# Natural Language Processing

## Basic Terminologies of NLP :-

(1) Corpus : A corpus is a large ~~and~~ structured set of texts used in NLP. It serves as a dataset for linguistic analysis ~~and~~ model training.

(2) Document :-

## # Applications of NLP :-

(1) Sentiment Analysis :-

Analyze text data from social media, reviews and feedback to determine sentiment expressed (+ve, -ve, neutral) towards the products, services (or) topics. This is widely used in market research to gauge public opinion.

(2) Chatbots and Virtual Assistants :-

Powering conversational agents that can understand and respond to human queries in natural language. Chatbots are used in customer service, while virtual assistants like Siri, Alexa, and Google Assistant help users perform tasks through voice commands.

(3) Machine Translation :-

Automatically translating text to speech from one language to another. Applications like Google Translate enable users to understand content in foreign language

(4) Named Entity

Identifying

into predef

people, o

This is

extraction

(5) Speech Rec

converting sp

(6) Text

categorizing

is used in

## # LEVELS OF

(1) Morpholog

(2) Syntacti

(3) Semanti

(4) Discours

(5) Pragm

(1) Morphological

Morphology is

words. At

smallest u

a "morpheme

Prefixes

(4) Named Entity Recognition (NER):

Identifying & classifying key elements into text into predefined categories such as names of people, organisation, locations dates & more.

This is useful in content classification, information extraction etc..,

(5) Speech Recognition

Converting spoken language into text.

(6) Text classification :-

categorizing text into predefined groups. This is used in email filtering.

# LEVELS OF NLP :- (or) STAGES :-

(1) Morphological Analysis
(2) Syntactic Analysis
(3) Semantic Analysis
(4) Discourse Analysis
(5) Pragmatic Analysis.

(1) Morphological Analysis :-

Morphology is study of structure & formation of words. At this level, NLP systems analyze the smallest units of meaning within words, known as "morphemes". This includes understanding root words, prefixes, suffixes.

(2) Syntactic Analysis :- (Parsing sentences)

Syntax is concerned with the arrangement of words in sentences ~~a grammat~~ grammatical relationship between them. ~~Syntatic~~ Syntactic level, analyzes sentence structure to understand how words combine to form a sentence. This involves parsing the sentences to identify subjects, predicates, objects.

(3) Semantic Analysis :-

Semantic Analysis deals with meaning of words, Phrases, sentences & texts. At Semantic level, NLP aims of understanding the indented meaning conveyed by a combination of words.

(4) Discourse Analysis :-

This analyses goes beyond the analysis of individual sentences. It involves analysing the coherence & structure of longer texts

(5) Pragmatic Level :-

Pragmatic focuses of use of language in context & how context influences the interpretation of meaning. It ~~uses~~ considers factors like the speaker's intention & the relationship b/w speaker & the listen

arrangement of words
antical relationship
level, analyses
how words
involves parsing
objects, products,
objects.

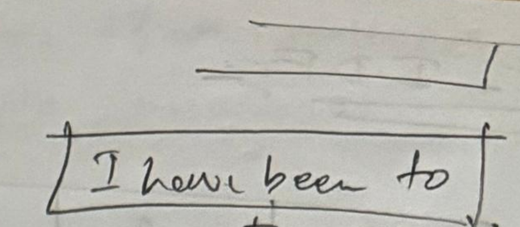
meaning of words,
Semantic level,
dentified means
words.

analysis of
analysing the
tags

rk in
flavours the
- words
eather's interby
d the

# STEPS IN NLP :-

1. Tokenization
2. Lemmatization
3. Stop Words
4. Stemming
5. Bag of Words
6. TF - IDF
7. N - Gram
8. Word 2 Vec
9. Avg Word 2 Vec.




→ Vocabulary creation
→ Vectorization

[Term Frequency — Inverse Document Frequency]

→ mapping semantically similar words to provide good input to DL-specific

calculate the repetition of word

$$T.F = \frac{\text{No. of times term t appears in document d}}{\text{Total no. of terms in document d}}$$

calculate the importance of ter

$$IDF = \log\left(\frac{\text{Total no. of documents D}}{\text{No. of documents containing term t}}\right)$$

Sent 1 → good boy
Sent 2 → good girl
Sent 3 → boy girl good

} IDF

total sentence / No. of word repeat

No. of w / Total wo

T.F

| | sent 1 | sent 2 | sent 3 |
|---|---|---|---|
| good | 1/2 | 1/2 | 1/3 |
| boy | 1/2 | 0 | 1/3 |
| girl | 0 | 1/2 | 1/3 |

IDF

| | |
|---|---|
| good | $\log(3/3)$ |
| boy | $\log(3/2)$ |
| girl | $\log(3/2)$ |

## T-F – I·D·F:-

| | | | |
|---|---|---|---|
| | | | |
| Sent 1 | 0 | | |
| Sent 2 | 0 | | |
| Sent 3 | 0 | | |

## Stream Processing

Stream processing is a computing [illegible] focuses on the "[illegible]

UNIT-IV

Semantic and Discourse Analysis

# Word Embedding :-

① Word2Vec :-

* Word2Vec creates vectors of the words that are "distributed numerical representions of word features" — these words features could comprise of words that represent the context of individual words present in our ~~Vocab~~ Vocabulary.

