

Orientation of Natural Language Processing (18CS743)

By

Prof Shruthi.J
Assistant Professor
Department of CSE
BMSITM

Course Objectives

- ▶ Learn the techniques in natural language processing
- ▶ Be familiar with the natural language generation
- ▶ Be exposed to Text Mining
- ▶ Understand the information retrieval techniques

Contents

- ▶ Overview and language modeling
- ▶ Word level and syntactic analysis
- ▶ Extracting relations from text
- ▶ A case study in natural language based web search
- ▶ Information retrieval and lexical resources

Course Outcomes

- ▶ Analyze the natural language text
- ▶ Generate the natural language
- ▶ Do Text mining
- ▶ Apply information retrieval techniques

Text Books:

1. Tanveer Siddiqui, U.S. Tiwary, “Natural Language Processing and Information Retrieval”, Oxford University Press, 2008.
2. Anne Kao and Stephen R. Poteet (Eds), “Natural LanguageProcessing and TextMining”, Springer-Verlag London Limited 2007.

Reference Books:

1. Daniel Jurafsky and James H Martin, “Speech and Language Processing: Anintroduction to Natural Language Processing, Computational Linguistics and SpeechRecognition”, 2nd Edition, Prentice Hall, 2008.
2. James Allen, “Natural Language Understanding”, 2nd edition, Benjamin/Cummingspublishing company, 1995.
3. Gerald J. Kowalski and Mark.T. Maybury, “Information Storage and Retrieval systems”, Kluwer academic Publishers, 2000.

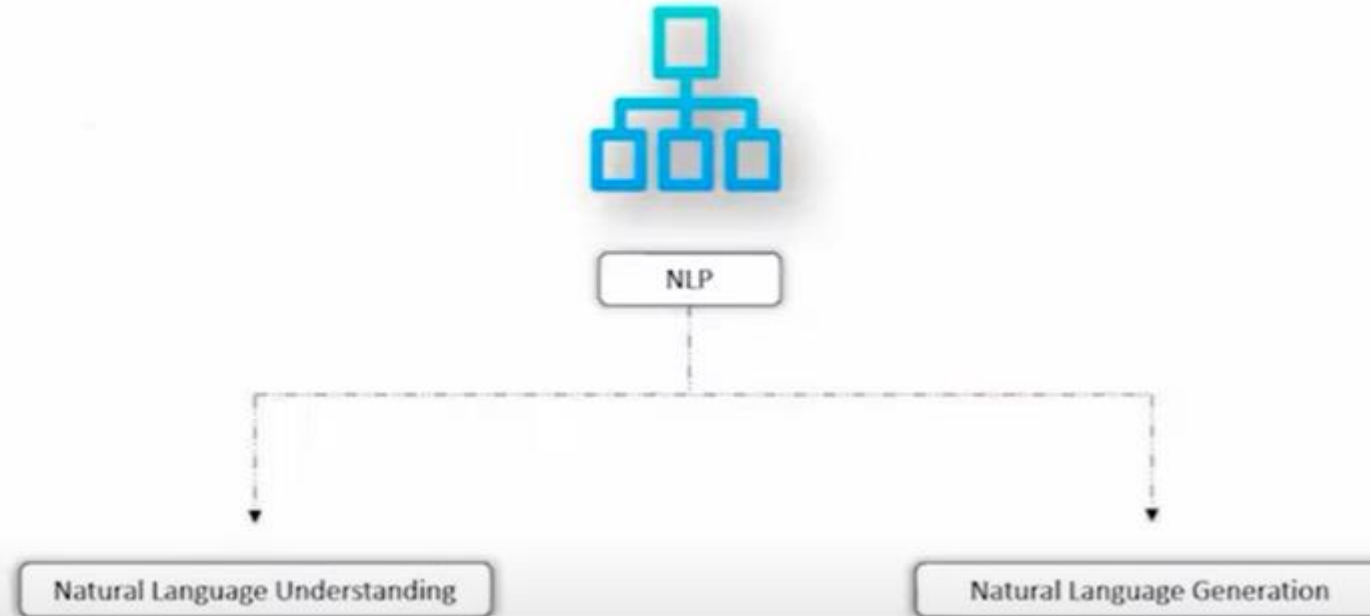
Understanding Natural Language



Natural Language is a language which has developed naturally in humans

Natural Language Processing is the ability of a computer program to understand human language as it is spoken

Components of Natural Language Processing



Natural Language Understanding

NLU deals with Understanding the input given by user as a part of natural language

After surprising the hosts in the first Test, Sri Lanka made a positive start to the second Test as well by bowling South Africa out for 222 before slightly losing their advantage towards the end of the day's play. The visitors lost three wickets in the final session before stumps, still trailing South Africa by 162 runs. Just as Sri Lanka's bowlers were on top of South Africa's batsmen for majority of their innings, South Africa's bowlers returned the favour. The only difference being Sri Lanka's batsmen found ways to negate their attack and keep them at bay, led by Lahiru Thirimanne's unbeaten 25, but only for a while as Dale Steyn and Kagiso Raba kept things



Natural Language Generation

NLG deals with producing written or spoken language from raw data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa



This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type

NLP

- ▶ It is concerned with the development of computational models of aspects of human language processing
- ▶ Two main reasons:
 1. To develop automated tools for language processing.
 2. To gain a better understanding of human communication.

Origins of NLP

- ▶ Natural Language Understanding
 - Originated from Machine Translation research
 - Involves only the interpretation of language
- Natural Language Processing
 - Involves both understanding (interpretation) and generation(production) of language
 - Includes Speech processing

Phases of analysis

- ▶ Lexical analysis
- ▶ Syntactic analysis
- ▶ Semantic analysis
- ▶ Discourse analysis
- ▶ Pragmatic analysis

Lexical analysis/word level analysis

- ▶ It involves in identifying and analysing the structure of words
- ▶ Lexion of a language means the collection of words and phrases in a language
- ▶ Deriving lexical analysis- the whole text is divided into paragraphs, sentences and words

Syntactic analysis[parsing]

- ▶ It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among them.
- ▶ Ex **The school goes to boy** {rejected by a syntactic analysis}

Semantic analysis

- ▶ It involves in identifying the exact or dictionary meaning from text i.e it involves meaningful formulation of sentences
- ▶ Ex: **hot ice-cream** { not correct}

Discourse analysis

- ▶ The meaning of any sentence depend on the meaning of the sentence that precedes it and may also influence the meaning of the sentence that follow it

Ex: **She wanted it** { not clear unless we are aware of the previous sentence}

Pragmatic analysis

- ▶ The structure representing what is said is reinterpreted to determine what was actually meant .
- ▶ It involves deriving aspects of language which requires real world knowledge.
- ▶ Ex: “Do you know who I am” { Machine may not understand the expression behind the sentence}

Challenges in NLP

1. Ambiguity:

It refers to any sentence in a language with multiple representation/multiple interpretation

Types:

a. Lexical ambiguity:

It represents the words that can have multiple lexical properties

Ex: In English the word "Back"

Backstage– back is noun

Back door ----Back is adjective

Challenges in NLP

b. Syntactic ambiguity:

The sentences can be parsed in multiple syntactical form

Ex: **Multiple parse tree representation of a single tree**

c. Semantic ambiguity:

Sentences with multiple meaning or interpretation

Ex :1. **Rama went to a bank**

which bank? Financial or river bank

2. **I saw the boy with a telescope**

Who has the telescope? Boy or I

Challenges in NLP

► 2. Inability of complete knowledge about NL:

- It is almost impossible to gain complete knowledge about a natural language.
- So it is difficult to write algorithms or programs

3. Quantifier scoping:

The scope of quantifier is often not clear so it creates problem in processing ex: small, big, some , large , the each etc)

NLP applications

1. Machine Translation: this refers to automatic translation of text from one human language to another
2. Speech Recognition: Process of mapping of acoustic speech signals to a set of words
3. Speech synthesis: automatic production of speech . Such systems can read out mails on telephone, or even read out a storybook for you.

