CMRIT
CELEBRATING 25 YEARS
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A+ GRADE BY NAAC

Internal Assessment Test 1 – Nov 2021

Sub:	Natural Language Processing					Sub Code:	18CS74, 15CS741 & 17CS741	Branch:	CSE		
Date:	13/11/2021	Duration:	90 mins	Max Marks:	50	Sem / Sec:	VII / A, B & C			OBE	

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1 (a)	What is Natural Language Processing? Explain two major approaches to NLP	[04]	CO1	L2
(b)	List and explain different levels of processing involved in NLP.	[06]	CO1	L2
2 (a)	Explain the components of transformational grammar	[8]	CO1	L2
(b)	Write the transformational grammar for the sentence – "The boy hit the girl."	[02]	CO1	L3
3 (a)	List and explain application of NLP	[10]	CO1	L2
4 (a)	List and explain two approaches of language modelling. List the various grammar-based model.	[06]	CO1	L2
(b)	Explain the challenges of NLP	[04]	CO1	L2
5 (a)	What is Government and Binding theory? Explain with an example	[04]	CO1	L2
(b)	Explain components and organization of GB	[06]	CO1	L2
6 (a)	Explain X-bar theory with example.	[10]	CO1	L2
7 (a)	Explain Lexical functional grammar (LFG) model for the sentence "She saw stars in the sky"	[10]	CO1	L3

6the X-bar theory.

1) It is the central concept of Government and Binding theory.

2) Instead of considering several phrase and sentence structure, X-bar theory considers Maximal projection of some head.
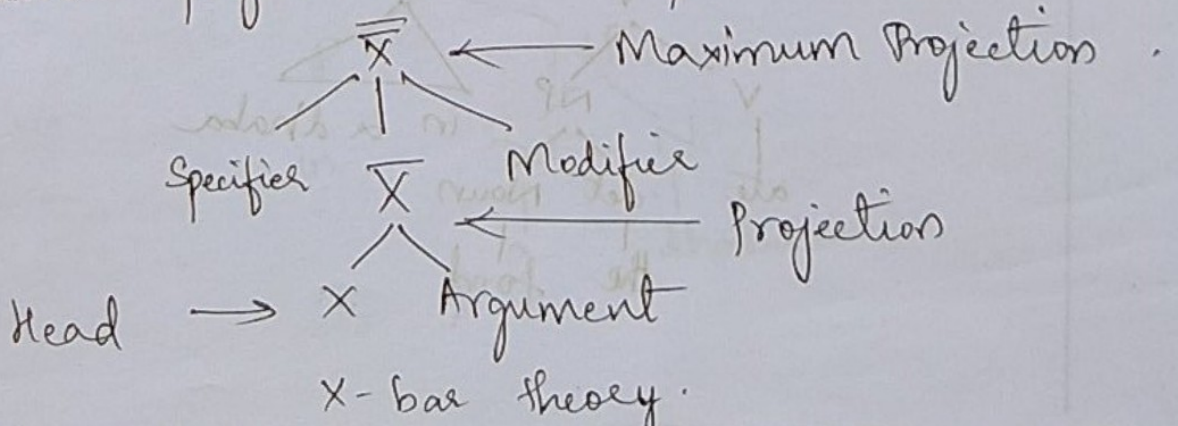
The maximal projections

(i) NP   Noun Phrase
(ii) VP   Verb Phrase
(iii) AP   Adjective Phrase
(iv) PP   Preposition Phrase

are the maximal projections of Noun, Verb, adjective and Preposition respectively.

3) Government and Binding theory considers pro two levels of projections namely:

(i) Projection of head at Semi-phrase level $(\bar{X})$
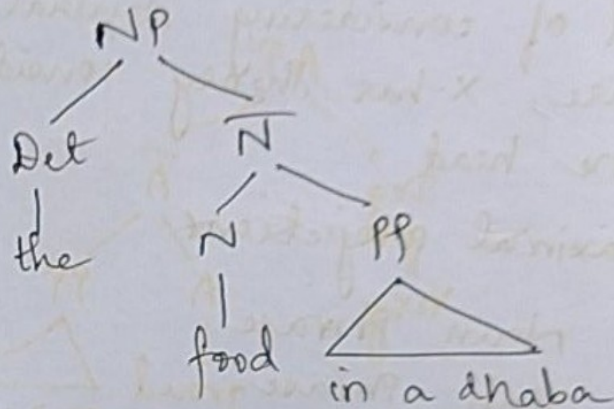(ii) Maximal projection at the phrase level $(\bar{\bar{X}})$

$$\bar{\bar{X}} \longleftarrow \text{Maximum Projection}$$

Specifier $\bar{X}$ ← Modifier

Projection

Head → X Argument

X-bar theory.

The components are:

(i) <u>NP</u> (Noun Phrase)

$$[_{NP} \text{ the } [_N \text{food}] [_{PP} \text{ in a dhaba}]]$$

```
              NP
             /  \
           Det    N
            |    / \
           the  N   PP
                |   / \
              food /   \
                  in a dhaba
```

Tree representation.