* Word embeddings eventually help in establishing the association of word with another (similar meaning word through the created vectors.

* It is a two layer neural network to generate word embeddings given a text corpus.

* Word Embeddings — Mapping of words in Vector Space

Man →

| |
|---|
| 0.52 |
| 0.76 |
| 1.21 |
| 0.22 |
| -1.36 |
| 0.49 |
| -3.69 |
| -0.07 |

Women →

| |
|---|
| 0.73 |
| 0.89 |
| -1.67 |
| 1.32 |
| 0.36 |
| -1.49 |
| 2.71 |
| 0.05 |

King - Man + Women = Queen

Analysis

are "distributed
these words
represent
ent in our
ng the
ar meaning
word
Vector

## Working of Word2Vec :-

* The word2Vec objective function causes the words that occur in similar contexts to have similar embeddings.

Ex- The kid said he would grow up to be superman
The child said he would grow up to be superman.

The words kid and child will have similar word vectors due to similar context.

## Types of Word2Vec :-

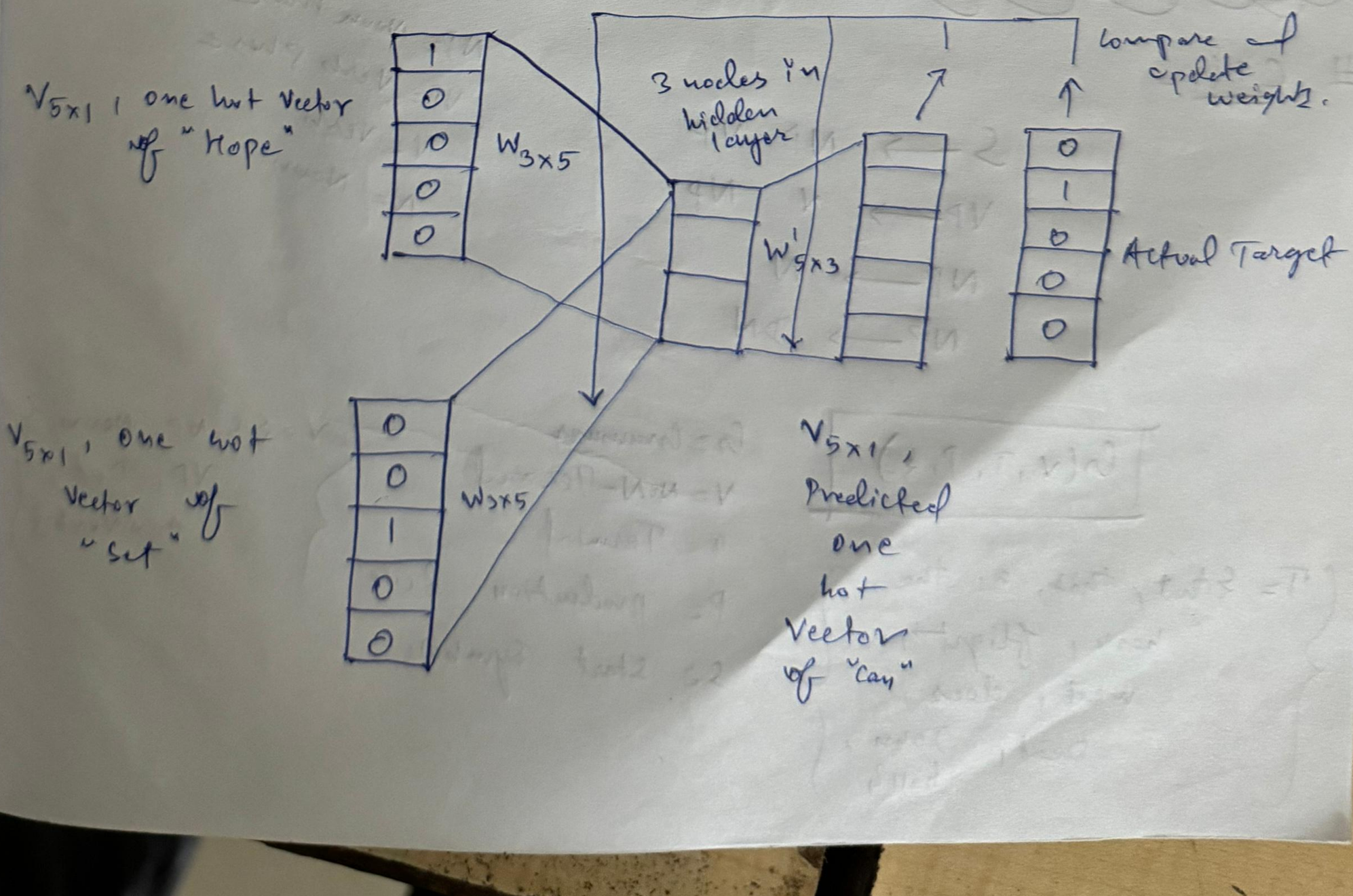
Two different model architectures that can be used by Word2Vec to create the word embeddings are:

(a) Continuous Bag of Words (CBOW) Model. ⟹ Predict target word from Context
(b) Skip Gram Model ⟹ Predict Context word from target.