NLP applications

4. Natural Language Interfaces to Databases: it allows querying a structured database using natural language sentences.
5. Informational Retrieval: It is identifying documents relevant to a user's query Indexing(stop word elimination,stemming, phrase extraction etc) , word sense disambiguation, query modification and knowledge bases have also been used in IR system to enhance performance,e.g..by providing methods for query expansion

NLP applications

- ▶ WordNet, Roget's Thesaurus are some of the useful lexical resources for IR research
- ▶ 6. Information Extraction: captures and outputs factual information contained within a document. It identifies a subset of documents in a large repository of text database eg- library scenario
- ▶ 7. Question answering: it attempts to find the precise answer, or at least the precise portions text in which the answer appears
- ▶ 8. Text Summarization: deals with the creation of summaries of documents and involves syntactic, semantic, and discourse level processing of text

Language Modeling

- ▶ Various types of languages and their modelling

- ▶ Language Model:

1. Automatic processing of language requires the rules and exceptions of a language to be explained to the computer

- In order to process NLs through computer based programs, some

representation model is required to be built called a language model

- ▶ Language model has 2 approaches:

1. Grammer based

2. Statistical based

1. Grammar Based LM

- ▶ It uses the grammar of a language to create its model.
- ▶ It represents the syntactic structure of language
- ▶ Grammar consists of hand coded rules defining the structure and ordering of different linguistic units in a language (phrase,sentence etc)
- ▶ Various computational grammars proposed are
 - A) Transformational grammar (Chomsky 1957)
 - (B) Lexical functional grammar(Kaplan & Bresnan 1982)
 - (C) Government and binding (Chomsky 1981)

A) Transformational grammar :

- ▶ It considers grammar to be a system of rules that generate exactly those combinations of words that form grammatical sentences in a given language
- ▶ It involves the use of defined operations to produce new sentences from existing ones
- ▶ Example: Transformational grammar relates the **active sentence** "John read the book" with its corresponding **passive** "The book was read by John"

► It assumes each sentences in a language has 2 levels of representation:

1. Surface level(S-level) structure

2. Deep level (D-level)structure

3. Example:

S1) I know a man who flies planes(S-structure)

S2) I know a man.The man flies aeroplanes(d-structure)

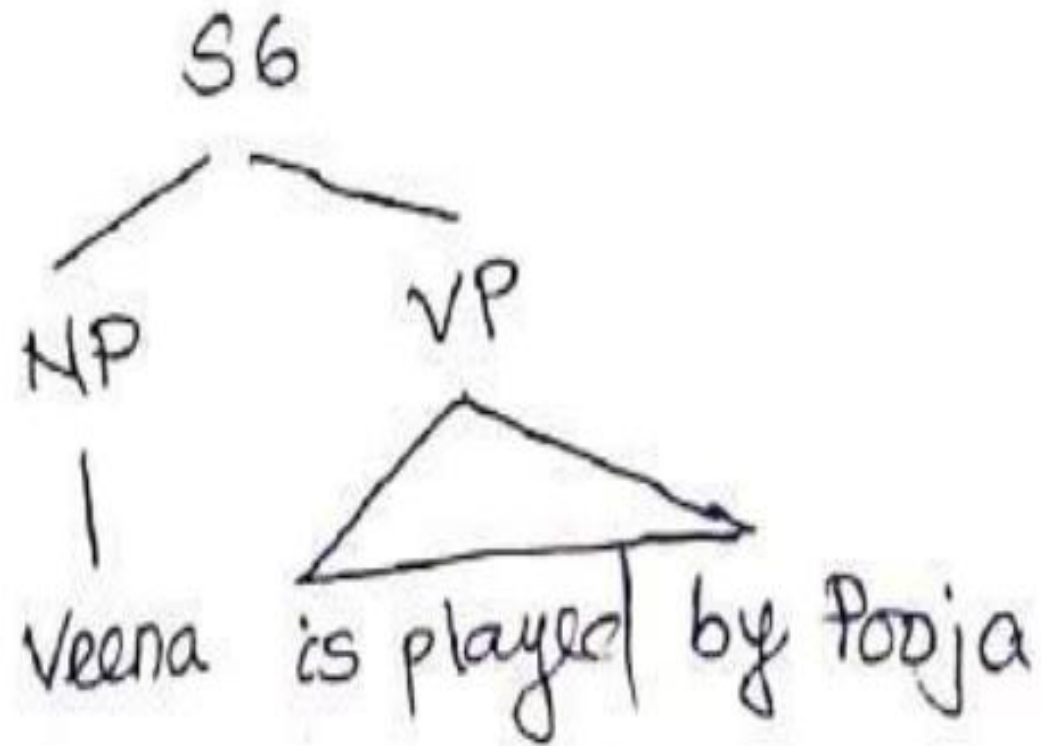
Another set of Ex:

S3) Flying airplanes can be dangerous(S-structure)

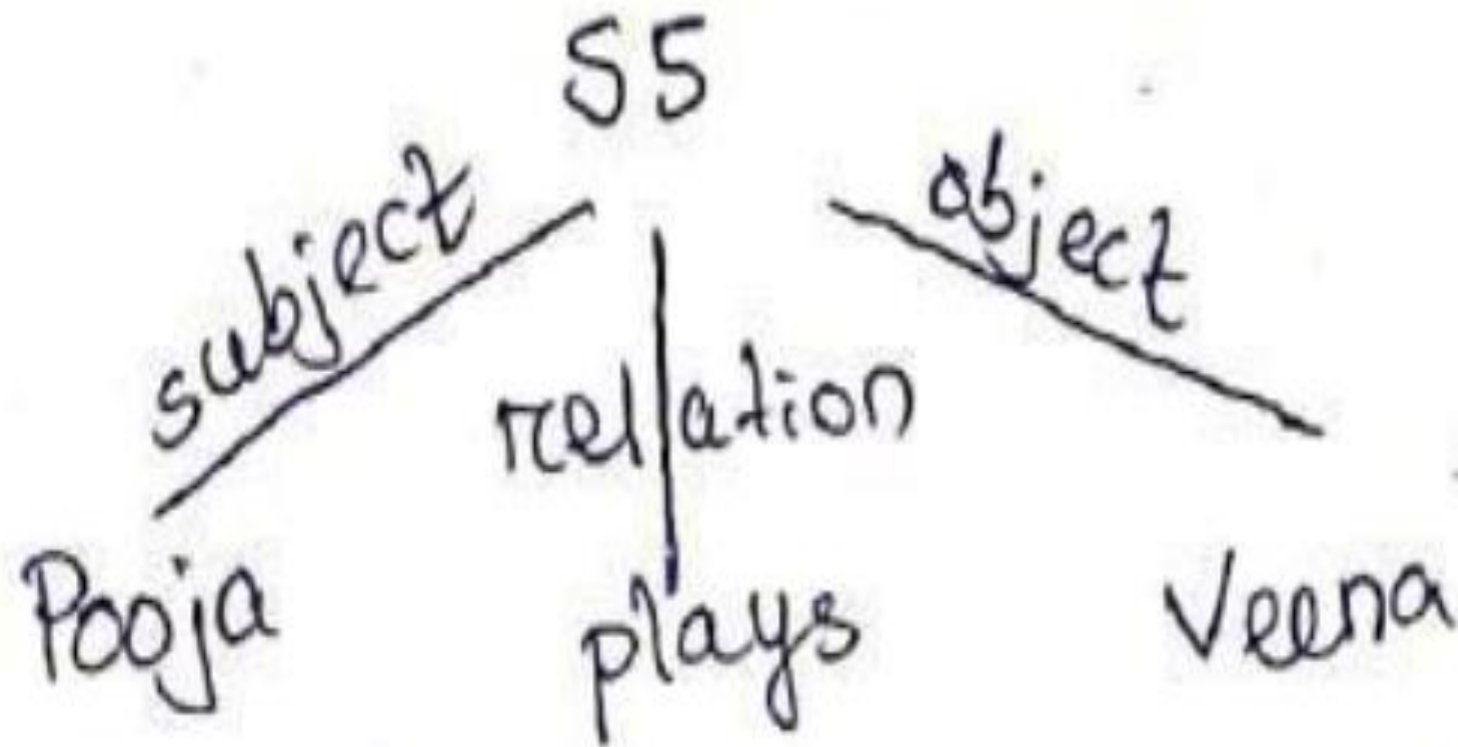
S4) Airplanes can be dangerous(d-structure)

