

CSE628 Natural Language Processing

Instructor: Yejin Choi

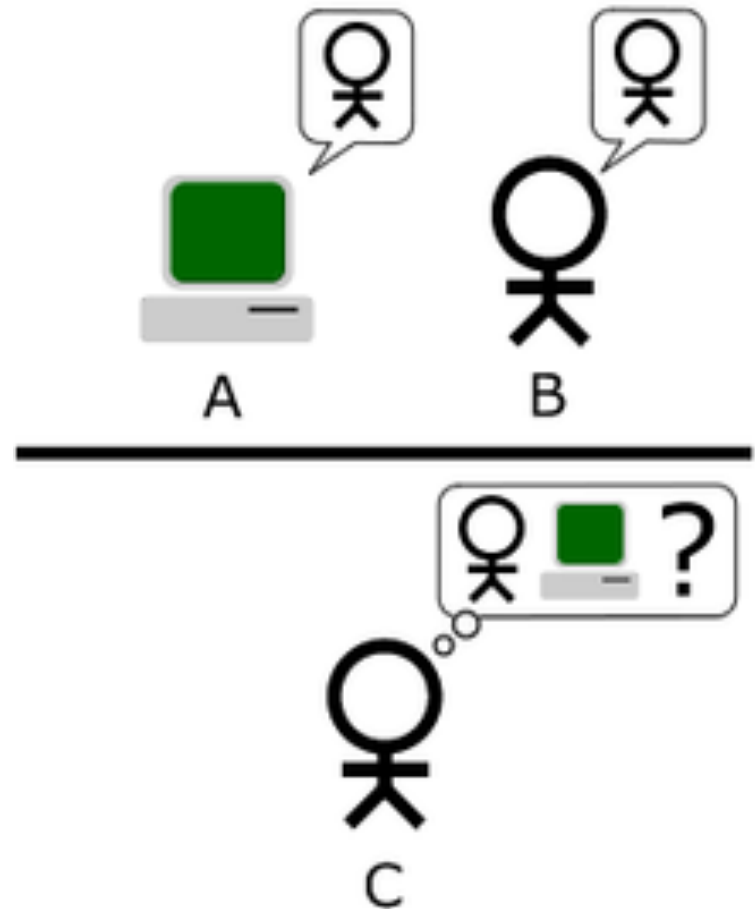
<http://www.cs.stonybrook.edu/~ychoi/cse628>

Please fill out answers for below questionnaire

1. Your name & email
2. Masters? Phd?
3. Planning to take the class? Or audit? Or haven't decided?
4. Have you taken either "artificial intelligence" or "machine learning" or other NLP or linguistics classes?
5. What is your area of interest outside NLP? e.g. systems, theory, etc
6. Why are you taking this class?

What is NLP?

- Artificial Intelligence dealing with human language.
- NLP is AI-Complete.
- **Turing Test:** Interrogator 'c' engages in a natural language conversation with 'a' and 'b' to determine which is a computer and which is a human.



What we say to dogs



What they hear



Why is NLP hard?

Reason (1) – human language is ambiguous.

- Task: Pronoun Resolution
 - Jack drank the wine on the table. **It** was red and round.
 - Jack saw Sam at the party. **He** went back to the bar to get another drink.
 - Jack saw Sam at the party. **He** clearly had drunk too much.

[Adapted from Wilks (1975)]

Why is NLP hard?

Reason (1) – human language is ambiguous

- Task: Preposition Attachment (aka PP-attachment)

- I ate the bread with pecans.



- I ate the bread with fingers.



Why is NLP hard?

Reason (2) – requires reasoning beyond what is explicitly mentioned **(A,B)** , and some of the reasoning requires world knowledge **(C)**

I couldn't submit my homework because my horse ate it.

Implies that...

- A. I have a horse.*
- B. I did my homework.*
- C. My homework was done on a soft object (such as papers) as opposed to a hard/heavy object (such as a computer).
– it's more likely that my horse ate papers than a computer.*

Why is NLP hard?

Reason (3) – Language is difficult even for human.

- Learning mother tongue (native language)
 - you might think it's easy, but...
 - compare 5 year old V.S. 10 year old V.S. 20 year old
- Learning foreign languages
 - even harder

Is NLP really that hard?

In the back of your mind, if you're still thinking...

