# Introduction to NLP

(Text Based Natural Language Processing)

# What is Natural Language Processing?

- NLP aims to analyze language not only as a sequence of words but for its meaning as well.

- Natural Language Processing is a field  at intersection of computer science , artificial intelligence, linguistics.

# Why should we care about NLP?

We want computers to understand human language.

So that:

1. Computers can  communicate with humans.
2. We want computers to perform  certain tasks using language data, like:
    a. Text Classification
    b. Information Retrieval
    c. Information extraction / Question Answering
    d. Language Translation

# NLP in Industry

1. Search engines
2. Grammarly
3. Autocompletion
4. Sentiment Analysis
5. Language Translation
6. Chatbots
7. .....

# What makes NLP challenging?

1. Natural languages are **ambiguous**, e.g
   print(1+1)  always prints 2
   I want to eat honey.  I want to eat honey. I made her duck. I made her duck

2. Natural language **assumes contextual information** is known, as a result it does not go deep explaining everything, eg
   I ask you to count the number of girls in the room. You get up and count.
   To a computer I need to define what is a room, what is a girl and what count means!

3. The corpus of natural languages is large and dynamic.

# Characteristics of Natural Languages:

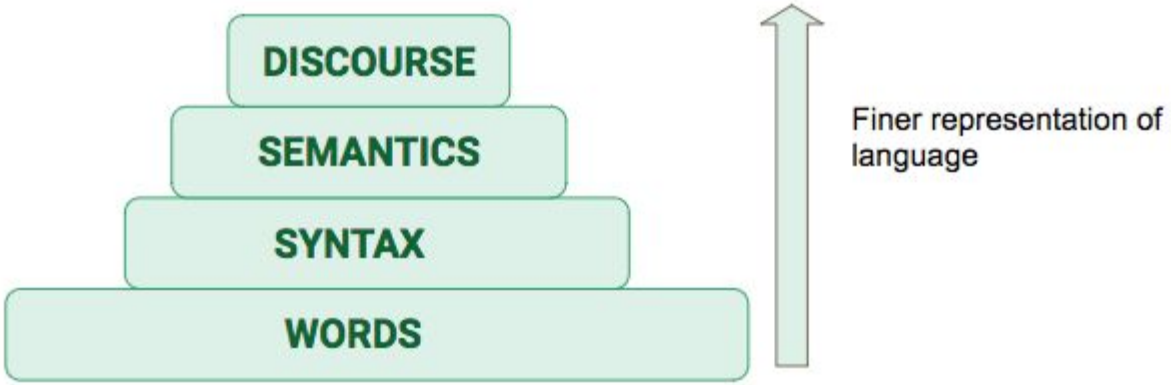
1. Construction units or words
2. Syntax
3. Semantics
4. Discourse

**Syntax:**The way in which words are put together to form phrases, clauses, or sentences; the part of grammar dealing with this.

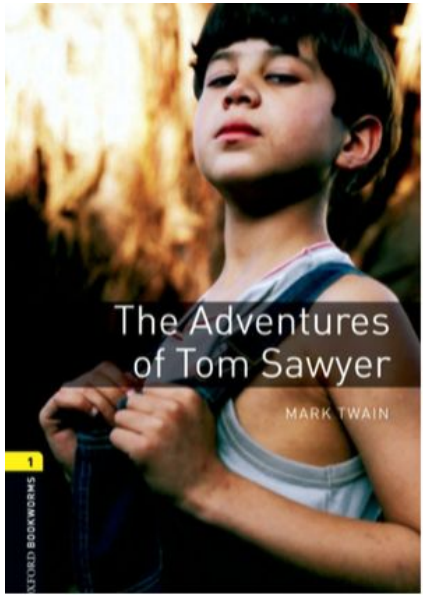
**Semantics:** The study of the meanings of words and phrases in language.

**Discourse:** The study the meanings of sentences in context with one another.

# Analysis levels



# Consider this text from a novel:




"TOM!"

No answer.

"TOM!"

No answer.

"What is gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Example text from David Bamman material

WORDS

# A. Represent the corpus as a collection of words:

“TOM!”
No answer.

