LEVELS OF LANGUAGE USED BY NATURAL LANGUAGE PROCESSING

adopted some materials developed in previous courses by Nancy McCracken, Liz Liddy and others; and some instructor resources for the book “Speech and Language Processing” by Daniel Jurafsky and James H. Martin

Course Admin

- Class absence policy: written explanation (accepted once), make-up essay

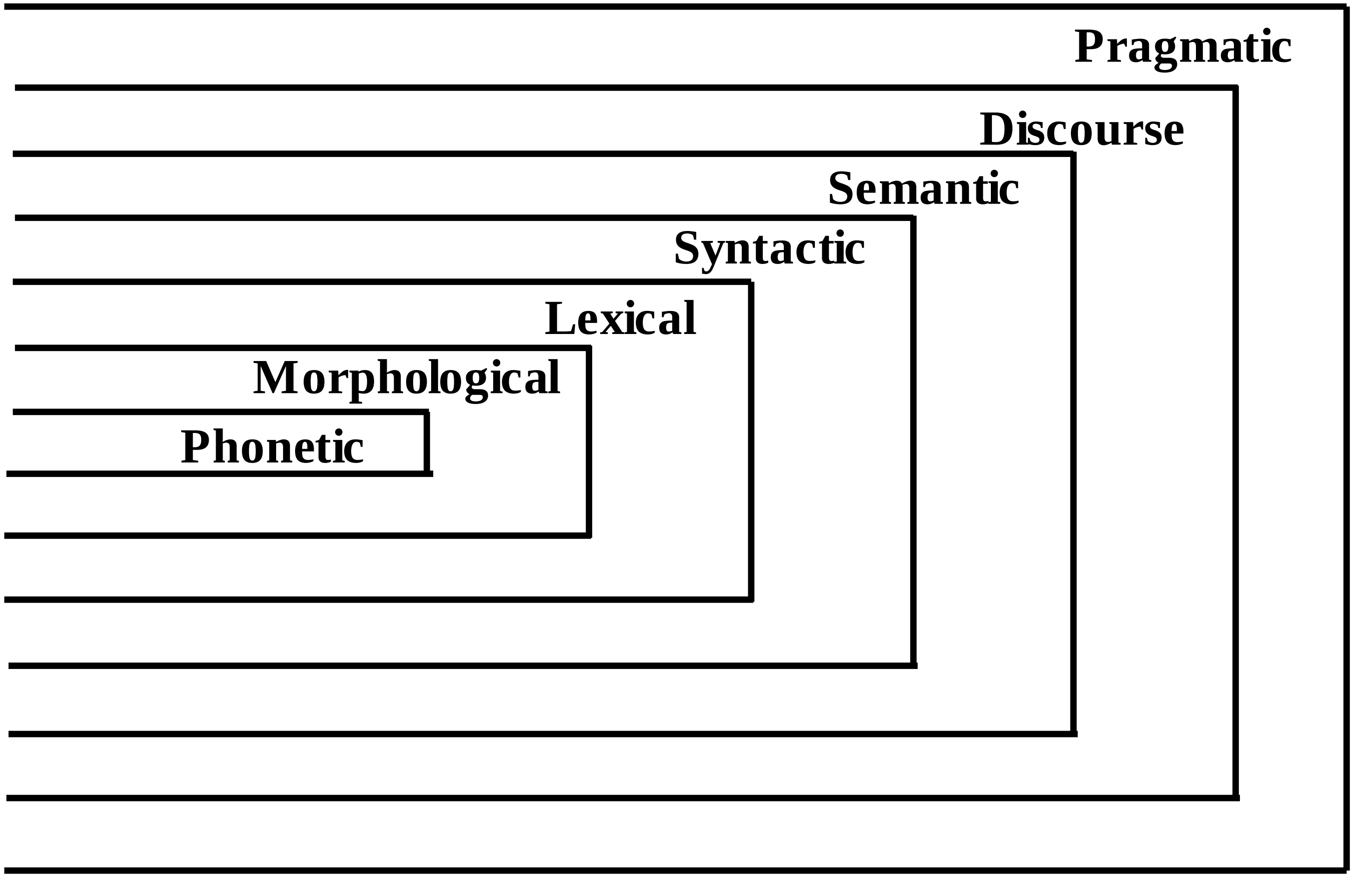
- No LATE submission will be accepted

- Poster Day (last day)

