NATURAL LANGUAGE PROCESSING (CS324)

Subject Scheme

L	T	P	C
3	1	0	4

References

- **Daniel Jurafsky, James H. Martin: "Speech and Language Processing", 2/E, Prentice Hall, 2008**.
- James Allen, "Natural Language Understanding", 2/E, Addison-Wesley, 1994
- Christopher D. Manning, Hinrich Schutze: "Foundations of Statistical Natural Language Processing", MIT Press, 1999
- Steven Bird, Natural Language Processing with Python, 1st Edition, O'Reilly, 2009.
- Jacob Perkins, Python Text Processing with NLTK 2.0 Cookbook, Packt Publishing, 2010.

Chapter 1 INTRODUCTION

Introduction

- Language is meant for **communicating about the world.**
- By studying language, we can come to **understand more about the world**. We can test out theories about world by **how well they support our attempt to understand language.**
- If we can succeed at building a computational model of language ,we will have a **powerful tool for communicating about the world.**

Introduction

- It is useful to divide the entire language processing problem into two tasks:
  - Processing **written text, using lexical ,syntactic ,and semantic knowledge of the language as well as the required real world information**.
  - Processing **spoken language, using all the information needed above plus additional knowledge about phonology** as well as enough added information to handle the further ambiguities that arise in speech .

Introduction

- English sentences are incomplete description of the information that they are intended to convey:
- Some dogs are outside →
  - Some dogs are on the lawn
  - Three dogs are on the lawn
  - Rover, Tripp, and spot are on the lawn
- Good Side : language allows speakers to be as precise as they like. It also allows speakers to leave out things they believe their hearers already know.

Introduction

- The same expression means different things in different contexts :
  - Where's the water ? (in a chemistry lab, it must be pure)
  - Where's the water ? (when you are thirsty, it must be potable)
  - Where's the water ? (dealing with a leaky roof)
- Good side : language lets us communicate about an infinite world using a finite number of symbol.

Introduction

- There are lots of way to say same thing :
  - Mary was born on October 11.
  - Mary's birthday is October 11.

- Good side when you know a lot. facts imply each other. Language is intended to be used by agent who know a lot .

Example

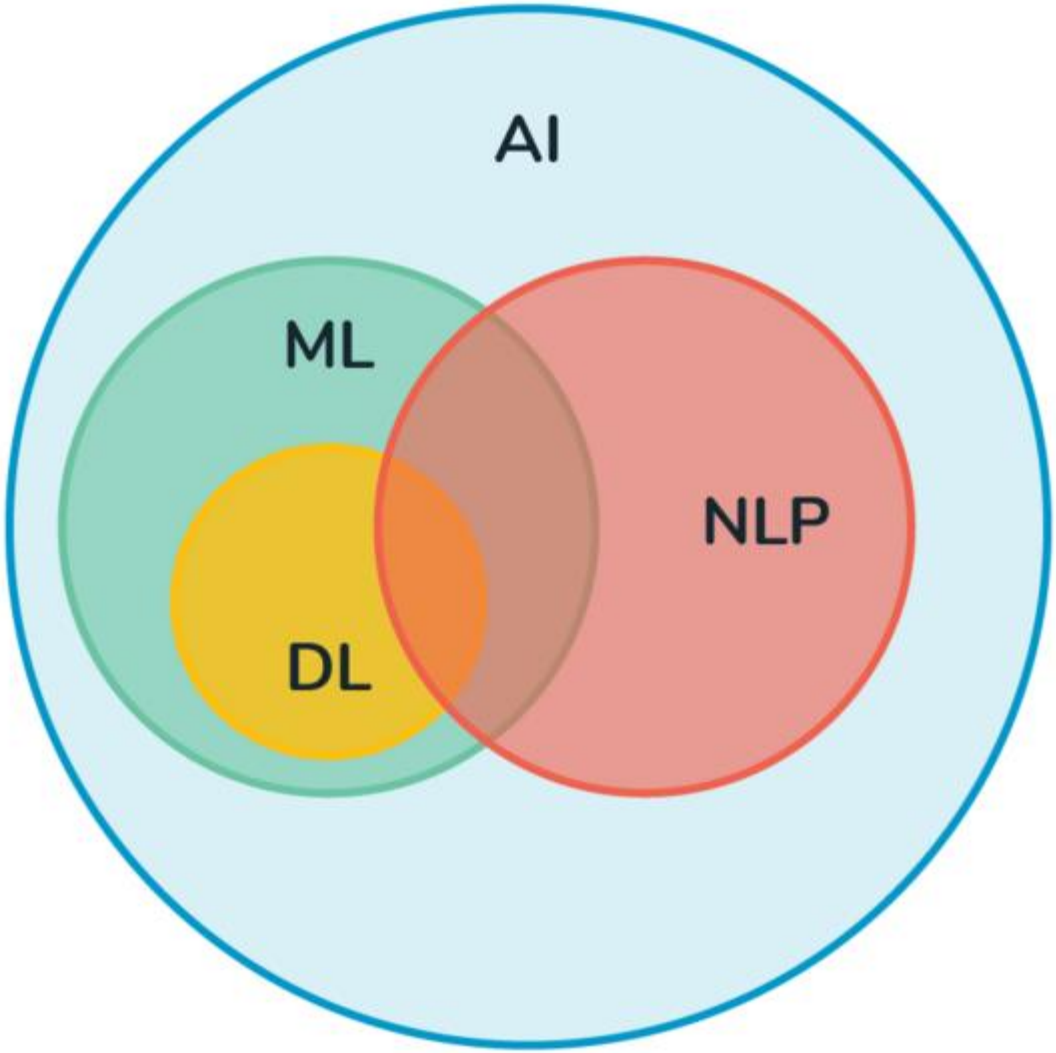
- 1968 film *2001: A Space Odyssey*, HAL (**H**euristically Programmed **Algorithmic** Computer)
- An intelligent computer
  - controls the systems of the spacecraft and interacts with the ship's crew

NLP

- Computer science and artificial intelligence which deals with human languages linguistics, computer science, information engineering, and artificial intelligence
- Processing of natural languages like English, Chinese etc. by computers to:
- Interact with people,
  - Follow natural language commands
  - Answer natural language questions
  - Provide information in natural language

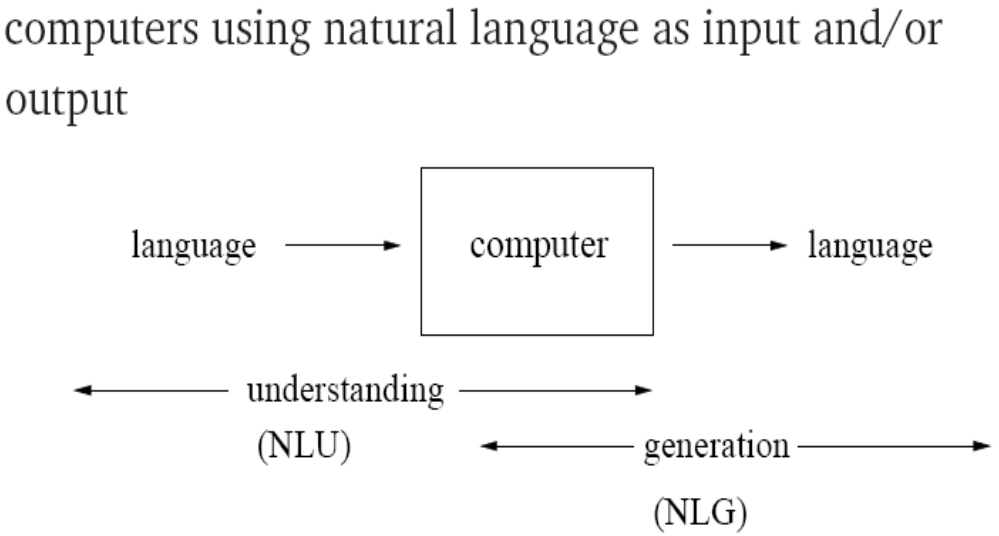
- Perform useful tasks,
  - Find required information from several documents
  - Summarize large or many documents
  - Translate from one natural language to another

NLP



NLP

Natural language processing (NLP) is a collection of techniques used to **extract grammatical structure and meaning from input in order to perform a useful task as a result, natural language generation builds output based on the rules** of the target language and the task at hand.



NLP

- NLU
  - Natural Language Understanding
  - understanding as the process of mapping from an input form into a more immediately useful form.
  - All truths are easy to understand once they red, the point is to discover them.
- NLG
  - Natural Language Generation
  - NLG makes data universally understandable and seeks to automate the writing of data-driven narratives like financial reports, product descriptions, meeting memos, and more.

NLP

- Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English

Applications

- Automatic summarization
- Foreign language reading aid
- Foreign language writing aid
- Information extraction
- Information retrieval (IR) - IR is concerned with storing, searching and retrieving information. It is a separate field within computer science (closer to databases), but IR relies on some NLP methods (for example, stemming). Some current research and applications seek to bridge the gap between IR and NLP.