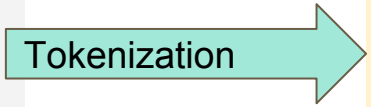
“TOM!”

No answer.

“What is gone with that boy, I wonder? You TOM!”

No answer.

The old lady pulled her spectacles down and looked over them about the room.




tom no answer tom no answer tom what is gone with that boy wonder you tom no answer the old lady pulled her spectacles down and looked over them about the room

# Count the word frequencies:

tom no answer tom no answer what
is gone with that boy wonder you
tom no answer the old lady pulled
her spectacles down and looked
over them about the room

**Bag of Words (BOW)** representation of this corpus.

| Word | count | Word | count |
|---|---|---|---|
| tom | 3 | old | 1 |
| no | 3 | lady | 1 |
| answer | 3 | pulled | 1 |
| what | 1 | her | 1 |
| is | 1 | spectacles | 1 |
| gone | 1 | down | 1 |
| with | 1 | and | 1 |
| that | 1 | looked | 1 |
| boy | 1 | over | 1 |
| wonder | 1 | them | 1 |
| you | 1 | about | 1 |
| the | 2 | room | 1 |

SYNTAX

WORDS

# B. Syntax Analysis:

We want to understand the **structural relationship** between words and how language is constructed.

1. Identify **parts of speech** -nouns, verbs, adjectives

2. Identify **named entities** - look for nouns.

3. Identify r**elations or structural phrases.**

## PARTS OF SPEECH TAGGING:

--- nouns
--- verbs
--- adjectives

"TOM!"
No answer.

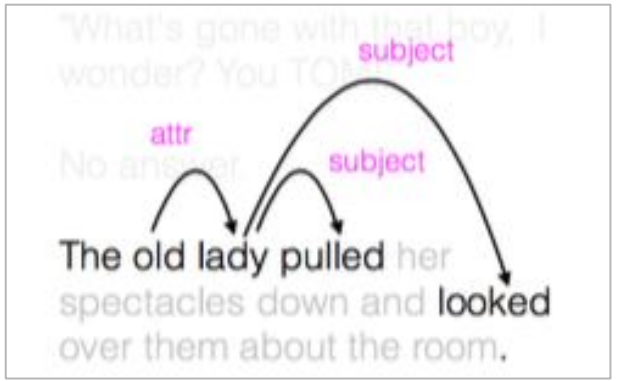
"TOM!"
No answer.

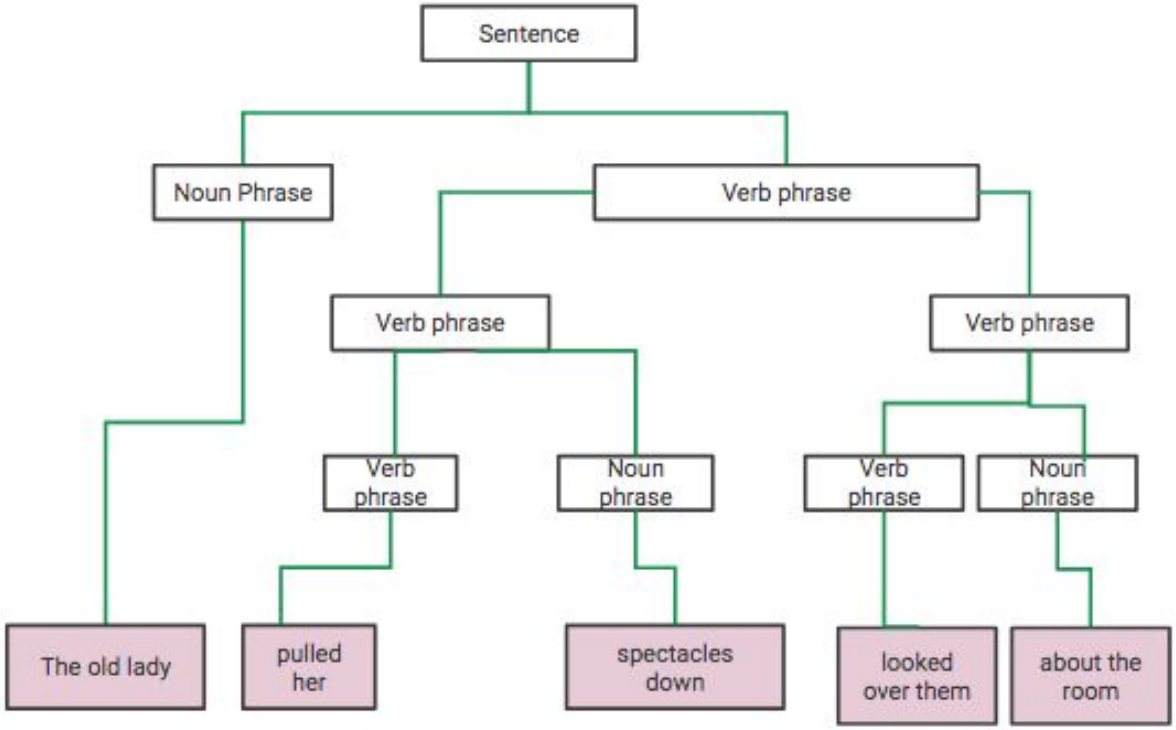
"What is gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

