# Introduction to Natural Language Processing

School of Information Studies
Syracuse University

# Outlines

- Course logistics
- Intro to NLP
- NLP applications
- Levels of language
- Intro to NLTK
- Lab

# Self-intro

- Glad to learn more about my students!

# Course Logistics

- **Instructor**: Mei Zhang
- **Course developing team**: Nancy McCracken, Lu Xiao, Jeff Stanton, Steven Wallace, Hernando Hoyos, Benjamin Nichols, Norma Palomino, Mei Zhang
- **Office**: Room 209, Hinds Hall
- **Office hours**: noon-1 pm on Wednesdays **office/online or** by appointment
- **Email**: mzhang@syr.edu (from Sunday to Friday)

# Course Logistics (2)

- Bring your own laptop to the class
- Slides/lab materials will be available on course website right **before** each class
- Reading for the next week will be available right **after** class.
- Share your favorite songs with us!

# Learning Objectives

- Demonstrate the levels of linguistic analysis, the computational techniques used to understand text at each level, and what the challenges are for those techniques.
- Process text through the language levels using the resources of the Natural Language Toolkit (NLTK) and some rudimentary use of the programming language Python.
- Understand basic terminology and concepts related to neural network computing models, deep learning approaches and platforms, and demonstrate processes for training new embedding models on unique/custom corpora
- Describe how NLP is used in many types of real-world applications.

# Assignments Overview

- Individual Assignments
  - Mini talk (10%)
  - Class participation (10%)--- available during 9:30-1pm on Mondays
    - Each student must post at least for 5 weeks.
  - Weekly lab exercise (10%)
  - 3 homework (30%)
    - corpus stats, CFG grammars and parsing, sentiment analysis
- Group Assignments– NLP Application Investigation
  - Proposal (5%)
  - Presentation (15%)
  - Final report (20%)

# Textbooks

- Jurafsky, D., & Martin, J. H. *Speech and language processing* (3rd ed. draft). Available from https://web.stanford.edu/~jurafsky/slp3/

- Bird, S., Klein, E., & Loper, E. *Natural language processing with Python*. Available from http://www.nltk.org/book

- Lane, H., Howard, C., & Hapke, H.. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. https://www.manning.com/books/natural-language-processing-in-action#toc

# Let's get started with questions!

Q1: in Blackboard "Discussion" – "week 1"

- What do you think is Natural Language Processing? *Or*

- What NLP can do for us?

# Natural Language Processing (NLP)

- Definition: "A range of **computational techniques**

for **analyzing and representing** naturally occurring texts at one or more levels of linguistic analysis

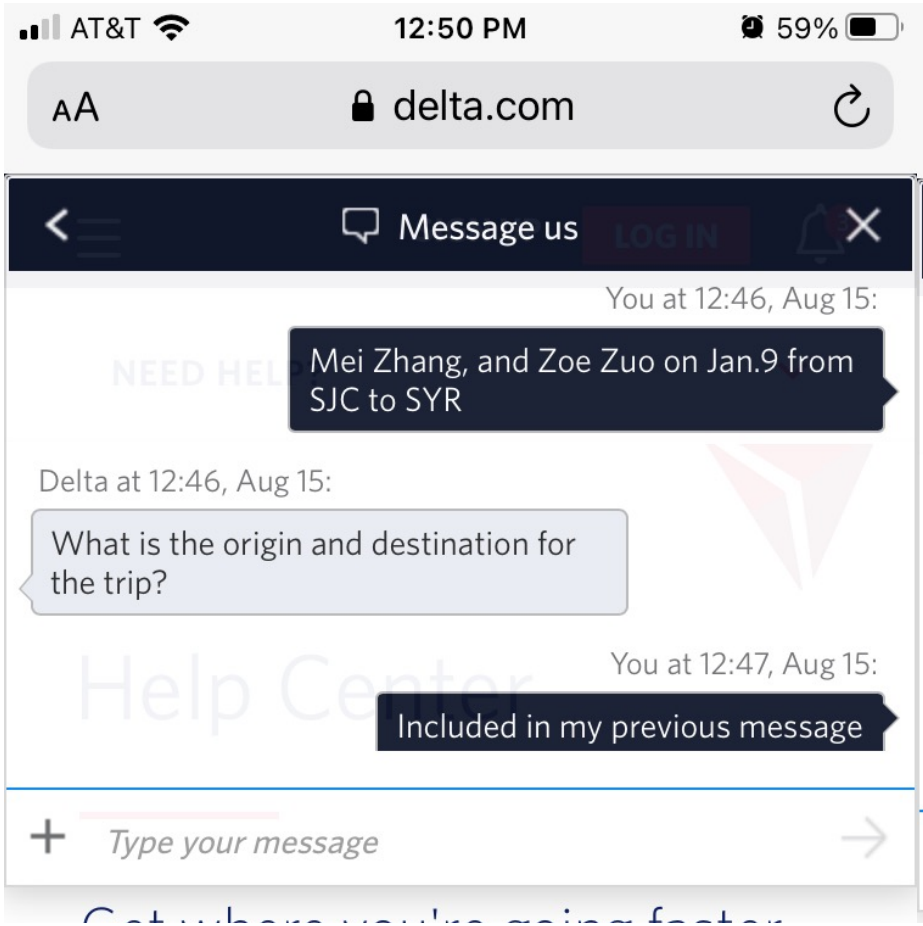
for the purpose of **achieving human-like language processing** for a range of particular tasks or applications."

