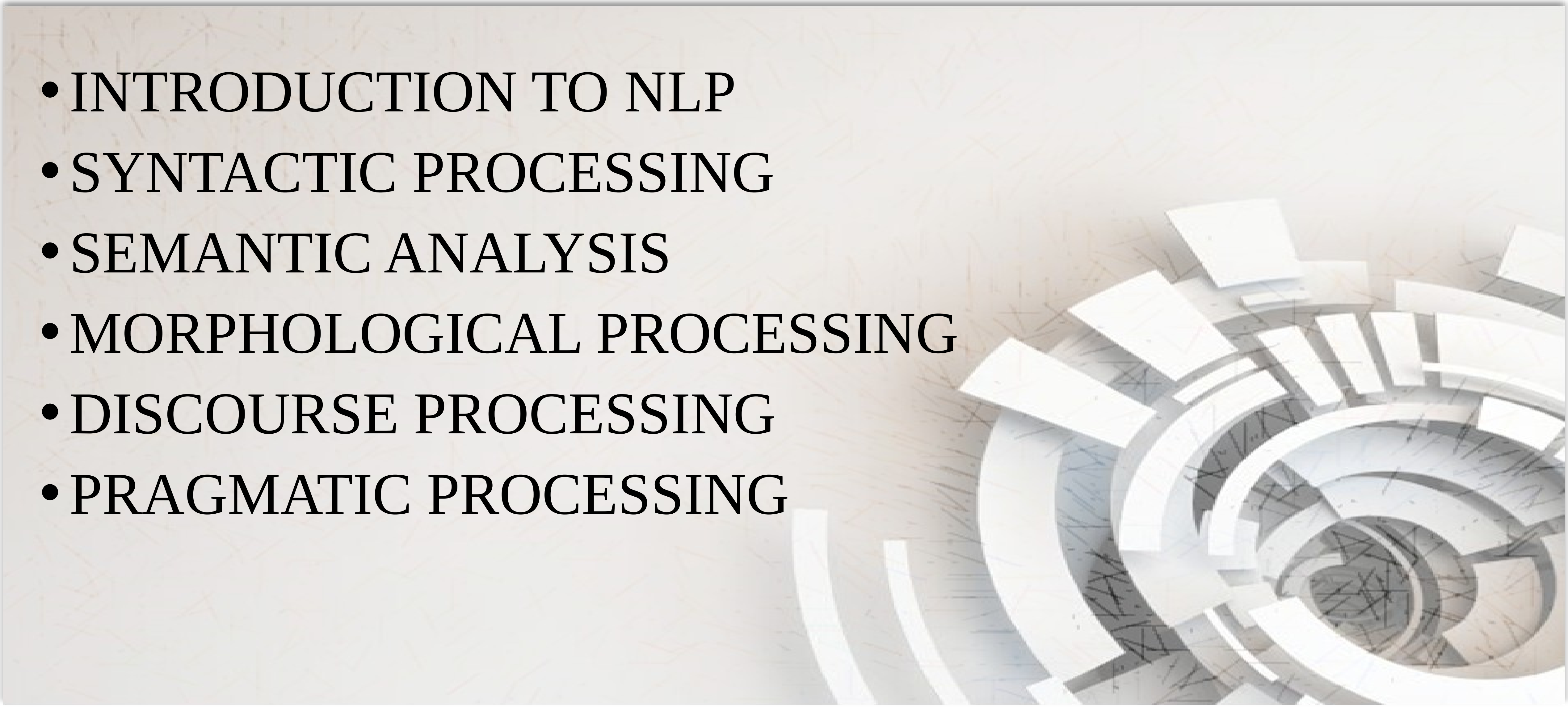
Syntactic, semantic, morphological, discourse &Pragmatic processing