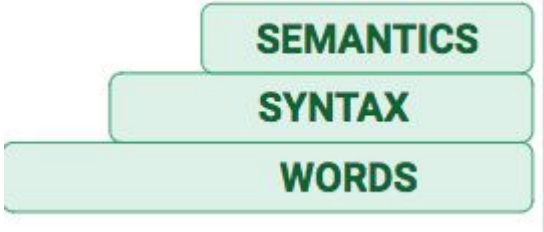
## Chunking:






Simple Syntax Parse Tree showing Noun and Verb phrases in a sentence

# C. Semantic Analysis:




We study the relationship between different parts of a corpus in creating meaning that is language independent..

**Lexical semantics** – meanings of component words of a corpus
I am good . I am ok. I am unwell.
Cat is a mammal.

**Compositional semantics** – how words or phrases combine to create meaning.

# Different semantic frames for a same word:

Apply_heat frame: “Michelle baked the potatoes for 45 minutes .”
Cooking_creation frame: “Michelle baked her mother a cake for birthday.”
Absorb_heat frame: “The potatoes have to bake for more than 30 minutes.”

Not_interesting: Too much theory makes a lecture dry.
Water_evaporated: Your jacket is dry now.

*Some definitions you might come across in NLP:

**Lemma:** The canonical form of an inflected word, e.g. baked ---->bake
**Morphemes:** The smallest linguistic unit within a word that can carry a meaning e.g. clueless--->clue, less

## Co-references Resolution

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

## Identify Speakers

"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

# D. Discourse Analysis:

Discourse is dialog or communication, which always has an underlying subject and tone. Here we study how sentences in a corpus affect each other and  overall meaning of the language.

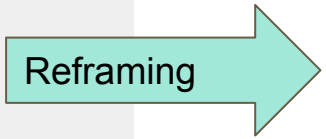
E.g

"TOM!"
No answer.
"TOM!"
No answer.

"What is gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.




An old woman is looking for a boy named TOM.

# How is NLP is done? - General Idea

Principles of the traditional NLP


| Word-Level | Phrase-Level | Sentence-Level | Discourse-Level |
|---|---|---|---|
| Tokenization | Chunking | Sentence Boundary Detection | TextTiling |
| Morphology / Stemming / Part of Speech Tagging | Chunking / Information Extraction | Parsing | Rhetorical Structure Theory Parsing |
| Thesauri / Word Similarity | Information Extraction | Sentiment Classification / QA / Word Sense Disam. | Summarization/ Categorization / Discourse Analysis |

Table fom Marti Hearst

**Imagine the scenario, which do you think is more likely?**

I want to go outside and take a [_____]

I want to go outside and take a walk  --- looks most likely

I want to go outside and take a picture.

I want to go outside and take a stone --- looks least likely

We want our computer to do this job for us!
How can we teach it?
We teach our machine  --LANGUAGE MODEL

# Language Model for Machines:

At high  level if we want our machines to do the specific tasks on language , they should be able to understand the language.

To that end we create **language models.**
Language models predict the probability distribution of language expressions given a set of vocabulary.