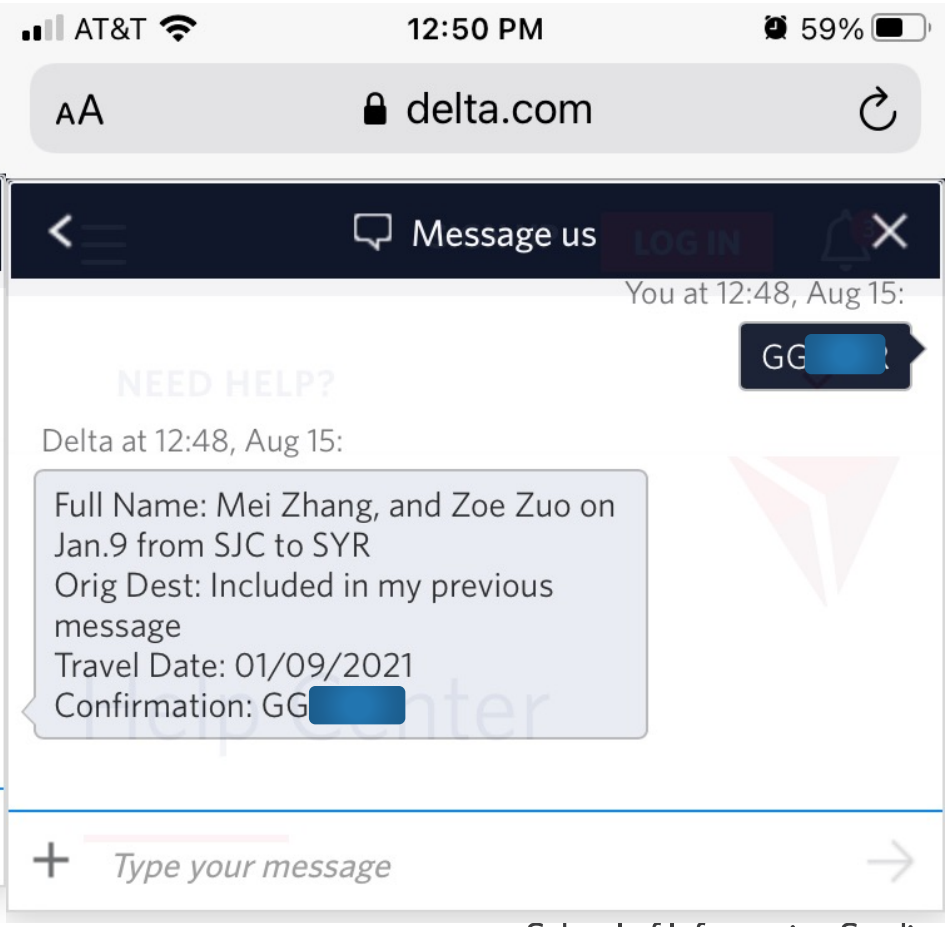
- Often used interchangeably with "computational Linguistics" – doing linguistics on computers

# Where is NLP now?

- Goals can be far-reaching
  - True text understanding
  - Reasoning about knowledge in text
  - Real-time participation in spoken dialogs
- Or very down-to-earth
  - Finding the price of products on the web
  - Context-sensitive spell-checking
  - Analyzing sentiment and opinions statistically
  - Extracting facts or relations from documents
- Currently, NLP is providing these practical applications (yet still dreaming of the AI goals: https://youtu.be/JvbHu_bVa_g