“My native language is so easy. How hard can it be to type all the grammar rules, and idioms, etc into a software program? Sure it might take a while, but with enough people and money, it should be doable!”

You are not alone!

Brief History of NLP

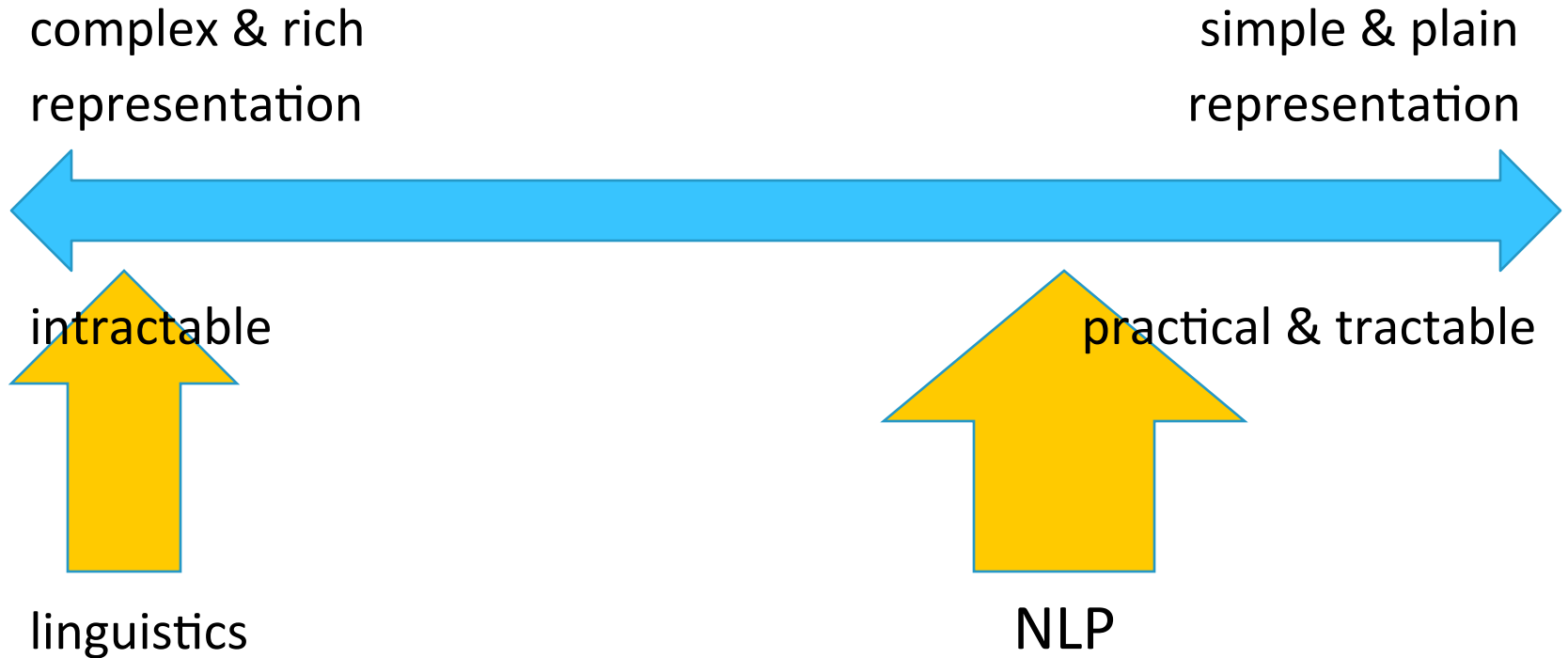
- Mid 1950's – mid 1960's: Birth of NLP and Linguistics
 - At first, people thought NLP is easy! Researchers predicted that “machine translation” can be solved in 3 years or so.
 - Mostly hand-coded rules / linguistics-oriented approaches
 - The 3 year project continued for 10 years, but still no good result, despite the significant amount of expenditure.
- Mid 1960's – Mid 1970's: A Dark Era
 - After the initial hype, a dark era follows -- people started believing that machine translation is impossible, and most abandoned research for NLP.

