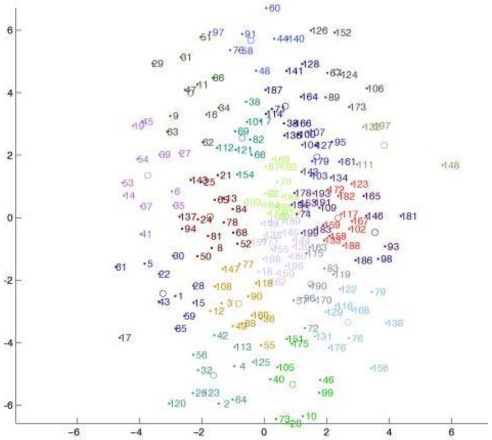
We describe language with various formal systems.





cat + z > cats

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



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

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

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

- In some sense, an idealization of spoken language.
- 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

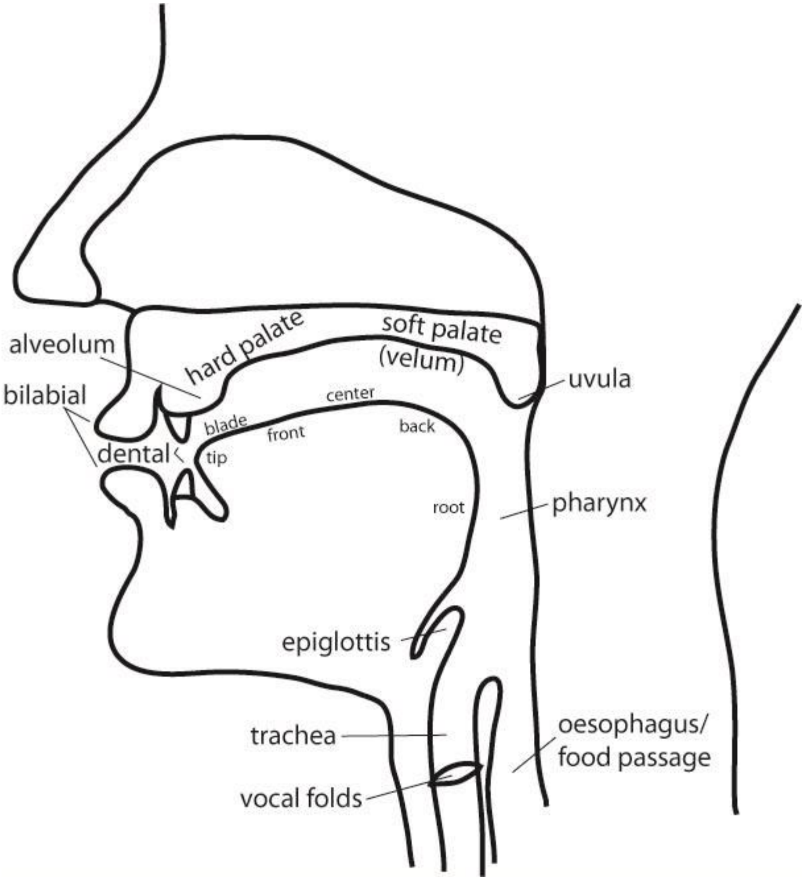
The grammar of a language has traditionally been divided into multiple levels.

