Lecture 1: Natural Language Processing

Instructor: Jackie CK Cheung & David I. Adelani

COMP-550

Fall 2024

J&M Chapter 1

About Jakie

Associate Professor at McGill	2021 -
• Associate Scientific Co-Director at Mila	
Assistant Professor at McGill	2015 – 2021
PhD in Computer Science (Toronto)	2014

Research topics in my lab

- Natural language generation
- Automatic summarization
- Computational semantics
- Computational pragmatics
- Applications of NLP

About David

Assistant Professor at McGill	2024 -
• Core Academic Member at Mila	
Senior Research Fellow at UCL	2022 – 2024
PhD in Computer Science (Saarland)	2023

Research topics in my lab

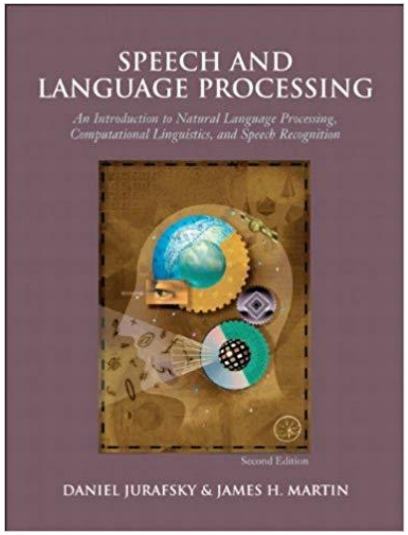
- Multilingual Natural language processing
- Machine translation
- Representation learning
- Speech processing

Preliminaries

Instructor:	Jackie Chi Kit Cheung & David I. Adelani
Time and Loc.:	11:35 – 12:55 Macdonald-Harrington, G-10
Office hours:	Mon. 14:00-15:30 MC 108N (Jackie)
	Wed. 14:00-15:30 MC 204N (David)
TAs:	Shira Abramovich
	Ziling Cheng
	Gaurav Iyer
	Xijuan Sun
	Zihan Wang
Evaluation:	2 programming assignments (20%)
	4 reading assignments (20%)
	1 midterm (25%)
	1 group project (35%)

Textbook

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/

Assignments

Two programming assignments (10% each x 2 = 20%)

- Hand in online through myCourses
- Programming to be done in Python 3.

Four reading assignments (5% each x 4 = 20%)

- Covers advanced material and applications

Midterm

Worth 25% of your final grade

To be completed online as a myCourses quiz

Time: November 6, 2024

More details as we approach the midterm date.

Final Project

Worth 35%.

Experiment on some language data set

Summarize and review relevant papers

Report on experiments

Must be done in teams of three

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

General Policies

Lateness policy for assignments:

- Grace period of 24 hours
- > 24 hours: accepted if it is convenient for us at our discretion

Plagiarism: just don’t do it—I regularly catch and submit cases.

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

Generative AI Usage

Fine to use in an assistive manner

- Help understand course content
- Search for information
- Brainstorm ideas
- Edit writing

Must acknowledge use of this technology.

Not okay to use as primary means to complete tasks

- Feed in assignment questions to generate solutions
- Generate project report from scratch on a topic

Platforms

ed

- Being adopted by many CS courses this term
- You’ll be added this week
- Most releases will be done via this platform

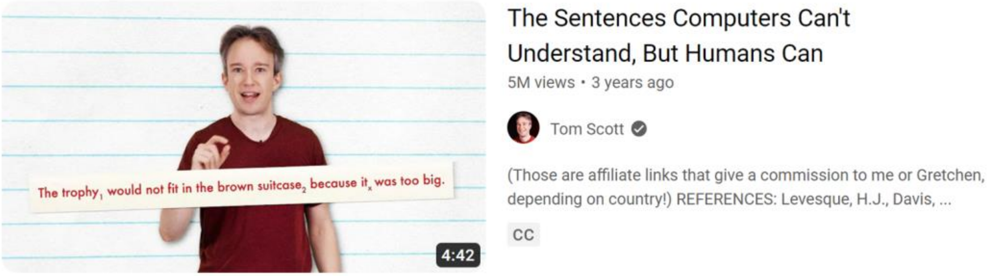
myCourses

- Assignment and project submissions
- Midterm
- Grade release

Computational Linguistics and Natural Language Processing

LLMs – Impressive Impact!

