CS 4705
Natural Language Processing
Fall 2019

Professor Kathy McKeown

Today

- What is NLP?

- Class Logistics
  - What will we cover
  - Helpful background
  - Class homework and exams

Class Size and In-class Discussion

- PollEverywhere to interact in class
  - Verbal interaction also welcome!
- Regular intervals: pause to answer questions
- I will also ask questions of you
  - Answers will be show up on the slide
  - May use an interesting answer to further discussion
  - **In-class participation will count towards your grade (10%)**
  - **Check that your answers are being recorded. Send a message on Piazza under PollEverywhere tag after a class where you think you participated.**

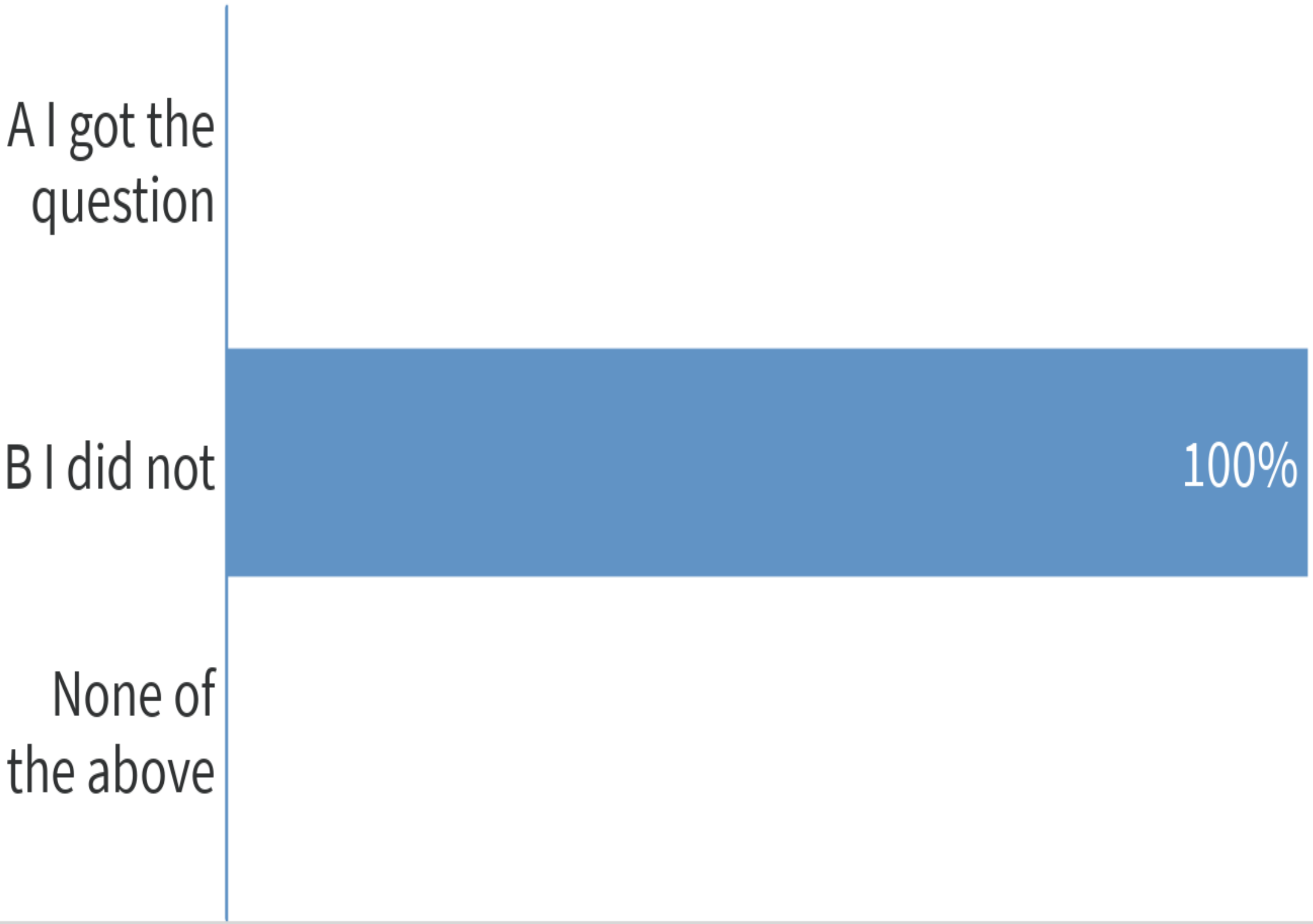
Participation: speaking up in class

- For credit from participation by speaking in class, we will post **a link to a Google Form on Piazza after every class**. There will be a new link for each class.
- In order to get a participation mark for the class based on verbal answers, **you must fill out this form**, which asks what question you answered and one thing you learned from class.
- The answers should be no more than one sentence long!
- Please fill it out in a timely manner -- submissions that occur 48 hours after class will not be counted. In fact, the Google Form may not be accessible after this point.
- We will count verbal participation in addition to polleverywhere responses. You can participate either way. No need to participate both ways!

