Module 1 - handwritten

Natural Language Processing (Visvesvaraya Technological University)

MODULE-1

Chapter-1 - Introduction

Natural Language Processing (NLP) :-

Language is the primary means of communication used 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 (NLP) is concerned with the development of computational models of aspects of human language processing."

The two main reasons for development :-

1. Develop automated tools for language processing.
2. Gain better understanding of human communication.

Building computational models with human language-processing abilities requires a knowledge of how humans acquire, store and process language.

CHETAN. R
Asst. Professor

Two major approaches to NLP :-

1. Rationalist approach.
2. Empiricist approach.

Rationalist approach:- Which assumes the existence of some language faculty in human brain. Supporters of this approach argue that it is not possible for children to learn complex thing like natural language from limited sensory inputs.

Empiricists approach:-

They do not believe in the existence of some language faculty. They believe in the existence of some general organization principles such as pattern recognition, generalization and association. Learning of detailed structures can, therefore, take place through the application of these principles on sensory inputs available to the child.

Origins of NLP:-

Natural language processing sometimes termed as natural language understanding - originated from machine translation research.

Natural Language Understanding involves in only interpretation of language.

Natural language processing includes both understanding and generation.

Computational linguistics is similar to theoretical and psycho-linguistics.

Theoretical Linguistics:

They mainly provide structural description of natural language and its semantics. They are not concerned with the actual processing of sentences or generation of sentences from structural description.

They are in a quest for principles that remain common across languages and identify rules that capture linguistic generalization.

Example, most languages have constructs like noun and verb phrases.

Theoretical linguists identify rules that describe and restrict the structure of languages.

Psycholinguistics:

They explain how humans produce and comprehend natural language. They are interested in the representation of linguistic structures as well as in the process by which these structures are produced. They rely primarily on empirical investigations to back up their theories.

Computational Linguistics :-

It is concerned with the study of language using computational models of linguistic phenomena. It deals with the application of linguistic theories and computational techniques for NLP.

In computational linguistics, representing a language is a major problem; most knowledge representations tackle only a small part of knowledge.

Computational models may be broadly classified under knowledge driven and data-driven categories.

Knowledge driven systems rely on explicitly coded linguistic knowledge, often expressed as a set of handcrafted grammar rules. Acquiring and encoding such knowledge is difficult.

Data driven approaches presume the existence of a large amount of data and usually employ some machine learning technique to learn syntactic patterns. The amount of human effort is less and performance of these systems is dependent on the quantity of the data. These systems are usually adaptive to noisy data.

Language and Knowledge :-

Language is the medium of expression in which knowledge is deciphered. Language, being a medium of expression, is the outer form of the content it expresses. The same content can be expressed in different languages. The meaning of one-language is written in the same language.

CHETAN. R
Asst. Professor

(1) Lexical Analysis :-

It is the simplest level, which involves in Analysis of words. Words are the most fundamental unit of any natural language text. Word level processing requires morphological knowledge. i.e. knowledge about the structure and formation of words from basic units.
The rules for forming words from morphemes are language specific.
This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences and words.

② Syntactic Analysis :-

It consists of sequence of words as a unit, usually a sentence and finds its structure.

It decomposes a sentence into its constituents and identifies how they relate to each other. It captures grammaticality or non-grammaticality of sentences by looking at constraints like word order, number and case agreement.

This level of processing requires syntactic knowledge, i.e. knowledge about how words are combined to form larger units such as phrases and sentences.

For example, 'I went to the market' is a valid sentence whereas 'went the I market to' is not.

③ Semantic Analysis :-

Semantics is associated with the meaning of the language. It is concerned with creating meaningful representation of linguistic inputs.

The general idea of semantic interpretation is to take natural language sentences and map them onto some representation of meaning.

Defining meaning components is difficult as grammatically valid sentences can be meaningless.

The starting point of semantic analysis, has been lexical semantics. A word can have a number of possible meanings associated with it. But in a given context, only one of these meanings participates. Finding out the correct meaning of a particular use of word is necessary to find meaning of larger units.

CHETAN. R
Asst. Professor

Consider the following sentences:

Kabir and Ayan are married. —— (a)
Kabir and Suha are married. —— (b)

Both sentences have identical structures, and the meanings of individual words are clear. But most of us end up with two different interpretations.

We may interpret the second sentence to mean that Kabir and Saha are married to each other, but this interpretation does not occur for the first sentence. Syntatic structure and compositional semantics fail to explain these interpretations. We make use of pragmatic information.

(4) Discourse Analysis :-

Discourse-level processing attempts to interpret the structure and meaning of even larger units, e.g. at the paragraph and document level, in terms of words, phrases, clusters

and sentences.

It requires the resolution of anaphoric references and identification of discourse structure. It also requires discourse knowledge, that is, knowledge of how the meaning of a sentence is determined by preceding sentences. In fact, pragmatic knowledge may be needed for resolving anaphoric references.

For example, in the following sentences, resolving the anaphoric reference 'they' requires pragmatic knowledge:

"The district administration refused to give the trade union permission for the meeting because they feared violence." — (a)

"The district administration refused to give the trade union permission for the meeting because they oppose government." — (b)

(5) Pragmatic Analysis :-

This is the highest level of processing which deals with the purposeful use of sentences in situations.

It requires knowledge of the world, i.e. knowledge that extends beyond the contents of the text.

CHETAN. R
Asst. Professor

The Challenges of NLP :-

There are number of factors that make NLP difficult.

① There relate to the problems of representation and interpretation. Language computing requires precise representation of content. Since natural languages are highly ambiguous and vague, achieving such representation can be difficult.

② The inability to capture all the required knowledge is another source of difficulty. It is almost impossible to embody all sources of knowledge that humans use to procen language.

③ The greatest source of difficulty in natural language is identifying its semantics. words alone do not make a sentence. Instead, it is the words as well as their syntactic and semantic relation that give meaning to a sentence.

A language keeps on evolving. New words are added continually and existing words are introduced in new context. For example, most newspapers and TV channels use 9/11 to refer to the terrorist act on the World Trade Centre in the USA in 2004.

④ Idioms, metaphor and ellipses add more complexity to identify the meaning of the written text. As an example consider the sentence:

"The old man finally kicked the bucket".

The meaning of this sentence has nothing to do with the words 'kick' and 'bucket' appearing in it.

⑤ Quantifier-scoping is another problem. The scope of quantifiers (the, each, etc), is often not clear and poses problem in automatic processing.

⑥ The ambiguity of natural languages is another difficulty. Lexical Ambiguity:-

→ The first level of ambiguity arises at the word level. Without much effort, we can identify words that have multiple meanings associated with them.

Example:-

Manya is looking for a match.

In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

Solution:-

Part-of-speech tagging and word sense disambiguation.

② Syntactic Ambiguity :-

Syntactic ambiguity exists in the presence of two or more possible meanings within the sentence.

Example:

I saw the girl with the binocular.

In the above example, did I have the binoculars? or did the girl have the binoculars?

CHETAN. R
Asst. Professor

③ Referential Ambiguity :

Referential ambiguity exists when you are referring to something using the pronoun.

Example: Kiran went to Sunita. She said, "I am hungry".

In the above sentence, you do not know that who is hungry either Kiran or Sunita.

Language and Grammar

Automatic processing of language requires the rules and exceptions of a language to be explained to the computer. Grammar consists of a set of rules that allow us to parse and generate sentences in a language. Rules relate information to coding devices at the language level not at the world-knowledge level.

The world knowledge affects both the coding and the coding convention blurs the boundary between syntax and semantics.

Types of Grammars

Transformational Grammar (Chomsky 1957)

Lexical functional Grammar (Kaplan and Bresnan 1982)

Government and binding (Chomsky 1981)

Generalized Phare structure Grammar.