- Question answering, code generation, essay writing, summarization
- Commercial uses: customer service, personal assistants, healthcare
- Many informal uses: entertainment, settling disputes


The trophy, would not fit in the brown suitcase, because it, was too big.
4:42

The Sentences Computers Can't Understand, But Humans Can

5M views • 3 years ago

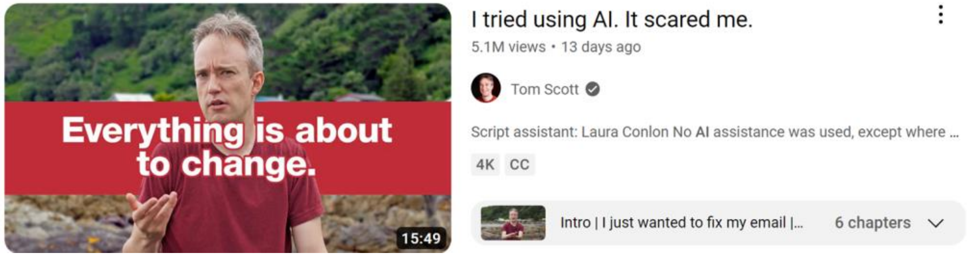
Tom Scott

(Those are affiliate links that give a commission to me or Gretchen, depending on country!) REFERENCES: Levesque, H.J., Davis, ...

CC

Tom Scott, 2020
"Artificial language processing remains 10 years away, just as it has for the last few decades."

Tom Scott, 2023
"… that this new technology, the thing that was going to change everything, was starting to actually change everything"



I tried using AI. It scared me.

5.1M views • 13 days ago

Tom Scott

Script assistant: Laura Conlon No AI assistance was used, except where ...

4K CC

Intro | I just wanted to fix my email |... 6 chapters

How Do Language Models Work?

Key insight: learn correlations between words in context

Language modelling:

Mary had a little _____

• *lamb*	GOOD
• *accident*	GOOD?
• *very*	BAD
• *up*	BAD

Do this at internet-scale with sophisticated statistical techniques (deep learning)!

What This Course Is About

- How did we get to large language models dominating NLP research?
- What was the progression of the field of NLP? Why did people try the methods that they did?
- What are some common tasks and paradigms involving natural language?
- How do we evaluate and analyze NLP systems?
- How are properties of natural language reflected in NLP research?

What This Course Is Not About

- The latest techniques in language modelling
- Deep learning / machine learning as a primary focus
    - We will touch on this, and you can do a final project that uses ML, but it is **not** the primary focus of the course.

Language is Everywhere

Canada
BC
News
Events
Weather
Programs
Video
Audio
Contact Us

NEW | Hiker Julien Landry rescued days after fleeing up a tree to avoid bear

Hiker climbed a tree after a mother bear charged him - with incredible unexpected consequences

Stay Connected with CBC News
4 shares
Facebook
Twitter
Reddit
Google+
Share
Email

A Quebec man is in a stable condition in a Kelowr
spending several days injured and alone in the fo
a mother bear attack.

After a day's work in the orchards around near S
B.C., Julien Landry, 25, of Trois-Rivières, Que., w
the Trout Creek canyon when a bear charged, fo

It is not clear whether the bear and her cubs wer
bears but as they circled the tree below, Landry
in the branches for hours, growing increasingly f

"Eventually he fell asleep because he'd been wor
orchards," said RCMP Const. Jacques Lefebvre. "When he fell asleep he fell down off the tree and landed on some rocks in the creek."

Lying unconscious in the creek, it was a day and a half before Landry awoke. He eventually managed to drag himself out of the water but was too weak to walk.

A search and rescue team including an RCMP helicopter and a plane could not find him.

It was three more days before another hik
badly injured Landry, who was unable to
he had buried himself in dirt to keep warm

Landry suffered a concussion, bleeding in
and broken vertebrae and was rushed to
underwent emergency surgery. Doctors
good recovery.

"I don't think he could have gotten himself
lucky," said Lefebvre.

Related Stories

- How to survive a bear encounter
- Outhouse bear attack survivor was grabbed from 'throne'
- Forestry worker survives bear attack by 'stomping' on claws
- B.C. man kills grizzly that attacked him
- 'He's eating my brains, I can feel it,' recalls bear

18.

Shall I compare thee to a Summers day?
Thou art more louely and more temperate:
Rough windes do ſhake the darling buds of Maie,
And Sommers leaſe hath all too ſhort a date:
Sometime too hot the eye of heauen ſhines,
And often is his gold complexion dimm'd,
And euery faire from faire ſome-time declines,
By chance,or natures changing courſe vntrim'd:
But thy eternall Sommer ſhall not fade,
Nor looſe poſſeſſion of that faire thou ow'ſt,
Nor ſhall death brag thou wandr'ſt in his ſhade,
When in eternall lines to time thou grow'ſt,
So long as men can breath or eyes can ſee,
So long liues this,and this giues life to thee,

Ouch! It hurts!

What hurts, sweetie?

Well, he told me to open wide so I did and he pushed it in and he said it wouldn't hurt but when he pulled it out it hurt so much!