Applications

- Machine translation - Automatically translating from one human language to another.
- Named entity recognition (NER) - Given a stream of text, determining which items in the text map to proper names, such as people or places. Although in English, named entities are marked with capitalized words, many other languages do not use capitalization to distinguish named entities.
- Natural language generation
- Natural language search
- Natural language understanding

Applications

- Optical character recognition
- anaphora resolution:
- Anaphora meaning: the use of a word referring back to a word used earlier in a text or conversation, to avoid repetition, for example the pronouns he, she, it, and they and the verb do in I like it and so do they.
- Query expansion
- Question answering - Given a human language question, the task of producing a human-language answer. The question may be a closed-ended (such as "What is the capital of Canada?") or open-ended (such as "What is the meaning of life?").
- Speech recognition - Given a sound clip of a person or people speaking, the task of producing a text dictation of the speaker(s). (The opposite of text to speech.)
- Spoken dialogue system
- Text simplification
- Text-to-speech

Knowledge of Language

- **Phonetics & Phonology** –
  - concerns how words are related to the sounds that realize them.
  - Analyzing and generating sequence of words or incoming signals
  - how words are pronounced in colloquial speech
- **Morphology**
  - Captures information about shape and behavior of words in context
- **Syntax**
  - concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- **Semantics** –
  - concerns what words mean and how these meaning combine in sentences to form sentence meaning.

Knowledge of Language

- **Pragmatics** –
  - concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

- **Discourse**
  - concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns

- **World Knowledge**
  - includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

Knowledge of Language

- Difference of **Language processing applications** from **other data processing systems** is their use of ***knowledge of language.***
- Eg:
  - wc program, which is used to count the total number of bytes, words, and lines in a text file.
    - When used to **count bytes and lines, wc is an ordinary data processing application.**
    - However, when it is used to **count the words in a file it requires *knowledge about what it means to be a word***, *and thus* becomes a **language processing system**.

Knowledge of Language

Six Categories:

- **Phonetics and Phonology** – The study of linguistic sounds.
- **Morphology** – The study of the meaningful components of words.
  - **Eg:** Recognizing "doors" as plural
  - Capable of producing contractions like *I'm and can't.*
- **Syntax** – The study of the structural relationships between words.
  - **Eg:** Basic English noun phrase consists of an **optional determiner**, some number of adjectives, and then a noun, do capture major patterns within the language.
- **Semantics** – The study of meaning.

- **Pragmatics** – The study of how language is used to accomplish goals.
  - **Eg:** embellishes the responses with the phrases *I'm sorry* and *I'm afraid.*
- **Discourse** – The study of linguistic units larger than a single utterance.
  - **Eg: correct structuring of conversation (eg:** using **that,** etc.)

Knowledge of Language

Six Categories:

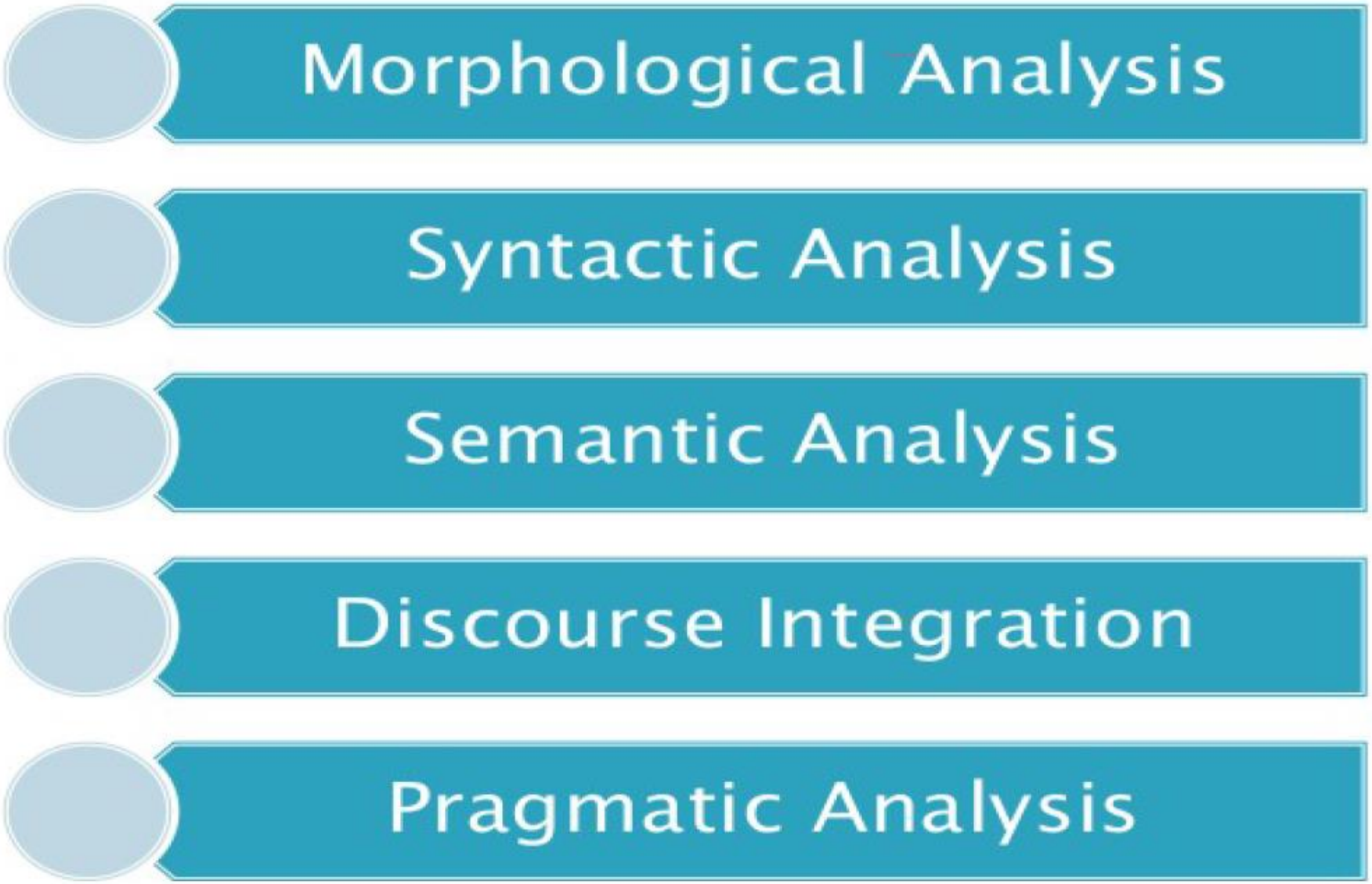
What kinds of things do people say?

- **Phonetics and Phonology** – The study of linguistic sounds.
- **Morphology** – The study of the meaningful components of words.
  - **Eg:** Recognizing "doors" as plural
  - Capable of producing contractions like *I'm and can't.*
- **Syntax** – The study of the structural relationships between words.
  - **Eg:** Basic English noun phrase consists of an **optional determiner**, some number of adjectives, and then a noun, do capture major patterns within the language.
- **Semantics** – The study of meaning.

What do these things say/ask/request about the world?

- **Pragmatics** – The study of how language is used to accomplish goals.
  - **Eg:** embellishes the responses with the phrases *I'm sorry* and *I'm afraid.*
- **Discourse** – The study of linguistic units larger than a single utterance.
  - **Eg: correct structuring of conversation (eg:** using **that,** etc.)

Steps


Morphological Analysis
Syntactic Analysis
Semantic Analysis
Discourse Integration
Pragmatic Analysis

Steps

- **Morphological Analysis:**
  - **Individual words are analyzed** into their components.
  - Non-word tokens, such as punctuation(.), are separated from the words
- **Syntactic Analysis:**
  - **Linear sequences of words are** transformed into structures that show how the words **relate to each other.**
  - Some word sequences may be rejected if they violate the language's rules for how words may be combined.
  - Example , an English syntactic analyzer would reject the sentence "boy the go the to store

Steps

- **Semantic Analysis :**
  - **The structures created by** the syntactic analyzer are assigned meanings.
  - A mapping is **made between the syntactic structures and objects in the task domain.**
  - Structures for which no such mapping is possible may be rejected .
  - Example : in most universes , the sentence "**Colorless green ideas sleep furiously"** would be rejected as semantically anomalous

Steps

- **Discourse integration :**
  - **The meaning of an** individual sentence may depend on the sentences the **precede it and may influence the meanings of the sentences that follow it.**
  - Example : the word **“it”** in the sentences, **“john wanted it,”** depends on the prior discourse context ,while the word “john” may influence the meaning of later sentences (such as “He always had”)

Steps

- **Pragmatic Analysis :**
  - **The structure** representing what was said is reinterpreted to **determine what was actually meant.**
  - Example : the sentence “Do you know what time it is ?” should be **interpreted as a request to be told the time.**
- The boundaries between these phases are often **very fuzzy.**

Morphological Analysis

- Let's start with one example:
  - I want to print Bill's .init file.

Morphological Analysis

- It must do the following things :
  - Pull apart the word “bill’s” into the proper noun “bill” and possessive suffix “ ’s ”
- Recognize the sequence **“.init”** as a file extension that is functioning as an adjective in the sentence
- This Process will usually assign syntactic categories to all the words in the sentence .
- This is usually done now because interpretations for affixes may depend on the syntactic category of the complete word .

Morphological Analysis

- Example : consider the word “prints”.
  - This word is either a plural noun(with the “-s” marking plural) or a third person singular verb (as in “he prints”)
  - in which case the “-s” indicates both singular and third person.
  - If this step is done now, then in our example, there will be ambiguity since “want,” “print” and “file” can all function as more than one syntactic category.

Syntactic Analysis

- Syntactic analysis **must exploit the results of morphological analysis** to build a **structural description of the sentence.**
- The goal of this process called **parsing** , is to convert the **flat list of words that forms the sentence into a structure that defines the units that are represented by that flat list.**

Syntactic Analysis

- The first issue concerns the formalism that is used to specify **what sentences are possible in a language. This information is specified by a set of rules called a grammar**.