# An example

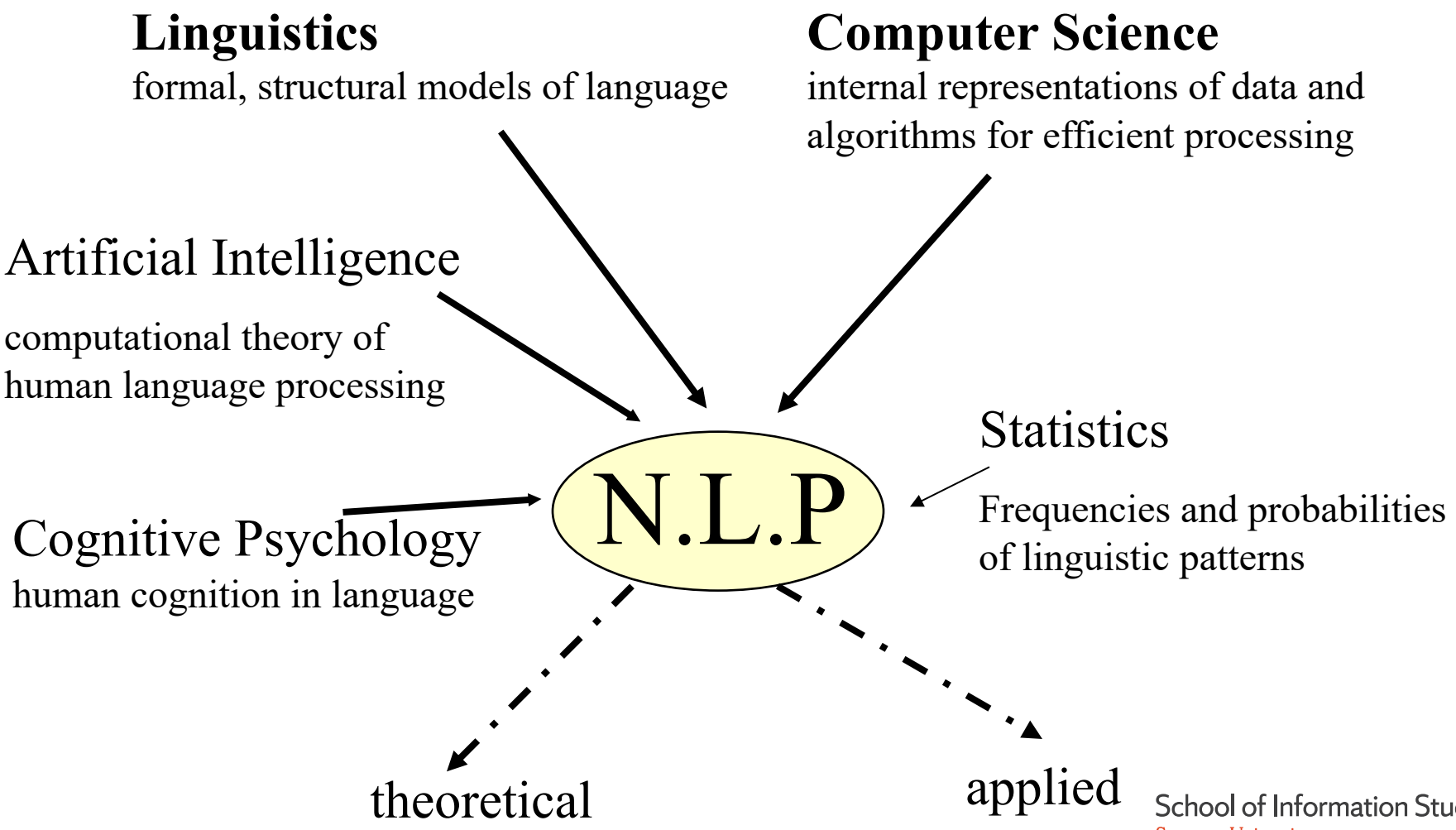

AT&T
12:50 PM
59%
AA
delta.com
Message us
You at 12:46, Aug 15:
Mei Zhang, and Zoe Zuo on Jan.9 from SJC to SYR
Delta at 12:46, Aug 15:
What is the origin and destination for the trip?
You at 12:47, Aug 15:
Included in my previous message
Type your message


AT&T
12:50 PM
59%
AA
delta.com
Message us
You at 12:48, Aug 15:
GG
Delta at 12:48, Aug 15:
Full Name: Mei Zhang, and Zoe Zuo on Jan.9 from SJC to SYR
Orig Dest: Included in my previous message
Travel Date: 01/09/2021
Confirmation: GG
Type your message

# Why is NLP so hard?

- **Seems pretty simple for humans**
  - Usually quite unaware of the complexity of the language tasks they perform so effortlessly
- Some reasons are
  - Ambiguity
  - Subtleties of meaning
    - Irony, sarcasm, humor, metaphor

# Fields contributing to NLP



# Two Sides of NLP: analysis and generation

1. paraphrase an input text
2. translate it to another language or representation
3. answer questions about it
4. draw inferences from it
5. phrase the results in natural language

