Unit I: Introduction to Natural Language Processing (NLP)

1. Introduction to Natural Language Processing

Natural Language Processing (NLP) is a branch of **Artificial Intelligence (AI)** that focuses on enabling computers to **understand, interpret, and generate human language** such as English, Marathi, Hindi, etc.

NLP lies at the intersection of:

- Computer Science
- Artificial Intelligence
- Linguistics
- Machine Learning

Examples of NLP Applications

- Machine Translation (Google Translate)
- Chatbots and Virtual Assistants
- Sentiment Analysis
- Speech Recognition
- Text Summarization
- Question Answering Systems

---

2. Why NLP is Hard?

NLP is difficult because **human language is complex and ambiguous**.

Major Reasons

1. **Ambiguity**
    - *Lexical ambiguity*: Same word, multiple meanings
      Example: *bank* (river bank / money bank)
    - *Syntactic ambiguity*: Sentence structure confusion
      Example: *I saw the man with a telescope*
2. **Context Dependency**
    - Meaning changes based on context
      Example: *He is running* (machine / person)
3. **Variability of Language**
    - Synonyms, slang, abbreviations, dialects
4. **Implicit Information**
    - Humans understand hidden meanings easily, machines don’t
5. **World Knowledge Requirement**
    - Understanding needs real-world knowledge

---

3. Programming Languages vs Natural Languages

Programming Languages	Natural Languages
Artificially designed	Naturally evolved
Unambiguous	Highly ambiguous
Strict grammar rules	Flexible grammar
Limited vocabulary	Very large vocabulary
Easy for machines	Difficult for machines

Example
Programming: `if x > 5:` → clear meaning
Natural language: *He saw her duck* → unclear meaning

---

4. Are Natural Languages Regular?

Natural languages **are NOT regular languages**.

Reason

- Regular languages can be handled by **Finite Automata**
- Natural languages have:
  - Nested structures
  - Long-distance dependencies
  - Agreement rules (subject-verb)

Example:

The boy who is standing near the tree is my friend.

Such structures cannot be fully captured using regular expressions.

---

5. Finite Automata for NLP

Finite Automata (FA) are used in **limited NLP tasks**, mainly at the **lexical level**.

Applications

- Token recognition
- Morphological analysis
- Pattern matching
- Lexical analysis

Limitations

- Cannot handle complex syntax
- Cannot represent deep linguistic structures

Hence, FA is useful only for **simple NLP problems**.

---

6. Stages of NLP

NLP processing is divided into multiple stages:

1. Lexical Analysis

- Breaking text into words (tokens)
- Removing punctuation

2. Morphological Analysis

- Analyzing word structure
- Root word identification

3. Syntactic Analysis (Parsing)

- Grammar checking
- Sentence structure analysis

4. Semantic Analysis

- Meaning of sentence
- Word sense disambiguation

5. Discourse Analysis

- Understanding relation between sentences

6. Pragmatic Analysis

- Understanding intent using context

---

7. Challenges and Issues (Open Problems) in NLP

1. **Ambiguity Resolution**
2. **Sarcasm and Irony Detection**
3. **Low-resource languages**

4. **Code-mixed language processing**
5. **Context and commonsense reasoning**
6. **Multilingual NLP**
7. **Bias and fairness in language models**

Basics of Text Processing

8. Tokenization

Tokenization is the process of **breaking text into smaller units called tokens**.

Types

- **Word Tokenization**
- **Sentence Tokenization**
- **Sub-word Tokenization**

Example:
Text: *I love NLP*
Tokens: `I`, `love`, `NLP`

---

9. Stemming

Stemming reduces words to their **root form** by removing suffixes.

Characteristics

- Fast
- Rule-based
- Root may not be a valid word

Example:

- Running → Run
- Studies → Studi

---

10. Lemmatization

Lemmatization reduces words to their **dictionary base form (lemma)**.

Characteristics

- Uses vocabulary and grammar
- Slower than stemming
- Produces meaningful root words

Example:

- Running → Run
- Better → Good

---

11. Stemming vs Lemmatization

Stemming	Lemmatization
Rule-based	Dictionary-based
Faster	Slower
May produce invalid words	Produces valid words
Less accurate	More accurate

---

12. Part of Speech (POS) Tagging

POS Tagging assigns **grammatical tags** to words.

Common POS Tags

- Noun (NN)
- Verb (VB)
- Adjective (JJ)
- Adverb (RB)
- Pronoun (PRP)

Example:
Sentence: *She is reading a book*

- She → Pronoun
- is → Verb
- reading → Verb
- book → Noun

Exam Tip (SPPU Pattern) ✍️

- Write **definitions + examples**
- Draw **stage diagram of NLP**
- Include **comparison tables**

- Use **simple real-life examples**