- The second issue concerns how to determine the structure of a given sentence once you know the grammar for the language. This **process is called parsing.**

Syntactic Analysis

- **The important thing here is that a flat sentence has been converted into a hierarchical structure** and that the structure correspond to meaning u**nits when semantic analysis is performed.**
- Reference markers are shown in the parenthesis in the parse tree
- Each one corresponds to **some entity** that has been mentioned in the sentence.
- These **reference markers(RM)** are useful later since they provide a place in which to accumulate information about the entities as we get it.

Syntactic Analysis

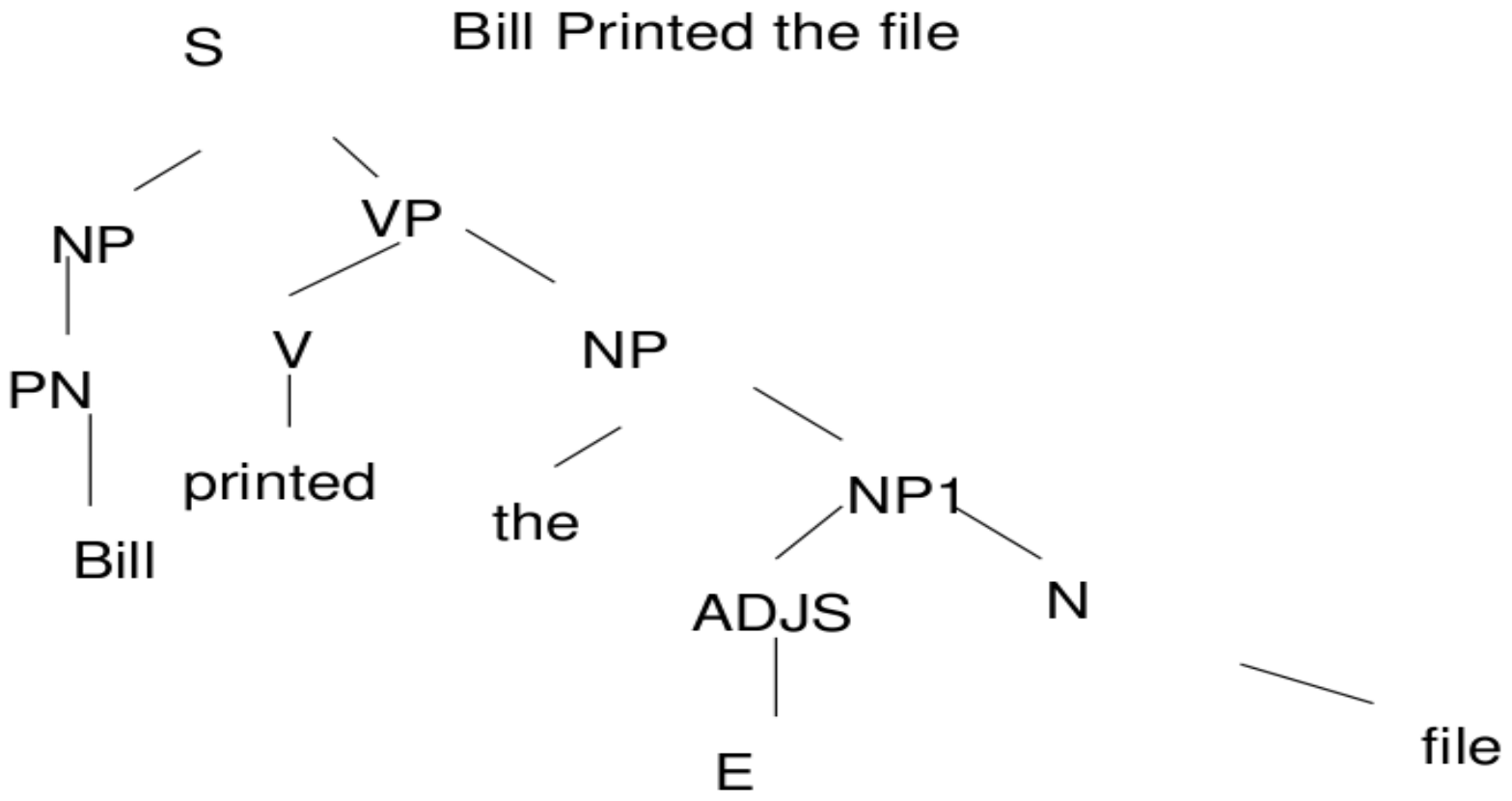
- Almost all the systems that are actually used have two main components:
  - A **declarative representation,** called a grammar, of the syntactic facts about the language.
  - A **procedure, called parser,** that compares the grammar against input sentences to produce parsed structures.

Syntactic Analysis

- The most common way to represent grammars is as a set of production rules.
- A simple Context-free phrase structure grammar fro English:
- S → NP VP
- NP → the NP1
- NP → PRO
- NP → PN
- NP → NP1
- NP1 → ADJS N
- ADJS → ε | ADJ ADJS
- VP → V
- VP → V NP
- **EG:**
- **N → file | printer**
- **PN → Bill**
- **PRO → I**
- **ADJ → short | long | fast**
- **V → printed | created | want**

Syntactic Analysis

A Parse tree for a sentence :



Syntactic Analysis

- It is the grammar that consists rules with a single symbol on the left-hand side of the rewrite rules. Let us create grammar to parse a sentence − “The bird pecks the grains”
- Articles (DET) − a | an | the
- Nouns − bird | birds | grain | grains
- Noun Phrase (NP) − Article + Noun | Article + Adjective + Noun
- = DET N | DET ADJ N
- Verbs − pecks | pecking | pecked
- Verb Phrase (VP) − NP V | V NP
- Adjectives (ADJ) − beautiful | small | chirping

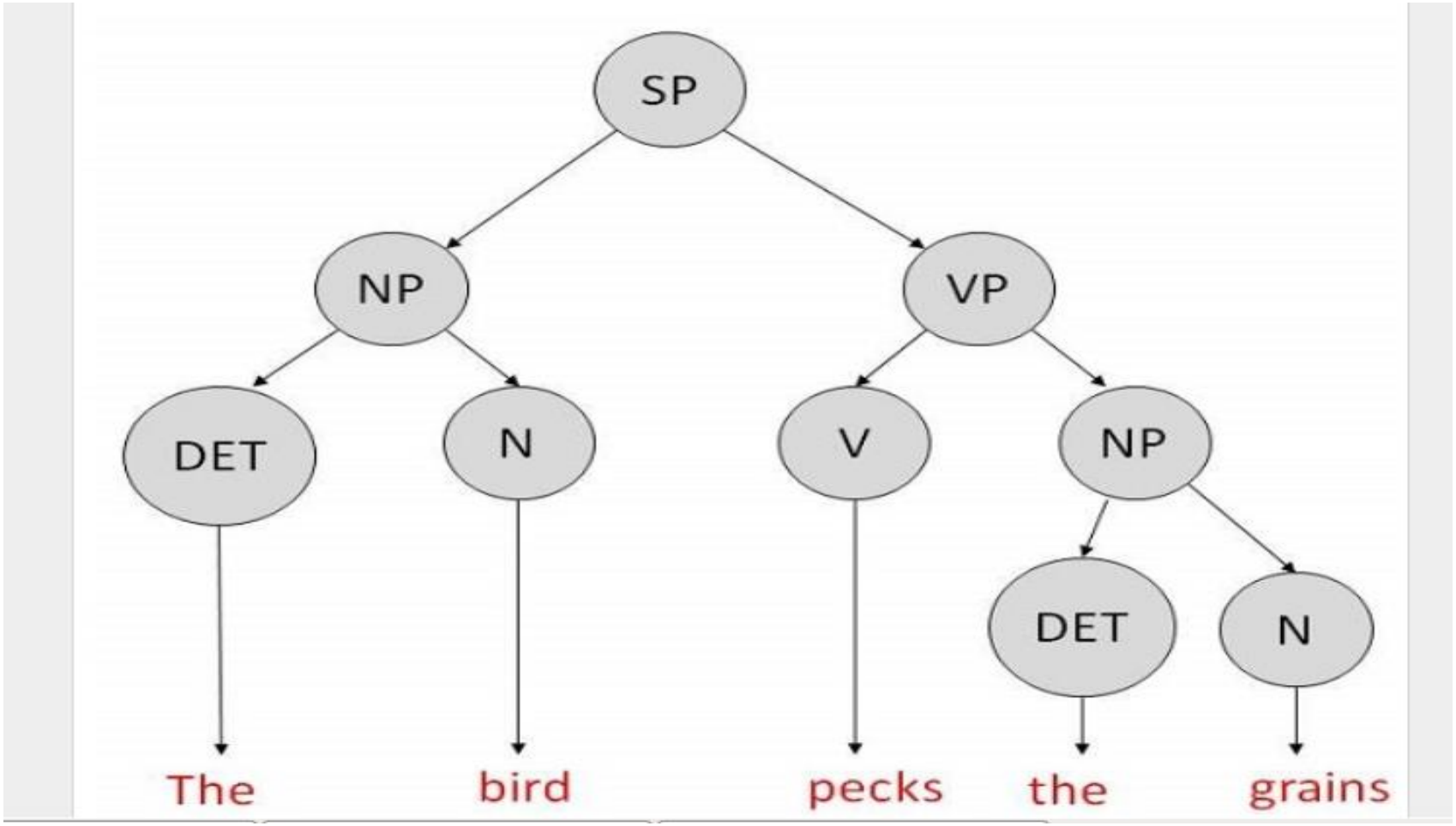
Syntactic Analysis

- The parse tree breaks **down the sentence into structured parts so that the computer can easily understand and process it.**
- In order for the **parsing algorithm to construct** this parse tree, a set of rewrite rules, which describe what tree structures are legal, need to be constructed.
- These **rules say that a certain symbol may be expanded in the tree by a sequence of other symbols**.
- According to first order logic rule, **if there are two strings Noun Phrase (NP) and Verb Phrase (VP), then the string combined by NP followed by VP is a sentence**. The rewrite rules for the sentence are as follows −