- Self-Introduction:
  - Name
  - Major
  - Something (interesting/unique/funny) about you
  - (optional) NLP topics that you are interested to investigate

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

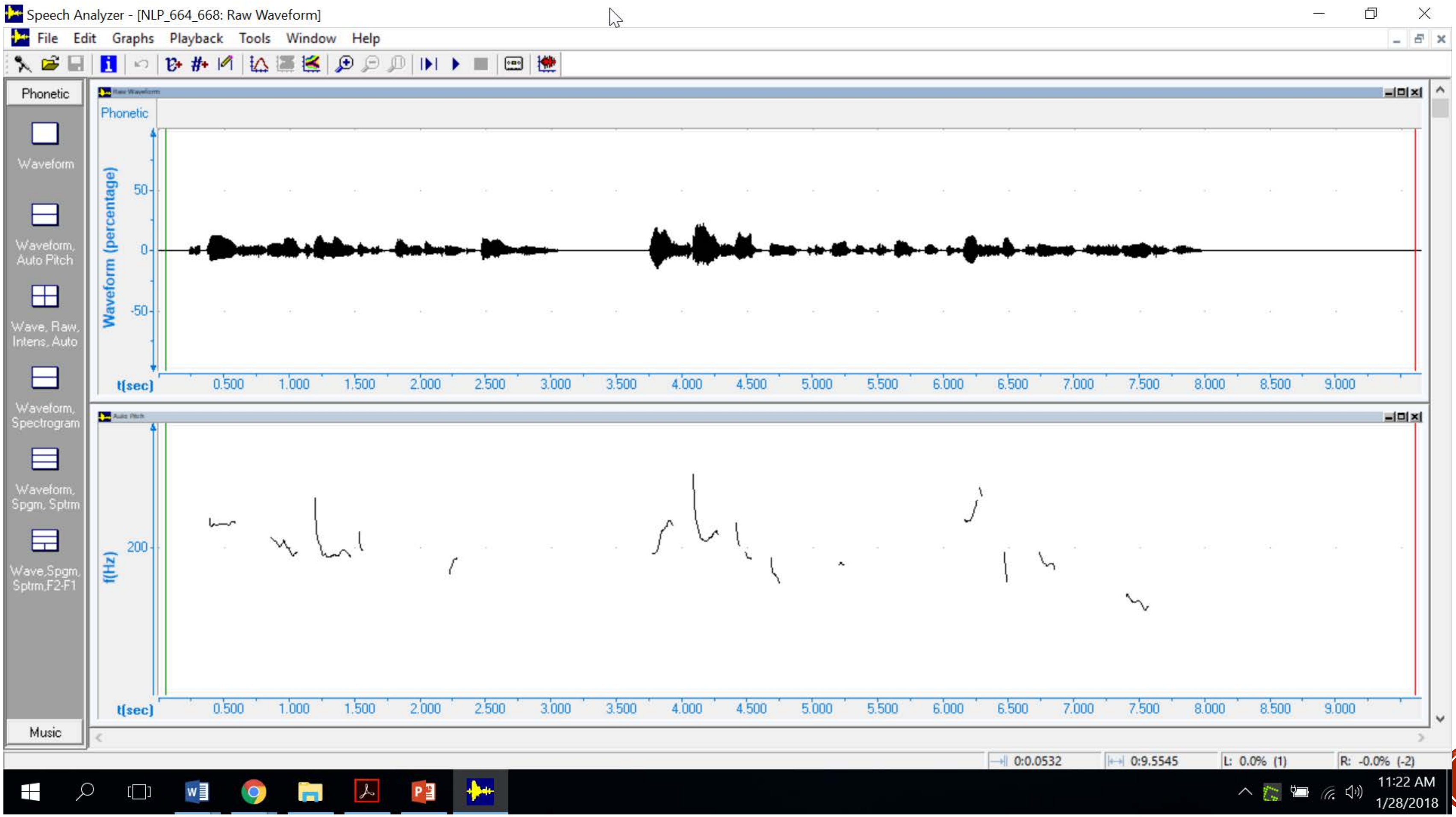
Rules used in Phonological Analysis

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

- Separating the spoken word "cat" into three distinct phonemes, /k/, /æ/, and /t/, requires phonemic awareness.
- The *prosodic rules* of communication tell what rhythm, volume, pitch, tempo, and stress is to be used during a conversation.