- 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

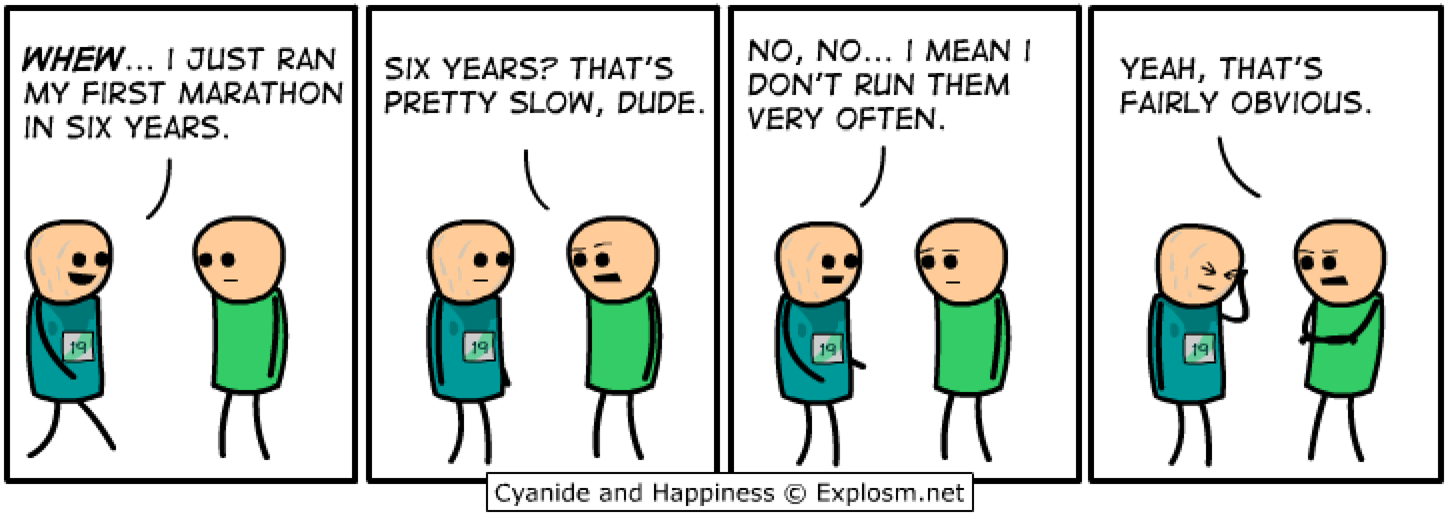
**I a woman saw park in the.*

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


WHEW... I JUST RAN MY FIRST MARATHON IN SIX YEARS.
19
SIX YEARS? THAT'S PRETTY SLOW, DUDE.
19
NO, NO... I MEAN I DON'T RUN THEM VERY OFTEN.
19
YEAH, THAT'S FAIRLY OBVIOUS.
19

Cyanide and Happiness © Explosm.net

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

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

Semantics

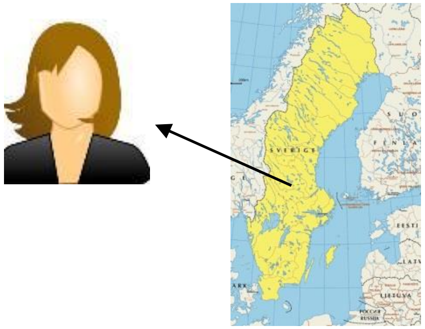
Study of the meaning of language

bank

Ambiguity in the **sense** of the word

Semantics

Ross wants to marry a Swedish woman.