Syntactic Analysis

- S → NP VP
- NP → DET N | DET ADJ N
- VP → V NP
- DET → a | the
- ADJ → beautiful | perching
- N → bird | birds | grain | grains
- V → peck | pecks | pecking

Syntactic Analysis



Semantic Analysis

- Semantic analysis must do two important things:
- It must map **individual words into appropriate objects in the knowledge base or database**
- It must **create the correct structures** to **correspond to the way the meanings of the individual words combine with each other.**

Semantic Analysis

- **The first step in any semantic processing system is to look up the individual words in a dictionary and extract their meanings.**
- Many words have several meanings, and it may not be possible to choose the correct one just by looking at the word itself.
- **The process of determining the correct meaning of an individual word is called word sense disambiguation or lexical disambiguation.**
- It is done by associating, with each word in lexicon, information about the contexts in which each of the word's senses may appear.
- Sometimes only very straightforward info about each word sense is necessary. For example, **baseball field could be marked as a LOCATION.**
- Some useful semantic markers are :
  - PHYSICAL-OBJECT
  - ANIMATE-OBJECT

Semantic Analysis

- WordNet is a lexical database
- It groups words into sets of synonyms called synsets.
- Synset: a set of synonyms, representing one underlying lexical concept
- Eg:- dog {domestic dog, wild dog,Canis familiaris }

Semantic Analysis

Knowledge Representation for NLP

- **Which knowledge representation will be used depends on the application** -- Machine Translation, Database Query System.
- Requires the **choice of representational framework, as well as the specific meaning vocabulary** (what are concepts and relationship between these concepts -- ontology)
- Must be computationally effective.
- Common representational formalisms:
  - first order predicate logic
  - conceptual dependency graphs
  - semantic networks
  - **Frame-based representations (let's take example of frame based)**

Semantic Analysis

- Suppose that we have a frame-based knowledge base that contains the units shown in fig below (in next slide).

Semantic Analysis

User
isa : Person

User068
instance: user
login-name: Susan-black
User073
instance: user
login-name: bill_smith
F1
instance: File-struct
name: Stuff
Extension: .init
Owner: user073
in-directory: /wsmith/

Semantic Analysis

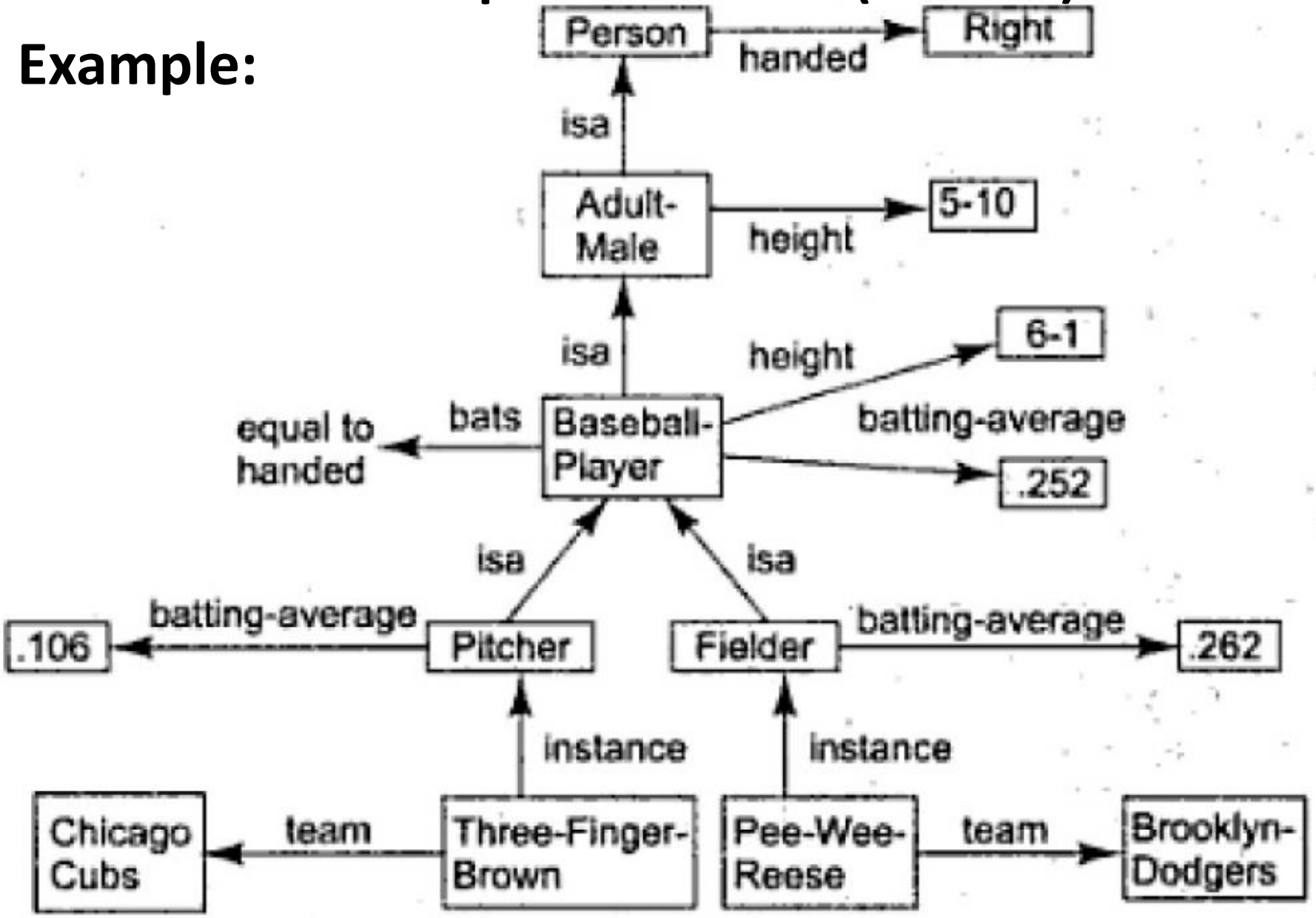
Frame-based representations (in brief):

- **Frames as Sets and Instances**

  – Each frame represents either a class or an instance (object).

- **Frame-based representations (in brief):**

Example:



- **Frame-based representations (in brief):**
  Example:

Person
- *isa* : *Mammal*
- *cardinality* : 6,000,000,000
- * *handed* : *Right*

Adult-Male
- *isa* : *Person*
- *cardinality* : 2,000,000,000
- * *height* : 5-10

ML-Baseball-Player
- *isa* : *Adult-Male*
- *cardinality* : 624
- **height* : 6-1
- * *bats* : equal to handed
- * *batting-average* : .252
- * *team* :
- * *uniform-color* :

Fielder
- *isa* : *ML-Baseball-Player*
- *cardinality* : 376
- **batting-average* : .262

Example(conti..):

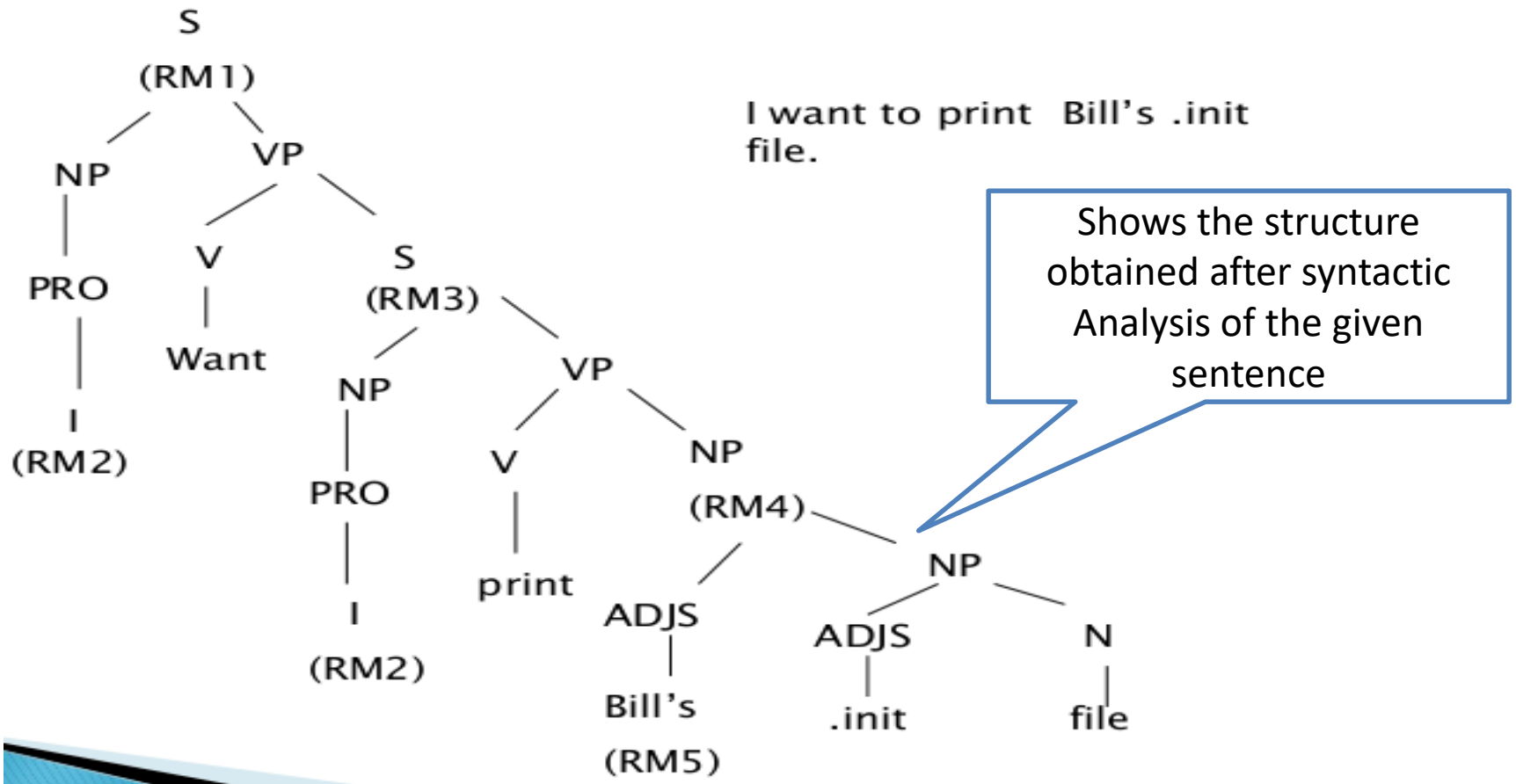
Pee-Wee-Reese	
instance :	*Fielder*
height :	5-10
bats :	*Right*
batting-average :	.309
team :	*Brooklyn-Dodge*
uniform-color :	*Blue*
ML-Baseball-Team	
isa:	*Team*
cardinality :	26
* *team-size* :	24
* *manager* :	
Brooklyn-Dodgers	
instance :	*ML-Baseball-Te*
team-size :	24
manager :	*Leo-Durocher*
players :	*{Pee-Wee-Rees*