# Language models or the grammar:

Language modeling is the task of estimating likelihood of a **sequence** or **a word** given a sequence.

$$P(x1, x2, x3) = P(x1)\, P(x2 \mid x1)\, P(x3 \mid x2, x1)$$

EXAMPLE:

P("sunday is a boring day") = P(sunday)* P( is|sunday)* P(a|sunday,is)* P(boring|sunday,is,a)* P(day|sunday,is,a,boring)

For longer sentences it becomes very hard to track large dependencies, we use **Markov Assumption and MLE of Probability.**

# N-grams

The Markov Assumption:

The probability of a future event depends only on a limited history of preceding events.

MLE:

$P(w_i | w_1 w_2 \ldots w_{i-1}) = \text{count}(w_1 \ldots w_i) / \text{count}(w_1 \ldots w_{i-1})$

An **n-gram model** is a statistical model of language in which the **previous n−1** words are used to predict the next word.

# Unigram Model

- Likelihood of a word is not dependent on contextual word
- Just multiply the probability of each word to get the probability of a sentence.
- $P(w1w2 \ldots wn) \approx \prod P(wi)$
- EXAMPLE:

  P("sunday is a boring day") = P(sunday)* P( is)* P(a)* P(boring)* P(day)

- P("day|sunday is a boring") = P(day) =1/5

# Bigram Model

- Likelihood of a word is dependent on one preceding word contextual word .
- $P(wi \mid w1w2...wi-1) \approx P(wi \mid wi-1)$
- EXAMPLE:

  P(“day|sunday is a boring ”) = P(day|boring) =C(boring day)/C(boring)

# Unigram and bag Of words Model

Doc1: I love dogs.
Doc2: I hate dogs and knitting.
Doc3: Knitting is my hobby d my passion

**Bag of words:**

| | I | love | dogs | hate | and | knitting | is | my | hobby | passion |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc 1 | 1 | 1 | 1 | | | | | | | |
| Doc 2 | 1 | | 1 | 1 | 1 | 1 | | | | |
| Doc 3 | | | | | 1 | 1 | 1 | 2 | 1 | 1 |

# Information Extraction

If we want the most relevant document to be delivered back we need to find the document that has our query as the signature word.

**IDF: Inverse Document Frequency,** which measures how rare a term is the in the documents.

IDF(term(t) = log_e(Total number of documents / Number of documents with term t in it).

**Instead of using, counts  we use tf-idf weights of terms as features.**

Doc1: I love dogs.
Doc2: I hate dogs and knitting.
Doc3: Knitting is my hobby d my passion

| | I | love | dogs | hate | and | knitting | is | my | hobby | passion |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc 1 | 1 | 1 | 1 | | | | | | | |
| Doc 2 | 1 | | 1 | 1 | 1 | 1 | | | | |
| Doc 3 | | | | | 1 | 1 | 1 | 2 | 1 | 1 |

| | I | love | dogs | hate | and | knitting | is | my | hobby | passion |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc 1 | 0.18 | **0.48** | 0.18 | | | | | | | |
| Doc 2 | 0.18 | | 0.18 | **0.48** | 0.18 | 0.18 | | | | |
| Doc 3 | | | | | 0.18 | 0.18 | **0.48** | **0.95** | **0.48** | **0.48** |

# How we can do  NLP:

At high  level we can have:

- **Rule Based  or Logical methods** -classification and information retrieval
- **Probabilistic Models or Language Model** - QA,Information extraction
  - Documents are ranked based on the probability of the query *Q* in the document's language model.
- **Distributional Approaches: Word2Vec**
  - Assumes: Meaning is related to the context, Words that appear in same context are similar
  - Words are represented as continuous representation or embeddings of corpus vocabulary
    - Requires large corpus to learn the relation between words

# Applications:

| TASK: | SIMPLE SOLUTIONS |
| --- | --- |
| a. Text Classification<br>b. Information Retrieval<br>c. Question Answering<br>d. Information extraction<br>e. Spelling correction<br>f. Machine Translation | BOW or any n-gram using a classifier<br>BOW with link ranking analysis<br>BOW with if-else<br>Information retrieval with rule based methods for association<br>Character n-grams<br>Rule based with POS tagging and semantic matching |