Introduction to Natural Language Processing

School of Information Studies
Syracuse University

Natural Language Processing (NLP)

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

Computational linguistics: doing linguistics on computers

- Closely related, often treated as synonymous with NLP
- "The field is Computational Linguistics, but what computational linguists do is Natural Language Processing"

Natural Language as the User Interface

Goal is complete natural language understanding

- Enables computers to interact with humans with natural language
- Vision of a future with HAL in *2001: A Space Odyssey*

Most common current approach is to craft human/computer interfaces that are in terms that the computer can understand

- XML, drop-down boxes, other forms of knowledge representation
- Cleverness is supplied by the human

Nascent natural language interfaces are being deployed

- Apple's Siri, the Google Assistant, Amazon's Alexa

Where Is NLP Now?

Goals can be far-reaching

- True text understanding
- Reasoning about knowledge in text
- Real-time participation in spoken dialogues

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
- Remembering previous searches and contexts to guide future interactions

Currently, NLP is providing these practical applications
(yet still dreaming of the AI goals)

Need for NLP

Huge amounts of data

- Internet and Intranet

Applications for processing large amounts of texts require NLP expertise

Data Science/Text Mining

Classify text into categories

Index and search large texts

Automatic translation of web documents in different languages

Speech understanding, e.g. phone conversations

Information extraction, e.g. extract useful information from resumes

Automatic summarization

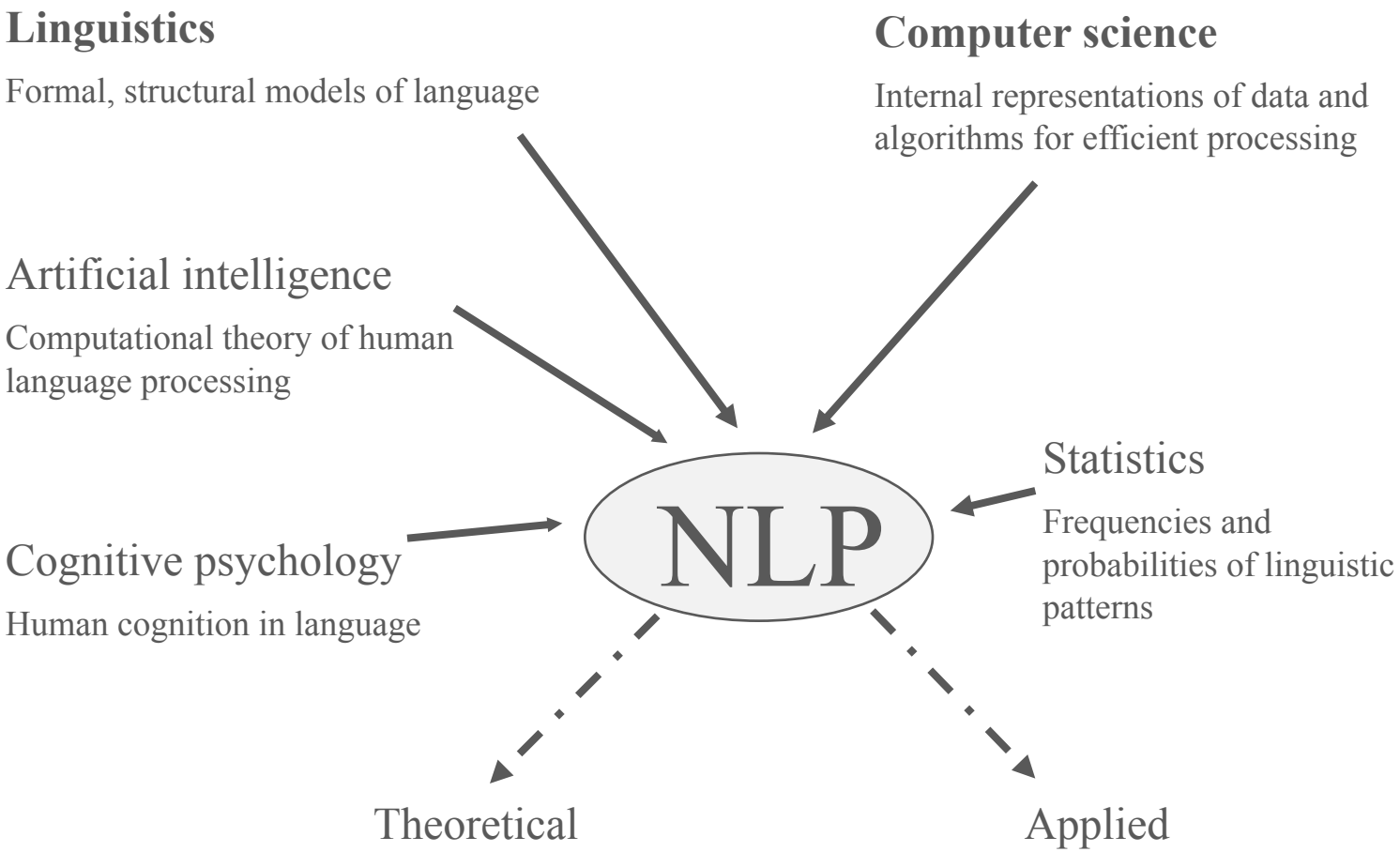
- Condense one book into one page
- Daily news summaries

