MENOUFIA UNIVERSITY
FACULTY OF COMPUTERS AND INFORMATION

Fourth Year (Second Semester)
CS Dept., (CS 436)

Natural Language Processing (NLP)

Lecture one

Dr. Hamdy M. Mousa

Definitions

- The goal of this field is to get computers to perform useful tasks involving human language, tasks like enabling
  - human-machine communication,
  - improving human-human communication, or
  - simply doing useful processing of text or speech.

Natural Language

- In this text we study the various components that make up modern conversational agents, including language input
  - **Automatic speech recognition**
  - **Natural language understanding**
  - Language output (dialogue and response planning and **speech synthesis.**

Definitions

- Modern conversational agents can:
  - answer questions,
  - book flights, or
  - find restaurants,
  - functions for which they rely on a much more sophisticated understanding of the user's intent,

Natural Language

- The goal of **machine translation** is to automatically translate a document from one language to another.

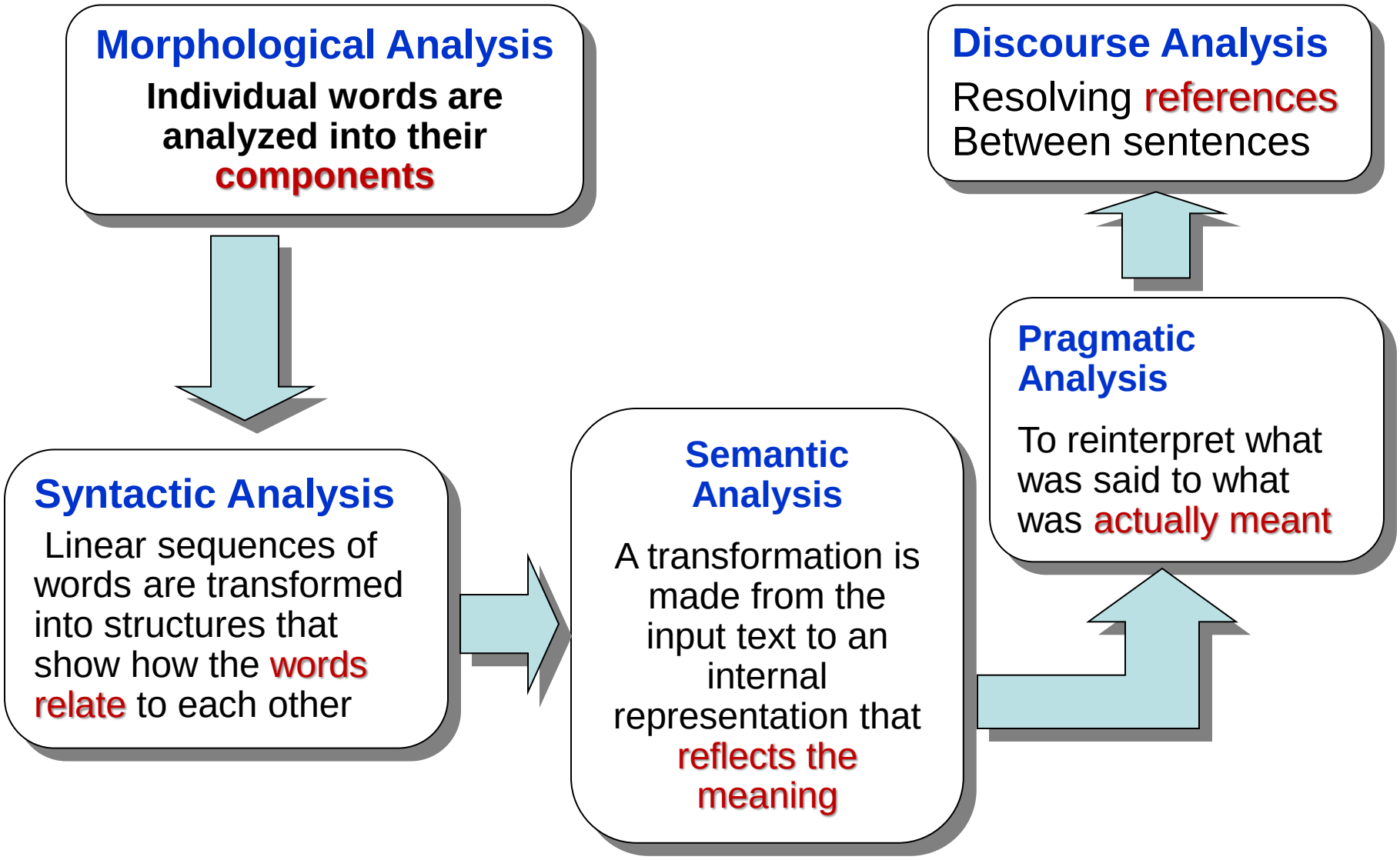
- Many other language processing tasks are also related to the Web.
  - such *Question* task is **Web-based question answering**.