FREE SPEECH ANALYSIS SOFTWARE

http://www.personal.rdg.ac.uk/~llsroach/phon2/freespeech.htm



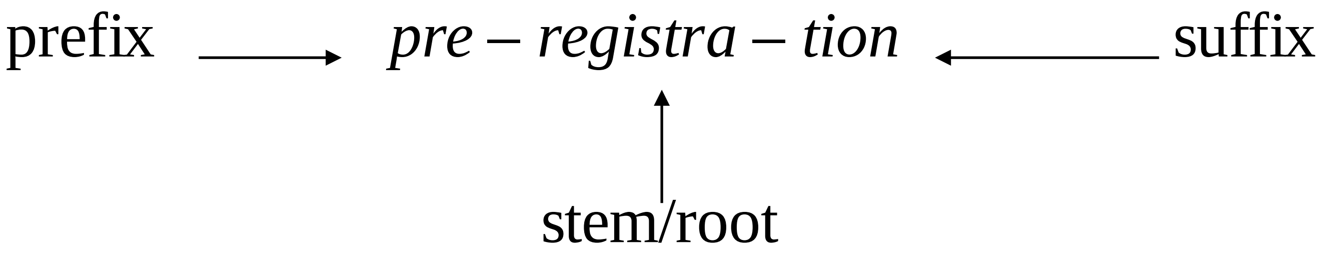
SPEECH RECOGNITION

- Google Speech Recognition API

https://docs.google.com/document/d/1NWnQzP5--7OufZtV5puW5IhClYceabpixAVADl7Rcy4/edit

Morphological Analysis

- deals with the componential nature of lexical entities:



prefix → *pre – registra – tion* ← suffix

stem/root

- What features do inflections reveal in English?

Verbs → tense & number

Nouns → single/plural

Adjectives → comparison features

Inflection: a change in the form of a word (typically the ending) to express a grammatical function or attribute such as tense, mood, person, number, case, and gender

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 …

In linguistics, a *treebank* is a parsed text corpus that annotates syntactic or semantic sentence structure. Penn Treebank is the first large-scale treebank.

Alphabetical List Of Part-of-speech Tags Used In The Penn Treebank Project

No.	Tag	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Word Level Meaning

- Usually given by online *lexicon* such as WordNet

“Linguistic theories generally regard human languages as consisting of two parts: a **lexicon**, essentially a catalogue of a language's words (its wordstock); and a **grammar**, a system of rules which allow for the combination of those words into meaningful sentences” --- from Wikipedia: https://en.wikipedia.org/wiki/Lexicon

Example in a lexicon:
launch:
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

LEXEME AND LEMMA

- **Lexeme** – a unit of lexical meaning that exists regardless of the number of inflectional endings it may have or the number of words it may contain
- **Lemma** – the canonical form, dictionary form, or citation form of a set of words
- Example: *run*, *runs*, *ran* and *running* are forms of the same lexeme, with *run* as the lemma.
    - *Lexeme*: the set of all the forms that have the same meaning
    - *Lemma*: the particular form that is chosen by convention to represent the lexeme.