Dependency Grammar

Panirian Grammar

Tree-adjoining Grammar (Joshi 1985)

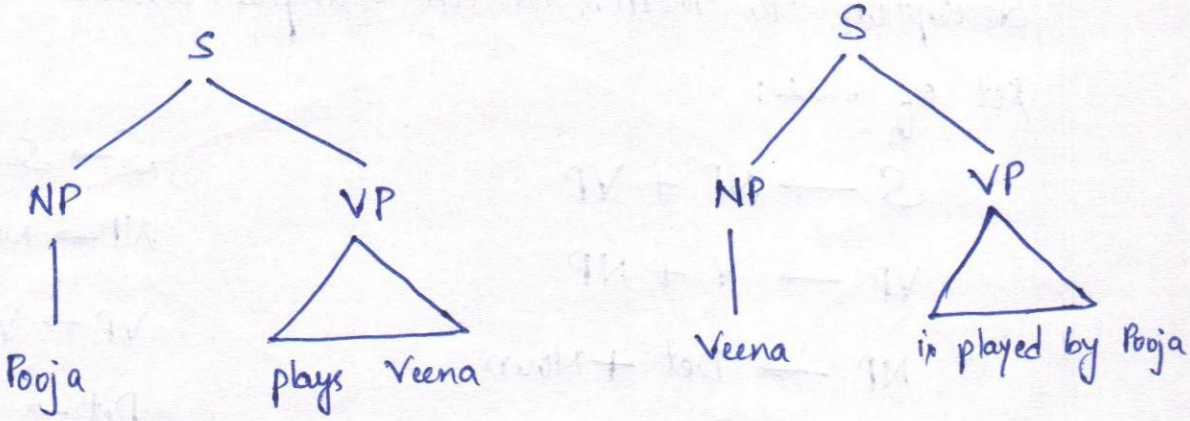
Some of there grammars focus on deeivation and other focus on relationships.

The greatest contribution to grammar comes from Noam Chomsky, who proposed a hierarchy of formal grammar based on level of complexity.

There grammars are phare structure rules.

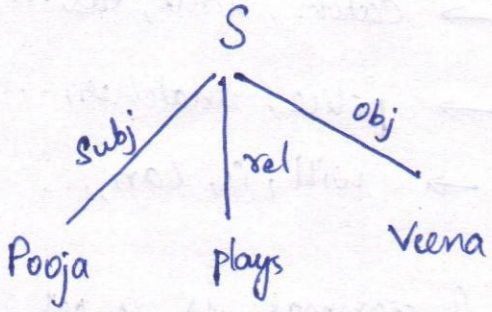
Generatic grammar basically refers to any grammar that uses a set of rules to specify or generate all and only grammatical sentences in a language.

Transformational grammar, each sentence in a language has two levels of representation, namely, a deep structure and surface structure.



Surface structure

CHETAN. R
Asst. Professor



Deep Structure

Transformational grammar has three components:

1. Phrase structure grammar.
2. Transformational rules
3. Morphophonemic rules - These rules match each sentences representation to a string of phonemes.

Each of there components consists of a set of rules. Phrase structure grammar consists of rules that generate natural language sentences and assign a structural description to them. As an example, consider the following set of rules:

$S \longrightarrow NP + VP$

$VP \longrightarrow V + NP$

$NP \longrightarrow Det + Noun$

$V \longrightarrow Aux + Verb$

$Det \longrightarrow$ the, a, an, ...

$Verb \longrightarrow$ catch, write, eat, ...

$Noun \longrightarrow$ police, snatcher, ...

$Aux \longrightarrow$ will, is, can, ...

S → Sentence

NP → Noun Phrase

VP → Verb Phrase

Det → Determiner.

Transformational grammar is a set of transformation rules, which transform one phrase-maker into another phrase-maker. There rules are applied on the terminal string generated by phrase structure rules. Unlike phrase structure rules, transformational rules are heterogeneous and may have more than one symbol on their left hand side. There rules are used to transform one surface representation into another. e.g. active sentence into passive one.

The rule relating active and passive sentences is:

$$NP_1 - Aux - V - NP_2 \rightarrow NP_2 - Aux + be + en - V - by + NP_1$$

Transformational rules can be obligatory or optional. An Obligatory transformation is one that ensures agreement in number of subject and verb.

An optional transformation is one that modifies the structure of a sentence while preserving its meaning.

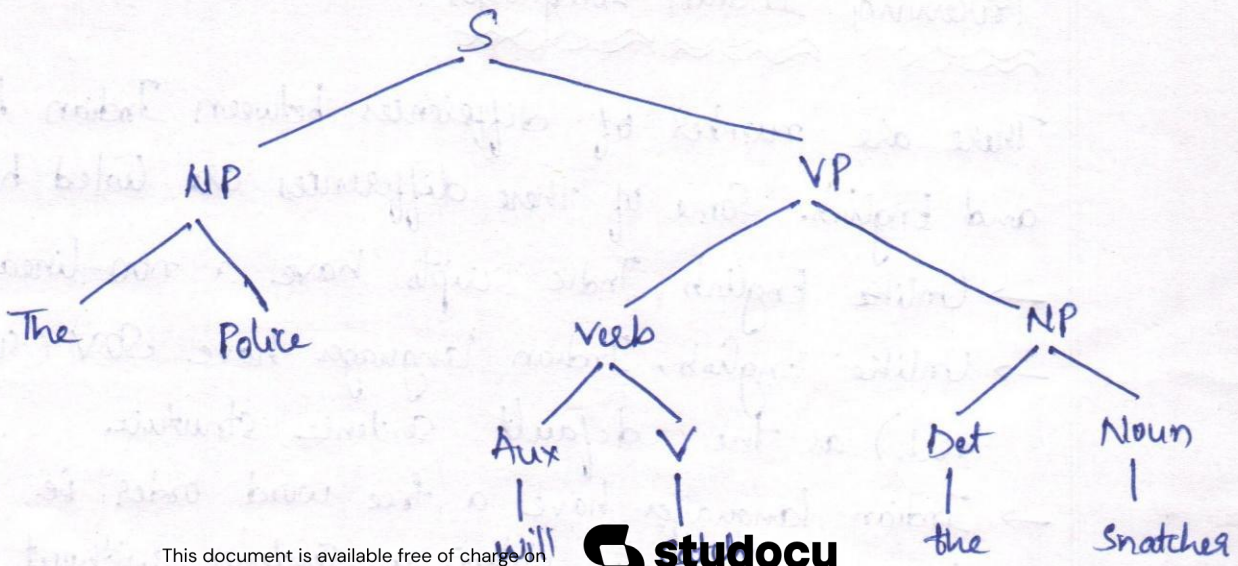
Morphophonemic rules match each sentence representation to a string of phonemes.

CHETAN. R
Asst. Professor

Consider the active sentence:

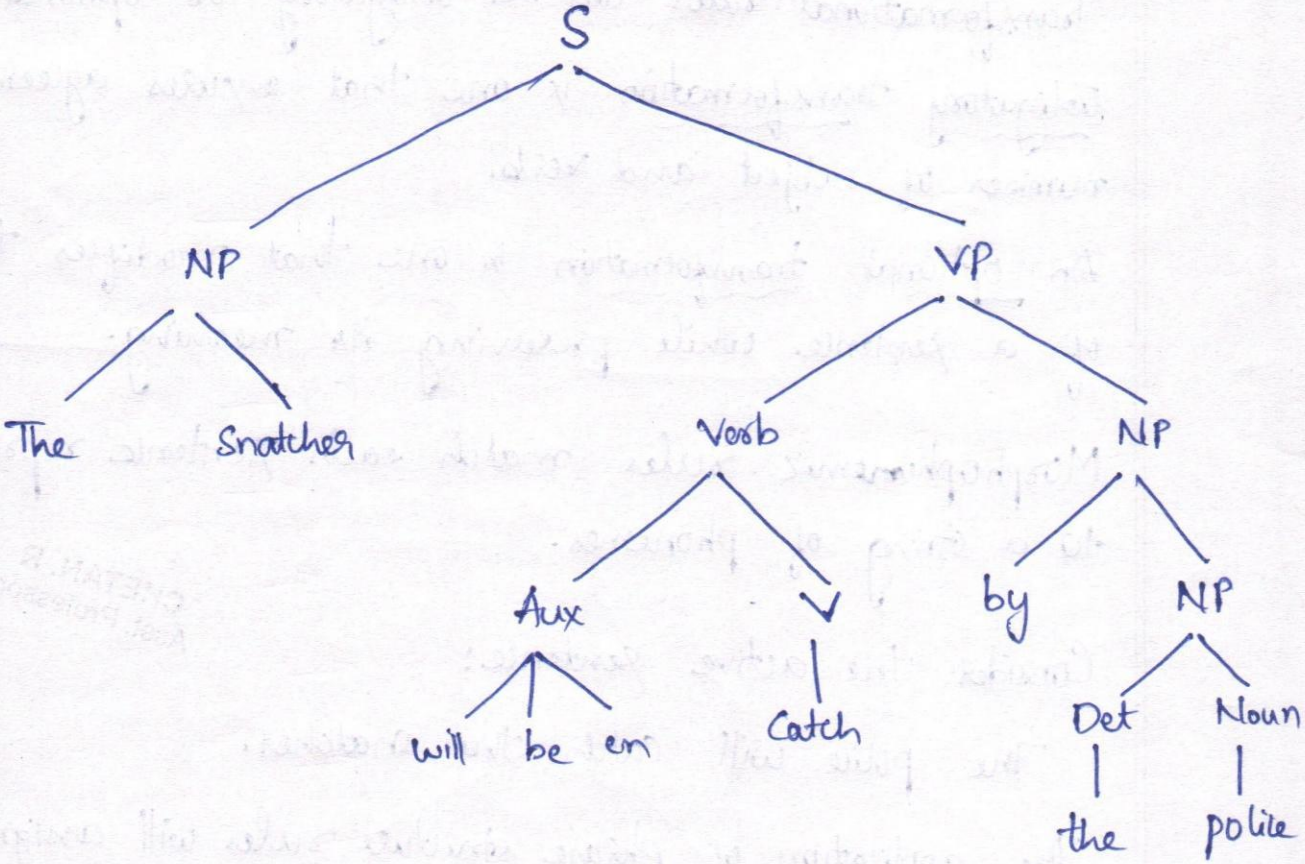
The police will catch the snatcher.