WHAT!? WE ARE GOING TO HAVE A SERIOUS TALK, YOUNG LADY! COME HOME, NOW!

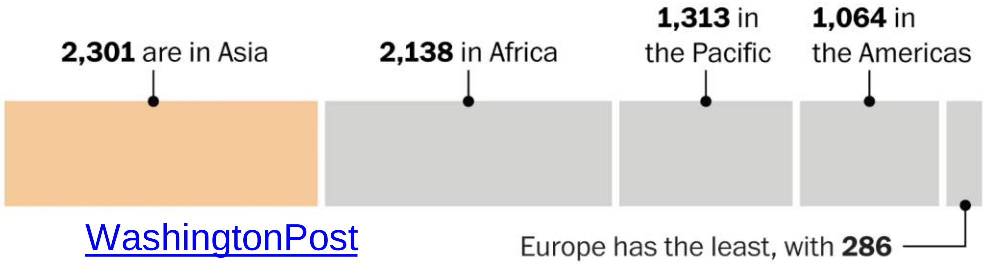
Mom, I went to the dentist. You even scheduled the fucking appointment!

Oh, right...

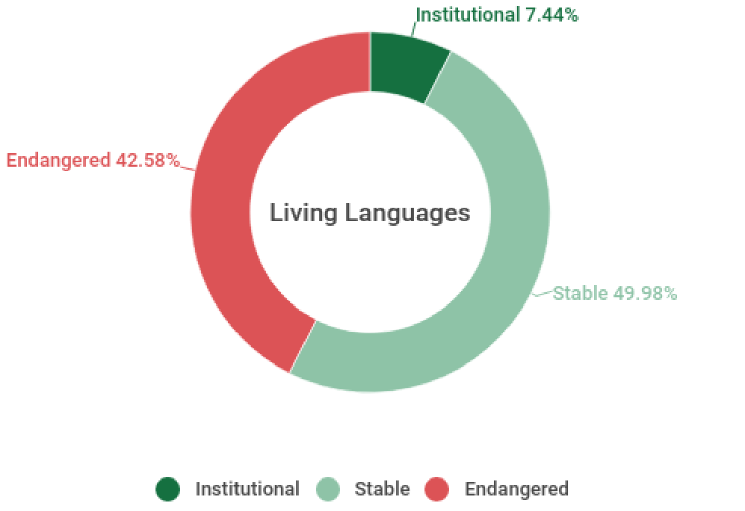
Languages Are Diverse

6000+ languages in the world

- language
- langue
- ਭਾਸ਼ਾ
- 語言
- idioma
- Sprache
- lingua



WashingtonPost



Ethnologue

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

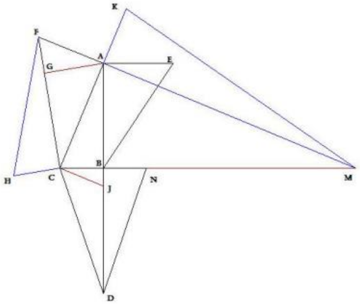
Besides new language technologies, there are other reasons to study CL and NLP as well.

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.

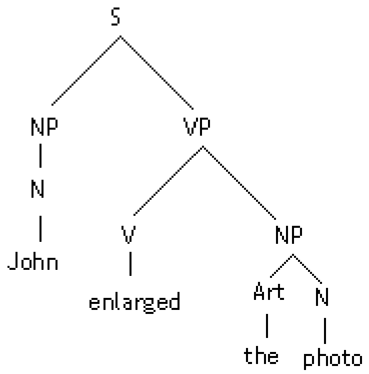
The rat escaped.

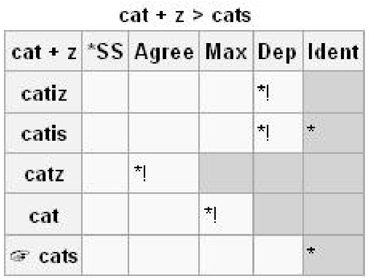
The rat <u>the cat caught</u> escaped.