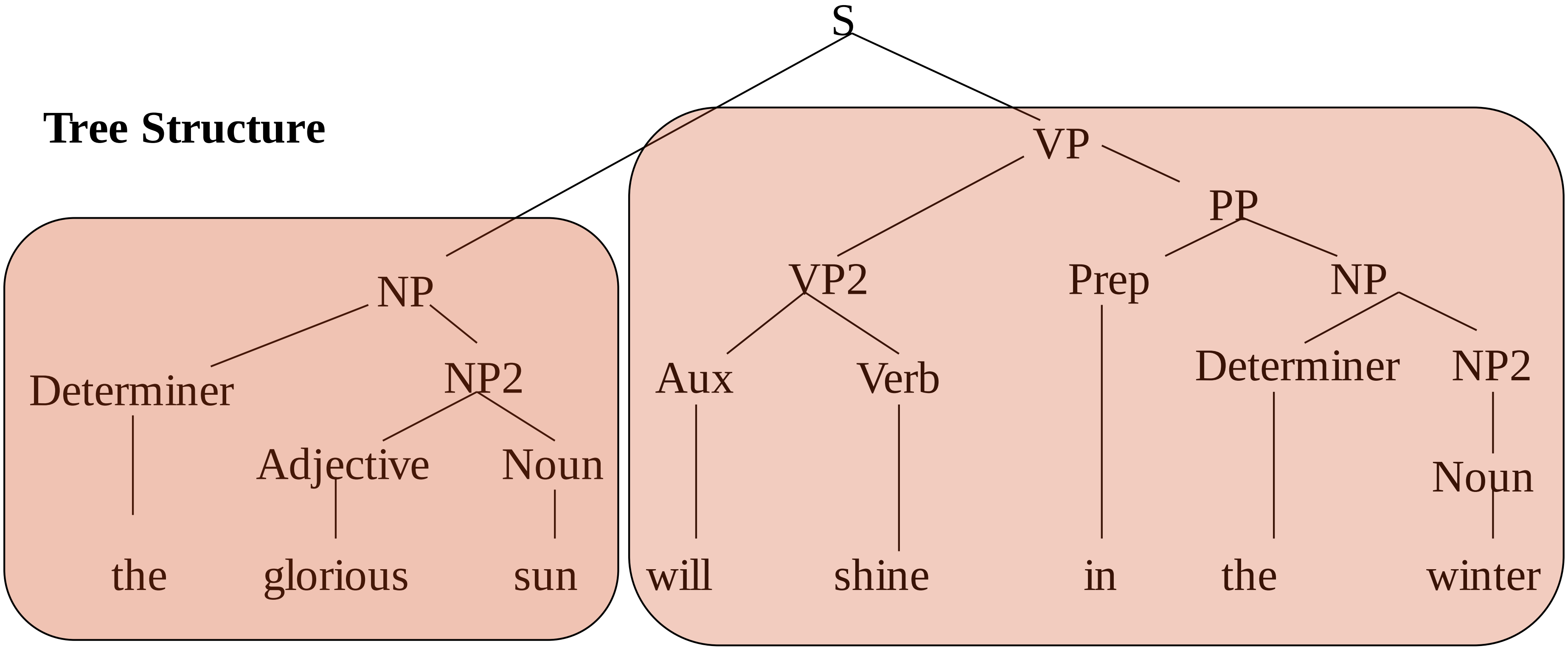
Syntactic Analysis

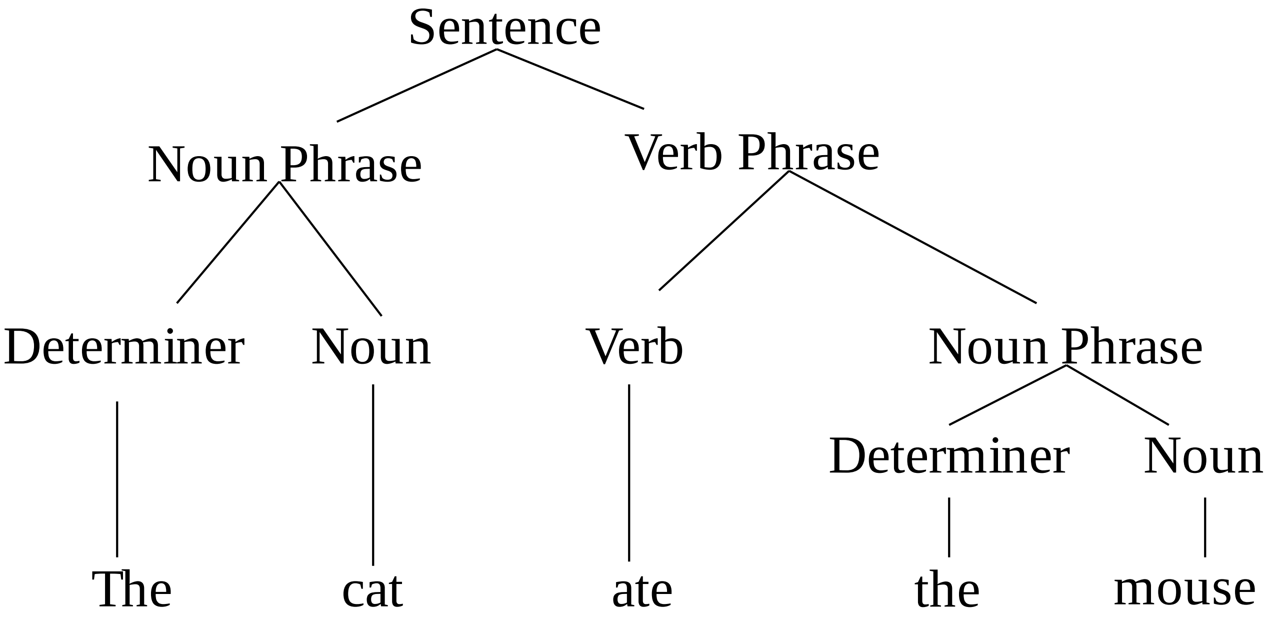
- analyzes words in a sentence so as to uncover the grammatical structure of the sentence