The application of phrase structure rules will assign the structure shown below:



The passive transformation rules will convert the sentence into:

The + culprit + will + be + en + catch + by + police.



Processing Indian Languages:-

There are number of differences between Indian Languages and English. Some of these differences are listed here:

→ Unlike English, Indic scripts have a non-linear structure.

→ Unlike English, Indian languages have SOV (Subject-Object-Verb) as the default sentence structure.

→ Indian languages have a free word order, i.e., words can be moved freely within a sentence without changing the meaning of the sentence.

→ Spelling standardization is more subtle in Hindi than in English.

→ Indian languages have a relatively rich set of morphological variants.

→ Indian languages make extensive and productive use of complex predicates (CPs).

→ Indian languages use post-position (Karakas) case markers instead of prepositions.

→ Indian languages use verb complexes consisting of sequences of verbs. e.g. गा रहा है and खेल रही है. The auxiliary verbs in this sequence provide information about tense, aspect, modality, etc.

NLP Applications :-

The applications utilizing the NLP include the following:

1. Machine Translation: This refers to automatic translation of text from one human language to another. In order to carry out these translation, it is necessary to have an understanding of words and phrases, grammars of the two languages involved, semantics of the language and world knowledge.

2. Speech Recognition :- This is the process of mapping acoustic speech signals to a set of words. The difficulties arise due to wide variations in the pronunciation of words, homonym and acoustic ambiguities.

3. Speech Synthesis :- It refers to automatic production of speech. Such systems can read out your mails on telephone, or even read out a story books.

4. Natural Language Interfaces to Databases :-
It allow quesying a structured database using natural language sentences.

5. Information Retrieval :- This is concerned with identifying documents relevant to a user's query. Indexing, word sense disambiguation, query modification, and knowledge bases have also been used in IR system to enhance performance.

6. Information Extraction :- It captures and outputs factual information contained within a document.

7. Question Answering :- It attempts to find the precise answer, or at least the precise portion of text in which the answer appears.

8. Text Summarization :- This deals with the creation of summaries of documents and involves syntactic, semantic and discourse level procesing of text.

Some Successful Early NLP Systems:-

1. ELIZA (Weizenbaum)

ELIZA is one of the earliest natural language understanding programs. It uses syntactic patterns to mimic human conversation with the user. Here is a sample conversation.

Eliza: Hello, I am Eliza. How may I help you?

User: I am feeling a little bit sleepy.

Eliza: How long have you been feeling a little bit sleepy?

User: For almost half an hour.

Eliza: Please go on.

CHETAN. R
Asst. Professor,

2. SysTran (System Translation)

The first SysTran machine translation system was developed in 1969 for Russian - English translation. SysTran also provided the first on-line machine translation service called Babel Fish, which is used by AltaVista search engines for handling translation requests from users.

3. TAUM METEO

This is a natural language generation system used in Canada to generate weather reports. It accepts daily weather data and generates weather reports in English and French.

4. SHRDLU (Winograd 1972)

This is a natural language understanding system that simulates actions of a robot in a block world domain. It uses syntactic parsing and semantic reasoning to understand instructions. The user can talk to robot to manipulate blocks, to tell the blocks configurations and to explain its reasoning.

5. LUNAR (Woods 1977)

This was an early question answering system that answered questions about moon rock.

INFORMATION RETRIEVAL :-

The availability of a large amount of text in electronic form has made it extremely difficult to get relevant information. Information retrieval system aims at providing a solution to this.

The term information is being used here to reflect 'subject matter' or the 'content' of some text. The focus is on the communication taking place between human beings as expressed through natural languages. Information is always associated with some data : we are concerned with text only.

The word 'retrieval' refers to operation of accessing information from some computer-based representation. Retrieval of information thus requires information to be processed and stored. Not all the information represented in computable form is retrieved. Instead, only the information relevant to the needs expressed in the form of query is located. Information retrieval is concerned with the organization, storage, retrieval and evaluation of information relevant to the query.

CHETAN. R
Asst. Professor

Information retrieval deals with unstructured data. The retrieval is performed based on the content of the document rather than on its structure. The IR systems usually return a ranked list of documents. The IR components have been traditionally incorporated into different types of information systems including database management systems, bibliographic text retrieval systems, question answering systems and more recently in search engines.

Current approaches for accessing large text collections can be broadly classified into two categories.

(1) Consists of approaches that construct topic hierarchy. e.g. Yahoo. This helps the user locate documents of interest manually by traversing the hierarchy.

(2) Consists of approaches that rank the retrieved documents according to relevance.

Issues in Information Retrieval:-

1. Choose a representation of the document.

Most human knowledge is coded in natural language which is difficult to use as knowledge representation language for computer systems.

Most of the current retrieval models are based on keyword representation. This representation creates problems during retrieval due to polysemy, homonymy and synonymy.

Polysemy: involves the phenomenon of a lexeme with multiple meaning.

Homonymy: is an ambiguity in which words that appear the same have unrelated meanings.

Synonymy : creates problem when a document is indexed with one term and the query contains a different term, and the two term share a common meaning.

Another problem with keyword-based retrieval is that it ignores semantic and contextual information in the retrieval process. This information is lost in the extraction of keywords from the text and cannot be recovered by the retrieval algorithms.

CHETAN. R
Asst. Professor

(2) Inappropriate characterization of queries by the user. The user may fail to include relevant terms in the query or may include irrelevant terms. Inappropriate or inaccurate queries lead to poor retrieval performance.

(3) Matching query representation with that of the document is another issue.

(4) Selection of the appropriate similarity measure is a crucial issue in the design of IR systems.

(5) Evaluating the performance of IR systems is also a major issue. Recall and precision are the most widely used measures of effectiveness.

⑥ The major goal of IR is to search a document in a manner relevant to the query, understanding what constitutes relevance is also an important issue.

⑦ The size of document collections and the varying needs of users also complicate text retrieval.

MODULE - 1

Chapter - 2 LANGUAGE MODELLING

A model is a description of some complex entity or process. A Language model is thus a description of language. Indeed, natural language is a complex entity and in order to process it through a computer-based program, we need to build a representation of it. This is known as Language Modeling.

Language modeling can be viewed either as a problem of grammar inference or a problem of probability estimation.

CHETAN. R
Asst. Professor

There are 2 approaches for Language Modeling:

1. Grammar-based Language model

It uses the grammar of a language to create its model. It attempts to represent the syntactic structure of a language. Grammar consists of hand-coded rules defining the structure and ordering of various constituents appearing in a linguistic unit.
For example, a sentence usually consists of noun phrase and a verb phrase. The grammar based approach attempts to utilize this structure and also the relationships between these structures.

(2) Statistical Language Modelling

It creates a language model by training it from a corpus. In order to capture regularities of a language, the training corpus needs to be sufficiently large.

"Statistical language modelling is the attempt to capture regularities of natural language for the purpose of improving the performance of various natural language applications."

Statistical language modelling is one of the fundamental tasks in many NLP applications, including speech recognition, spelling correction, handwriting recognition and machine translation.

Various Grammar-Based Language Models

(1.) Generative Grammar

Noam Chomsky in 1957 in his book Syntactic Structures wrote that we can generate sentences in a language if we know a collection of words and rules in that language. Only those sentences that can be generated as per the rules are grammatical. This point of view has dominated computational linguistics and is called generative grammar.