Or To fly airplanes can be dangerous(d-structure)

- ▶ A deep structure can be transformed into a number of ways to generate different surface level representation
- ▶ The mapping from deep structure to surface structure is carried out by transformational grammar
- ▶ Example:
- ▶ S5) Pooja plays Veena
- ▶ S6) Veena is played by Pooja



[s-level structure of S5 & S6]



[d-level structure of S5]

► Transformational grammar has 3 components:

1) Phase- structure grammar:

It consists of rules that generate natural language sentences and assign a structural description to them

Example: Set of rules to form sentences

$S \rightarrow NP + VP$

$VP \rightarrow V + NP$

$NP \rightarrow Det + Noun$

$V \rightarrow Aux + Verb$

$Det \rightarrow the, a, an, \dots$

$Verb \rightarrow catch, write, eat, \dots$

$Noun \rightarrow police, snatcher, \dots$

$Aux \rightarrow will, is, can, \dots$

► 2) Transformational rules:

- It is the set of rules which transform one surface representation into another
- Example: active to passive voice
- Plays-→ is played by
- Apply -→ applied

3) Morpho-Phonemic rules:

These rules match each sentence representation to a string of phonemes

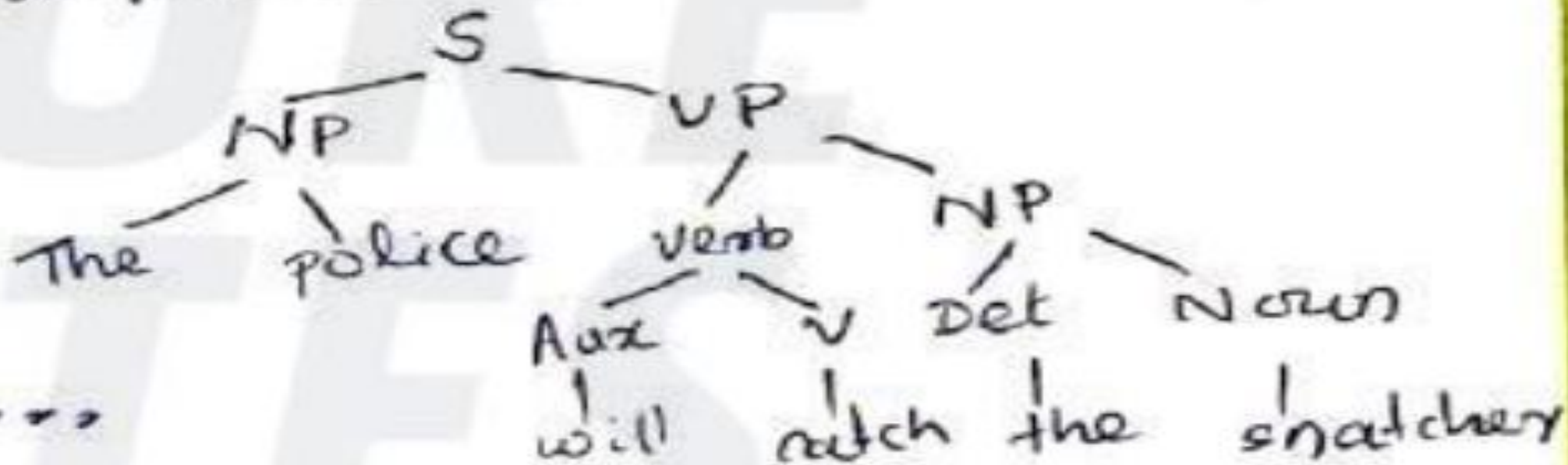
Rule relating active & passive sentences is

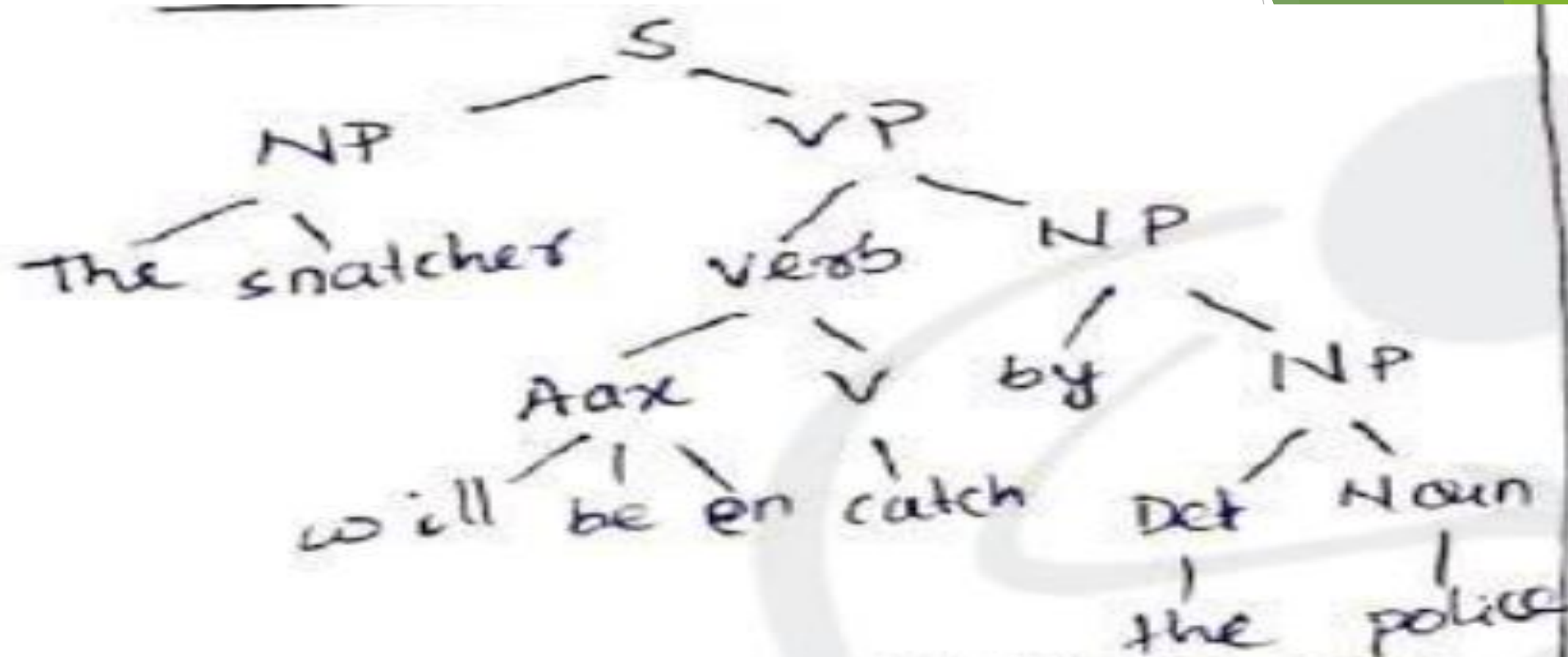
$NP_1 - Aux - V - NP_2 \rightarrow NP_2 - Aux +$

$be + en - V - by + NP_1$

Eg:

The police will catch the snatchers





the + snatcher + will + be + en +
catch + by + police

the + snatcher + will + be
catch + by + police



another transformation 2)
rule will recoder
ent + catch to
catch + en



catch + en will be
converted to caught
by morphophonemic
rule.