Natural Language Processing

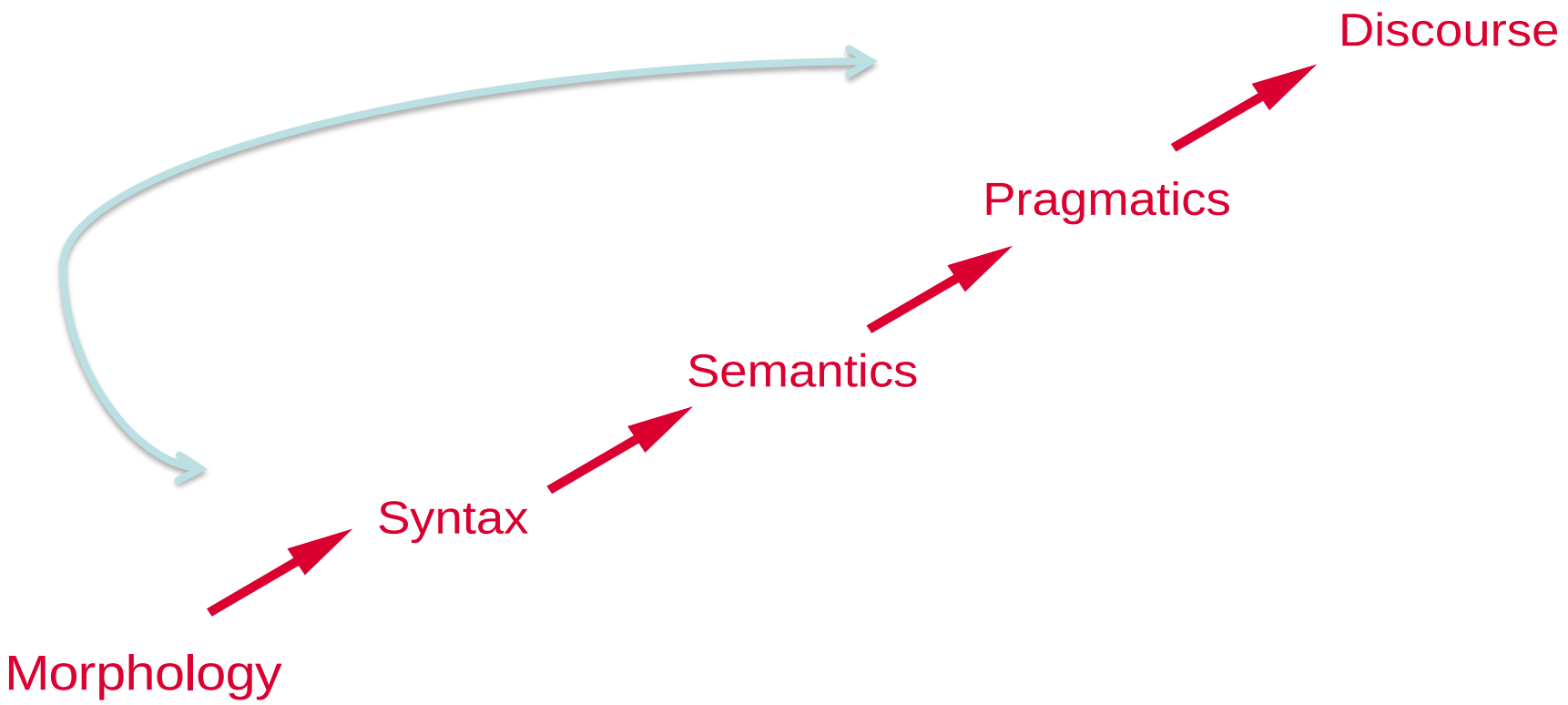
- The sub-domain of artificial intelligence concerned with the task of developing programs possessing some capability of 'understanding' a natural language in order to achieve some specific goal.

