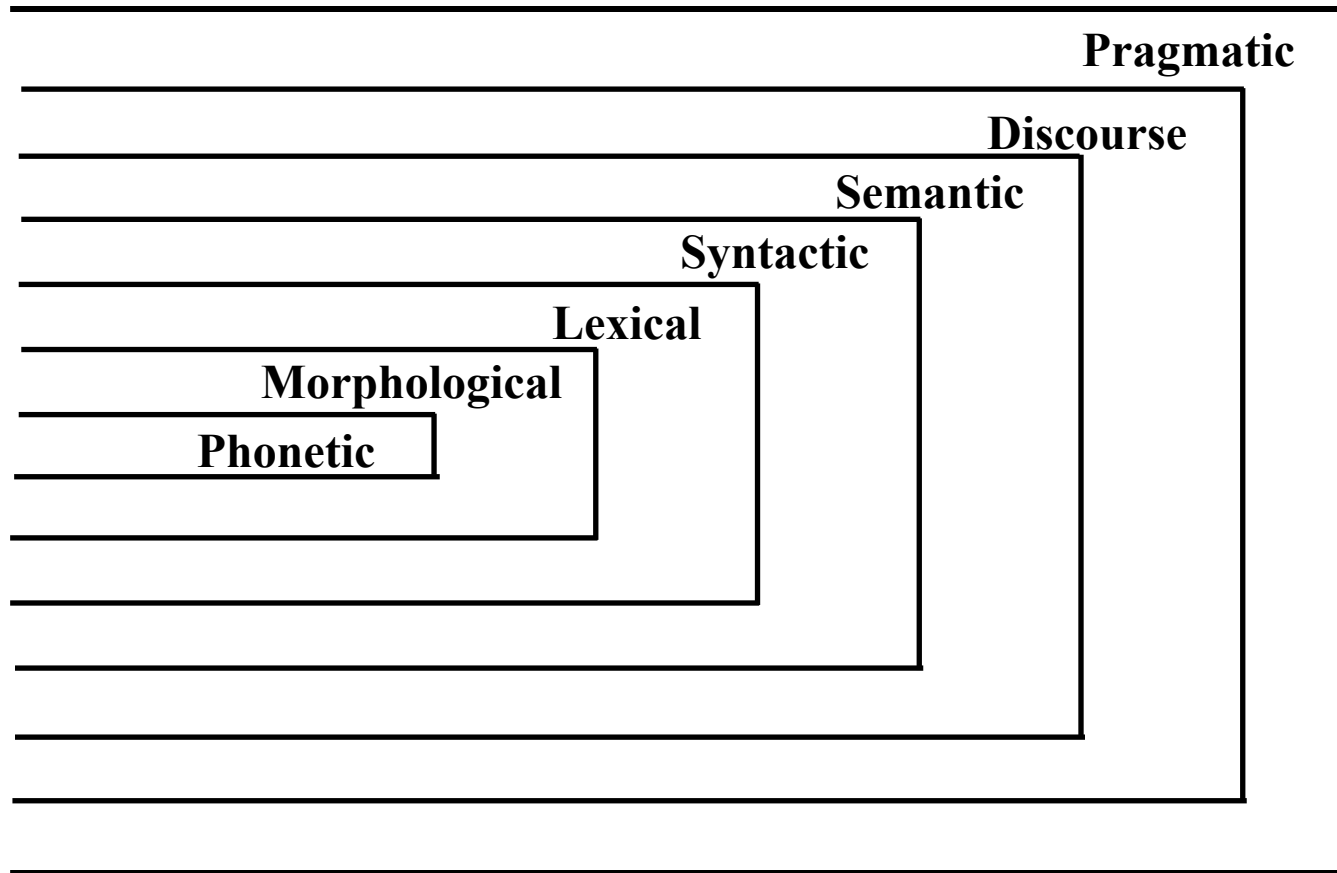

Levels of Language used by Natural Language Processing

Levels of Language Analysis

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



Synchronic Model 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 & variability
 - 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

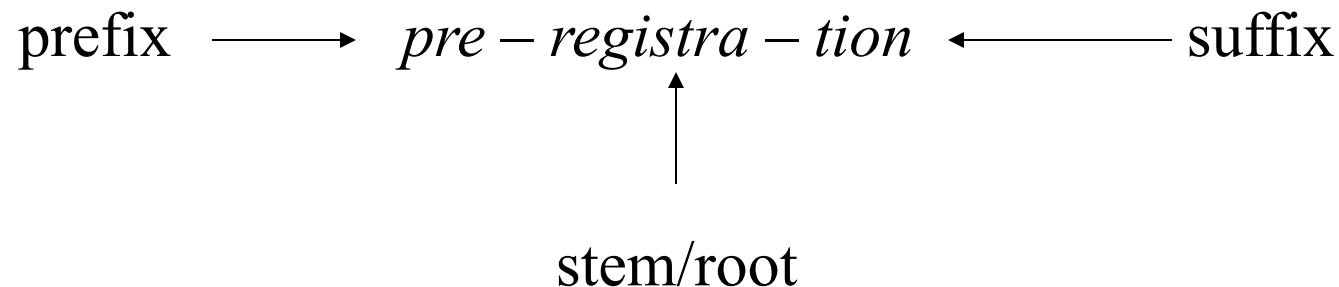
- Interpretation of speech sounds within & across words
- sound waves are analyzed and encoded into a digitized signal

