CS-585

Natural Language Processing

Sonjia Waxmonsky, Ph.D.
swaxmonsky@iit.edu

Slides based in part on material from Derrick Higgins (IIT)

Today

1. About the course
2. About me
3. About you
4. About language and linguistics

THIS COURSE

About this Course: Goals

- Learn about **core concepts** and tasks in the field of *statistical n*atural language processing

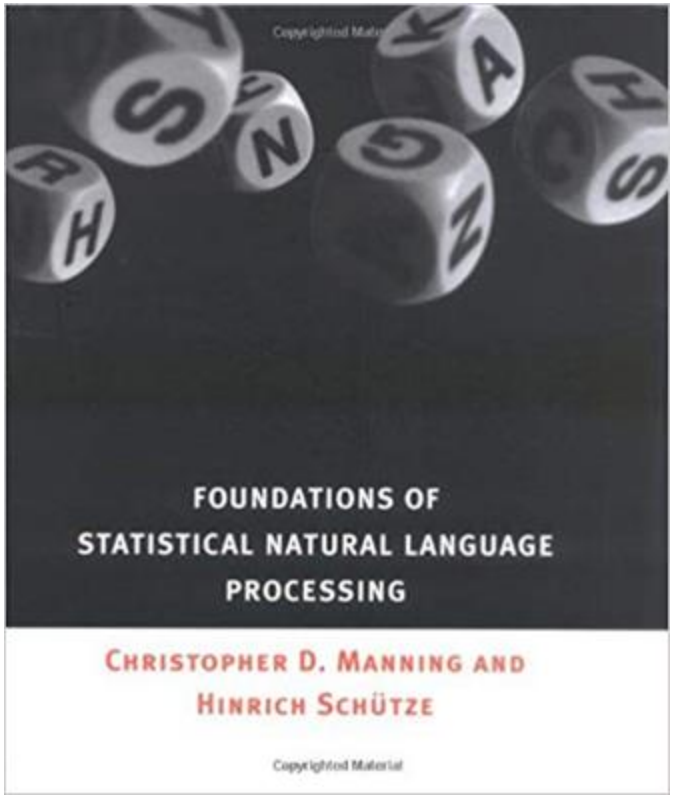
- Gain experience processing, analyzing, and building with **human language text data**

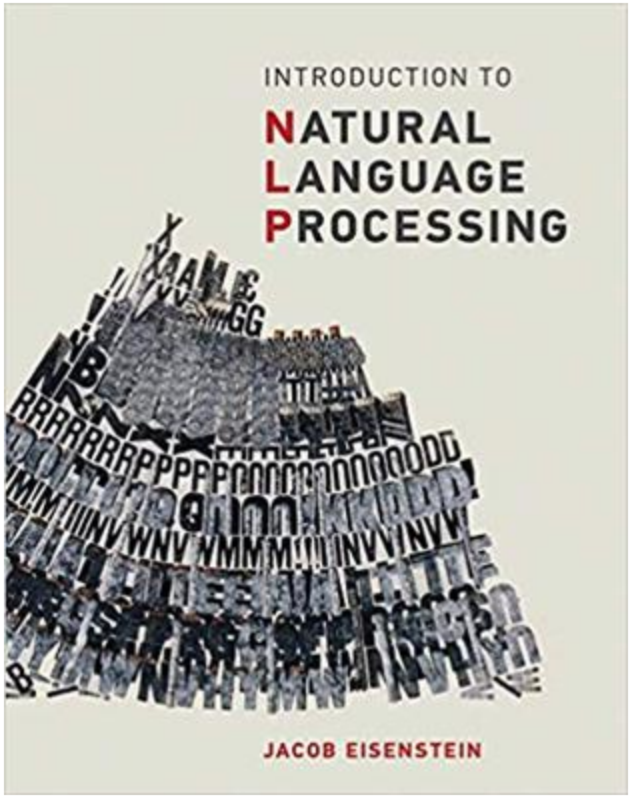
- **Prepare for further study** and project work in machine learning, data science, and deep learning.

About This Course: Prerequisites

- For whom is this course designed?
  - Senior undergraduate and graduate students
- Prerequisites
  - Math: Linear algebra and probability
  - Courses: CS 430 Intro to Algorithms
  - Programming:
    - Python 3
    - Algorithms & Data Structures
    - Access to a Linux or UNIX system

Readings


Copyrighted Material
FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
Christopher D. Manning and Hinrich Schütze
Copyrighted Material


INTRODUCTION TO
NATURAL LANGUAGE PROCESSING
JACOB EISENSTEIN

Grading

- 5 Homework Assignments (10% each → 50%)
- Tentative Schedule:
  - 2 HW before Fall break (Monday October 9)
  - 3 HW between Fall Break and Thanksgiving

