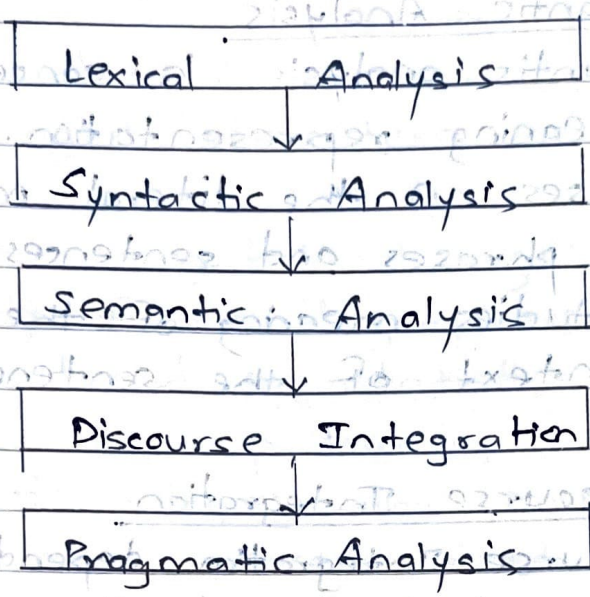
Q.B. P.T - I

Q.1 Explain about the phases of NLP?

⇒

- There are five phases of NLP :-



1) Lexical Analysis

⇒ In this phase ~~the~~ scans the source code as a stream of characters.

- Then it converts into meaningful lexemes.
- It divides the whole text into paragraphs, sentences and words.

2) Syntactic Analysis

⇒ Syntactic analysis is used to check grammer, word arrangements, and shows the relationship among the words.

- Hence the words are collected to form phrases, phrases gets converted to clauses and clauses form sentences.
- It shows relationship among words.
- For e.g., Pune goes to gopal.

The above sentence does not make any

sense, so this sentence is rejected by the syntactic analyzer.

3) Semantic Analysis

⇒ Semantic analysis is concerned with the meaning representation.

- It focuses on the literal meaning of the words, phrases and sentences.
- It studies meaning of the words independent of context of the sentences.

4) Discourse Integration

⇒ Discourse Integration depends upon the sentences that proceeds it and also involve the meaning of the sentences that follow it.

- It connects sentences.

5) Pragmatic Analysis

⇒ Pragmatic analysis is the last phase of NLP. It helps one to discover the intended effect by applying a set of rules that characterstics cooperative dialogues.

- It is mainly concerned with how the sentences are used and what the inner meaning of the sentence

Q.2 What is Natural language processing? Explain generic NLP system.

⇒

- NLP is a field of computer science and a subfield of artificial intelligence that aims to ~~find~~ make computers understand human language.
- NLP uses computational linguistics, which is the study of how language works, and various models based on statistics, machine learning and deep learning.
- These technologies allow computers to analyze and process text or voice data, and to grasp their full meaning.

Q.3 What are different ambiguities which needs to be handled by NLP?

⇒

- Ambiguity is the capability of being understood in more than one way.
- Natural language is very ambiguous.
- NLP has the following types of ambiguity :-

1) Lexical Ambiguity

⇒ The ambiguity of single word is called lexical ambiguity..

- There are two forms of lexical ambiguity :-

a) Polysemy : This is when two words are the same but have a different meaning depending on usage.

b) Homonym : This occurs when a word has the same spelling or pronounciation, but has different meanings overall.

2) Syntactic Ambiguity

⇒ This kind of ambiguity occurs when a sentence is parsed in different ways.

- For e.g., the sentence "The man saw girl with the telescope".
- It is ambiguous whether the man saw girl carrying a telescope or he saw her through his telescope.

3) Semantic Analysis

⇒ This kind of ambiguity occurs when the meaning of the words themselves can be misinterpreted.

- In other words, semantic ambiguity happens when a sentence contains an ambiguous word or phrase.
- For e.g., the sentence "The car hit the pole while it was moving." is having the semantic ambiguity because the interpretations can be "The car, while moving, hit the pole" and "The car hit the pole while the pole was moving."

4) Anaphoric ambiguity

⇒ This kind of ambiguity arises due to the use of anaphora entities in discourse.

- For example, the horse ran up the hill. It was very steep. It soon got tired.
- Here, the anaphoric reference of "it" in two situations cause ambiguity.

5) Pragmatic ambiguity

⇒ Such kind of ambiguity refers to the situation where the context of a phrases gives it multiple interpretations.