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 & number

Nouns → single/plural

Adjectives → comparison features

Lexical

1. 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** ...

2. Productive rules which explain how new words are formed

highchair

egghead

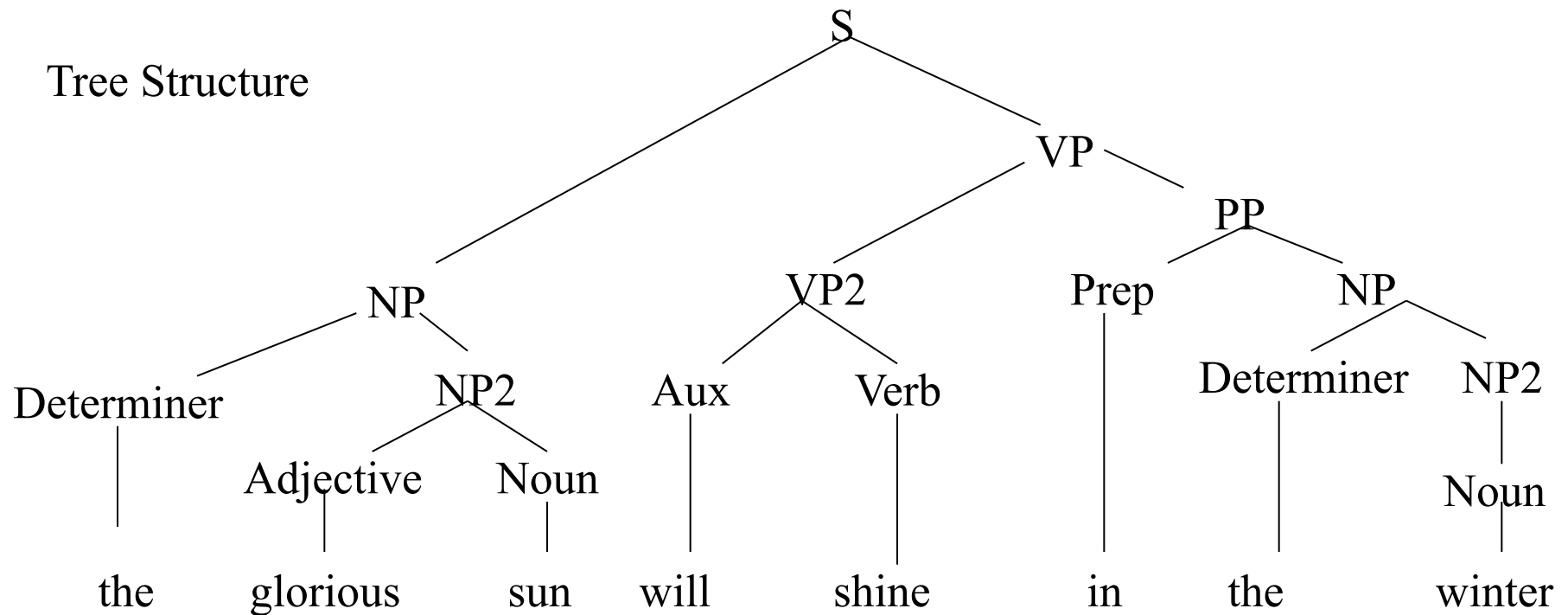
Word Level Meaning

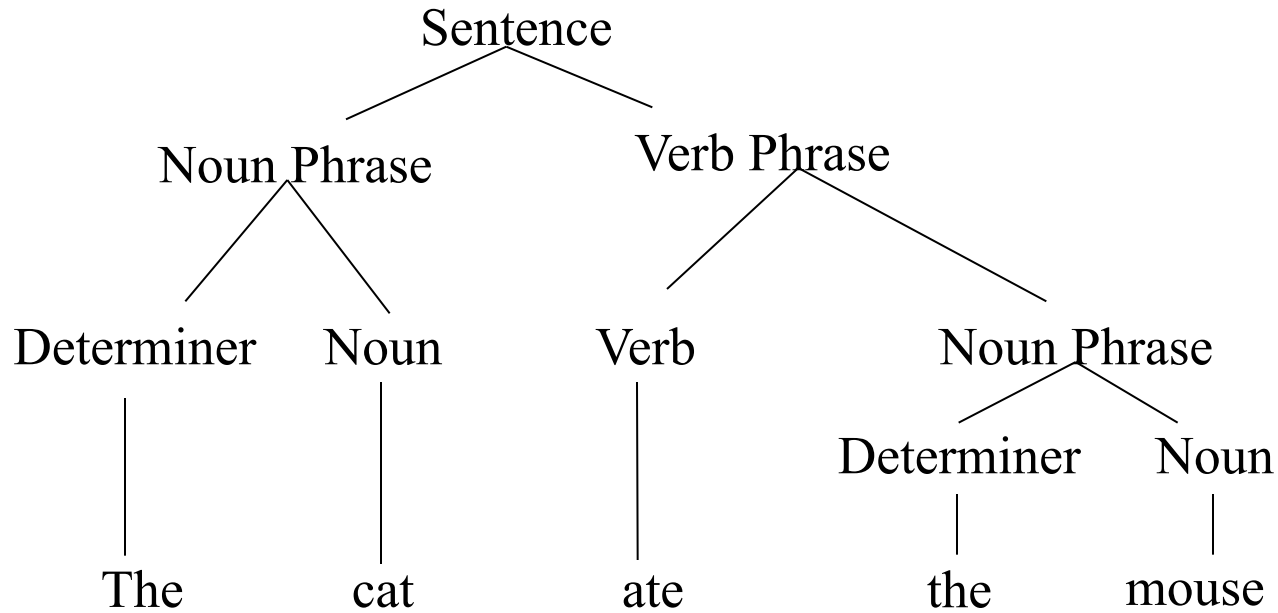
- Usually given by online lexicon such as WordNet
- Word with senses
 - Example: launch
- Definitions