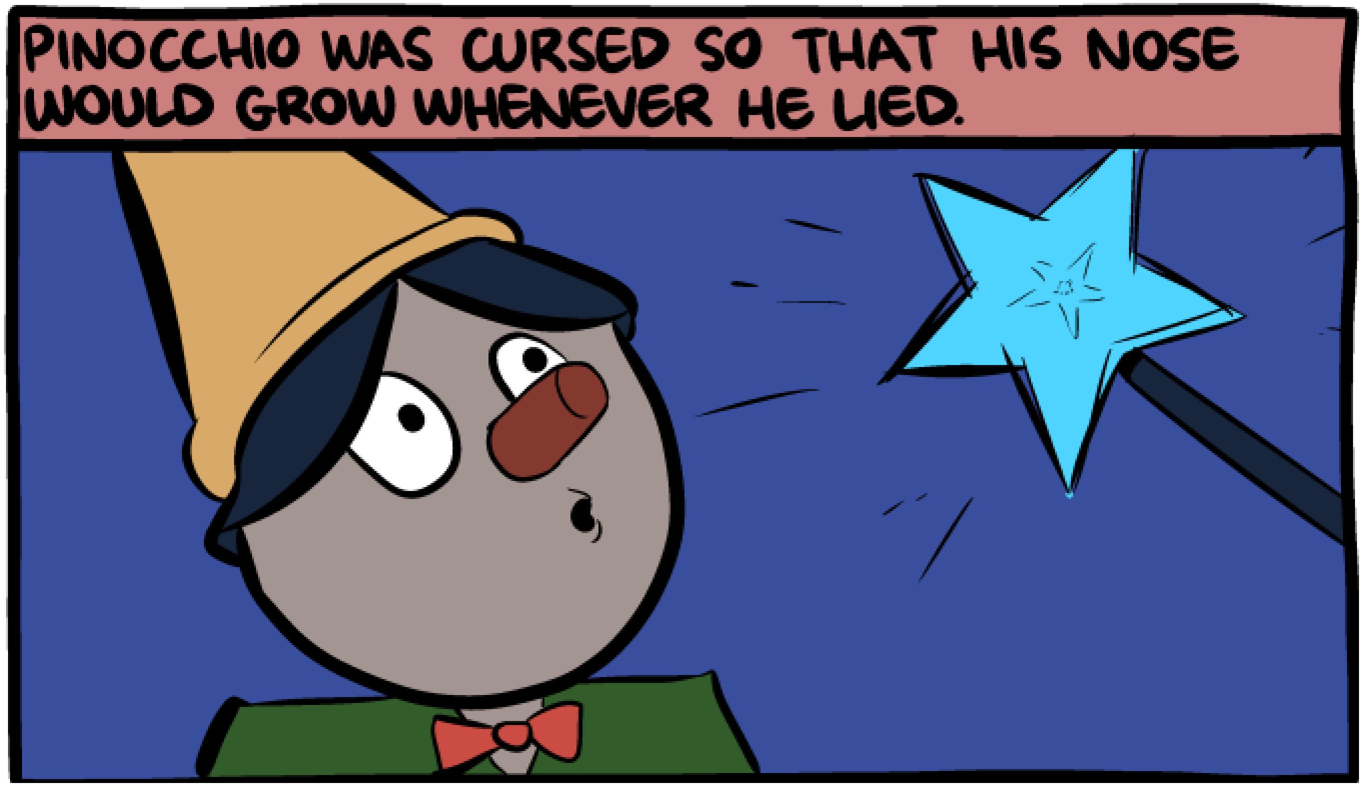


Pragmatics

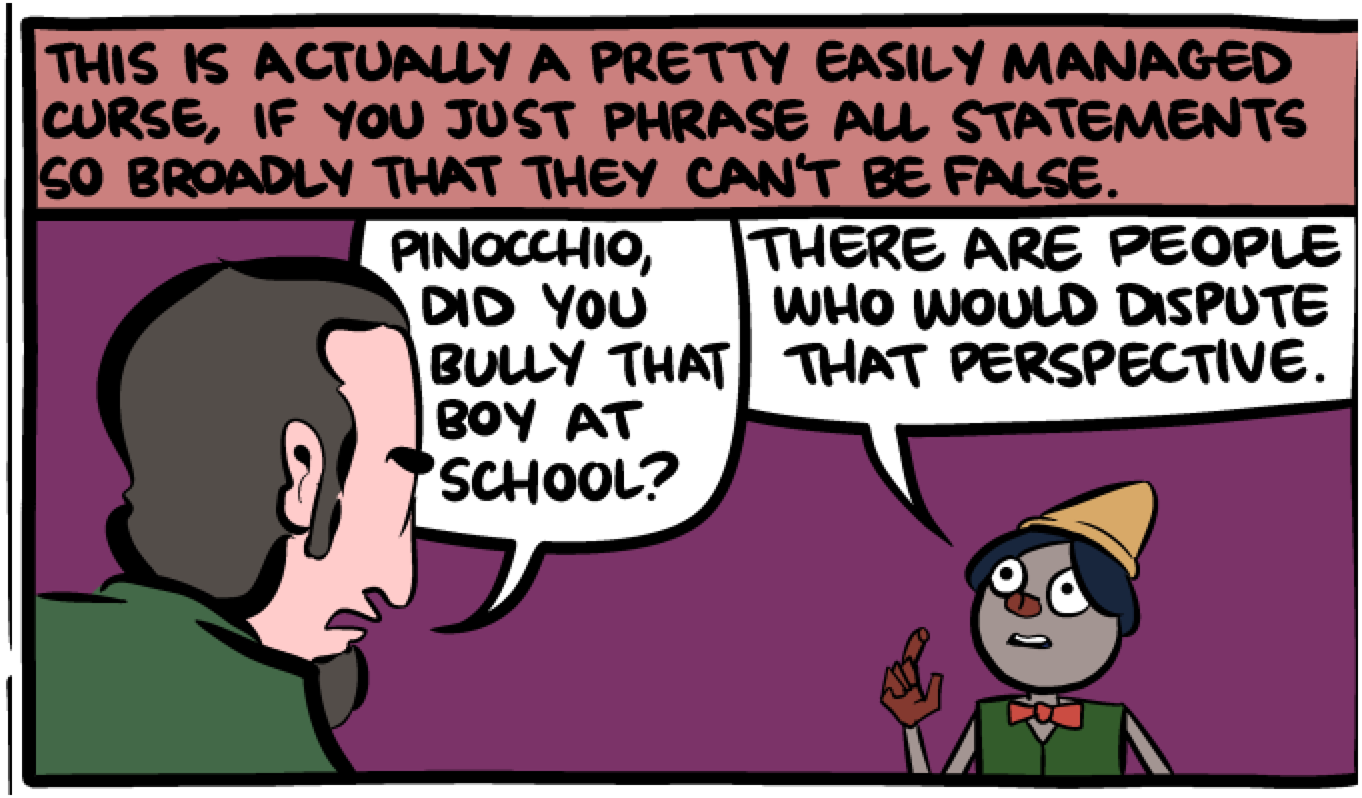