(ii) <u>VP</u> (Verb Phrase)

$$[[_V \text{ ate}] [_{NP} \text{ the food}] [_{PP} \text{ in a dhaba}]]$$

```
                VP
               /  \
              V     PP
             / \    / \
            V   NP /   \
            |  /\  in a dhaba
           ate det Noun
              |    |
             the  food
```

(iii) AP (Adjective Phrase)

[AP [Deg Very] [A proud] [PP of his country]]

AP
├─ Deg
│   └─ Very
└─ A
    ├─ A
    │   └─ proud
    └─ PP
        └─ of his country

(iv) PP (Preposition phrase)

[PP [P in] [NP [Det a] [H dhaba]]]

P
└─ P
   ├─ P
   │ └─ in
   └─ NP
      ├─ Det
      │   └─ a
      └─ N
          └─ dhaba
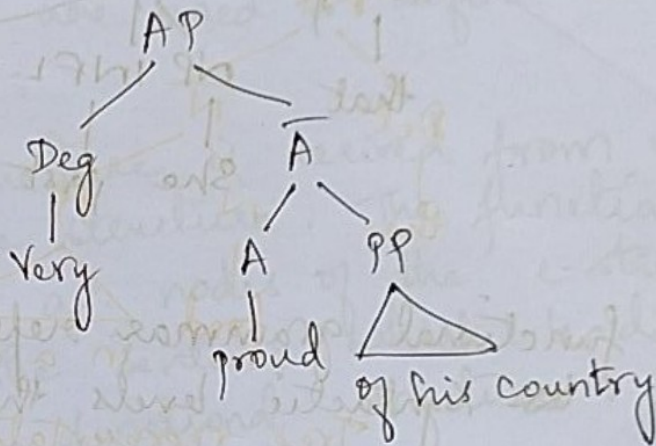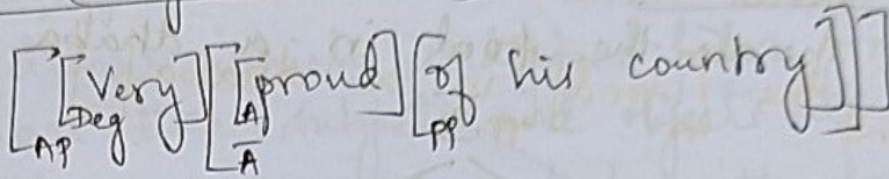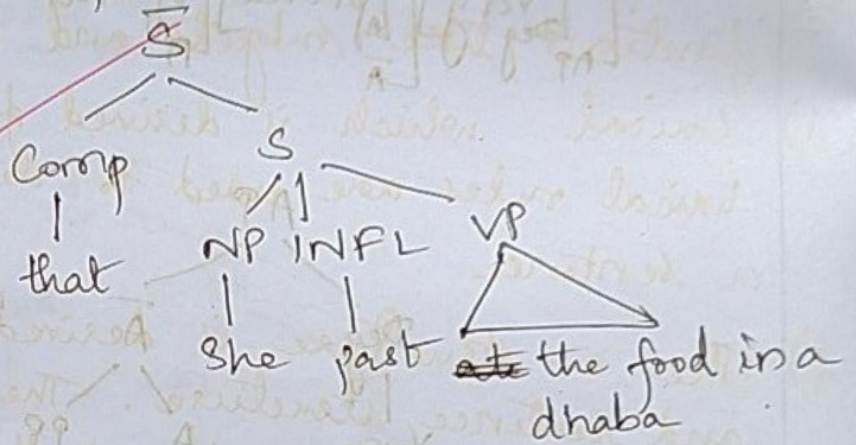
(v)

S (Sentence)

that she ate the food in a dhaba

7a) Lexical functional grammar represents a sentence in two syntactic levels that is

(i) c-structure (constituent structure)

(ii) f-structure (functional structure)

b) Kaplan proposed a concrete model for register of names and values which led to the formation of f-structure.

c) Bresnan was more concerned with dealing with linguistic issues. She said that this problem can be solved by lexical recursive rule.

d) The combination of both the theories led to the formation of lexical functional grammar.

e) Lexical functional has two parts

(i) Functional which is derived from grammatical functions like subjects and objects in a sentence.

(ii) Lexical which is derived from the fact that lexical rules are used to define the structure of a sentence.

f) The c-structure is derived from usual phrase and sentence structure. The functional specification is annotated as nodes of the c-structure. It is applied to the sentence which leads to the formation of functional structure.
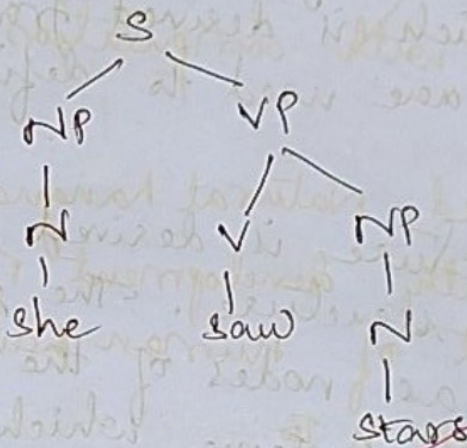
The CFG rules for 'She saw stars in sky' is

$$S \rightarrow NP \ VP$$
$$VP \rightarrow V \ \{NP\} \ \{NP\} \ PP^* \ \{S'\}$$
$$PP \rightarrow P \ NP$$
$$NP \rightarrow Det \ N \ \{PP\}$$
$$S' \rightarrow Comp \ S$$

S → Sentence, NP → Noun Phrase, VP → Verb Phrase V → Verb, P → Preposition, * - represents that the phrase can appear any number of times of can be blank also.

S' → clause, N → Noun

Det → Determiner, { } - optional

c-structure

She saw stars

```
            S
          /   \
        NP      VP
        |      /  \
        N     V    NP
        |     |    |
       she   saw   N
                   |
                 stars
```

(10)

f-structure

$$
\begin{bmatrix}
\text{subj} & \begin{bmatrix} \text{Pers} & 3 \\ \text{Num} & \text{SG} \\ \text{Gen} & \text{FEM} \\ \text{Case} & \text{NOM} \\ \text{Pred} & \text{'PRO'} \end{bmatrix} \\
\\
\text{Obj} & \begin{bmatrix} \text{Pers} & 3 \\ \text{Num} & \text{PL} \\ \text{Pred} & \text{'star'} \end{bmatrix} \\
\\
\text{Pred} & \text{'seé} < (\uparrow\text{subj}) \; (\uparrow\text{obj}) >
\end{bmatrix}
$$

1a) * NLP is called Natural Language Processing.