Brief History of NLP

- 1970's and early 1980's – Slow Revival of NLP
 - Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation
- Late 1980's and 1990's – Statistical Revolution!
 - By this time, the computing power increased substantially .
 - Data-driven, statistical approaches with simple representation win over complex hand-coded linguistic rules.
 - *“Whenever I fire a linguist our machine translation performance improves.” (Jelinek, 1988)*
- 2000's – Statistics Powered by Linguistic Insights
 - With more sophistication with the statistical models, richer linguistic representation starts finding a new value.

Why is NLP hard?

Reason (4) – representation v.s. computability



Why learn NLP?

- Because it's fun.
 - It's a field that is relatively young and growing rapidly
=> a lot of opportunities for being creative and making contributions.

Why learn NLP?

- Because you can make the world better.
 - Computer system that can help with your writing/ composition
 - beyond spell checker or grammar checker
 - Computer system that reads all the important blogs and news and provides you the summary
 - Product review analysis

Why learn NLP?



- Because your future employer will love it.



IBM Research



YAHOO!
LABS



Powerset
NATURAL LANGUAGE SEARCH

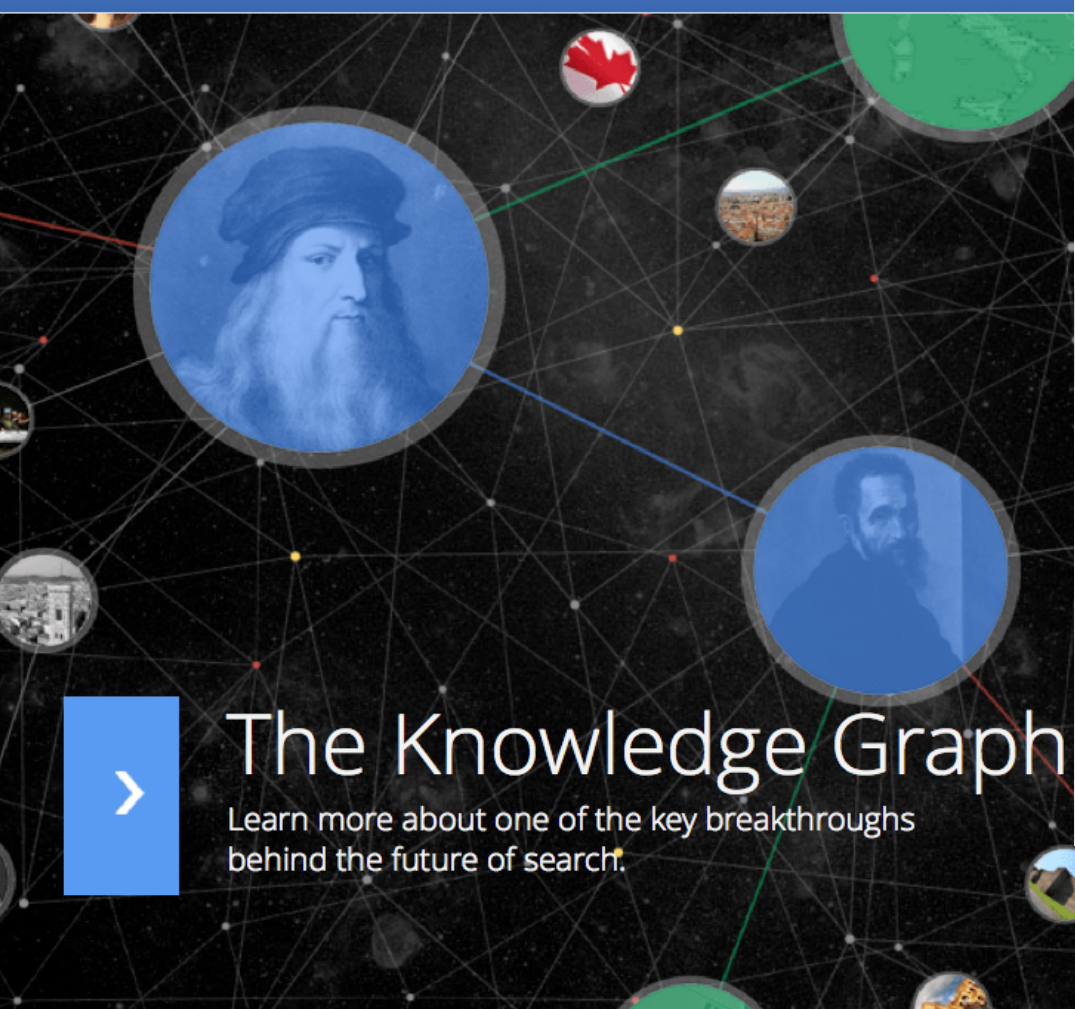


Microsoft

Knowledge & Information Extraction

Google Inside Search

Home How Search Works Tips & Tricks **Features** Search Stories Playground



The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

Leonardo da Vinci



leonardo.net

Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, ...

