Natural Language Processing
CS690

Lecture 01

Razvan C. Bunescu

School of Electrical Engineering and Computer Science

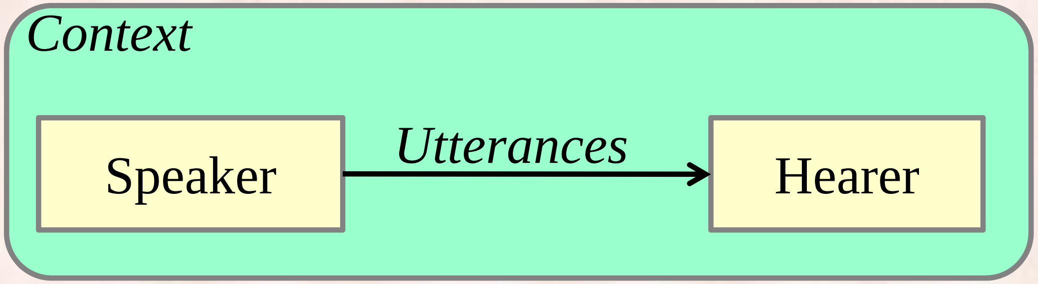
bunescu@ohio.edu

What is Natural Language Processing?

- **Natural Language Processing** = developing computer systems that can *process*, *understand*, or *communicate* in natural language:
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - Natural Languages are significantly more difficult to process than Artificial Languages!
- i.e. **Computational Linguistics**.

Communication

- **Communication** = intentional exchange of information through the production and perception of *signs* drawn from a shared system of conventional signs.
  - The main goal of generating and processing natural language.
  - In natural language, communication through *utterances*:
    - Speech
    - Writing
    - Facial expression
    - Gestures



Context

Speaker

Utterances

Hearer

Communication for the Speaker

- **Intention**:
  - Speaker decides that there is some proposition P worth saying to hearer H.
    - May require planning and reasoning about goals and beliefs.
- **Generation**:
  - Speaker transforms proposition P into an utterance, i.e. sequence of words W_1 in the desired natural language.
- **Synthesis**:
  - Speaker produces the words W_1 in the desired physical modality, e.g. text or speech, as T.