- Exams (50%)
  - Midterm will cover material up to Fall Break (20%)
  - Final will cover material from the entire course (30%)

Homework Late Policy

Late Policy:

- HW will be due on 11:59pm on posted due date, Central time (Chicago time)
- Multiple submissions allowed and encouraged. Only last submission graded (No penalty for repeat submission)
- Up to 24 hours late: 50% penalty
- After 24 hours late: Not accepted
- Medical emergencies: Please contact CS department Associate Chair with supporting documentation

Communication Channels

- Online discussion is encouraged. We will use **Blackboard** discussion groups
- Please direct general interest questions on course contents and homework to online forums so others may benefit
- Instructor Office Hours: Refer to syllabus on Blackboard
- TA Office Hours: TBA

Academic Honesty

- If you violate the academic honesty policy (such as unauthorized/undocumented collaboration, cheating, etc.), I **have to** report it to the university
- Depending on the severity of the violation, it can result in
  - zero points on the respective assignment,
  - E in the course,
  - suspension from the university,
  - expulsion from the university
- Full guidelines: https://web.iit.edu/student-affairs/handbook/fine-print/code-academic-honesty

About Me

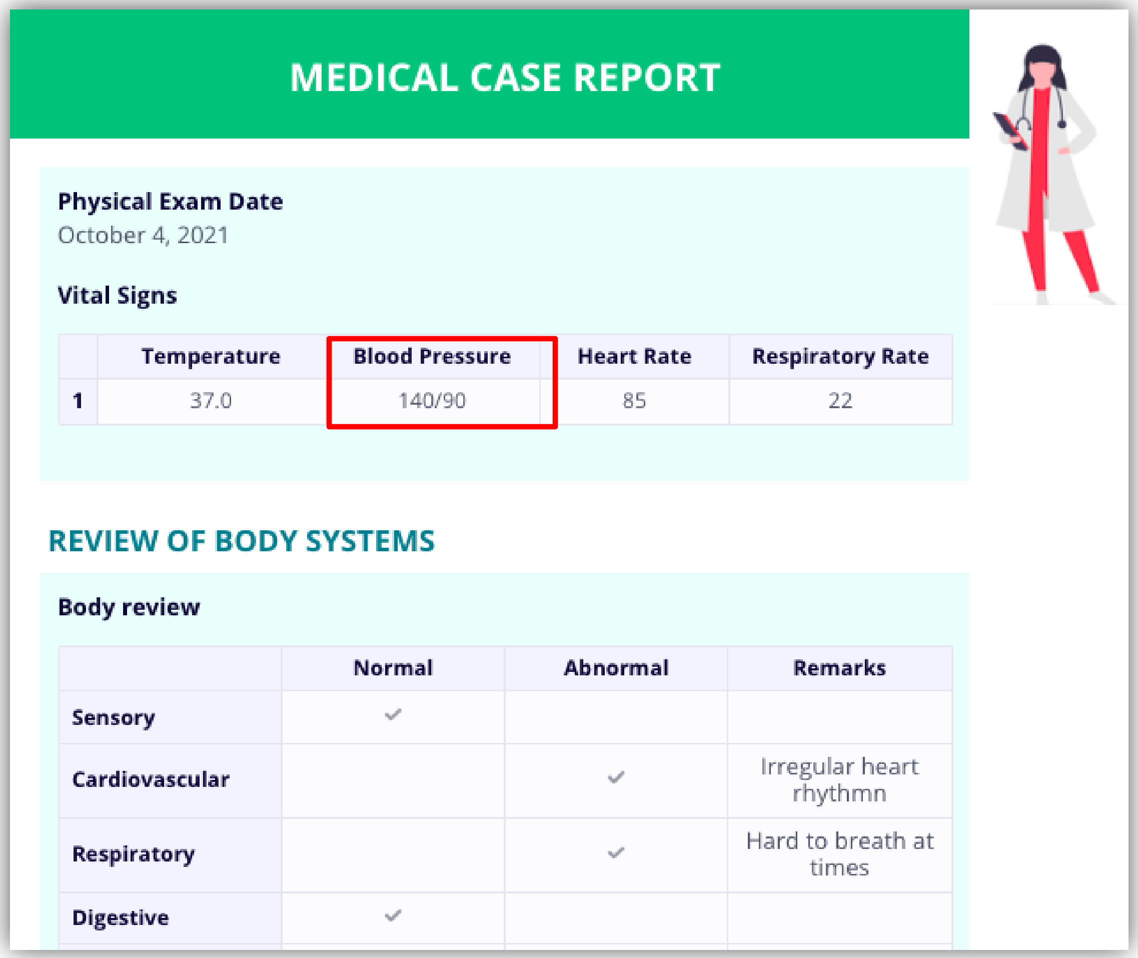
NLP and Machine Learning for Insurance

"The total cost of insurance fraud (non-health insurance) is estimated to be more than $40 billion per year. That means insurance fraud costs the average U.S. family between $400 and $700 per year in the form of increased premiums."