CONTENTS

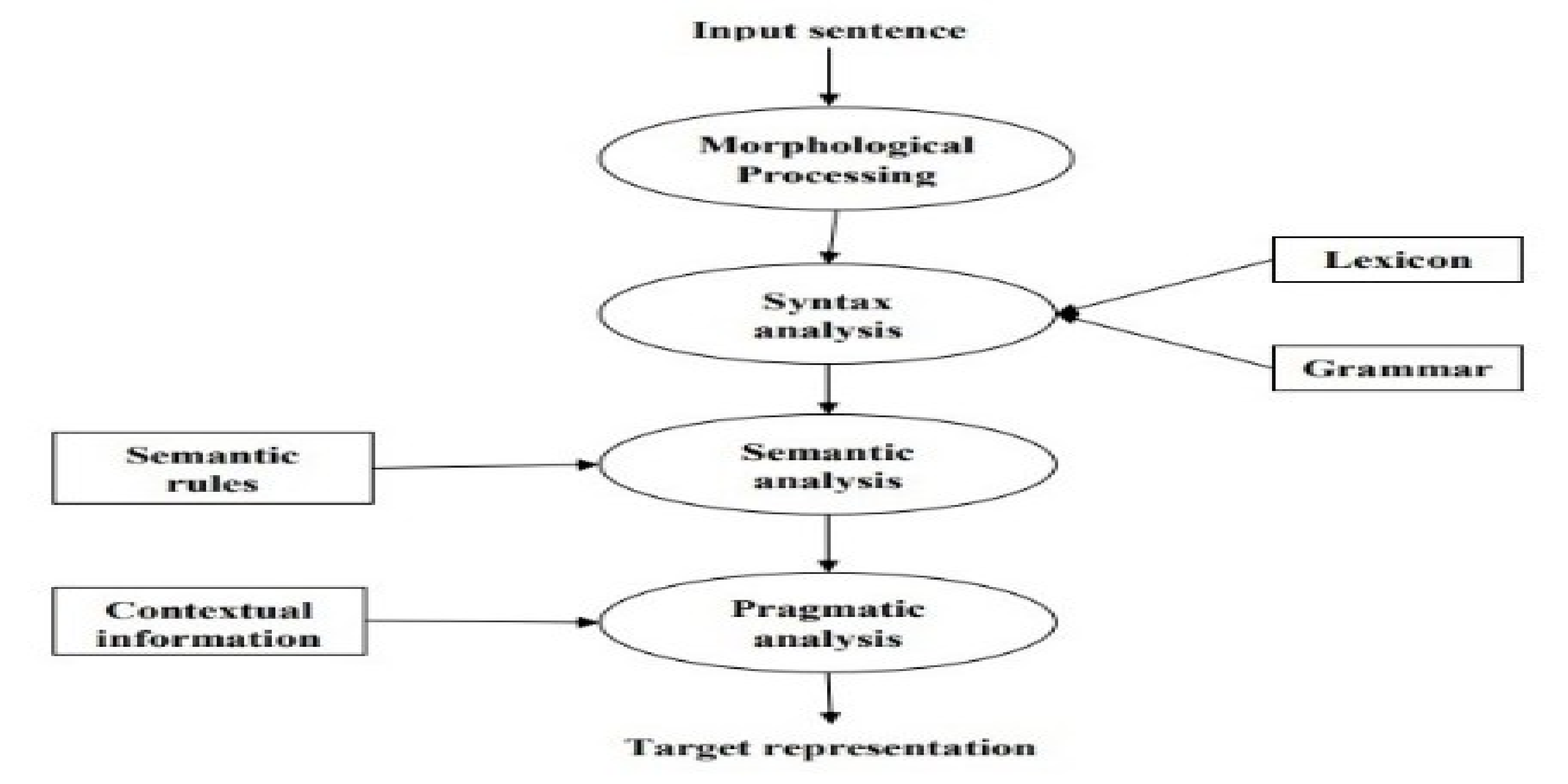


- INTRODUCTION TO NLP
- SYNTACTIC PROCESSING
- SEMANTIC ANALYSIS
- MORPHOLOGICAL PROCESSING
- DISCOURSE PROCESSING
- PRAGMATIC PROCESSING

INTRODUCTION

- Language is a method of communication with the help of which we can speak, read and write.
- Natural Language Processing (NLP) is a subfield of Computer Science that deals with Artificial Intelligence (AI), which enables computers to understand and process human language.
- Technically, the main task of NLP is to program computers for analyzing and processing huge amount of natural language data.