- Computers use (analyze, understand, generate) natural language

Steps of NLP


Morphological Analysis
Individual words are analyzed into their components
Syntactic Analysis
Linear sequences of words are transformed into structures that show how the words relate to each other
Semantic Analysis
A transformation is made from the input text to an internal representation that reflects the meaning
Pragmatic Analysis
To reinterpret what was said to what was actually meant
Discourse Analysis
Resolving references Between sentences

The Steps in NLP



** we can go up, down and up and down and combine steps too!!
** every step is equally complex

The steps in NLP

- Morphology: Concerns the way words are built up from smaller meaning bearing units.
- Syntax: concerns how words are put together to form correct sentences and what structural role each word has
- Semantics: concerns what words mean and how these meanings combine in sentences to form sentence meanings

Cont., The steps in NLP

- Pragmatics: concerns how sentences are used in different situations and how use affects the interpretation of the sentence
  - knowledge of the relationship of meaning to the goals and intentions of the speaker

- Discourse: concerns how the immediately preceding sentences affect the interpretation of the next sentence

Parsing (Syntactic Analysis)

- Assigning a syntactic and logical form to an input sentence
  - uses knowledge about word and word meanings (lexicon)
  - uses a set of rules defining legal structures (grammar)

Ahmad ate the apple.

```
(S  (NP (NAME Ahmad))
    (VP  (V ate)
         (NP  (ART the)
              (N apple))))
```

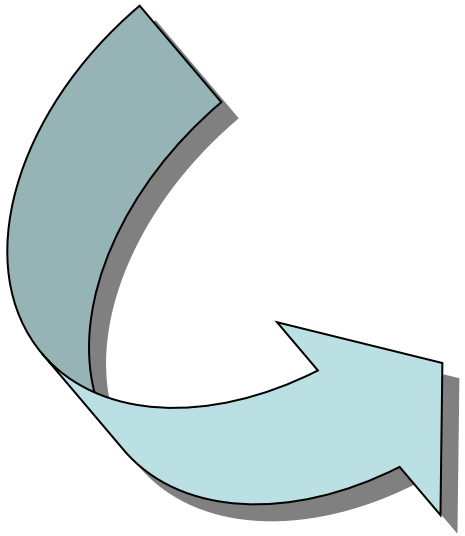
Word Sense Resolution

- Many words have many meanings or senses
- We need to resolve which of the senses of an ambiguous word is invoked in a particular use of the word
- I made her duck.
  - made her a bird for lunch or
  - made her move her head quickly downwards?