### (a) CBOW :-

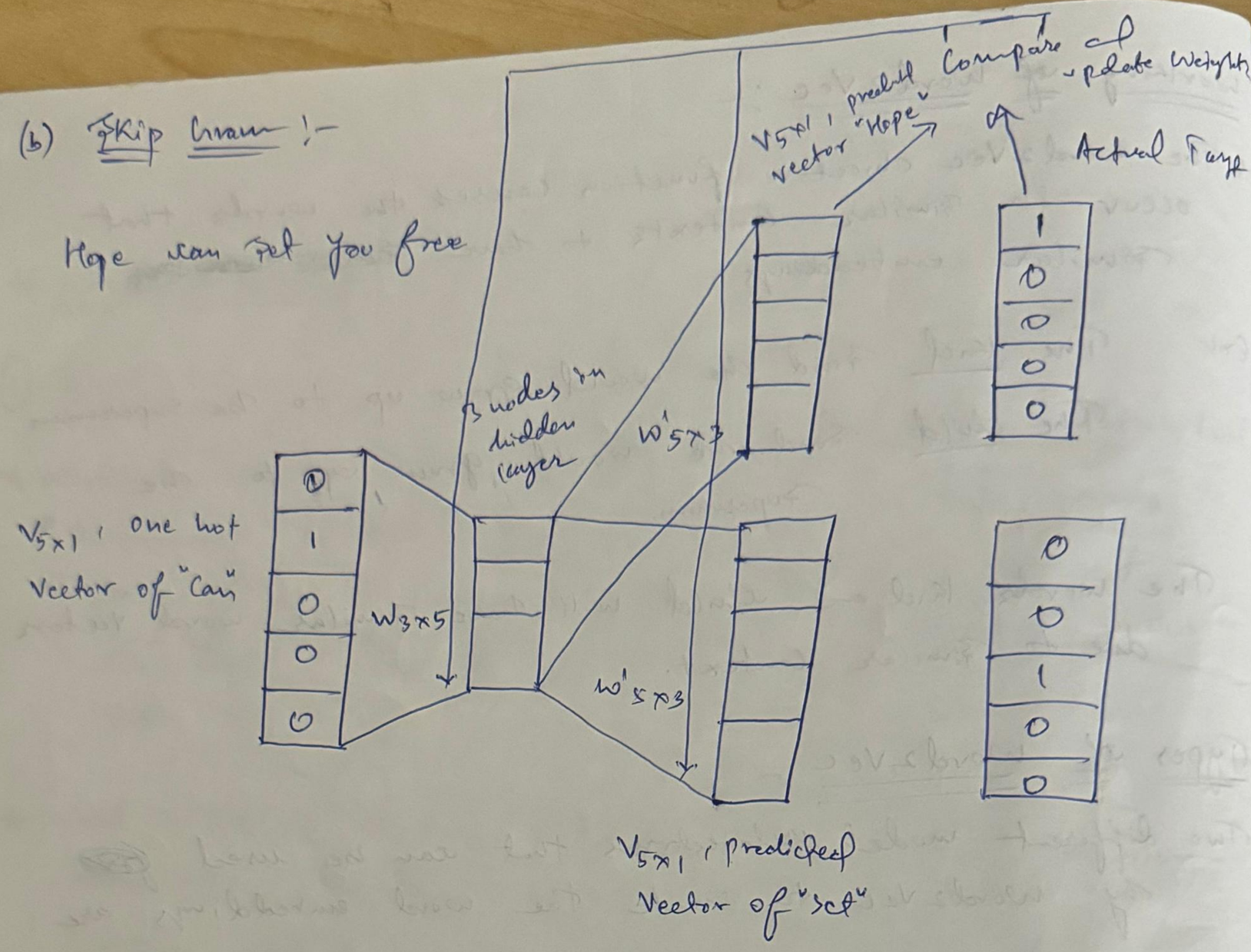
Hope can set you free

window size = 3



(b) Skip Gram :-

Hope can set you free




SKIP GRAM is used to predict the context word for a given target word. It's reverse of CBow algorithm. Here, target word is input while context words are output

# DISCOURSE AND COHERENCE

# CFG :-

S ⟶ NP VP
VP ⟶ V NP
NP ⟶ N
NP ⟶ DN

NP = Noun phase
VP = verb phase
V = Verb
N = Noun.

G(V, T, P, S)

G = Grammar
V = NON-Terminal
T = Terminal
P = production Rule
S = Start Symbol.