~~It is defined as~~

~~It deals with computation~~

* NLP deals with computation of

1a) * NLP is called Natural Language processing.

* It deals with development of computational modds as an aspect of Human language processing.

The main aim is to:

(i) Develop automated tools for language processing

(ii) To gain better understanding of Human communication.

The two approaches to NLP are:

(i) Rationalist - They follow early NLP. According to them humans are born with some language faculty. But the supporters argue that it is not possible for a child to learn such complex thing like Natural language.

(ii) Empericist - They are opposite to Rationalist. They donot believe in inborn language faculty. Instead they believe in some general organizational principles which include pattern recognition, generalization and association.

(4)

b) The different levels of processing in NLP are:

1. Lexical Analysis — It is the first and basic level of in language processing.

* It deals with analysis of words.
* Words are fundamental unit of Natural Language text.
* Lexical Analysis needs Morphological knowledge.
* Morphological knowledge — knowledge of structure and formation of words from basic units called morphemes.
* Morphemes — Meaningful morphological units that cannot be further divided.

Eg: -un, -break, -able → unbreakable.

2. Sytactic Analysis — It considers representation sequence of words as a single unit usually a sentence and finds its structure.

* It breaks down a sentence into words and tries to find the relationship between them.
* It captures the gramaticality and non-gramaticality of a sentence seeing the word order and number.
* Not all sentences are valid.

Eg: 'I went to the market' is a valid sentence.

whereas 'went I the to market' is not.

* Therefore it needs
* Syntactic Analysis requires syntactic knowledge
  - It considers sequence of words as a single unit and finds it structure. It also considers the constraints.

3. **Semantic Analysis** - It deals with the meaning of the language.

Eg. Kabir and Ayan are married
    Kabir and Suha are married

* In the first sentence, it states that both are married.
* But in the second sentence it states that both Kabir and Suha are married to each other.
  This implies, the meaning of the sentence not only depends on the meaning of the words alone.

* It requires pragmatic knowledge
* Pragmatic knowledge - It deals with how the language is used in different ways for different purposes.

4. **Discourse Analysis** - It deals with interpreting the structure and meaning of even larger units like paragraphs, documents which consists of words, sentences etc.

* Discourse analysis needs:

(i) Resolution of anaphoric references.

(ii) Discourse knowledge - It tells how the one sentence is determined by the preceeding sentence.

Eg: Pronoun refers to the preceeding noun.

* Resolution of Anaphoric references need pragmatic knowledge.

Eg: 1) The district administration refused to give the trade union to give permission for the meeting as they feared violence.

2) The district administration refused the trade union to give permission for the meeting as they opposed government.

* In the 1st sentence, they refers to the district administration whereas in the 2nd sentence, they refers to the trade union.

⑥

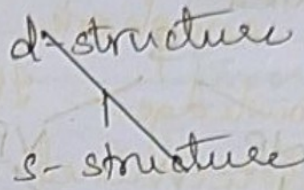5a) * Government and Binding theory was proposed by Noam Chomsky.

* Transformational grammar has two levels

(i) surface structure level

(ii) deep root structure level

* GB theory then renamed these surface str

level as s-level and deep root level as d-level.

$$d\text{-structure}$$
$$\diagdown$$
$$s\text{-structure}$$

* In addition to these two levels, Government and Binding theory identified of defined two more levels called phonetic form and logical form.

d-structure
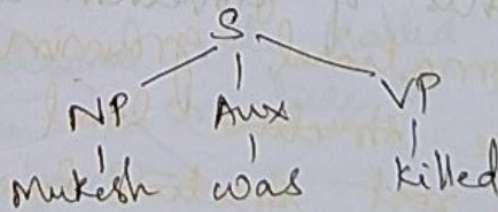|
s-structure

phonetic   logical
form   form

(4)

* Language is som representation of some meaning in sound form.

meaning - Logical from
sound - phonetic form.
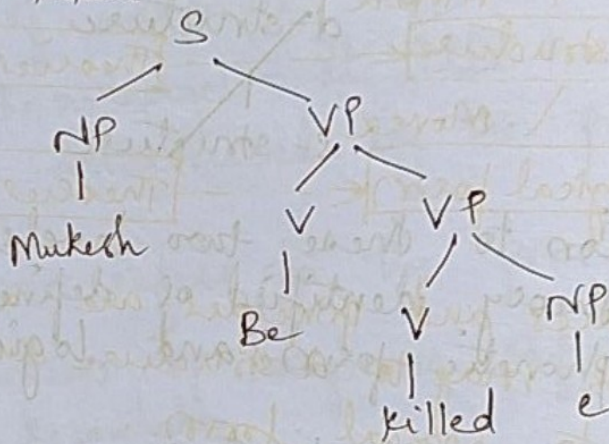
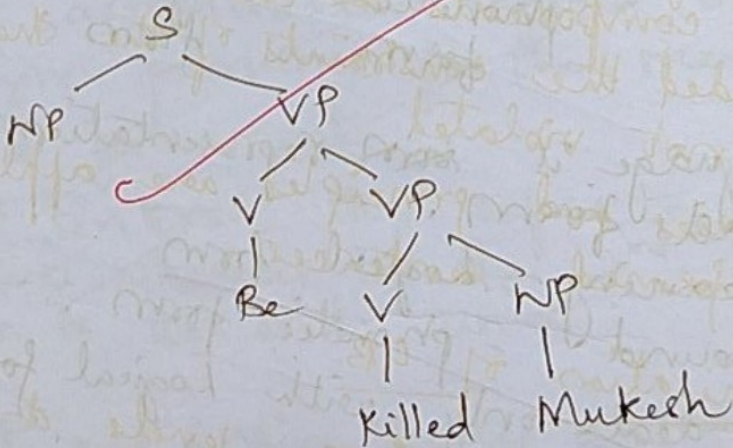* GB deals only with Logical form.