Ex. Morphological analysis

Surface form

I want to print Ali's .init file



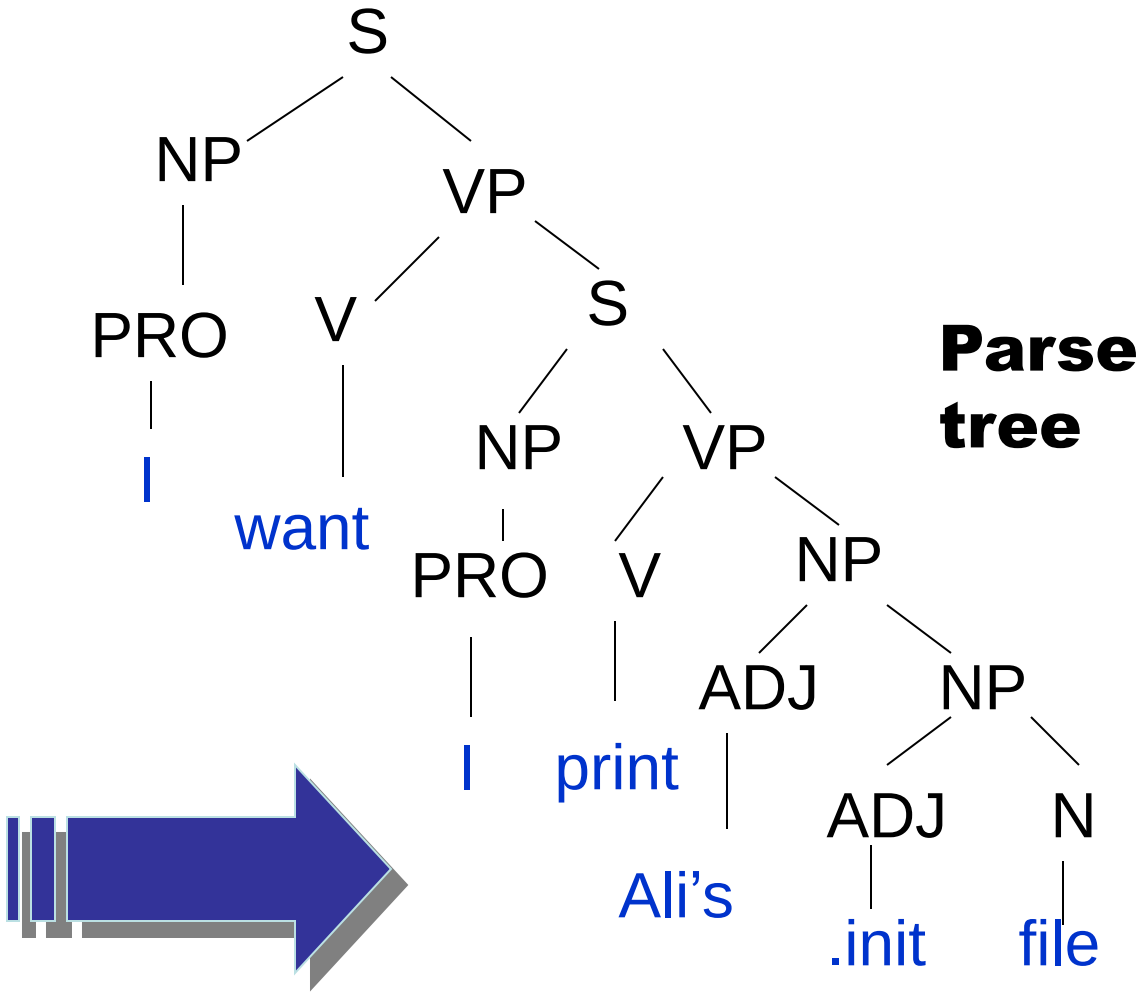
stems

I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
's (possessive)
.init (adj)
file (noun)
file (verb)

