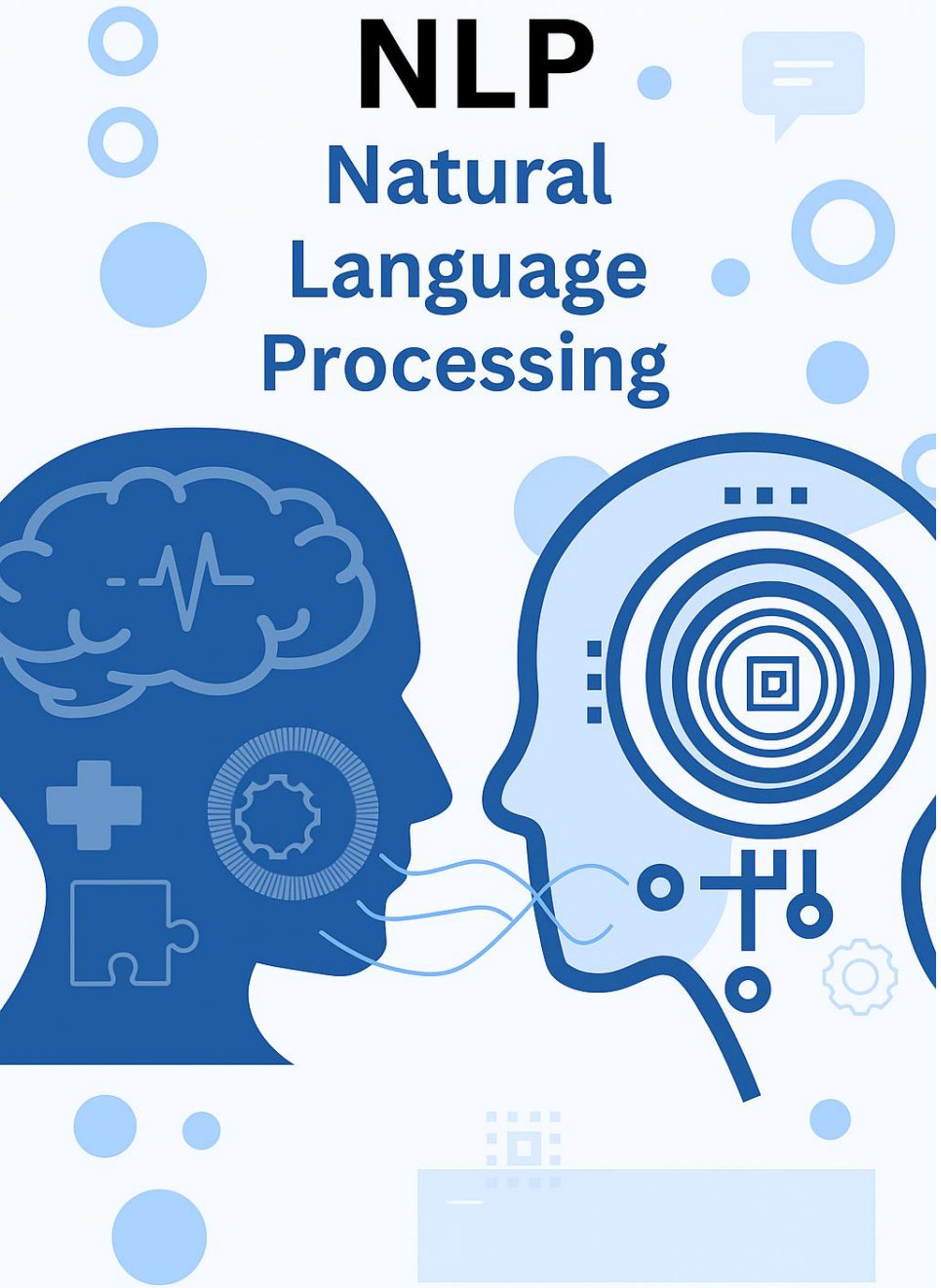

NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 1

LECTURE 3

Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT

AMBIGUITIES

- Lexical Ambiguity
- Syntactic Ambiguity
- Semantic Ambiguity
- Pragmatic Ambiguity
- Anaphoric (Referential) Ambiguity

LEXICAL AMBIGUITY

- Meaning: A single word has multiple meanings.