NLP PHASES


Input sentence
Morphological Processing
Lexicon
Syntax analysis
Grammar
Semantic rules
Semantic analysis
Contextual information
Pragmatic analysis
Target representation

CONTD…

- **MORPHOLOGICAL PROCESSING :**

The purpose of this phase is to break chunks of language input into sets of tokens corresponding to paragraphs, sentences and words.

For example, a word like **“uneasy”** can be broken into two sub-word tokens as **“un-easy”**.

CONTD…

- **SYNTAX ANALYSIS :**

The purpose of this phase is two folds:

- to check that a sentence is well formed or not and
- to break it up into a structure that shows the syntactic relationships between the different words.

For example, the sentence like **“The school goes to the boy”** would be rejected by syntax analyzer or parser.

CONTD…

- **SEMANTIC ANALYSIS :**

The purpose of this phase is to draw exact meaning, or you can say dictionary meaning from the text. The text is checked for meaningfulness.

For example, semantic analyzer would reject a sentence like **“Hot ice-cream”**.

CONTD…

- **PREGMATIC ANALYSIS :**

  Pragmatic analysis simply fits the actual objects/events, which exist in a given context with object references obtained during the last phase (semantic analysis).

  For example, the sentence **“Put the banana in the basket on the shelf”** can have two semantic interpretations and pragmatic analyzer will choose between these two possibilities.