Download the PollEverywhere App on your Phone

- Enter my user name: kathleenmckeown
- Log in to your Columbia account
- When you respond to a poll, you should get a thank you message from me.
- Check on courseworks that we are getting answers from you.

Test



Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Class Policy on Electronics

- Cell phone in class OK and needed for PollEverywhere interaction

- Keep laptops closed or don’t bring to class

What will we study in this class

- How can machines *understand* and *generate* language?
  - Examples drawn from naturally occurring corpora
  - Theories about language
  - Algorithms
  - Statistical and neural methods
  - Applications

Knowledge Needed

- Morphology: word formation
- Syntax: word order
- Semantics: word meaning and composition
- Pragmatics: Influence of context and situation

Goal: discover what the speaker meant (or communicate what the system intends)

Morphology

- Important for search, machine translation, summarization

- *Union Activities in New York*
  - Singular/plural
    - Union/unions
    - Activity/activities
  - Other languages are morphologically rich
    - Arabic: definite embedded in the word (clitics): The union (Al+) vs. a union, unions
    - German: case part of the word (subj vs obj)
  - *Are there examples in your language?*

News article titles

- Stud tires out
- Eye drops off shelf
- Teacher strikes idle kids
- Drunk gets nine months in violin case
- Enraged cow injures farmer with ax
- Ban on nude dancing on Governor’s desk
- Hospitals are sued by seven foot doctors
- Red tape holds up new bridges
- Government head seeks arms
- Patient at death’s door – doctors pull him through
- In America a woman has a baby every 15 minutes