?? *The rat <u>the cat **the dog chased** caught</u> escaped.*

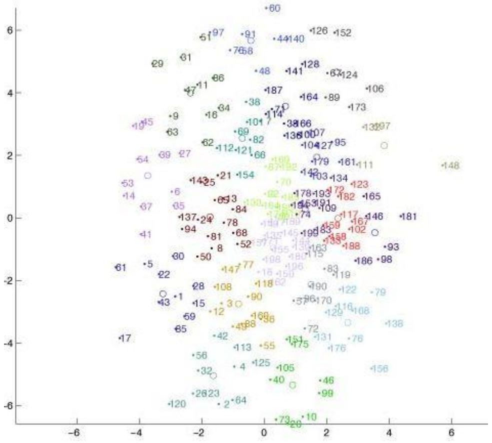
Mathematical Foundations of CL

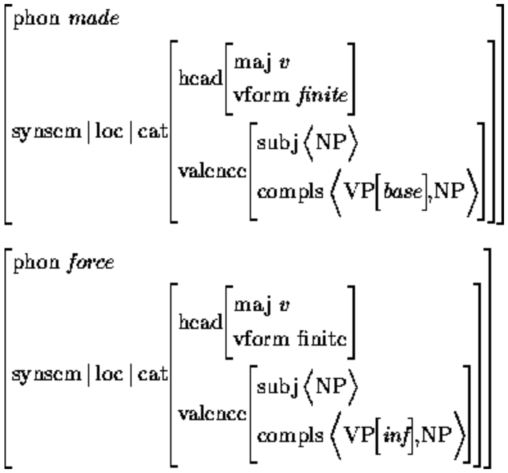
We describe language with various formal systems.




cat + z > cats

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





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

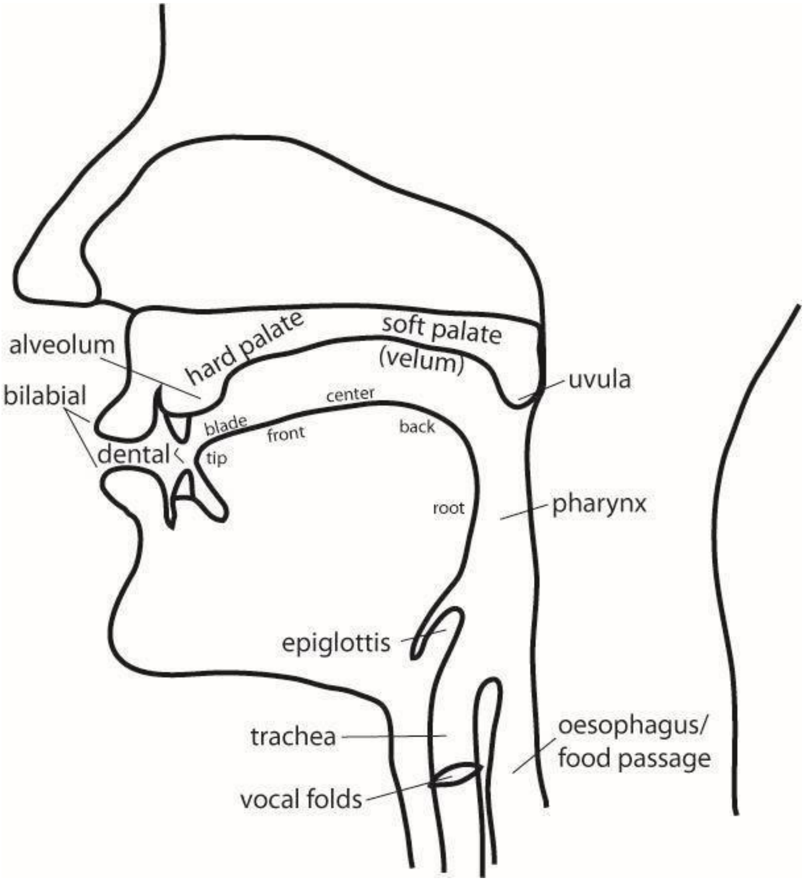
The grammar of a language has traditionally been divided into multiple levels.

