
📘 Module 1 — Introduction to Natural Language Processing

🧠 1. Introduction to Natural Language Processing

Definition:
Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language.
It bridges human communication and machine understanding.

Goal:
To make computers process natural languages (like English, Hindi, etc.) in a meaningful way.

Example Applications:

- Chatbots and virtual assistants (e.g., Alexa, Siri).
- Google Translate and speech recognition.
- Sentiment analysis of reviews and social media.
- Automatic summarization and question answering.

NLP combines:

- Linguistics ➔ understanding grammar and meaning.
- Computer Science ➔ algorithms and data processing.
- Statistics / ML ➔ learning from data and patterns.

🧩 2. Key Terminologies in NLP

Term	Meaning	Example
Corpus	A large collection of text used for training or evaluating NLP models.	Wikipedia articles, tweets, or news datasets.
Token	The smallest meaningful unit of text (word, punctuation, etc.).	“NLP is fun!” ➔ [‘NLP’, ‘is’, ‘fun’, ‘!’]

Term	Meaning	Example
Tokenization	Splitting text into tokens (words or sentences).	“I love NLP.” → [“I”, “love”, “NLP”]
Stem	The crude root form of a word after removing suffixes.	“running”, “runs”, “ran” → “run” (sometimes “runn”).
Lemma	The correct dictionary (base) form of a word.	“studies” → “study”, “better” → “good”.
Stop Words	Common words with little meaning in text analysis.	“the”, “is”, “in”, “and”, “to”.
POS Tag	Part-of-speech tag (noun, verb, adjective, etc.) assigned to each word.	“The dog runs” → The/DT, dog/NN, runs/VBZ.
Dictionary / Vocabulary	The unique set of valid words in the corpus.	{“apple”, “banana”, “fruit”}.

📊 3. Empirical Rules in NLP

Empirical rules are data-driven linguistic patterns derived from large corpora, rather than manually defined grammar rules.

🔹 Key Ideas:

- Word Frequency Distributions:
  A few words occur very often, most occur rarely — Zipf’s Law.
  Example: “the” is much more frequent than “beautiful”.
- Co-occurrence Patterns:
  Words appearing together often have contextual relationships.
  Example: “doctor” and “hospital” co-occur more often.
- Statistical Modeling:
  Modern NLP uses probability and statistics instead of handcrafted grammar.

Comparison:

Aspect	Rule-Based NLP	Empirical NLP
Based On	Handwritten grammar and lexicon	Data and statistical models
Advantage	Linguistically precise	Scalable and adaptive
Example	Syntax rules (S → NP VP)	Word2Vec, Transformer models

---

⚙️ 4. Why NLP is Hard

Challenge	Explanation	Example
Ambiguity	Same sentence or word can have multiple meanings.	"I saw her duck." → (bird or action?)
Context Dependence	Meaning depends on situation or nearby words.	"He sat by the bank."
Idioms & Figurative Language	Non-literal usage is hard for machines.	"Kick the bucket" = "die".
Morphological & Syntactic Complexity	Word forms and grammar vary across languages.	"run", "running", "ran".
Multilingual Variation	Word order and structure differ by language.	English (SVO) vs Hindi (SOV).
Sarcasm & Emotion	Hidden tone is difficult for algorithms.	"I just love getting stuck in traffic."

💡 5. Why NLP is Useful

Use Case	Description
Human–Computer Communication	Enables chatbots, virtual assistants, and voice systems.
Automation of Text Tasks	Summarization, translation, question answering.
Information Extraction	Finds key facts from large text datasets.

Use Case	Description
Sentiment and Trend Analysis	Understands opinions from reviews, social media.
Search and Recommendation	Improves search engines and content personalization.

🔁 6. NLP Processing Pipeline

The NLP workflow moves step-by-step from raw text to machine understanding.

Stage	Description
1. Text Collection	Gather text data (from files, web, APIs, etc.).
2. Text Preprocessing	Clean and normalize text (tokenization, stop-word removal, etc.).
3. Feature Extraction	Convert text into numerical form (BoW, TF-IDF, embeddings).
4. Model Building	Train models for tasks like classification or tagging.
5. Evaluation	Assess accuracy, precision, recall, F1-score.
6. Deployment	Integrate model into real-world applications.