Language is a relation between the sound and its meaning. Thus any model of a language should also deal with the meaning of its sentences.

(2) Hierarchical Grammar

Chomsky described classes of grammars in a hierarchical manner, where the top layer contained the grammars represented by its sub classes.

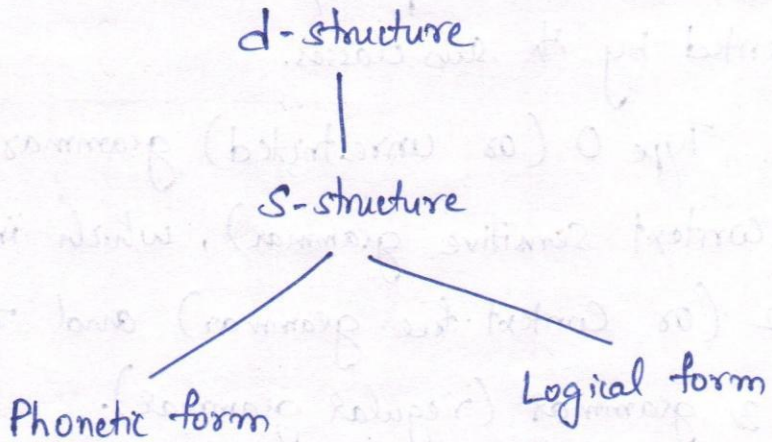
Hence, Type 0 (or unrestricted) grammar contains Type 1 (or context sensitive grammar), which in turn contains Type 2 (or Context-free grammar) and that again contains Type 3 grammar (regular grammar).

(3) Government and Binding

A common viewpoint taken by linguists is that the structure of a language can be understood at the level of its meaning, particularly while resolving structural ambiguity. However, the sentences are given at the syntactical level and the transformation from meaning to syntax or vice versa is not well understood.

Transformational grammars assume two levels of existence of sentences one at the surface level and the other at the deep root level.

Government and Binding theories have renamed them as S-level and d-level and identified two more levels of representation called phonetic form and logical form.



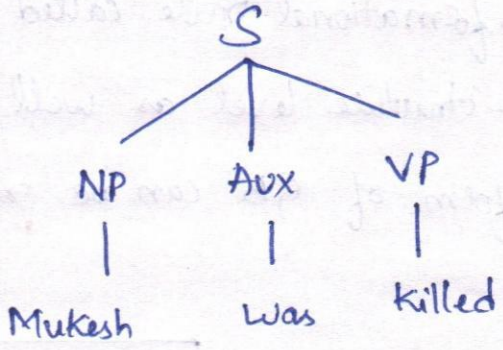
Transformational grammars have hundreds of rewriting rules, which are generally language-specific and also construct-specific. Generation of a complete set of coherent rules may not be possible.

In GB if we define rules for structural units at the deep level, it will be possible to generate any language with fewer rules. There deep-level structures are abstractions of noun-phrase, verb-phrase, etc. common to all languages.

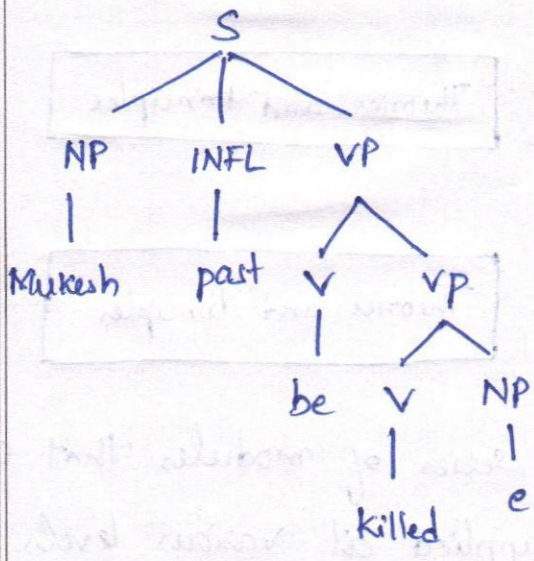
Let us take an example to explain d- and s-structures.

"Mukesh was killed"

(i) In transformational grammar, this can be represented as S-NP Aux VP as given below:



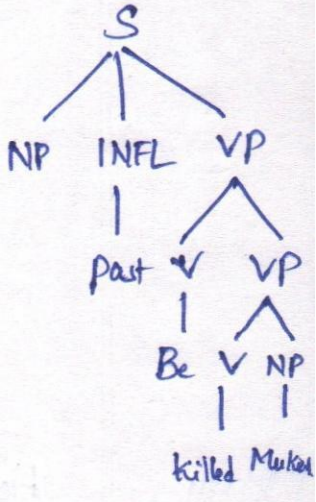
(ii) In GB, the s-structure and d-structure are as follows:



(a) Surface Structure

Mukesh was killed
(e) killed Mukesh
e) past kill Mukesh

INFL: Inflection
NP: Noun phrase
VP: Verb phrase
e: empty.



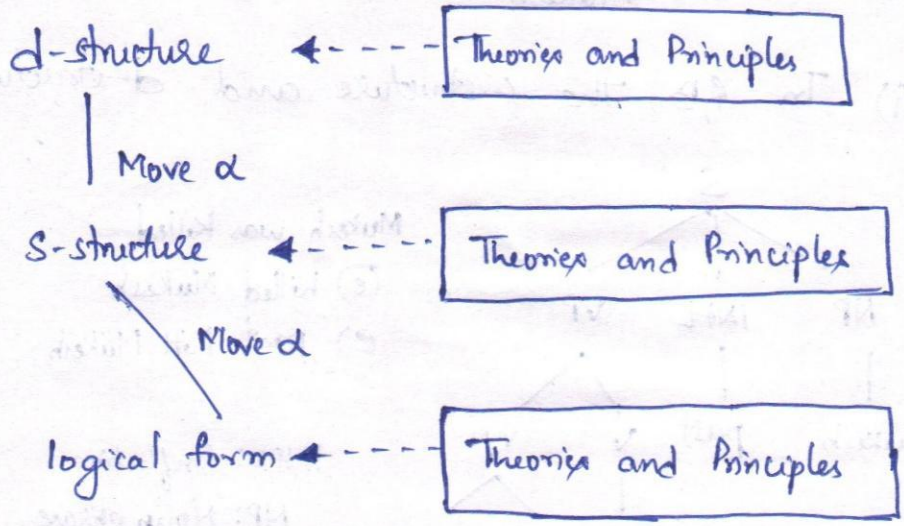
(b) Deep Structure.

Components of GB

Government and binding (GB) comprises a set of theories that map the structures from d-structure to s-structure and to logical form (LF).

A general transformational rule called 'Move α' is applied at d-structure level as well as at s-structure level.

The simplest form of GB can be represented by below figure:



Hence GB consists of 'a series of modules that contain constraints and principles applied at various levels of the representations and transformational rule, Move α.

The GB considers all three levels of representations (d-, s-, and LF) as syntactic and LF is also related to meaning or semantic-interpretive mechanisms.

The GB applies the same Move α transformation to map d-levels to s-levels or s-levels to LF Level. LF level helps in quantifier scoping and also in handling various sentence constructions such as passive or interrogative constructions.

Example:-

"Two countries are visited by most travellers".

Its two possible logical forms are:

LF1: $[_S$ Two countries are visited by $[_{NP}$ most travellers$]]$

LF2: Applying Move α

$[_{NP}$ Most travellers$_i$ $][_S$ two countries are visited by $e_i]$

In LF1, the interpretation is that most travellers visit the same two countries.

In LF2, when we move [most travellers] outside the scope of the sentence, the interpretation can be that most travellers visit two countries, which may be different for different travellers.

One of the important concepts in GB is that of constraints. It is the part of the grammar which prohibits certain combinations and movements; Otherwise Move α can move anything to any possible position. Thus GB. is basically the formulation of theories or principles which create constraints to disallow the construction of ill-formed sentences.