- 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

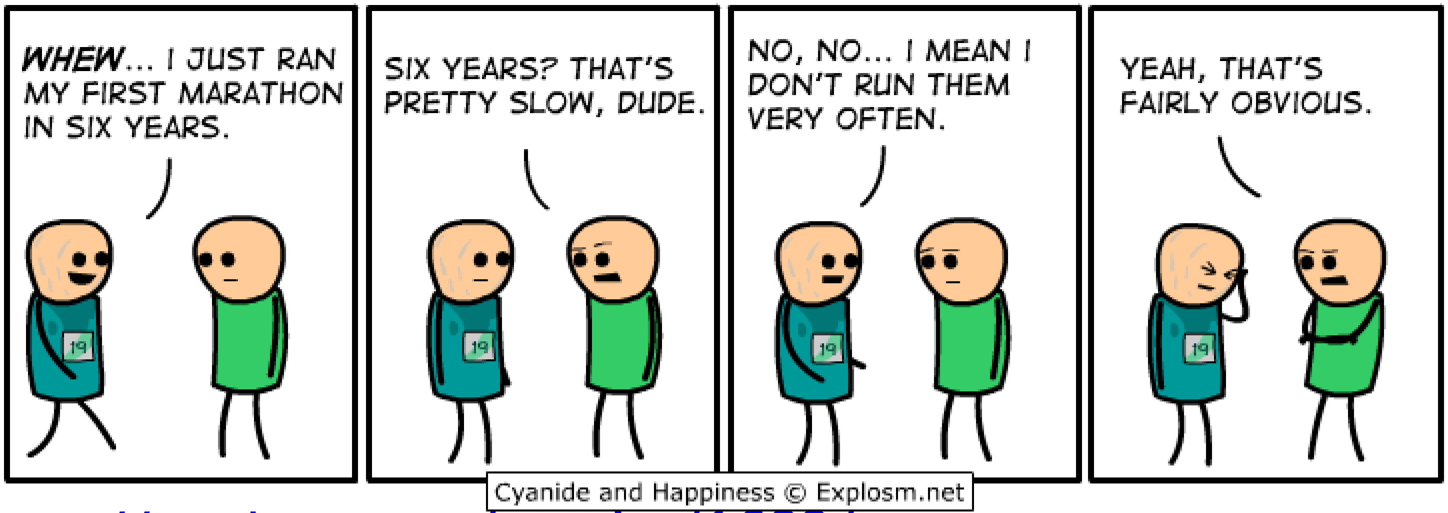
**I a woman saw park in the.*

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.
SIX YEARS? THAT'S PRETTY SLOW, DUDE.
NO, NO... I MEAN I DON'T RUN THEM VERY OFTEN.
YEAH, THAT'S FAIRLY OBVIOUS.
19

Cyanide and Happiness © Explosm.net

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

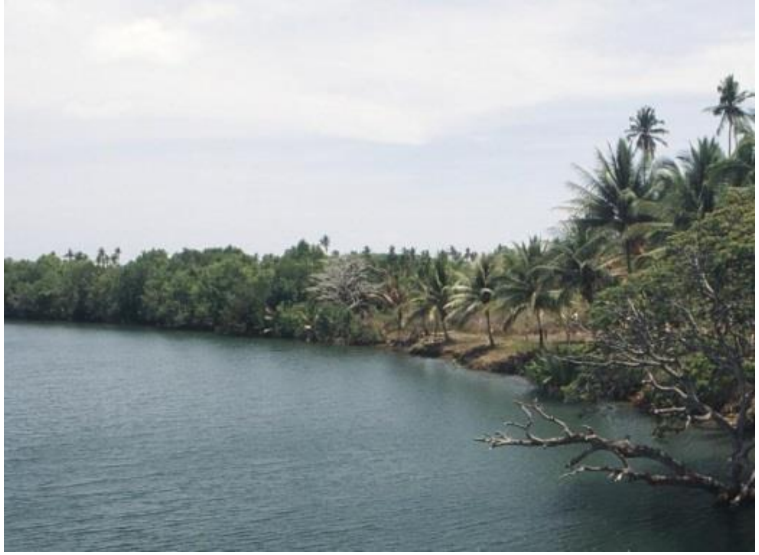
There are two meanings for the first sentence in the comic! What are they? This is called **ambiguity**.

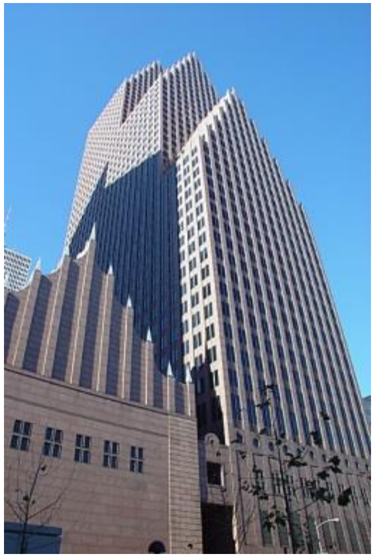
Semantics

Study of the meaning of language

bank

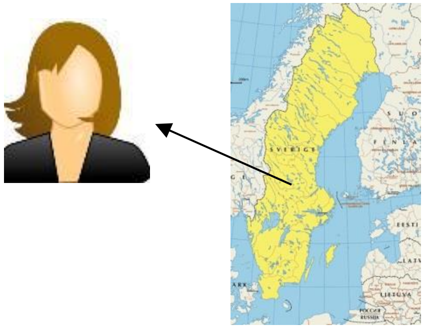
Ambiguity in the **sense** of the word





Semantics

Ross wants to marry a Swedish woman.