Eg:-   Mukesh was killed.

```
              S
         ╱    |    ╲
       NP    Aux    VP
        |     |      |
     Mukesh  was   killed
```

*

Representation in Transformational grammar

S-structure

Mukesh was Killed.

```
                S
              /   \
            NP      VP
            |      /  \
         Mukesh   V    VP
                  |   /  \
                  Be V    NP
                     |    |
                  killed  e
```
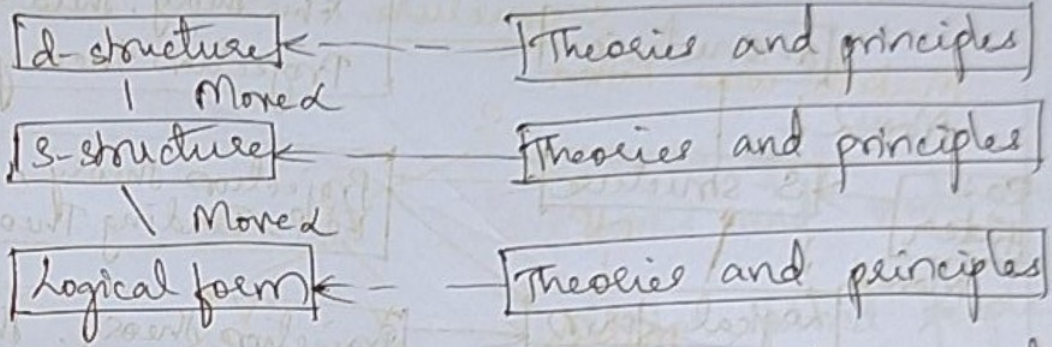
d-structure

Killed Mukesh (e)

```
            S
          /   \
        NP      VP
              /   \
             V     VP
             |    /  \
            Be   V    NP
                 |    |
              Killed Mukesh
```

b) Components of Government and Binding theory.

d-structure	←	-	-	Theories and principles

| Moved

| s-structure | ← | - | Theories and principles |

| Moved
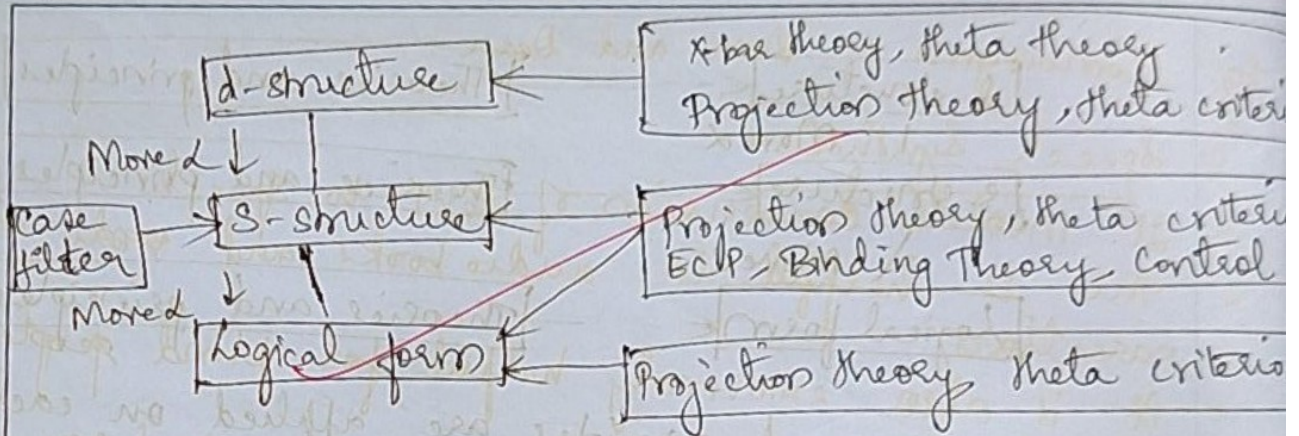
| Logical form | ← | - | - | Theories and principles |

* The theories and principles are applied on each level or map to d-structure and s-structure and further to logical form.

* Transformational rule called 'Move α' is used to transform the d-structure to s-structure and then to logical form.

* The components can be transformed anywhere provided the constraints of the theories and principles are not violated.

* Theories and principles are applied to prevent the ill-formed sentences.

(2) organisation of GB.

* It consists of three levels d-structure, s-structure and logical form.

* The transformational rule called 'Moved' is used to transform the d-structure to s-structure and then to logical form.

* Case filter is applied to s-structure

* The theories and principles are further defined.

```
                    ┌──────────────┐         ┌─ *bar theory, theta theory
                    │  d-structure │◄────────┤  Projection theory, theta criteri
                    └──────────────┘         └─
        More α ↓
  ┌──────┐   ┌──────────────┐               ┌─ Projection theory, theta criteri
  │ Case │──►│  S-structure  │◄──────────────┤  ECP, Binding Theory, Control
  │filter│   └──────────────┘               └─
  └──────┘
        More α ↓
        ┌──────────────┐                    ┌─ Projection theory, theta criterio
        │ Logical form │◄───────────────────┤
        └──────────────┘                    └─
```

3. The applications of NLP are:

a) Machine translation

b) Speech Recognition

c) Speech synthesis

d) Natural language inferance to data base

e) Information Retrival

f) Information extraction

g) Question Answering System

h) Text summarization.

a) Machine Translation — * It involves automated translation of text from one language to other. For this the knowledge of words, phrases and grammar in both the languages is necessary.

b) Speech Recognition — It is the conversion of ~~text to speech~~. Speech to text.

