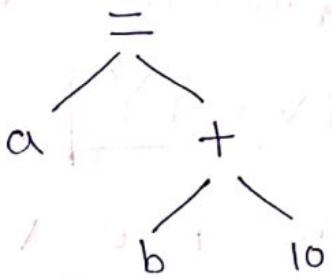
UNIT II : Syntax Analysis (Part-I)

* **Parser** : A Parsing or syntax analysis is a process which takes the input string 'w' & Produces either a Parse tree (syntactic structure) or generates the syntactic errors.

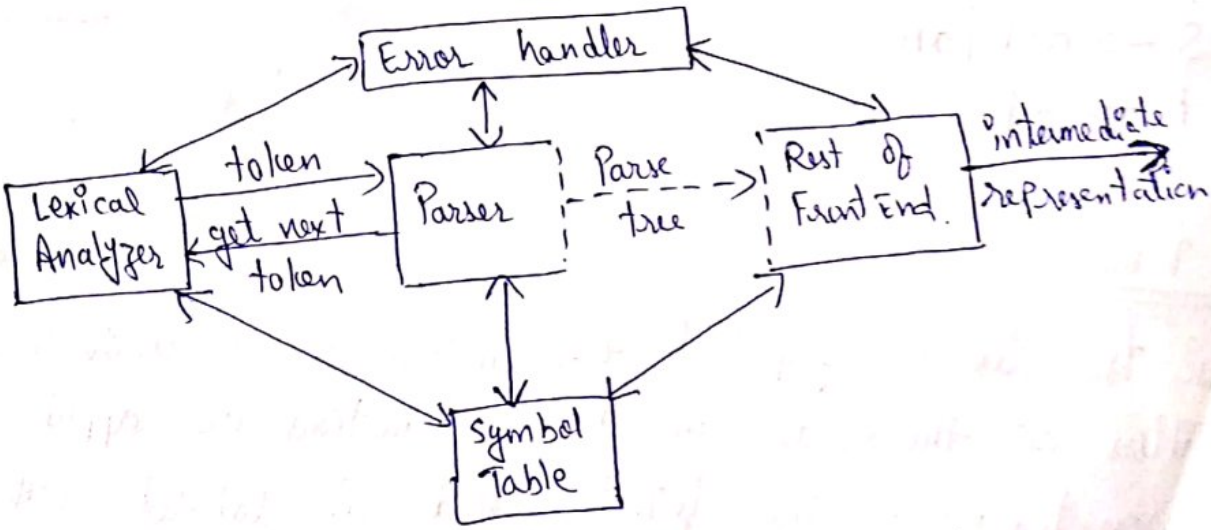
→ The syntax analyzer takes the tokens as input & generates a tree like structure called '**Parse tree**'.

Eg : $a = b + 10$



Parse tree for $a = b + 10$.

(*)



Position of Parser in compiler model

→ Specification of input can be done by using "Context-Free Grammar" (CFG).

Scanned by CamScanner

* Context-Free Grammar:

The CFG, 'G' is a collection of the following:

$$G_1 = (V, T, S, P)$$

- 'V' is a set of non-terminals (All upper-case letters) (Syntactic variables that denote sets of strings)
- 'T' " " " " terminals (basic symbols from which strings are formed)
- 'S' is a start symbol
- 'P' is a set of Production rules.

→ The Production rules are must be in the format:

$$\boxed{\text{Non-terminal} \longrightarrow (V \cup T)^*}$$

(or) $\boxed{A \longrightarrow \alpha}$ where $A \in V$

$\alpha \in (V \cup T \cup \epsilon)^*$.

eg: $S \rightarrow aSb \mid aA$

$A \rightarrow a \mid \epsilon$.

* Parse trees:

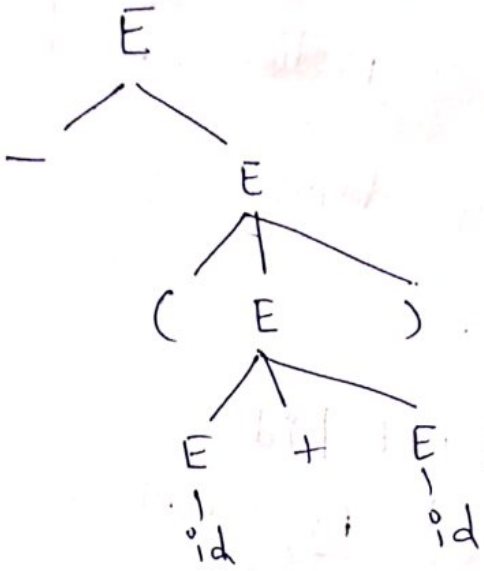
A Parse tree is a graphical representation of a derivation that filters out the order in which productions are applied to replace non-terminals. The interior node is labeled with the non-terminal.