Issues in transformational grammar

- ▶ They have hundreds of re-writing rules, which are language-specific and also construct specific (different rules for active/passive voice sentence)
- ▶ Generation of a complete set of rules covering all language is a challenging task

B) Government and Binding (GB)

- ▶ What are the benefits of GB over transformational grammar?

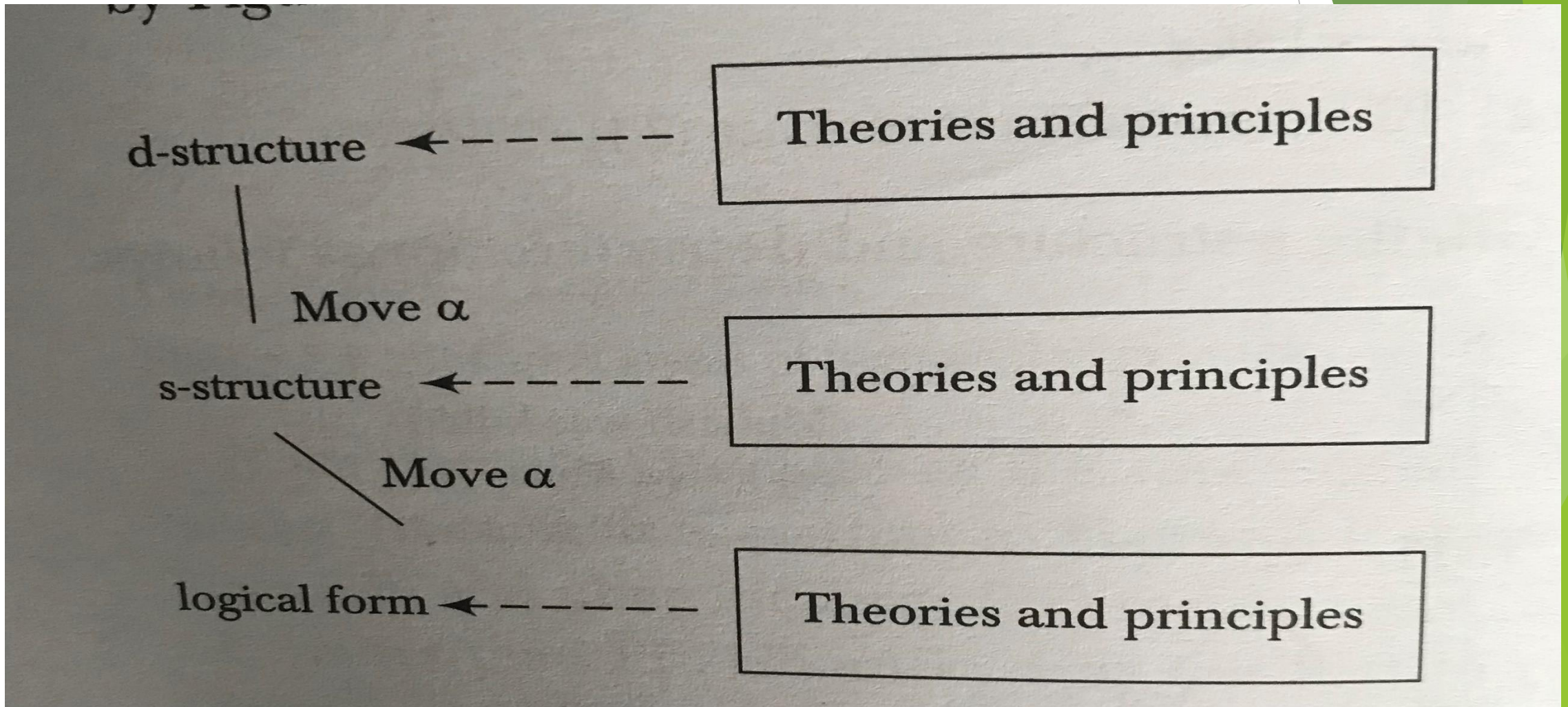
Transformational grammars consists of hundreds of rewriting rules and these rules are:

1. Language dependent
2. Construct dependent

Ex: Rules are different for active and passive sentences

- GB defines universal structural definition of Noun phrase(NP) and verb phrase (CVP), adjective phrase (AP),prepositional phrase(PP)etc.
- GB defines language independent grammar(Universal grammar)

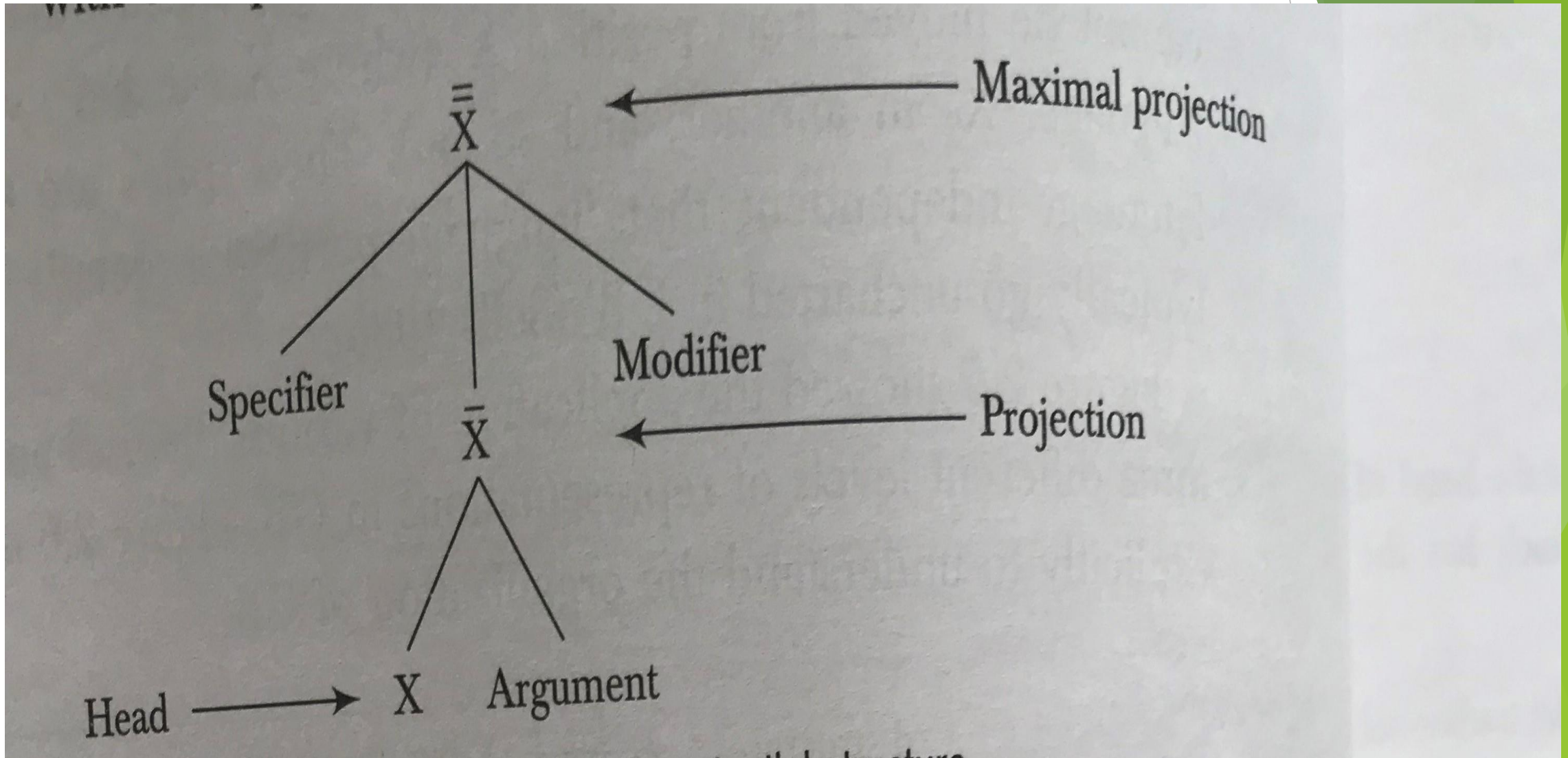
Organisation of GB (Pelter Shells, 1985)