The organization of GB is given below:

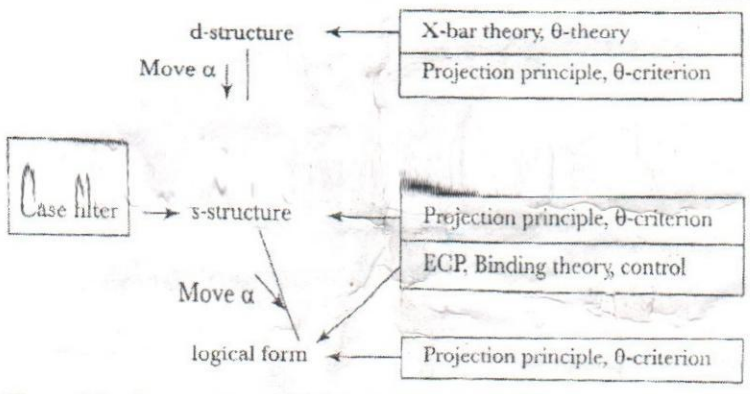


Figure 2.6 Organization of GB (adapted from Peter Sells 1985)

$\bar{X}$ Theory :-

The $\bar{X}$ theory is one of the central concepts in GB. Instead of defining several phrase structures and the sentence structure with separate sets of rules. $\bar{X}$ theory defines them both as maximal projections of some head. Noun Phrase (NP), Verb Phrase (VP), adjective phrase (AP) and prepositional phrase (PP) are maximal projections

of noun (N), verb (V), adjective (A) and preposition (P) respectively and can be represented as head X of their corresponding phrases (where $X = \{N, V, A, P\}$).

Even the sentence structure can be regarded as the maximal projection of inflection (INFL).

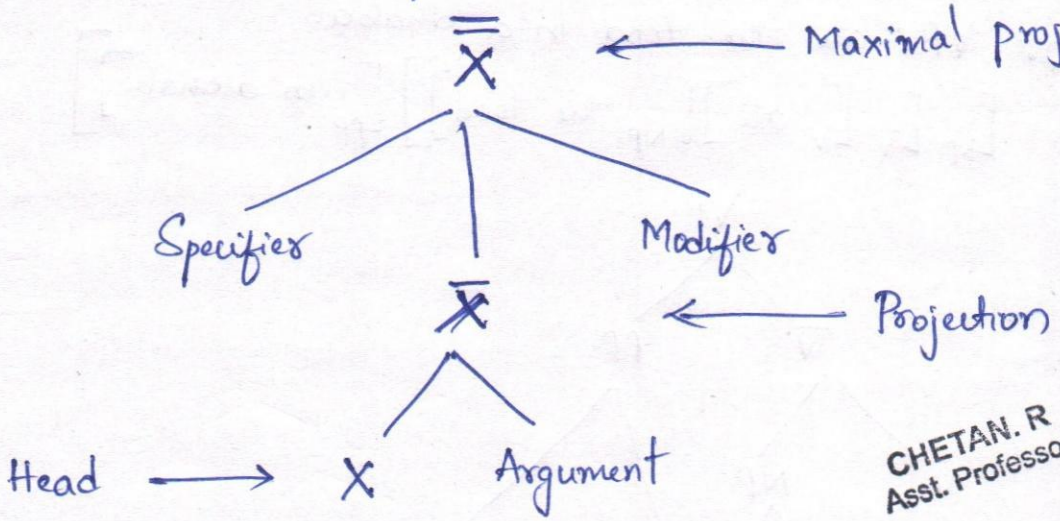
The GB envisages projections at two levels

→ projection of head at semi-phrasal level denoted by $\overline{X}$

→ maximal projection at the phrasal level denoted by $\overline{\overline{X}}$.

The first level projection is denoted as S and second level projection is denoted by S'.

The below figure depicts the general and particular structures with examples.

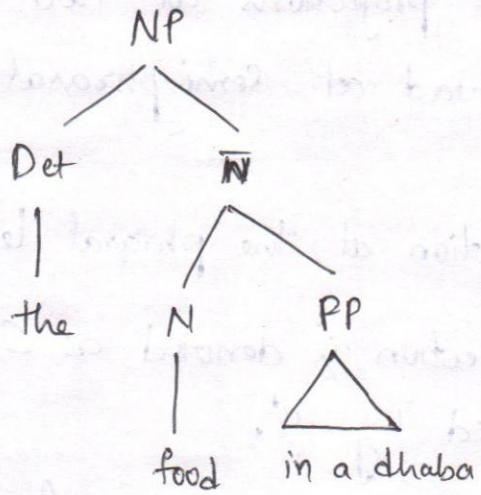


CHETAN. R
Asst. Professor

Next, we consider the representation of the NP, the food in a dhaba. This is followed by the representation of VP, AP and PP structure.

1. NP: the food in a dhaba

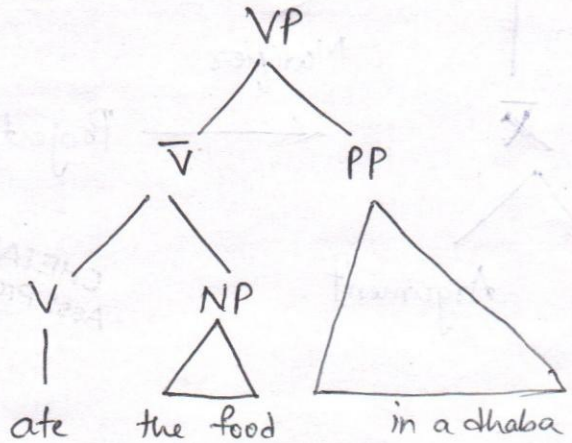
$[_{NP}$ the $[_{N}$ food $]_{pp}$ $[$in a dhabha$]$



NP structure

2. VP: ate the food in a dhaba.

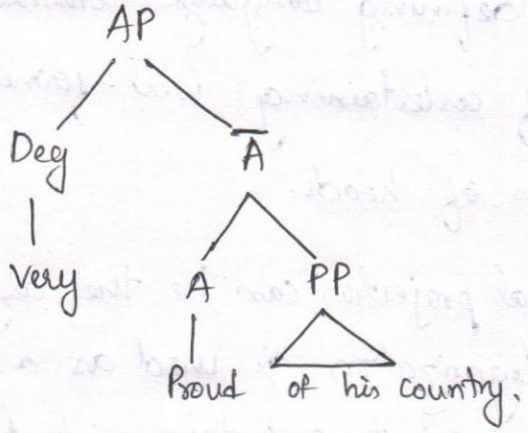
$[_{VP}$ $[_{\bar{V}}$ $[_{V}$ ate$][_{NP}$ the food$]]$ $[_{PP}$ in a dhaba$]]$



VP structure

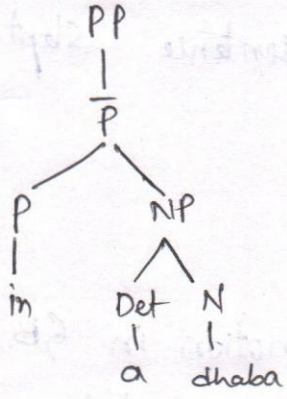
3. AP: very proud of his country.

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



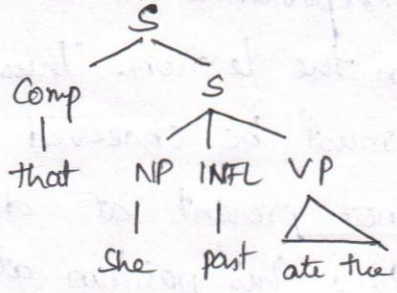
4. PP: in a dhaba

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



5. S: that she ate the food in a dhaba.

$[_{\bar{S}}\ [_{COMP}\ that][_{S}\ [_{Det}\ She][_{INFL}\ past][_{VP}\ ate\ the\ food\ in\ a\ dhaba]]]$



Sub-categorization :-

GB does not consider traditional phrase structure as an appropriate device for defining language constructs. It places the burden of ascertaining well-formedness to sub-categorization frames of heads.

In principle any maximal projection can be the argument of a head, but subcategorization is used as a filter to permit various heads to select a certain subset of the range of maximal projections.

Example:- (i) the verb "eat" can sub-categorize for NP, whereas the verb 'sleep' cannot.

(ii) "ate food" is well-formed but sentence "slept the bed" is not.

Projection Principle :-

The projection principle is a basic notion in GB, places a constraint on the three syntactic representations and their mapping from one to the other.

The principle states that representations at all syntactic levels are projections from the lexicon. Thus, lexical properties of categorical structure must be observed at each level.

Suppose "the object" is not present at d-level, then another NP cannot take this position at S-level.

Theta-Theory (θ-Theory) or Theory of Thematic Relations

Sub-Categorization only places a restriction on syntactic categories which a head can accept.
GB puts another restriction on the lexical heads through which it assigns certain roles to its arguments. These roles are pre-assigned and cannot be violated at any syntactical level as per the projection principle. These role assignments are called theta-roles and are related to 'Semantic-Selection'.

