MSc (AI)
Semester II and IV
NLP

UNIT I
Introduction to NLP

Definition of Natural Language

- In simple terms, **a natural language** is a language developed and evolved by humans through natural use and communication, rather than constructed and created artificially, like a computer programming language. They can be communicated in different forms, including speech, writing, symbols or even signs.
- Human languages like Hindi, Marathi, English, Japanese, and Sanskrit …etc are natural languages.
- **Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to hear decision from a dialogue based clinical expert system, etc.**
- Natural language processing(NLP) makes computers learn, understand, analyze, manipulate and interpret natural(human) languages. It is combination of Computer Science, Human languages or **Linguistics,** and Artificial Intelligence.
- The ability of machines to interpret human language is used in every day applications - chatbots, Email classification and spam filters, search engines, grammar checkers, voice assistants, and social language translators.
- The input and output of an NLP system can be Speech or Written Text.

NLP Terminology derived from Linguistics

The origins of linguistics can be dated back to the 4th century BCE, when Indian scholar and linguist Panini formalized the Sanskrit language description.

Linguistics is defined as the scientific study of

- the nature of "meaning" in language, including form and syntax of language, meaning, and
- semantics depicted by the usage of language and context of use.

Different distinctive components of linguistics are

- Phonetics, Phonology
- Syntax, Semantics
- Morphology
- Lexicon, Pragmatics
- Discourse Analysis
- Stylistics, Semiotics

Semantics in NLP is the study of the meaning of words, phrases, and sentences in a language.

NLP Terminology

Phonetics	• It is the study of the **acoustic properties of sounds** produced by the human vocal tract during speech. • It includes studying the properties of sounds as well as how they are created by human beings. • **A smallest individual unit of human speech(sound) in a specific language is called a phoneme or phone.**
Phonology	• It is a study of sound patterns as interpreted in the human mind and used to find the difference between different phonemes to find out which ones are significant. • The structure, combination, and interpretations of phonemes are studied to study the language in detail. • English language consists of around 45 phonemes. • Besides phonemes, Phonology also includes accents, tone, and syllable structures.

NLP Terminology

Morphology	• **A morpheme is the smallest unit of language that has distinctive meaning.** • This includes things like words, prefixes, suffixes, and so on which have their own distinct meanings. • Morphology is the study of the structure and meaning of these distinctive units or morphemes in a language. Specific rules and syntaxes usually govern the way morphemes can combine together.
Lexicon	• It is a study of properties of words and phrases used in a language and how they build the vocabulary of the language. • It include what kinds of sounds are associated with meanings for words, the parts of speech words belong to, and their morphological forms.

Examples

Morphology is the domain of linguistics that analyses the internal structure of words and explores the structure of words

Words are built up of minimal meaningful elements **called morphemes**

Played → play – ed

Cats → cat – s

Unfriendly → un – friend – ly

Two types of morphemes:

1. Stems: play , cat , friend
2. Affixes: -ed , -s , un-, -ly

Two main types of affixes:

a) Prefixes precede the stem: un

b) Suffixes follow the stem: ed s ly

Stemming means finding the stem by stripping off affixes

play = play

replayed = re play ed

computerized = comput er ize d

NLP Terminology

Syntax	• **It is a study of sentences, phrases, words, and their structures.** • It includes researching how words are combined together grammatically to form phrases and sentences. • Syntactic order of words used in a phrase or a sentence matter because the order can change the meaning entirely.
Semantics	• It is a study of meaning in language and can be further subdivided into lexical and compositional semantics. • **Lexical semantics :** The study of the meanings of words and symbols using morphology and syntax. • **Compositional semantics :** Studying relationships among words and combination of words and understanding the meanings of phrases and sentences and how they are related.

NLP Terminology

Pragmatics	• It is a study of how both linguistic and nonlinguistic factors like context and scenario might affect the meaning of an expression of a message or an utterance. • This includes trying to infer whether there are any hidden or indirect meanings in the communication.
Discourse analysis	• Discourse is Study of the structure of larger spans of language. • It deals with how the 'immediately preceding sentence' can affect the 'interpretation of the next sentence' • It analyzes language and exchange of information in the form of sentences across conversations among human beings. • These conversations could be spoken, written, or even signed.
Stylistics	• It is a study of language with a focus on the style of writing, including the tone, accent, dialogue, grammar, and type of voice.
Semiotics	• It is a study of signs, symbols, and sign processes and how they communicate meaning. • Things like analogy, metaphors, and symbolism are covered in this area.

