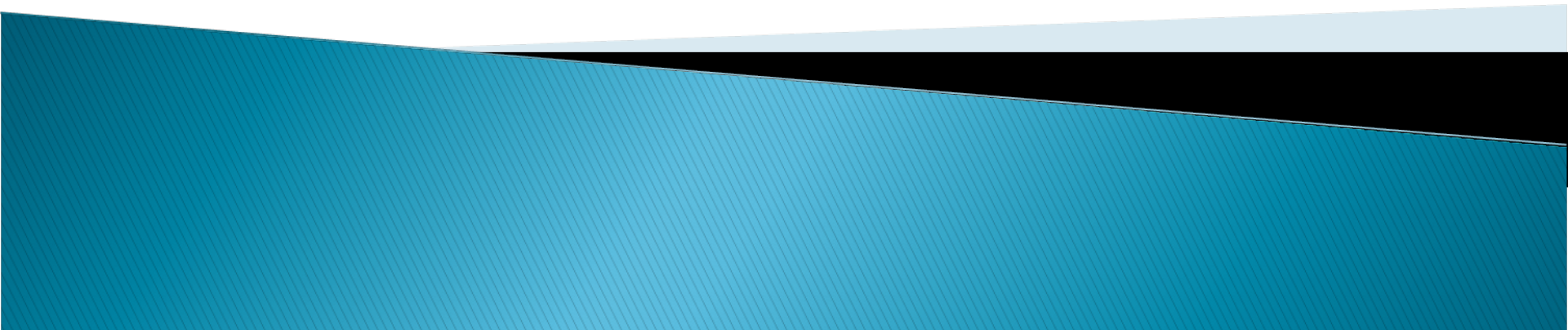


CS4705

Natural Language Processing

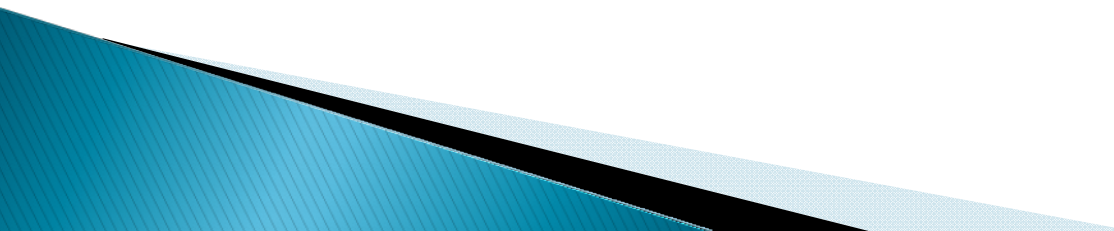
Fall 2009



What will we study in this course?

- ▶ How can machines **recognize** and **generate** text and speech?
 - Human language phenomena
 - Theories, often drawn from linguistics, psychology
 - Algorithms
 - Applications

Newspaper Titles

- "Something Went Wrong In Jet Crash, Expert Says"
 - "Police Begin Campaign To Run Down Jaywalkers"
 - "Drunk Gets Nine Months In Violin Case"
 - "Farmer Bill Dies In House"
 - "Iraqi Head Seeks Arms"
 - "Enraged Cow Injures Farmer With Ax"
 - "Stud Tires Out"
 - "Eye Drops Off Shelf"
 - "Teacher Strikes Idle Kids"
 - "Squad Helps Dog Bite Victim"
- 

Knowledge Needed

- ▶ Morphology: word formation
- ▶ Syntax: word order
- ▶ Semantics: word meaning and word composition
- ▶ Pragmatics: influence of context/situation

Goal: Discover what the speaker meant



Morphology

- ▶ “Stud tires out”
 - “Tires”: a noun or a verb?
- ▶ Internet search: *union activities in New York*
 - Union/unions; activities/activity
 - Active? Action? Actor?
 - New vs. New York