- In simple words, we can say that pragmatic ambiguity arises when the statement is not specific.

Q. 4 Why handling ambiguities important in NLP applications?

⇒

- Ambiguity, generally used in NLP, can be referred as the ability of being understood in more than one way.
- In simple terms, we can say that ambiguity is the capability of being understood in more than one way.
- Natural language is very ambiguous. NLP has the following types of ambiguities -
  1) Lexical
  2) Syntactic
  3) Semantic
  4) Anaphoric
  5) Pragmatic
- The key reasons why resolving ambiguities is crucial :-

1) Ensuring accurate communication

⇒ Ambiguities can lead to misinterpretation of meaning.

- For NLP applications like virtual assistant or chatbots, accurately understanding the user's intent is critical for providing the correct response or action.

2) Enhancing user experience

⇒ Users expect NLP systems to understand their inputs accurately and respond

appropriately.
- Resolving ambiguities ensures that interactions are smooth and natural, leading to higher user satisfaction and trust in the system.

3) Precision in Information Retrieval
⇒ In search engines, question-answering systems, and other information retrieval applications, resolving ambiguities ensures that the results are precise and relevant, avoiding irrelevant or incorrect information.

4) Improving contextual relevance
⇒ Ambiguities often arise from the lack of content or multiple possible interpretations of a phrase or sentence.
- Handling these ambiguities allow NLP system to better grasp the context, leading to more relevant and accurate outputs.

Q.5 What are the components of NLP?

⇒

- There are two components of NLP :-

1) Natural Language Understanding (NLU)

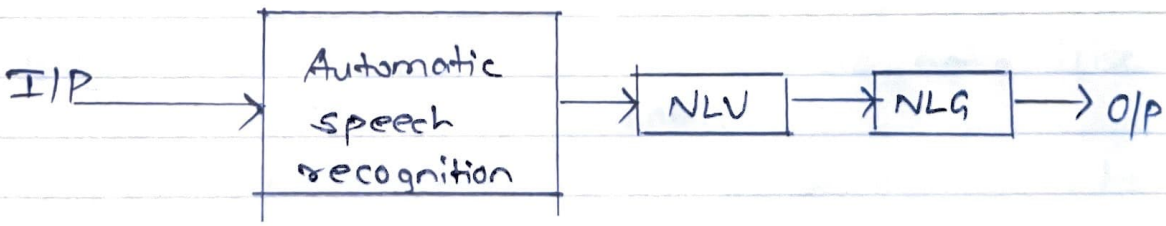
⇒ NLU helps the machine to understand and analyze human language by extracting the metadata from content such as concepts, entities, keywords, emotions and semantic roles.

- NLU is mainly used in Business applications to understand the customer's problem in both spoken and written language.
- NLU involves the following tasks :-
  a) It is used to map the given input into useful representation.
  b) It is used to analyze different aspects of the language.

2) Natural Language Generation

⇒ NLG acts as translator that converts the computerized data into natural language representation.

- It mainly involve Text planning, Sentence planning and Text realization.



Q.6 What are the applications of NLP?

⇒

- The applications of NLP are :-

1) Machine translation

⇒ In machine translation, the translation of the text in one human language to another human language is performed automatically.

- For performing the translation, it is important to have the knowledge of the words, and phrases, grammers of two language that are involved in Translation, semantics of the languages and the knowledge of the words.

2) Speech recognition

⇒ Speech recognition is the process where the acoustic speech signals are mapped to the set of words.

- As there is a wide variation in the pronounciation of the word, honoinym for example, sea and see, acoustic ambiguities like in the rest and interest.

3) Sentiment Analysis

⇒ Sentiment analysis uses NLP to interpret and analyze emotions in subjective data like news articles and tweets.

- Positive, negative and neutral opinions can be identified to determine a customer's

sentiment towards a brand, product, or service.

4) Chatbots

=> Chatbots are programs used to provide automated answers to common customer queries.

- They have pattern recognition system with heuristic response, which are used to hold conversations with humans.
- AI-powered chatbots are designed to handle more complicated request making conversational experiences increasingly original.

5) Automatic Text Summarization

=> Automatic text summarization is the task of condensing a piece of text to a shortest version.

- It extracts its main ideas and preserving the meaning of content.
- This application of NLP is used in news headlines result snipers in web search and buttentins of market reports.

Q.7 What are challenges of NLP?

⇒

- Challenges in NLP are :-

1) Language differences

⇒ If we speak english and if we are thinking of reaching an international and/or multicultural audience, we shall need to provide support for multiple languages.