- requires both a grammar and a parser

- produces a de-linearized representation of a sentence which reveals dependency relationships between words

Tree Structure





Sentence

Noun Phrase Verb Phrase

Determiner Noun Verb Noun Phrase

Determiner Noun

The cat ate the mouse

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

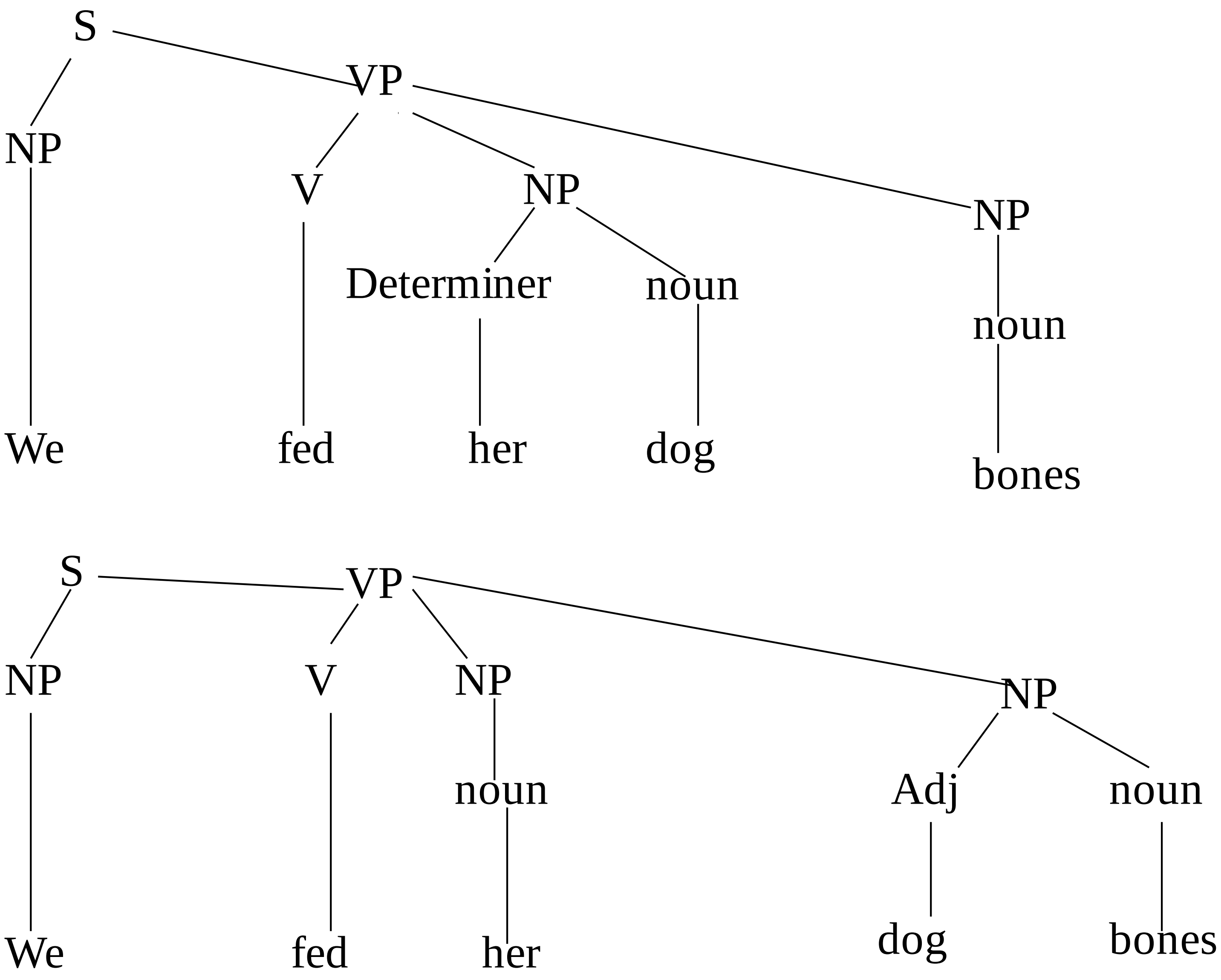
Noun = cat

Noun = mouse

Verb = ate

Parsing a sentence using simple phrase structure rules

Syntactic Ambiguity: We fed her dog bones



S
VP
NP
V
NP
NP
Determiner
noun
noun
We
fed
her
dog
bones

S
VP
NP
V
NP
NP
noun
Adj
noun
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 [$_{\text{payer}}$ Disney] offered to [$_{\text{V}}$ pay] [$_{\text{recipient}}$ Mr. Steinberg] [$_{\text{money}}$ a premium] for [$_{\text{commodity}}$ 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 similar