T = {that, this, a, the, book, flight, meal, does, Dead, John, hall}

V = {S, NP, Noun, VP, Det, Aux verb}

P = {
S ⟶
S ⟶
S ⟶
NP
NP
NP
}

Det ⟶
Noun ⟶
Verb ⟶
Aux ⟶

Rule :-
α ⟶
α ⟶

Q.1 The ma

$P = \{$

$S \longrightarrow NP\ VP$

$S \longrightarrow Aux\ NP\ VP$

$S \longrightarrow VP$

$NP \longrightarrow Det\ Noun$

$VP \longrightarrow Verb$

$VP \longrightarrow Verb\ NP$

$\}$

$Det \longrightarrow this \mid that \mid a \mid the$

$Noun \longrightarrow book \mid flight \mid John \mid ball \mid meal$

$Verb \longrightarrow book \mid include \mid Read.$

$Aux \longrightarrow does \mid is$

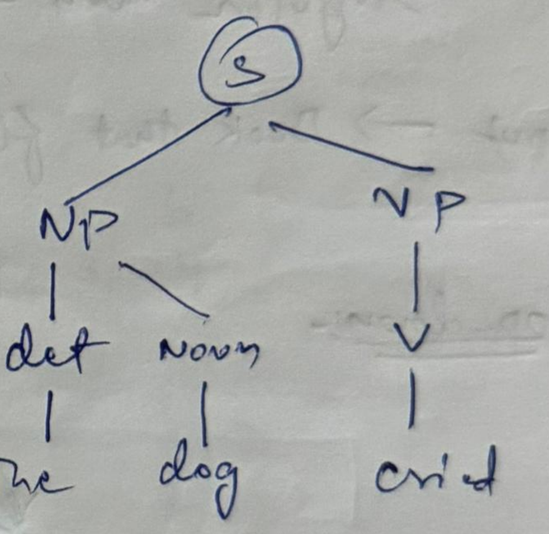
**Rule :-**

$\alpha \longrightarrow \beta$

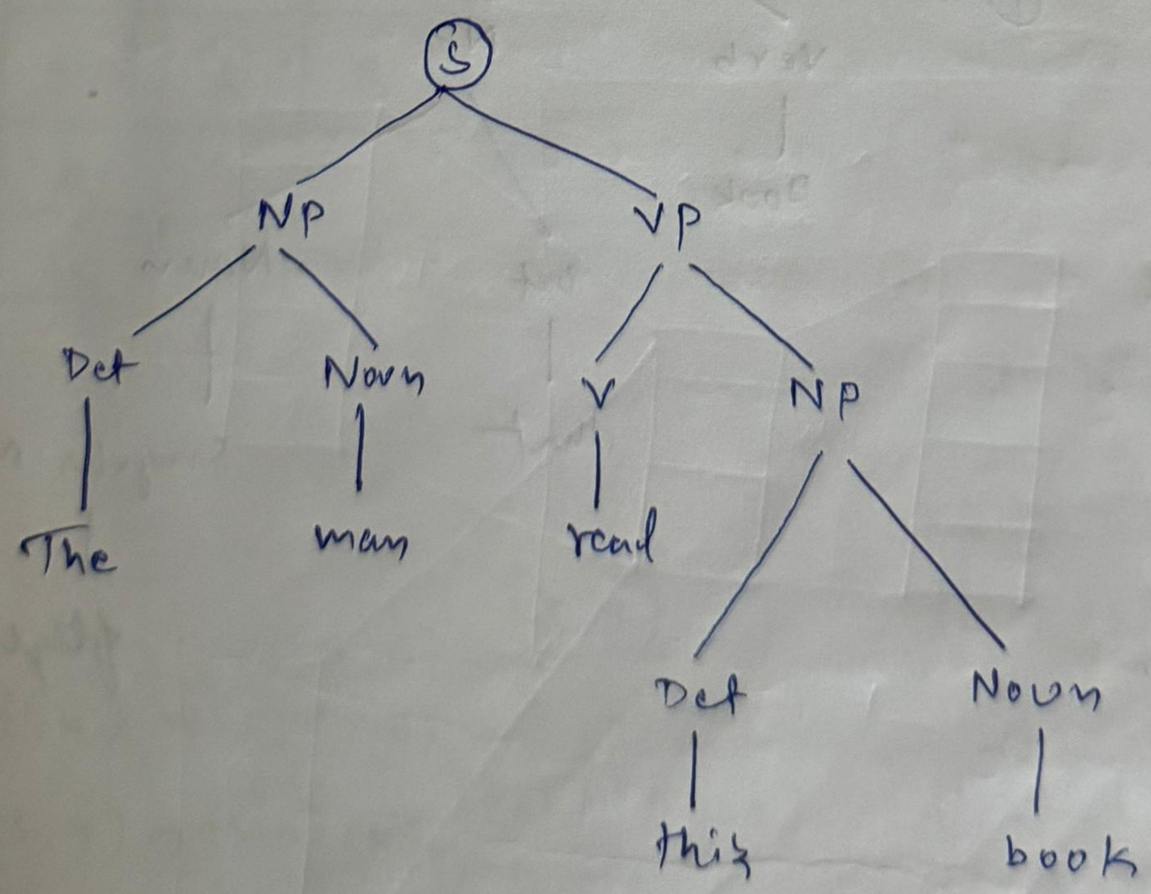
$\alpha \longrightarrow$ Single variable