Question answering

Knowledge acquisition

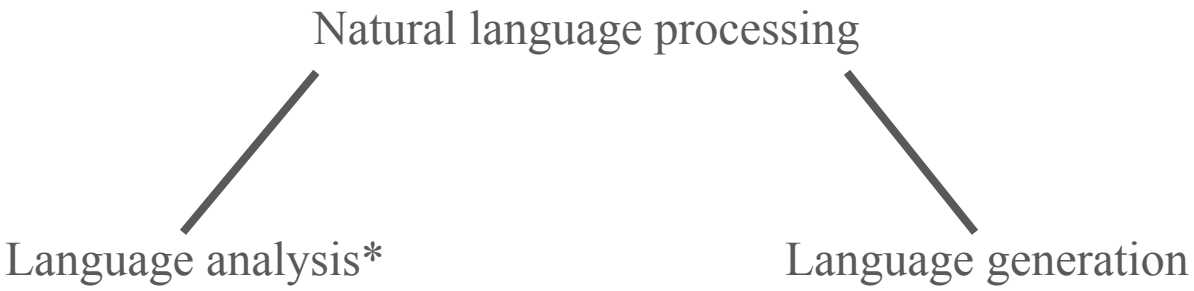
Text generations/dialogues

Fields Contributing to NLP



Two Sides of NLP: Analysis and Generation

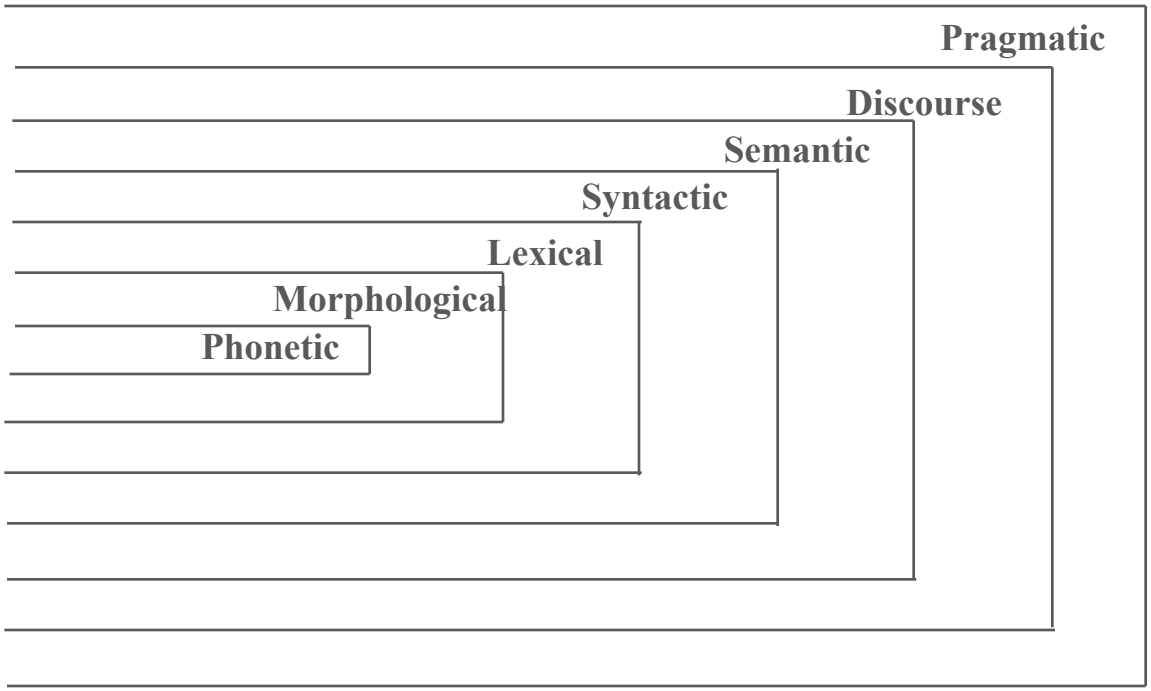
1. Paraphrase an input text
2. Translate it to another language or representation
3. Answer questions about it
4. Draw inferences from it
5. Phrase the results in natural language



*Main emphasis in this course

Synchronic Model of Language

The synchronic model postulates levels of language to understand the use of language at this point in time.