Components and Steps of NLP

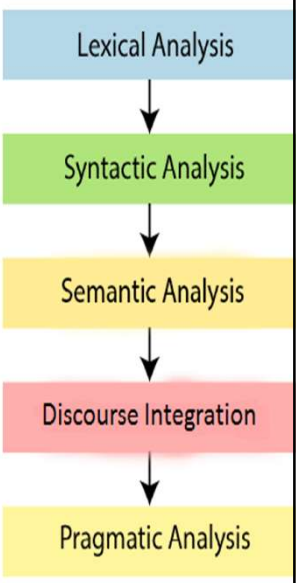
There are two components of NLP,

1. **Natural Language Understanding (NLU),**
   - It involves transforming human language into a machine-readable format
   - It helps the machine to understand and analyze human language by extracting the text from large data such as keywords, emotions, relations, and semantics
2. **Natural Language Generation (NLG)**
   - It acts as a translator that converts the computerized data into natural language representation
   - It mainly involves Text planning, Sentence planning, and Text realization. (The NLU is harder than NLG)

Steps of NLP

1. **Lexical Analysis**

- The first phase of NLP
- It divides the whole text into paragraphs, sentences, and Words.
- In linguistics, the abstract unit of morphological analysis that corresponds to a set of forms taken by a single word is called lexeme.
- Lexeme is a basic unit of meaning.
- The way in which a lexeme is used in a sentence is determined by its grammatical category.
- It scans the source code as a stream of characters and converts it into meaningful lexemes.
- Lexeme can be individual word or multiword.

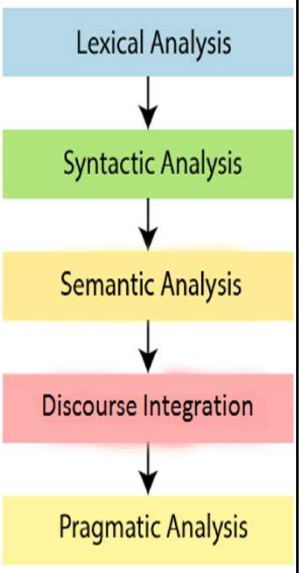


Steps of NLP

- For example, the word talk is an example of an individual word lexeme, which may have many grammatical variants like talks, talked and talking.
- Multiword lexeme can be made up of more than one orthographic word.
- For example, speak up, pull through, etc. are the examples of multiword lexemes.

2. Syntactic Analysis(Parsing)

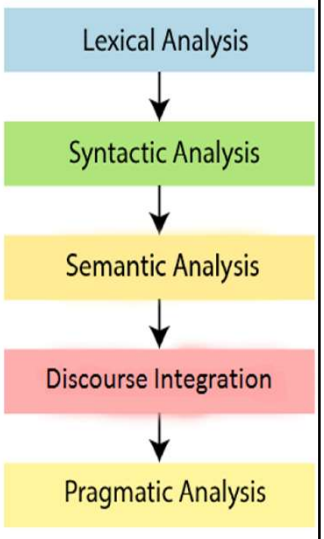
- It is used to check grammar, word arrangements, and shows the relationship among the words
- The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.



Steps of NLP

3. Semantic Analysis

- Semantic analysis is concerned with the meaning representation.
- It mainly focuses on the literal meaning of words, phrases, and sentences
- The semantic analyzer disregards sentence such as "hot ice cream"
- Another Example is "India calls out to Sachin" passes a syntactic analysis because it is grammatically correct sentence.
- However, it fails a semantic analysis, because India is a place (and can not literally call out to people), the sentence's meaning doesn't make sense.



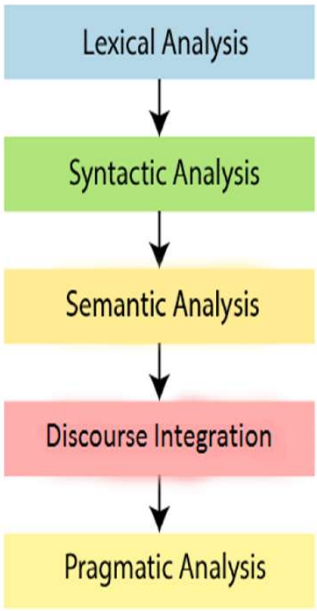
Steps of NLP

4. Discourse Integration

- Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.
- For instance, if one sentence reads, "India speaks to all its people," and the following sentence reads, "It calls out to Sachin," discourse integration checks the first sentence for context to understand that "It" in the latter sentence refers to India.

