CS-585
Natural Language Processing

Prof. Derrick Higgins
dhiggins1@iit.edu

Today

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

THIS COURSE

Goals

- **Breadth of coverage**: Familiarity with a wide range of task types and methods in natural language processing
- **Depth in key areas**: Mastery of critical concepts and algorithms for NLP
- **Preparedness for further study**: introduction to deep learning frameworks as applied to NLP, such that current research papers will be accessible

Prerequisites

- Math
  - Linear algebra
  - Probability
- Programming
  - Python 3
  - Basic algorithms and data structures
  - Access to a Unix system

Methods

- Exams
  - Open-book, multiple choice (mostly)
  - No electronics allowed
  - Midterm will cover material through October 5
  - Final will cover material from the entire course
- Class Project
  - We will create a new dataset for text categorization
  - Three parts
    1. Data labeling
    2. Annotation analysis
    3. NLP Modeling

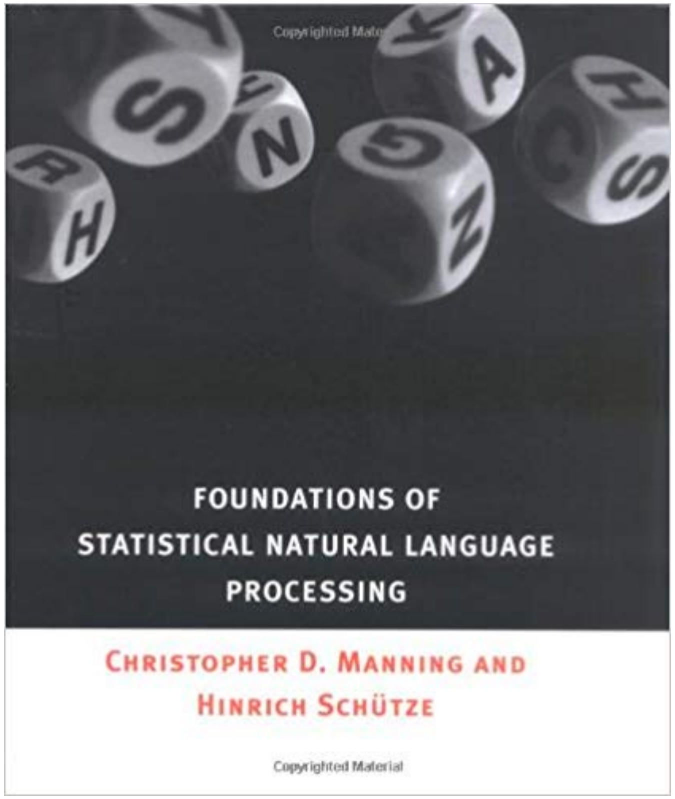
Grading: Available Points

Assignment	Points
Project part 1: Annotation	175
Project part 2: Analysis	150
Project part 3: Modeling	175
Midterm	200
Final	300
Total:	**1000**

Grading: Letter Grades

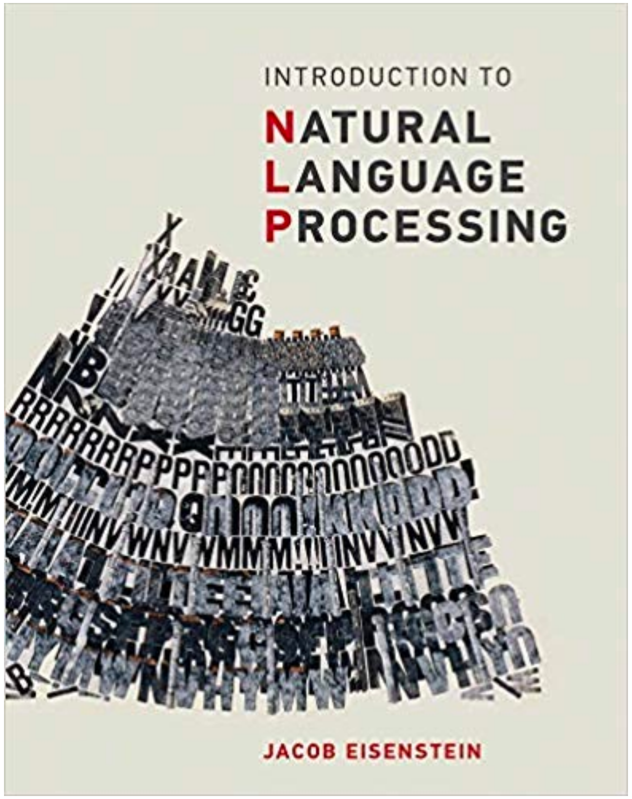
Points	Grade
900-1000	A
750-899	B
550-749	C
0-550	E

Readings



FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE



INTRODUCTION TO NATURAL LANGUAGE PROCESSING

JACOB EISENSTEIN

Lecture plan

1. Foundational concepts
2. Consideration of progressively higher levels of linguistic structure
3. Connection to neural networks at end of each unit

Academic Honesty

- If you violate the academic honesty policy (such as unauthorized/undocumented collaboration, cheating, etc.), I **will** 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