2) Training Data

⇒ NLP is all about analysing languages to better understand it.

- One must spend years constantly to become fluent in language.
- One must spend a significant amount of time reading, listening to, and utilising a language.
- The abilities of NLP system depends on the training data provided to it.

3) Development time

⇒ One also must think about the development time for an NLP system.

- With a distributed deep learning mode and multiple GPUS working in coordination, one can trim down the training time to just a few hours.

4) Misspelling

=> Miscpelling are a simple problem for human beings, but for a machine, misspelling can be harder to identify.

- One should use an NLP tool with capabilities to recognize common misspelling of words, and move beyond them.

5) Words with multiple meaning

=> Most of the languages have words that could have multiple meanings, depending on the context.

- For e.g., a user who asks "how are you" has a totally different goal than a user.

6) Phrases with multiple intentions

=> Some phrases and questions actually have multiple intentions, so the NLP system cannot over simplify the situation by interpreting only one of those intentions.

- For eg., a user may prompt the Chatbot with something like, "I need to cancel any previous order and update my card on file."
- The AI needs to be able to distinguish these intentions separately.

Q.8 Write short note on Indian Language Processing.

⇒

- One might think that people who are aquainted with computers are already familiar with the English interface.
- However, it's worth nothing that majority of the Indian population in India is still based in rular areas where teaching and learning would be in local languages, where communities are literate, but still are not familiar with English.
- So, yes, it is a worthwhile effort to upscale NLP research in India.
- The challenges in Indian Language Processing:

1) Script diversity

⇒ Indian languages uses various scripts, such as Devanagari (Hindi, Marathi), Bengali script, Tamil script, etc.

- This diversity poses challenge for text encoding and processing.

2) Word Order Flexibility

⇒ Indian languages often exhibit free word order, which complicates syntactic parsing and machine translation.

3) Morphological Complexity

⇒ Many Indian languages are morphologically rich, meaning words can have numerous

inflected forms.
- This makes tasks like stemming, lemmatization and part-of-speech tagging more challenging

- Applications:

1) NLP has huge application in helping people with disabilities - interpretation of sign language, text to speech, speech to text, etc.
2) Digitisation of Indian Manuscripts to preserve knowledge contained in them.
3) Signboard translation from Vernacular Languages to make travel more accessible.

Q.9 What is morphological analysis?

⇒

- Morphological analysis is the process of studying and analyzing the structure of words to understand their formation and the relationship between different word forms.
- It aims to break down words into their constituent parts such as roots, prefixes and suffixes and understand their roles and meanings.
- Morphological analysis is a critical step in NLP for several reasons:

1) Understanding word formation

⇒ It helps in identifying the basic building blocks of words, which is crucial for language comprehension.

2) Improving text analysis

⇒ By breaking down words into their roots and affixes, it enhances the accuracy of text analysis tasks like sentiment analysis and topic modeling

3) Facilitate Multilingual Processing

⇒ Morphological analysis provides detailed insights into word formation

It aids in handling the morphological diversity of different languages, making NLP system more robust and versatile.

- Key techniques used in Morphological Analysis :-

1) Stemming

⇒ Stemming reduces words to their base or root form, usually by removing suffixes.

- The resulting stems are not necessarily valid words but are useful for text normalization.
- For e.g., the stem of these three words, connections, connected, connects is "connect".

2) Lemmatization

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

- It considers the context and part of speech, providing valid words.
- For e.g.,

studying
studies
study

Lemmatization →

study
study
study

- Applications :

1) Facilitates accurate translation by understanding and generating correct word forms in different language.

2) Information Retrieval

⇒ Enhances search engines by improving the matching of query terms with relevant documents, even if they are in different morphological forms.

Q.10 What is significance of FSA in morphological analysis?

⇒

- Finite State Automata (FSA) play a significant role in morphological analysis, especially in tasks that involve recognizing and generating valid words forms in natural languages.
- An automation having a finite ~~state or~~ number of states is called Finite Automata or FSA.
- Mathematically, an automation can be represented by 5-tuple $(Q, \Sigma, \delta, q_0, F)$

Where:

Q : Finite set of states
Σ : Finite set of symbols
δ : transition function
q_0 : Initial state
F : Final state

- Significance

1) Efficient word recognition

⇒ FSAs can be used to recognize and validate morphological patterns in words.

- For e.g., FSA can be constructed to recognize different forms of verbs like "walk", "walked", "walking", etc. by encoding the suffix rules.