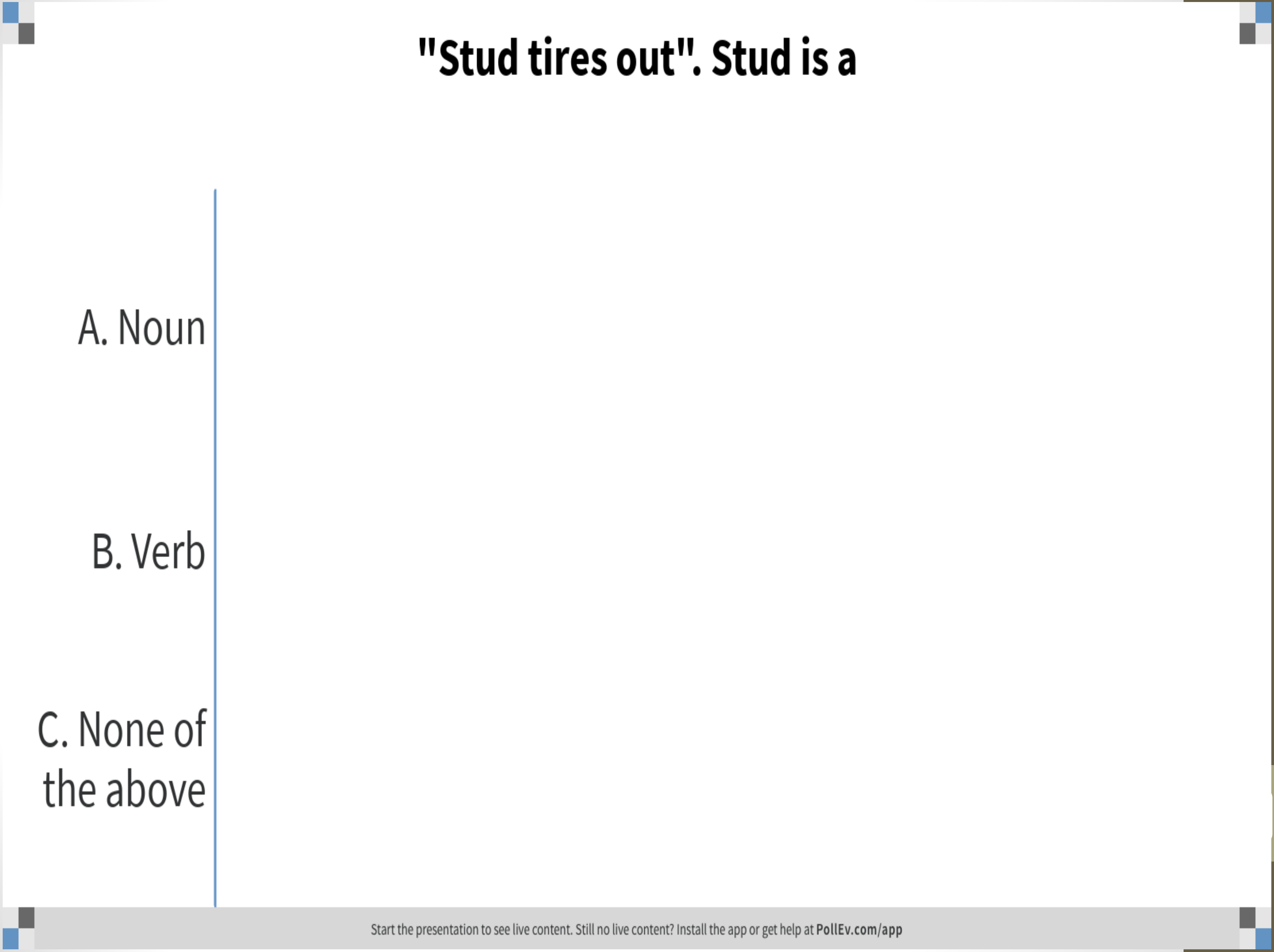
Syntax

- Part of speech tagging: is a word a noun, verb, adverb, adjective, etc?

- Parsing
  - Identifying constituents
    - NP: *Kathy McKeown, a man in the park*
    - VP: *was looking up, had risen*
  - Identifying subjects and objects
    - *Bill hit John* vs *John hit Bill*
  - Modification
    - *John saw the man in the park with a telescope*

Part of Speech tagging

- *Stud tires out*
  - *Tires: a noun or a verb?*
- *Eye drops off shelf*
  - *Drops: a noun or a verb?*
- *Teacher strikes idle kids*
  - *Strikes: a noun or a verb?*



"Stud tires out". Stud is a

A. Noun

B. Verb

C. None of the above

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

"Eye drops off shelf". Drop is a

A. Noun

B. Verb

C. None of the above

"Teacher strikes idle kids" strikes is a

A. Noun

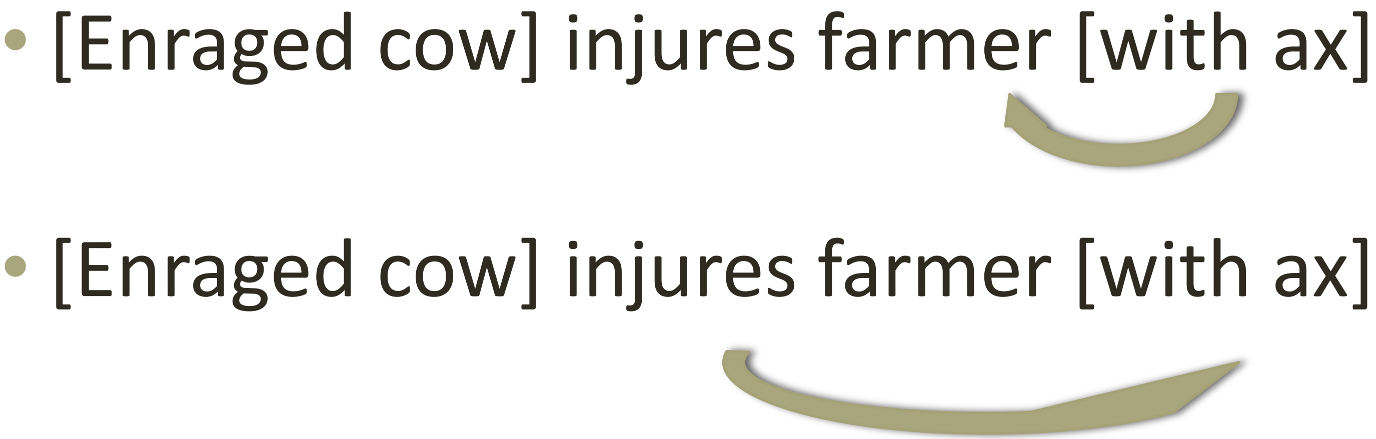
B. Verb

C. None of the above

Constituent Structure and Modification

- The problem of PP attachment

  Enraged cow injures farmer with ax



- [Enraged cow] injures farmer [with ax]

- [Enraged cow] injures farmer [with ax]

Representing modification with brackets

- [Enraged cow] [injures [farmer [with ax]]]
- [Enraged cow] [injures [farmer] [with ax]]]

Constituent Structure and Modification

- The problem of PP attachment

  Ban on nude dancing on governor’s desk

  NP NP PP

- [Ban] on [nude dancing] [on governor’s desk]

- There are two possible modifications? What are they?
- Which one is correct?