Ex. Syntactic analysis

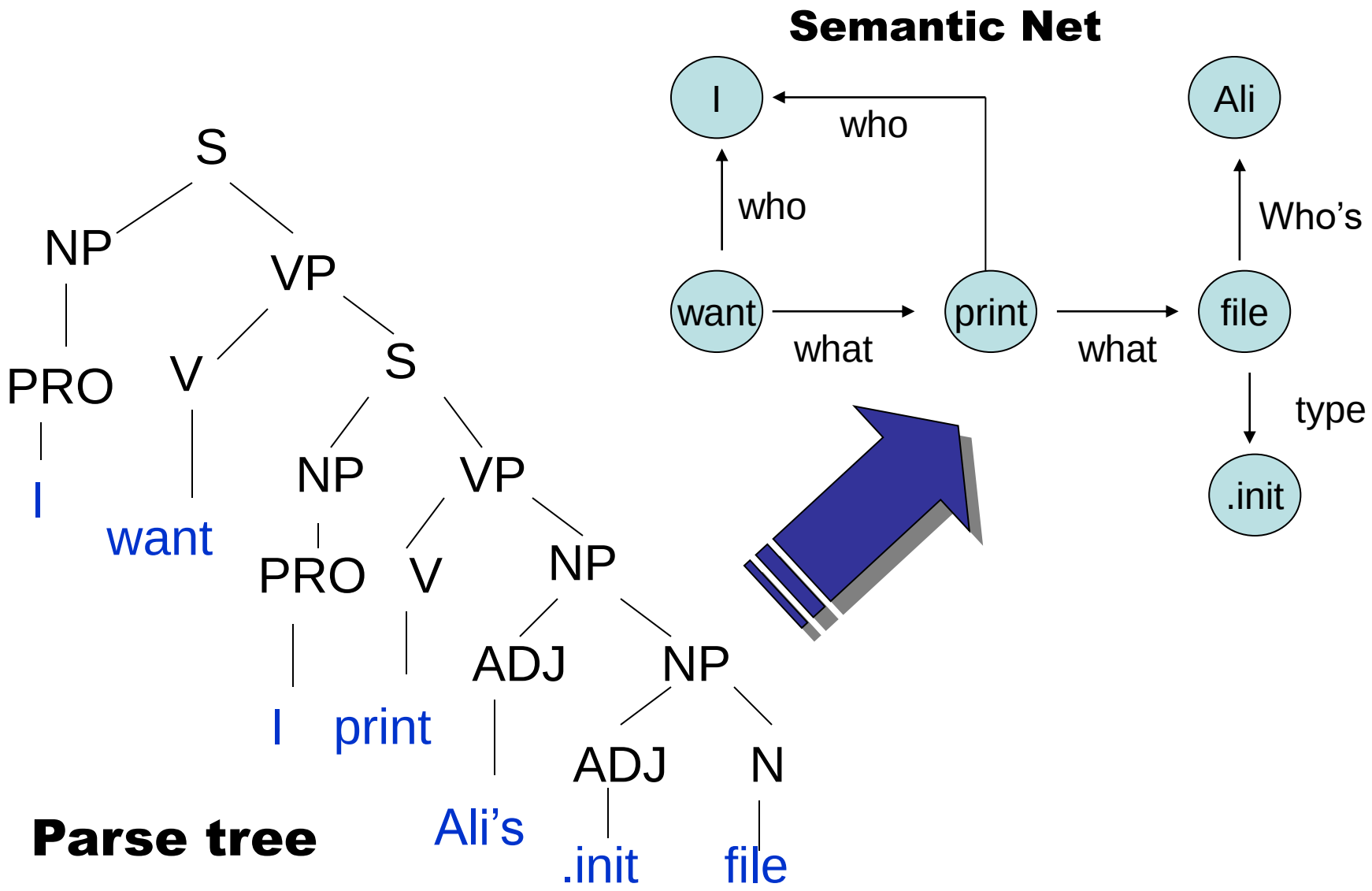
stems

I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
‘s (possessive)
.init (adj)
file (noun)
file (verb)