[Read more on en.wikipedia.org](#)

Born: April 15, 1453, [Anchiano](#)

Died: May 2, 1519, [Clos Lucé](#)

Buried: [St Florentin's Church](#)

Inventions: [Viola organista](#), [Double hull](#)

Parents: [Caterina da Vinci](#), [Piero da Vinci](#)

Explore your search



[Mona Lisa](#)
1507



[The Last Supper](#)
1498



[Virgin of the Rocks](#)
1508



[Lady with an Ermine](#)
1490



[The Battle of Anghiari](#)
1505

People also search for



[Michela...](#)



[Raphael](#)



[Vincent van Gogh](#)



[Pablo Picasso](#)



[Rembra...](#)

[Report a problem](#)

Question Answering



Twitter Sentiment US Election Candidates Oct 22 5-9 pm

Overall Twitter Sentiment by Candidate:



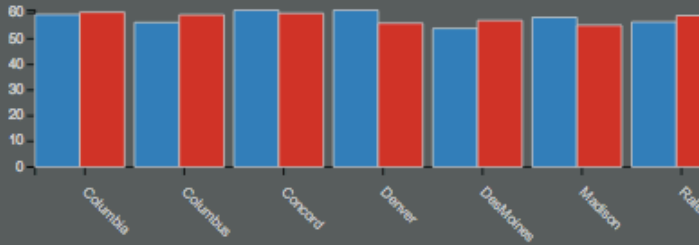
Obama
57.92



Romney
58.10

Overall Sentiment by City

Sentiment by City:



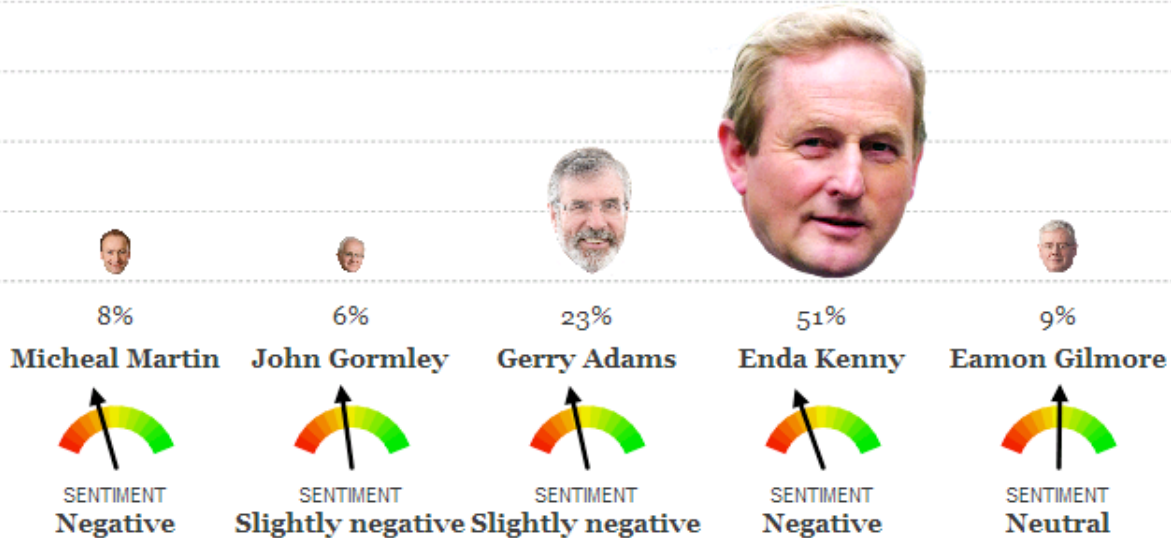
Candidates	Columbia	Columbus	Concord	Denver	DesMoines	Madison	Raleigh
Obama	59.50	56.40	61.20	61.20	54.10	58.40	57.20
Romney	60.40	59.40	60.00	56.10	57.20	55.30	58.10

Sentiment Analysis

Most Talked About Leader on Twitter


Last Few Hours Last Day Whole Campaign

Total Tweets: 22,365



Machine Translation



From: English ▼  To: French ▼

Type text or a website address or [translate a document](#).