2) Morphological Parsing

⇒ FSAs can be used in morphological parsers to break down complex words into their constituent morphemes.
- This is done by traversing the automaton to identify prefixes, roots, suffixes and other morphological components.
- For e.g., For the word "unhappiness" an FSA could parse it into "un" (prefix), "happy" (root) and "-ness" (suffix).

3) Modelling Morphological rules

⇒ FSAs are well-suited for modelling the rule-based aspects of morphology, such as inflectional or derivational process.
- For e.g., An FSA can represent the rule that adding "-ing" to a verb creates its present participle form

4) Scalability and Efficiency

⇒ FSAs are computationally efficient and can process large volumes of text quickly, making them suitable for real-time application in NLP.
- They are particularly effective for regular and predictable morphological patterns

that occur frequently in natural language.

Q.11 Explain Tokenization, Stemming and Lemmatization.

⇒

- **Tokenization**

⇒ Tokenization is the process of dividing a text into smaller units known as tokens.

- Tokens are typically words or sub-words in the context of NLP.
- There are two types of tokenization:

a) **Word tokenization**: Splitting the text into words based on a certain delimiter.

- For e.g., "I love NLP!" becomes

["I", "love", "NLP", "!"]

b) **Sentence tokenization**

⇒ Splitting the text into sentences.

- For e.g., "I love NLP! It's Facinating."

["I love NLP!", "It's Facinating."]

c) **Character tokenization**

⇒ Character tokenization splits a piece of text into a set of characters.

- It overcomes the drawbacks of the word tokenization.

- **Stemming**

⇒ Stemming reduces words to their base or root forms, usually by removing suffixes.

- The resulting stem are not necessarily valid words but are useful for text normalization.
- There are two kinds of stemming algorithm:

a) **Porter Stemmer**: One of the most popular stemming algorithms, which uses a set of rules to iteratively reduce words to their stems.

b) **Snowball Stemmer**

⇒ An improvement on the Porter stemmer with more aggressive stemming rules.

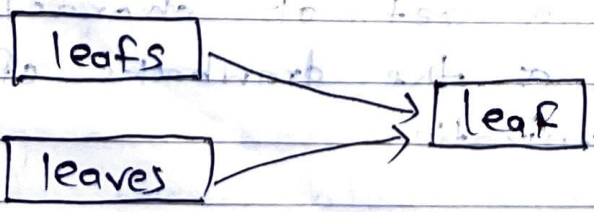
- For e.g.,

The words connections, connected, connects might all be stemmed to "connect" using stemming.

- **Lemmatization**

⇒ Lemmatization reduces words to their base or dictionary forms (lemma).

- It considers the context and part-of-speech producing valid words.



- For e.g.,

studying	Lemmatization →	study
studies		study
study		study

- Lemmatization has application in:

1) Biomedicine: Using lemmatization to parse biomedicine literature may increase the efficiency of data retrieval tasks.
2) Search engines.

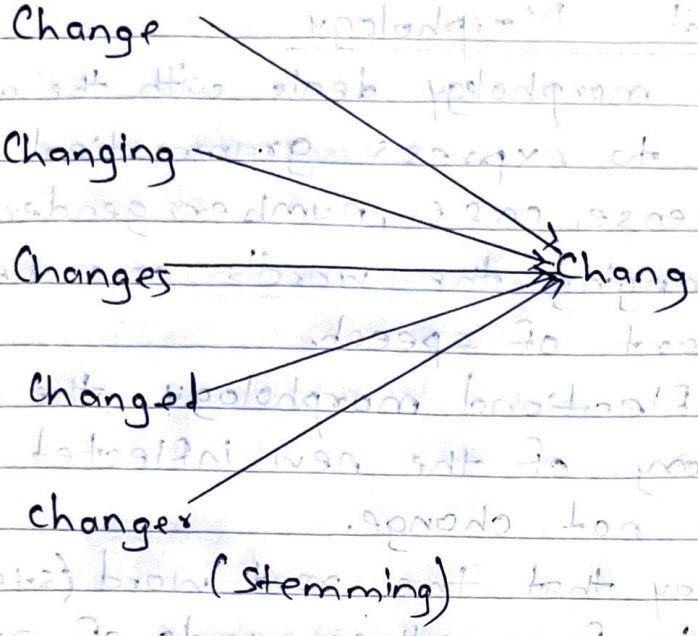
Q. 12 Compare and contrast Lemmatization and stemming.