→ Each interior node of a Parse tree represents the application of a Production.

Scanned by CamScanner

Eg: $E \rightarrow E+E \mid E*E \mid -E \mid (E) \mid id$.

generate Parse tree for $-(id+id)$

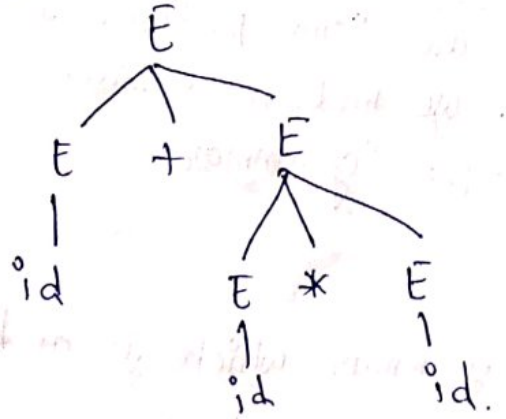


Parse tree

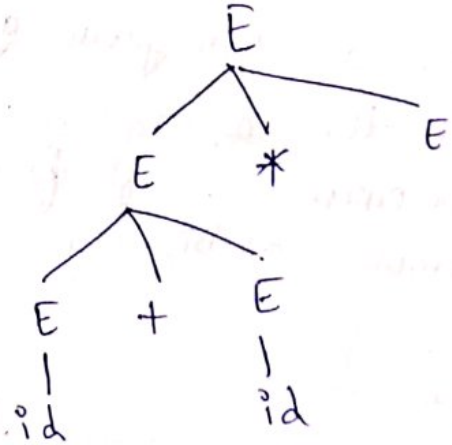
* **Ambiguous Grammar:** A grammar is said to be ambiguous if it generates more than one parse trees for same sentence of language $L(G)$.

Eg: $E \rightarrow E+E \mid E*E \mid id$.

id+id*id:



Parse tree 1



Parse tree 2

Since we have more than one Parse tree for id+id*id, the given grammar is Ambiguous Grammar.

Scanned by CamScanner

* **Derivation** : Derivation from 'S' means generation of string 'w' from 'S'. For constructing derivation two things are important.
  - choice of Non-terminal from several others.
  - choice of rule from Production rules for corresponding non-terminal.

→ we have two types of derivations:
  a) Left most derivation
  b) Right most derivation.

* **Example** : $E \rightarrow E+E \mid E*E \mid id$

Leftmost derivation

$E \rightarrow E+E$
$E \rightarrow id+E$
$E \rightarrow id+E*E$
$E \rightarrow id+id*E$
$E \rightarrow id+id*id$.

Right most derivation

$E \rightarrow E*E$
$E \rightarrow E*id$
$E \rightarrow E+E*id$
$E \rightarrow E+id*id$
$E \rightarrow id+id*id$

* **Eliminating ambiguity** :

Sometimes an Ambiguous grammar can be rewritten to eliminate the ambiguity. There are some problems like left recursion, left factoring. we need to eliminate these to remove ambiguity in the grammar.

(1) **Left Recursion** :

The left recursive grammar is a grammar which is as below :

$A \overset{+}{\Longrightarrow} A\alpha$ where A is a non-terminal
α denotes some i/p string.

(A ... same) $\overset{+}{\Longrightarrow}$ means deriving the i/p in one or more steps.

Scanned by CamScanner

→ Because of left recursion, the top down parser can enter into infinite loop. So we need to eliminate it.

* Removing left recursion

> $A \longrightarrow A\alpha \mid \beta$ is rewritten as :
>
> $A \longrightarrow \beta A'$
> $A' \longrightarrow \alpha A' \mid \epsilon$

Example :

$E \longrightarrow E+T \mid T$
$T \longrightarrow T*F \mid F$
$F \longrightarrow id \mid (E)$

we have left recursion in the following Production rules.

$E \longrightarrow E+T \mid T$ (same A; α = $+T$; β = T)

$T \longrightarrow T*F \mid F$ (same A; α = $*F$; β = F)

removing left recursion we get :

$E \longrightarrow TE'$
$E' \longrightarrow +TE' \mid \epsilon$
$T \longrightarrow FT'$
$T' \longrightarrow *FT' \mid \epsilon$
$F \longrightarrow id \mid (E)$

$\frac{E}{A} \longrightarrow \frac{E}{A}\frac{+T}{\alpha} \mid \frac{T}{\beta}$

$\frac{T}{A} \longrightarrow \frac{T}{A}\frac{*F}{\alpha} \mid \frac{F}{\beta}$

Scanned by CamScanner

② Left factoring :