- Example: “Bank” → could mean a riverbank or a financial institution.

Ambiguous Word	Possible Meanings	Example Sentences
Bank	1. Financial institution 2. River edge	- "I deposited money in the bank." - - "They sat on the river bank."
Bat	1. Flying animal 2. Sports equipment	- "The bat flew out of the cave." - - "He hit a six with the bat."
Seal	1. Animal 2. To close something tightly	- "The seal clapped its flippers." - - "Seal the envelope, please."
Pitch	1. Throw 2. Sales talk 3. Musical note	- "He made a great sales pitch." - - "The pitcher threw the pitch."
Light	1. Not heavy 2. Brightness	- "This bag is light." - - "Turn on the light."
Right	1. Correct 2. Direction 3. Legal claim	- "You're right." - "Turn right." - - "You have a right to speak."
Rock	1. Stone 2. Genre of music 3. Sway	- "He threw a rock." - - "I love rock music." - - "The boat began to rock."
Well	1. In good health 2. A water source	- "She is doing well." - - "They dug a well."
Date	1. A calendar day 2. A romantic meeting 3. Fruit	- "What's today's date?" - - "He went on a date." - - "I ate a date."
Watch	1. To observe 2. A timepiece	- "Watch the road!" - - "He looked at his watch."

SYNTACTIC AMBIGUITY

- Meaning: Sentence structure allows multiple interpretations.

- Example: “I saw the man with the telescope.”

→ Did I use the telescope, or did the man have it?

SYNTACTIC AMBIGUITY

- **"She watched the man on the hill with the binoculars."**
- **Meaning 1:** She used binoculars to watch the man who was on the hill.
- **Meaning 2:** She watched the man who was on the hill and had the binoculars.
- **Meaning 3:** She was on the hill, watching the man with binoculars.

SYNTACTIC AMBIGUITY

- **“Visiting relatives can be annoying."**
- **Meaning 1:** The act of visiting relatives is annoying.
- **Meaning 2:** Relatives who visit can be annoying.

SEMANTIC AMBIGUITY

- Meaning: Sentence meaning is unclear, even if structure is correct.

- Example: “The chicken is ready to eat.”

→ Is the chicken going to eat, or be eaten?

SEMANTIC AMBIGUITY

- **"He saw her duck."**
- **Meaning 1:** He saw the woman lower her head quickly (verb: duck).
- **Meaning 2:** He saw the duck that belonged to her (noun: duck).

PRAGMATIC AMBIGUITY

- Meaning: Depends on speaker's intention or context.

- Example: "Can you open the door?"

→ Literally asking for ability, but meant as a request.

PRAGMATIC AMBIGUITY

- **"Do you know what time it is?"**
- **Meaning 1:** A question about your knowledge of the time.
- **Meaning 2:** A polite way of asking for the current time.

ANAPHORIC (REFERENTIAL) AMBIGUITY

- Meaning: Uncertainty in what a pronoun refers to

- Example: “Rita told Sita that she won.”

→ Who won?

ANAPHORIC (REFERENTIAL) AMBIGUITY

- **"When Sarah met Priya, she was very nervous."**
- **Who is "she"?**
  - Sarah was nervous.
  - Priya was nervous.

- **"Ravi called Arjun while he was driving."**
- **Who was driving?**
  - Ravi could be driving.
  - Arjun could be driving.

VARIETIES IN NATURAL LANGUAGE

- Language Diversity
- Dialects and Regional Diversity
- Code-Mixing and Code-Switching
- Social Diversity
- Styles and Registers
- Temporal Diversity
- Evolving Language

LANGUAGE DIVERSITY

Structural and Grammatical Variations

- English follows Subject-Verb-Object: *"She eats rice."*
- Japanese follows Subject-Object-Verb: *"She rice eats."* (*Kanojo wa gohan o tabemasu.*)

Script and Writing Systems:

- Different languages use different writing systems: Latin (English), Devanagari (Hindi), Cyrillic (Russian), Hanzi (Chinese).
- Some languages (like Arabic or Hebrew) are written right-to-left, while others are left-to-right.