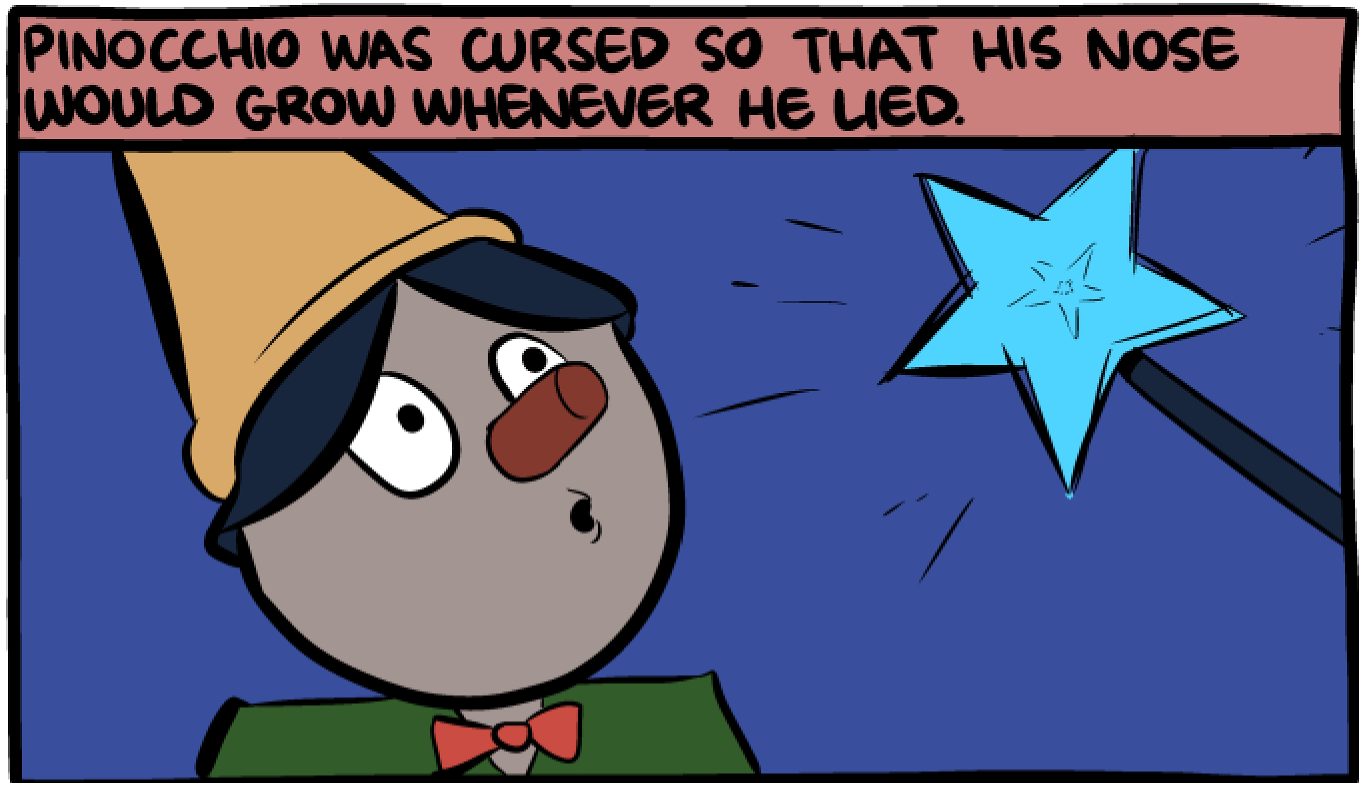


Pragmatics

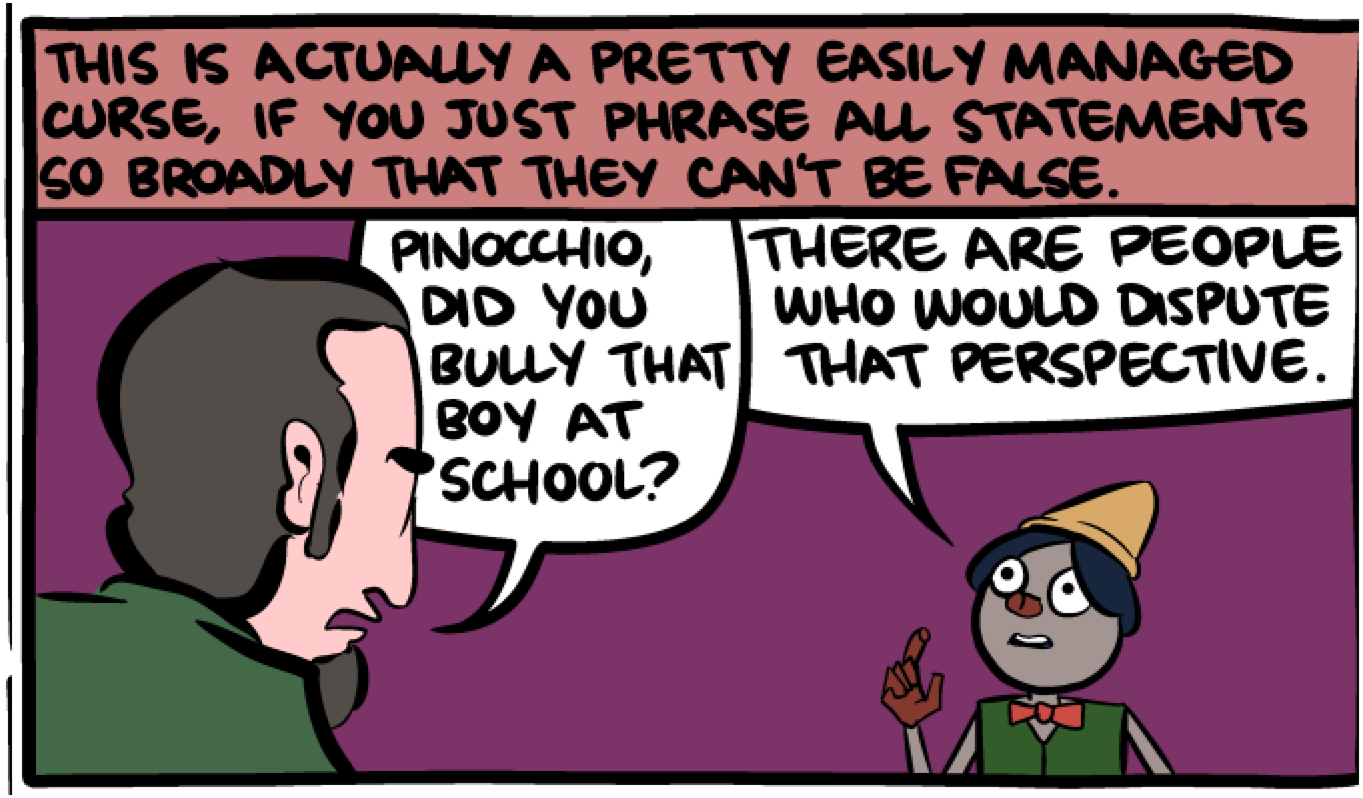
Study of the meaning of language in context.