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

Ambiguous Newspaper Headlines

Ban on Nude Dancing on Governor's Desk

Iraqi Head Seeks Arms

Juvenile Court to Try Shooting Defendant

Teacher Strikes Idle Kids

Stolen Painting Found by Tree

Local High School Dropouts Cut in Half

Red Tape Holds Up New Bridges

Clinton Wins on Budget, but More Lies Ahead

Hospitals Are Sued by 7 Foot Doctors

Kids Make Nutritious Snacks

- Examples collected by Chris Manning

Introduction to NLP Applications

School of Information Studies
Syracuse University

NLP Application Areas

Machine translation (MT): conversion of text from one language to another

- Google, Yahoo, and Bing all have language translators
- MT techniques use context, not just word-for-word substitution
- Often statistically based patterns of word usage and context

Google Translate

School of Information Studies
Syracuse University

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

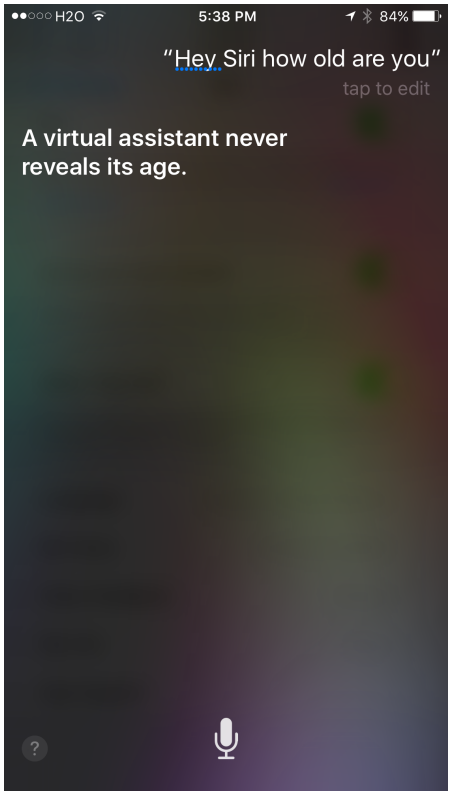
- Competitive intelligence analyzes news text and web blogs for:
  - Names of people, companies, and other entities
  - Relations between them (e.g., corporate roles) or events such as mergers

Weblog Analytics
Data-mining of weblogs, discussion forums, message boards, user groups, and other forms of user-generated media

- Product marketing information
- Political opinion tracking, Social network analysis
- Buzz analysis (what's hot, what topics are people talking about now)

NLP Application Areas

Human–computer interfaces:
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 question-answering system trained to play *Jeopardy*

Extensive development of NLP techniques

Trends

An enormous amount of knowledge is now available in machine-readable form as natural language text.

Conversational agents are becoming an important form of human–computer communication.

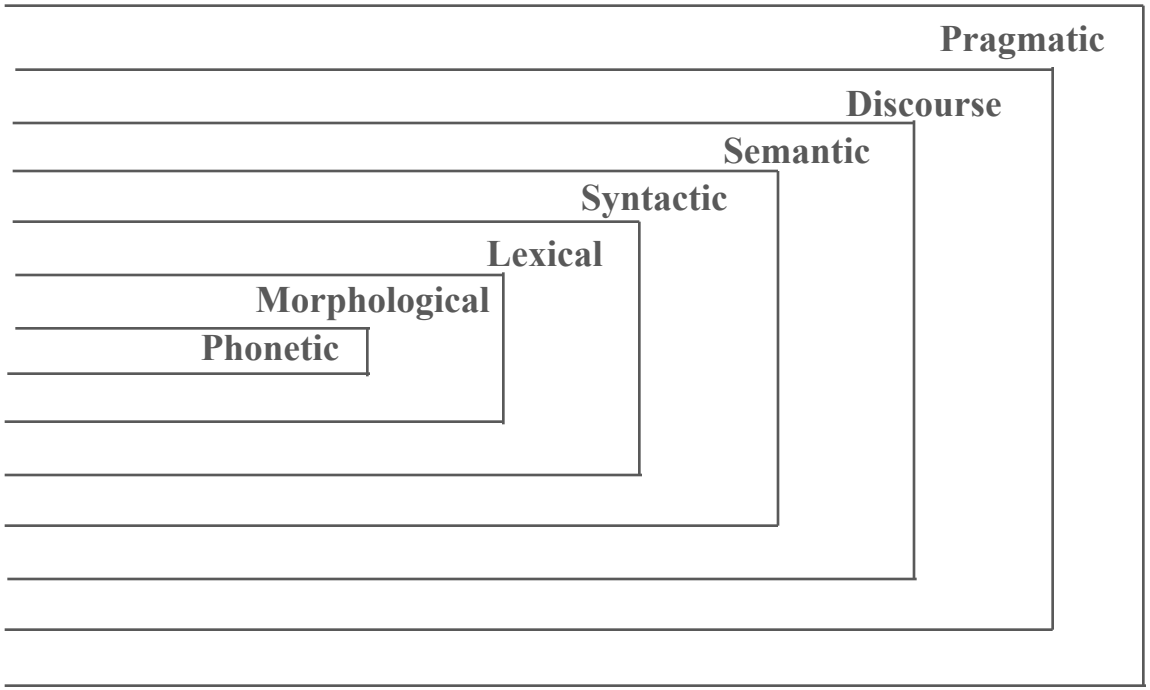
Much of human–human communication is now mediated by computers.