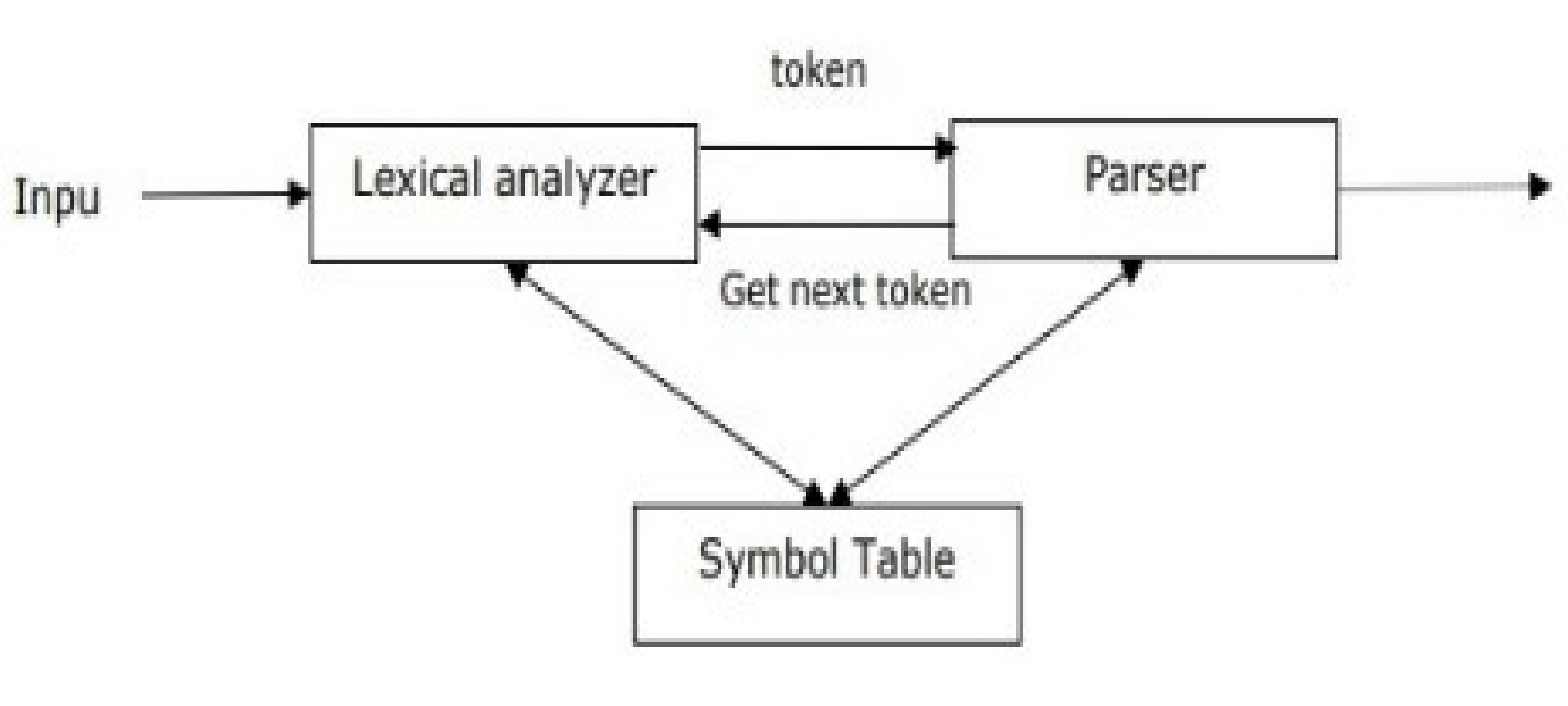
SYNTACTIC ANALYSIS

- The purpose of this phase is to draw exact meaning, or you can say dictionary meaning from the text.

- Syntax analysis checks the text for meaningfulness comparing to the rules of formal grammar.

- Syntactic analysis or parsing may be defined as the **process of analyzing the strings of symbols in natural language conforming to the rules of formal grammar.**

CONCEPT OF PARSER



CONTD…

- **The main roles of parser include :**
  - To report any syntax error.
  - To recover from commonly occurring error so that the processing of the remainder of program can be continued.
  - To create parse tree.
  - To create symbol table.
  - To produce intermediate representations (IR).

TYPES OF PARSING

Derivation divides parsing into the followings two types :

- **Top-down Parsing :**

It begind by hypothesizing a sentence and successively predicting lower level constituents until individual preterminal symbols are written.

- **Bottom-up Parsing :**

  It begins with the actual words appearing in the sentence and is, therefore, data driven.

CONTD…

- **Top-down Parsing :**

  For eg : “Kathy jumed the horse”

S > NP VP
NAME VP
Kathy VP
Kathy V NP
Kathy jumped NP
Kathy jumped ART N
Kathy jumped the N
Kathy jumped the horse

CONTD…

- **Bottom-up Parsing :**

  For eg : “Kathy jumed the horse”

Kathy jumped the horse

NAME jumped the horse

NAME V the horse

NAME V ART horse

NAME V ART N

NP V ART N

NP V NP

NP VP

S

TYPES OF DERIVATION

In order to get the input string, we need a sequence of production rules. Derivation is a set of production rules. During parsing, we need to decide the non-terminal, which is to be replaced along with deciding the production rule with the help of which the non-terminal will be replaced.

In this we will learn about the two types of derivations, which can be used to decide which non-terminal to be replaced with production rule :

- **Left-most Derivation :**

  In the left-most derivation, the sentential form of an input is scanned and replaced from the left to the right. The sentential form in this case is called the left-sentential form.

- **Right-most Derivation :**

  In the right-most derivation, the sentential form of an input is scanned and replaced from right to left. The sentential form in this case is called the right-sentential form.

CONCEPT OF GRAMMAR

A mathematical model of grammar was given by **Noam Chomsky** in 1956, which is effective for writing computer languages.

Mathematically, a grammar G can be formally written as a 4-tuple **(N, T, S, P)** where :

N or $\mathbf{V_N}$ = set of non-terminal symbols, i.e., variables.

T or $\sum$ = set of terminal symbols.

S = Start symbol where $S \in N$

P denotes the Production rules for Terminals as well as Non-terminals. It has the form $\alpha \rightarrow \beta$, where α and β are strings on $V_N \cup \sum$ and least one symbol of α belongs to V_N