Do more with Google Translate

Translate

Zamalek

Nord-Cin

Onet.pl -

Focus Or



Speak now

Language Processing & HCI

What can I help you with?

“ Siri do you still have the greatest enthusiasm and confidence in the mission ”

A fine question, Benjie. Now can we get back to work?

What can I help you with?

“ What is the answer to life the universe and everything ”

OK. I found this:

Input interpretation

Answer to the Ultimate Question of Life, the Universe, and Everything

Result

42

(according to Douglas Adams' humorous science-fiction novel *The Hitchhiker's Guide to the Galaxy*)

WolframAlpha

What can I help you with?

“ What do you think of android ”

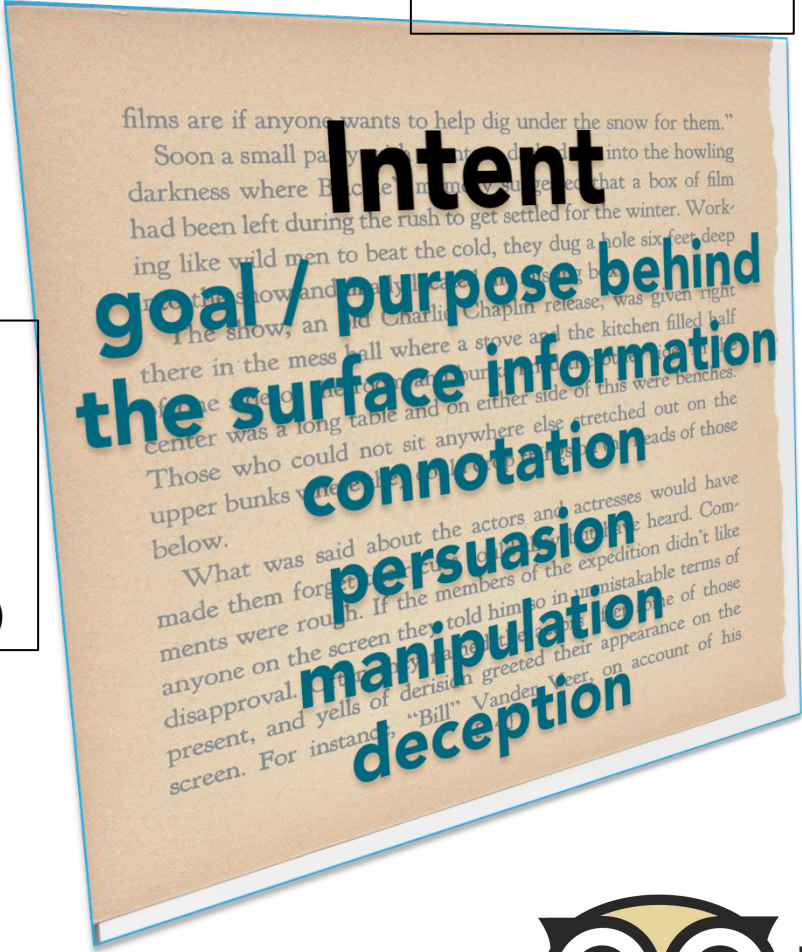
It's your opinion that counts, Benjie.



dodging
(Nguyen et al 2013)

hedging
(Choi et al. 2012)
(Ganter and Strube, 2009)
(Kilicoglu and Bergler 2008)

framing in media
& political discourse
(Yano et al., 2010)
(Recasens et al., 2013)



deception

fake online reviews



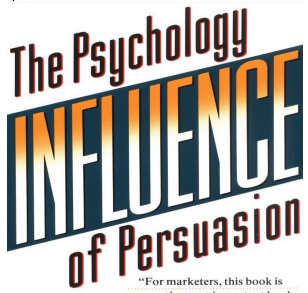
tripadvisor

syntactic packaging

"My toy broke"
instead of

"I broke my toy"

(Greene and Resnik 2009)



"For marketers, this book is among the most important books written in the last ten years."
— Journal of Marketing Research

ROBERT B. CIALDINI, PH.D.