Study of the meaning of language in context.

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

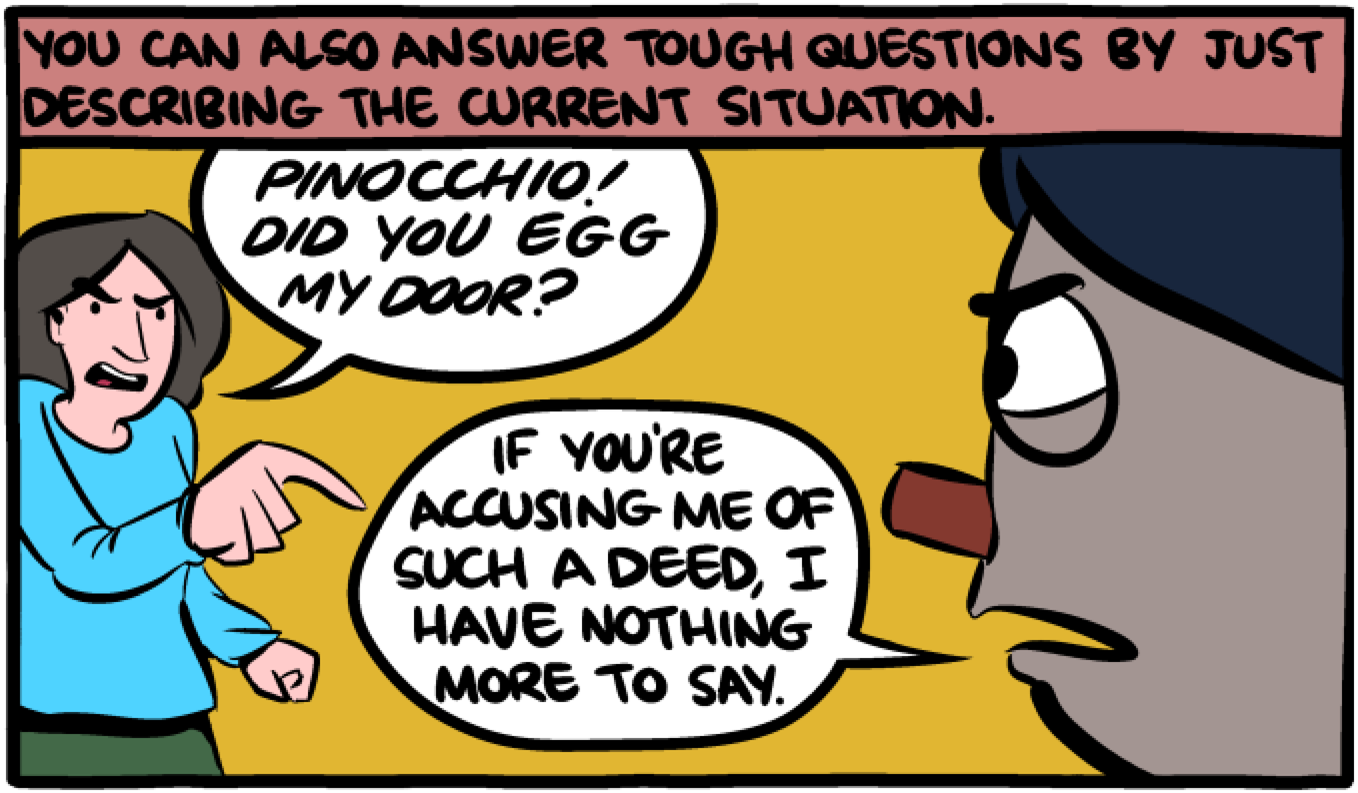
http://www.smbc-comics.com/index.php?id=3730