Theta-role and Theta-criterion :-

There are certain thematic roles from which a head can select. These are called θ-roles and they are mentioned in the lexicon. Example, the verb 'eat' can take arguments with θ-roles "('Agent', 'Theme')".

Agent is a special type of role which can be assigned by a head to outside arguments whereas other roles are assigned within its domain.
Hence in "Mukesh ate food", the verb 'eat' assigns the 'Agent' role to 'Mukesh' and 'Theme' role to 'food'.
Theta-Criterion states that 'each argument bears one and only one θ-role and each θ-role is assigned to one and only one argument'.

C-Command and Government :-

C-command - defines the scope of maximal projection. It is a basic mechanism through which many constraints are defined on Move α.

If any word or phrase (say α or β) falls within the scope of and is determined by a maximal projection, we say that it is dominated by the maximal projection.

If there are two structures α and β related in such a way that 'every maximal projection dominating α dominates β' we say that α C-commands β, and this is the necessary and sufficient condition (iff) for C-commands.

Government

α governs β iff:

α C-commands β

α is an X and every maximal projection dominating β dominates α.

Movement, Empty Category, and Co-indexing :-

In GB, Move α is described as 'move anything anywhere' though it provides restrictions for valid movements.

In GB, the active to passive transformation is the result of NP movement as shown in sentence below: Another well-known movement is the wh-movement, where wh-phrase is moved as follows.

What did Mukesh eat?
[Mukesh INFL eat what]

In the projection principle, lexical categories must exist at all the three levels. This principle, when applied to some cases of movement leads to the existence of an abstract entity called empty category.

In GB, there are 4 types of empty categories, two being empty NP positions called wh-trace and NP trace and the remaining two being pronouns called small 'pro' and big 'PRO'. This division is based on two properties - anaphoric (+a or -a) and pronominal (+p or -p).

Wh-trace	-a, -p
NP-trace	+a, -p
Small 'pro'	-a, +p
big 'PRO'	+a, +p

CHETAN. R
Asst. Professor

Co-indexing is the indexing of the subject NP and AGR at d-structure which are preserved by Move α operations at s-structure.

When an NP-movement takes place, a trace of the movement is created by having an indexed empty category (e_i) from the position at which the movement began to the corresponding indexed NP.

For defining constraints to movement, the theory identifies two positions in a sentence. Positions assigned θ-roles are called θ-positions, while others are called $\bar{\theta}$-positions.

Core grammatical positions are called A-positions and the rest are called $\bar{A}$-positions.

Binding Theory

Binding is defined by Sells (1985) as follows:

α binds β iff

α C-commands β, and

α and β are co-indexed

[e_i INFL Kill Mukesh]

[Mukesh$_i$ was killed (by e_i)]

Mukesh was killed.

Empty clause (e_i) and Mukesh (NP_i) are bound. This theory gives a relationship between NPs.

Binding theory can be given as follows:

(a) An anaphor (+a) is bound in its governing category.

(b) A pronominal (+p) is free in its governing category.

(c) An R-expression (-a, -p) is free.

Example:-

A: Mukesh$_i$ knows himself$_i$

B: Mukesh$_i$ believes that Amrita knows him$_i$

C: Mukesh believes that Amrita$_j$ knows Nupur$_k$

Similar rules apply on empty categories also:

NP-trace: +a, -p: Mukesh$_i$ was killed e_i

wh-trace: -a, -p: who$_i$ does he$_i$ like e_i

Empty Category Principle (ECP)

The proper government is defined as:

α properly governs β iff:

α governs β and α is lexical (i.e. N, V, A or P) or

α locally A-binds β

The ECP says 'A trace must be properly governed.'

This principle justifies the creation of empty categories during NP-trace and Wh-trace and also explains the subject/object asymmetries to some extent. As in the following sentences:

(a) What$_i$ do you think that Mukesh ate e_i?

(b) What$_i$ do you think Mukesh ate e_i?

Bounding and Control Theory:

In English, the long distance movement for complement clause can be explained by bounding theory if NP and S are taken to be bounding nodes. The theory says that the application of Move α may not cross more than one bounding node. The theory of control involves syntax, semantics and pragmatics.

Case Theory and Case Filter :-

In GB, Case theory deals with the distribution of NPs and mentions that each NP must be assigned a case. In English, we have the nominative, objective, genitive etc., cases which are assigned to NPs at particular positions. Indian languages are rich in case-markers, which are carried even during movements.

Case Filter :-

An NP is ungrammatical if it has phonetic content or if it is an argument and is not case-marked.

Phonetic content here, refers to some physical realization, as opposed to empty categories. Thus, case filters restrict the movement of NP at a position which has no case assignment. It works in a manner similar to that of the θ-criterion.

GB presents a model of the language which has three levels of Syntactic representations.

→ It assumes phrase structures to be the maximal projection of some lexical head and in a similar fashion, explains the structure of a sentence or a clause.

→ It assigns various types of roles to these structures and allows them a broad kind of movement called Move α.

→ It then defines various types of constraints which restrict certain movements and justifies others.

CHETAN. R
Asst. Professor

(4) Lexical Functional Grammar (LFG) Model

LFG represents sentences at two syntactic levels - Constituent structure (c-structure) and functional structure (f-structure).

Kaplan proposed a concrete form for the register names and values which became the functional structures in LFG. On the other hand, Bresnan was more concerned with the problem of explaining some linguistic issues such as active/passive and dative alternations, in transformational approach. She proposed that such issues can be dealt with by using lexical redundancy rules.

The term 'Lexical Functional' is composed of two terms: the 'functional' part is derived from 'grammatical functions', such as subject and object, or roles played by various arguments in a sentence.
The 'lexical' part is derived from the fact that the lexical rules can be formulated to help define the given structure of a sentence and some of the long distance dependencies, which is difficult in transformational grammars.

C-structure and f-structure in LFG

The C-structure is derived from the usual phrase and sentence structure syntax as in CFG. The grammatical functional role cannot be derived directly from phrase and sentence structure. functional specifications are annotated on the nodes of C-structure, which when applied on sentences, results in f-structure.

Hence f-structure is the final product which encodes the information obtained from phrase and sentence structure rules and functional specifications.

Example :-

"She saw stars in the sky".

CFG rules to handle this sentence are:

$S \longrightarrow NP\ VP$

$VP \longrightarrow V\ \{NP\}\ \{NP\}\ PP^*\ \{S'\}$

$PP \longrightarrow P\ NP$

$NP \longrightarrow Det\ N\ \{PP\}$

$S' \longrightarrow Comp\ S$

where
S: Sentence
P: preposition
S': clause
{}: optional
*: Phrase can appear any number of times including blank
V: Verb
N: noun
Comp: Complement

When annotated with functional specifications, the rules become:

Rule 1: $S \longrightarrow$ NP VP
$\uparrow subj = \downarrow$ $\uparrow = \downarrow$

Rule 2: $VP \rightarrow$ V {NP} {NP} PP* {S'}
$\uparrow obj = \downarrow$ $\uparrow obj2 = \downarrow$ $\uparrow (\downarrow case) = \downarrow$ $\uparrow comp = \downarrow$

Rule 3: $PP \longrightarrow$ P NP
$\uparrow obj = \downarrow$

Rule 4: $NP \longrightarrow$ {Det} N {PP}
$\uparrow Adjunct = \downarrow$

Rule 5: $S' \longrightarrow$ Comp S
$\uparrow = \downarrow$

Here $\uparrow$ refers to the f-structure of the mother node that is on the left hand side of the rule.
The $\downarrow$ symbol refers to the f-structure of the node under which it is denoted.

Hence, in Rule 1, $(\uparrow subj = \downarrow)$ indicates that the f-structure of the first NP goes to the f-structure of the subject of the sentence, while $(\uparrow = \downarrow)$ indicates that the f-structure of the VP node goes directly to the f-structure of the Sentence. VP.

Consistency:- In a given f-structure, a particular attribute may have at the most one value. Hence, while unifying two f-structures, if the attribute Num has value SG in one and PL in the other, it will be rejected.

Completeness: When an f-structure and all its subsidiary f-structures contain all the functions that their predicates govern, then and only then is the f-structure complete.