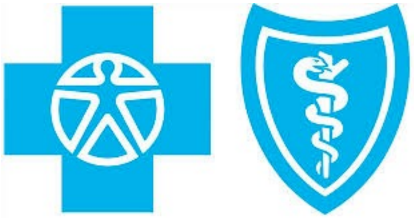
Americans With Disabilities Act

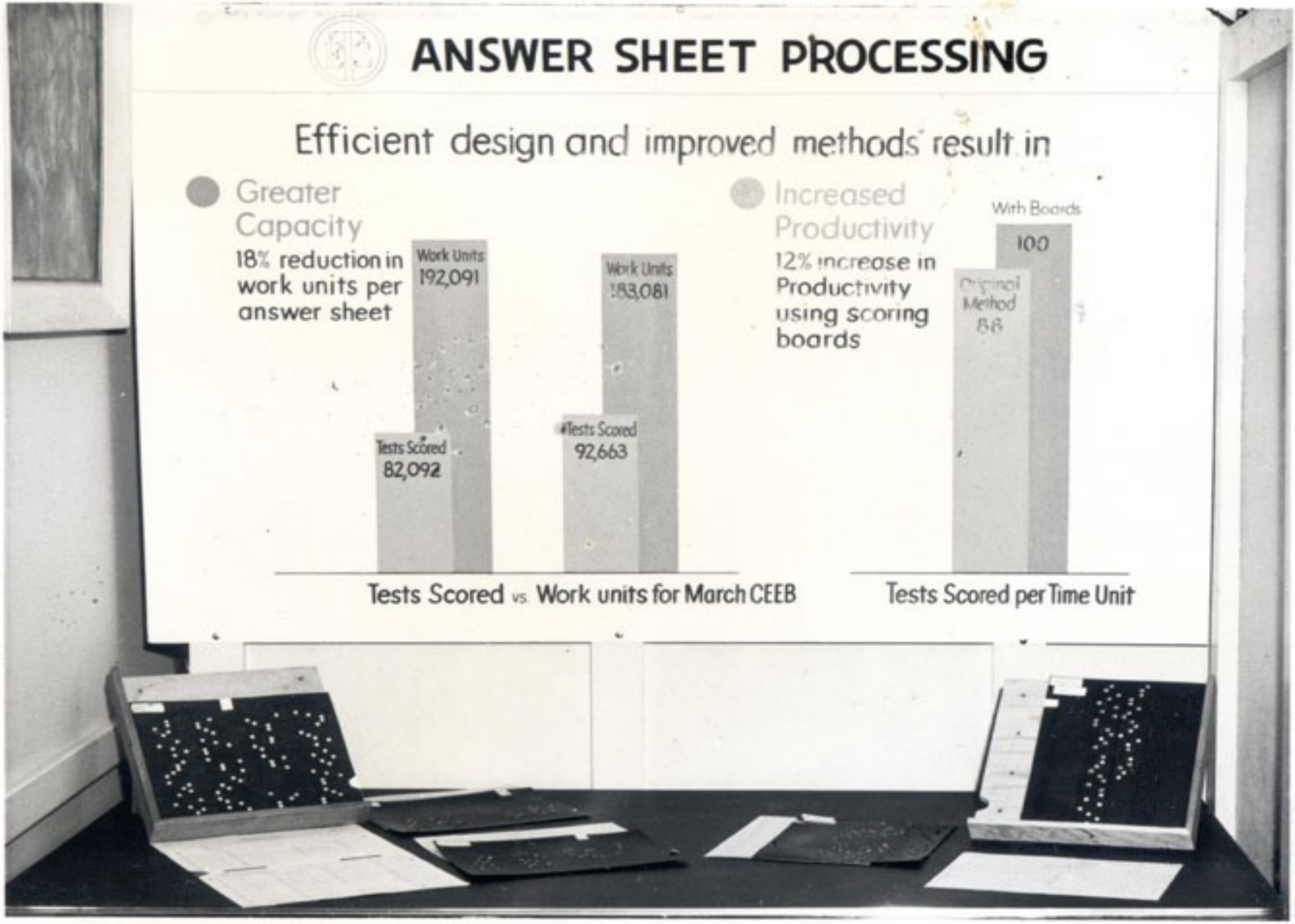
Reasonable accommodations will be made for students with documented disabilities. In order to receive accommodations, students must obtain a letter of accommodation from the Center for Disability Resources.

The Center for Disability Resources (CDR) is located in 3424 S. State St., room 1C3-2 (on the first floor), telephone: 312.567.5744 or disabilities@iit.edu

INTRODUCTION

Me





ANSWER SHEET PROCESSING

Efficient design and improved methods result in

Greater Capacity

18% reduction in work units per answer sheet

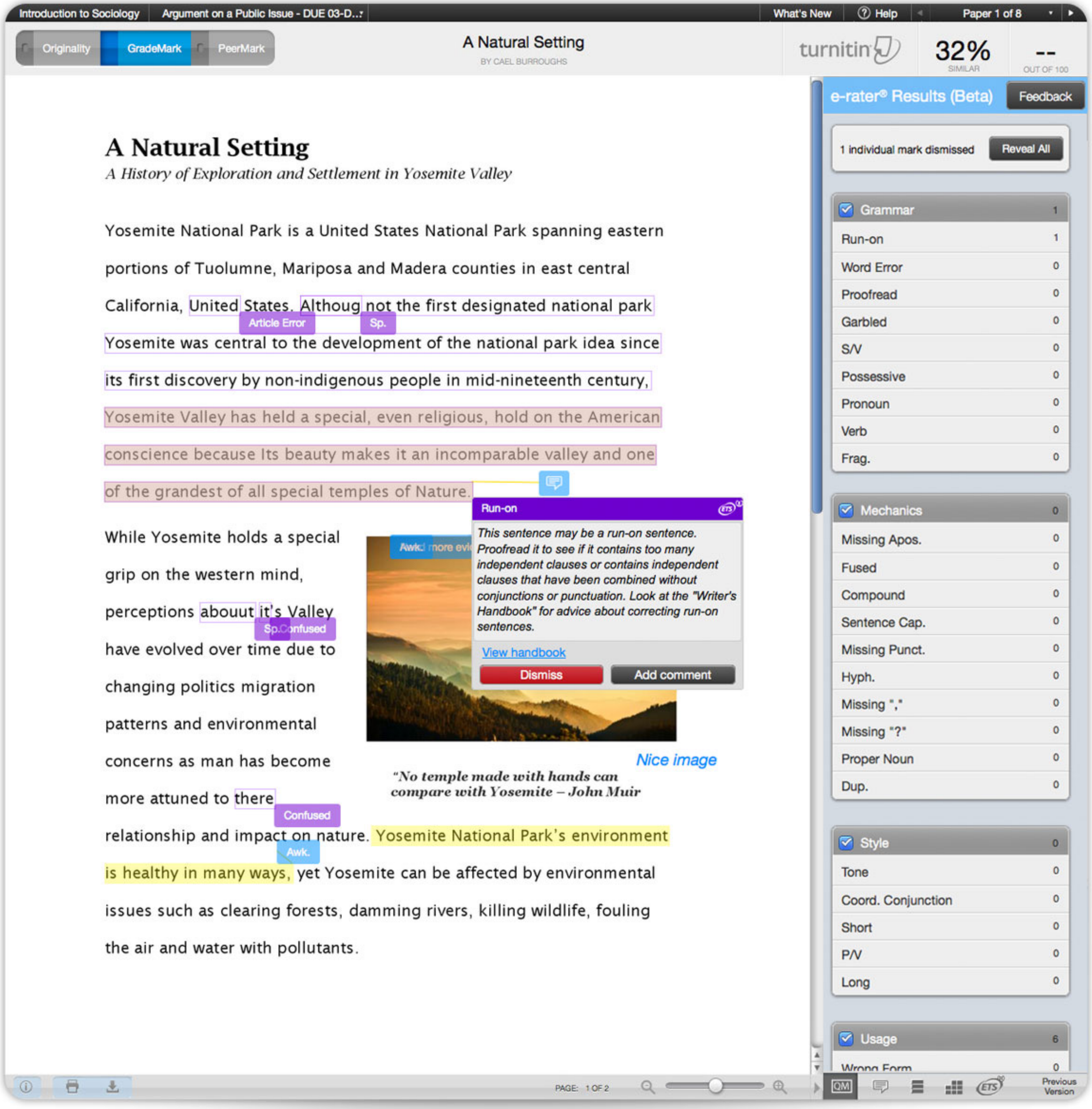
Tests Scored vs. Work units for March CEEB