Natural Language Processing

Language Analysis*

Language Generation

*Main emphasis in this course

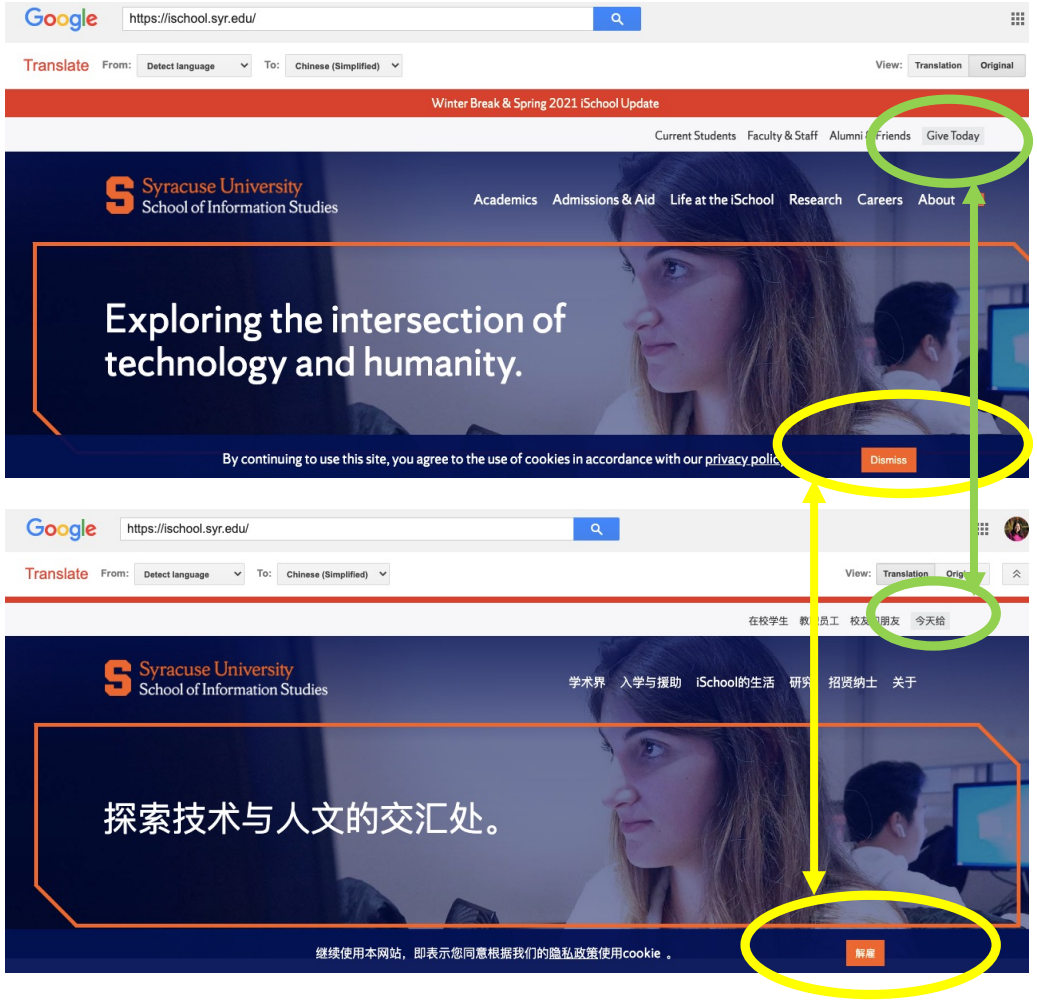
# Introduction to NLP applications

School of Information Studies
Syracuse University

# NLP Application Areas

Machine Translation – conversion of text from one language to another

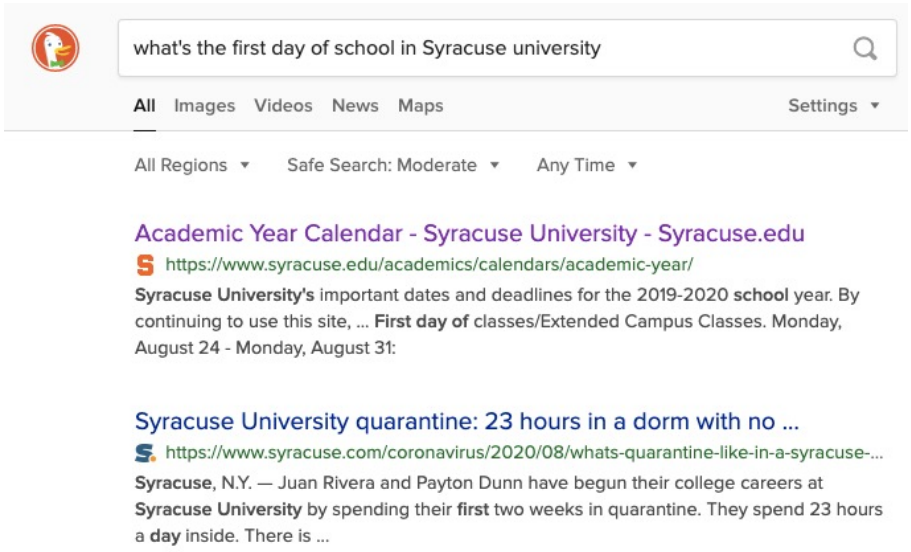
- Google, Mircrosoft, Amazon and IBM all have language translators
- MT techniques use context , not just word for word substitution

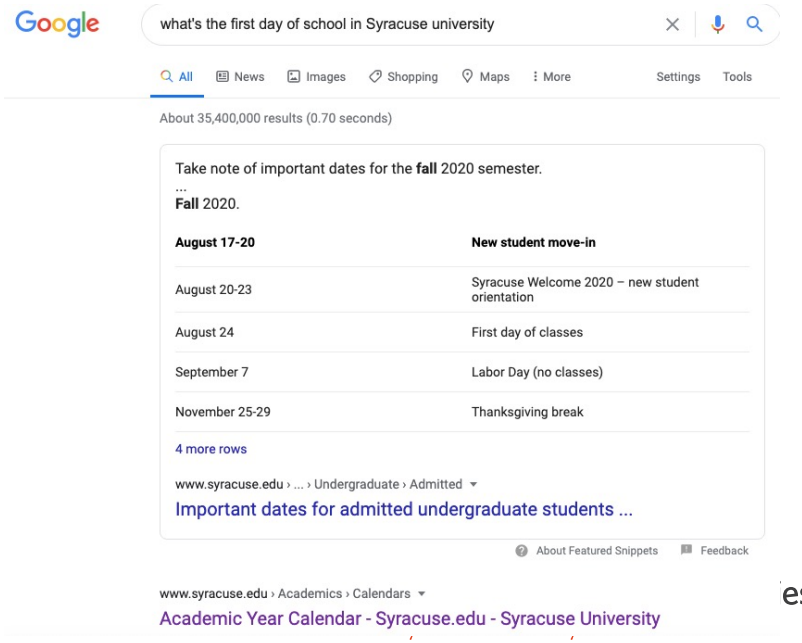


# NLP Application Areas

Information Retrieval / Search Engines – provision of documents containing requested information

- Google, many other search engines
- Use lowest levels of NLP to stem words, find phrases for indexing documents
- Users conform to keyword query restriction, but many search engines will now accept questions in natural language form



# NLP Application Areas

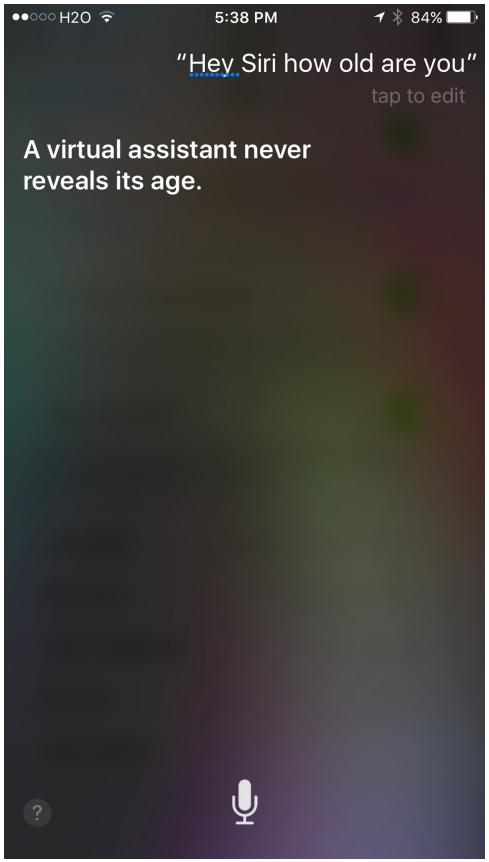
Information Extraction / Text-mining – populating a structured database with specific bits of information found in text