Levels of Language: Part 1

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).



Levels of Language

The more exterior the level of language processing:

- The larger the unit of analysis
  - Phoneme -> morpheme -> word -> sentence -> text -> world
  - The less precise the language phenomena
- The more free choice and variability
  - The less rule-oriented, more exceptions to regularities
- The more levels it presumes a knowledge of or reliance on
- Theories used to explain the data move more into the areas of cognitive psychology and AI

Lower levels of the model have been more thoroughly investigated and incorporated into NLP systems

Speech Processing

Interpretation of speech sounds within and across word
sound waves are analyzed and encoded into a digitized signal.

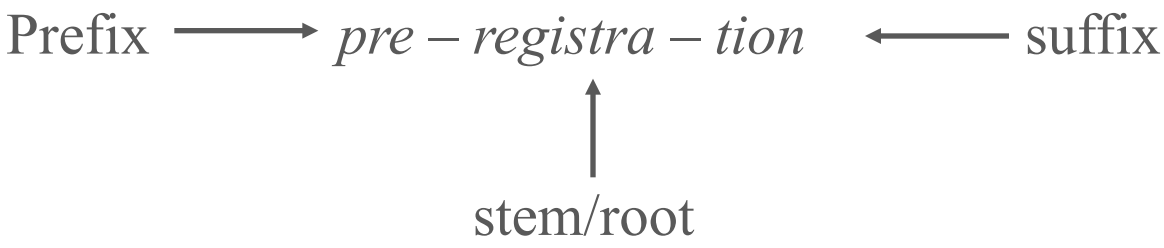
Rules used in Phonological Analysis

1. Phonetic rules: sounds within words
2. Phonemic rules: variations of pronunciation when words are spoken together
3. Prosodic rules: fluctuation in stress and intonation across a sentence—rhythm, volume, pitch, tempo, and stress

Separating the spoken word "cat" into three distinct **phonemes,** /k/, /æ/, and /t/, requires phonemic awareness.

Morphological Analysis

Deals with the componential nature of lexical entities:



What features do inflections reveal in English?

- Verbs → tense and 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…

The POS tags in this example are taken from the Penn Treebank tag set.