DIALECTS AND REGIONAL DIVERSITY

Pronunciation (Accent)

- The way words are pronounced varies greatly across regions.
- Example: *“Schedule”* pronounced as /ˈʃedjuːl/ (UK) vs. /ˈskedʒuːl/ (US)
- Tamil spoken in Chennai vs. Coimbatore has noticeable accentual differences.

Vocabulary

- Different regions use distinct words for the same object or concept.
- UK: *Lift*, *biscuit*, *flat*
- US: *Elevator*, *cookie*, *apartment*
- India (English): *Prepone* (not standard in US/UK English)

DIALECTS AND REGIONAL DIVERSITY

Sentence Formation

- Dialects may alter sentence structure or verb usage.
- Standard English: *He doesn't have any money.*
- African American Vernacular English (AAVE): *He don't got no money.*

Same Word Different Meaning

- The same word might have different meanings in different dialects.
- *Chips* in UK means **French Fries or Potato Wedges** in US English.

CODE-MIXING AND CODE-SWITCHING

Code-Mixing

- The blending of words, phrases, or morphemes from one language into another within the same sentence or utterance.
- “I am going to bazaar for some shopping.”

Code-Switching

- The practice of shifting between two languages or dialects depending on the context, audience, or topic.
- “I can’t attend the meeting today, kal mera exam hai.”

WHY USE CODE-MIXING AND CODE-SWITCHING?

- **Ease of Expression**: Some ideas are more naturally or effectively expressed in one language than another. Example: “I’m not feeling well, *mann nahi lag raha hai*.”
- **Social Identity and Belonging**: Reflects group membership, bilingual fluency, or cultural belonging. Used to build rapport or show solidarity.
- **Filling Lexical Gaps**: When a word doesn't exist or is hard to recall in one language. Example: “I went to the *mandap* and it was beautifully decorated.”
- **Stylistic or Emphatic Purposes**: Used for emphasis, humor, or dramatic effect. Example: “This movie was so boring, *pura time barbaad ho gaya!*”

PROBLEMS IN NLP

- **Language Detection Issues**: Identifying which part of a sentence belongs to which language is non-trivial. Example: Tokenization fails if the script changes (e.g., English and Devanagari).
- **Lack of Annotated Datasets**: Code-mixed corpora are limited, especially for low-resource language pairs.
- **Syntax Ambiguity**: Mixed grammatical structures confuse parsers and language models.
- **Speech Recognition Errors**: Code-mixed speech may confuse voice assistants.
- **Machine Translation Difficulties**: Translating code-mixed text into a single target language while preserving meaning is complex.

NLP TECHNIQUES TO HANDLE CODE-MIXING AND CODE-SWITCHING

- **Language Identification (LangID) at token level**: Classifies each word based on its language.
- **Transliteration modules**: To handle romanized text (e.g., “namaste” instead of “नमस्ते”).
- **Joint embeddings**: Represent multilingual tokens in a shared semantic space.
- **Transfer learning from multilingual pre-trained models**: Like mBERT, XLM-R, IndicBERT.
- **Development of code-mixed corpora**: Projects like GLUECoS and LINCE benchmark code-mixed NLP.

SOCIAL DIVERSITY

- Language varieties influenced by social factors like class, education, profession, age, or ethnicity.
- Youth slang vs formal adult speech.
- Professional jargon used by doctors or lawyers.
- People from different socioeconomic backgrounds may speak differently.
- Highly educated individuals may use more complex or standardized forms of a language while less formally educated speakers might rely on more colloquial or regional expressions.
- Research suggests that men and women may use language differently in terms of politeness, intonation, or topic preference

STYLES AND REGISTER

Style

- Unique language style or usage specific to an individual.
- Examples**:** A poet’s lyrical style vs a scientist’s technical tone

Register

- Changes in language depending on the context or situation.
- Formal register (e.g., academic writing, speeches)
- Informal register (e.g., chats with friends)

TEMPORAL DIVERSITY

Lexical Change (Vocabulary Evolution)

- New words are introduced (neologisms), old ones fall out of use (archaisms).
- Old English: *“hwaet”* (listen!) → obsolete
- Modern: *selfie*, *emoji*, *internet*, *ghosting*

Semantic Shift (Meaning Change)

- Word meanings broaden, narrow, or shift entirely.
- *Nice* → originally meant “ignorant” in Middle English, now means “pleasant.”
- *Awful* → once meant “full of awe,” now means “terrible.”