$\beta \in (V+T)^*$

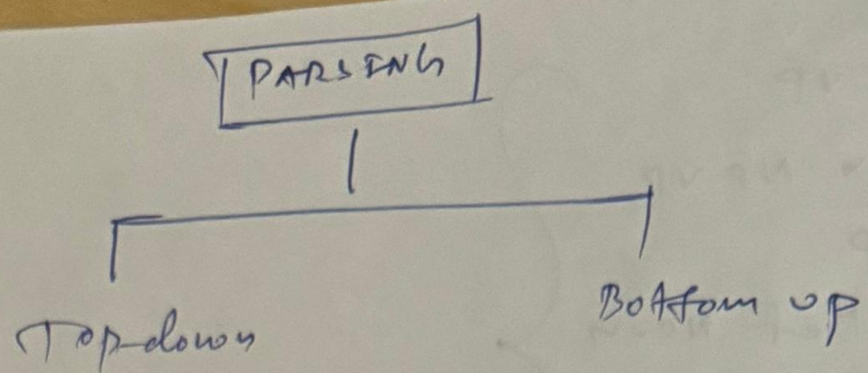
The dog cried
Det Noun Verb




Q.1 The man read this book



# Parsing :-




Q. GRAMMER:

S ⟶ VP

VP ⟶ Verb NP

NP ⟶ Det Nom

ND ⟶ Det Nom

Det ⟶ that

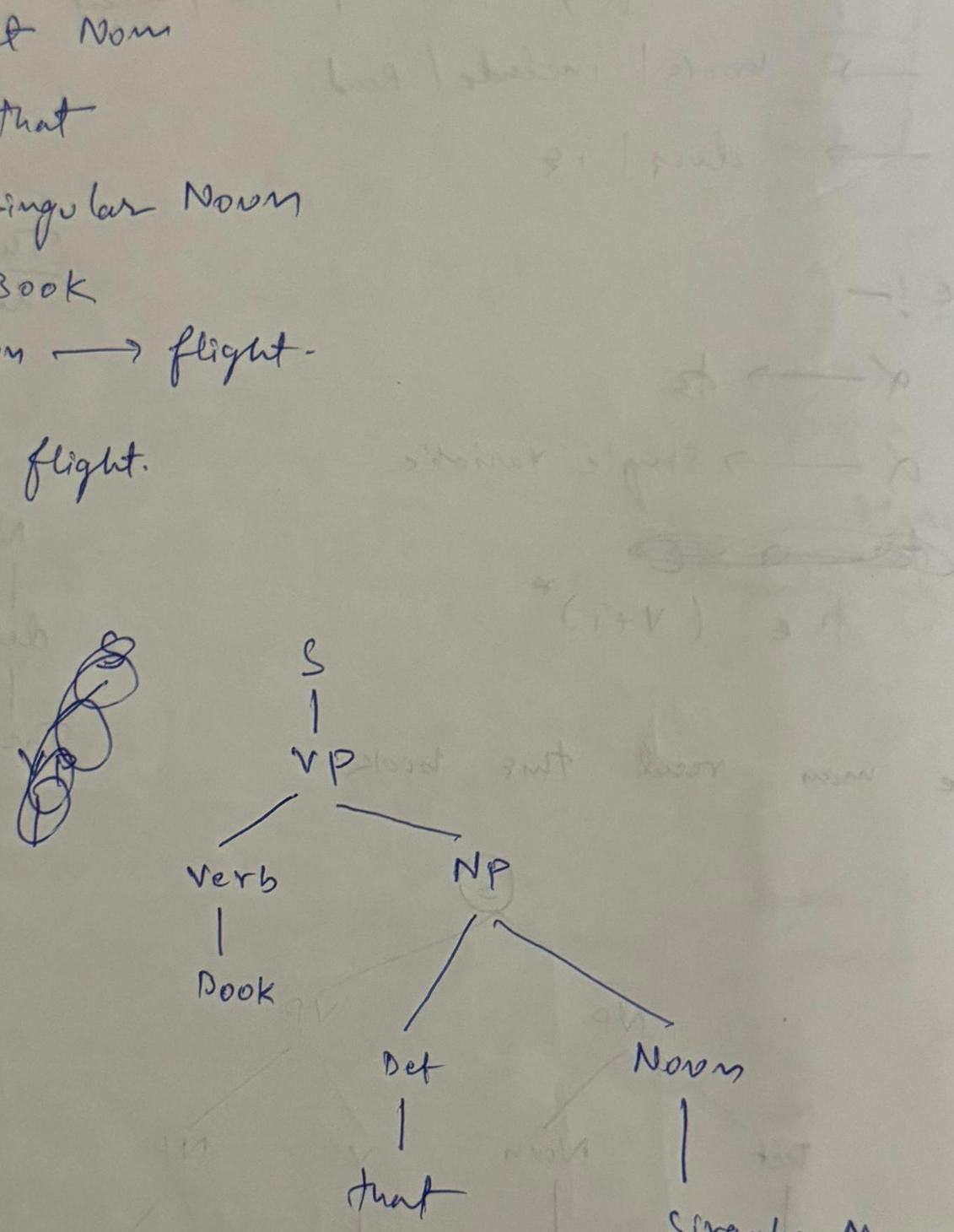
Nom ⟶ Singular Noun

Verb ⟶ Book

Singular noun ⟶ flight.

Input ⟶ Book that flight.