Lexical: Word Meanings

Usually given by online lexicon such as WordNet

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

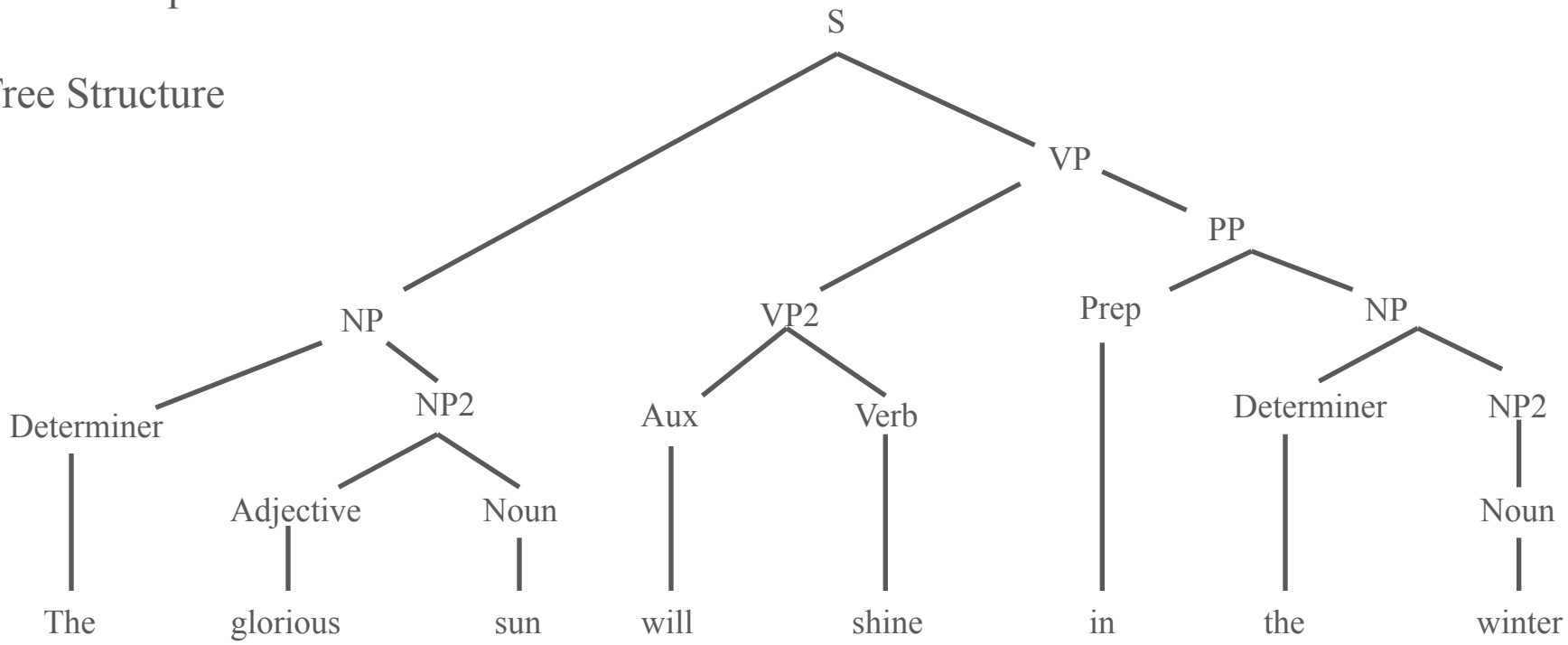
Analyzing of words in a sentence so as to uncover the grammatical structure of the sentence:

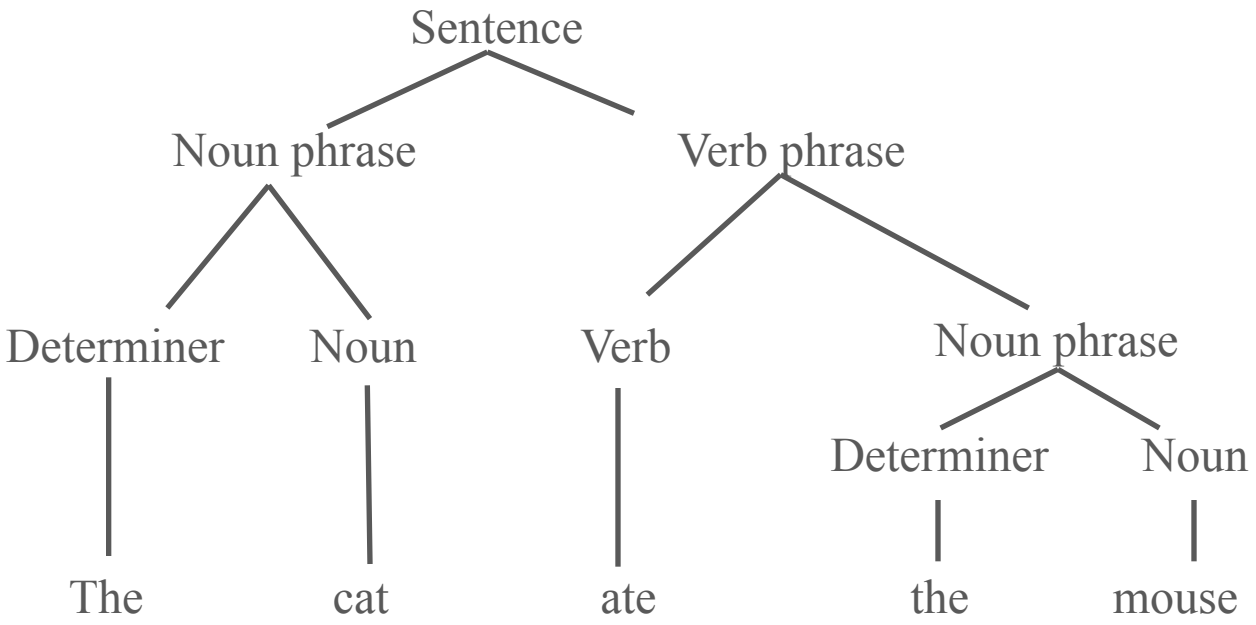
How words are grouped to form phrases and the order in which they occur