https://www.bankrate.com/insurance/car/fraud-facts-statistics/

About Me

Text Extraction on Medical Charts



MEDICAL CASE REPORT

Physical Exam Date
October 4, 2021

Vital Signs

	Temperature	Blood Pressure	Heart Rate	Respiratory Rate
1	37.0	140/90	85	22

REVIEW OF BODY SYSTEMS

Body review

	Normal	Abnormal	Remarks
Sensory	✓		
Cardiovascular		✓	Irregular heart rhythmn
Respiratory		✓	Hard to breath at times
Digestive	✓		

GETTING TO KNOW YOU

Questions for you

How many of you

- …know python?
- …have worked with Unix shell?
- …have taken a data mining/machine learning/social media analysis course?
- …have built a statistical or machine learning model?

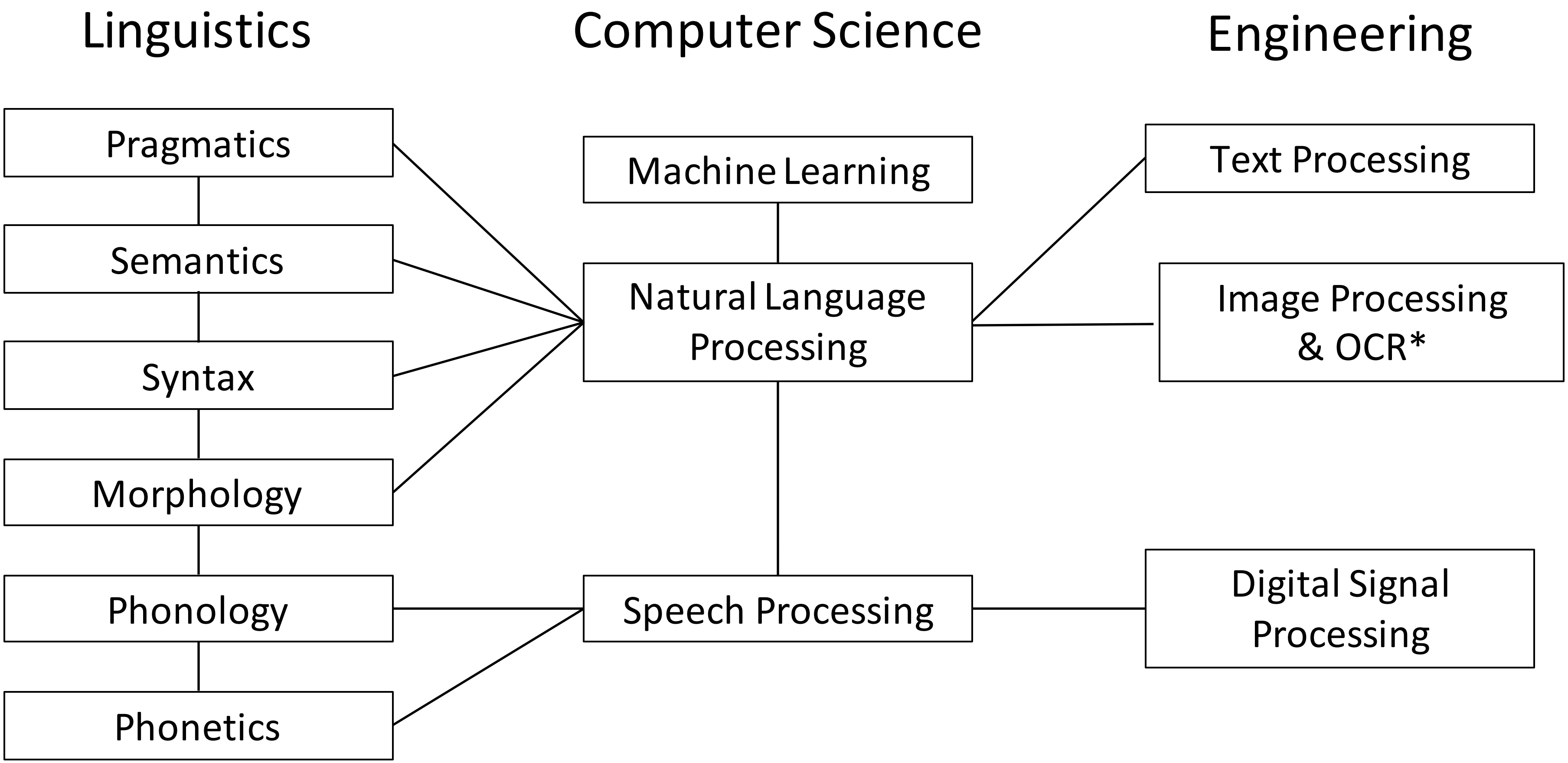
LANGUAGE, LINGUISTICS AND NLP

Some terminology

- **Natural language processing**: The study of methods for *exploiting or generating* language represented as text, for practical tasks