CONSTITUENCY GRAMMAR

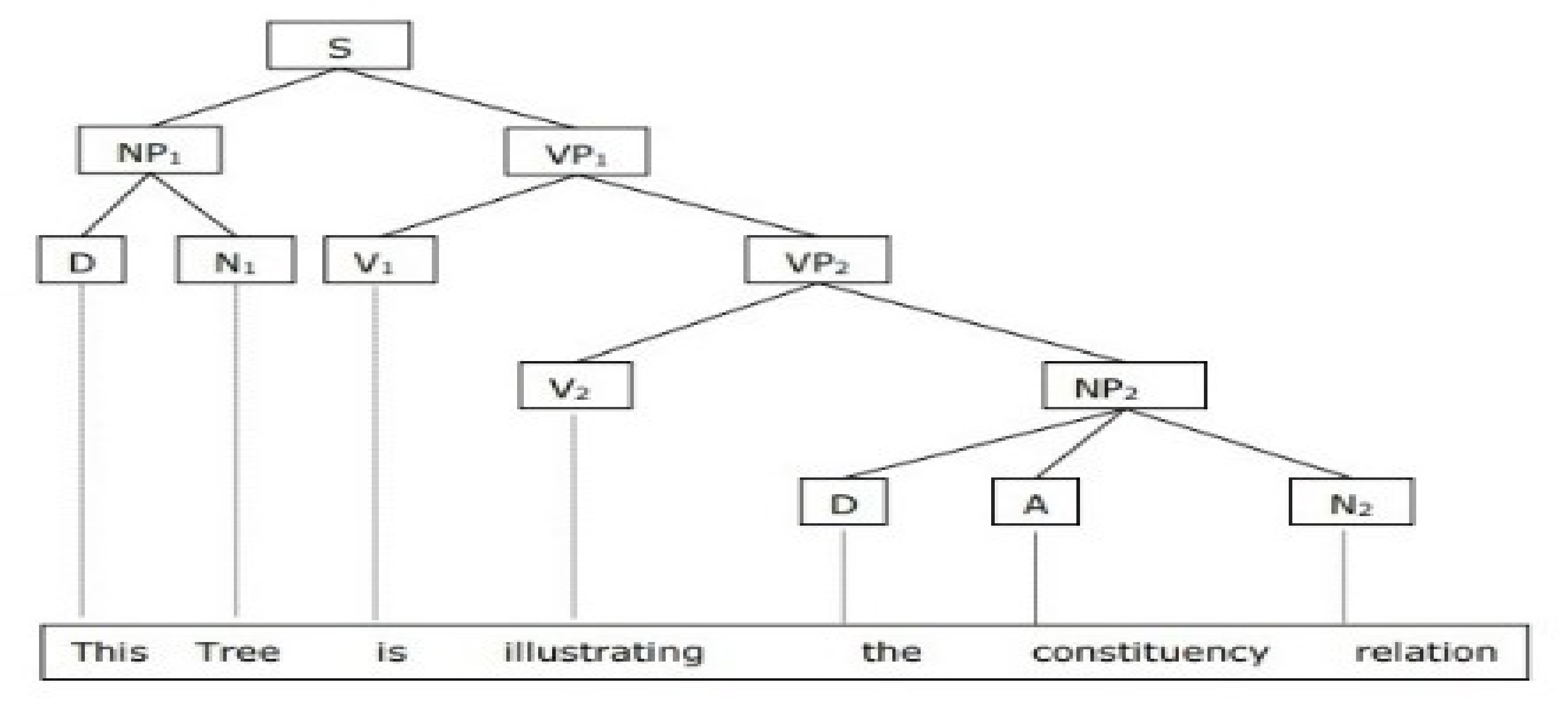
Phrase structure grammar, is based on the constituency relation. That is why it is also called constituency grammar. It is opposite to dependency grammar.

The fundamental points about constituency grammar and constituency relation :

- All the related frameworks view the sentence structure in terms of constituency relation.
- The constituency relation is derived from the subject-predicate division of Latin as well as Greek grammar.
- The basic clause structure is understood in terms of **noun phrase NP** and **verb phrase VP**.

CONTD…

For eg. : “This tree is illustrating the constituency relation” can be written as :