- Competitive Intelligence: analyzes news text and social media text for
  - Names of people, companies and other entities
  - Relations between them, e.g. corporate roles, or events such as mergers
  - Translation of professional analysis into common languages
- Sentiment analysis
- Identifying influencers on social media

# NLP Application Areas

Human-computer Interfaces –

- Voice/speech recognition
- Information assistants
- Chatbots
- Automatic phone agents
- Interactive querying of databases



# NLP Application Areas

Summarization – abstraction and condensation of text's major points

- Current systems select a set of significant sentences from the document as a summary
- Example summarizer: http://textsummarization.net/text-summarizer

# NLP Application Areas

Question & Answering Systems – focused information provision

- Find answers to questions in documents or other resources
- Must be able to handle many different phrasings of desired answer and to provide justification

Watson

- IBM's QA system trained to play Jeopardy
- Extensive development of NLP techniques

# Discussion Question

Q2 in Blackboard “Discussion” – “week 1”

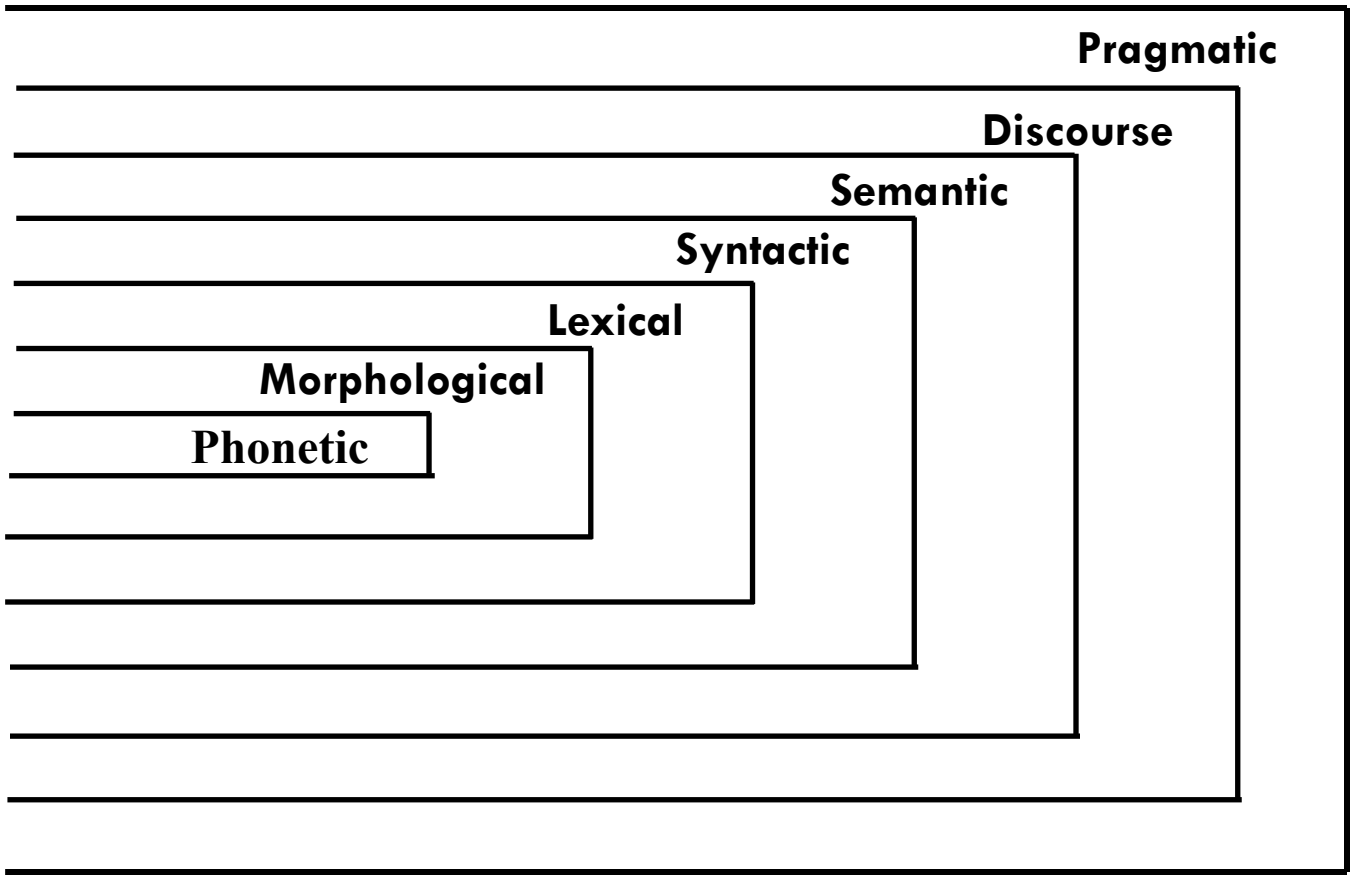
Share your experience with some form of NLP application. Are you satisfied with the interaction?