Communication for the Hearer

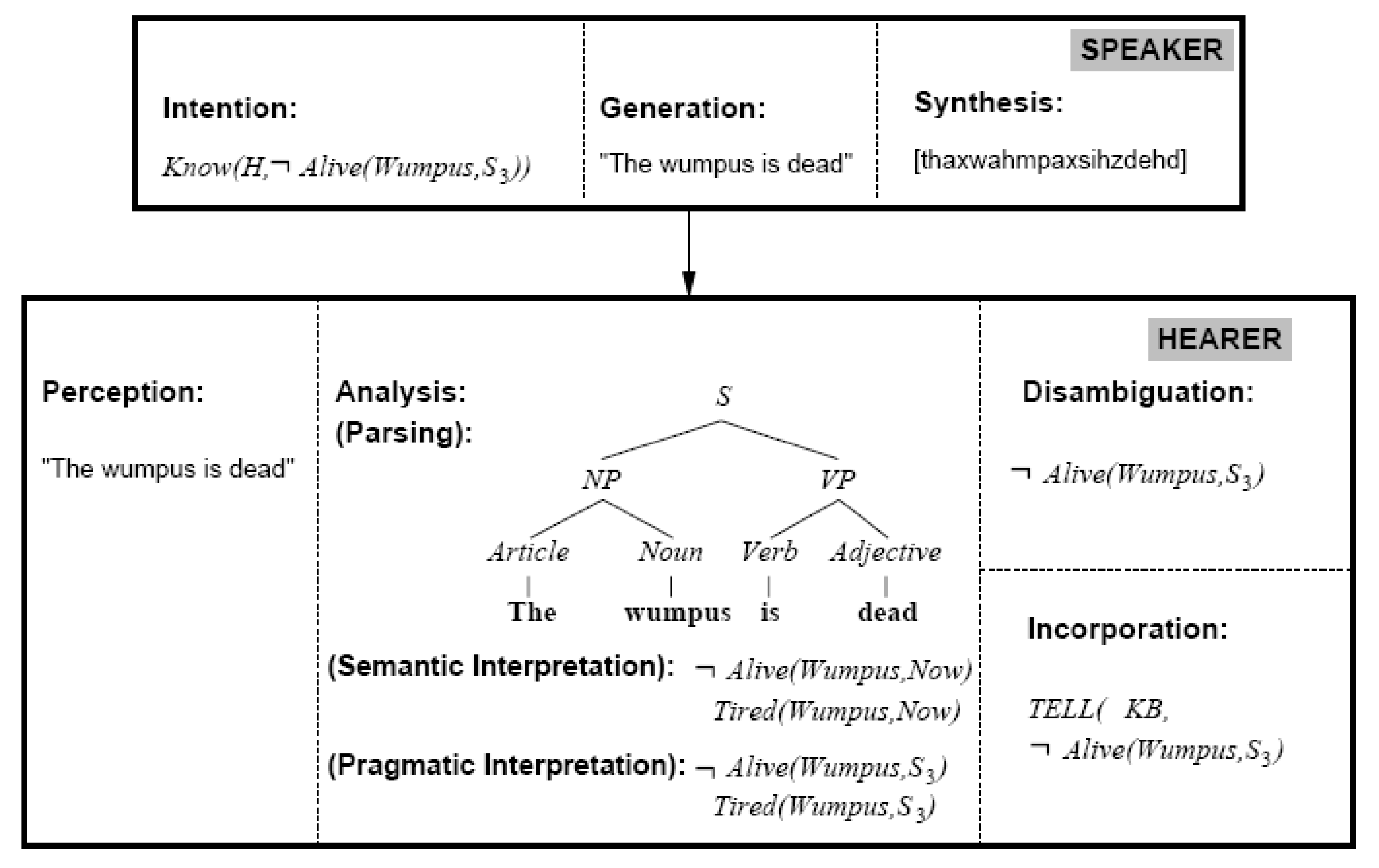
- **Perception**:
  - Hearer perceives physical realization and decodes it as the words W_2:
    - *speech recognition*, *optical character recognition*.
    - ideally $W_2 = W_1$.
- **Analysis**:
  - Hearer determines W_2 has possible meanings $P_1, P_2, \ldots, P_n$.
    - **Syntactic Interpretation**: find the *parse tree* showing the phrase structure of the word sequence.
    - **Semantic Interpretation**: find the meaning, e.g. *logical form*, of the word sequence.
    - **Pragmatic Interpretation**: consider effect of the overall *context* on altering the literal meaning of a sentence

Communication for the Hearer

- **Disambiguation**:
  - Hearer infers that Speaker intended to convey P_i.
  - Ideally $P_i = P$.

- **Incorporation**:
  - Hearer decides whether to believe P_i:
    - Incorporate P_i into Hearer's knowledge base KB.

Communication in the Wumpus World



SPEAKER

Intention:

$Know(H, \neg Alive(Wumpus, S_3))$

Generation:

"The wumpus is dead"

Synthesis:

[thaxwahmpaxsihzdehd]

HEARER

Perception:

"The wumpus is dead"

Analysis:
(Parsing):

(Semantic Interpretation): $\neg Alive(Wumpus, Now)$
$Tired(Wumpus, Now)$

(Pragmatic Interpretation): $\neg Alive(Wumpus, S_3)$
$Tired(Wumpus, S_3)$

Disambiguation:

$\neg Alive(Wumpus, S_3)$

Incorporation:

$TELL(\ KB,$
$\neg Alive(Wumpus, S_3))$

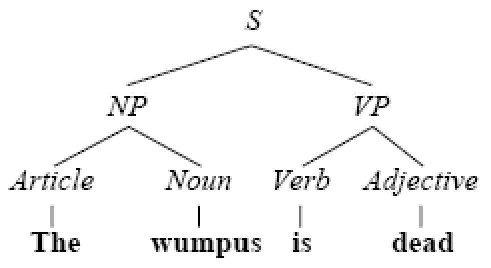
sound waves →

Phonetics

words → “The wumpus is dead”

Syntax

parse trees →



Semantics

logic forms → ¬Alive(Wumpus, Now)

Pragmatics

meaning in context → ¬Alive(Wumpus$_{101}$, Time$_{646}$)

What is an NLP Application?

- What makes an application an NLP application, as opposed to any other piece of software?
  - An application that requires the use of knowledge about human languages:

- Is Unix wc (word count) an example of a language processing application?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities.

Big NLP Applications

- These kinds of applications require a tremendous amount of knowledge of language:
  - Question answering.
  - Conversational agents.
  - Summarization.
  - Machine translation.