* During this process problems arise due to pronoun ciation.

Eg: Homonyms :- Deer and Dear.

c) **Speech Synthesis**

* It involves conversion of text to speech.
* This is helpful in audio books and video narrations.
* It is also extremely helpful for blind people.

d) **Natural Language Inference to database.**

* This involves writing queries for the database.

e) **Information Retrieval**

* Retrieving the information or document based on users query is called information retrieval.
* It is used to retrive information from large set of text in electronic form.
* It returns a hierarchical set of list of documents.
* It is unstructured.

f) **Information Extraction**

* Extracting the information from a document is called information extraction.
* Given a
* It involves searching an extracting information from a particular document.

g) Question Answering System

* Given a question and set of documents, the system tries to find the answer.

* This involves comparison of the given documents.

h) Text summarization

* Summaries of the documents are prepared.
* This requires syntactic, semantic level of processing.

2a) The components of transformational grammar are:

i) Phrase structure grammar.

* This contains grammar contains a set of rules which generate sentences and assign structural description to them.

* Rules for phrase structure grammar.

$$S \rightarrow NP + VP$$
$$VP \rightarrow V + NP$$
$$NP \rightarrow Det + Noun$$
$$V \rightarrow Aux + Verb$$

Noun - cat, bat etc
Verb - eat, catch etc
Det - the, a, an
Aux - will is can

NP - Noun Phrase
VP - Verb Phrase
V - Verb
N - Noun
S - Sentence

Eg:- Pooja Plays Veena

$$S \rightarrow NP + VP$$
$$VP \rightarrow V + NP$$
$$NP \rightarrow Det + Noun$$
$$V \rightarrow Aux + Verb$$
$$NP \rightarrow Pooja, Veena$$
$$VP \rightarrow Plays$$

2) <u>Transformational rules</u>.
* It is used to transform one phrase maker to another.
* It is at the terminal of the string generated by Phrase structure grammar.
* It contains more than one symbol on left hand side of the rule.

Eg:- The police will catch the snatcher
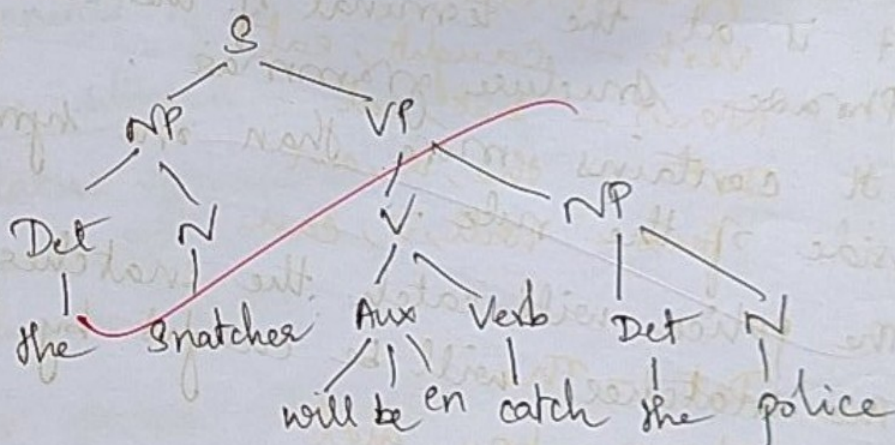The snatcher will be caught by the police

Rules

$$NP1 + Aux - V - NP2 \rightarrow NP2 - Aux + be + en - \cancel{of} V$$
$$- by - NP1$$

Eg:- The Police will catch the snatcher
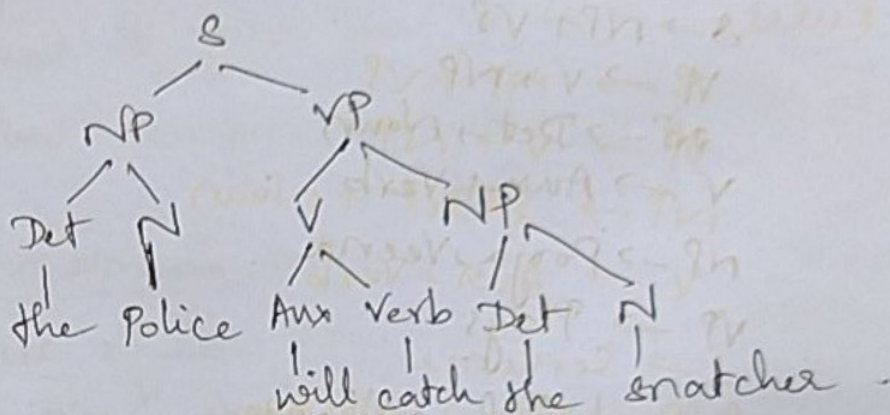
The snatcher will be caught by the police

3) Morphophonemic rules.