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

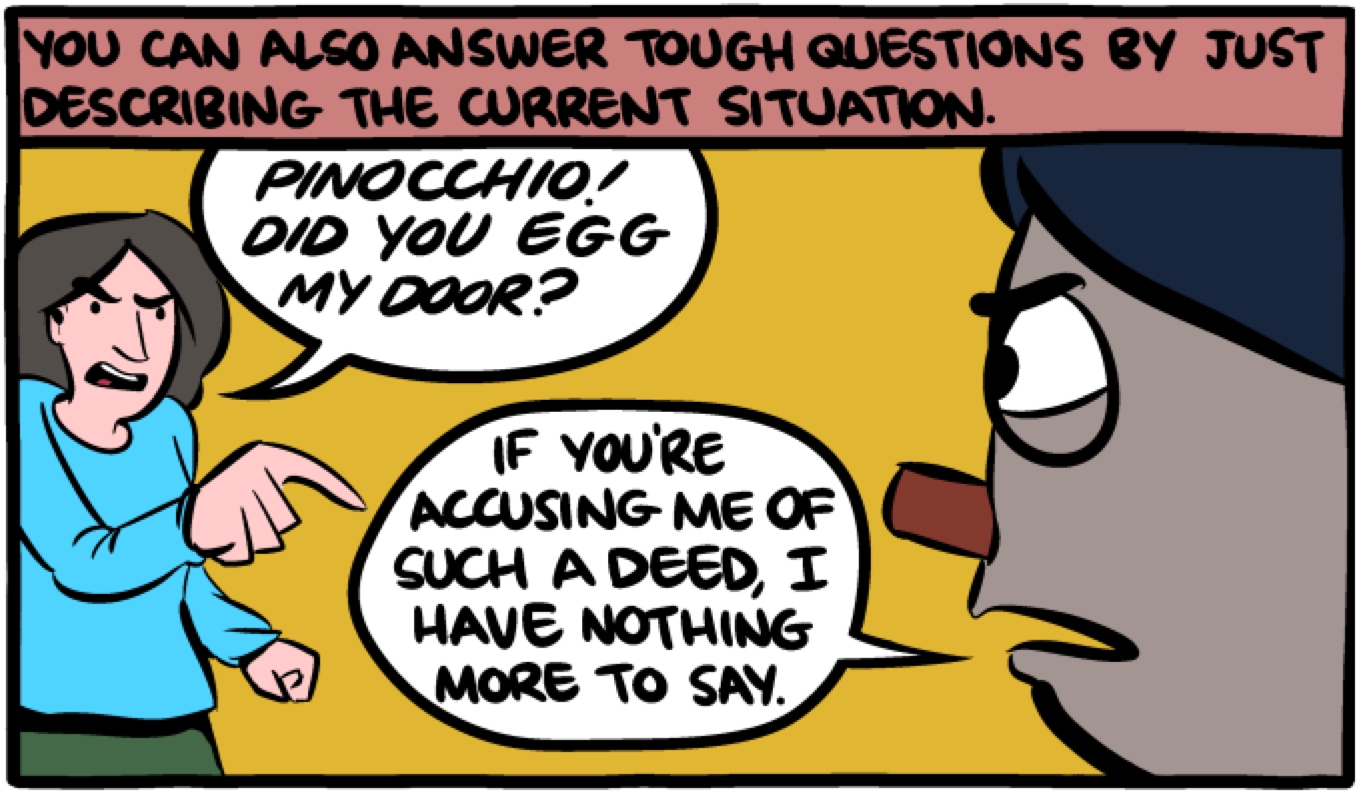
http://www.smbc-comics.com/index.php?id=3730