🧹 7. Corpus Cleaning Techniques

Raw text often contains noise — punctuation, numbers, symbols, or inconsistent casing.
Cleaning improves data quality for NLP models.

🧩 7.1 Word Tokenization

Splits text into individual words.
Example (Python NLTK):

```
from nltk.tokenize import word_tokenize

word_tokenize("NLP makes machines intelligent.")

# ['NLP', 'makes', 'machines', 'intelligent', '.']
```

🧩 7.2 Sentence Tokenization

Splits a paragraph into sentences.

```
from nltk.tokenize import sent_tokenize

sent_tokenize("AI is powerful. It helps computers learn.")

# ['AI is powerful.', 'It helps computers learn.']
```

🧩 7.3 Word Frequency Distribution

Counts occurrences of each word in a corpus.

```
from nltk import FreqDist

tokens = ["NLP", "is", "fun", "NLP", "is"]

FreqDist(tokens).most_common(2)

# [('NLP', 2), ('is', 2)]
```

Usage: Keyword extraction, topic modeling, text summarization.

🧩 7.4 Stemming

Removes suffixes to get the crude base form of a word.
Rule-based, may produce non-words.

```
from nltk.stem import PorterStemmer

ps = PorterStemmer()

ps.stem("studies")  # 'studi'

ps.stem("running")  # 'run'
```

Example: “playing”, “played” → “play”.

🧩 7.5 Lemmatization

Reduces words to their valid dictionary (lemma) form using grammar.
More accurate than stemming.

```
from nltk.stem import WordNetLemmatizer

lm = WordNetLemmatizer()

lm.lemmatize("studies")  # 'study'

lm.lemmatize("running", pos='v')  # 'run'
```

Difference:

Feature	Stemming	Lemmatization
Basis	Rules	Dictionary (WordNet)
Output	May not be valid	Always valid
Example	studies → studi	studies → study

🧩 7.6 Dictionary / Vocabulary

A dictionary or lexicon is the list of unique valid words used for processing.
Example:
Corpus: "NLP makes machines intelligent."
Vocabulary = {NLP, makes, machines, intelligent}

🏷️ 8. Introduction to Part of Speech (POS) Tagging

Definition:
POS tagging assigns each word a grammatical category such as noun, verb, adjective, etc.
This helps in understanding sentence structure and meaning.

Example:

Sentence: "The quick brown fox jumped over the lazy dog."
Tags:
The/DT quick/JJ brown/JJ fox/NN jumped/VBD over/IN the/DT lazy/JJ dog/NN

Common Tags:

Tag	Meaning	Example
NN	Noun	cat
VB	Verb	eat
JJ	Adjective	beautiful
RB	Adverb	quickly
IN	Preposition	on
DT	Determiner	the

📜 9. Textual Preprocessing Techniques

Text preprocessing standardizes raw text before feature extraction or modeling.

Technique	Description	Example
Lowercasing	Converts all characters to lowercase.	"NLP IS FUN" → "nlp is fun".
Stop Word Removal	Removes common meaningless words.	"This is an NLP example" → ["NLP", "example"].
Regular Expressions (Regex)	Removes unwanted symbols, numbers, etc.	"Email me at test123@gmail.com!" → "Email me at testgmailcom".
Text Standardization	Replaces informal abbreviations or slang.	"bcoz" → "because", "gr8" → "great".

Example Combined Flow:

1. Lowercase → 2. Remove punctuation → 3. Tokenize → 4. Remove stop words → 5. Lemmatize

Final Clean Output Example:

Input: "I'm happy bcoz NLP is gr8!"

→ Output: ['i', 'am', 'happy', 'because', 'nlp', 'is', 'great']

🧾 10. Summary

Step	Technique	Purpose
Text Cleaning	Remove noise, lowercase	Normalize text
Tokenization	Split text	Build vocabulary
Stop Word Removal	Filter out common words	Reduce redundancy
Stemming / Lemmatization	Normalize word forms	Reduce dimensionality
POS Tagging	Identify grammatical roles	Structural understanding

---

✅ In short:
NLP converts messy human text into structured data through cleaning, tokenization, normalization, and tagging — enabling machines to perform intelligent language-based tasks.