Semantic Analysis

Frame-based representations (in brief):

- *Isa: subset relation*
  - For example: The set Adult male is a subset of the set of persons.
- *Instance: element-of relation*
  - For example: Pee-Wee-Reese is an element of the set of fielders.

Semantic Analysis



Semantic Analysis

- Then, we can generate a partial meaning, with respect to knowledge base, as shown in fig below (in next slide).

Semantic Analysis

RM1 **{the whole sentence}**
- instance : wanting
- Agent : RM2 {I}
- Object : RM3 {a printing event}

RM2 **{I}**

RM3 **{a printing event}**
- instance : Printing
- agent : RM2 {I}
- object: RM4 {Bill's .init file}

RM4
- instance : File -struct
- extension: .init
- owner: RM5 {bill}

RM5 **{bill}**
- instance: person
- first-name: bill

Discourse Integration

- Specifically we do not know whom the pronoun “I” or the proper noun “Bill” refers to.
- To pin down these references requires an appeal to a model of the **current discourse context,** from which we can learn that the **current user is USER068 and that the only person named “Bill” about whom we could be talking is USER073**.
- Once the correct referent for Bill is known, we can also determine exactly which file is being referred to.

Pragmatic Analysis

- The final step toward **effective understanding is to decide what to do as a results.**
- One possible thing **to do is to record what was said as a fact and be done with it.**
- For some sentences, whose intended effect is clearly **declarative**, that is **precisely correct thing to do.**
- But for other sentences, including this one, the intended effect is different.
- We can discover this **intended effect by applying a set of rules that characterize cooperative dialogues.**
- **The final step in pragmatic processing is to translate, from the knowledge based representation to a command to be executed by the system.**
- The results of the understanding process is
- Lpr/wsmith/stuff.init file where “lpr” is the operating system's file print command

Pragmatic Analysis

- Knowledge about the kind of actions that speakers intend by their use of sentences
- REQUEST: HAL, open the pod bay door.
- STATEMENT: HAL, the pod bay door isopen.
- INFORMATION QUESTION: HAL, is the pod bay door open?
- Speech act analysis (politeness, irony, greeting, apologizing...)

Stages/Steps in NLP

- Phonology and Phonetics (processing of sound)
- Morphology (processing of word forms)
  - Lexicon (Storage of words and associated knowledge)
- Syntax/ Parsing (Processing of structure)
- Semantics (Processing of meaning)
- Pragmatics (Processing of user intention, modeling etc.)
- Discourse (Processing of connected text)

Stages/Steps in NLP

- Phonology and Phonetics (processing of sound)
- Morphology (processing of word forms)
- Syntax/ Parsing (Processing of structure)
- Semantics (Processing of meaning)
  - Lexicon (Storage of words and associated knowledge)
- Pragmatics (Processing of user intention, modeling etc.)
- Discourse (Processing of connected text)

Ambiguity

- Speech and language processing can be viewed as resolving **ambiguity at all levels.**
- Some input is ambiguous
  - if there are **multiple alternative linguistic structures than can be built for it**

Ambiguity

- Phonology and Phonetics:
  - At this stage utterances are processed
  - Challenges here:
    - Noise
    - *Homophony*
    - *Word boundary recognition*

Ambiguity

- Phonology and Phonetics:
  - At this stage utterances are processed
  - Challenges here:
    - *Homophony*
      - **Homophony arises when two words sound the same**
      - **Though their meanings are widely different**
      - e.g., *bank* (embankment of a water body) and *bank (an institution where* financial transactions are held).

Ambiguity

- Phonology and Phonetics:
  - At this stage utterances are processed
  - Challenges here:
    - *Word boundary recognition*
      - Word boundary detection is a challenge in **case of rapid speech**
      - e.g**. *fox and folks.***
      - ***I got up late vs. I got a plate,*** both of which sound very much the same.
      - **aajaayenge (aaj aayenge or aa jaayenge)**

Ambiguity

- Morphology
  - Words form from **root words or lexemes (**A **lexeme** is a unit of lexical **meaning** that underlies a set of words that are related through inflection**)** through processes of: inflexion, derivation, etc.

Ambiguity

- Morphology
  - English is one example of relatively simpler morphology.
  - The main ambiguity at the level of morphology arises from choices available **in breaking the word into stem and suffix as well as from choices of features**

Ambiguity

- Lexicon
  - Words are stored in the **lexicon with a variety of information that facilitates** the **further stages of NLP,** like question answering, information extraction *etc.*
  - *For example, the word **dog** might be* stored in the lexicon with information like:
    - *POS (Noun)*
    - *Semantic Tag (Animate, 4-legged)*
    - *Morphology (takes 's' in plural)*
  - Words typically have **multiple meanings even in the same part of speech.**
  - ***Dog**, for example, means an animal and a very* detestable person.

Ambiguity

- lexicon
  - **Lexical Ambiguity**
  - A word is ambiguous w.r.t. its syntactic class .
  - **Eg:**
    - The word silver can be used as a noun, an adjective, or a verb.
      - She bagged two silver medals. [Noun]
      - She made a silver speech. [Adjective]
      - His worries had silvered his hair. [Verb]
  - Soln: POS Tagging: process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence.

Ambiguity

- lexicon
  - **Lexical Semantic Ambiguity**
  - Lexical **Semantic** ambiguity can occur when a word is **polysemous**, i.e. has more than one meaning, and the sentence in which it is contained can be interpreted differently depending on its correct sense.
  - Polysemy, which implies shades of meaning,
    - **Eg.,**
      - ***falling of snow***
      - ***falling of a kingdom. (fall –verb)***
    - Eg:
      - The **tank** was full of water.
      - I saw a military **tank**. ***(tank-noun)***
  - **Soln:**
    - **Word sense or lexical disambiguation** refers to the identification of the meaning of an ambiguous word from clues in the context. (by help of repository like the wordnet)
    - The predominant approach in WSD is supervised learning, where a machine is trained with sense annotated corpora.

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - refers to uncovering the **hierarchical structure behind a linear sequence of words**.
  - **Syntactic Ambiguity**: The structural ambiguities were syntactic ambiguities.
- Eg: , the noun phrase (NP) ***flight from Mumbai to Delhi via Jaipur on Air India***

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - *flight from Mumbai to Delhi via Jaipur on Air India*
  - *has following structure:*

$[NP_4$
  $[NP_3$
    $[NP_2$
      $[NP_1$ [NN flight]]
        $[PP_1$ [P from][NP [NNP Mumbai]]]
      ]
      $[PP_2$ [P to] [NP [NNP Delhi]]]
    ]
    $[PP_3$ [P via][NP [NNP Jaipur]]]
  ]
  [PP_4 [P on][NP [NNP Air-India]]]
]

Ambiguity

- Syntactic Processing/ Analysis (P
  - *flight from Mumbai to Delhi via Jaipur on Air Inc*
  - *has following structure:*

$[NP_4$
  $[NP_3$
    $[NP_2$
      $[NP_1$ [NN flight]]
        $[PP_1$ [P from][NP [NNP Mumbai]]]
      ]
      $[PP_2$ [P to] [NP [NNP Delhi]]]
    ]
    $[PP_3$ [P via][NP [NNP Jaipur]]]
  ]
  $[PP_4$ [P on][NP [NNP Air-India]]]
]

- Such bracketed structures are created by a **Grammar of the language.**
- **Eg: *PP -> P NP* is a grammar rule expressing the fact that a preposition phrase is composed of a preposition and a noun phrase.**