DEPENDENCY GRAMMAR

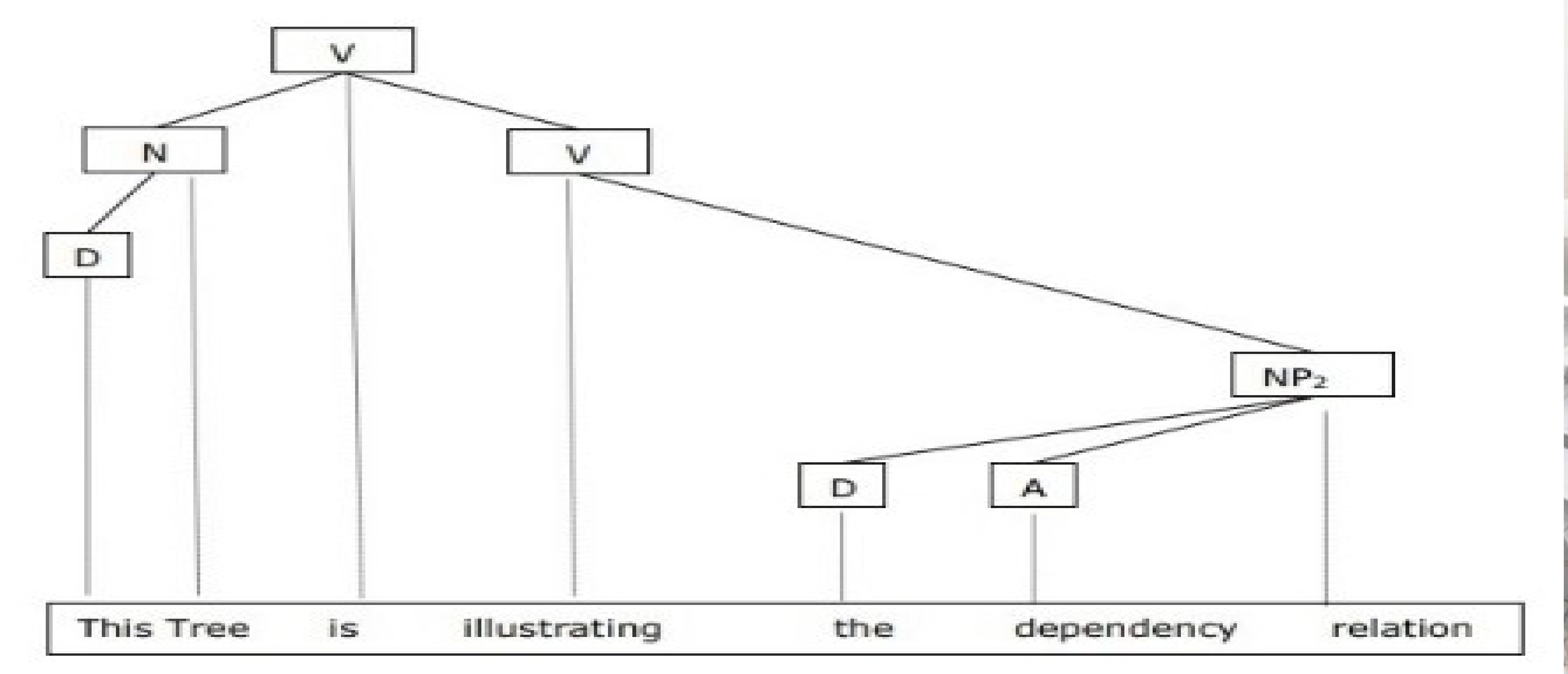
It is based on dependency relation. Dependency grammar (DG) is opposite to the constituency grammar because it lacks phrasal nodes.

The fundamental points about Dependency grammar and Dependency relation :

- In DG, the linguistic units, i.e., words are connected to each other by directed links.
- The verb becomes the center of the clause structure.
- Every other syntactic units are connected to the verb in terms of directed link. These syntactic units are called *dependencies*.