- Enabled by the solutions to more basic, fundamental NLP tasks.

Fundamental NLP Tasks in Text Analysis

- Tokenization
- Morphological Analysis
- Part of Speech Tagging
- Syntactic Parsing
- Word Sense Disambiguation
- Semantic Role Labeling
- Semantic Parsing
- Anaphora/Coreference Resolution

Tokenization

- **Tokenization** = segmenting text into words and sentences.
  - A crucial first step in most text processing applications.

- Whitespace indicative of word boundaries?
  - Yes: English, French, Spanish, …
  - No: Chinese, Japanese, Thai, …

- Whitespace is not enough:
  - ‘What’re you? Crazy?’ said Sadowsky. ‘I can’t afford to do that.’

  ⇒ *‘what’re you? crazy? Sadowsky. ‘I can’t that.*

Tokenization: Word Segmentation

- In English, characters other than whitespace can be used to separate words, e.g. , ; . - : ( )”

- But punctuation often occurs inside words:
  - m.p.h., Ph.D., AT&T, 01/02/06, google.com, 62.5

- Expansion of clitic constructions:
  - he’s happy ⇒ he is happy
  - Need ambiguity resolution between clitic construction, possessive markers, quotative markers:
    - he’s happy vs. the book’s cover vs. ‘what are you? crazy?’

Tokenization: Sentence Segmentation

- Generally based on punctuation marks: **? ! .**
  - Periods are ambiguous, as sentence boundary markers and abbreviation/acronym markers:
    - *Mr., Inc., m.p.h.*
  - Sometimes they mark both:
    - SAN FRANCISCO (MarketWatch) – Technology stocks were mostly in positive territory on Monday, powered by gains in shares of Microsoft Corp. and **IBM Corp.**

- Tokenization approaches:
  - Regular Expressions.
  - Machine Learning (state of the art).

Morphological Analysis

- **Morphology** = the field of linguistics that studies the internal structure of words.
  - **Morpheme** is the smallest linguistic unit that has semantic meaning:
    - **stems**: “carry”, “depend”, “Google”, “lock”
    - **affixes**: “pre”, “ed”, “ly”, “s”
- **Morphological analysis** = segmenting words into morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ? (un + lock) + able ?

Morphological Analysis: Stemming

- In **IR** applications such as **Web search**, only need to know if two words have the same stem:
  - Boolean Query: “marsupial OR kangaroo OR koala”.
  - Document contains: “marsupials”

  ⇒ **stemming**, i.e. given a word, extract the stem:
    - marsupials => marsupial
    - played, playing, player, plays => play
- **Porter stemmer** – a series of simple cascaded rewrite rules:
  - ATIONAL => ATE (e.g. relational => relate)
  - ING => ε (e.g. motoring => motor)
  - SSES => SS (e.g. grasses => grass)

Part of Speech (POS) Tagging

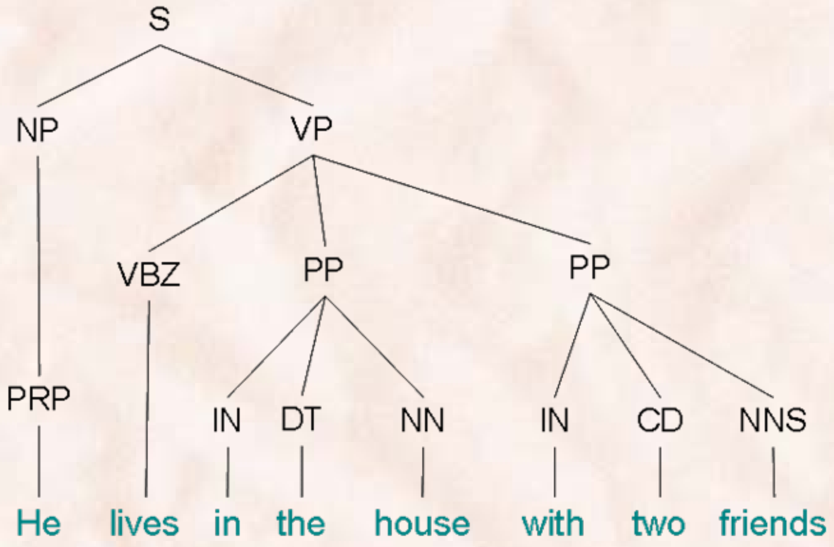
- Annotate each word in a sentence with its POS:
  - nouns, verbs, adjectives, adverbs, pronouns, prepositions, …