* The transformational rule converts 'en + catch' to 'catch + en' which is the converted to caught by morphophonemic rule

'en + catch' + 'catch + en' ⟶ caught

b)   The boy hit the girl

Rules

$$S \rightarrow NP + VP$$
$$VP \rightarrow V + NP$$
$$NP \rightarrow \cancel{N} Det + Noun$$
$$V \rightarrow Aux + Verb$$

S → Sentence
NP — Noun phrase
VP — Verb phrase
~~V → Verb~~
~~N — Noun~~
Verb — caught, eat
Noun — bat, eat
Det — the, a, an
Aux — will, is, can

②

Det — The
Noun — boy, girl
Verb — hit

Q1. a. Language is the primary means of communication by humans. It is the tool we use to express the greater part of our ideas and emotions. It shapes thought, has a structure and carries meaning.

Natural Language Processing is concerned with the development of computational models of human language processing. There are two reasons for such development.

★   To develop automated tools for language understanding

★   To gain a better understanding of human communication.

Development of computational models with human natural language processing abilities requires a knowledge of how humans acquire, store and process language. It also requires a knowledge of the world and of language.

Historically there have been two major approaches to NLP — rationalist and empiricist approach. Early NLP research took a rationalist approach which assumes the existence of a language faculty in the human brain. Supporters of this approach believe that it is not possible for a child to learn a complex thing like natural language with limited sensory inputs. Empiricists do not believe in the existence of a language faculty. They believe in the

evidence of general organisation principles such that pattern recognition, generalisation and association. Learning of detailed structures

(4). can be done through the application of these principles on sensory inputs available to the child.

Q1.b. Language Processing has different levels of processing each associated with a different set of knowledge. They are:

1. Lexical Analysis 2. Syntactic analysis 3. Semantic analysis 4. Discourse analysis 5. Pragmatic analysis

1. **Lexical Analysis**

* It is the simplest analysis involving the analysis of words.

* Words are the fundamental unit of any natural language.

* Word level analysis requires morphological knowledge i.e. how words are formed from morphemes.

* Building of words from morphemes are language specific.

2. **Syntactic analysis**

* It is the analysis of a group of words usually a sentence and finds its meaning.

* A sentence is decomposed to words and are found how words are

related to each other.

* It finds grammatically correct and incorrect sentences by looking at constraints such as word order, etc.

* It requires syntactic knowledge.

* eg: 'I went to the market' is a valid sentence whereas 'went the market I to' is not.

3. ## Semantic analysis

* It is concerned with a meaning of a sentence.

* It's main idea is to create a meaningful representation of linguistic inputs.

* The concern is to take natural language utterances and map them into some representation of particular meaning

eg: Kahia and Ayan are married
Kahia and Suha are married

* In the second sentence we can see that Kahia and Suha are married but this meaning does not come in the first sentence.

* Syntactic structure and compositional semantics is not sufficient and also requires pragmatic knowledge.

4. <u>Discourse analysis</u>

* It is the analysis of paragraphs and document level in terms of sentences, words and clusters.

* The idea is to resolve araphoric references and identification of discourse knowledge.

* It decides how the meaning of a sentence depends on the previous sentence. eg: a pronoun referencing noun.

* eg: The district administration refused to give the trade unione permission for the meaning because they feared violence.

The district administration refused to give the trade unions permission for the meaning because they opposed government.

* Here 'they' is an araphoric reference.

5. <u>Pragmatic analysis</u>

* It deals with the purposeful use of sentences. It requires world knowledge - knowledge beyond context of text.

* The Cyc Project at the University of Austin finds the use of world knowledge

Q3.   The main applications of NLP are:

1.   Machine Translation

2.   Speech Recognition

3.   Speech Synthesis

4.   Natural Language Interfaces to databases

5.   Information Retrieval

6.   Information Extraction

7.   Question Answering

8.   Text Summarisation

1.   Machine Translation

★   It deals with the automatic translation of text from one human language to another.

★   It requires an understanding of words and phrases, grammars, semantics of the two languages and also of world knowledge.

2.   Speech Recognition

★   It deals with the mapping of acoustic speech signals into a set of words.

★ Boblems may arise due to wide variation in pronunciation and acoustic differentations in hearing.

3. Speech Synthesis

★ It deals with the production of speech on analysis of text.

★ It carried out mails from telephone or even a story book.

★ Natural Language Processing is required for this because for producing speech, text has to be processed.

4. Natural Language Interfaces for Databases

☆ It involves querying a structured database using natural language sentences.

5. Information Retrieval

★ The main idea is to identify documents matching users query.

★ Indexing, wordsense disambiguation, querying modifying and knowledge bases are effectively used for this.

6. Information Extraction

* An IE system captures and outputs factual information from a document.

7. Question Answering

* Given a question and set of documents, this system finds the precise answer or the relevant portion of text.

* It requires more NLP than IR or IE system

8. Text Summarisation

* It is concerned with the summarisation of text and requires syntactic, semantic and discourse knowledge.

Q.4a. The two approaches of language modelling are:

1. Grammar based model

2. Stastical based model

The explanation are as follows:

1. ## Grammar based language modelling

   ★ Language modelling is viewed as a problem of grammar rules

   ★ We define a grammar that can handle language.

   ★ We can identify grammatical sentences from non-grammatical ones.

   ♦ It uses grammar of a language to create its model.
   Its grammar is represented as syntactic structures of the language.

   ★ For example, a sentence consists of noun, verb phrases. This model uses these phrases and the relationships between each other to create a grammar of the sentence.

2. ## Stastical language modelling

   ♦ Language modelling is viewed as a problem of probability estimation.

   ★ We can identify pattern rules in a sentence statistically.

   ♦ It is the ability to identify a sentence as a probability measure.

*   It involves building a model by training it from a corpus.

*   To capture the regularities of a language, the training corpus must be sufficiently large.

Statistical language modelling aims at capturing regularities of a language to improve its performance in various natural language applications

▶   It has found many applications in machine translation, information retrieval, question answer finding and so on.