Authorship Attribution

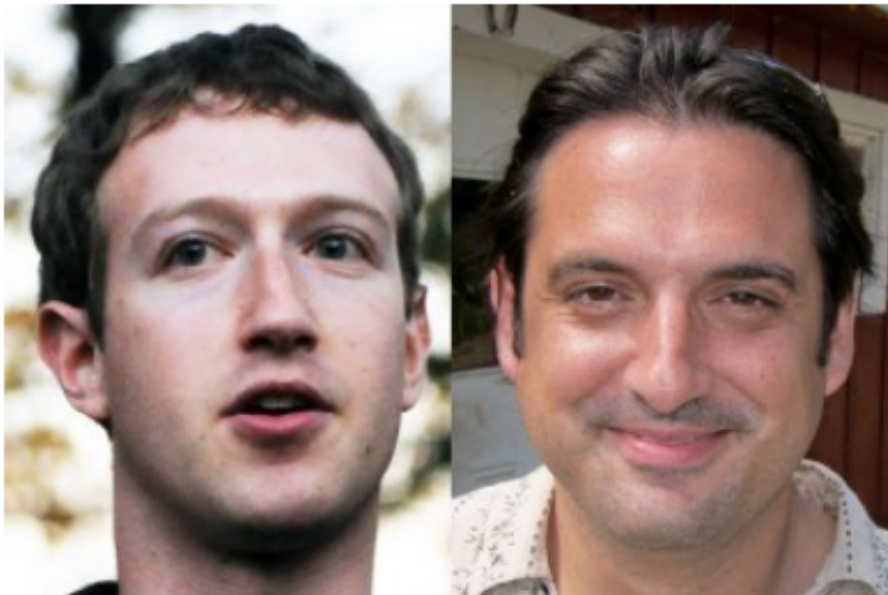


[TIME](#) | [Magazine](#) | [Video](#) | [LIFE.com](#) | [Lists](#)

[NEWSFEED](#) | [U.S.](#) | [POLITICS](#) | [WORLD](#) | [BUSINESS](#) | **[TECH](#)** | [HEALTH](#) | [SCIENCE](#) | [ENI](#)

[Home](#) | [Gadgets](#) | [Video Games](#) | [Apps & Web](#) | [News](#) | [Reviews & Features](#) | [Vi](#)

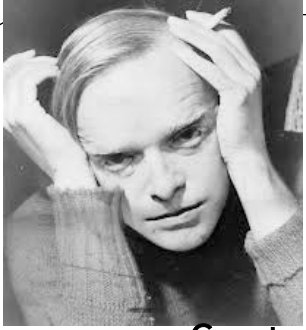
paul ceglia



POLITICS & LAW

Facebook Asks Court to Throw Out 'Fraudulent' Paul Ceglia Lawsuit

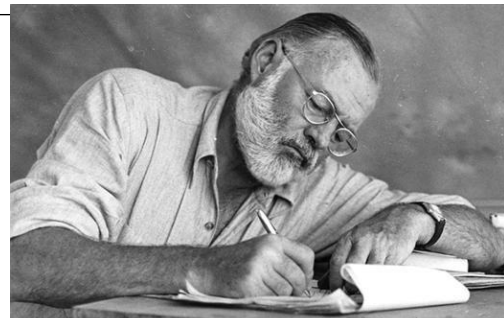
By Sam Gustin



Capote



Hempel



Hemingway



Woolf

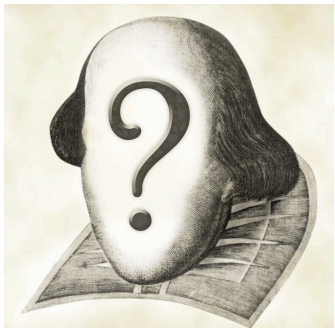
authorship verification

authorship obfuscation

demographics: gender, nationality, age, vocation

personality, psychological state: happy, authoritative, depressed...

intellectual traits & development: literary success



films are if anyone wants to help dig under the snow for them."

Soon a small party with a lantern dashed out into the howling darkness where Blackie's memory suggested that a box of film had been left during the rush to get settled for the winter. Working like wild men to beat the cold, they dug a hole six feet deep into the snow and finally they had found the film.