- ▶ So, GB comprises a set of theories that map d-structure to s-structure and to logical form through a transformation rule (move α)
- ▶ move α moves a constituent “ α ” from one place to another place in a sentence if it does not violate the constraints put by GB theories or principles
- ▶ Components of GB:
 - ▶ 1. \bar{x} -theory
 - ▶ 2. Projection principle
 - ▶ 3. C-command
 - ▶ 4. Government
 - ▶ 5. Binding theory

- ▶ 1. \bar{x} -theory:
- ▶ It is one of the central concept in GB
- ▶ It defines phrase structure and sentence structure as “maximal projection of some head
- ▶ Noun Phrase(NP) is the maximal projection of noun (N).In NP, head is noun
- ▶ Verb phrase(CVP) is the maximal projection of verb(V) is the head in VP
- ▶ $\alpha = \{ V, N, A, P \}$
- ▶ Sentence structure is the maximal projection of inflection(INFL) i.e. is the modification of a word to explain different category
- ▶ Ex: Tense, Voice,gender etc

- General phrase and sentence structure as per X-theory is as follows:



1. NP: the food in a dhaba

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

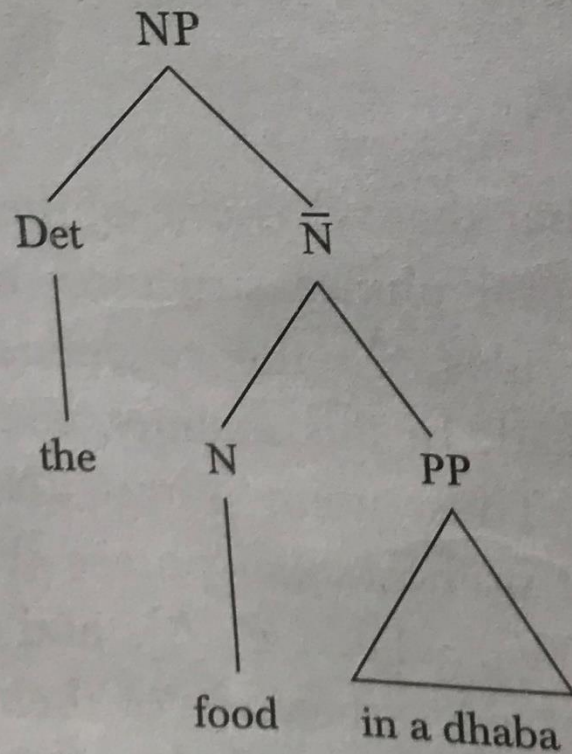


Figure 2.7(b) NP structure

2. VP: ate the food in a dhaba

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

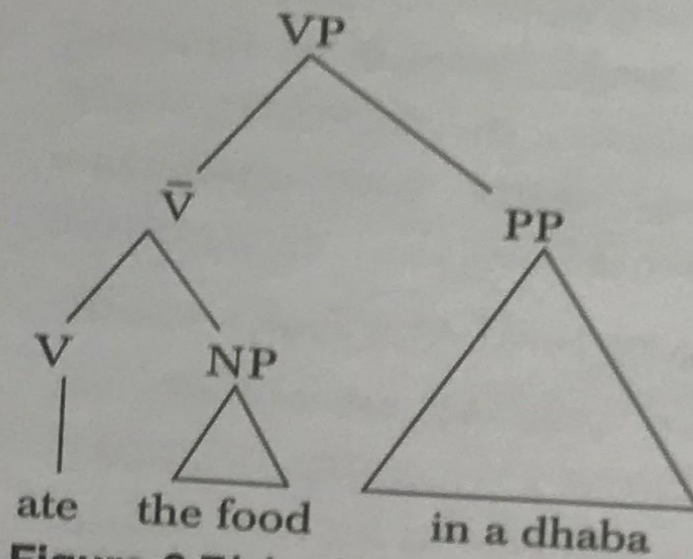


Figure 2.7(c) VP structure

3. AP: very proud of his country

$[_{AP} [_{Deg} \text{very}] [_{\bar{A}} [_{A} \text{proud}] [_{PP} \text{of his country}]]]$

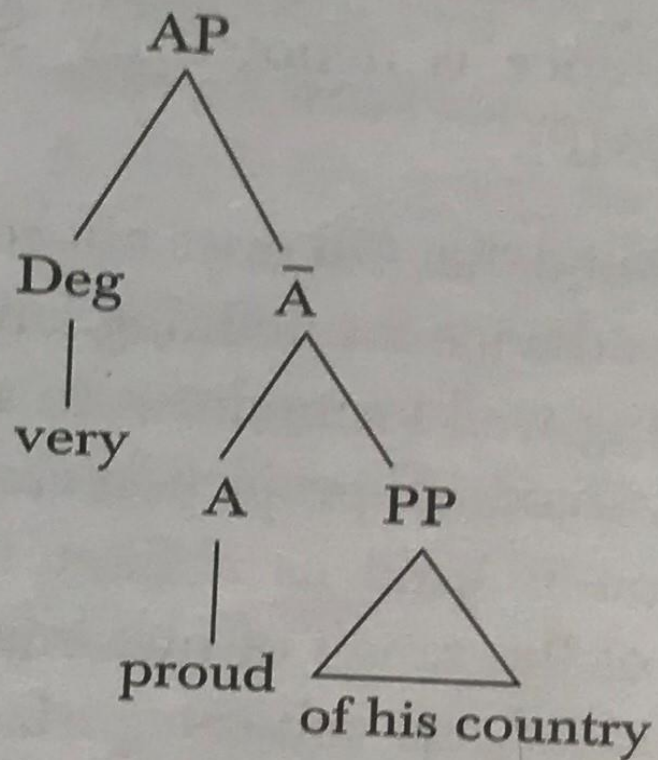


Figure 2.7(d) AP structure

4. PP: in a dhaba

$$[{}_{PP} [{}_{\bar{P}} [{}_P in] [{}_{NP} [{}_{Det} a] [{}_N dhabha]]]]$$

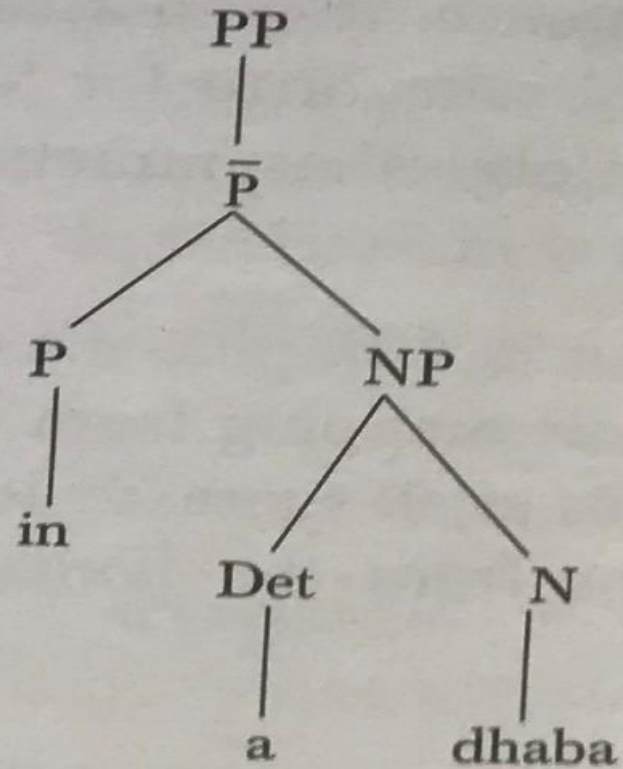


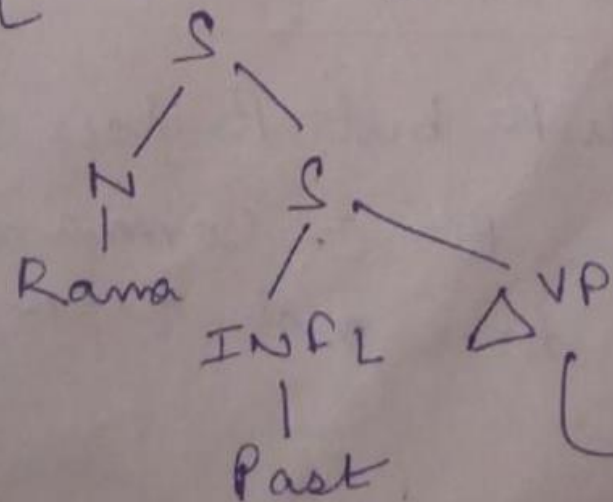
Figure 2.7(e) PP structure

Sentence Structure

Rama ate the food in a dhaka.