① Top-down:-

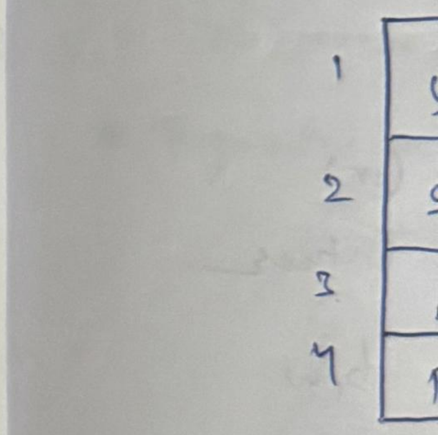



② Bottom u

# CKY Pars

Q. Check whe
the follow

Soln



1 4

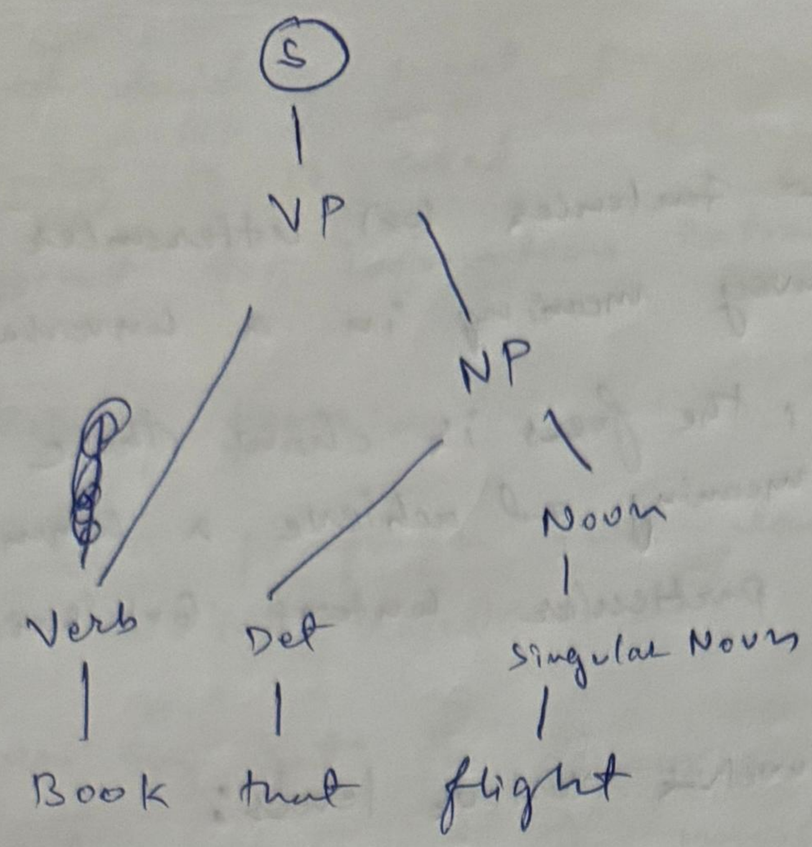
1 2 3

(1,1) (2,

(1,2) (3

(1,3) (4

(2) Bottom UP




# CKY Parsing :-

Q. Check wheather a string "abbb" is a valid member of the following CFG

$S \longrightarrow AB$

$A \longrightarrow BB/a$

$B \longrightarrow AB/b$

a b b b
1 2 3 4

Soln

| | 4 | 3 | 2 | 1 |
|---|---|---|---|---|
| 1 | S,B | A | S,B | A |
| 2 | S,B | A | B | |
| 3 | A | B | | |
| 4 | B | | | |

1 2
(1,1) (2,2)
A . B
(A . B)
S , B

2 3
(2,2) (3,3)
B B
(BB)
A

1 3
1 2 3
(1,1) (2,3)
A A ⇒ ∅

3 4
(3,3) (4,4)
B B
(BB)
A

1 4
1 2 3 4
(1,1) (2,4) ⇒ (A) (S.B)
(1,2) (3,4) ⇒ (S,B) (A)
(1,3) (4,4) ⇒ (A) (B)
(S,B)

(∅) (S,B)
(S,B)

(2,4)
2 3 4
(2,2) (3,4)
B A
AB
(2,3) (4,4)
A B
AB
⇒ S,B

1 2
(1,1) (2,2)
A . B
(A . B)
S , B

2 3
(2,2) (3,3)
B B
(BB)
A

1 3
1 2 3
(1,1) (2,3)
A A ⇒ ∅

3 4
(3,3) (4,4)
B B
(BB)
A

1 4
1 2 3 4
(1,1) (2,4) ⇒ (A) (S.B)
(1,2) (3,4) ⇒ (S,B) (A)
(1,3) (4,4) ⇒ (A) (B)
(S,B)

(∅) (S,B)
(S,B)

(2,4)
2 3 4
(2,2) (3,4)
B A
AB
(2,3) (4,4)
A B
AB
⇒ S,B

(1,4)
1 2 3 4
(1,2) (3,3) (3,3) (4,4)
(S,B) (B) B B
(S,B) (B,B) (BB)
∅ A A
∅ ∪ ∅ ∪ A ⇒ A

# DISCOURSE AND COHERENCE

(a) Discourse :-

Discourse refers to how sentences (or) utterances are connected ed organized to convey meaning in a conversation (or) text.