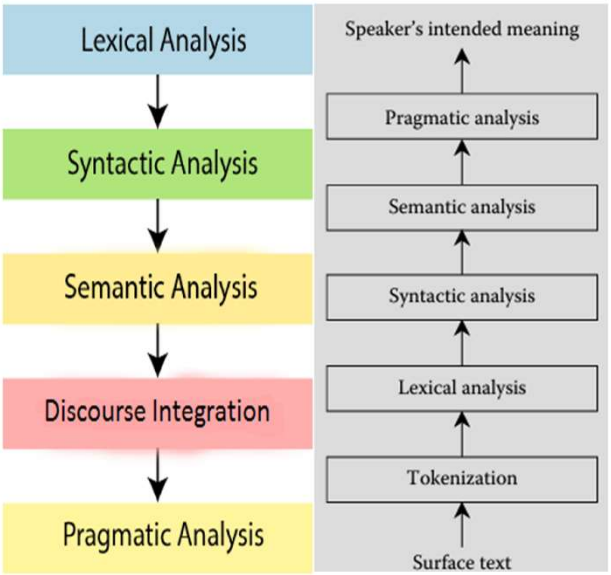
5. Pragmatic Analysis

- Here, what was said is reinterpreted on what it actually meant.
- It involves deriving those aspects of language which require real world knowledge.

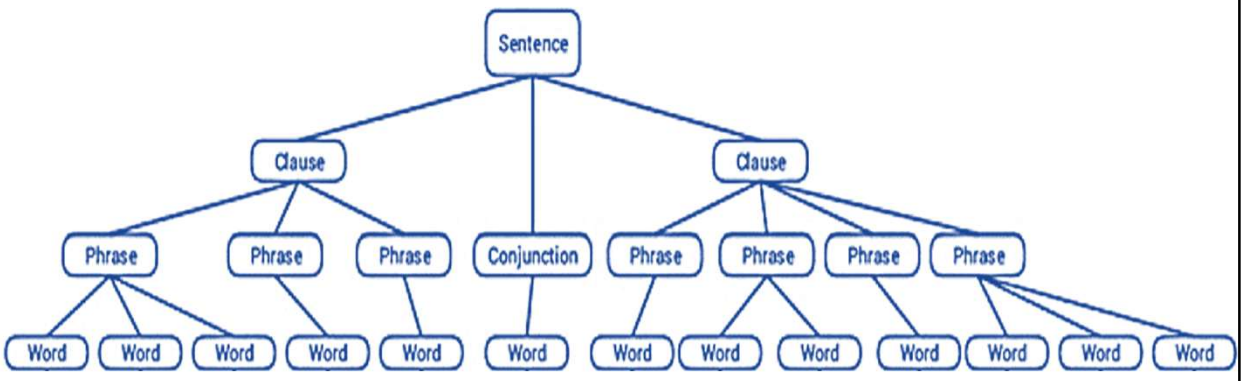


Steps of NLP

- Example: "Open the door" is interpreted as a request instead of an order.
- For instance, a pragmatic analysis can uncover the intended meaning of "India speaks to all its people."
- A pragmatic analysis deduces that this sentence is a metaphor for how people emotionally connect with place.



Language Structure



Words

- Words are the smallest units in a language that are independent and have a meaning of their own.
- Although morphemes are the smallest distinctive units, morphemes are not independent like words, and a word can be madeuo of several morphemes.
- It is useful to annotate and tag words and analyze them into their parts of speech (POS) to see the major syntactic categories.
- Different methods are used to generate POS tags programmatically.

The main categories and significance of the various POS tags are given as follows:

Words

N(oun) : This usually denotes words that depict some object or entity which may be living or nonliving.

- E.g. fox , dog , book , and so on.
- The POS tag symbol for nouns is N .

V(erb) : Verbs are words that are used to describe certain actions, states, or occurrences.

- There are a wide variety of further subcategories, such as auxiliary, reflexive, and transitive verbs (and many more).
- E.g. running , jumping , read , and write .
- The POS tag symbol for verbs is V .

Words

Adj(ective) : Adjectives are words used to describe or qualify other words, typically nouns and noun phrases.

- The phrase beautiful flower has the noun (N) flower which is described or qualified using the adjective (ADJ) beautiful .
- The POS tag symbol for adjectives is ADJ .

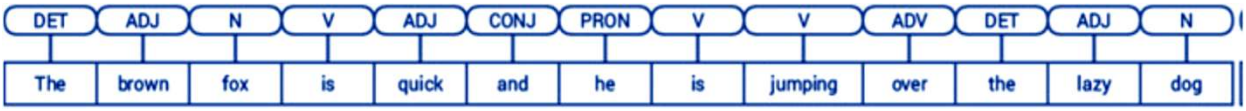
Adv(erb) : Adverbs usually act as modifiers for other words including nouns, adjectives, verbs, or other adverbs.