Ambiguity

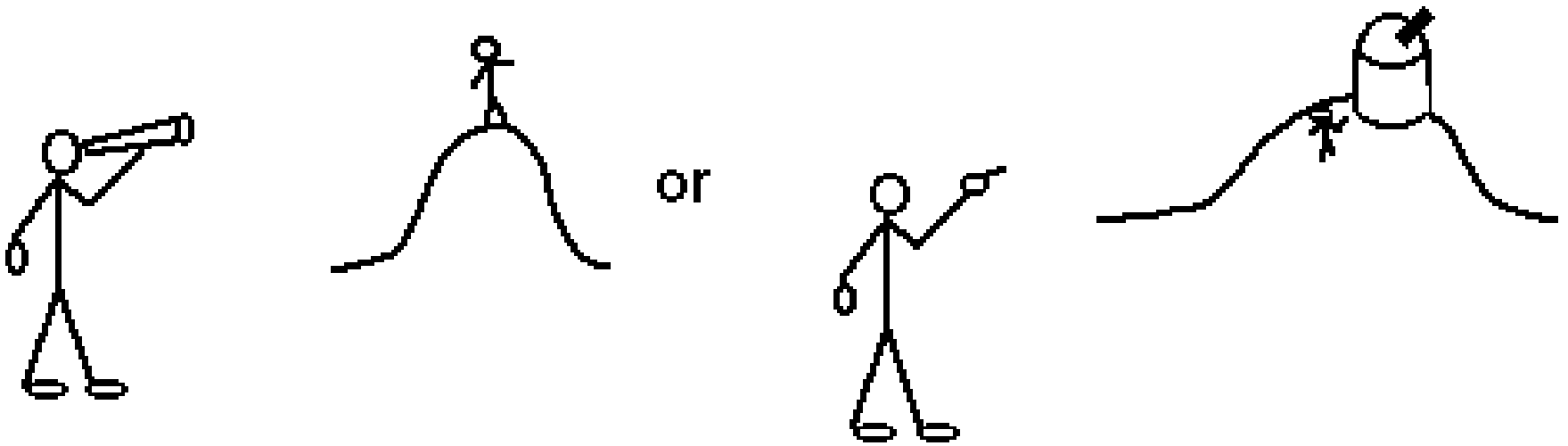
- Syntactic Processing/ Analysis (Parsing):
  - refers to uncovering the **hierarchical structure behind a linear sequence of words**.
  - **Syntactic Ambiguity**: The structural ambiguities were syntactic ambiguities.
  - Two types of *structural ambiguities:*
    - *Scope ambiguity*
    - *Attachment ambiguity*.

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - *Scope ambiguity*
    - **Scope** ambiguity involves operators and **quantifiers**
    - *Eg: Old men and women were taken to safe locations.*
      - *Here,* The scope of the adjective (*i.e.*, the amount of text it qualifies) is ambiguous.
      - That is, is the structure *(old men and women)* or *((old men) and women)*?
    - Eg: Every man loves a woman.
      - The interpretations can be, For every man there is a woman and also it can be there is one particular woman who is loved by every man.

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - if given Sentence-
      - **fits more than one position in a parse tree.**
      - arises from uncertainty of **attaching a phrase or clause to a part of a sentence**
      - **Eg:** *I saw the boy with a telescope.*



Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - **Eg:** *I saw the boy with a telescope.*
      - 1. the preposition phrase ***with a telescope*** attaches with the verb *saw* with the instrumental case.
      - 2. the PP attaches to *the boy* as a modifier.
    - **PP-attachment is a classical problem in NLP**
      - Given the structure : *V-NP $_1$-P-NP$_2$*
      - *Where does NP2 attach, V or NP1?*
      - ***Soln:*** ***Both*** rule based and machine learning based approaches.

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - **Eg:** *I saw the boy with a telescope.*
      - 1. the preposition phrase ***with a telescope*** attaches with the verb *saw* with the instrumental case.
      - 2. the PP attaches to *the boy* as a modifier.
    - ambiguity of attachment arises from the **dual role of** prepositions, *viz.,* assigning case to nouns with respect to a verb and for modifying a noun phrase.
    - In example, *with* can assign instrument case to *telescope* or specify a particular boy *having* a telescope

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - **PP-attachment** for Hindi language:
      - *post positions* instead of prepositions,
      - entities that assign case roles follow the noun, and do not precede.
      - “दूरबीन से लड़के को देखा”
      - The postposition ***se*** assigns case role to ***duurbiin*** and follows it.
      - Attachment ambiguity of the type pp-attachment is not so common in Indian languages which are as a rule SOV (subject object- verb) languages.
      - As, Postpositions follow this pattern:
        - » ***NP1 P NP2 V***

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - Phrase
    - Clause

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - Phrase
      - Eg: I saw a tiger running across the field
        - » Who was running: *I* or *the tiger*? The attachment of the phrase *running across the field* is ambiguous.

Ambiguity

- Syntactic Processing/ Analysis (Parsing):
  - **Attachment Ambiguity**
    - Clause
      - *Eg: I told the child that I liked that he came to the playground early.*
        - The sentence has two meanings:
          - (a) *I told the child the FACT that I liked his coming early to the ground*
          - (b) *I told the child WHOM I liked that he came early to the ground.*
- *constituency parsing* that identifies phrases in a sentence,
- *dependency parsing* that finds heads and modifiers.

Ambiguity

- Semantics
  - After word forms and structure have been detected, sentence processing devotes itself to **meaning extraction.**
  - the sentence needs to be **represented in one of the unambiguous forms like predicate calculus, semantic net, frame, conceptual dependency, conceptual structure etc.**
  - **Semantic Ambiguity**: This occurs when the meaning of the words themselves can be **misinterpreted. Even after the syntax and the meanings of the individual words have been resolved.**

Ambiguity

- Semantics
  - **Semantic Ambiguity**
    - Eg: Seema loves her mother and Sriya does too.
      - The interpretations can be Sriya loves Seema's mother or Sriya likes her own mother.
      - Semantic ambiguities born from the fact that generally a computer is not in a position to distinguishing what is logical from what is not.

Ambiguity

- Semantics
  - **Semantic Ambiguity**
    - आपको मुझे मीठाई खीलानी पड़ेगी (aapko mujhe mithaai khilaanii padegii)
      - *You will have to feed me sweets*

        Or
      - *I will have to feed you sweets*
      - Both *I* and *you* have semantic role ambiguity *(agent vs. beneficiary).*

Ambiguity

- **Pragmatics Processing**
  - Pragmatic ambiguity refers to a situation where the context of a phrase gives it **multiple interpretation**
  - The problem involves processing **user intention, sentiment, belief world, modals *etc.*- all of which are highly complex tasks.**
  - Pragmatic ambiguity
    - when the **statement is not specific, and the context does not provide the information** needed to clarify the statement.
    - Information is missing, and must be inferred.

Ambiguity

- **Pragmatics Processing**
  - I like you too.
  - This can be interpreted as
    - I like you (*as if "* you like me")
    - I like you ("I like you and someone else")

Ambiguity

- **Ellipsis:**
  - The **omission** (**leaving out)** of words that are **needed for grammatical completion, and are "understood".**
  - Ellipsis causes problems for NLP since it is necessary to infer the rest of the sentence from the context.
  - "ellipsis" is also the name of the symbol "..." used when something is omitted from a piece of text.
  - Eg. 1:
    - She said, "I like apples, oranges and bananas because they are all fruits."
    - She said, **"I like apples, oranges and bananas … .“**
  - Eg. 2 :
    - So…what happened?
  - *Eg. 3:*
    - ***Should you bunk?*** *is a sentence with ellipsis*
      - *i.e.,* ***gap or*** **omission** which needs to be filled with text from an earlier sentences. Bunk what? Bunk the *school.*
      - *Additionally, Bunk* is ambiguous with respect to both POS and sense.

Ambiguity

- **Discourse**
  - processing connected sentences
  - processing needs a **shared world or shared knowledge and the interpretation** is carried out using this context.
- *Example:*
  - *Sentence-1: John was coming dejected from the school*
  - (who is John: most likely a student?)
  - *Sentence-2: He could not control the class*
  - (who is John now? Most likely the teacher?)
  - *Sentence-3: Teacher should not have made him responsible*
  - (who is John now? Most likely a student again, albeit a
  - special student- the monitor?)
  - *Sentence-4: After all he is just a janitor*
  - (all previous hypotheses are thrown away!).

Ambiguity

- **Discourse**
  - **Anaphoric Ambiguity**
    - Anaphoras are the entities that have been previously introduced into the discourse.
    - “The horse ran up the hill. It was very steep. It soon got tired.”
      - The anaphoric reference of ‘it’ in the two situations cause ambiguity.
        - » 1. Steep applies to surface hence ‘it’ can be hill.
        - » 2. Tired applies to animate object hence ‘it’ can be horse.

Ambiguity

- **Discourse**
  - **Anaphoric Ambiguity**
  - Eg:
    - "Margaret invited Susan for a visit, and she gave her a good lunch." (she = Margaret; her = Susan)
    - "Margaret invited Susan for a visit, but she told her she had to go to work" (she = Susan; her = Margaret.)

Ambiguity

- **Discourse**
  - *Cataphora*
    - When the reference to an entity is in the forward direction
    - *e.g. "*in *that he will win was clear to Sam"*,
      - where *he* is forward bound to *Sam*, the binding is called *cataphora*
    - " When **he** arrived home, **John** went to sleep"
      - *he* appears earlier than the noun *John* that it refers to.