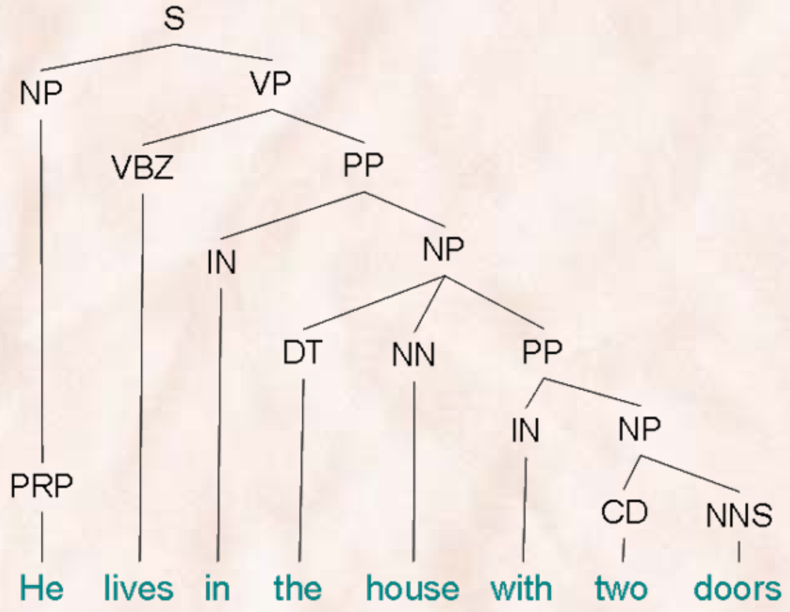
PRP	VBD	TO	VB	TO	DT	NN	IN	NN	VBD	VBG
They	used	to	object	to	the	use	of	object	oriented	programming
			obJECT					OBject		

- Useful for many other NLP tasks:
  - speech recognition and synthesis
  - syntactic parsing
  - word sense disamgiguation
  - information retrieval, …

Syntactic Parsing

- Output the correct *phrase structure (parse tree)* of a sentence.





Word Sense Disambiguation

- Words in natural language may have multiple meanings:
  - he cashed a check at the bank
  - he sat on the bank of the river and watched the currents
  - they built a large plant to manufacture automobiles
  - chlorophyll is generally present in plant leaves

- Identifying the meaning of a word is useful for:
  - machine translation
  - information retrieval
  - question answering
  - text classification

Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb:

  agent   patient   source   destination   instrument

  - John drove Mary from Athens to Columbus in his Toyota Prius.
  - The hammer broke the window.

- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing".

Semantic Parsing

- Map natural language sentences to a formal semantic representation (*logic form*).
- In GeoQuery, map sentences to Prolog queries:
  - *How many states does the Mississippi run through?*
  - answer(A, count(B, (state(B), const(C, riverid(mississippi)), traverse(C, B)), A))
- In RoboCup, map coaching advice to Clang:
  - *If the ball is in our penalty area, all our players except player 4 should stay in our half.*
  - ((bpos (penalty-area our)) (do (player-except our {4}) (pos (half our))))

Coreference Resolution

- Determine which noun phrases refer to the same discourse entity.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

Big NLP Applications

- These kinds of applications require a tremendous amount of knowledge of language:
  - **Question answering**.
  - Conversational agents.
  - Summarization.
  - Machine translation.

- Enabled by the solutions to more basic, fundamental NLP tasks.

Web Question Answering

- Web queries:
  - “Which *companies* were bought by *Google*.”
  - “What *proteins* interact with *cyclin D1*?”
  - “List the past presidents of the *Harvard Law Review*?
- Need automated information extraction to locate *companies*, *people*, and *proteins* in documents and identify *relationships* between them.
  - Named Entity Recognition
  - Relation Extraction

Sample Sentences from the Web

Search engine giant Google has bought video-sharing website YouTube in a controversial $1.6 billion deal.

The companies will merge Google's search expertise with YouTube's video expertise, pushing what executives believe is a hot emerging market of video offered over the Internet.

Drug giant Pfizer Inc. has reached an agreement to buy the private biotechnology firm Rinat Neuroscience Corp., the companies announced Thursday.