head = [past]
INFL

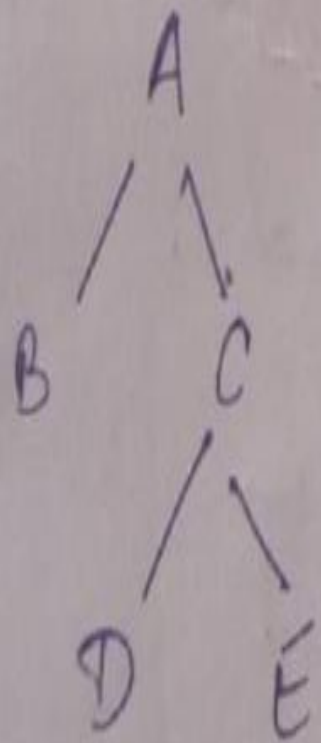
[Rama] [[past]
INFL] [ate the food in a dhaka]]



(Sentence Structure)

Government

- ▶ It refers to the relationship between a word and its dependence
- ▶ α governs β iff:
- ▶ α C-commands β
- ▶ α is the X(head eg: noun, verb, preposition, adjective, inflection)
- ▶ Government occurs between any 2 words connected by a dependency
- ▶ The dominate word (governor) is represented as parent node in the dependency tree
- ▶ The subordinate words(governees) are represented as child nodes



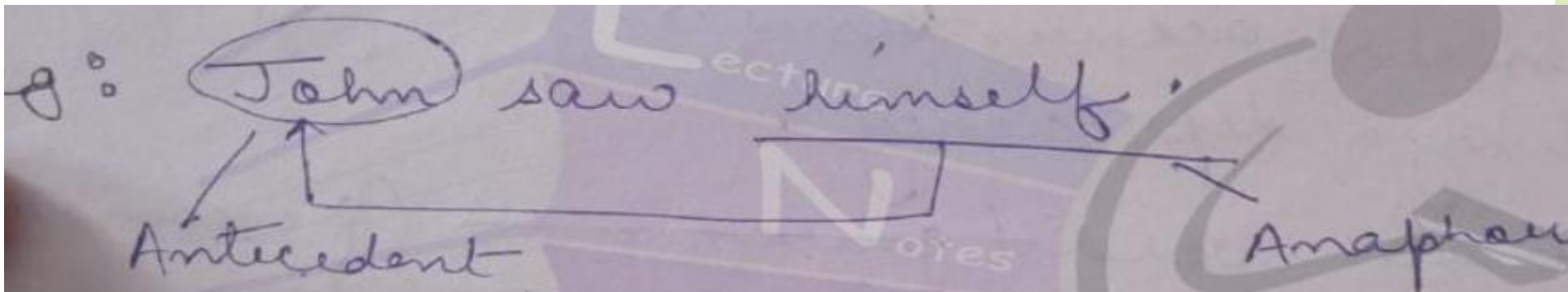
A governs B and C

(In other words,

A is the governor and
B and C are the governees.

Binding Theory

- ▶ It describes the relationship between NPs
- ▶ Noun Phrase may include:
- ▶ R- expressions(proper nouns, common nouns...)
- ▶ Pronouns (he ,she,it,him...)
- ▶ Anaphors (himself,herself,itself....)
- ▶ Antecedent: NP that gives its meaning to an anaphor or a noun



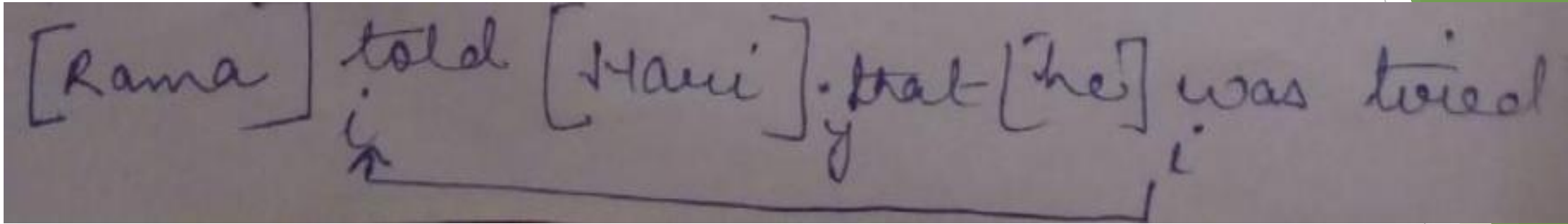
- John and Mary saw themselves on TV

Rama told Hari that he was tired.

↑ ↑ ↓
pronoun pronoun pronoun
may refer may refer may refer.

- Indexing:- The process that helps to find the proper antecedent

[Rama] told [Hani] that [he] was tired
i' y i'



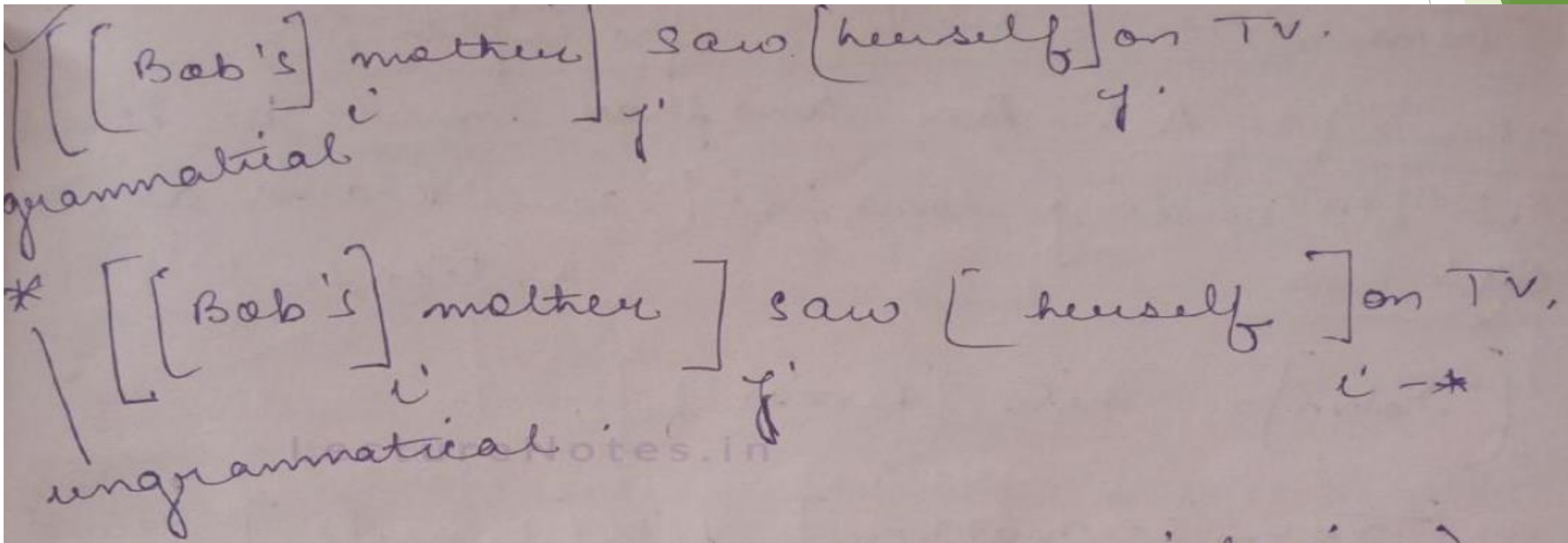
[[[Bob's] mother] saw [himself] on TV.
i' y'

grammatical

* [[[Bob's] mother] saw [himself] on TV.
i' y' i' - *

ungrammatical

Notes.in



► Binding is used, along with binding principles to determine whether a sentence is grammatically correct or not