PINOCCHIO WAS CURSED SO THAT HIS NOSE WOULD GROW WHENEVER HE LIED.

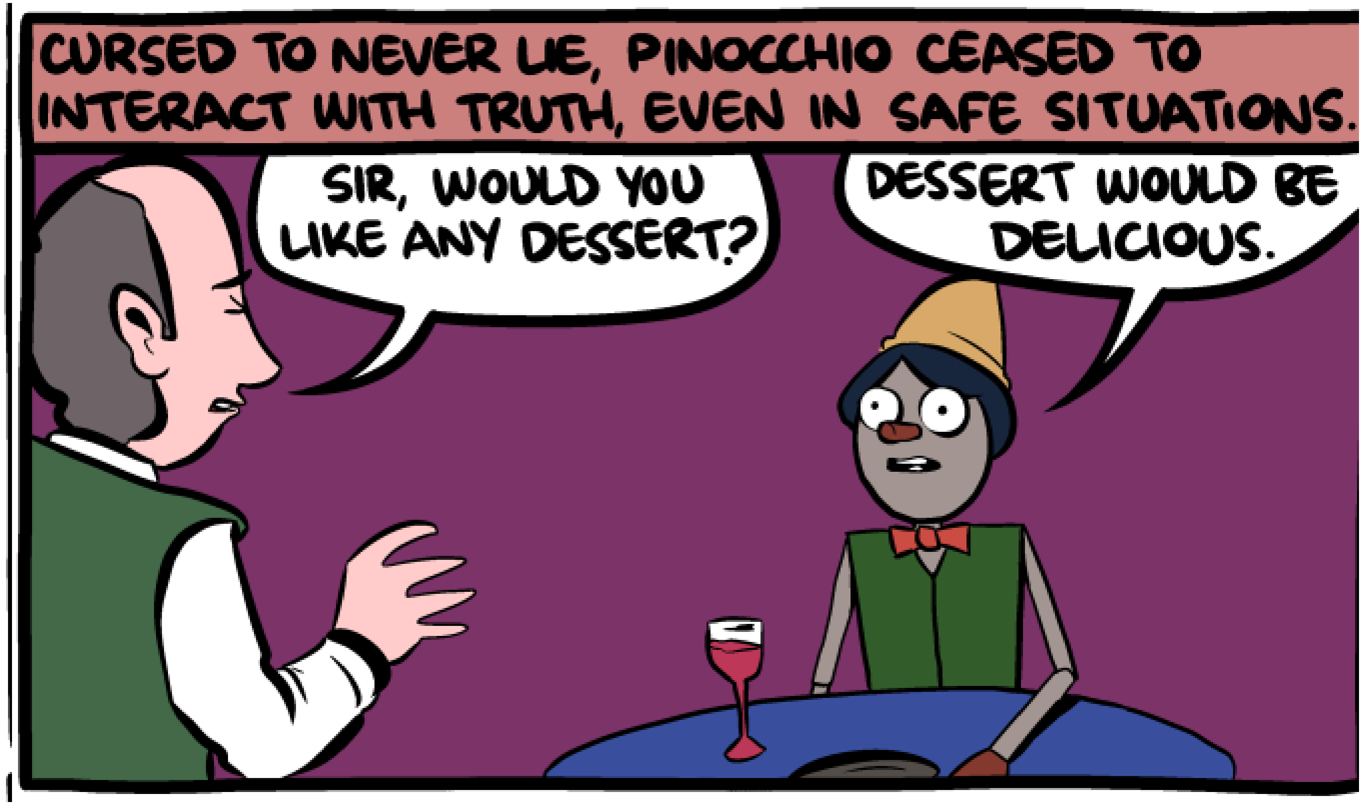
Pragmatics


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.

NLP – the Technological Perspective

A combination of **pre-specified knowledge** and **machine learning from data**

Problem specification

Machine learning algorithms

Human annotations

Linguistic knowledge

…

'The Expanse': 7 Big Questions I Had While Watching Syfy's Uber-Addicting Drama

By Meghan O'Keefe @megsokay | May 25, 2017 at 3:00pm 125 Shares

The Expanse

"Caliban's War"

Season 2 , Episode 13

Community Grade (60 Users)

Your Grade

I take notes on my laptop while I watch screeners, but my notes for tonight's season 2 finale are relatively short. That's because every time I jot something down—a quote, a plot point, a character name I keep forgetting—I have to pause the screener and switch tabs. It's not an ideal solution, but it usually works, except when I'm watching something I don't really want to pause. Several things happen in "Caliban's War," and a few of them are pretty damn important, but most of the episode is spent hoping people we've come to care about aren't going to die. It's hard to pay attention to the smaller details when a monster is ripping a hole in the ship's hull, I guess.

The Expanse Season 2: Entertaining if uneven, reaching ever closer to greatness.

What it is: The second season of SyFy's increasingly popular sci-fi political thriller. The show is based on books by Daniel Abraham and Ty Franck, who collectively write under the pen name James S.A. Corey.

What I've watched: Took my time with season 2, which I finally finished up late last week.

Is it good? Yeah. It still needs some work before it can rival the likes of *Battlestar Galactica*, but the second season took some steps in the right direction.

Quick thoughts: I started watching *The Expanse* while the first season was still underway, and it's been cool to see more people talking about it. I dug it well enough from the start, particularly how hard-boiled it was about killing off

Websites

News articles

Discussions

Knowledge bases

…

NLP Tools and Techniques

Major paradigms for NLP, not mutually exclusive:

Rule-based systems

- Often hand-engineered knowledge about language
- E.g., *heureux* -> *happy*

Machine learning

- Model learns about language through examples
- **Classification**: e.g., is this e-mail spam?
- **Sequence models**: make series of decisions
- Many other paradigms

Knowledge representation

- Formal structure to encode what model knows
- Logic? A large set of continuous-valued numbers?

Topics in COMP-550

Organized roughly by level of linguistic analysis and a corresponding technical approach (ML or otherwise)

NLP Topic	Linguistic layer	Techniques
Text classification	Words	Classification
Language modelling, POS tagging	Words (esp. syntactic structure of words)	Sequence models
Syntactic parsing	Syntactic structure	Structure prediction, dynamic programming
Computational semantics, coreference resolution	Meaning (semantics, discourse)	Logic, semi-supervised learning, neural models
Applications: MT, summarization, etc.	Various	Various

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…”

Questions?

For lecture content, go to myCourses:

Discussions > Lectures

For general questions about NLP, go to myCourses:

Discussions > NLP