He has also received consulting fees from Alpharma, Eli Lilly and Company, Pfizer, and Rinat Neuroscience,

Named Entity Recognition

Company Names

Search engine giant **Google** has bought video-sharing website **YouTube** in a controversial $1.6 billion deal.

The companies will merge **Google**'s search expertise with **YouTube**'s video expertise, pushing what executives believe is a hot emerging market of video offered over the Internet.

Drug giant **Pfizer Inc.** has reached an agreement to buy the private biotechnology firm **Rinat Neuroscience Corp.**, the companies announced Thursday.

He has also received consulting fees from **Alpharma**, **Eli Lilly and Company**, **Pfizer**, and **Rinat Neuroscience**,

Relation Extraction

Company Acquisitions

Search engine giant **Google** has bought video-sharing website **YouTube** in a controversial $1.6 billion deal.

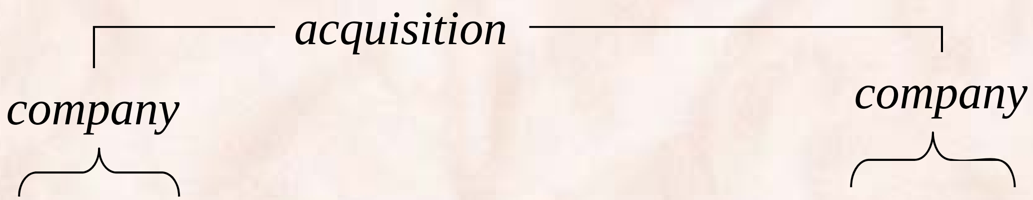
The companies will merge **Google**'s search expertise with **YouTube**'s video expertise, pushing what executives believe is a hot emerging market of video offered over the Internet.

Drug giant **Pfizer Inc.** has reached an agreement to buy the private biotechnology firm **Rinat Neuroscience Corp.**, the companies announced Thursday.

He has also received consulting fees from **Alpharma**, **Eli Lilly and Company**, **Pfizer**, and **Rinat Neuroscience**,

Relation Extraction (RE)

- Task: extract relations only between entities mentioned in the same sentence.
- Input: text with relevant named entities already tagged.

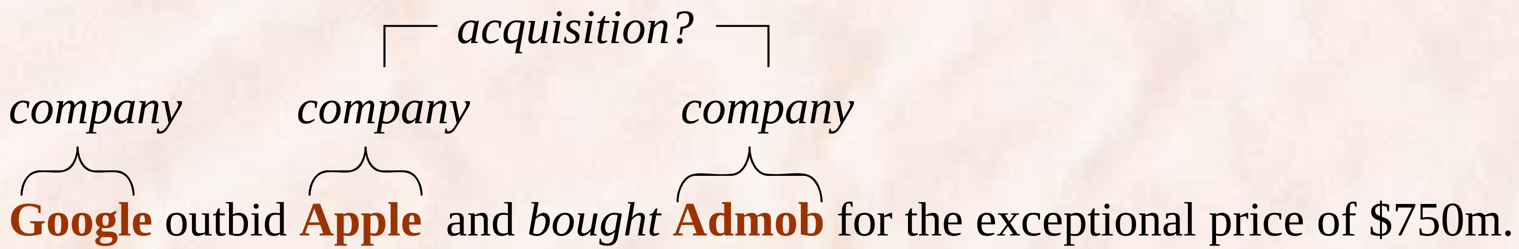


Search engine giant **Google** has *bought* video-sharing website **YouTube** in a controversial $1.6 billion deal.

- Relevant extraction pattern:
  - $\langle \mathrm{C}_1 \rangle$ … *bought* … $\langle \mathrm{C}_2 \rangle$

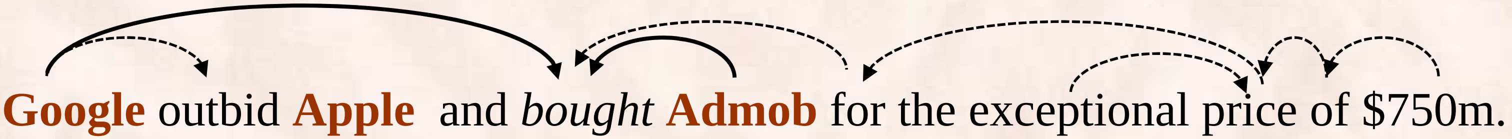
When Word Patterns Fail

- In many instances, rules based on word patterns extract the wrong pairs:



Google outbid **Apple** and *bought* **Admob** for the exceptional price of $750m.