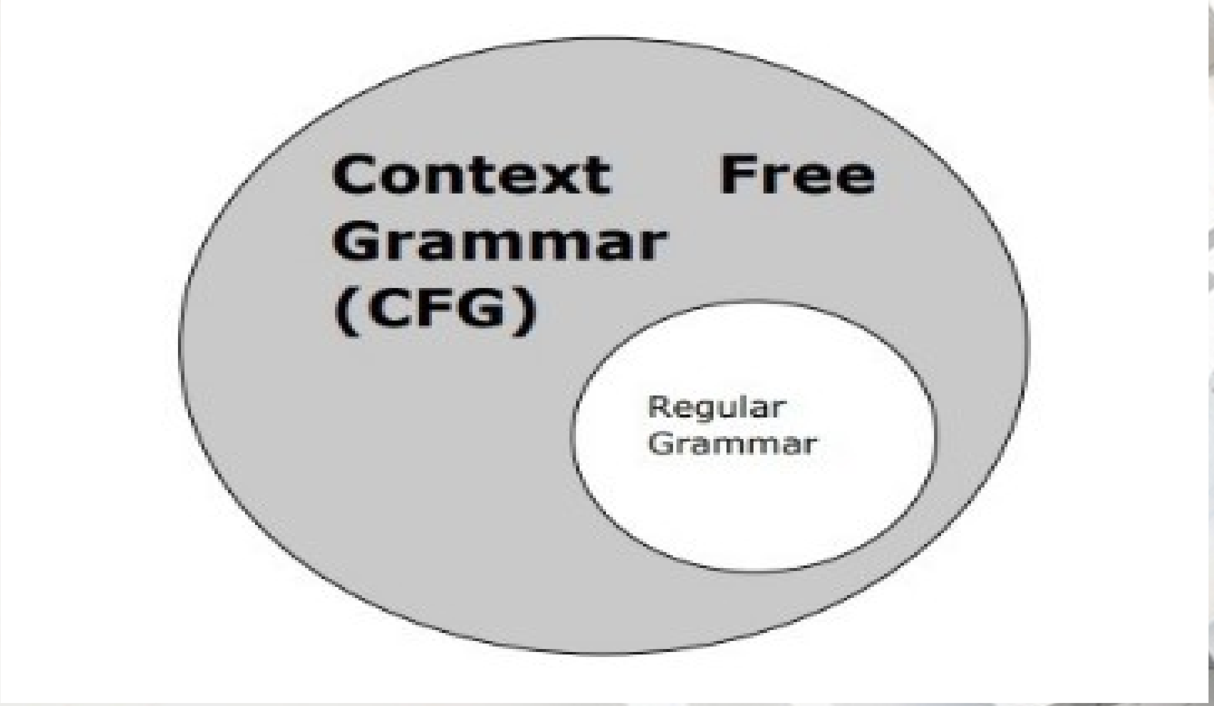
CONTD…

For eg. : “This tree is illustrating the constituency relation” can be written as :



CONTEXT FREE GRAMMAR

Context free grammar, also called **CFG,** is a notation for describing languages and a superset of Regular grammar. It can be seen in the following diagram :



CONTD…

CFG consists of finite set of grammar rules with the following **four** components :

- **Set of Non-terminals :**

It is denoted by V. The non-terminals are syntactic variables that denote the sets of strings, which further help defining the language, generated by the grammar.

- **Set of Terminals :**

It is also called tokens and defined by Σ. Strings are formed with the basic symbols of terminals.

CONTD…

- **Set of Productions :**

It is denoted by P. The set defines how the terminals and non-terminals can be combined. Every production(P) consists of non-terminals, an arrow, and terminals (the sequence of terminals). Non-terminals are called the left side of the production and terminals are called the right side of the production.

- **Start Symbol :**

The production begins from the start symbol. It is denoted by symbol S. Non-terminal symbol is always designated as start symbol.

SEMANTIC ANALYSIS

- The purpose of semantic analysis is to draw exact meaning, or you can say dictionary meaning from the text. The work of semantic analyzer is to check the text for meaningfulness.

- We already know that lexical analysis also deals with the meaning of the words, then how is semantic analysis different from lexical analysis?

- Lexical analysis is based on smaller token but on the other side semantic analysis focuses on larger chunks.

CONTD…

- **That is why semantic analysis can be divided into the following two parts :**
  - Studying meaning of individual word :

    It is the first part of the semantic analysis in which the study of the meaning of individual words is performed. This part is called lexical semantics.
  - Studying the combination of individual words :

    In the second part, the individual words will be combined to provide meaning in sentences.

The most important task of semantic analysis is to get the proper meaning of the sentence.

For example, analyze the sentence **“Ram is great.”**

ELEMENTS OF SEMANTIC ANALYSIS

- **Hyponymy :**

It may be defined as the relationship between a generic term and instances of that generic term. Here the generic term is called hypernym and its instances are called hyponyms.

For example, the word **color** is hypernym and the color blue, yellow etc. are hyponyms.

- **Homonymy :**

It may be defined as the words having same spelling or same form but having different and unrelated meaning.