- **Computational linguistics**: The use of computational tools to understand or learn the *structure of human languages*

Related fields



*OCR: Optical Character Recognition

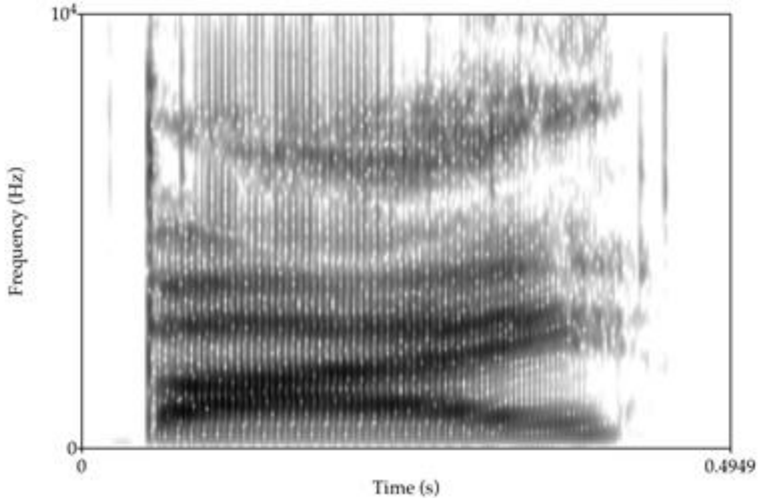
Phonetics

The study of speech sounds

- ***Articulatory*** phonetics deals with the physiological speech process
- ***Acoustic*** phonetics deals with the sound waves produced

Applications:

- Speech recognition
- Speech synthesis
- Clinical speech pathology



https://en.w ikiversity.org/wiki/Psycholinguistics/Acoustic_Phonetics #/media/File:Spectrogram-buy.png

Phonology

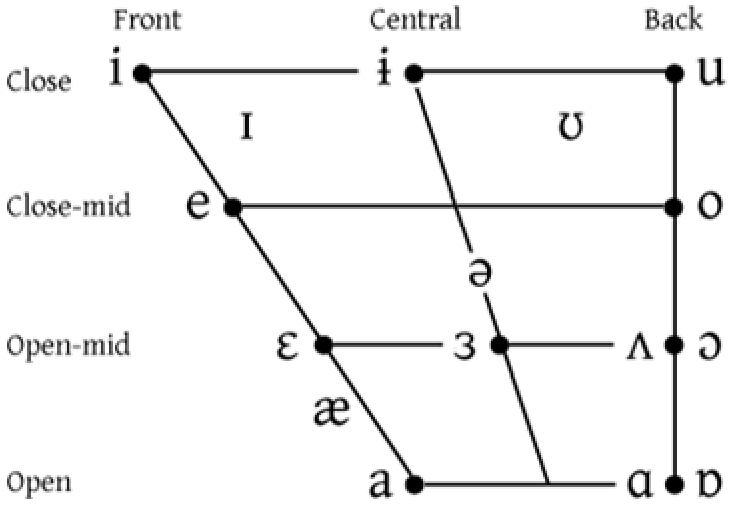
The structure and patterning of ***sounds*** within a language

- ***Segmental*** phonology deals with phonemes (minimal contrastive units)
- ***Supra-segmental*** phonology deals with tones, prosody and stress accent
- ***Sub-segmental*** phonology deals with features of phonemes

Applications:

- Speech recognition
- Speech synthesis

Pen	/pɛn/
Pan	/pæn/

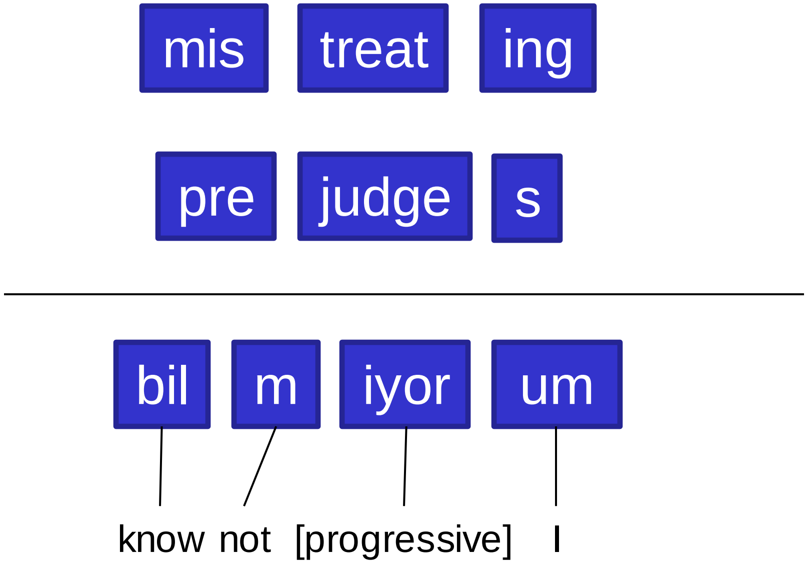


https://commons.wikimedia.org/w/index.php?curid=18555461

Morphology

The internal structure of words

- ***Morphemes*** include stems, prefixes, suffixes and infixes



mis treat ing

pre judge s

bil m iyor um

know not [progressive] I

Applications

- Stemming / lemmatization
- Compound breaking
- Inflection generation (NLG)

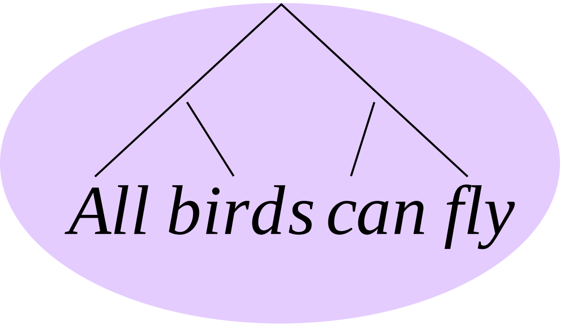
Syntax

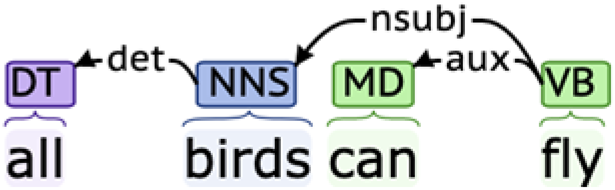
The structure of ***words and phrases*** within a sentence

- Different formalisms, coming from the American (***phrase structure***) and European (***dependency grammar***) structuralist traditions

Applications

- Part-of-speech tagging
- Entity extraction
- Syntactic parsing (CFG)
- Syntactic parsing (dependencies)





Semantics

The representation of ***meaning*** in language

- At different levels: lexical, sentential, textual
- Logical formalisms: reference and truth conditions

Applications:

- Word embedding/encoding
- Lexical resources
- Semantic role labeling

$\forall x(\text{bird}(x) \rightarrow \text{fly}(x))$

$\text{kill}(x, y) :=$

$\text{Cause}(x, \text{Become}(\neg\text{Alive}(y)))$

Pragmatics

How language is used to achieve specific ***intentions***

- Conversational implicatures: how I **interpret** what you say because of what I assume you're trying to do
- Speech acts

Applications:

- Speech act labeling
- Discourse structure parsing
- Dialogue systems

"I ate <u>most</u> of your cookies"

$\models$

I did not eat <u>all</u> of your cookies

"Where does your brother live?"

$\models$

I do not know where your brother lives

Sociolinguistics

Language use patterns associated with particular ***groups***, or language used to communicate status relative to a ***group***

Applications:

- Stylometrics / authorship attribution
- Forensic linguistics
- Natural language generation

With friends

In class

With family

With strangers

With professional colleagues

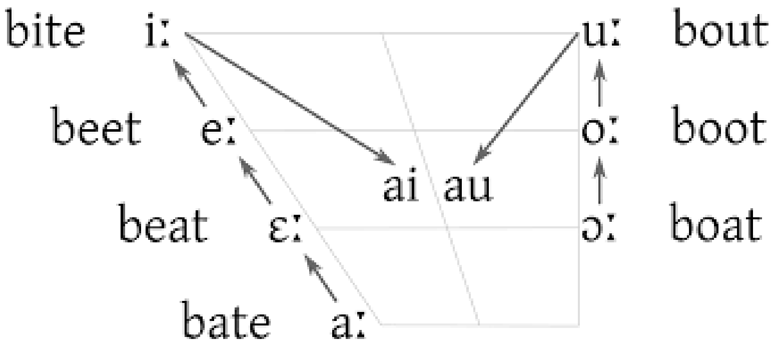
Historical Linguistics

Language change over time

- Lexical innovation
- Phonological change
- Language contact

Applications:

- Linguistic typology
- Digital humanities



https://en.wikipedia.org/wiki/File:Great_Vowel_Shift2c.svg

Psycholinguistics

Language as a ***cognitive*** function

- Role of brain areas in language production and processing
- Language learning

Applications:

- Language pathology
- Assistive technology

http://arikaokrent.com/bio.html

And of course…

Not all NLP tasks relate to a single linguistic domain.