Increased Productivity

12% increase in Productivity using scoring boards

Tests Scored per Time Unit

Open-ended Written Responses



A Natural Setting

BY CAEL BURROUGHS

A Natural Setting

A History of Exploration and Settlement in Yosemite Valley

Yosemite National Park is a United States National Park spanning eastern portions of Tuolumne, Mariposa and Madera counties in east central California, United States. Althoug not the first designated national park Yosemite was central to the development of the national park idea since its first discovery by non-indigenous people in mid-nineteenth century, Yosemite Valley has held a special, even religious, hold on the American conscience because Its beauty makes it an incomparable valley and one of the grandest of all special temples of Nature.

Run-on

This sentence may be a run-on sentence. Proofread it to see if it contains too many independent clauses or contains independent clauses that have been combined without conjunctions or punctuation. Look at the "Writer's Handbook" for advice about correcting run-on sentences.

View handbook

Dismiss | Add comment

While Yosemite holds a special grip on the western mind, perceptions abouut it's Valley have evolved over time due to changing politics migration patterns and environmental concerns as man has become more attuned to there relationship and impact on nature. Yosemite National Park's environment is healthy in many ways, yet Yosemite can be affected by environmental issues such as clearing forests, damming rivers, killing wildlife, fouling the air and water with pollutants.

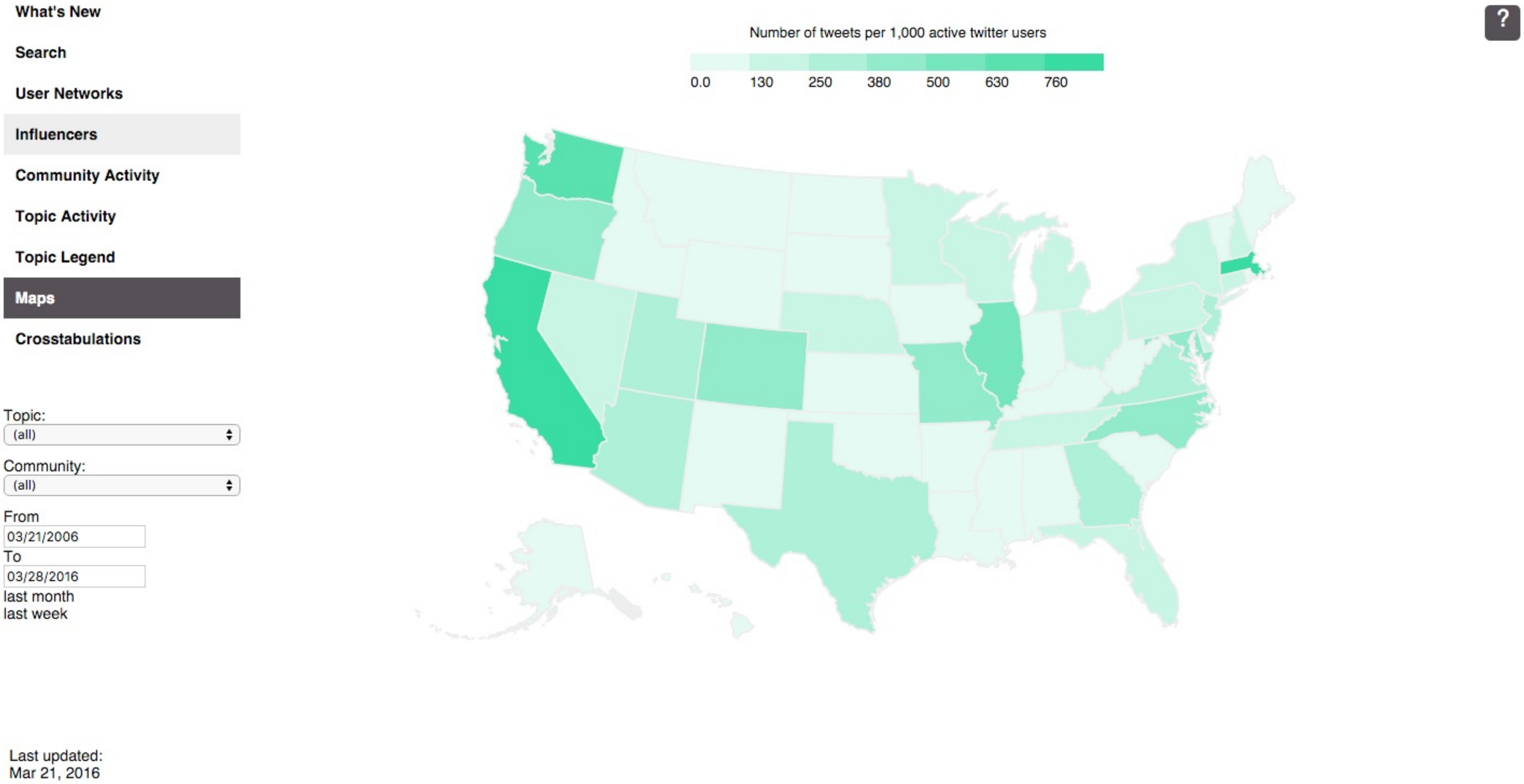
"No temple made with hands can compare with Yosemite – John Muir