set of queries with Siri. First, she correctly identified the current president. When I asked, "Who is his wife?" she said, "Hmmm.. Let me think", then she gave me the correct answer

Anaphora (Coreference) Resolution

The city councilors refused the demonstrators a permit because **they** feared violence.

The city councilors refused the demonstrators a permit because **they** advocated revolution.

Pragmatics

- The purposeful use of language in situations

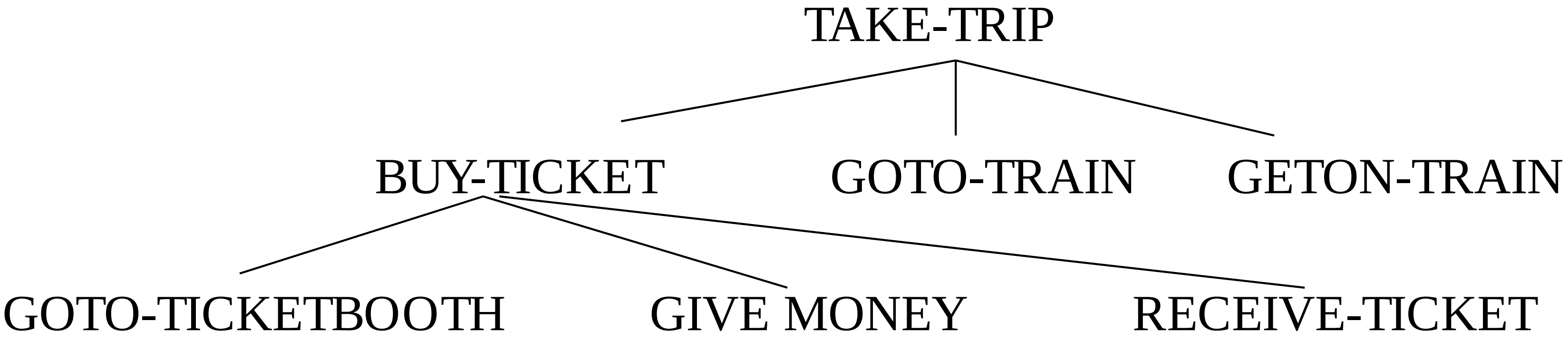
  - A functional perspective

- 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



TAKE-TRIP

BUY-TICKET

GOTO-TRAIN

GETON-TRAIN

GOTO-TICKETBOOTH

GIVE MONEY

RECEIVE-TICKET

Sketch of a commonsense task plan to take a trip

The “Non-level” 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 LINGUISTICS USING WORD FREQUENCIES

WHAT IS CORPUS LINGUISTICS?