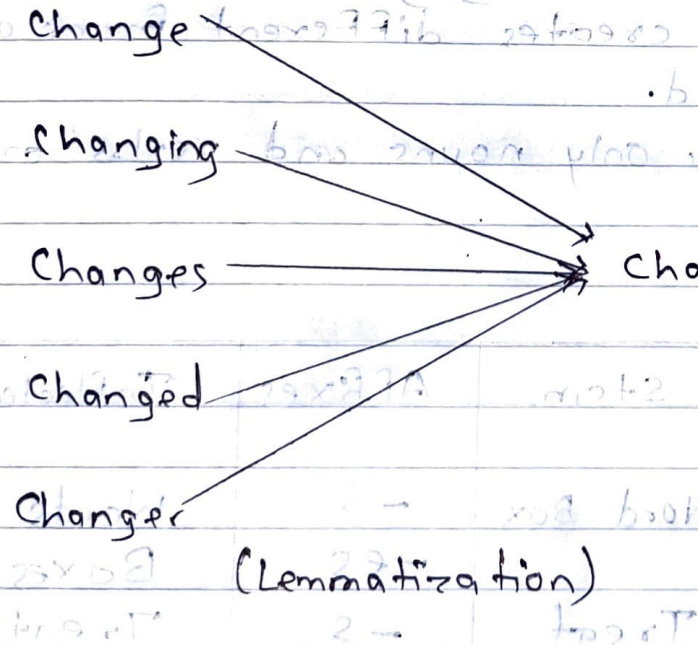
⇒

Stemming	Lemmatization
1) Stemming attempts to reduce inflectional form of each word into a common base or root.	1) Lemmatization also attempts to reduce inflectional form of each word into a common base or root.
2) In stemming the end or beginning of a word is cut off, keeping common prefixes and suffixes.	2) Lemmatization uses dictionaries to conduct a morphological analysis of the word and link it to the lemma.
3) Stemming tends to be faster process because it chops words without knowing the context of word in sentence.	3) Lemmatization is slow process, it is because it knows the context of the word before processing.
4) Stemming is a rule-based approach.	4) Lemmatization is a dictionary based approach
5) The process of stemming has a lower degree of accuracy.	5) The process of lemmatization has comparatively higher degree of accuracy.
6) Typically language independent with simple rule sets.	6) Language dependent, requires linguistic knowledge and resources like dictionary.

For e.g.,

Stemming vs Lemmatization:





Q.13 Explanation on Inflectional and derivational morphology.

⇒

- **Inflectional Morphology**

⇒ Inflection morphology deals with the modification of a word to express grammatical information such as tense, case, number, gender, etc. without changing the word's core meaning or its part of speech.

- Due to inflectional morphology, the meaning and category of the new inflected words usually do not change.
- One can say that the root word (stem) is inflected to form other words of same meaning and category.
- Inflection creates different forms of the same word.
- In English, only nouns and verbs can be inflected.
- For e.g.,

Category	Stem	Affixes	Inflected Word
Noun	Word Box	-s	Words
		-es	Boxes
Verb	Treat	-s	Treats
		-ing	Treating
		-ed	Treated

- **Derivational Morphology:**

⇒ Derivational morphology involve the creation of new words by adding affixes to the base word.

- This process often changes the meaning of the word and may also change its part of speech.
- One of the most common ways to derive new words is to combine derivational affixes with root words.
- The purpose of derivation is to create new words with different meanings or different syntactic roles, allowing for greater lexical variety.
- Less productive than inflection; derivational rules are more specific to certain words.
- For e.g.;

Category	Stem	Affixes	Derived word	Target category
Noun	Vapour	-ize	Vaporize	Verb
Verb	Read	-er	Reader	Noun
Adjective	Real	-ize	Realize	Verb
Noun	Mouth	-ful	Mouthful	Adjective

Q.14 What is significance of porter stemmer in NLP applications?

⇒

- The porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English.

Significance:

1) Text Normalization

⇒ In NLP applications, text data often contains ~~various~~ variations of the same word, which need to be normalize to a common form.

- The porter stemmer helps by reducing words to their base form, thus simplifying the text data and reducing redundancy.
- For e.g., running, ran, runs to their base form "run".

2) Information Retrieval Improved

⇒ In information retrieval system, stemming allows for better matching of queries with documents.

- For e.g.,

a user searching for "running" should ideally retrieve documents that contains related term like "run" or "ran".

- The porter stemmer helps achieve this by reducing ~~their~~ words to their base forms.