TEMPORAL DIVERSITY

Phonological Change (Pronunciation Shift)

- Sounds change over time due to ease of articulation or social influence.
- The **Great Vowel Shift** (15th–17th century) drastically altered English pronunciation.

Grammatical Change

- Syntax, word endings, and sentence structure evolve.
- Old English: *Ic geseah hine* (I saw him)
- Modern English: *I saw him* (word order became more fixed)

Orthographic Change (Spelling and Writing)

- Spellings become standardized or change due to technology and reform.
- Example: *Musick* → *music*; *publick* → *public*

EVOLVING LANGUAGE

- **Cultural Shifts**: Social movements, pop culture, and global events influence vocabulary and usage. Example: *Woke*, *cancel culture*, *climate emergency*.
- **Technology and the Internet**: Social media platforms like Twitter, TikTok, and Reddit accelerate language change. Emojis, hashtags, abbreviations, and memes become integral to communication. Example: *DM me*, *hashtag goals*, *LOL*, *ROFL*, *#ThrowbackThursday*.
- **Youth and Generational Trends**: Younger generations often innovate slang and digital communication styles. Example: *Ghosting*, *FOMO* (Fear of Missing Out), *Yeet*, *Slay*.
- **Globalization and Language Contact**: Words from one language are borrowed or adapted into another. Example: *Pizza* (Italian) used worldwide; *Guru or Karma* (Sanskrit) in English.

PROBLEMS IN NLP

- **Rapid Vocabulary Expansion**: New words and abbreviations emerge quickly, often without formal definitions. NLP models trained on older corpora fail to recognize or interpret new terms.
- **Non-standard Grammar and Spelling**: Internet language often breaks traditional grammar rules. Example: *I'm sooo tired rn lol* (contains elongated spelling, abbreviation "rn" for "right now").
- **Informal and Multimodal Communication**: Use of emojis, GIFs, and memes adds layers of non-textual meaning. Example: "I'm fine 😐" might express sarcasm or frustration - not neutrality.
- **Contextual and Dynamic Meaning**: Some words shift meanings based on current events or subcultures. Example: *Karen* originally a name, now slang for an entitled, demanding person (often in a meme format).

NLP TECHNIQUES TO HANDLE EVOLVING LANGUAGE

- **Continual Model Updates**: Regularly updating language models with recent corpora (e.g., Twitter, Reddit).
- **Social Media-Aware Embeddings**: Training word embeddings on social media text (e.g., GloVe-Twitter).
- **Fine-tuning on Domain-Specific Data**: Custom models trained on chat, gaming, or youth-slang corpora.
- **Context-Aware and Sentiment Models**: To better capture sarcasm, tone, and informal usage.

CHALLENGES IN NLP

- Multilingual Model Development
- Context Understanding
- Data Scarcity for Low-Resource Languages
- Ambiguity Resolution
- Sarcasm and Irony
- World Knowledge and Common Sense

MULTILINGUAL MODEL DEVELOPMENT

- Building models that understand and translate hundreds of languages is complex.

- A model trained on standard English may fail to understand regional variants or dialectal constructions.

DATA SCARCITY

- Many languages lack sufficient digital data for AI training. Example: Indigenous or tribal languages in India, Africa, and South America.

- Lack of annotated corpora for many regional and minority languages. Hence, very difficult to apply supervised learning methods without sufficient training examples.

OPPORTUNITIES AND PROGRESS

- Multilingual NLP frameworks like mBERT, XLM-RoBERTa, and IndicNLP are being developed to bridge language gaps.
- UNESCO and local governments support initiatives to digitize and preserve endangered languages.
- Advances in zero-shot and few-shot learning enable systems to understand new languages with minimal data.

WORLD KNOWLEDGE AND COMMON SENSE

- “She dropped the glass on the floor. What will happen to the glass?”
- “John put the ice in the oven. What will happen to the ice?.”