E.g., machine translation involves morphology, syntax, semantics and pragmatics, …

Why is NLP hard?

- The "hidden structure" of language is **ambiguous** at all levels!

- Consider the simple proverb:

Time flies like an arrow

Word sense ambiguity

Time: “abstract time”, “a specific point in time”, “to measure time”

flies: “moves through the air”, “little pesky insects”

like: “similar to”, “have affection for”

arrow: “pointy stick shot from a bow”, “to move straight towards a target”

Time flies like an arrow

Part of speech ambiguity

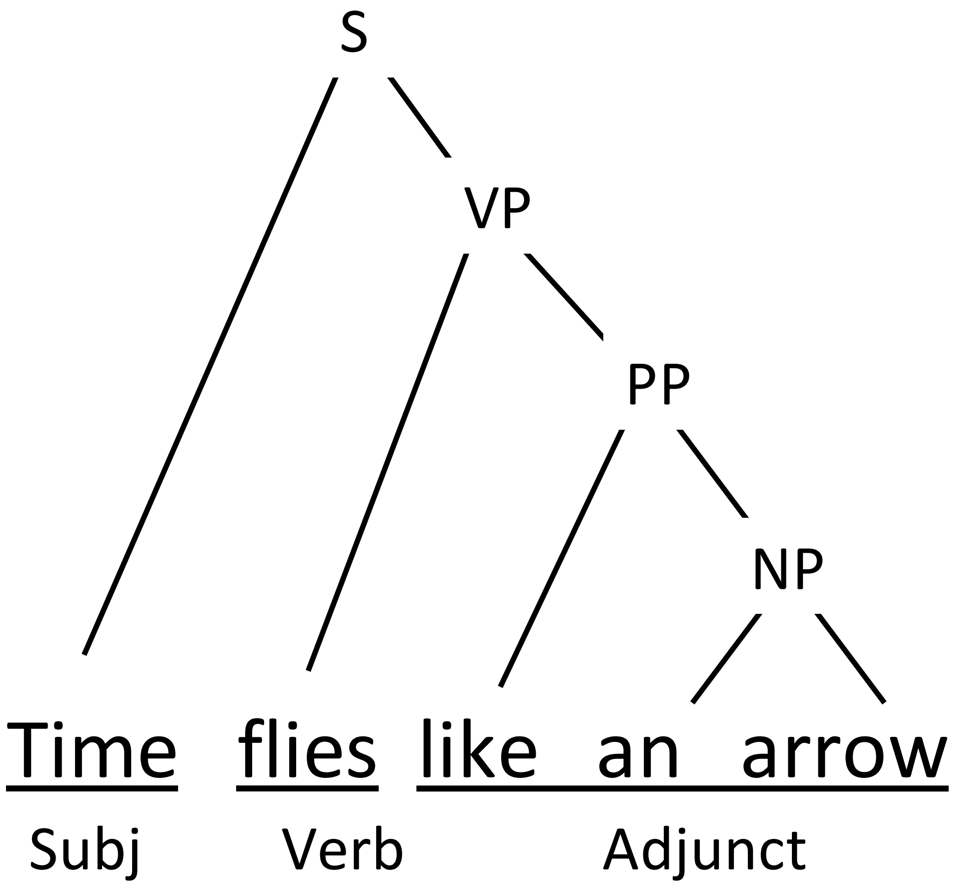
		JJ		
		VB		
VB	NNS	NN		VB
NN	VBZ	IN	DT	NN
<u>Time</u>	<u>flies</u>	<u>like</u>	<u>an</u>	<u>arrow</u>

```
Part-of-speech tags

JJ: Adjective
VB: Verb, base form
NN: Noun, singular
DT: Determinant
...
```