# Levels of Language

School of Information Studies
Syracuse University

# Levels of Language Analysis

Use the synchronic model to guide computational techniques to analyze text (as much as possible)



# Speech Processing

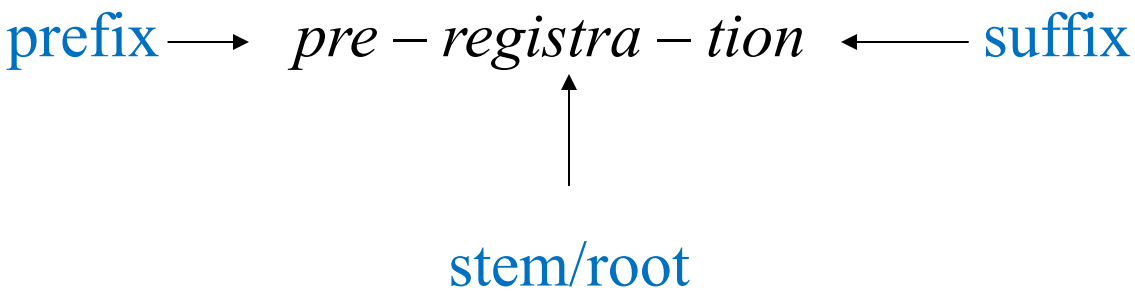
- Interpretation of speech sounds within & across words
- sound waves are analyzed and encoded into a digitized signal

Rules used in Phonological Analysis

1. Phonetic rules – sounds within words
   - e.g. When a vowel stands alone, the vowel is usually long
2. Phonemic rules – variations of pronunciation when words are spoken together
   - e.g. "r" in "part" vs. in "rose"
3. Prosodic rules – fluctuation in stress and intonation across a sentence: rhythm, volume, pitch, tempo, and stress
   - e.g. High pitch vs. low pitch

# Morphological Analysis

- Deals with the componential nature of lexical entities:




- What features do inflections reveal in English?

Verbs ⟶ tense & number

Nouns ⟶ single/plural

Adjectives ⟶ comparison features

# Lexical

- Adding lexical class information to words:

  Part-of-speech (POS) tagging tags words with specific noun, verb, adjective and adverb types

*03/14/1999 (AFP)*… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

… the|DT extremist|JJ Harkatul_Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama_bin_Laden|NP …

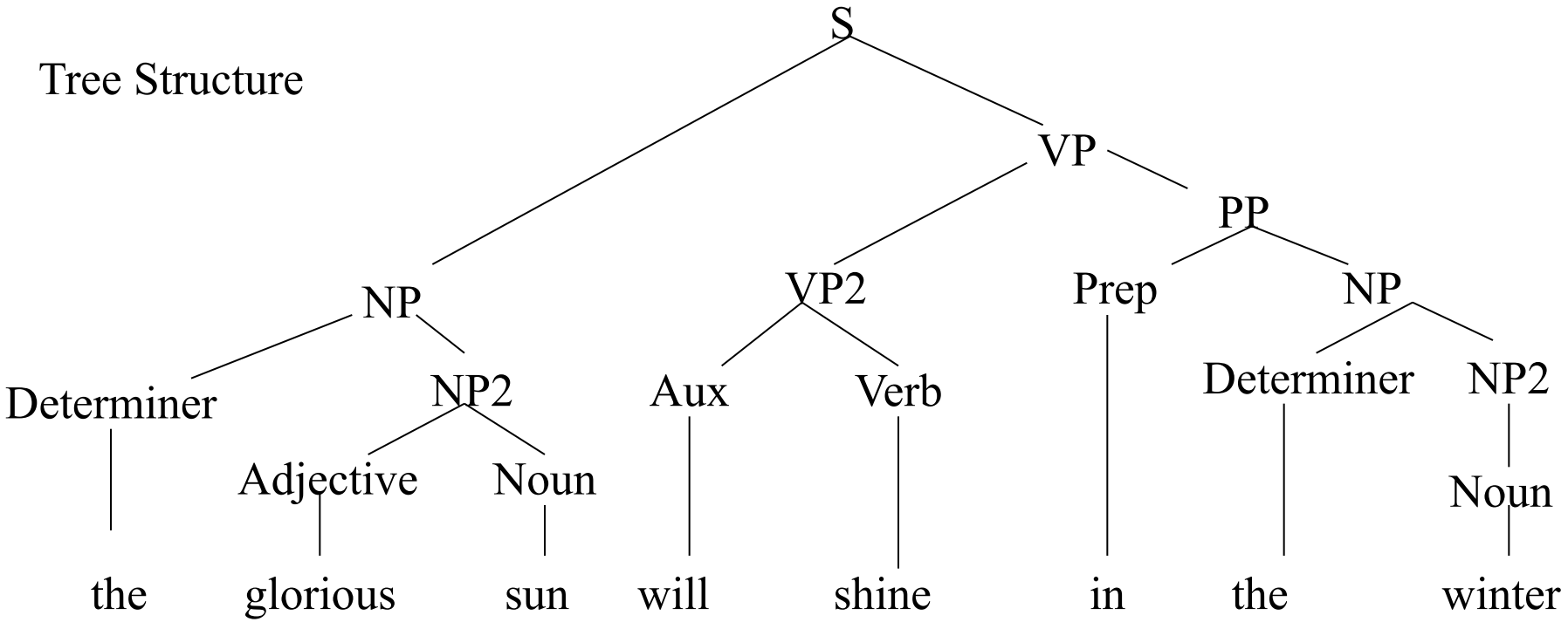
- These POS tags are taken from the Penn Treebank tag set.