- Need syntactic/dependency parsing.



Google outbid **Apple** and *bought* **Admob** for the exceptional price of $750m.

⇒ dependency patterns: ⟨C_1⟩ … *bought* … ⟨C_2⟩

When Patterns are Insufficienct

- Many sentences use *anaphoric* phrases that refer back to a previously introduced entity:

Obama is a graduate of **Columbia University** and **Harvard Law School**, where *he* was the president of the **Harvard Law Review**.

  - Q: Who was the president of the Harvard Law Review?
  - A: he ???

- Need coreference resolution.

The Curse of Ambiguity

- Computational Linguists are obsessed by ambiguity in NL:
  - unlike compiler writers.
- Ambiguity happens at all basic levels of natural language processing.
- Find at least 5 meanings of the following sentence:
  - I made her duck.

Ambiguity: “I made her duck”

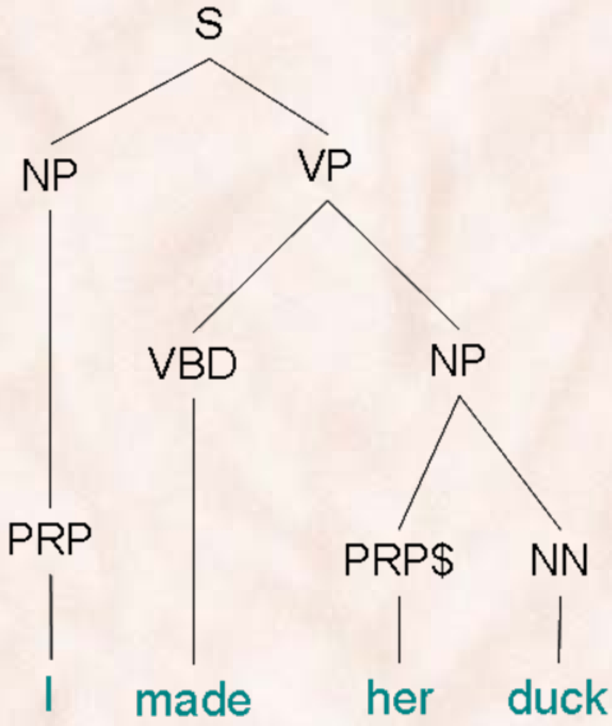
1) I cooked waterfowl for her benefit (to eat).
2) I cooked waterfowl belonging to her.
3) I created the (plaster?) duck she owns.
4) I caused her to quickly lower her head or body.
5) I waved my magic wand and turned her into undifferentiated waterfowl.

Ambiguity: “I made her duck”

- **POS tagging**: “duck” can be a N or V:
  - V: I caused her to quickly lower her head or body
  - N: I cooked waterfowl for her benefit (to eat).
- **POS tagging:** “her” can be a *possessive* (“of her”) or *dative* (“for her”) or *accusative* pronoun:
  - Possessive: I cooked waterfowl belonging to her.
  - Dative: I cooked waterfowl for her benefit (to eat).
  - Accusative: I waved my magic wand and turned her into waterfowl.
- **WSD:** “make” can mean “create” or “cook”:
  - Create: I made the (plaster) duck statue she owns
  - Cook: I cooked waterfowl belonging to her.

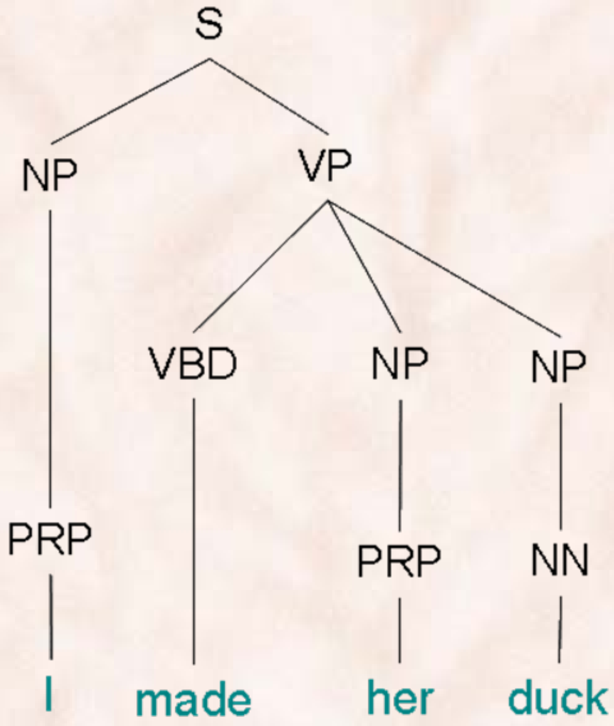
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Transitive (verb has a noun direct object):**
    - I cooked [waterfowl belonging to her]