The various grammar based models are:

1.   Generative grammar

2.   Heirachial grammar

3.   Government and Binding

4.   Lexical Functional grammar

5.   Paninian grammar.

Q.4b.   The challenges of NLP are as follows:

1. Problem of representation and identification

2. The inability to capture all the sources of required knowledge is another challenge.

3. Greatest challenge is to identify the meaning of its semantics

4. Idioms, metaphors and ellipses add to the complexity of a sentence.

5. Quantifier scoping - The scope of a quantifier may not be clear and may create confusions.

6. Ambiguity

   There are two levels of ambiguity:

i. First level - Word level eg:- bat, bark

   Ambiguity may be in its meaning or subject verb agreement

ii. Second level - sentence level eg: Stolen rifle found by tree

   The words have meaning but the sentence does not. called structural ambiguity. Verb subcategorisation may solve this.

Word level — eg: bat, bark ⟶ Word sense disambiguation

(4)

Sentence level — eg: Stolen rifle found by tree ⟶ Verb subcategorisation

---

Q6.

x bar theory is the central concept of Government and Binding.

* Instead of defining several phrase structure rules and sentence structure defined by a different set of rules, it can be maximally projected of some head.

* The maximal projections are as follows:

i. The Noun Phrase is maximal projection of a noun.

ii. The Verb Phrase is maximal projection of verb.

iii. Adjective Phrase is maximal projection of adjective

iv. Preposition Phrase is a maximal projection of preposition.

* The heads of maximal projection, can be represented by X. where X = noun, verb, adjective, preposition.

* There are two levels of projection:

1. The maximal projection at the phrase level (represented by $\overline{\overline{x}}$)

2. The projection of head at the subphrase level (represented by $\overline{x}$)

← maximal projection

$\overline{\overline{X}}$

Specifier $\overline{X}$ Modifier ← Projection of head

Head ⟶ X   Argument
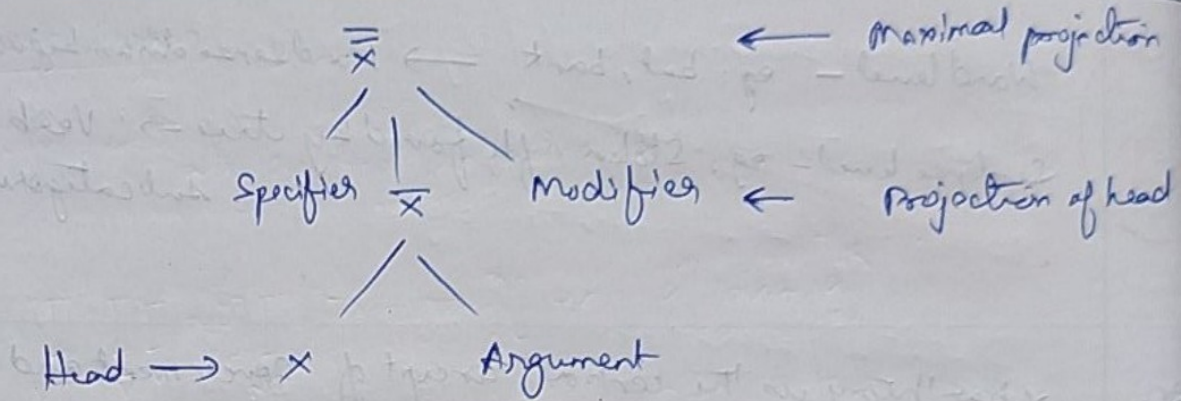
The projection of various types are as follows:

i. **Noun Phrase**

the food in a dhaba

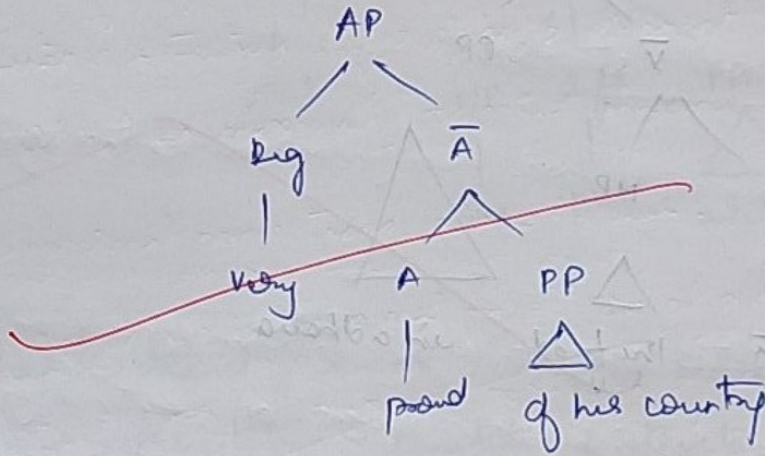Therefore,

[[the] [[food] [in a dhaba]]]
NP Det   N̄ N   PP

NP
/ | \
Det   N̄
|
The   N   PP
| △
food   in a dhaba
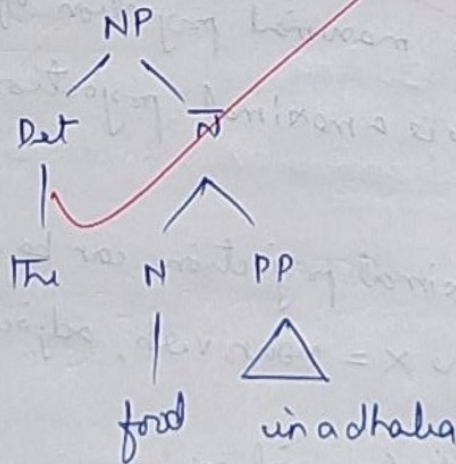
ii. Adjective Phrase