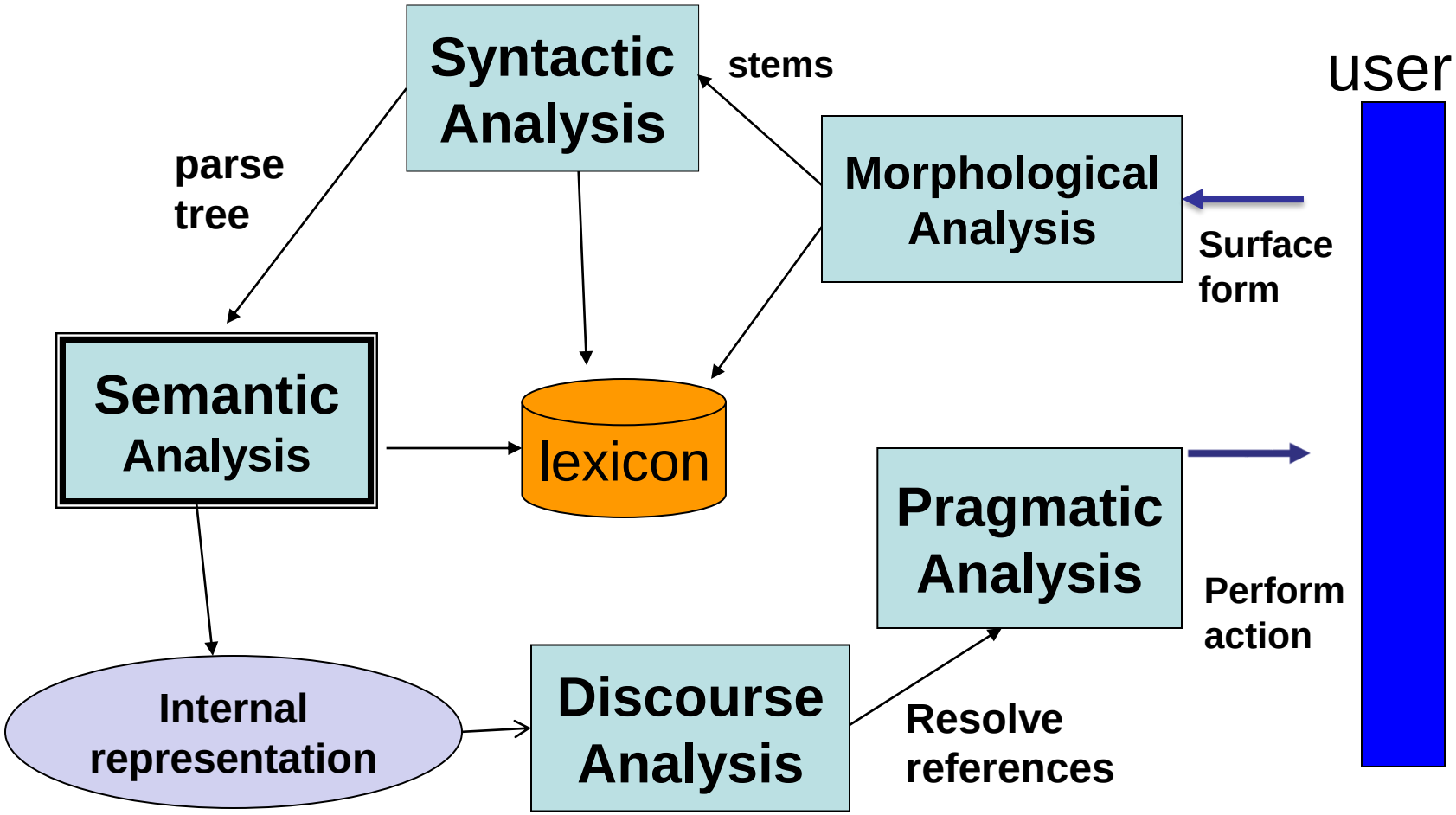


Parse tree

Semantic analysis



The steps of NLP



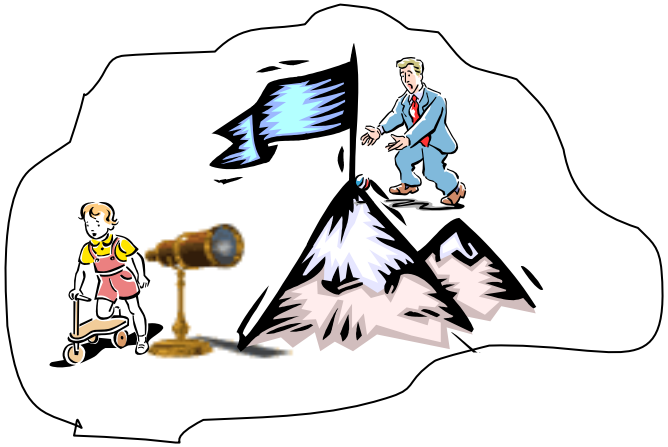
Ambiguity

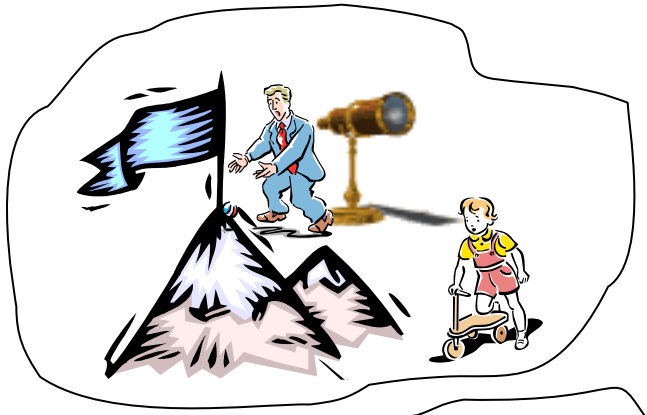
- These categories of linguistic knowledge is that most tasks in speech and language processing can be viewed as resolving **ambiguity** at one *Ambiguous* of these levels.
- We say some input is **ambiguous** if multiple, alternative linguistic structures can be built for it.
- **It more than one meaning for the same sentence**