Requires both a grammar and a parser

Produces a delinearized representation of a sentence, which reveals dependency relationships between words

Tree Structure





The phase structure rules underlying this analysis are as follows:

Sentence → Noun phrase Verb phrase

Noun phrase → Determiner Noun

Verb phrase → Verb Noun phrase

Determiner = The

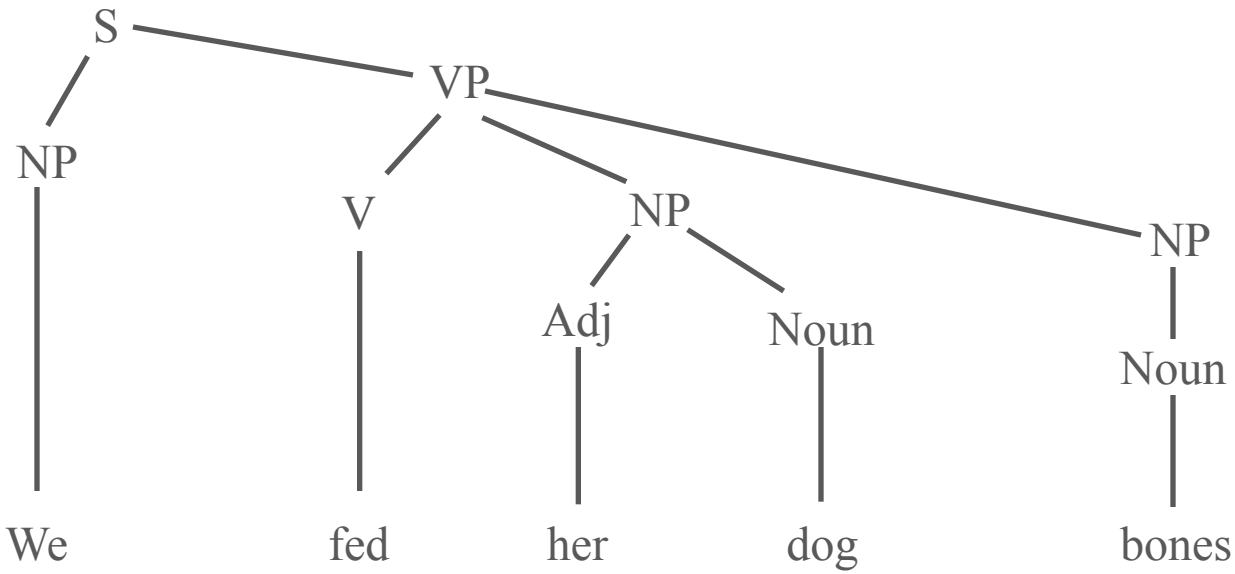
Noun = cat

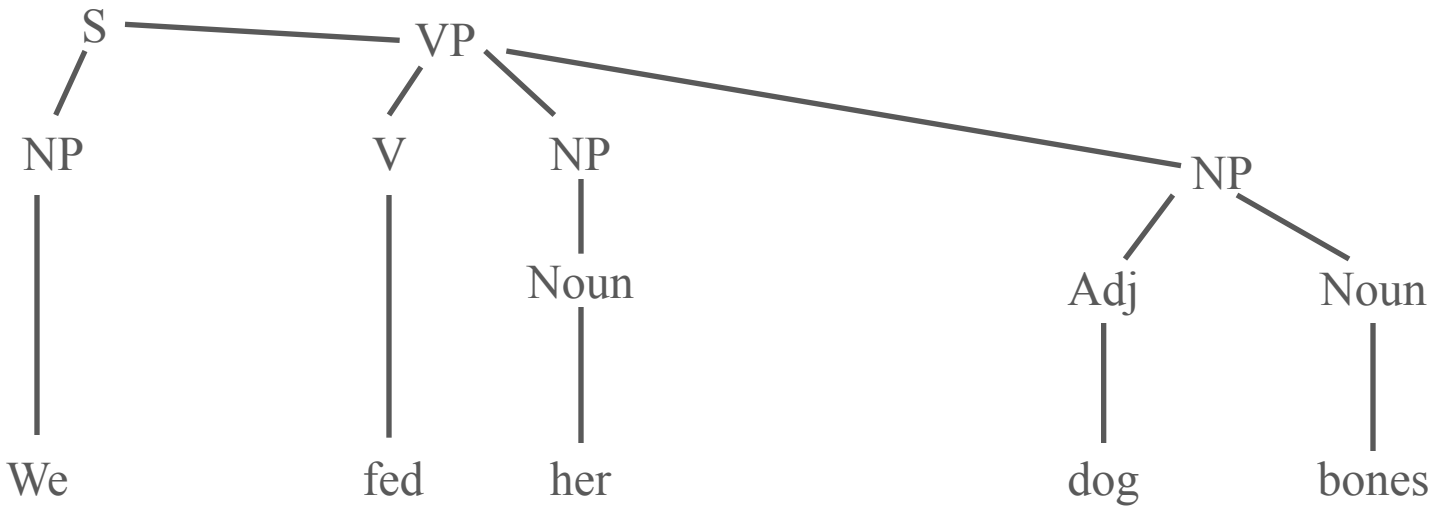
Noun = mouse

Verb = ate

Parsing a sentence using simple phrase structure rules

Syntactic Ambiguity: We Fed Her Dog Bones





Levels of Language: Part 2

School of Information Studies
Syracuse University

Semantics

Determining possible meanings of a sentence

- Interactions among words affect lexico-semantic interpretation

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.
- Text in dialogues has distinct forms according to position in the dialogue

Interpretation of later-mentioned entities depends on interpretation of earlier-mentioned entities—"coreference"

Anaphora (Coreference) Resolution

Excerpt from story by Farhad Manjoo of *Slate,* "Siri vs. Google," 2014. The coreference is to understand the meaning of the word "his" in the second sentence.