# Lexical: Word Meanings

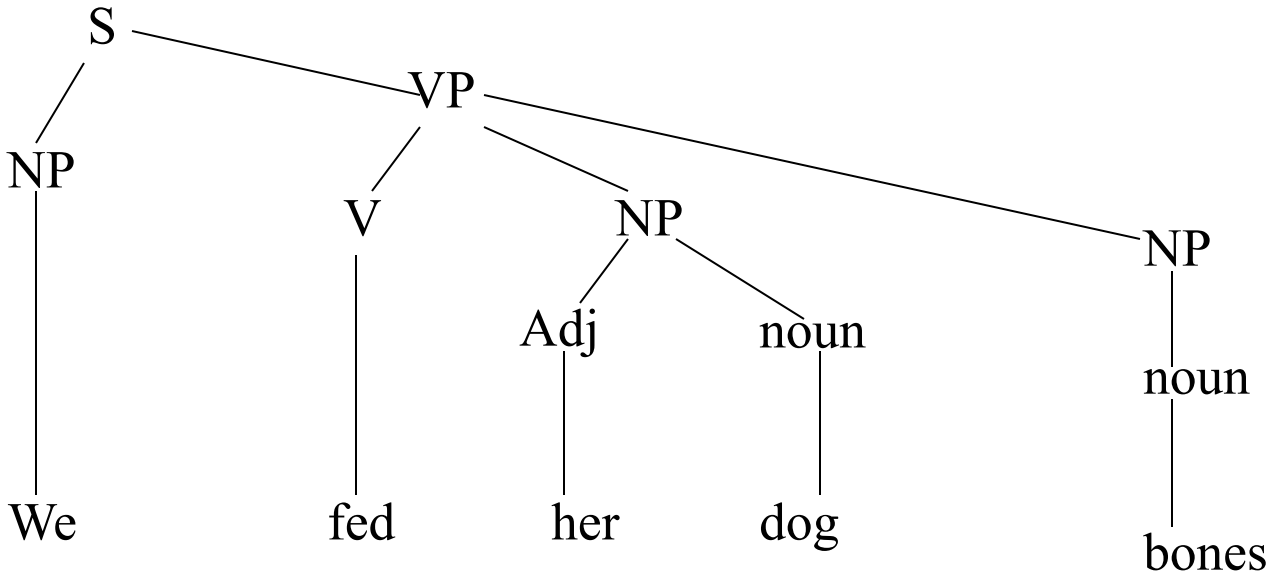
- Usually given by online lexicon such as WordNet, including:
  - Word with senses
    - Example: launch
  - Definitions
    - Noun sense 1: a large, usually motor-driven boat used for carrying people on rivers, lakes harbors, etc.
    - Verb sense 1: set up or found
  - Synonyms
    - Verb sense 1: establish, set up, found

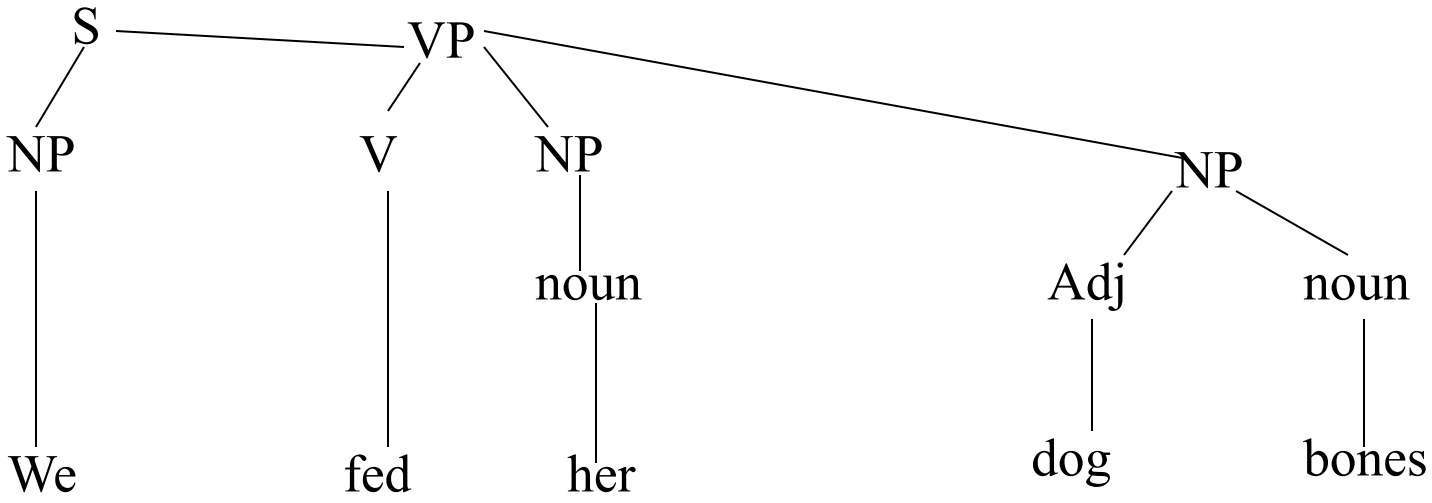
# Syntactic Analysis

- Analyzing of words in a sentence so as to uncover the grammatical structure of the sentence
- Requires both a grammar and a parser



# Syntactic Ambiguity: We fed her dog bones



# Semantics

- Determining possible meanings of a sentence

- Goal: Capturing meaning of a sentence in a knowledge representation formalism

# Semantic Role Labeling (SRL) Problem

- In a sentence, a verb and its semantic roles form a proposition; the verb can be called the *predicate* and the roles are known as *arguments*.
- Given a target verb, the Semantic Role Labeling task is to identify and label each semantic role present in the sentence.
  - *"When Disney offered to pay Mr. Steinberg a premium for his shares, the New York investor didn't demand the company also pay a premium to other shareholders."*
    - Example roles for the verb "pay", using roles more specific than theta roles:
    - When [$_{payer}$ Disney] offered to [$_{V}$ pay] [$_{recipient}$ Mr. Steinberg] [$_{money}$ a premium] for [$_{commodity}$ his shares], the New York investor …

# Semantic Relation Extraction

"Coca-Cola Enterprises, Inc. said its Atlanta Coca-Cola Bottling Co. unit and its CEO, John Smith, is a target of an investigation into alleged antitrust violations in the soft-drink industry by a federal grand jury in Atlanta."