Syntax

▶ Word Order

- John hit Bill
- Bill was hit by John
- Bill hit John
- Bill, John hit
- Who John hit was Bill

▶ Constituent Structure

- "Teacher Strikes Idle Kids"
- "Enraged Cow Injures Farmer With Ax"

Syntax

▶ Word Order

- John hit Bill
- Bill was hit by John
- Bill, John hit
- Who John hit was Bill

▶ Constituent Structure

- “[Teacher Strikes] [Idle] [Kids]“
- “Enraged Cow Injures Farmer With Ax”

Syntax

▶ Word Order

- John hit Bill
- Bill was hit by John
- Bill, John hit
- Who John hit was Bill

▶ Constituent Structure

- “[Teacher] [Strikes] [Idle Kids]“
- “Enraged Cow Injures Farmer With Ax”

Syntax

▶ Word Order

- John hit Bill
- Bill was hit by John
- Bill, John hit
- Who John hit was Bill

▶ Constituent Structure

- "Teacher Strikes Idle Kids"
- "[Enraged Cow] [Injures] [Farmer With Ax]"

Syntax

▶ Word Order

- John hit Bill
- Bill was hit by John
- Bill, John hit
- Who John hit was Bill

▶ Constituent Structure

- "Teacher Strikes Idle Kids"
- "[Enraged Cow] [Injures] [Farmer] [With Ax]"



Semantics

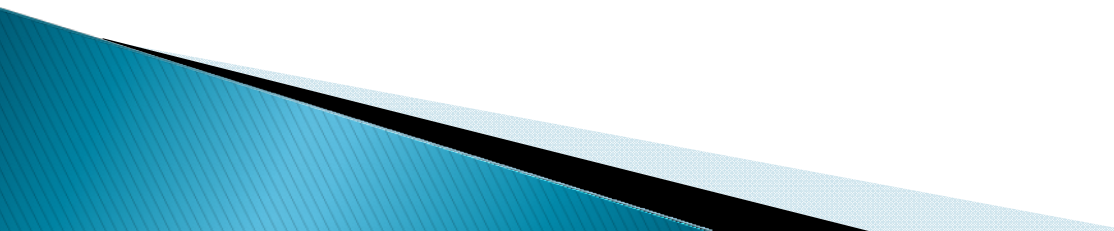
- ▶ **Word meaning**
 - John picked up a bad cold.
 - John picked up a large rock.
 - John picked up Radio Netherlands on his radio.
- ▶ **Composition of meaning**
 - Squad helps dog bite victim
 - Enraged cow injures farmer with ax