For example, the word **“Bat”** is a homonymy word because bat can be an implement to hit a ball or bat is a nocturnal flying mammal also.

CONTD…

- **Polysemy :**

Polysemy means “many signs”. It is a word or phrase with different but related sense. In other words, we can say that polysemy has the same spelling but different and related meaning.

For example, the word **“bank”** is a polysemy word having the following meanings −

  - A financial institution.
  - The building in which such an institution is located.
  - A synonym for “to rely on”.

- **Difference between Polysemy and Homonymy :**

The main difference between them is that in polysemy, the meanings of the words are related but in homonymy, the meanings of the words are not related.

For example, if we talk about the same word **“Bank”,** we can write the meaning ‘a financial institution’ or ‘a river bank’. In that case it would be the example of homonym.

CONTD…

- **Synonymy :**

It is the relation between two lexical items having different forms but expressing the same or a close meaning. Examples are ‘author/writer’, ‘fate/destiny’.

- **Antonymy :**

It is the relation between two lexical items having symmetry between their semantic components relative to an axis. The scope of antonymy is as follows −

  - **Application of property or not** − Example is ‘life/death’, ‘certitude/incertitude’
  - **Application of scalable property** − Example is ‘rich/poor’, ‘hot/cold’
  - **Application of a usage** − Example is ‘father/son’, ‘moon/sun’.

BUILDING BLOCKS

In word representation or representation of the meaning of the words, the following building blocks play an important role :

- **Entities** − It represents the individual such as a particular person, location etc. For example, Haryana. India, Ram all are entities.
- **Concepts** − It represents the general category of the individuals such as a person, city, etc.
- **Relations** − It represents the relationship between entities and concept. For example, Ram is a person.
- **Predicates** − It represents the verb structures. For example, semantic roles and case grammar are the examples of predicates.

BUILDING BLOCKS

In word representation or representation of the meaning of the words, the following building blocks play an important role :

- **Entities** − It represents the individual such as a particular person, location etc. For example, Haryana. India, Ram all are entities.
- **Concepts** − It represents the general category of the individuals such as a person, city, etc.
- **Relations** − It represents the relationship between entities and concept. For example, Ram is a person.
- **Predicates** − It represents the verb structures. For example, semantic roles and case grammar are the examples of predicates.

DISCOURSE PROCESSING

- The most difficult problem of AI is to process the natural language by computers or in other words *natural language processing* is the most difficult problem of artificial intelligence.
- If we talk about the major problems in NLP, then one of the major problems in NLP is discourse processing − building theories and models of how utterances stick together to form **coherent discourse**.
- Actually, the language always consists of collocated, structured and coherent groups of sentences rather than isolated and unrelated sentences like movies.
- These coherent groups of sentences are referred to as discourse.

CONCEPT OF COHERENCE

- Coherence and discourse structure are interconnected in many ways.
- Coherence, along with property of good text, is used to evaluate the output quality of natural language generation system.
- The question that arises here is what does it mean for a text to be coherent?
- Suppose we collected one sentence from every page of the newspaper, then will it be a discourse? Of-course, not. It is because these sentences do not exhibit coherence.

CONCEPT OF COHERENCE

The coherent discourse must possess the following properties :

- **Coherence relation between utterances :**

The discourse would be coherent if it has meaningful connections between its utterances(ucharan). This property is called coherence relation.

For example, some sort of explanation must be there to justify the connection between utterances.

- **Relationship between entities :**

Another property that makes a discourse coherent is that there must be a certain kind of relationship with the entities. Such kind of coherence is called entity-based coherence.

ALOGRITHMS FOR DISCOURSE SEGMENTATION

- **Unsupervised Discourse Segmentation :**

Often represented as linear segmentation. For example, there is a task of segmenting the text into multi-paragraph units; the units represent the passage of the original text. These algorithms are dependent on cohesion that may be defined as the use of certain linguistic devices to tie the textual units together.

- **Supervised Discourse Segmentation :**

It needs to have boundary-labeled training data. It is very easy to acquire the same. In supervised discourse segmentation, discourse marker or cue words play an important role. Discourse marker or cue word is a word or phrase that functions to signal discourse structure. These discourse markers are domain-specific.