Using bracketing show two possible constituent structures for "[Ban] on [nude dancing] [on governor's desk]"

Constituent Structure and Modification

- The problem of PP attachment

  Ban on nude dancing on governor’s desk

  NP **PP**

- [[Ban] on [nude dancing]] [on governor’s desk]

  NP **PP**

- [Ban] on [[nude dancing] [on governor’s desk]]

NP

Noun noun modification

- *Water fountain:* a fountain that *supplies* water
- *Water ballet:* a ballet that *takes place* in water
- *Water meter:* a device (called a meter) that *measures* water
- *Water barometer:* a barometer that *uses* water (instead of mercury) to measure air pressure
- *Water glass:* a glass that is *meant to hold* water

Which constituent structure best represents the meaning of country song platinum album?

A [country [song [platinum album]]]

B [country [[song platinum] album]]

C [[country song] [platinum album]]

D [[country [song platinum]] album]

[[[country song] platinum] album]

None of the above

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Noun noun modification and headlines

- *Hospitals are sued by seven foot doctors*

- *Hospitals are sued by [[seven foot] doctors]*

- *Hospitals are sued by [seven [foot doctors]]*

Word Meaning

- *Red tape **holds up** new bridges*
  - *Holds up*:
    1. [TRANSITIVE] to support someone or something so that they do not fall down
       - *Her legs were almost too shaky to hold her up*.
    2. [TRANSITIVE] [OFTEN PASSIVE] to cause a delay for someone or something, or to make them late
       - *Sorry I'm late, but my flight was held up.*
- *Government **head** seeks **arms***
  - *Head:*
  - *Arms:*

Pragmatics

- Discourse context
  - *John went to the store.* ***He*** *bought bread and butter.*
- Situational (real world) context
  - Day after the Texas highway shooting
  - ***His*** *irresponsible actions took the life of a young woman who was just beginning her adult life.*
- Commonsense knowledge
  - ***Boston*** *called and left a message for Joe*.

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What data is available for learning?

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What features yield good predictions?

Machine learning framework

- Data (often labeled)
- Extraction of “features” from text data
- Prediction of output

What representations and architectures yield good predictions?

Machine Learning Methods

- Supervised
  - Support vector machine, Naïve Bayes, Logistic regression
  - Sequence labeling: Hidden Markov Modeling (HMM), Conditional Random Fields (CRF)
  - Neural networks
- Unsupervised
  - Clustering
- Semi-supervised
  - Boot-strapping, self-training, co-training
  - Distant learning

Where does the data come from?

- Manually labeled

- Naturally occurring

- A noisy, but plentiful substitute

Core NLP

- Morphological analysis
- Part of speech tagging
- Parsing

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., Chinese and English
- Summarizing text: e.g., your email, the news, reviews
- Sentiment analysis
- Generating texts
- Dialog systems: Amtrak's Julie

Logistics

Homework 0

- Must pass with 40/50 to stay in the class
- Tests and refreshes background knowledge in related math
- Set up programming environment (google cloud)
- Due this Saturday, 10AM. (Both registered and waitlist participants)

Syllabus

- Available at: http://www.cs.columbia.edu/~kathy/NLP/2019

Instructor

- Kathy McKeown
  - Office: 722 CEPSR
  - NLP Group
  - 37 years at Columbia
  - Experience: Dept chair, SEAS Vice Chair for Research, Founding Director Data Science Institute
- Research
  - Summarization
  - Question Answering
  - Language Generation
  - Sentiment analysis
  - Social media analysis
  - Multilingual applications

TAs

- Elsbeth Turcan (head TA)
- Emily Allaway
- Katy Gero
- Alyssa Hwang
- Fei-Tzin Lee

Background

- Programming. We will use Python
- In addition, at least one:
  - Artificial Intelligence
  - Machine learning
  - Programming Languages and Translators
  - Statistics

Math background

- Calculus
- Linear algebra
- Probability and statistics

Textbooks

- Natural Language Processing, Jacob Eisenstein
- Speech and Language Processing, 3rd Edition, by Jurafsky and Martin.
- Neural Network Methods for Natural Language Processing by Yoav Goldberg.

Assignments

- 4 homework assignments: 3 programming, 1 written
  - We will be using Google Cloud
  - HW0:
    - You must pass HW0 with 40/50 to stay in the class
    - Due Sept 7th, 10AM.
    - Sets up your google cloud account properly
  - Four free late days
  - After that 10% off for each day late
- Midterm and final
- Evaluation: 50% homework + 40% exams + 10% class participation (via PollEverywhere)

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

- Read Chapter 1 of NLP, Chapter 1 of Speech and Language
- Questions?

Questions