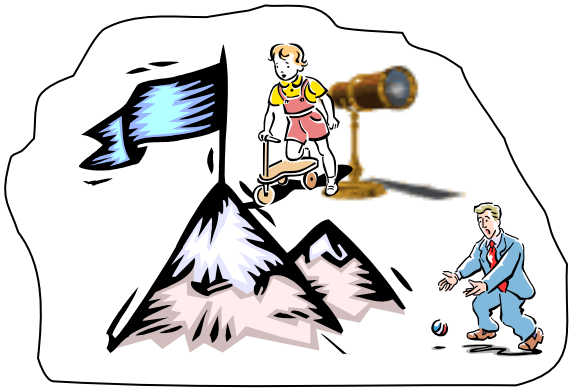
Ambiguity

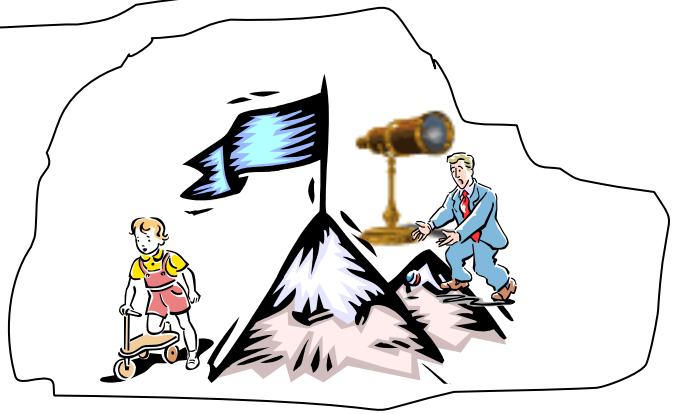
The boy saw the man on the mountain with a telescope

Prepositional phrase attachment









Ambiguity

Short men and women

Visiting relatives can be boring

The chicken is ready to eat

Ali knows a richer man than Ahmad

I made her duck

We eat what we can, and what we can not we can

Basic NLP Terminology

- Token
- Sentence
- Tokenization
- Corpus
- Part-of-speech (POS) tag
- Parse tree

Basic NLP Terminology

- **Token**: Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, numbers or alphanumeric. These units are known as tokens.
- **Sentence**: An ordered sequence of tokens.
- **Corpus:** A body of text, usually containing a large number of sentences.

Basic NLP Terminology

- **Tokenization:** The process of splitting a sentence into its constituent tokens.
- For segmented languages such as English, the existence of white space makes tokenization relatively easier and uninteresting.
  - However, for languages such as Arabic and Chinese, the task is more difficult since there are no explicit boundaries.
  - Furthermore, almost all characters in such non-segmented languages can exist as one-character words by themselves but can also join together to form multi-character words.

Basic NLP Terminology

- **Part-of-speech (POS) tag :** A word can be classified into one or more of a set of lexical or part-of-speech categories such as Nouns, Verbs, Adjectives and Articles, to name a few.
  - A POS tag is a symbol representing such a lexical category - NN(Noun), VB(Verb), JJ(Adjective), AT(Article).
    - One of the oldest and most commonly used tag sets is the Brown Corpus tag set.