. Very proud of his country

[AP [Deg very] [Ā A [proud] [PP of his country]]]

AP
├ Deg
│ └ very
└ Ā
   ├ A
   │ └ proud
   └ PP
      └ of his country
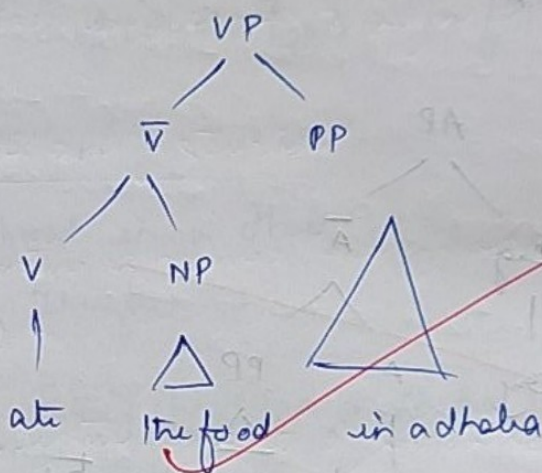
iii). **Verb Phrase**

ati the food in a dhaba

[[[ati] [the food]] [in a dhaba]]
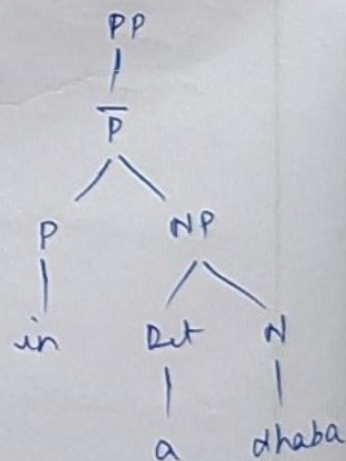VP V̄ V   NP   PP

iv. **Preposition Phrase**

in a dhaba

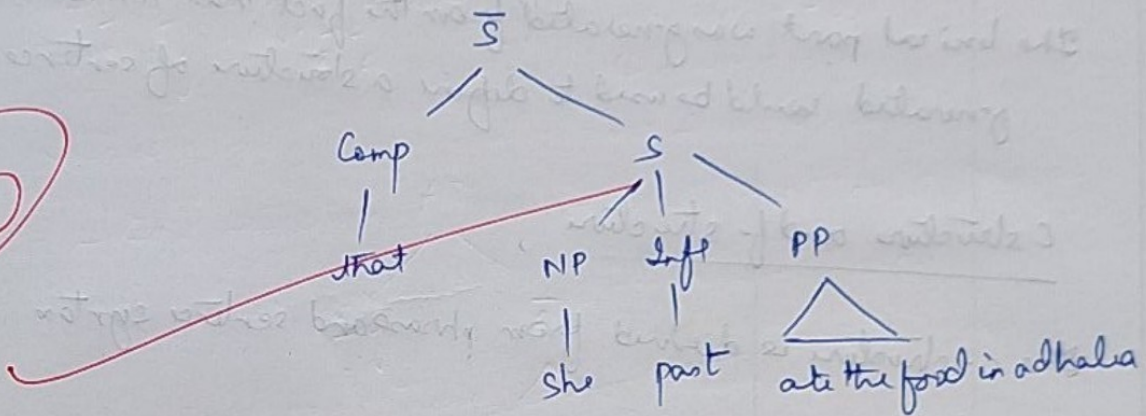[[[in) [[a] [dhaba]]]]
PP P̄ P   NP Det N

v. Sentence

that she ate the food in a dhaba

[[That] [She] [ate] [The food in a dhaba]]]
S   Comp   S^{NP} Inf^l PP

10

Comp
|
that

NP
|
she

Inf^l
|
past

PP
ate the food in a dhaba

Q7. Lexical Functional grammar is represented by two syntactic
levels of representation — constituent structure (c-structure)
and functional structure (f-structure)

Kaplan introduced a very complex form of register names and
values used in the ATN Theory which formed the basis
for functional part of the sentence

Bresnan on other hand was interested to solve problems of
linguistic issues such as active and passive sentences.
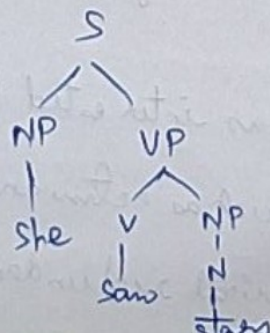She clarified that lexical redundancy rules can be used to

Solve each problems.

* The functional structure is derived from the various grammatical functions such as subject and verb, object and also the roles played by each argument in a sentence.

* The lexical part was generated from the fact that lexical rules generated could be used to define a structure of sentence.

c-structure and f-structure

* The c-structure is derived from phrase and sentence syntax constructs.

* The f-structure consists of annotations of node on the c-structure which when applied on a sentence gives us the required structure.

. The c structure and f-structure are as follows.

She saw stars in the sky

```
            S
          /   \
        NP     VP
        |     /  \
       She   V    NP
             |    |
            saw   N
                  |
                 stars
```

$$
\begin{bmatrix}
\text{subj} & \begin{bmatrix}
\text{Pers} & 3 \\
\text{Num} & SG \\
\text{GEN} & FEM \\
\text{Case} & Nom \\
\text{Pred} & \text{'PRO'}
\end{bmatrix} \\[2em]
\text{obj} & \begin{bmatrix}
\text{Pers} & 3 \\
\text{Num} & PL \\
\text{Pred} & \text{'Star'}
\end{bmatrix} \\[2em]
\text{Pred 'see'} < (\uparrow\text{subj})(\uparrow\text{obj}) >
\end{bmatrix}
$$

(10)