"Google Voice Search isn't close to realizing that vision, but it's not impossibly far off either. Huffman points out that Google's app can already hold very small conversations. It understands pronouns, so if you ask, 'Who is Barack Obama?' and then ask, 'Who is his wife?', it knows that his refers to Obama. And most important, it gives you the correct answer.

I just tried the same set of queries with Siri. First, she correctly identified the president. But when I asked, 'Who is his wife?' she shot back, 'What is your wife's name?' That's not what I asked."

Pragmatics

The purposeful use of language in situations

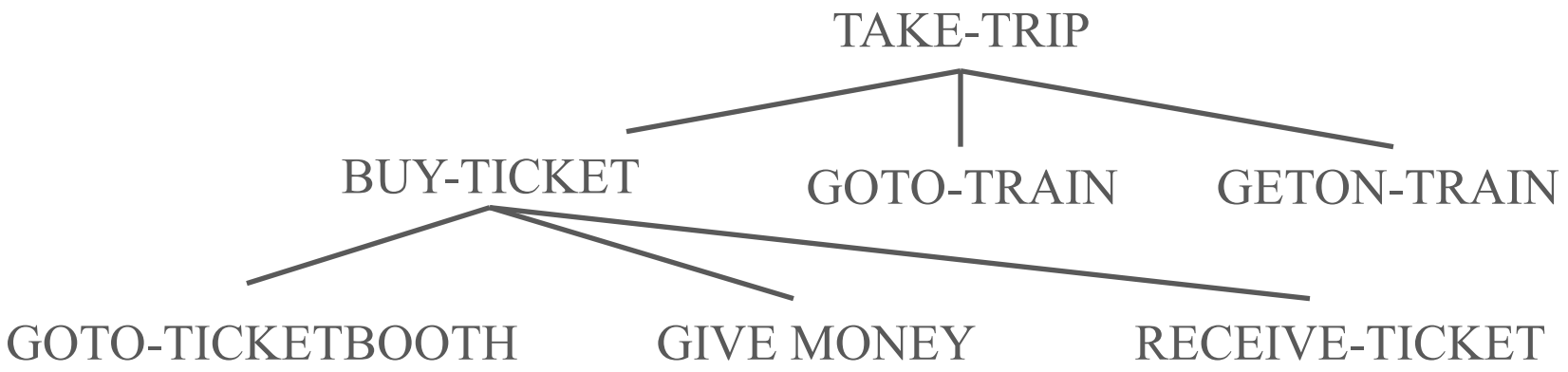
- A functional perspective: understanding the person's purpose in using the text

Those aspects of language that require context for understanding

Goal is to explain how extra meaning is *read into* texts without actually being encoded in them

- Requires much world knowledge
- Understanding of intentions/plans/goals

Pragmatics



Sketch of a commonsense task plan to take a trip

Techniques for NLP Analysis

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

Corpus Words

School of Information Studies
Syracuse University

First Text-Processing Task: Word Counts

Preliminary Text Processing Required

Define the words so that you can count them.