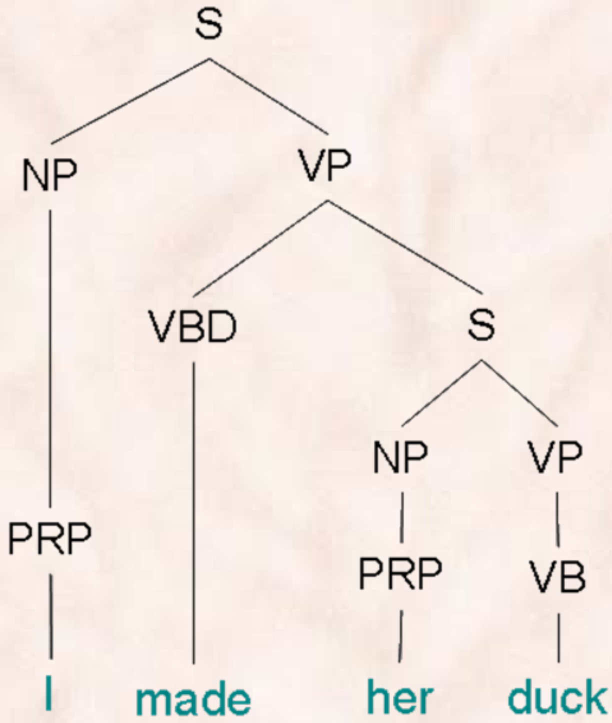
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Ditransitive (verb has 2 noun objects):**
    - I made [her] (into) [undifferentiated waterfowl]



Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Action-transitive:**
    - I caused [her] [to move her body]



Ambiguity: “I made her duck”

- **Speech Recognition**:
  - I mate or duck
  - I’m eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I’m aid her duck
  - I mate her duck
  - I’m ate her duck
  - I’m ate or duck

Ambiguity and Machine Translation

- English ⇒ Italian:
  - Mary **plays** the piano ⇒ Maria **suona** il pianoforte.
  - Mary **plays** with her cat ⇒ Maria **gioca** con il suo gatto.

- "Lost in translation" jokes from supposedly early MT system output (English ⇒ Russian ⇒ Italian):
  - "The spirit is willing, but the flesh is weak".

    ⇒ The vodka is good, but the meat is spoiled.
  - "Out of sight, out of mind".

    ⇒ Invisible idiot.

Modality and Ambiguity: What does Nancy want?

- **“Nancy wants to mary an analytic philosopher”**

  [Eco, “Kant and the Platypus”, 2000]

- Semantic interpretations:
  - [*de re*]: Nancy wants to marry a determined individual X, who is an analytic philosopher. $\exists x \Box Ax$
  - [*de dicto*]: Nancy wants to marry anybody, as long as he is an analytic philosopher. $\Box \exists x Ax$

- Pragmatic Interpretations (speaker’s intentions):
  - Nancy wants to marry a determined individual, an analytic philosopher: she knows who he is, but the speaker doesn’t, because she hasn’t told him the name.
  - Nancy wants to marry a determined individual X, an analytic philosopher: she has also given the speaker the name and introduced them to each other, but out of discretion the speaker has thought it more fitting to avoid going into details.
  - …

Ambiguity is Pervasive in Natural Language

- Computational Linguists are obsessed with ambiguity:
  - unlike compiler writers.
- Ambiguity happens at all basic levels of language processing.
- [Pros] Allows for significant compression of utterances:
  - people use context and knowledge about the world to disambiguate.
- [Cons] Very challenging for NLP.

Knowledge Involved in Resolving Ambiguity

- Syntax:
  - An agent is typically the subject of the verb (SRL).
- Semantics:
  - John and Mary are names of people.
  - Columbus and Athens are city names.
- Pragmatics:
  - If she is hungry and she is not vegetarian, it is likely she will enjoy cooked duck.
- Word knowledge:
  - Houses have a (variable number of) doors.
  - An individual may leave with other people (friends) in the same house.