3) Reducing Dimensionality

=) In NLP tasks like text classification, sentiment analysis or topic modelling, the dimensionality of the feature space can be very high.

- Stemming reduces the number of unique words by grouping words variations under a single root form.

4) Efficient and Lightweight

=) The porter stemmer is an efficient and lightweight algorithm.

- It uses simple, rule-based steps to reduce words to their stems, making it computationally inexpensive and fast to implement.

Q.15 What is regular expression? What is the significance in word level analysis?

⇒

- A regular expression (RE) is language for specifying text search strings.
- RE helps us to match or find other strings or sets of strings, using a specialized syntax held in a pattern.
- Regular expression are used to search texts in UNIX as well as in MS WORD in identical way.
- Mathematically, a RE can be defined as follows :-

i) ε is a RE,
ii) $\emptyset$ is a RE,
iii) IF X and Y are RE, then
   a) X, Y
   b) X.Y
   c) X+Y
   d) X*, Y* are also a RE
   e) If a string is derived from above rules then that would also be a RE.

- For e.g.,

RE	Regular set
(0 + 10*)	{0, 1, 10, 100, 1000, 10000, ...}
(0+ε)(1+ε)	{ε, 0, 1, 01}
(a+b)*	{ε, a, ab, b, aa, bb, ba, ...}
(11)*	{ε, 11, 1111, 1111, ...}

- In word level analysis, RE are highly significant.
- They allow you to perform various tasks related to text manipulation and analysis at the word level.

1) Pattern Matching

⇒ Regular expressions enable searching for specific patterns within text, such as finding all instances of dates, email or phone numbers.
- For instance, if you're searching for all occurences of email addresses in a document, you can define a regex that matches email address.

2) Text Extraction

⇒ RE enable you to extract specific pieces of information from text.
- For e.g., you can extract all the dates mentioned in a document by using a RE that identifies date formats.

3) Tokenization

⇒ Tokenization is a process of splitting text into individual units, often words or phrases.
- RE can help tokenize text by identifying spaces, punctuation or other delimeter.

Q.16 What is difference between free and bound morpheme?

⇒

Free Morpheme	Bound Morpheme
1) A morpheme that can stand alone as a word.	1) A morpheme that cannot stand alone and must be attached to another morpheme.
2) Can function independently as a word.	2) Depends on another morpheme to form a word.
3) Carries full meaning by itself.	3) Adds or changes the meaning of the word when attached to a free morpheme.
4) Typically consists of root words.	4) Includes Affixes and some roots.
5) Can represents nouns, verbs, adjectives,	5) Alter the form's or function of base word.
6) Can form a word on its own.	6) Must combine with a free morpheme to form a complete word.
7) For e.g., book, run, happy, cat	7) For e.g., "-ed" (walked), "-ing" (walking), "-s" (cats), "-un" (undo)

Q.17 Explain the following terms

1) Morpheme

⇒

- A morpheme is the smallest grammatical unit in a language that carries meaning.
- It can be word or part of word, such as a prefix, root or suffix.
- Morphemes are the building blocks of words, and they can't be further divided into smaller meaningful units without losing their meaning.
- Morphemes can be classified into two main types :-

a) Free Morpheme : Can stand alone as a word. (e.g., "cat", "run").

b) Bound Morpheme : Cannot stand alone and must be attached to another morpheme. (e.g., "-ed" (walked)).

- For e.g.,

"unhappiness" consists of three morphemes : ("un-") bound, "happy" (free) and "-ness" (Bound).

2) Morphotactics

⇒

- Morphotactics refers to the rules and constraints that govern the permissible arrangement of morphemes within a word in a particular language.
- It dictates how morphemes can be combined to form valid words.
- These rules determines the order, ~~management~~ arrangement, and structure of morphemes within words.
- For e.g.,

In English, the word "unhappily" follows morphotactic rules:
prefix ("un-") + root ("happy") + suffix ("-ly").

- However, "happilyun" would violate English morphotactics rules.

3) Orthographic Rules

⇒

- Orthographic rules are general rules. They are used when breaking a word into its stem and modifiers.
- These rules includes guidelines for spelling, punctuation, capitalization, and other aspects of written language.
- Different languages have their own orthographic rules, which can vary widely based on historical, phonological, and grammatical factors.
- Consider an example: singular English words ending with -y, when it is pluralised, it ends with -ies.
- For e.g.,
  In English, a common orthographic rule is that the letter 'i' comes before 'e' except after 'c'.
  E.g. "receive", "believe".