- Filter out "junk data."
  - Formatting/extraneous material
  - First, be sure it doesn't reveal important information
- Deal with upper-/lowercase issues.
  - Ignore capitalization at beginning of sentence? Is "They" the same word as "they"?
  - Ignore other capitalization? In a name such as "Unilever Corporation" is "Corporation" the same word as "corporation"?

Preliminary Processing Required

Tokenization (or Word Segmentation)

- Decide how to separate the characters in the sentence into individual words
  - Words separated by "white space" or by special characters in English
  - No white space in Japanese language
  - In some languages, there are complex compound words: "*Lebensversicherungsgesellschaftsangestellter*"
- Requires decisions on recognizing and dealing with punctuation
  - Apostrophes (one word it's vs. two words it 's)
  - Hyphens (*snow-laden* vs. *New York-New Jersey*)
  - Periods (kept with abbreviations vs. separated as sentence markers)

Preliminary Processing Required

Morphology (to stem or not to stem?)

- Depends on the application
- With stemming
  - "Cat" is the same word as "cats"
  - "Computing" is the same word as "compute"

Additional issues if OCR'd data or speech transcripts in order to correct transcription errors

Word Counting in Corpora

Terminology for word occurrences

- Tokens: the total number of words
- Distinct tokens (sometimes called word types): the number of distinct words, not counting repetitions
- The following sentence from the Brown corpus has 16 tokens and 14 distinct tokens:

  They picnicked by the pool, then lay back on the grass and looked at the stars.

Word Frequencies

Count the number of each token appearing in the corpus (or sometimes single document)

A frequency distribution is a list of all tokens with their frequency, usually sorted in the order of decreasing frequency

Used to make "word clouds"

- For example: http://www.tumblr.com/tagged/word+cloud

Natural Language Processing ToolKit (NLTK)

School of Information Studies
Syracuse University

Processing Text With NLTK

NL ToolKit provides libraries of many of the common NLP processes at various language levels.

- Leverage these libraries to process text.

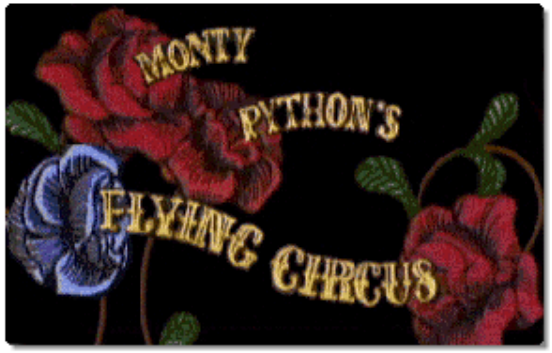
The goal is to learn about and understand how NLP can be used to process text without programming all processes.

- However, some programming is required to:
  - Call libraries
  - Process data
  - Customize NLP processes
- Programming language is Python.

Python and NLP

Python is freely available for many platforms from the Python Software Foundation:

- http://www.python.org/
- Named for the group Monty Python
- We are using Python version 3.x
  - Not backward compatible with Python 2.x



The group in 1969

Characteristics of Python

Easy-to-learn scripting language, similar in many aspects to Perl

- But with WYSIWYG block structure

Object-oriented, with modules, classes, exceptions, high-level dynamic data types, similar to Java

Strongly typed, but without type declarations (dynamic typing)

Regular Expressions and other string-processing features

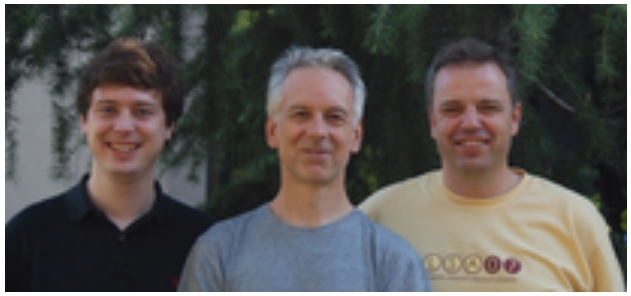
Many libraries offer wide functionality

- https://xkcd.com/353/

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

- Data types are packaged as classes
- Goal of code is to be clear, rather than fastest performance
  - But, increasingly, production-level software is made available through wrappers

Natural Language ToolKit (NLTK)

Latest version is compatible with Python 3.x

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

Authors:
Edward Loper, Ewan Kline, and Steven Bird