Ambiguity

- **Textual Humour and Ambiguity**
  - Arise from **incongruity** of views
  - Computational humour is an actively researched area.
  - **Computational humor (Eg: mood/temp detection)** is a branch of computational linguistics and artificial intelligence which uses computers in humour research

Ambiguity

- **Textual Humour and lexical Ambiguity**
  - *Eg: A car owner after coming back from a party finds the sticker "parking fine" on his car. He goes and thanks the policeman for appreciating his parking skill.*
  - The ambiguity of the word *fine* (*nice vs. penalty*) and the two different meanings picked by the car owner and **the policeman give rise to the humour**

Ambiguity

- Textual Humour and structural ambiguity:
  - *Eg:*
    - *Teacher: What do you think is the capital of Morocco?*
    - *Student: What do you think?*
    - *Teacher (Angrily): I do not think, I know.*
    - *Student: I … do not think I know.*
  - The attachment ambiguity of *I know* (standalone sentence vs. getting attached to *think*) and the two different attachments picked by the teacher and the student give rise to humour

Ambiguity

I made her duck.

- How many different interpretations does this sentence have?
- What are the reasons for the ambiguity?
- The categories of knowledge of language can be thought of as ambiguity resolving components.
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?
  - Yes – deciding word boundaries

Ambiguity (cont.)

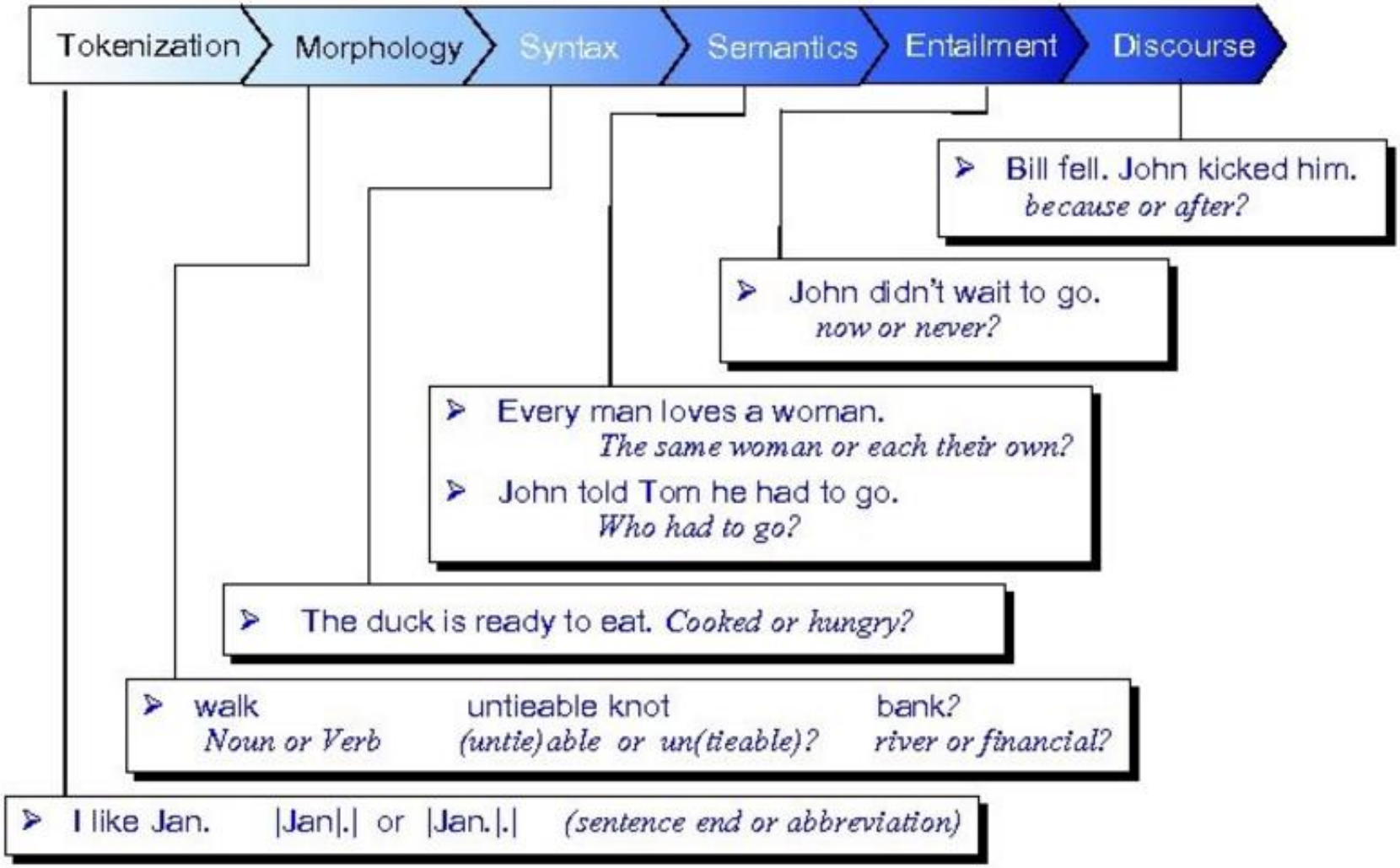
- Some interpretations of : **`I made her duck.`**
  1. I cooked *duck* for her.
  2. I cooked *duck* belonging to her.
  3. I created a toy duck which she owns.
  4. I caused her to quickly lower her head or body.
  5. I used magic and turned her into a *duck*.
- `duck` – **morphologically and syntactically** ambiguous: noun or verb.
- `her` – **syntactically ambiguous**: dative (third person) or possessive (owner).
- `make` – **semantically ambiguous**: cook or create.

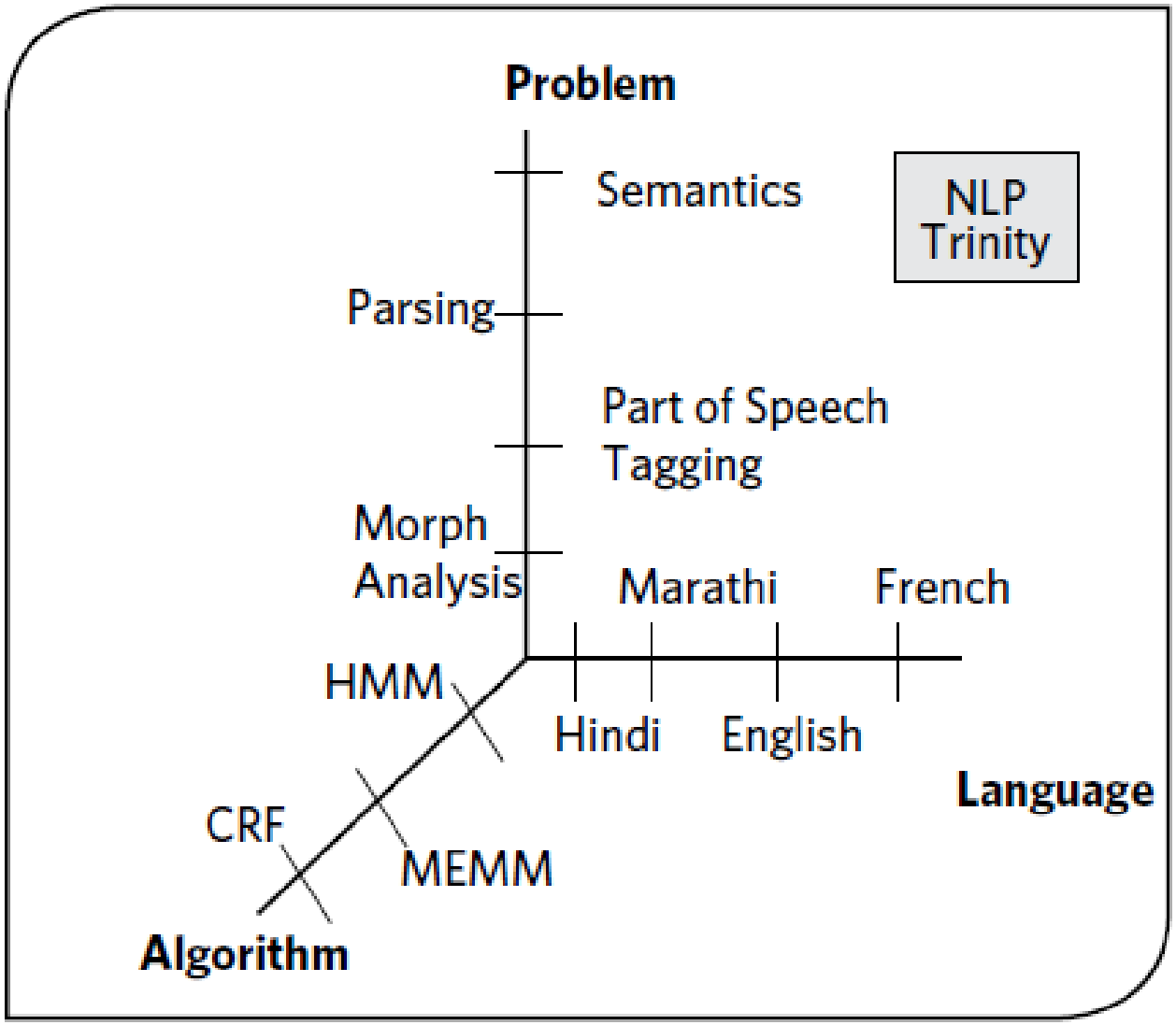
Ambiguity

- **Phonetics!**
  - 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
  - I mate or duck

Ambiguity is Pervasive!

Language has pervasive ambiguity





Three dimensions of NLP

Models to Represent Linguistic Knowledge

- We will use certain formalisms (*models*) to represent the required linguistic knowledge.
- **State Machines** -- FSAs, FSTs, HMMs, ATNs, RTNs
- **Formal Rule Systems** -- Context Free Grammars, Unification Grammars, Probabilistic CFGs.
- **Logic-based Formalisms** -- first order predicate logic, some higher order logic.
- **Models of Uncertainty** -- Bayesian probability theory.