https://cs.nyu.edu/~grishman/jet/guide/PennPOS.html

Syntactic ambiguity



S

VP

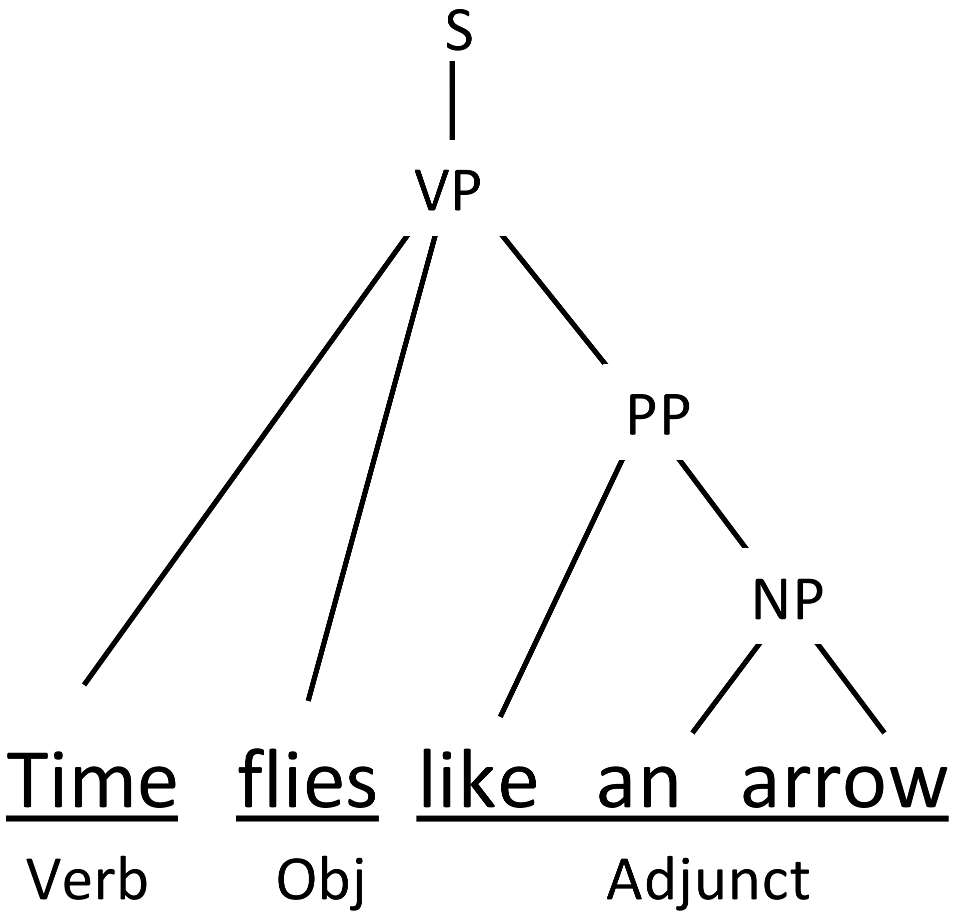
PP

NP

Time flies like an arrow

Subj Verb Adjunct

Syntactic ambiguity



S

VP

PP

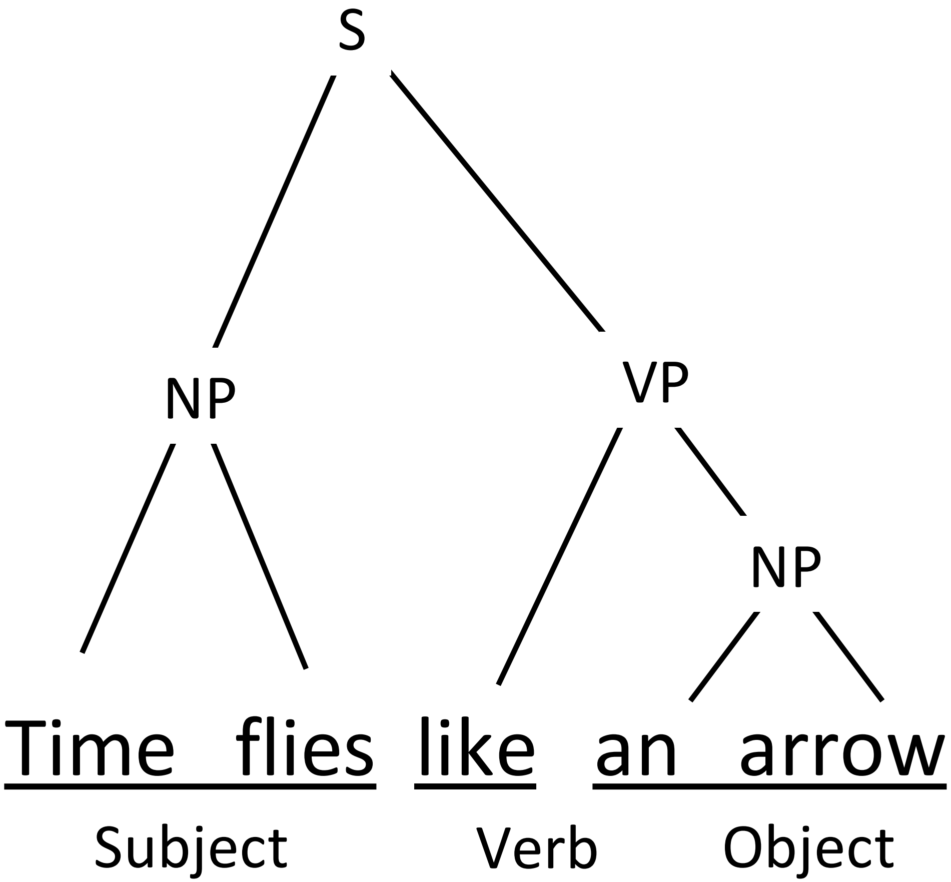
NP

Time flies like an arrow

Verb Obj Adjunct

...instead of timing them like a snail!