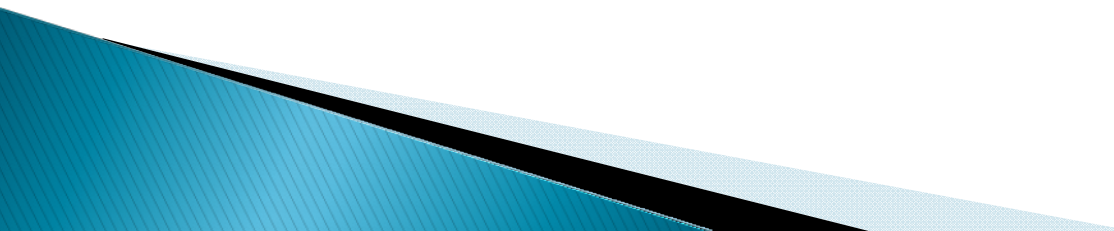
Pragmatics – The influence of context

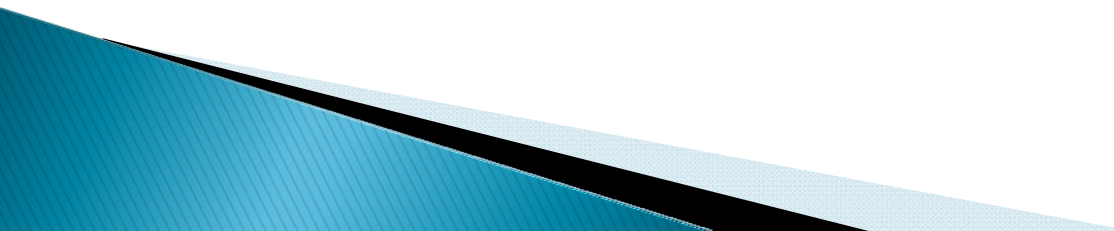
“Going Home” – A play in one act

- ▶ Scene 1: Pennsylvania Station, NY
- ▶ Bonnie: Long Beach?
- ▶ Passerby: Downstairs, LIRR Station.

- ▶ Scene 2: Ticket Counter, LIRR Station
 - ▶ Bonnie: Long Beach?
 - ▶ Clerk: \$4.50.
- 

- ▶ Scene 3: Information Booth, LIRR Station
 - ▶ Bonnie: Long Beach?
 - ▶ Clerk: 4:19, Track 17.
- 

- ▶ Scene 4: On the train, vicinity of Forest Hills
 - ▶ Bonnie: Long Beach?
 - ▶ Conductor: Change at Jamaica.
- 

- ▶ Scene 5: On the next train, vicinity of Lynbrook
 - ▶ Bonnie: Long Beach?
 - ▶ Conductor: Right after Island Park.
- 

Algorithms

- ▶ Rule-based/Symbolic
 - Parsers
 - Finite state automata
- ▶ Probabilistic
 - Learned from observation
 - Predicting best guess
 - Statistical

Current Real World Applications

Searching very large text and speech corpora:
e.g. the Web

Question answering over the web

Translating between one language and
another: e.g. Arabic and English

Summarizing very large amounts of text: e.g.
your email, the **news**, **reviews** (**mobile version**)

Sentiment analysis: **NYT article**

Generating texts

Dialogue systems: e.g. Amtrak's '**Julie**'

Instructor

- ▶ Kathy McKeown
- ▶ Office: 722 CEPSR
- ▶ Head NLP Group
- ▶ 25 years at Columbia, Department Chair for 6
- ▶ Research
 - Summarization
 - Question Answering
 - Language Generation
 - Multimedia Explanation

Logistics

- ▶ Instructor: Kathy McKeown
 - (kathy@cs.columbia.edu)
 - Office and hours: CEPSR 722, Tues 4–5, Wed 4–5
- ▶ Teaching Assistants:
 - Sara Rosenthal
 - ss3067@columbia.edu
 - <http://www.cs.columbia.edu/~ss3067>
 - Office and hours: 726 CEPSR, M 4:30–5:30, Thurs 1:30–2:30
 - Kaushal Lahankar
 - kn12102@columbia.edu
 - Office and hours: NLP Lab, 7LW, Thurs 4–6
- ▶ Syllabus available at
<http://www.cs.columbia.edu/~kathy/NLP>

- ▶ Text: Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, 2nd edition, Prentice–Hall, 2000 (available at CU Bookstore)
- ▶ Assignments:
 - 4 homework assignments
 - Midterm and final exams
 - **Four** ‘free’ **late days** for homework assignments
 - After that, 10% off per day late
 - You must get a CS account
- ▶ Evaluation: 50% homework + 40% exams + 10% class participation

Academic Integrity

Copying or paraphrasing someone's work (code included), or permitting your own work to be copied or paraphrased, even if only in part, is forbidden, and will result in an automatic grade of 0 for the entire assignment or exam in which the copying or paraphrasing was done. Your grade should reflect your own work. If you are going to have trouble completing an assignment, talk to the instructor or TA in advance of the due date please. Everyone: Read/write protect your homework files at all times.

For Next Class

- ▶ Look at syllabus
 - ▶ Read Chapters 1–2 of J&M
 - ▶ Questions?
- 