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

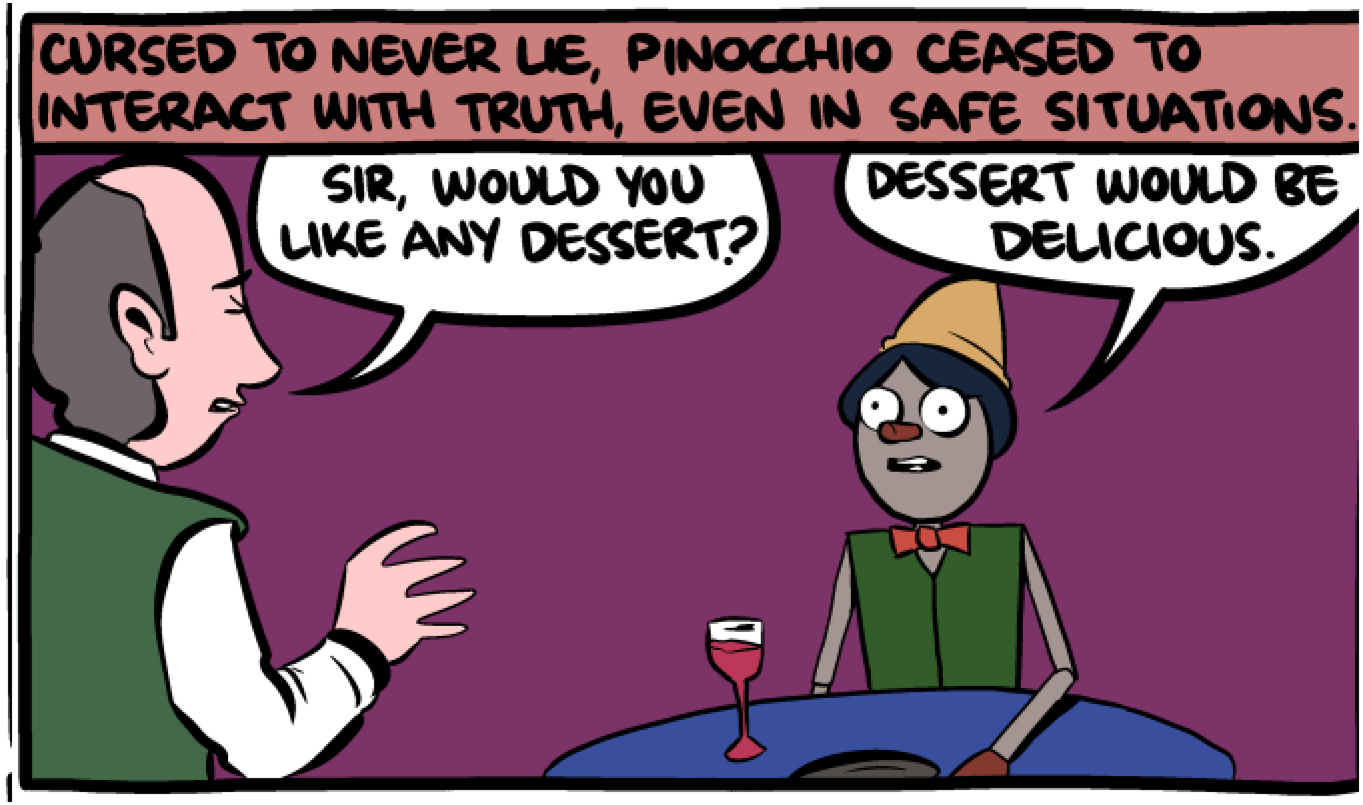
Pragmatics


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 @meganokay | 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.

I know I'm behind (a few years), but … drama, *The Expanse*.

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

Possible topics:

- Vision and language
- Automatic summarization
- Machine translation
- Evaluation issues in NLP

Accompanied by reading assignments!

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…”

Next Lecture

The next lecture is Wednesday, Sept 4

Monday, Sept 2 is Labour Day – enjoy!