 - Noun sense 1: a large, usually motor-driven boat used for carrying people on rivers, lakes harbors, etc.
 - Verb sense 1: set up or found
- Synonyms
 - Verb sense 1: establish, set up, found

Syntactic Analysis

- analyzing of words in a sentence so as to uncover the grammatical structure of the sentence
- requires both a grammar and a parser
- produces a de-linearized 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

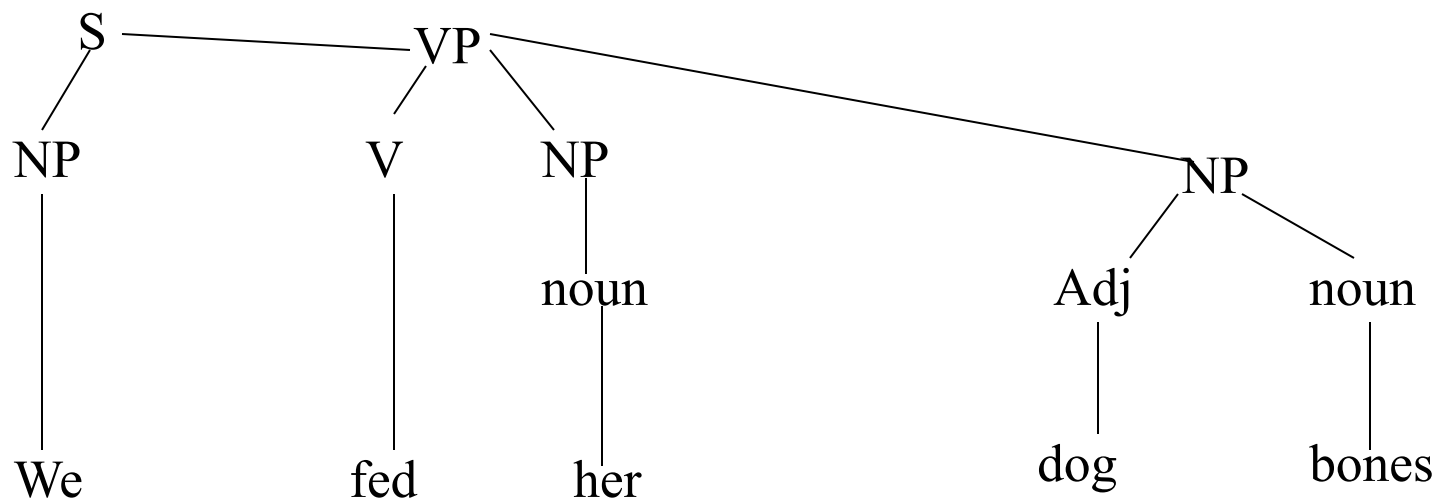
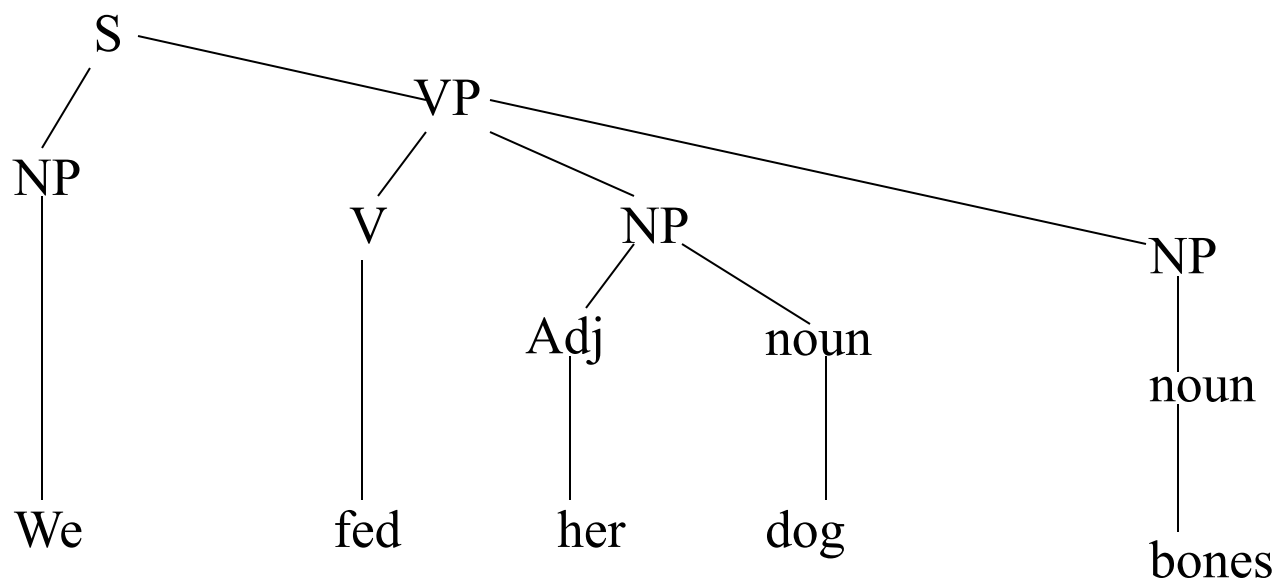
Noun = cat