- The phrase very beautiful flower has the adverb (ADV) very , which modifies the adjective (ADJ) beautiful , indicating the degree to which the flower is beautiful.
- The POS tag symbol for adverbs is ADV .

Words

Besides the above four major categories of parts of speech , there are other categories that occur frequently in the English language.
They include pronouns, prepositions, interjections, conjunctions, determiners, and many others.
Furthermore, each POS tag can be further subdivided into categories. E.g. the noun (N) is subdivided into singular nouns (NN), singular proper nouns (NNP), and plural nouns (NNS).
Considering previous example sentence **(The brown fox is quick and he is jumping over the lazy dog)** where we built the hierarchical syntax tree, we can annotate it using basic POS tags as follows:

Annotated words with their POS tags



Examples : The slow tortoise is persistent and he is crossing the sleeping lazy rabbit.

The - Det
slow - Adj
tortoise - Noun (N)
is - Verb (V)
persistent - Adj

and - Conjuction (Conj)

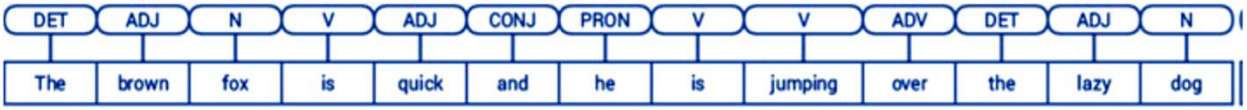
He - (Pronoun) PRON
is - Verb (V)
crossing - Verb (V)
the - Determinet
sleeping - verb (Det) (V)
lazy - Adj
rabbit. - Noun (N)

Words

Here the tag

- DET stands for determiner, which is used to depict articles like a , an , the , and so on.
- CONJ indicates conjunction, which is usually used to bind together clauses to form sentences.
- PRON tag stands for pronoun , which represents words that are used to represent or take the place of a noun.
- N, V, ADJ and ADV are typical open classes and represent words belonging to an open vocabulary.
- Open classes are word classes that consist of an infinite set of words and commonly accept the addition of new words to the vocabulary which are invented by people.

Words



- Words are usually added to open classes through processes like morphological derivation, invention based on usage, and creating compound lexemes.
- Some popular nouns added recently include Internet and multimedia.
- Closed classes consist of a closed and finite set of words and do not accept new additions.
- Pronouns are a closed class.
- Words have their own lexical properties like parts of speech.
- Using these words, we can order them in such a way that they give meaning to the words such that each word belongs to a corresponding phrasal category and one of the words is the main or head word.

Phrases

- Groups of words make up phrases , which form the third level in the syntax tree.
- Phrases are assumed to have at least two or more words, considering the order : words ← phrases ← clauses ← sentences.
- However, a phrase can be a single word or a combination of words based on the syntax and position of the phrase in a clause or sentence.
- For example, the sentence "Movie is good" has only three words, and each of them can be treated as three phrases.
- The word **Movie** is a noun as well as a noun phrase, **is** depicts a verb as well as a verb phrase, and **good** represents an adjective as well as an adjective phrase describing **Movie.**

Phrases

There are five major categories of phrases:

Noun phrase (NP) :

- These are phrases where a noun acts as the head word.
- Noun phrases act as a subject or object to a verb.
- Usually a noun phrases can be a set of words that can be replaced by a pronoun.
- For example: *movie*, *the lazy dog*, and *the brown fox*

Verb phrase (VP) :

- These phrases are lexical units that have a verb acting as the head word.

Usually there are two forms of verb phrases.

- One form has the verb components as well as other entities such as nouns, adjectives, or adverbs as parts of the object.

Phrases

- The verb here is known as a *finite verb*.
- It acts as a single unit in the hierarchy tree and can function as the root in a clause. This form is prominent in *constituency grammars*.
- The other form is where the finite verb acts as the root of the entire clause and is prominent in *dependency grammars*.

Adjective phrase (ADJP) :

- These are phrases with an adjective as the head word. Their main role is to describe or qualify nouns and pronouns in a sentence, and they will be either placed before or after the noun or pronoun.
- The sentence *The cat is too quick* has an adjective phrase, *too quick* , qualifying *cat*, which is a noun phrase.

Phrases

Adverb phrase (ADVP) :

- These phrases act like adverbs since the adverb acts as the head word in the phrase.
- Adverb phrases are used as modifiers for nouns, verbs, or adverbs themselves by providing further details that describe or qualify them.
- In the sentence *The train should be at the station pretty soon* , the adjective phrase *pretty soon* describes when the train would be arriving.