► 3 binding principles:

1. Principle A: an anaphor must be bound in its binding domain

Ex: 1. John hits himself

2. John's mother hits himself

2. Principle B: A pronoun must be free (not bound) in the binding domain(governing domain)

Ex: John saw him (bound)

instead John saw Bob (not bound)

3. Principle C: R-expression must be free (not bound)

John saw John -(bound and ungrammatical)

John saw Bob (free)

► Lexical Functional Grammar

- It provides precise algorithms to address various linguistic issues of languages
- The lexical part contains various lexical rules and dependencies to define the structure of a sentence
- The functional part includes grammatical functions such as subjects, objects, roles played by various argument in a sentence etc

She saw stars.

She N (\uparrow Pred) = 'PRO'

(\uparrow Pers) = 3

(\uparrow Num) = SG

(\uparrow Gen) = FEM

(\uparrow Case) = NOM

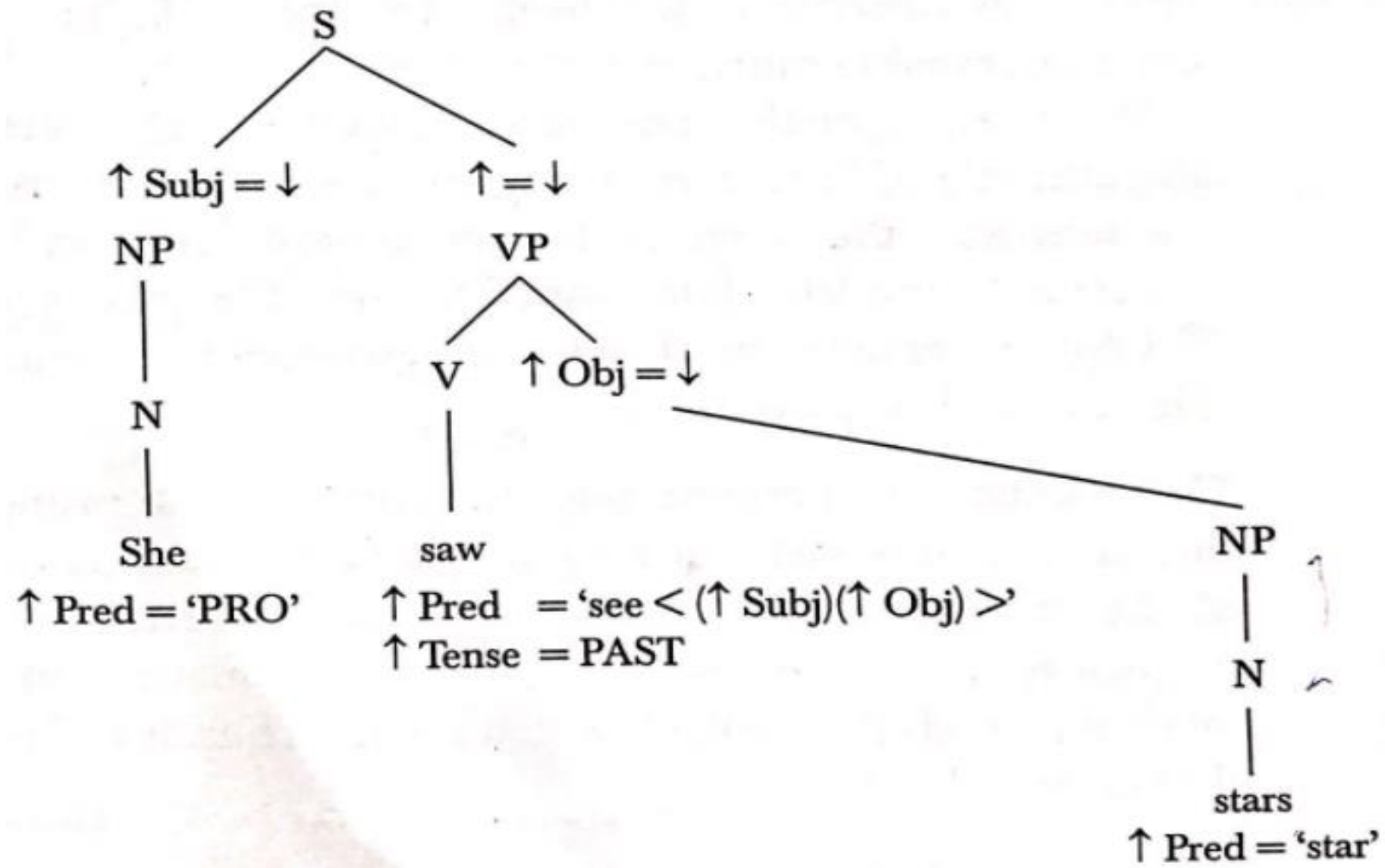
Saw V \uparrow Pred = 'see < (\uparrow Subj) (\uparrow Obj) >'

(\uparrow Tense = PAST)

Stars N \uparrow Pred = 'Star'

\uparrow Pers = 3

\uparrow Num = PL



Finally, the f-structure is the set of attribute–value pairs, represented a

subj	Pers	3
	Num	SG
	Gen	FEM
	Case	NOM
	Pred	'PRO'
obj	Pers	3
	Num	PL
	Pred	'Star'
Pred	'see' <(\uparrow subj) (\uparrow obj)>	

- ▶ Corpus: It is a large collection of text
- ▶ The plural form of corpus is corpora which are used in the development of NLP tools
- ▶ Statistical Language Model
 1. It is the probability distribution p over all possible word sequences
 2. It is a stochastic process model for word sequences(based on random variable and probability)
 3. $P(w)=p(w_1,w_2,\dots,w_n)$
 4. The role of statistical model is to estimate the probability of the likelihood
 5. A popular statistical language modelling is n-gram model

- ▶ N-gram model
- ▶ It uses previous $n-1$ words in a sentence to predict the next word(n th word)
- ▶ N-grams are the sequence of “ n ” tokens(word/term)
- ▶ N- stands for number of terms
- ▶ Ex : unigram:1 word
- ▶ bigram: 2 word
- ▶ trigram: 3 word
- ▶ 4-gram: 4 word
- ▶ n-gram: n word sequence
- ▶ The task of predicting the next word can be stated as
- ▶ $P(w_n | w_1, w_2, w_3, \dots, w_{n-1})$

- Bag of words with n-grams
- Unigram $n=1$

This is a sentence .
Unigram
($n=1$)

This is a sentence

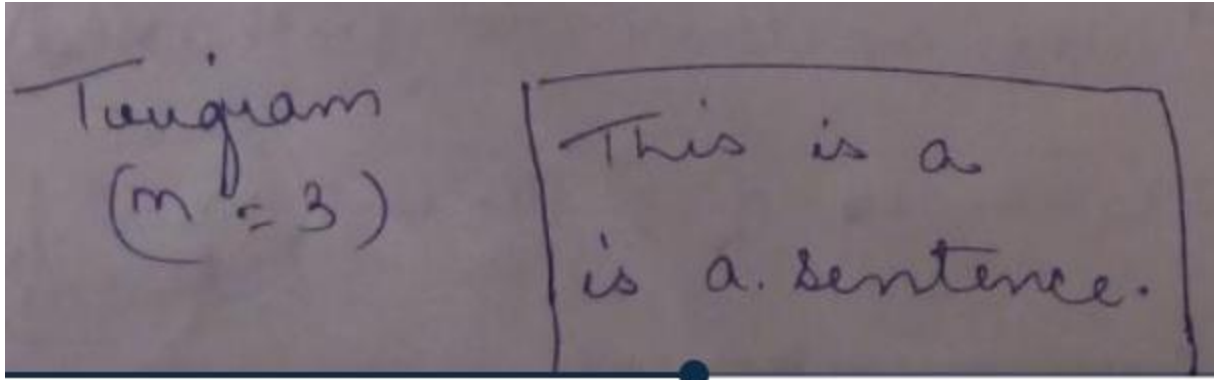
Bag - of - words.