Noun = mouse

Verb = ate

Parsing a sentence using simple phrase structure rules

Syntactic Ambiguity: We fed her dog bones



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

- 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 dialogs has distinct forms according to position in the dialog
- interpretation of later-mentioned entities depends on interpretation of earlier-mentioned entities – ‘anaphora’

Anaphora (coreference) resolution

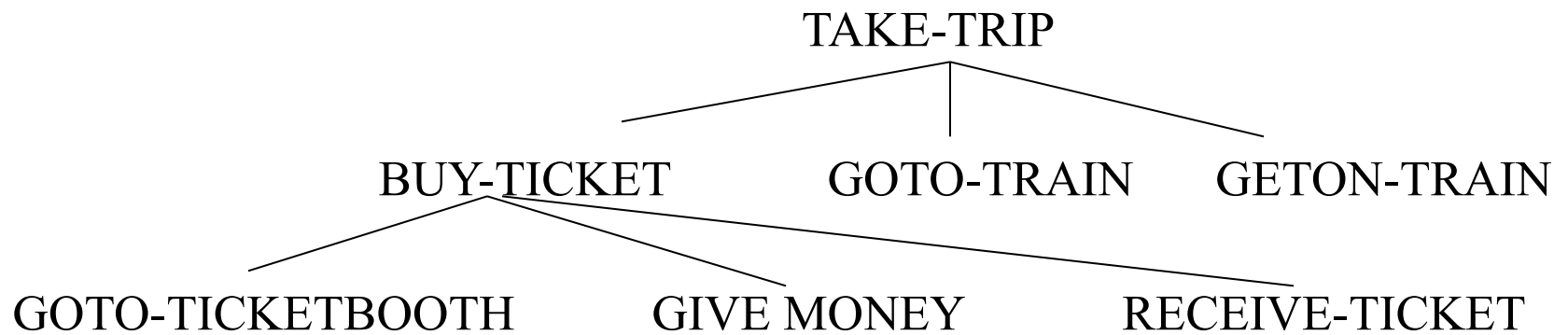
- Excerpt from story by Farhad Manjoo of Slate “Siri vs. Google” 2014

“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
- Those aspects of language which 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



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

NLP Demo

- The Stanford parser and related NLP software are some of the most widely used.
 - Stanford Core NLP demo:
 - <http://nlp.stanford.edu:8080/corenlp/>