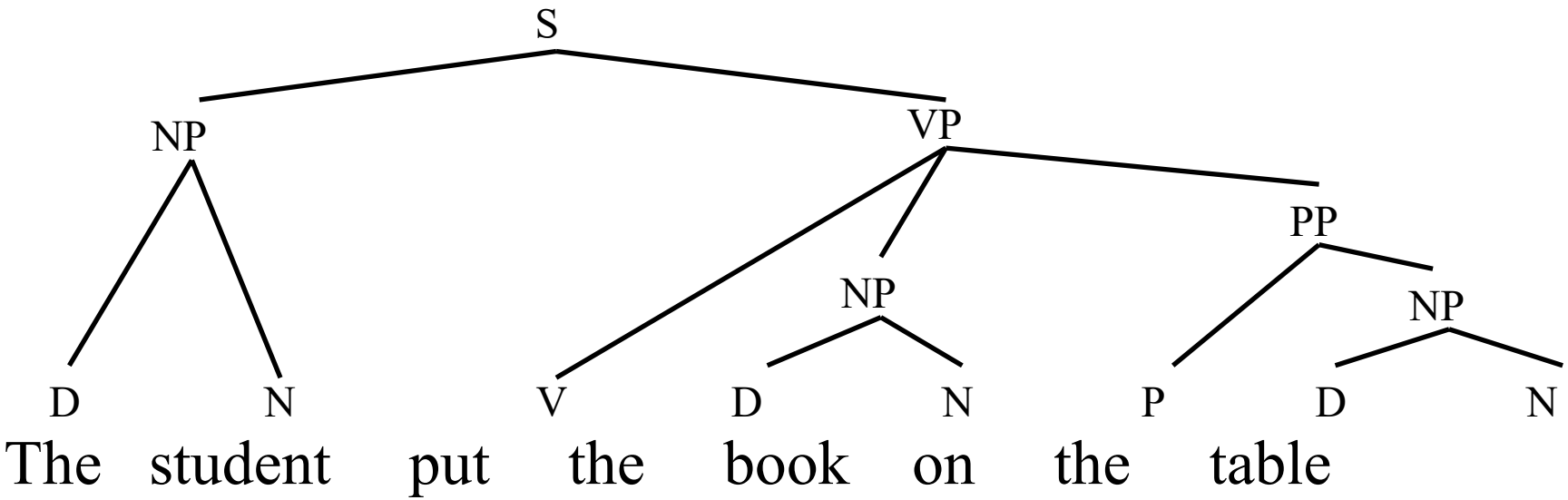
Basic NLP Terminology

- **Part-of-speech (POS) tag :**

	Tag	Arabic Name	Description
Nouns	N	اسم	Noun
	PN	اسم علم	Proper noun
	IMPN	اسم فعل أمر	Imperative verbal noun
Pronouns	PRON	ضمير	Personal pronoun
	DEM	اسم اشارة	Demonstrative pronoun
	REL	اسم موصول	Relative pronoun
Nominals	ADJ	صفة	Adjective
	NUM	رقم	Number
Adverbs	T	ظرف زمان	Time adverb
	LOC	ظرف مكان	Location adverb

Basic NLP Terminology

- **Parse tree :** A tree defined over a given sentence that represents the syntactic structure of the sentence as defined by a formal grammar.



Some common NLP tasks

- POS tagging
- Computational morphology
- Parsing

Some common NLP tasks

- **POS tagging**: Given a sentence and a set of POS tags, a common language processing task is to automatically assign POS tags to each word in the sentences.
  - For example, given the sentence;
    - " The ball is red",
    - the output of a POS tagger would be 'The/AT ball/NN is/VB red/JJ'.
  - Tagging text with parts-of-speech turns out to be extremely useful for more complicated NLP tasks such as parsing and machine translation.

Some common NLP tasks

- Natural languages consist of a very large number of words that are built upon basic building blocks known as **morphemes** (**stems**); *the smallest linguistic units possessing meaning*.
- **Computational morphology:** is concerned with the discovery and analysis of the internal structure of words using computers.

Some common NLP tasks

- **Parsing :** In the parsing task, a parser constructs the parse tree given a sentence.
  - Some parsers assume the existence of a set of grammar rules in order to parse.
  - recent parsers are smart enough to deduce the parse trees directly from the given data using complex **statistical** models.
- Most parsers also operate in a supervised setting and require the sentence to be POS-tagged before it can be parsed.
  - Statistical parsing is an area of active research in NLP.

Lexicon

- Lexicon is a vocabulary data bank, that contains the language words and their linguistic information.
- There are many on-line lexicon

- **WordNet** is a lexical database that contains English vocabulary words
- **COULD WE HAVE ONE FOR ARABIC?**

Some Applications

- Intelligent computer systems
- NLU interfaces to databases
- Computer aided instruction
- Information retrieval
- Intelligent Web searching
- Data mining
- Machine translation
- Automatic Speech recognition
- Natural language generation
- Question answering
- Text Summarization
- Information extraction
- ……..