- Extracted Relations:
  - Owns Coca-cola Enterprises, Inc. Coca-cola Bottling Co.
  - Employs Coca-cola Enterprises, Inc. John Smith
  - Location Coca-cola Bottling Co. Atlanta
  - Location federal grand jury Atlanta

# Discourse Level

- Determining meaning in texts longer than a sentence
- Making connections between component sentences
  - multi-sentence texts are not just concatenated sentences to be interpreted singly
  - Documents may have distinct patterns in different sections: introduction, conclusions, methodology, etc.
  - eg. "Please use the toilet, not the pool," & "Pool for members only."

# Pragmatics

- The purposeful use of language in situations
  - A functional perspective
- Those aspects of language which require context for understanding
- Goal: explain how extra meaning is *read into* texts without actually being encoded in them
  - Requires much world knowledge
  - Understanding of intentions / plans / goals

# Pragmatics

- ‘What time do you call this?’
    - Literal Meaning: What time is it?
      Literal Response: A time (e.g. ’11 am.’)
    - (Pragmatic Meaning: a different question entirely, e.g. Why are you so late?
      Pragmatic Response: Explain the reason for being so late.)

# More Techniques for NLP

- Corpus Statistics
  - Frequencies of words
  - Frequencies of word pairs, using co-occurrence or semantic measures
- Classification or other Machine Learning
  - Use NLP to produce features, also known as attributes, of the text
  - Classify the text according to a set of labels
    - Classify customer reviews as positive or negative
    - Classify news articles according to topic

# Natural Language Processing ToolKit (NLTK)

School of Information Studies
Syracuse University

# Processing Text with NLTK

- NL Tool Kit provides libraries of many of the common NLP processes at various language levels
  - Leverage these libraries to process text
- Goal: learn about and understand how NLP can be used to process text without programming all processes
  - However, some programming is required to
    - Call libraries
    - Process data
    - Customize NLP processes
  - Programming language is Python

# Characteristics of Python

- Easy-to-learn scripting language
- Object-oriented, with modules, classes, exceptions, high-level dynamic data types, similar to Java
- Strongly typed, but without type declarations (dynamic typing)
- Regular Expressions (REs) and other string processing features
- Many libraries offer wide functionality
- Case sensitive: e.g. text1 ≠ Text1

# Natural Language Toolkit (NLTK)

- A suite of Python libraries for symbolic and statistical natural language programming
  - Developed at the University of Pennsylvania
- Developed to be a teaching tool and a platform for research NLP prototypes
  - Goal of code is to be clear, rather than fastest performance
- Latest version is compatible with Python 3.x

Online book:
http://www.nltk.org/book/

Authors:
Edward Loper, Ewan Kline
and Steven Bird

# Lab: Hands-on of NLTK

School of Information Studies
Syracuse University

# Lab Session: Tasks (1)

- Get started by using Jupyter Notebook
  - Figure out where to find the nltk
  - Arithmetic expression and strings
    - define/add two strings variables

- Work with the *book* collection in NLTK
  - Download the nltk *book* collection
  - Import the *book* collection
  - String/ List:
    - try to add one string variable with one list variable

# Lab Session: Tasks (2)

- Search text
  - display the context of the word “monstrous” in text1
  - display the context of the word “affection” in text2
  - find **similar** words to “monstrous” in text1
  - display the same context for “monstrous” and “delightfully” in text 1

# Lab Session: Tasks (3)

- Counting vocabulary:
  - how many words in text3?
  - how many unique words in text 3?
  - sort the unique words in text 3
  - calculate the lexical richness of text3
  - first/last 10 words in text 3
  - how many times the word ‘he’ used in text3?
  - % of ‘he’ used in text3
  - % of another randomly picked word used in text3
  - difference between text3.concordance vs. text3.count

# Lab Session: Tasks (4)

- Control in Python:
  - Count/display the total number of tokens in text 1 that have more than 8 letters
  - Count/display the tokens in text5 that end with “ion”.

| Operator | Relationship |
|---|---|
| < | less than |
| <= | less than or equal to |
| == | equal to (note this is two "=" signs, not one) |
| != | not equal to |
| > | greater than |
| >= | greater than or equal to |

| Function | Meaning |
|---|---|
| s.startswith(t) | test if s starts with t |
| s.endswith(t) | test if s ends with t |
| t in s | test if t is a substring of s |
| s.islower() | test if s contains cased characters and all are lowercase |
| s.isupper() | test if s contains cased characters and all are uppercase |
| s.isalpha() | test if s is non-empty and all characters in s are alphabetic |
| s.isalnum() | test if s is non-empty and all characters in s are alphanumeric |
| s.isdigit() | test if s is non-empty and all characters in s are digits |
| s.istitle() | test if s contains cased characters and is titlecased (i.e. all words in s have initial capitals) |

# Your tasks before next class

- No class on next Monday!
- Install Jupyter Notebook and NLTK on your personal computer
- Weekly lab exercise: uploaded in "discussion" under "Lab1" (due on 11:59 pm EST on this Sunday)
  - In the body of the post, you should upload a screenshot of your coding along with the results.
- [Optional]: Week 1 in-class activity– in Discussion (available before 1 pm today)
- Readings for Week 2