- A methodology to process text and provide information about the text
- The Corpus is a collection of text
  - Utilizes a representative sample of machine-readable text of a language or a particular variety of text or language
  - Many contain linguistic annotations, such as POS tags, named entities, syntactic structures, semantic roles, etc.
- Statistical analysis
  - Word frequencies
  - Collocations
  - Concordances
- Often used in "Digital Humanities" as ways to characterize properties of corpora
  - Where the "properties" of interest may govern choices of words to highlight

Text Corpus Structure

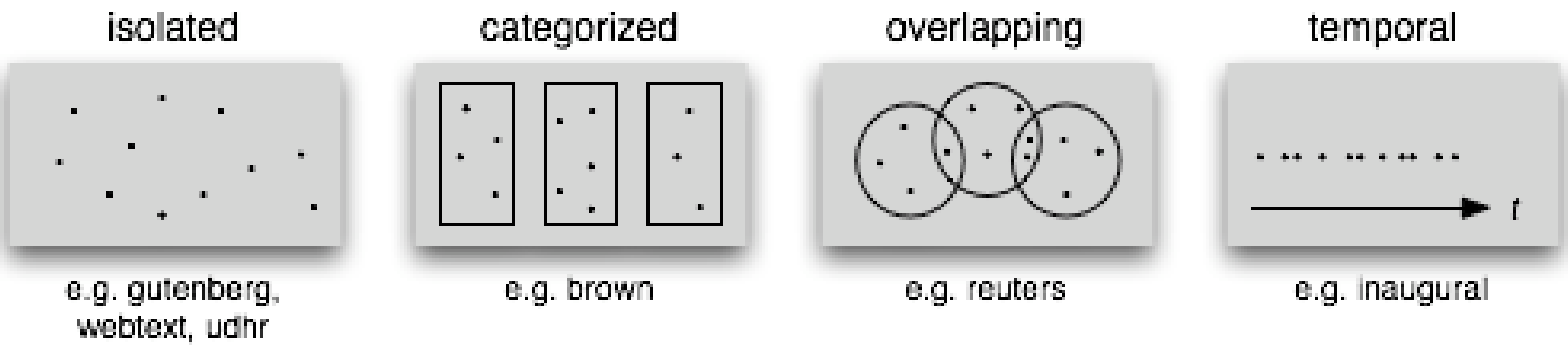


Image from: http://www.nltk.org/book/ch02.html

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 / lower case 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 Text Processing Required (Cont' d):

- Tokenization (or word segmentation):
  - Decide how to separate the characters in the sentence into individual words
    - Words are 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 how to recognize and deal 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: (Cont’d)

- 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

OCR (Optical Character Recognition): the recognition of printed or written text characters by a computer

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 – sometimes called vocabulary

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.

Note: we did not consider punctuations here. In NLTK, word_tokenize(text) function considers punctuations as well.

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, http://stateoftheunion.onetwothree.net/#
- Used for comparison and characterization of text
  - See the State of the Union (SOTU) Speeches by Nate Silver http://fivethirtyeight.com/features/obamas-sotu-clintonian-in-good-way/
  - Methodology: choose topic words of interest and plot frequencies of these words vs. different speeches

How Many Words In A Corpus?

- Let N be the number of tokens
- Let V be the size of the vocabulary (the number of distinct tokens)