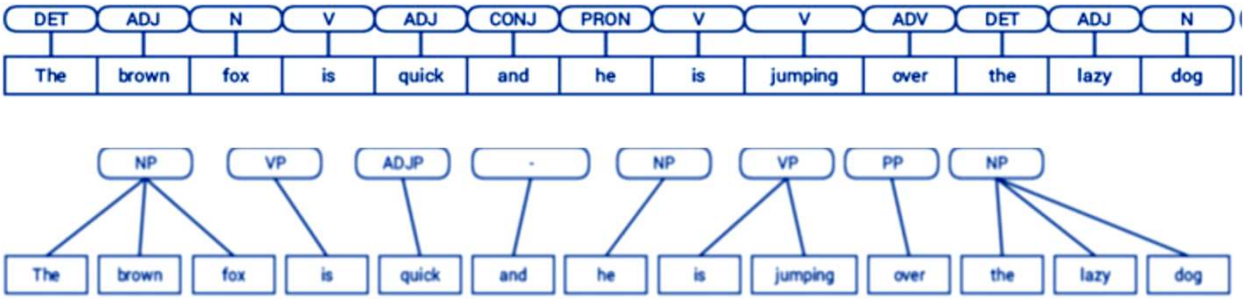
Prepositional phrase (PP) :

- These phrases usually contain a preposition as the head word and other lexical components like nouns, pronouns, and so on.
- It acts like an adjective or adverb describing other words or phrases.
- The phrase *going up the stairs* contains a prepositional phrase *up* , describing the direction of the stairs.

These five major syntactic categories of phrases can be generated from words using several rules.

Shallow parsing is a popular natural language processing technique to extract these constituents, including POS tags as well as phrases from a sentence.

For example for the sentence *The brown fox is quick and he is jumping over the lazy dog*, we have obtained seven phrases from shallow parsing, as shown in Figure



Annotated phrases with their tags

Clauses

- A clause is a group of words with some relation between them and can act as independent sentence,

e.g. The white dress is very beautiful.

The white dress is more graceful than the red one.

- Several clauses can be combined together to form a sentence.

e.g. The white dress is very beautiful and is more graceful than the red one.

- Clauses consist of **two parts**:

1. **The main clause or independent clause which** can form a sentence by itself and act as both sentence and clause.
2. **The subordinate or dependent clause** cannot exist just by itself and depends on the main clause for its meaning.

They are usually joined with other clauses using dependent words such as conjunctions.

Clauses

Example:

The brown fox is quick	**and**	**he is jumping over the lazy dog**
Main clause	conjunction	Subordinate Clause

With regard to syntactic properties of language, clauses can be subdivided into several categories based on syntax:

Declarative :

- These are just standard statements, which are declared with a neutral tone and which could be factual or non-factual.
- They occur quite frequently and denote statements having no specific tone associated with them.
- For example - Grass is green. Sky is blue. Mountains are big.

Clauses

Imperative :

- These clauses are usually in the form of a request or command or rule, or advice.
- The tone in this case would be a person issuing an order (request/instruction) to one or more people to carry out that order(request/instruction).
- For example - Please do not talk in class. Please pick up the phone. Please do the needful.

Relative :

- Relative clauses are subordinate clauses and hence dependent on another part of the sentence that usually contains a word, phrase, or even a clause.
- They act as the **antecedent(precursor)** to one of the words from the relative clause and relates to it.

Clauses

- For example: Consider a clause "Nilesh just mentioned that he wanted a water-bottle".
- Here the antecedent proper noun **Nilesh**, referred as **he** in the relative clause **he wanted a water-bottle**.

Interrogative :

- These clauses usually are in the form of questions.
- The questions can be either affirmative or negative.
- For example: Did you get my mail? Didn't you go to college?

Exclamative :

- These clauses are used to express shock, surprise, or even compliments.
- These expressions fall under exclamations, and these clauses often end with an exclamation mark.
- For example: What an amazing match!

Assignment 1: Applications of NLP

Gr.	Topic
1.	Machine Translation
2.	Speech Recognition Systems
3.	Question Answering Systems
4.	Contextual Recognition and Resolution
5.	Text Summarization
6.	Text Categorization
7.	Email filtering
8.	Smart assistants
9.	Document analysis
10.	Social media monitoring

Instructions

1. You will be assigned a group.
2. Prepare a group presentation on the topic given to you.
3. Presentation can include a brief introduction, how NLP used in the application, demonstration in Python.
4. Date of Presentation: **11.02.2025**