1 individual mark dismissed | Reveal All

☑ Grammar	1
Run-on	1
Word Error	0
Proofread	0
Garbled	0
S/V	0
Possessive	0
Pronoun	0
Verb	0
Frag.	0
☑ Mechanics	0
Missing Apos.	0
Fused	0
Compound	0
Sentence Cap.	0
Missing Punct.	0
Hyph.	0
Missing ","	0
Missing "?"	0
Proper Noun	0
Dup.	0
☑ Style	0
Tone	0
Coord. Conjunction	0
Short	0
P/V	0
Long	0
☑ Usage	6
Wrong Form	0

- Cf. Shermis, Mark, Jill Burstein, Derrick Higgins & Klaus Zechner. (2009). Automated essay scoring: Writing assessment and instruction. International Encyclopedia of Education, Third Edition. United Kingdom, Elsevier.

Data Science for Data Science



Topic: (all)

Community: (all)

From 03/21/2006

To 03/28/2016

last month

last week

Last updated:
Mar 21, 2016

OUR VALUE

02 /04

AMERICAN FAMILY INSURANCE

Claims / Underwriting

American Family utilizes Arturo data throughout multiple parts of their business to improve underwriting performance, identify risk within their book by identifying changes, and predicting the right resources necessary to respond to claim events.

GETTING TO KNOW YOU

Questions for you

How many of you

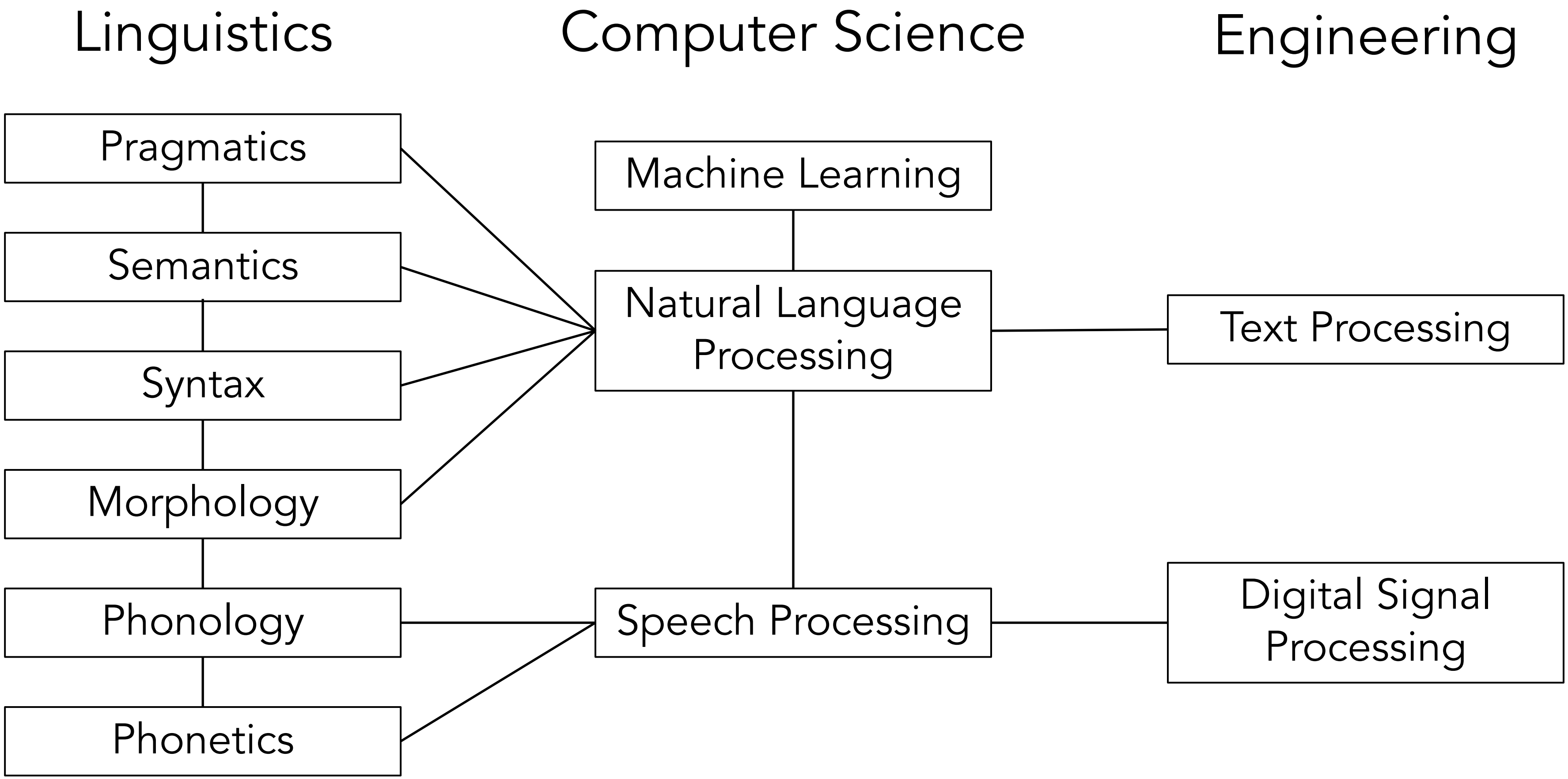
- …know python?
- …have worked with Unix shell?
- …have taken a machine learning course?