* In discourse analysis, the focus is how these sequences of language convey meaning ed achieve a communicative purpose within a particular context (or) social setting.

Structure of discourse involves various levels:

(i) Cohesion :-

The linguistic devices used to link sentences ed clauses within a text, such as pronouns, conjunctions, ed lexical ties.

(ii) Coherence :-

The logical ed semantic relationship that allows the text to be understood as a whole

(iii) Discourse markers :-

words (or) phrases like "however", furthermore (or) "for example", which guide the reader (or) listner through the discourse, indicating relations b/w segments of discourse

(b) Coherence :-

* Coherence is about underlying connections that make a text (or) speech understandable ed meaningful

* It is the way components of the discourse fit together logically ed semantically.

# coreference

The goal
phrases

* It is imp

# Coherent di

* If we re
paragraph

* Hence, we
if we
consecuti
just a

# Entity - base

# Discourse Seg

When we de
discourse,

* Segmentation
(1) info
(2) Te
(3)

Two types :-

(1) Unsupervi
(2) Supervise

* We
* We ma
ma
Se

# Coreference resolution :-

The goal of deciding what phrases refer to is called "coreference resolution".

* It is important for information extraction, summarization

# Coherent discourse texts :-

* If we read a paragraph, we can see that entire paragraph is interrelated.

* Hence, we can say that the discourse is coherence, but if we only combine the newspaper headlines consecutively, then it is not discourse, it is just a group of sentences that is non-coherence

# Entity-based coherence :-

Relationship b/w entities, is also coherent

# Discourse Segmentation :-

When we determine the types of structures for a large discourse, we term its segmentation.

* Segmentation is difficult, but necessary as :-

1. info Retrieval
2. Text summarisation
3. Information extraction.

Two types :-

1. Unsupervised Discourse Segmentation
   - ⇒ linear segmentation
   - ⇒ clustering or grouping up of similar texts with help of coherent discourse
2. Supervised Discourse Segmentation.
   - We deal with labeled boundary
   - We make use of Discourse markers to do the segmentation

Discourse Markers :-

① Transition Words and Phrases :-

words like "however", "despite" and "as" indicate shifts in the narrative

② Punctuation :-

Periods (.) are often used to denoted end of sentences and hence end of segment. Not all

all periods denote end of segment.

③ Temporal and Spatial indicator :-

Phrases like "this time", "soon enough" and "at the" provide temporal (or) spatial context, which can signal the beginning (or) continuation of new segment.

④ Contextual Cues :-

The change of topic from one subject to another signifies significant cue for segment boundary

# Text Coherence :-

* To achieve coherent discourse, we must focus on coherence relations in specific.

* As we know that coherence relation defines the possible connection b/w utterances in a discourse.

① Result
② Explanation
③ Parallel
④ Elaboration
⑤ occasion

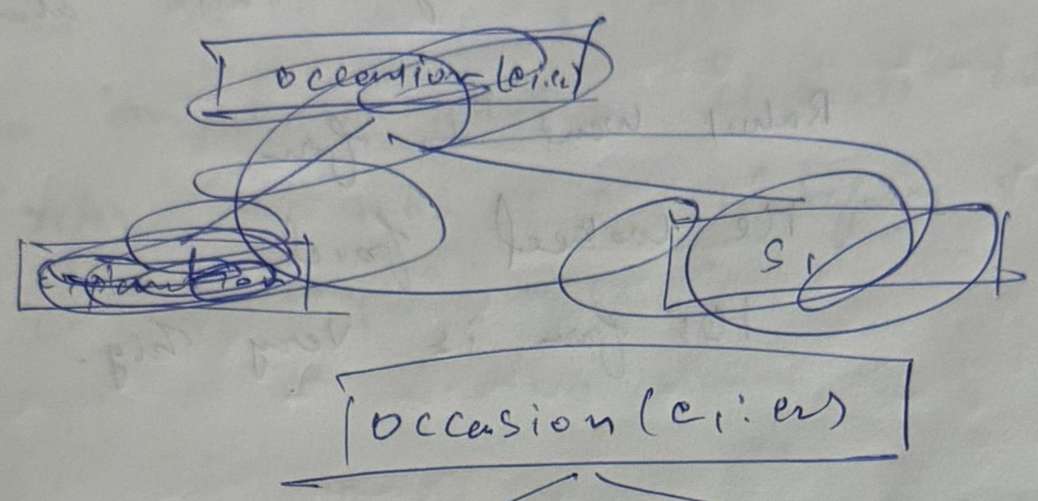
$S_1$ = John went to bank to deposite his check