Church and Gale (1990): $|V| > O(N^{½})$

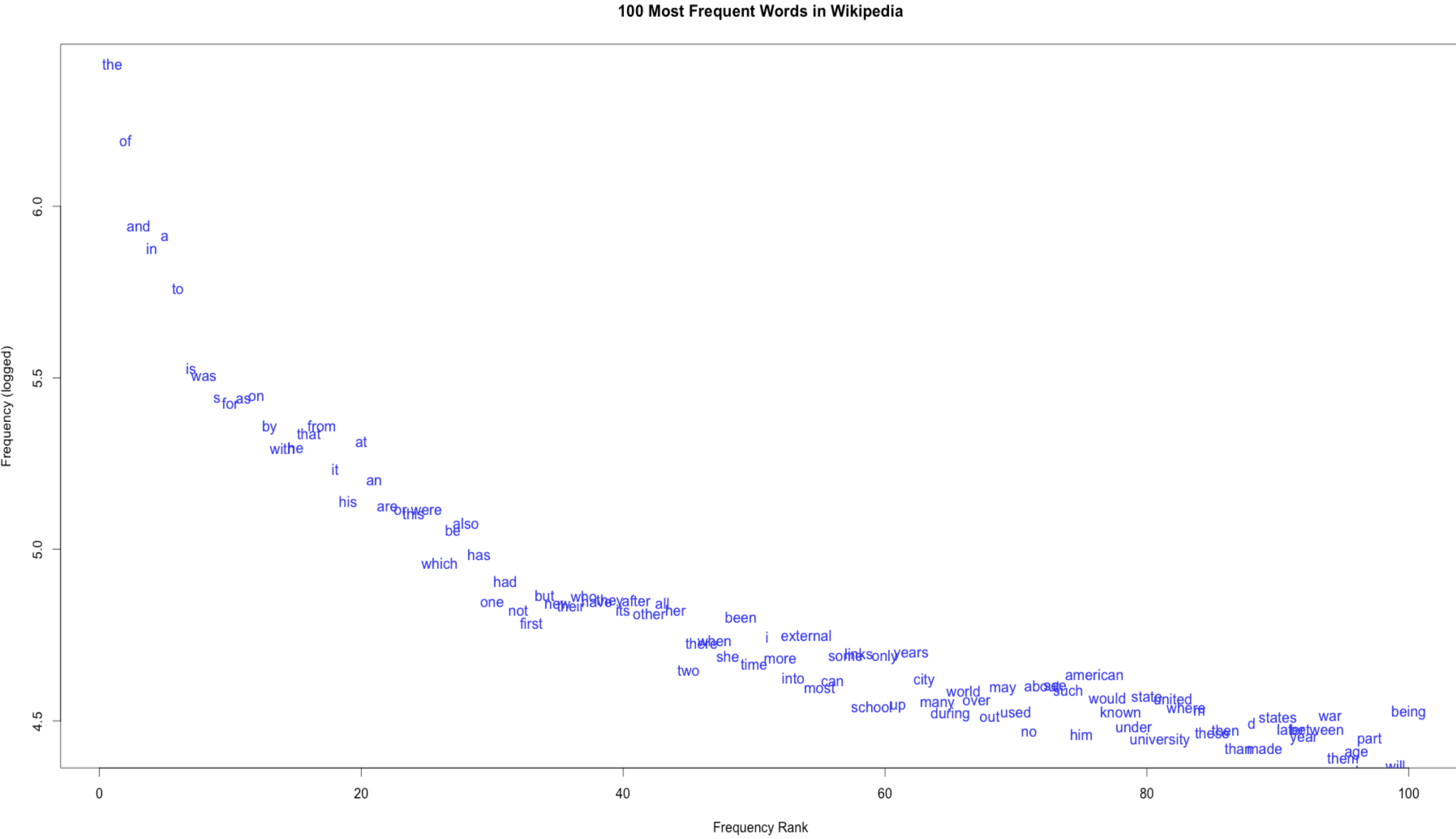
	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

from Dan Jurafsky

Zipf’s Law

- In a natural language corpus, the frequency of any word is inversely proportional to its rank in a frequency table
- Rank (r): The numerical position of a word in a list sorted by decreasing frequency (f).
- Zipf (1949) “discovered” that: $f \cdot r = k$ (for constant k)
  - Examples if k is 1:
    - Most frequent word (r = 1) is twice as frequent as 2nd most frequent
    - Most frequent (r = 1) is 3 times as frequent as 3rd most frequent, etc.

For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). ----- from Wikipedia



a sample of 36.8 million words from Wikipedia, over 580,000 word types, nearly half (280,000) occur just once in the sample. --- image and this data from http://wugology.com/zipfs-law/

Zipf' s Law (Cont'd)

- The highly frequent words tend to be short, grammatical (i.e., function words)

- The infrequent words tend to be longer, lexical (words like nouns and verbs which have some sort of referent or meaning).

Good News: Stopwords (commonly occurring words such as "the") will account for a large fraction of text so eliminating them greatly reduces the number of words in a text

Bad News: For most words, gathering sufficient data for meaningful statistical analysis is difficult since they are extremely rare.