Syntactic ambiguity



S

NP

VP

NP

Time flies like an arrow

Subject Verb Object

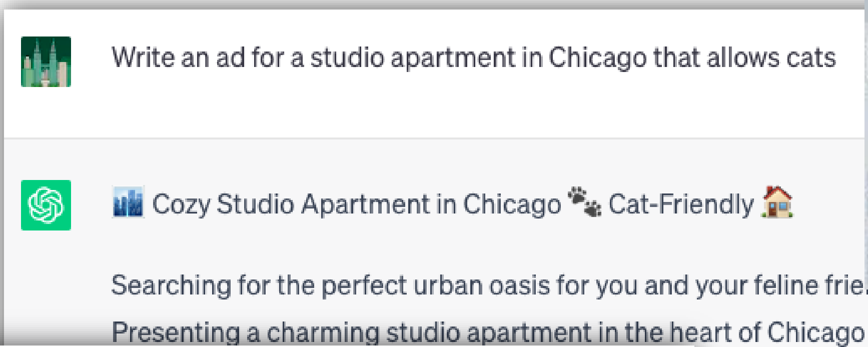
...but fruit flies like a banana!

A Changing Target

- Neologisms (= new words/phrases):
  - *cosmocrat, technocrat, davos man*
  - *megacryometeor*
  - *flash mob, carjack*
  - *googling, spam, blogger, wi-fi*
  - *kleptocracy, identity theft*
  - *just-in-time learning, egoboo*
- Also sentence structure, though it’s subtler...

Such a great time to get into NLP!

There is so much we can do now!

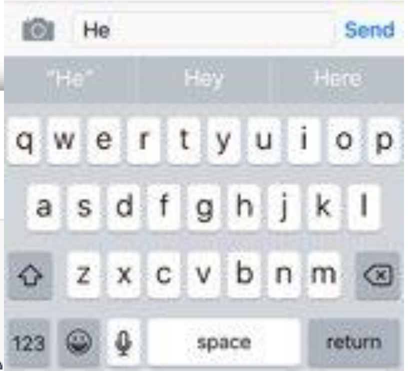


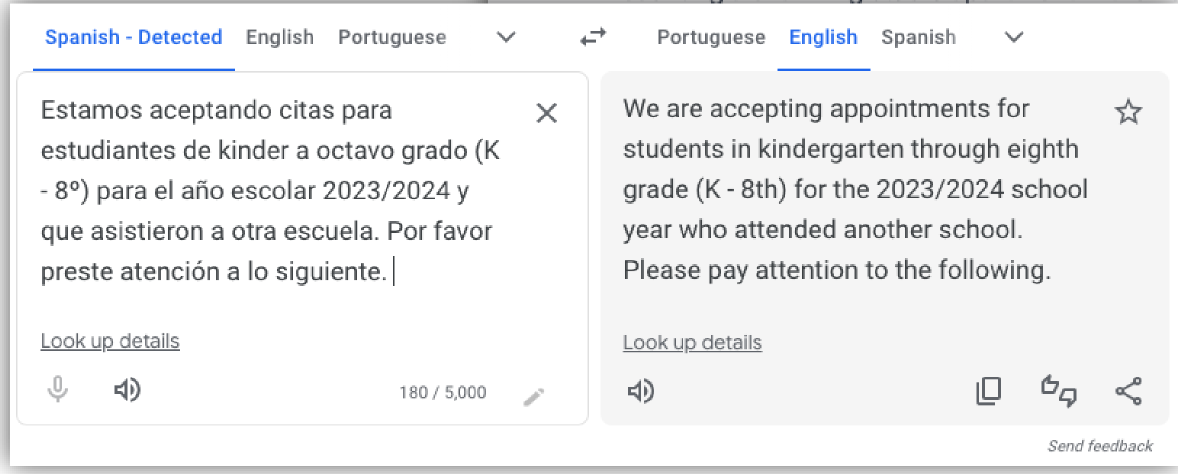
Write an ad for a studio apartment in Chicago that allows cats

🏙️ Cozy Studio Apartment in Chicago 🐾 Cat-Friendly 🏠

Searching for the perfect urban oasis for you and your feline frie

Presenting a charming studio apartment in the heart of Chicago that's ready to become your





Estamos aceptando citas para estudiantes de kinder a octavo grado (K - 8º) para el año escolar 2023/2024 y que asistieron a otra escuela. Por favor preste atención a lo siguiente.

We are accepting appointments for students in kindergarten through eighth grade (K - 8th) for the 2023/2024 school year who attended another school. Please pay attention to the following.