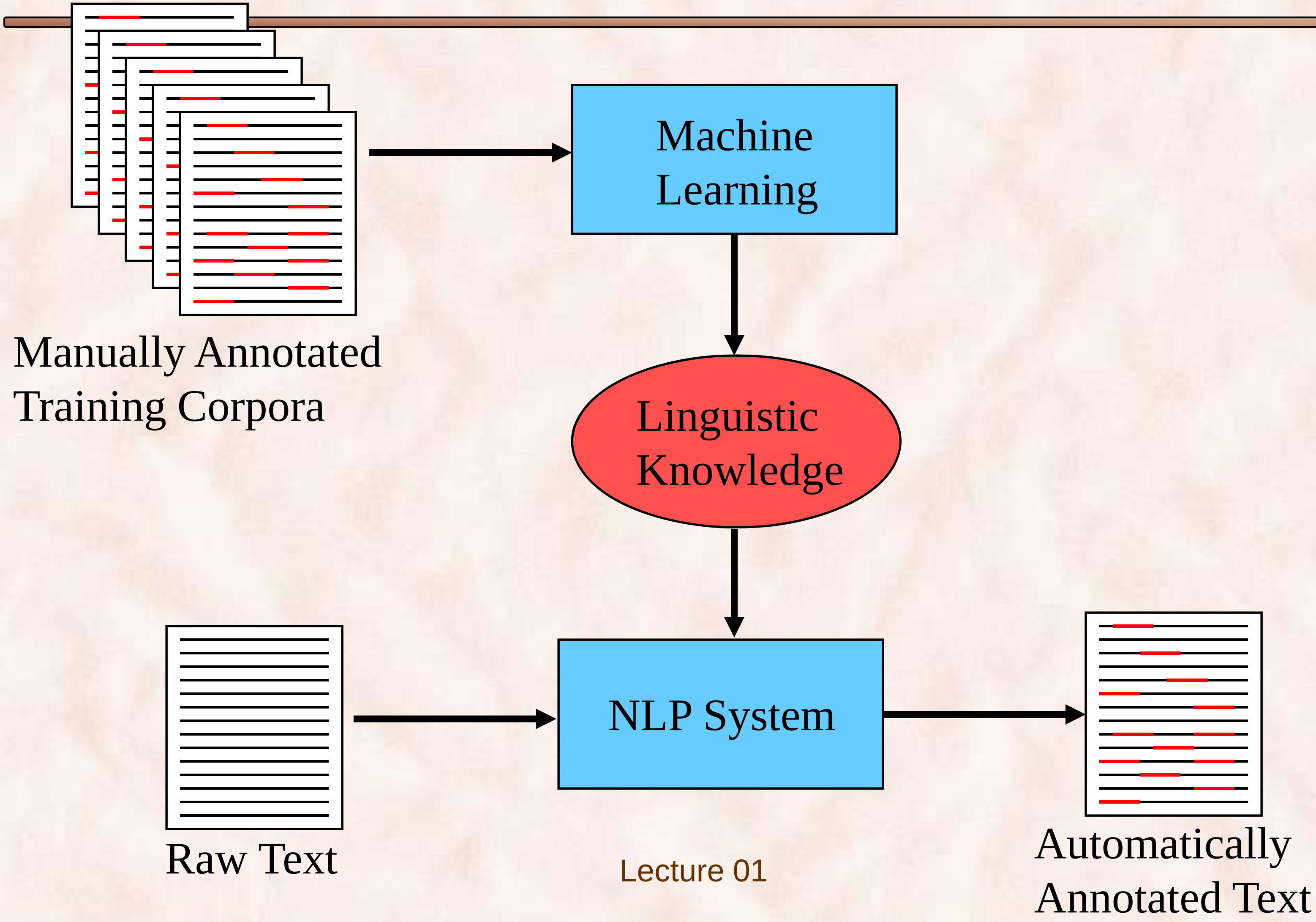
Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
  - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

Machine Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Variously referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

Machine Learning Approach


Machine Learning
Linguistic Knowledge
Manually Annotated Training Corpora
Raw Text
NLP System
Automatically Annotated Text

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - “Time flies like an arrow” has 4 parses, including those meaning:
    - Insects of a variety called “time flies” are fond of a particular arrow.
    - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.