The show, an old Charlie Chaplin release, was given right there in the mess hall where a stove and the kitchen filled half of one side of the room and the benches. In the center was a long table and on either side of this were benches. Those who could not sit anywhere else stretched out on the upper benches and held their feet up to the heat of those below.

What they said to each other and how they would have made their feet forget their toes could they but have heard. Comments were rough if the members of the expedition didn't like anyone of the songs. From the time the film started a chorus of disapproval. Often they named the actors after some of those present, and yells of derision greeted their appearance on the screen. For instance, "Bill" Vander Veer, on account of his [14]

Identity
social identity
group identity
personal traits
intellectual traits

What to learn?

- Fundamental concepts and techniques
 - Language Models, Sequence Tagging, Trees
- Skills to embark on new research projects
 - How to read research papers
 - Useful even if you do not consider pursuing a ph.d. degree
 - How to approach unclear problems and make progress
 - Extremely useful both in industry and in research
 - ***“A problem clearly stated is a problem half solved.”***
-- Dorthea Brande

Prerequisites

- Basic probability
 - Basic statistics
 - Basic linear algebra
 - Machine learning helps *a lot*
 - Algorithms
 - Artificial Intelligence
-
- You must know how to code in some programming language

NLP Conferences to look out

- <ACL> Association for Computational Linguistics
 - <NAACL> North American chapter of the ACL
 - <EACL> European Chapter of the ACL
- <EMNLP> Empirical Methods in Natural Language Processing
- <CoNLL> Conference on Computational Natural Language Learning
- <COLING> International Conference on Computational Linguistics

➔ **ACL Anthology** <http://aclweb.org/anthology-new/>

- A Digital Archive of Research Papers in Computational Linguistics

Grading

- Paper Discussion 20%
- Homework 30%
- Final Project 40%
- Class Participation 10%.

Grading – Paper Discussion (20%)

1. Active Participation

- One must attend 3 sessions out of 3 or 4 Sessions in total.
- You may attend more than 3 sessions, then your score will be based on your best 3 sessions.
- Everyone must actively participate in each session. You can either make a comment, or ask a question, or answer a question.
- It is recommended that you will bring a printed copy of the paper to the class.

Grading – Paper Discussion (20%)

2. Written Critique

- Submit the written critique at the beginning of each session as a print-out. You cannot submit your critique after the session.
- length: about 1 page
- suggested content:
 - summary of the paper in your own words
 - the summary portion should be less than 25% of your critique
 - do not copy sentences in the paper as is. (zero grade for copy and paste)
 - your own thought, criticism, suggestions, new research ideas
 - do not make empty statements: e.g., very interesting, I learned a lot
 - do not make trivial, shallow comments

Homework (30%)

- 2 programming oriented assignments

Grading

- Paper Discussion 20%
 - Homework 30%
 - Final Project 40%
 - Class Participation 10%.
-
- Second chance – If your final project is substantial enough to turn into a paper submission to a reputable conference, then your grade may be considered for an A.

Grading – Final Project (40%)

- Project proposal submission (10%)
 - Due: TBD
- Project update presentation & submission (10%)
 - 10 minute presentation
 - Due: TBD
- Final project presentation & submission (20%)
 - 15 minute presentation
 - Due: TBD

Grading – Final Project

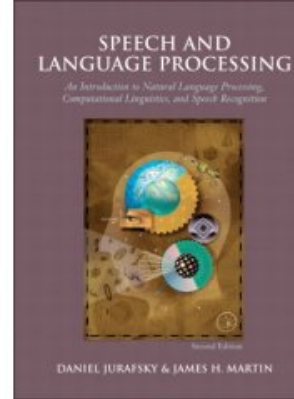
- Work in groups (2~4 students)
- You may work alone only if
 - you are a phd student
 - or you can achieve at least 70% of what typical 2-4 people groups can achieve.
- Most successful projects came out of students who worked in groups.
- Find partners no later than the end of Sep. If you need help in finding partners, send me an email.

Late Submission

- Each student may adjust his/her homework deadline upto 7 days throughout the semester without a penalty. (not 7 days for each assignment, but 7 days cumulatively for the entire semester).
- Fractional values will be rounded up - for instance, late submission by 1 hour is counted as late by 1 day.
- After then, 10% of score will be subtracted each day.
- This rule does not apply to the critique submission.
- This policy is to encourage students to submit quality work, rather than poorly composed work in a hurry.
- If there are situations where the application of this rule can be ambiguous, I have the right to apply the rule as I see appropriate.