LANGUAGE, LINGUISTICS AND NLP

Some terminology

- **Text processing**: Engineering practices for transforming, normalizing, compressing or accessing textual data
- **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
- **Speech processing**: The study of methods for exploiting or generating language represented as audible waveforms, for practical tasks

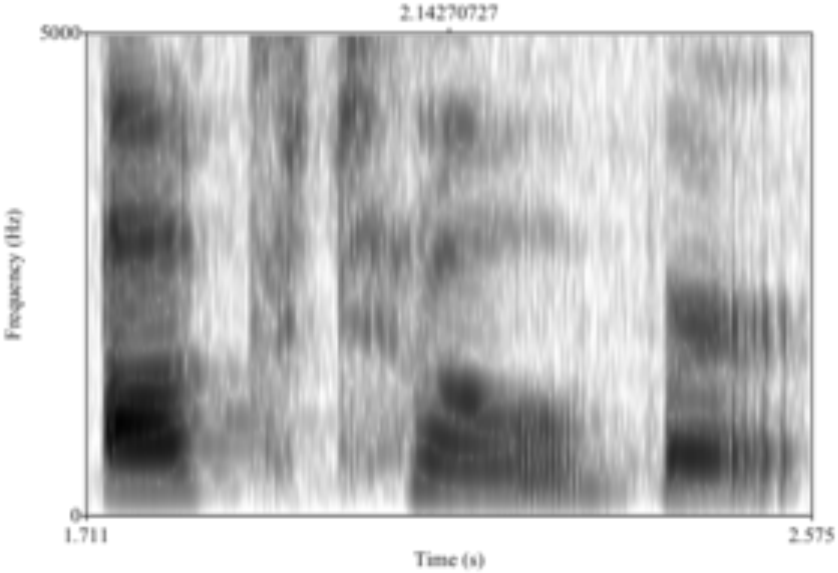
Adjacent fields



Phonetics

The study of speech sounds

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



https://commons.wikimedia.org/w/index.php?curid=14508443

Applications:

- Speech recognition
- Speech synthesis
- Clinical speech pathology

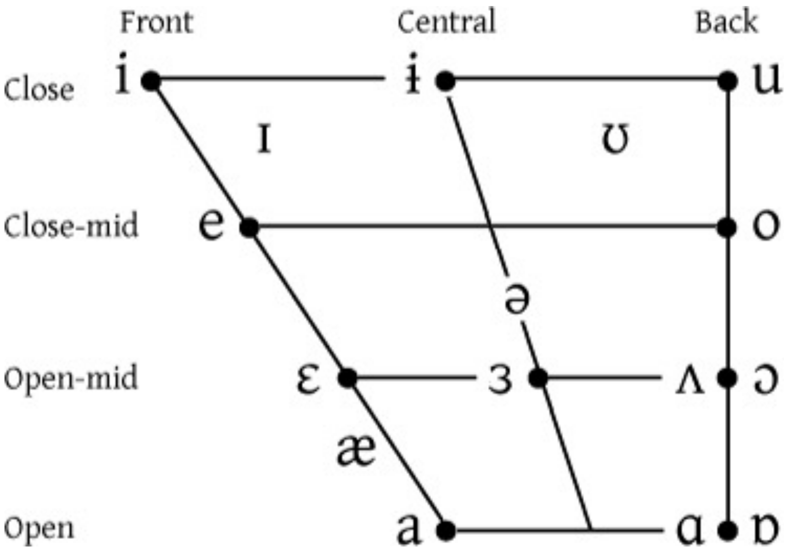
Phonology

The structure and patterning of sounds within a language

- Segmental phonology deals with phonemes (minimal contrastive units)
- Suprasegmental phonology deals with tones, prosody and stress accent
- Subsegmental phonology deals with features of phonemes

Applications:

- Speech recognition
- Speech synthesis

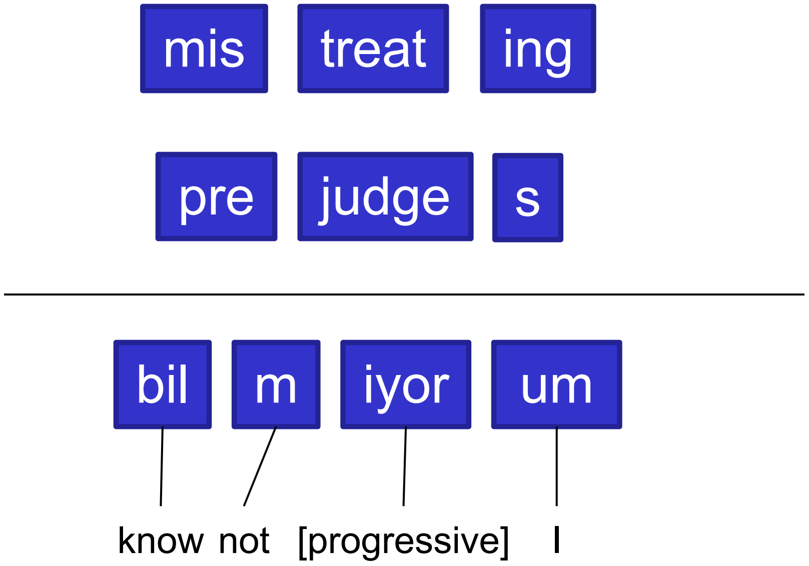


https://commons.wikimedia.org/w/index.php?curid=18555461

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

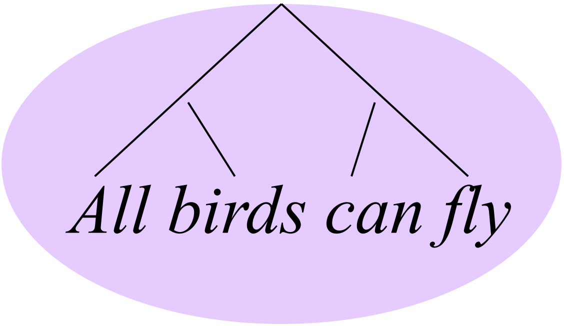
Applications

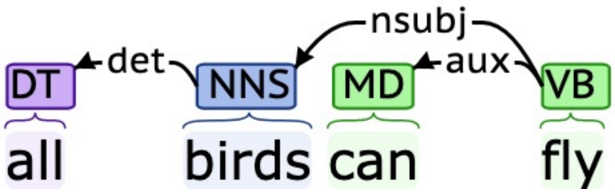
- Stemming / lemmatization
- Compound breaking
- Inflection generation (NLG)

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)

ILLINOIS INSTITUTE OF TECHNOLOGY

Transforming Lives.Inventing the Future.www.iit.edu

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 most of your cookies"

$\models$

I did not eat all of your cookies

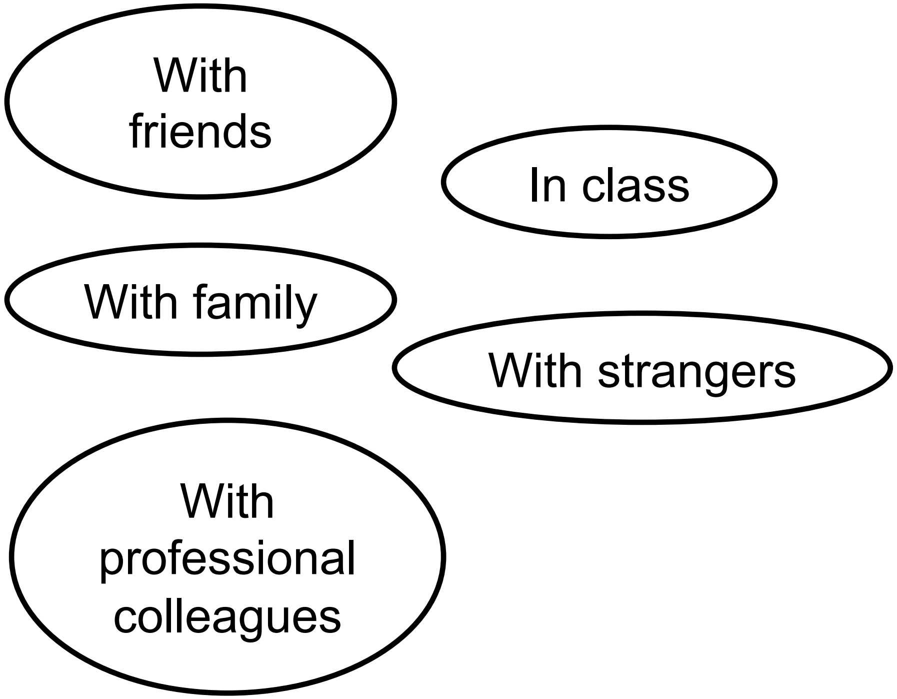
"Where does your brother live?"

$\models$

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



With friends

In class

With family

With strangers

With professional colleagues

Applications:

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

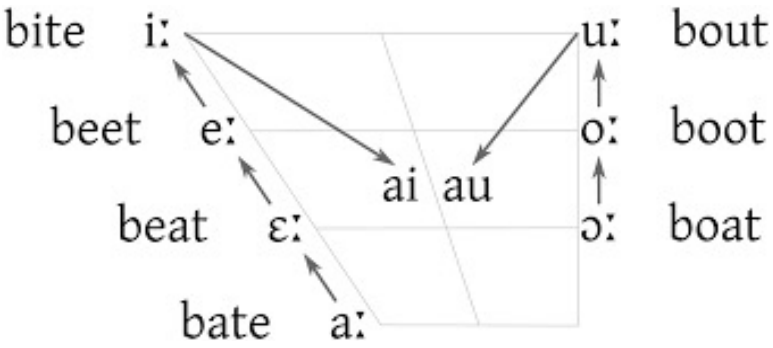
Historical Linguistics

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 (at least)

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"

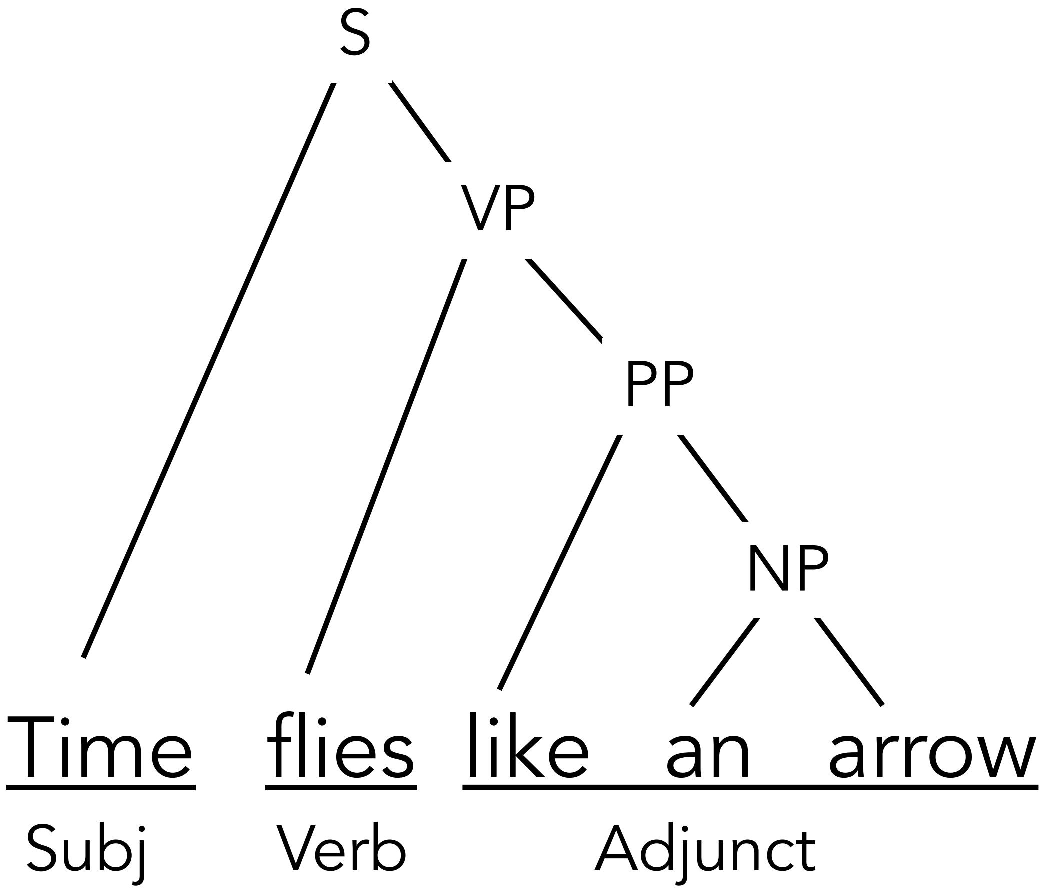
arrow: "pointy stick shot from a bow", "to move straight towards a target"