Phonology , Morphology :
State Machines , Formal Rule Systems, etc.

Syntax :
Trees can be generated using CFGs; depth-first search, as well as heuristic variants such as **best-first**, and **A* search, etc.**

Semantics, Discourse and Pragmatics:
capturing knowledge of language is logic.
first order logic, **predicate calculus**, etc.

- Probability theory is the final element in our set of techniques for capturing linguistic knowledge.
- One major use of probability theory is to solve the many kinds of ambiguity problems probabilistic models - are one of a class of **machine learning** models.

Brief History of NLP

- 1940s –1950s: Foundations
  - Development of **formal language theory** (Chomsky, Backus, Naur, Kleene)
  - Probabilities and information theory (Shannon)
- 1957 – 1970s:
  - Use **of formal grammars** as basis for natural language processing (Chomsky, Kaplan)
  - Use **of logic and logic based programming** (Minsky, Winograd, Colmerauer, Kay)
- 1970s – 1983:
  - **Probabilistic methods for early speech recognition** (Jelinek, Mercer)
  - Discourse modeling (Grosz, Sidner, Hobbs)
- 1983 – 1993:
  - **Finite state models (morphology)** (Kaplan, Kay)
- 1993 – present:
  - Strong integration of different techniques, different areas.

Natural Language Understanding

Words

Morphological Analysis

Morphologically analyzed words *(another step: POS tagging)*

Syntactic Analysis

Syntactic Structure

Semantic Analysis

Context-independent meaning representation

Discourse Processing

Final meaning representation

Natural Language Generation

Meaning representation

Utterance Planning

Meaning representations for sentences

Sentence Planning and Lexical Choice

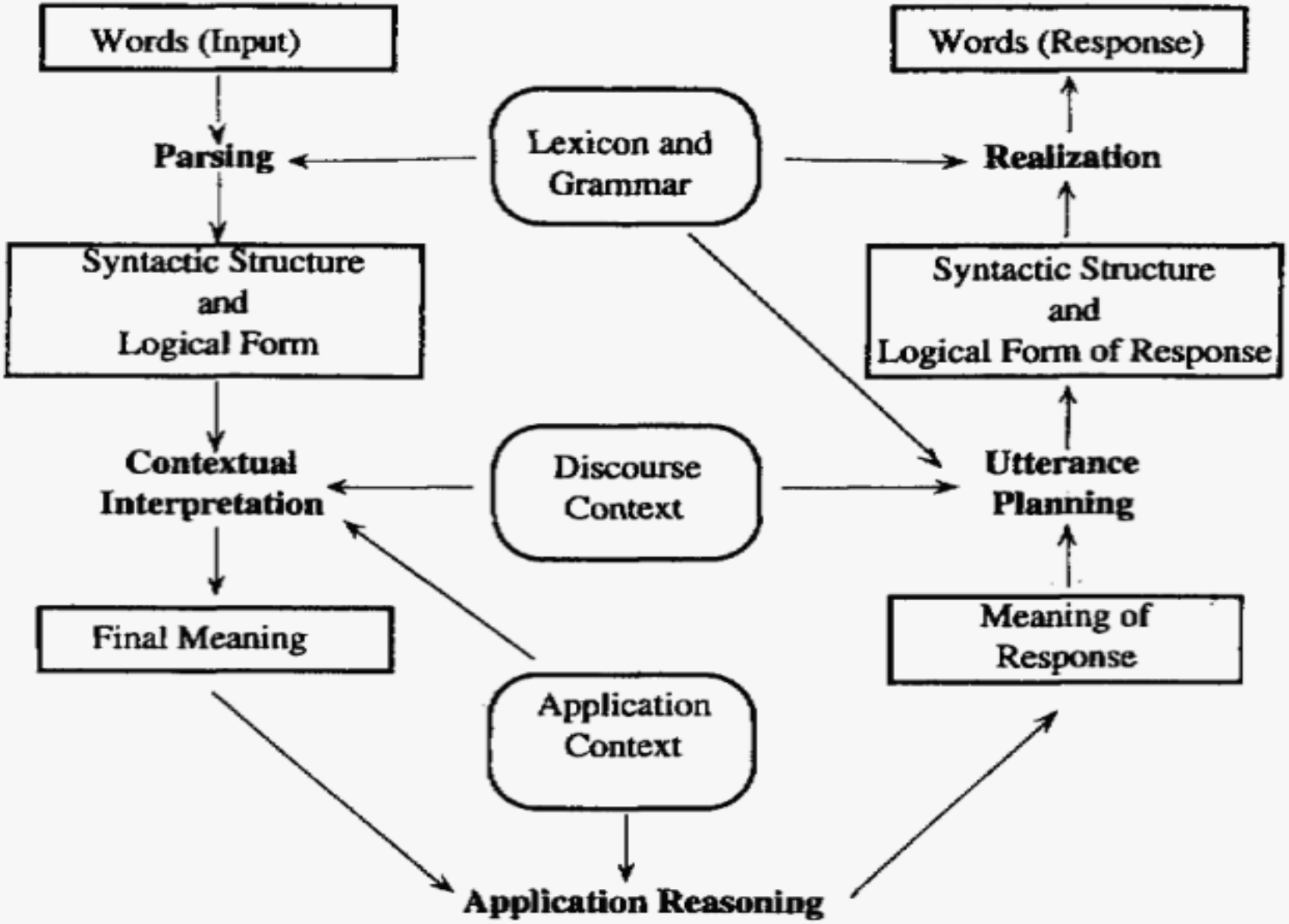
Syntactic structures of sentences with lexical choices

Sentence Generation

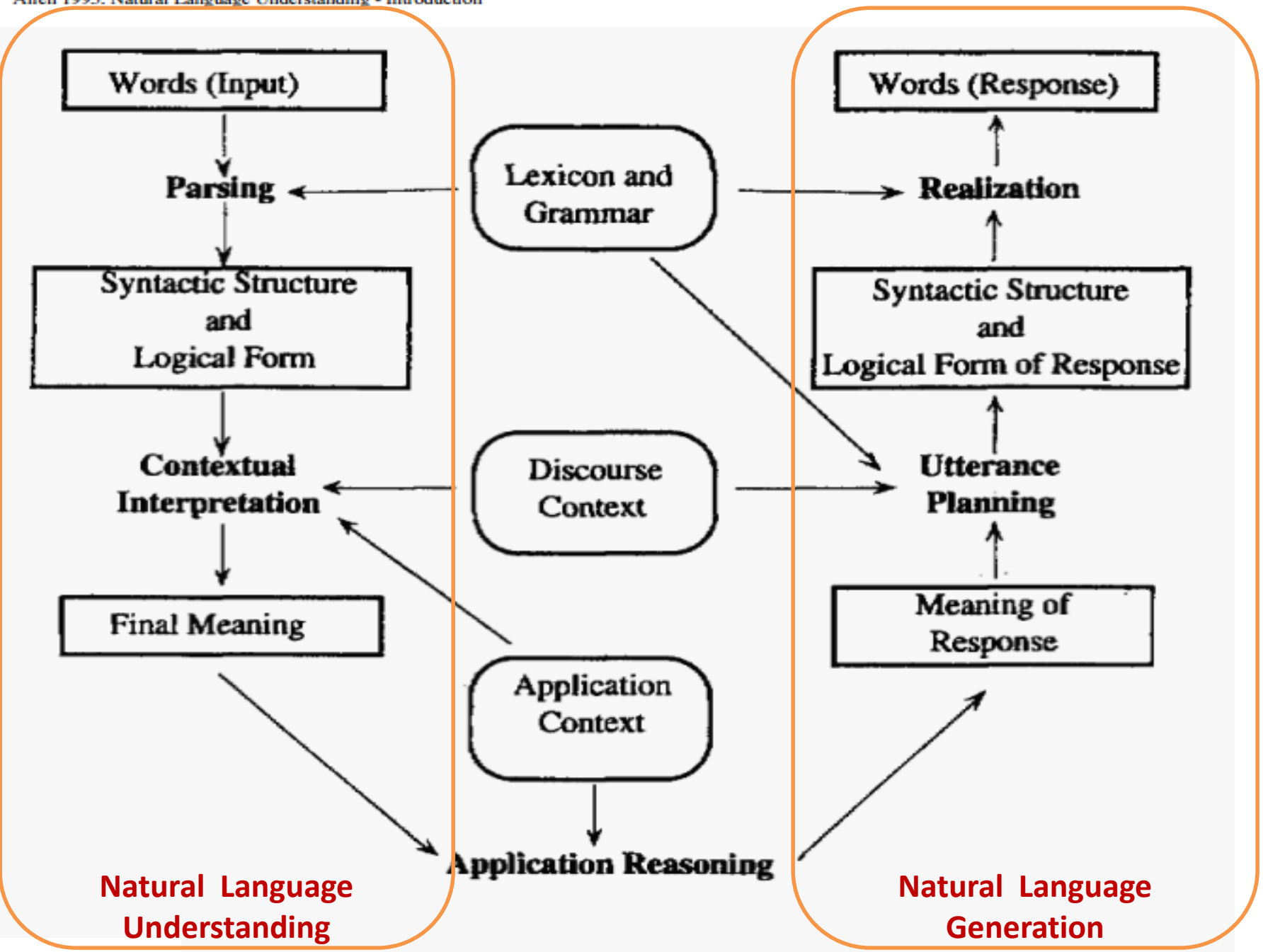
Morphologically analyzed words

Morphological Generation

Words



The flow of information



The flow of information

Thank You!!