Consider a grammar :

$$A \longrightarrow \alpha\beta_1 \mid \alpha\beta_2$$

Here it is not possible for us to take a decision whether to chose first rule or second. The above grammar can be left factored as : rewritten as :

$$A \longrightarrow \alpha A'$$
$$A' \longrightarrow \beta_1 \mid \beta_2$$

* Example : Grammar : $S \longrightarrow iEtS \mid iEtSeS \mid a$
$E \longrightarrow b$

The above grammar after left factoring becomes as :

$S \longrightarrow iEtSS' \mid a$
$S' \longrightarrow eS \mid \epsilon$
$E \longrightarrow b.$

$S \longrightarrow \underset{\alpha}{iEtS} \mid \underset{\alpha}{iEtSeS} \mid \underset{\beta_2}{a}$

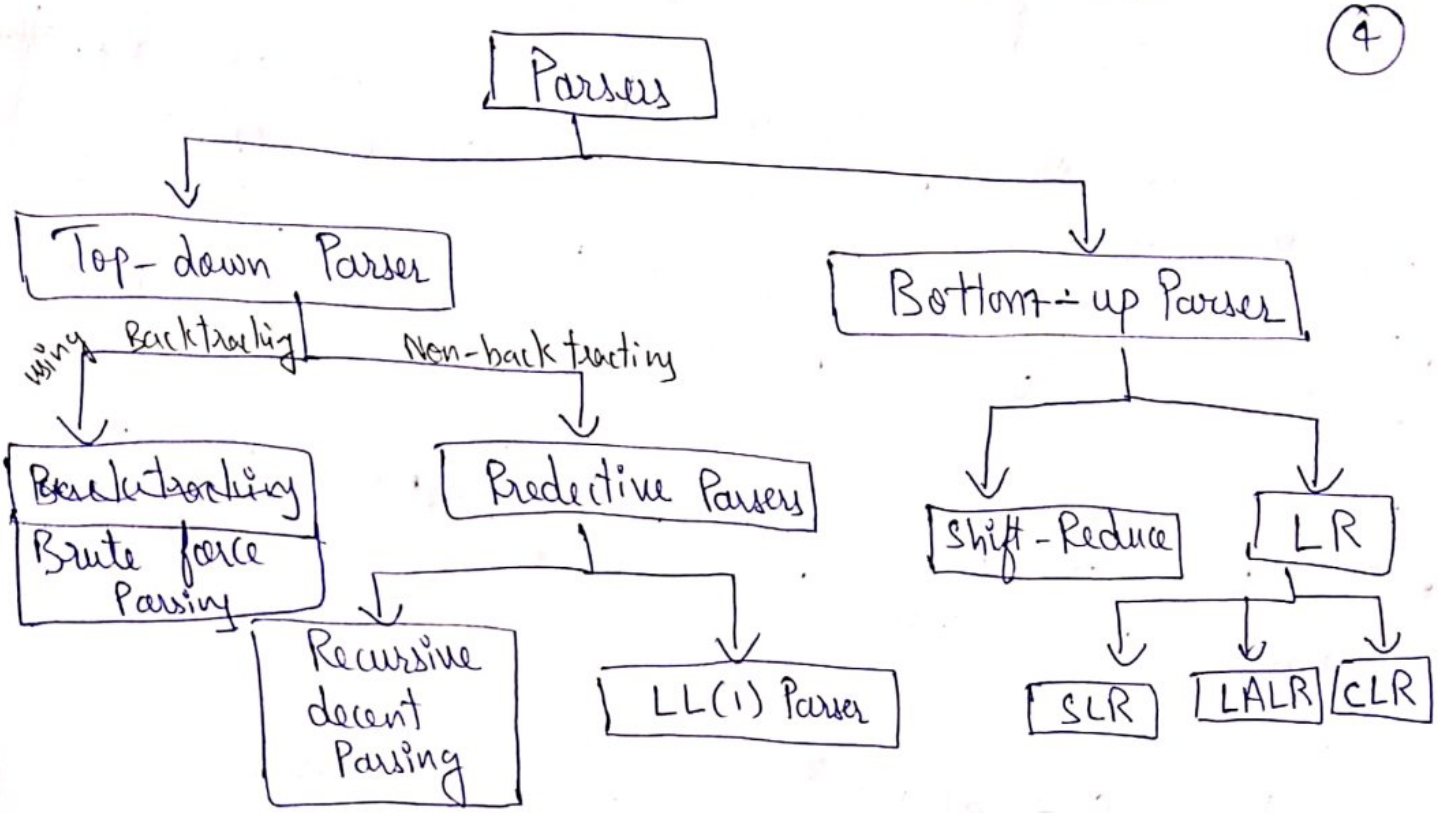
→ By left factoring we may be able to rewrite the Production in which the decision can be deferred until enough of the input is seen to make the right choice.