Time flies like an arrow

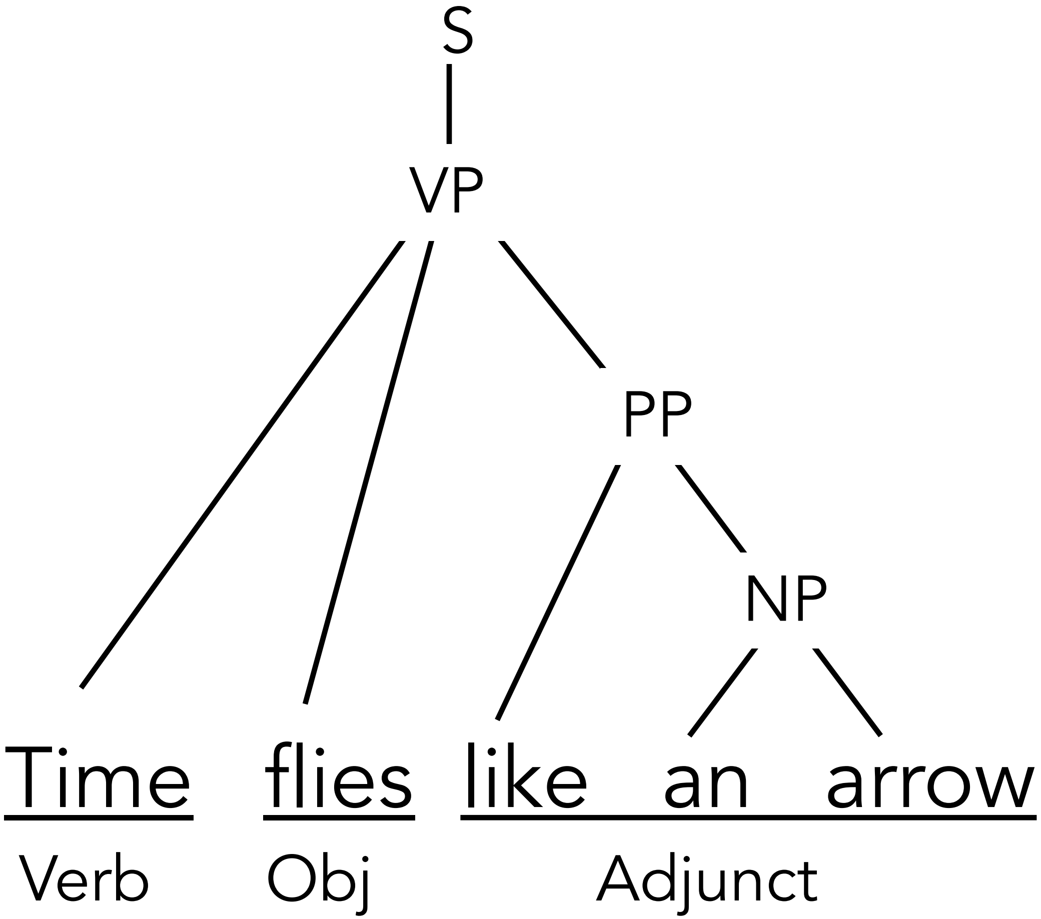
Part of speech ambiguity

		JJ		
		VB		
VB	NNS	NN		VB
NN	VBZ	IN	DT	NN
Time	flies	like	an	arrow

Syntactic ambiguity

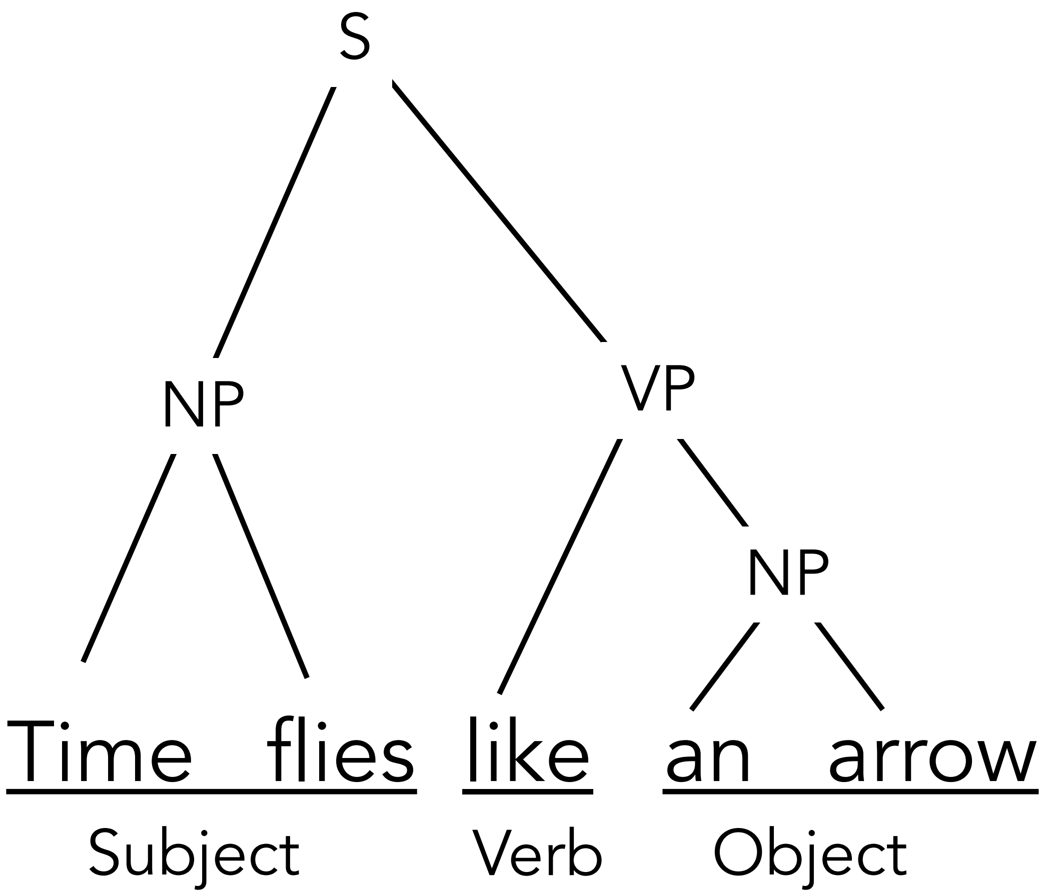


Syntactic ambiguity



...instead of timing them like a snail!

Syntactic ambiguity



...but fruit flies like a banana!

Newspaper Headlines

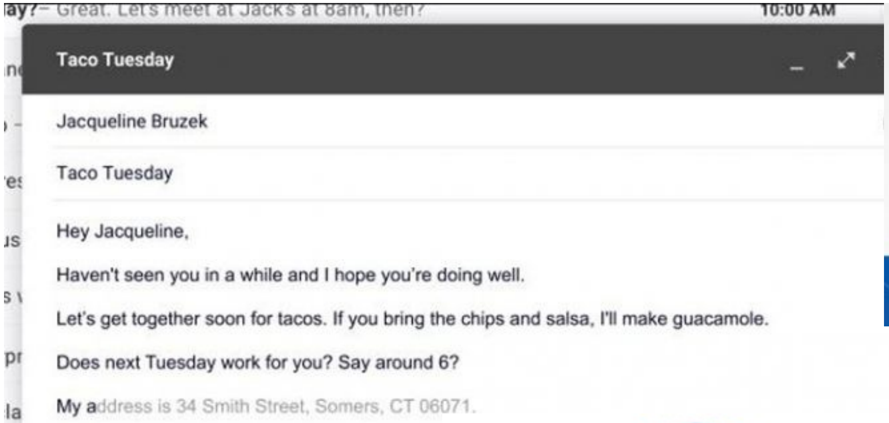
- Ban on Nude Dancing on Governor's Desk
  from a discussion of current legislation
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Kids Make Nutritious Snacks
- Hospitals Sued by Seven Foot Doctors

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!



ay?– Great. Let's meet at Jack's at 8am, then? 10:00 AM

Taco Tuesday

Jacqueline Bruzek

Taco Tuesday

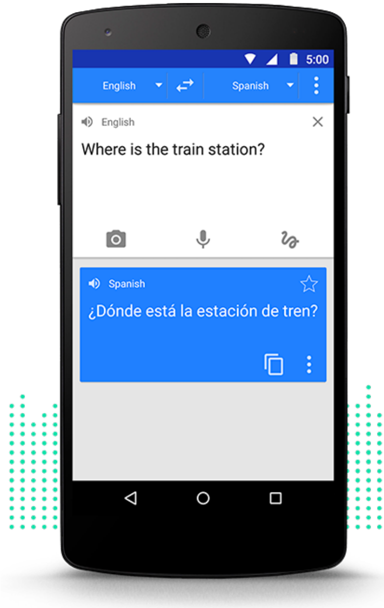
Hey Jacqueline,

Haven't seen you in a while and I hope you're doing well.