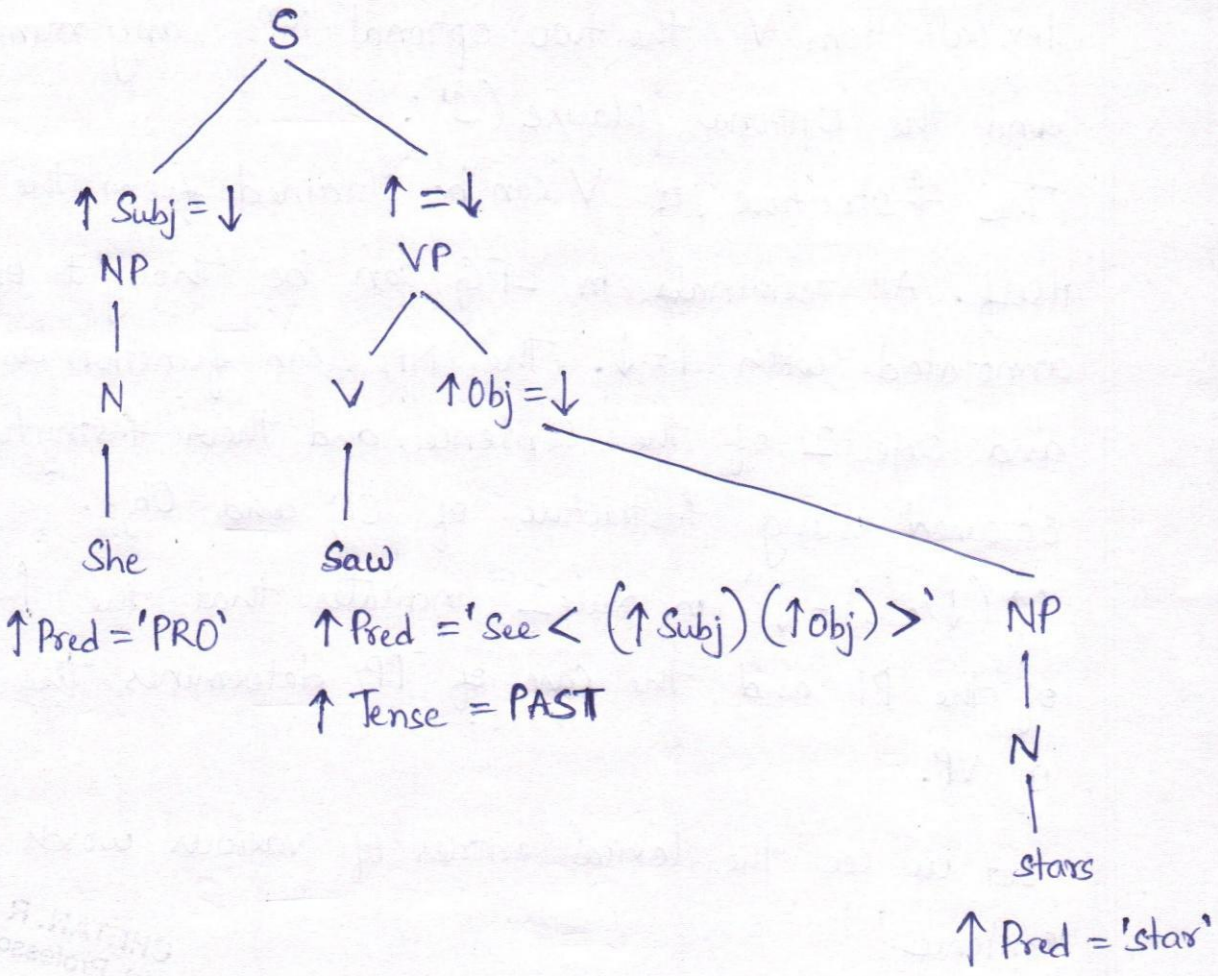
Coherence: Coherence maps the completeness property in the reverse direction. It requires that all governable functions of an f-structure and all its subsidiary f-structures must be governed by their respective predicates.
Hence in f-structure of a sentence, an object cannot be taken if its verb does not allow that object.

Lexical Rules in LFG

Different theories have different kinds of lexical rules and constraints for handling various sentence-constructs.

→ In GB, to express a sentence in its passive form, the verb is changed to its participial form and the ability of the verb to assign case and external θ-role is taken away.

C-structure of sentence is given below:



The f-structure is the set of attribute-value pairs, represented as

$$\begin{bmatrix} \text{Subj} & \begin{bmatrix} \text{Pers} & 3 \\ \text{Num} & \text{SG} \\ \text{Gen} & \text{FEM} \\ \text{Case} & \text{NOM} \\ \text{Pred} & \text{'PRO'} \end{bmatrix} \\ \text{Obj} & \begin{bmatrix} \text{Pers} & 3 \\ \text{Num} & \text{PL} \\ \text{Pred} & \text{'Star'} \end{bmatrix} \\ \text{Pred} & \text{'See'} \langle \uparrow \text{subj}) \ (\uparrow \text{obj}) \rangle \end{bmatrix}$$

Example :-

Active : तारा हँसी
Taaraa hansii
Taara laughed

Causative : मोनिका ने तारा को हँसाया
Monika ne Taara ko hansaayaa
Monika Subj Taara Obj laugh-cause-past
Monika made Taara to laugh.

Active : ↑ Pred = 'Laugh <↑ Subj>'

Causative : ↑ Pred = 'cause <(↑ Subj) (↑ Obj) (Comp)>'

Long Distance Dependencies and Coordination

In GB, when a category moved, it creates an empty category.

In LFG, Unbounded movement and coordination is handled by the functional identity and by correlation with the corresponding f-structure.

Example : Consider the wh-movement in the following sentence.

Which picture does Taara like-most?

The f-structure can be represented as

$$\begin{bmatrix} \text{Focus} & \begin{bmatrix} \text{Obl}_{th} & \begin{bmatrix} \text{pred} & \text{'PRO'} \\ \text{Refl} & + \end{bmatrix} \\ \text{Pred} & \text{'picture. <(Obl}_{th}\text{)>'} \end{bmatrix} \\ \text{Subj} & [\text{Pred} \quad \text{'Taara'}] \\ \text{Obj} & [\quad] \\ \text{Pred} & \text{'like <(↑ Subj) (↑ Obj)>} \end{bmatrix}$$

→ In LFG, the verb is converted to the participial form but the sub-categorization is changed directly.

Consider the following example:

Active: Tara ate the food.

Passive: The food was eaten by Tara.

Active: ↑ Pred = 'eat < (↑ Subj) (↑ Obj) >'

Passive: ↑ Pred = 'eat < (↑ Obl_{ag}) (↑ Subj) >'

Here, Oblag represents Oblique agent phrase. Similar rules can be applied in active and dative constructs for the verbs that accept two objects.

Active: Tara gave a pen to Monika.

Passive: Tara gave Monika a pen.

Active: ↑ Pred = 'give < (↑ Subj) (↑ Obj_2) (↑ Obj) >'

Passive: ↑ Pred = 'give < (↑ Subj) (↑ Obj) (↑ Obl_{go}) >'

Here, Oblgo stands for Oblique goal phrase. Similar rules are also applicable to the procen of causativization. This can be seen in Hindi where the verb form is changed as follows:

हँसना → (Causativization) → हँसाना

Laugh — Laugh-cause-past

made to laugh

The auxilary verbs follow the main verb. In Hindi, they remain as separate words, whereas in south Indian languages they combine with the main verb.
For example:

खा रहा है	करता रहा है
Khaa raha hai	Kartaa rahaa hai
eat-ing	doing been has
eating	has been doing

In Hindi, some verbs (main), e.g., give (देना), take (लेना), also combine with other verbs (main) to change the aspect and modality of the verbs.

CHETAN. R
Asst. Professor

Example

उसने खाना खाया।	उसने खाना खा लिया।
Usne Khanaa Khaayaa	Usne Khaanaa Kha Liyaa
He (Subj) food ate	He (Subj) food eat taken
He ate food	He ate food
वह चला	वह चल दिया
He moved.	He move given
	He moved.

⑤ Paninian Framework

Paninian grammar was written by Panini in 500 BC in Sanskrit, the framework can be used for other Indian languages and possibly some Asian languages as well.

Unlike English, Asian languages are SOV (Subject-Object-Verb) ordered and inflectionally rich. The inflections provide important syntactic and semantic cues for language analysis and understanding. The Paninian framework takes advantages of there features.

Some Important Features of Indian Languages

Indian languages have traditionally used oral communication for knowledge propagation. In Hindi, we can change the position of subject and object. For example:

(a) माँ बच्चे को खाना देती है।
Maan Bachche ko khanaa detii hai
Mother child to food give-(s)
Mother gives food to the child.

(b) बच्चे को माँ खाना देती है।
Bachche ko Maan khanaa detii hai
child to mother food give-(s)
Mother gives food to the child.