- Bi-gram $n=2$

Bigram
($n=2$)

This is is a a sentence .

► Tri-gram n=3



- N-gram is achieved by decomposing a sentence probability into a product of conditional probability using the chain rule as below:
- $P(s) = P(w_1, w_2, w_3 \dots w_n)$
- Chain rule = $p(w_1) \cdot p(w_2|w_1) \cdot P(w_3|w_1 w_2) \dots p(w_n|w_1 w_2 \dots w_{n-1})$
- Hence

$$= \prod_{i=1}^n P(w_i | h_i)$$

where h_i is history of word w_i defined as
 $w_1, w_2 \dots w_{i-1}$

$$\begin{aligned}
 P(s) &= P(\omega_1, \omega_2, \omega_3, \dots, \omega_n) \\
 &= P(\omega_1) P(\omega_2 | \omega_1) P(\omega_3 | \omega_1, \omega_2) \dots P(\omega_n | \omega_1, \omega_2, \dots, \omega_{n-1}) \\
 &= \prod_{i=1}^n P(\omega_i | h_i)
 \end{aligned}$$

where h_i is history of word ω_i defined as
 $\omega_1, \omega_2, \dots, \omega_{i-1}$

eg: $P(\text{the red apple}) = P(\text{the}) \cdot P(\text{red} | \text{the}) \cdot P(\text{apple} | \text{the red})$

- ▶ Markov approximation of order $n-1$
- ▶ As per Markov, the probability of next word depends only on the previous $(n-1)$ word
- ▶ N-gram model is the Markov approximation model of order $n-1$
- ▶ How to estimate the probability?
- ▶ It is done by training n-gram model on training corpora or (corpus)
- ▶ Using maximum likelihood estimation, using relative frequency.

- n-gram model calculates $P(w_i | h_i)$ by modelling language as Markov Model of order $n-1$ i.e. looking at previous $n-1$ words only
- Similarly bi-gram model looks at previous one word only

$$P(s) = \prod_{i=1}^n P(w_i | w_{i-1})$$

and tri-gram model looks at previous two words only

$$P(s) = \prod_{i=1}^n P(w_i | w_{i-2} w_{i-1})$$

... at P/east | The

- Estimating Probability

The probability is estimated by training the n-gram model on training corpora and using the maximum likelihood estimate (MLE) technique

$$P(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

- ▶ Training set:
- ▶ <S>The Arabian knights
- ▶ <S>These are the fairy tales of the east
- ▶ <S>The stories of the Arabian knights are translated in many languages.
- ▶ Find the sentence probability of the following sentences using bi-gram model.
- ▶ Given <S> The Arabian knights are the fairy tales of the east.
- ▶ As per bigram model
- ▶ $P(S) = P(\text{the} | \text{<S>}) * P(\text{Arabian} | \text{the}) * P(\text{knight} | \text{Arabian}) * P(\text{are} | \text{knight}) * P(\text{the} | \text{are}) * P(\text{fairy} | \text{the}) * P(\text{tales} | \text{fairy}) * P(\text{of} | \text{tales}) * P(\text{the} | \text{of}) * P(\text{east} | \text{the})$
- ▶ $\frac{2}{3} * \frac{2}{5} * \frac{2}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{5} * \frac{1}{1} * \frac{1}{1} * \frac{2}{2} * \frac{1}{5}$
- = 0.00268

2) Consider the following corpus of three sentences

there is a big garden

children play in a garden

they play inside beautiful garden

Calculate $P(\text{they play in a big garden})$ assuming a
bi-gram language model.

Soluⁿ $P(\text{they play in a big garden})$
 $= P(\text{they} | \langle s \rangle) \times P(\text{play} | \text{they}) \times P(\text{in} | \text{play}) \times P(a | \text{in}) \times$
 $P(\text{big} | a) \times P(\text{garden} | \text{big}) \times \cancel{P(\text{big} | a)}$
 $= \frac{1}{3} \times \frac{1}{1} \times \frac{1}{2} \times \frac{1}{1} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{12}$

3) Find the probability of a sentence "I am not sam" using both bi-gram and tri-gram language model given the following training data:

I am Sam

Sam I am

I am not Sam

Using bi-gram model

$$P(\text{I am not Sam}) = P(\text{I} | \langle s \rangle) \times P(\text{am} | \text{I}) \times P(\text{not} | \text{am}) \times P(\text{Sam} | \text{not})$$
$$= \frac{2}{3} \times \frac{3}{3} \times \frac{1}{3} \times \frac{1}{1} = \frac{2}{9}$$

Using tri-gram model

$$P(\text{I am not Sam}) = P(\text{I} | \langle s_1 \rangle \langle s_2 \rangle) \times P(\text{am} | \langle s_2 \rangle \text{I}) \times$$
$$P(\text{not} | \text{I am}) \times P(\text{Sam} | \text{am not})$$
$$= \frac{2}{3} \times \frac{2}{2} \times \frac{1}{3} \times \frac{1}{1} = \frac{2}{9}$$

Note:

- ▶ 1. A special pseudo word $\langle s \rangle$ is introduced to mark the beginning of the sentence in bigram estimation
- ▶ Similarly in trigram estimation we introduce 2 pseudo – words $\langle s1 \rangle$ and $\langle s2 \rangle$
- ▶ As each probability is necessarily less than 1, multiplying the probabilities might cause a numerical underflow, particularly in long sentences
- ▶ To avoid this calculations are made in log space, where a calculation corresponds to adding log of individual probabilities and taking antilog of the sum
- ▶ In a uni-gram model the probability of each depends on its own probability in the corpus
- ▶ $P(W_n) = \frac{\text{number of times } W_n \text{ appears in the corpus}}{\text{total number of words in the corpus}}$

Handling Data Sparseness problem

- ▶ The n-gram model suffers from data sparseness problem
 - ▶ An n-gram that does not occur in the training data is assigned zero probability
 - ▶ A number of smoothing techniques have been developed to handle the data sparseness problem
 - ▶ Smoothing refers to the task of re-evaluating zero-probability or low-probability n-grams and assigning them non-zero values
1. Adding -one smoothing:
 2. It adds a value of 1 to each n-gram frequency before normalizing them into probabilities
 3.
$$\text{Padd-1}(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + 1}{C(w_{i-1}) + V}$$
 V- vocabulary size(unique words in corpus)

Good-Turing smoothing

- ▶ Good -Turing adjusts the frequency of an n-gram using the count of n-grams having a frequency $f+1$
- ▶ It converts the frequency of an n-gram from f to f^* using the following expression:

$$f^* = (f+1) \frac{n_{f+1}}{n_f}$$

where n_f : no of n-grams that occur exactly f times

Caching Technique

- ▶ The frequency of n-gram is not uniform across the corpus
- ▶ The basic n-gram model ignore this sort of variation of n-gram frequency
- ▶ The cache model combines the most recent n-gram frequency with the standard n-gram model to improve its performance locally
- ▶ It is based on the assumption that the recently discovered words are more likely to be repeated

Uses/Application of n-gram model

- ▶ Probability
- ▶ Communication theory
- ▶ Computational linguistics
- ▶ Data compression
- ▶ Information retrieval