Let's get together soon for tacos. If you bring the chips and salsa, I'll make guacamole.

Does next Tuesday work for you? Say around 6?

My address is 34 Smith Street, Somers, CT 06071.

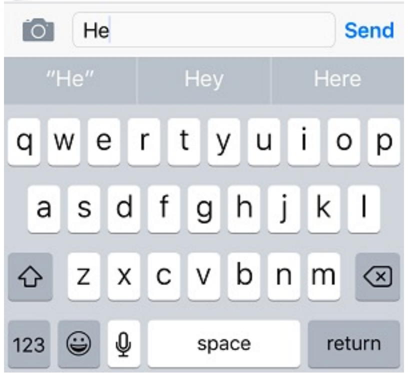


English

Where is the train station?

Spanish

¿Dónde está la estación de tren?



He

Such a great time to get into NLP!

- There is so much we <u>still can't</u> do!
  - Handle real-world knowledge and logical inferences
  - Deal with limited-data contexts and low resource languages
  - Transfer learning across tasks and domains (although we're getting better)
  - Integrate information across modalities: text, imagery, action sequences
  - Infer linguistic structure without manual labeling based on human judgements

DEMO: SLACK

Slack Channel Usage

- `#general` channel
  - Ask clarification questions publicly so that everyone can benefit from the answers
  - Email OK for personal concerns