Vibhakti literally means inflection, it refers to word (noun, verb or other) groups based on either on case endings, or post-positions or compound verbs or main and auxiliary verbs etc. Instead of talking about NP, VP, AP, PP, etc., word groups are formed based on various kinds of markers.

Karaka literally means Case. Paninian Grammar has its own way of defining Karaka relations. These relations are based on the way the word groups participate in the activity denoted by the verb group.

Karaka Theory :-

It is the central theme of PG framework. Karaka relations are assigned based on the roles played by various participants in the main activity. These roles are reflected in the case markers and post-position markers.

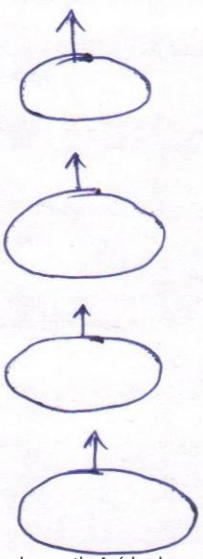
We will discuss the various karakas, such as Karta (subject) Karma (Object), Karana (instrument), Sampradana (beneficiary) Apadan (Separation) and Adhikaran (locus).

In Indian languages, the nouns are followed by post-positions instead of prepositions. They generally remain as separate words in Hindi, except in the case of pronouns, for example

रेखा के पिता	उसके पिता
Rekha ke pita	Uske pita
Rekha of father	Her (His) father.
Father of Rekha	

Layered Representation in PG

The GB theory represents three syntactic levels: deep structure, surface structure and logical form, where the LF is nearer to semantics. Paninian grammar framework is said to be syntactico-semantic, that is one can go from surface layer to deep semantics by passing through intermediate layers. The language can be represented as follows:



for example,

माँ ने थाली से खाना उठाकर बच्चे को दिया।

Maan ne thaali se khana uthakar bachche ko diyaa.

The mother gave food to the child taking it up from the plate.

Here thaali is the Apaadan.

'Adhikaran' is the locus of Karta or karma. aangan (courtyard) is the Adhikaran.

Issues in Paninian Grammar

The two problems challenging linguists are:

(i) Computational implementation of PG and

(ii) Adaptation of PG to Indian, and other similar languages.

The approach 'Utsarga-Apvada' where rules are arranged in multiple layers in such a way that each layer consists of rules which are in exception to rules in higher layer. Thus, as we go down the layer, more particular information is derived.

To explain various Karaka relations, let us consider the example.

माँ बच्ची को आँगन में हाथ से रोटी खिलाती है।

Maan bachchi ko aangan mein hath se rotii khilaathi hei

Mother child-to courtyard-in hand-by bread feed(s)

The mother feeds bread to the child by hand in the courtyard.

The first important Karak is subject, called 'Karta' in PG.

Karta is defined as the noun group which is most independent. Karta has generally 'ne' or 'ϕ' case marker. It is an independent entity in the activity denoted by the main verb. In the above sentence 'maan' (mother) is Karta.

Kaema is similar to object and is the locus of the result of the activity. In sentence rotii (bread) is the Kaema.

Another Karaka relation is 'Karan' (instrument) which is a noun group through which the goal is achieved. In the sentence haath (hand) is the Karan.

'Sampradan' is the beneficiary of the activity.

e.g. bachchi (child).

'Apaadaan' denotes separation and the marker is attached to the part that serves as a reference point.

A model that limits the history to the previous one word only is termed as bi-gram ($n=1$) model.

A model that conditions the probability of a word to the previous two words, is called a tri-gram model ($n=2$).

bi-gram: $$P(s) \approx \prod_{i=1}^{n} P(w_i / w_{i-1})$$

tri-gram: $$P(s) \approx \prod_{i=1}^{n} P(w_i / w_{i-2}, w_{i-1})$$

As an example, the bi-gram approximation of P(east / The Arabian knights are fairy tales of the) is

P(east/the)

Whereas a tri-gram approximation is

P(east/ of the).

We estimate n-gram parameters using the maximum likelihood estimation technique. We count a particular n-gram in the training corpus and divide it by the sum of all n-grams that share the same prefix.

$$P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{\sum_{w} C(w_{i-n+1}, \ldots, w_{i-1}, w)}$$

STATISTICAL LANGUAGE MODEL :-

A statistical language model is a probability distribution $P(s)$ over all possible word sequences.

n-gram model :-

The goal of statistical language model is to estimate the probability of a sentence. This is achieved by decomposing sentence probability into a product of conditional probabilities using the chain rule. as follows:

$$P(s) = P(w_1, w_2, w_3, \ldots, w_n)$$

$$= P(w_1)\, P(w_2/w_1)\, P(w_3/w_1, w_2)\, P(w_4/w_1\, w_2\, w_3) \ldots P(w_n/w_1\, w_2 \ldots w_{n-1}))$$

$$= \prod_{i=1}^{n} P(w_i/h_i)$$ where h_i is history of word w_i defined as $w_1\, w_2 \cdots w_{i-1}$

In order to calculate sentence probability we need to calculate the probability of a word, given the sequence of words preceding it.

An n-gram model simplifies the task by approximating the probability of a word given all the previous words by the conditional probability given previous $n-1$ words only.

$$P(w_i/h_i) = P(w_i / w_{i-n+1} \cdot w_{i-1})$$

Test Sentence(s): The Arabian Knights are the fairy tales of the east.

$P(\text{The}/\langle s\rangle) \times P(\text{Arabian}/\text{the}) \times P(\text{knights}/\text{Arabian}) \times P(\text{are}/\text{knights})$
$\times P(\text{the}/\text{are}) \times P(\text{fairy}/\text{the}) \times P(\text{tales}/\text{fairy}) \times P(\text{of}/\text{tales})$
$\times P(\text{the}/\text{of}) \times P(\text{east}/\text{the})$

$= 0.67 \times 0.5 \times 1.0 \times 1.0 \times 0.5 \times 0.2 \times 1.0 \times 1.0 \times 1.0 \times 0.2$

$= 0.0067$

CHETAN. R
Asst. Professor

The n-gram model suffers from data sparseness problem. An n-gram that does not occur in the training data is assigned zero probability, so that even a large corpus has several zero entries in its bi-gram matrix.
This is because of the assumption that the probability of occurence of a word depends only on the preceding word which is not true in general.

A number of smoothing techniques have been developed to handle the data sparseness problem, the simplest of these being add-on smoothing.

"Smoothing in general refers to the task of re-evaluating zero-probability or low-probability n-grams and assigning them non-zero values."

The sum of all n-grams that share first n-1 words is equal to the count of the common prefix $w_{i-n+1}, \ldots, w_{i-1}$.

So, we rewrite the previous expression as follows:

$$P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})}$$

Example :-

Training set:

The Arabian Knights

These are the fairy tales of the east

The stories of the Arabian knights are translated in many languages.

Bi-gram model:

P(the/<s>) = 0.67

P(Arabian/the) = 0.4

P(knights/Arabian) = 1.0

P(are/there) = 1.0

P(tales/fairy) = 1.0

P(east/the) = 0.2

P(are/knights) = 1.0

P(many/in) = 1.0

P(languages/many) = 1.0

P(the/are) = 0.5

P(of/tales) = 1.0

P(stories/the) = 0.2

P(translated/are) = 0.5

P(fairy/the) = 0.2

P(the/of) = 1.0

P(of/stories) = 1.0

P(in/translated) = 1.0

Example, Consider that the number of n-grams that occur 4 times is 25,108 and the number of n-grams that occur 5 times is 20,542. Then the Smoothed count for 5 will be

$$\frac{20542}{25108} \times 5 = 4.09$$

CHETAN. R
Asst. Professor

Caching Technique:

Another improvement over basic n-gram model is caching. The frequency of n-gram is not uniform across the text segments or corpus.

Certain words occur more frequently in certain segments and rarely in others.

The cache model combines the most recent n-gram frequency with the standard n-gram model to improve its performance locally.

Add one Smoothing :-

This is the simplest smoothing technique. It adds a value of one to each n-gram frequency before normalizing them into probabilities. Thus the conditional probability becomes:

$$P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1}) + V}$$

where V is the vocabulary size, i.e., size of the set of all the words being considered.

The add-one smoothing is not considered as good smoothing technique. It assigns the same probability to all missing n-grams, even though some of them could be more intuitively appealing than others.

Good-Turing Smoothing

It adjusts the frequency f of an n-gram using the count of n-grams having a frequency of occurence 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 is the number of n-grams that occur exactly f times in the training corpus.