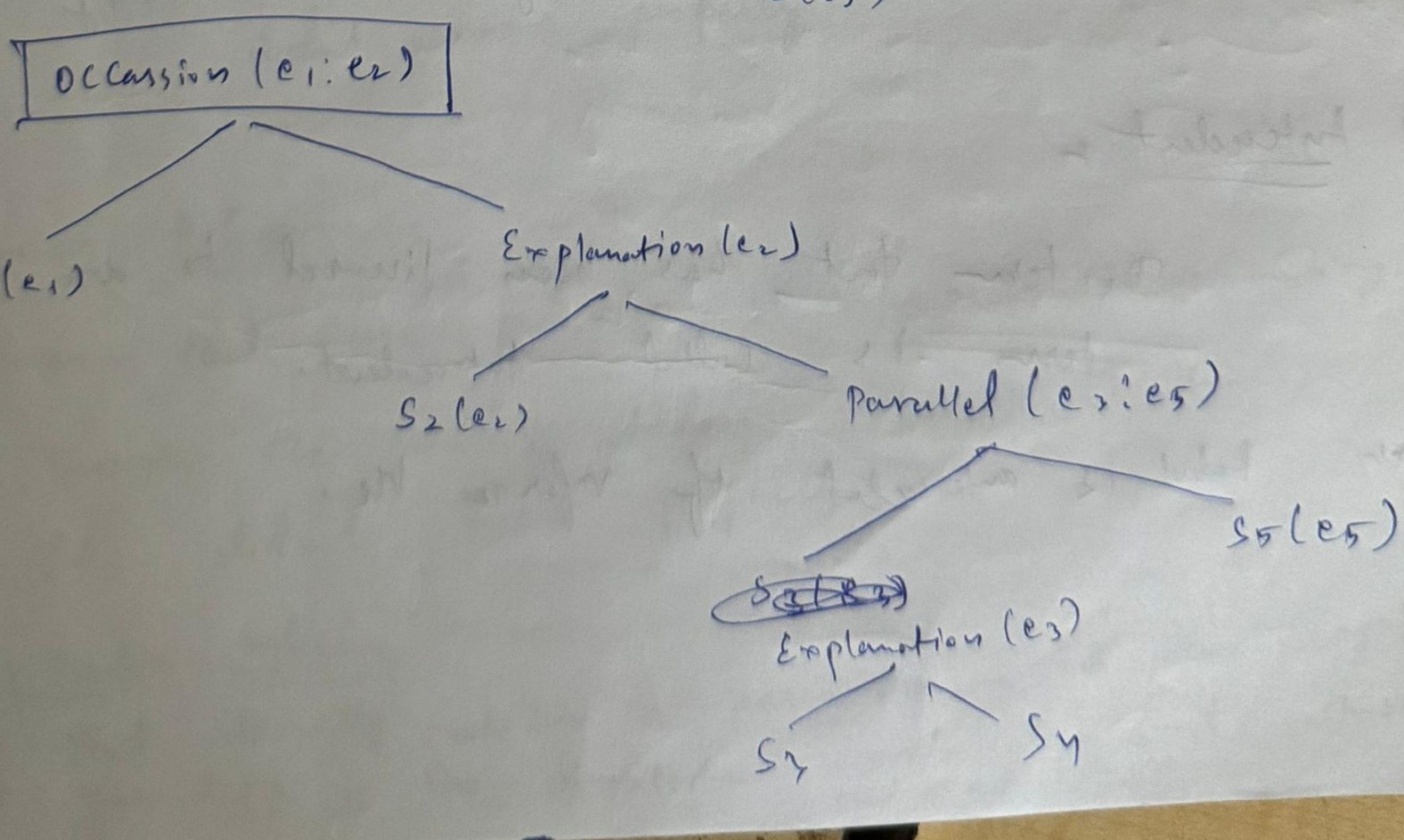
$S_2$ = He then took a train to Bill's car dealership

$S_3$ = He needed to buy a car

$S_4$ = The company he works for is not near an public trasportation

$S_5$ = He also wanted to talk with bill about soft ball team



# Reference Resolution :-

* The extraction of meaning (or) interpretations of the sentences of discourse is one of the most important tasks in natural language processing, to do so, we need to "what (or) who the entity" is.

* Reference resolution means understanding the type of entity that is being talk about

Rahul went to farm
He cooked food
His farm is very big.

① Referent :-
Entity which we refer to
Ex Rahul is referent

② to-refer :-
It is a term used for an entity if two (or) more expressions are referring to the entity.

Ex: Rahul and he used for the entity

③ Antecedent :-
The term that has been licensed to use another term is called antecedent.
Ex Rahul is antecedent of reference He.

(4) Anaphora and Anaphoric :-

can be said to be term or reference used for an entity that has previously been introduced in the text/sentence.

# Reference :-

Linking a referring expression to another referring expression in surrounding text.

Ex:- Suha brought a printer. It costs her ₹ 20,000

* "Her" refers to Suha and "it" refers to printer.
* also called "anaphoric reference".

## Five types :-

(1) Indefinite :-

introduces a new object to discourse context.
most commonly with a/an/some.

Ex:- Some printers make noise while printing
I met this girl earlier in surprise

(2) Definite :-

Refers to an object that already exists in a discourse context.

Ex:- I brought a printer today. The printer didn't work properly.

(3) Pronominal Reference :-

References that use pronouns to refer to some entity.

Ex:- I brought the printer today. on installation, it didn't work properly

(4) Demonstrative reference :-

* word that directly indicates a person/thing or few people & few things. The demonstrative words are that, those this, & these

Ex:- This is my books, Give me that blue water bottle.

(5) Ordinal Reference :-

Uses an ordinal (like first, one etc..,

Ex:- I visited a computer store to buy a printer. I have seen many of now I have to select one.