Textbook

Jurafsky and Martin,
[SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition](#) , Second Edition, McGraw Hill, 2008.



- **[Library]** Some copies of the textbook are available in the reserve shelf of the North Reading Room (NRR) at the library.
- **[eBook]** Download [NookStudy eBook](#) platform and purchase the eBook license of the textbrook at \$62.10. Note that the license is valid only for one machine, and good for 180 days.
- **[Hard Copy]** The list price is \$138.00, Amazon.com price is \$104.27. Used books start at \$88.99 as of Jan/22/2011. Apparently, you can [sell your used book to Amazon.com](#) at the end of the semester.
- **[Renting]** [Stony Brook University Book Store \(www.whywaitforbooks.com\)](#) offers *Rental Program* at \$67.60.
- Go to [Stony Brook University Book Store \(www.whywaitforbooks.com\)](#) to explore above options

Reference Material

- Required text book:
 - Jurafsky and Martin, *Speech and Language Processing*, Prentice-Hall, **2nd edition**.
- Other useful text book:
 - Manning and Schutze. *Foundations of Statistical NLP*, MIT Press, 1999.

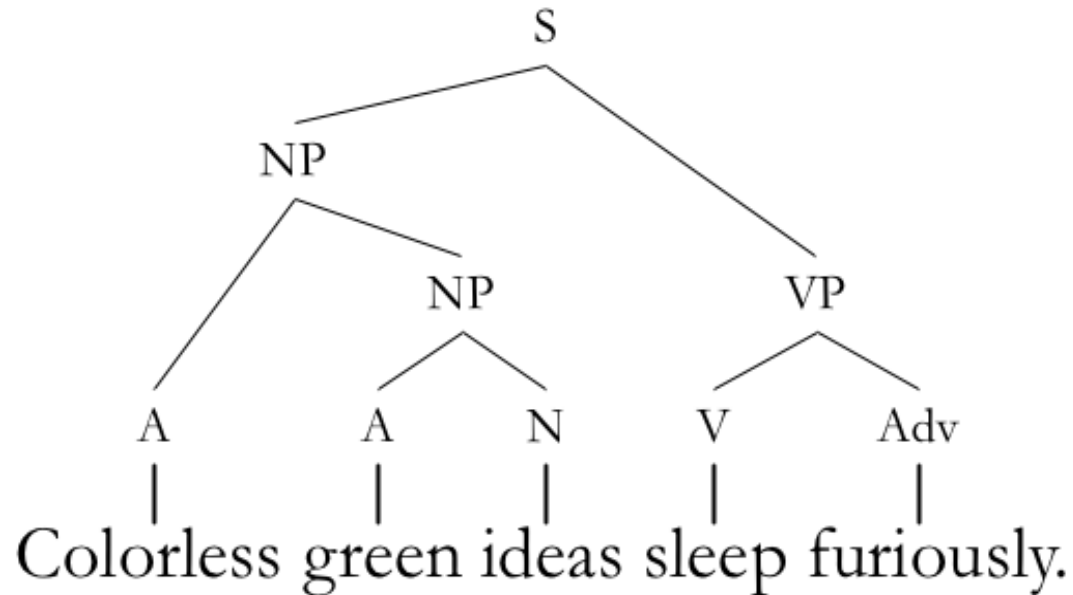
NLP 101: Syntax, Semantics, Pragmatics

- **Syntax** – grammatical ordering of words
- **Semantics** – meaning of words, phrases, sentences
- **Pragmatics** – meaning of words, phrases, sentences based on situational and social context

Syntax V.S. Semantics

(example by Noam Chomsky 1957)

- *Colorless green ideas sleep furiously.*
- *Furiously sleep ideas green colorless*



Semantics v.s. Pragmatics

What does "You have a green light" mean?

- You are holding a green light bulb?
- You have a green light to cross the street?
- You can go ahead with your plan?

Please fill out answers for below questionnaire

1. Your name & email
2. Masters? Phd?
3. Planning to take the class? Or audit? Or haven't decided?
4. Have you taken either "artificial intelligence" or "machine learning" or other NLP or linguistics classes?
5. What is your area of interest outside NLP? e.g. systems, theory, etc
6. Why are you taking this class?