* Parsing Techniques :

There are two Parsing techniques :

1. Top - Down Parsing (we will Parse from Root to leaf node)
2. Bottom - Up Parsing. (" " " leaf to root node)

Scanned by CamScanner



Types of Parsing Techniques

* (1) Top Down Parsing :-

The top-down construction of a Parse tree is done by starting with the root, labeled with the starting non-terminal, & repeatedly Performing the following two steps

- At node n, labeled with non-terminal A, select one of the Productions for 'A' & construct children at 'n' for the Symbols on the right side of the Production.
- Find the next node at which a sub-tree is to constructed.

→ Top-down Parsing can be viewed as an attempt to find a left-most derivation for an input string. A general form of top-down Parsing is called "Recursive descent Parsing".

Scanned by CamScanner

(a) Brute-force Parsing :

- Given a particular non-terminal that is to be expanded, the first Production for this non-terminal can be applied.
- Then, within this newly expanded string, the left most non-terminal is selected for expansion & its first Production is applied.
- This Process of applying Productions is repeated for all subsequent non-terminals that are selected untill the Process cannot be continued.

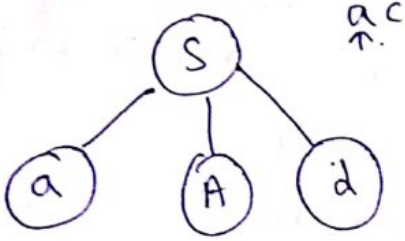
* Example : S → aAd | aB , w = accd.

A → b | c

B → ccd | ddc

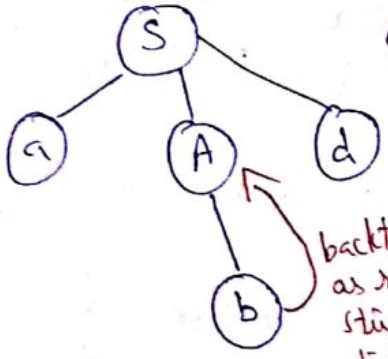
step 1 : S — accd.

step 2 :



accd

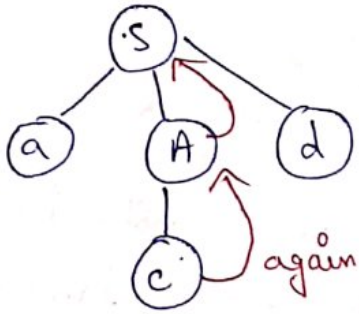
step 3 :



accd

Generated : ab.

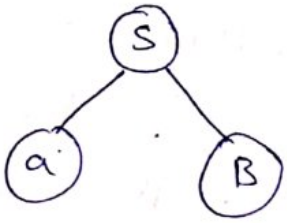
step 4 :



acd we got

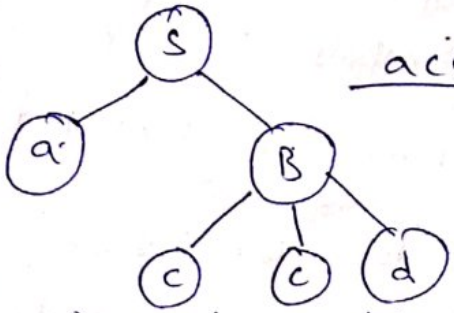
accd

step 5 :



accd

step 6 :



accd

here the given string match with generated string is done so it is ~~success~~ successful Passing.

Scanned by CamScanner

(b) Recursive Descent Parsing:

A Recursive-descent parsing program consists of a set of Procedures, one for each non-terminal. Execution begins with the Procedure for the start symbol, which halts & announces success if its Procedure body scans the entire input string.

→ CFG is used to build the recursive routines. The RHS of the Production rule is directly converted to a Program. For each non-terminal a separate Procedure is written & body of the Procedure (code) is RHS of the corresponding non-terminal.

* Basic steps for construction of Recursive descent Parser:

The RHS of the rule is directly converted into program code symbol by symbol

(1) If the input symbol is non-terminal then a call to the Procedure corresponding to the non-terminal is made.

(2) If the input symbol is terminal then it is matched with the lookahead input. The lookahead pointer has to be advanced on matching of the input symbol.

(3) If the Production rule has many alternates then all these alternates has to combined into a single body of Procedure

(4) The Parser should be activated by a Procedure corresponding to the start symbol.

Scanned by CamScanner

```
Void A() {
    Choose an A-production, A → X1 X2 .... Xk ;
    for (i = 1 to k) {
        if (Xi is a nonterminal)
            call Procedure Xi() ;
        else if (Xi equals the current input symbol a)
            advance the input to the next symbol ;
        else /* an error has occured */ ;
    }
}
```

A typical procedure for a non-terminal in a top-down Parser

(c) Predictive LL(1) Parsing :

Predictive LL(1) Parsing is of non-recursive type. In this type of Parsing a table is built.

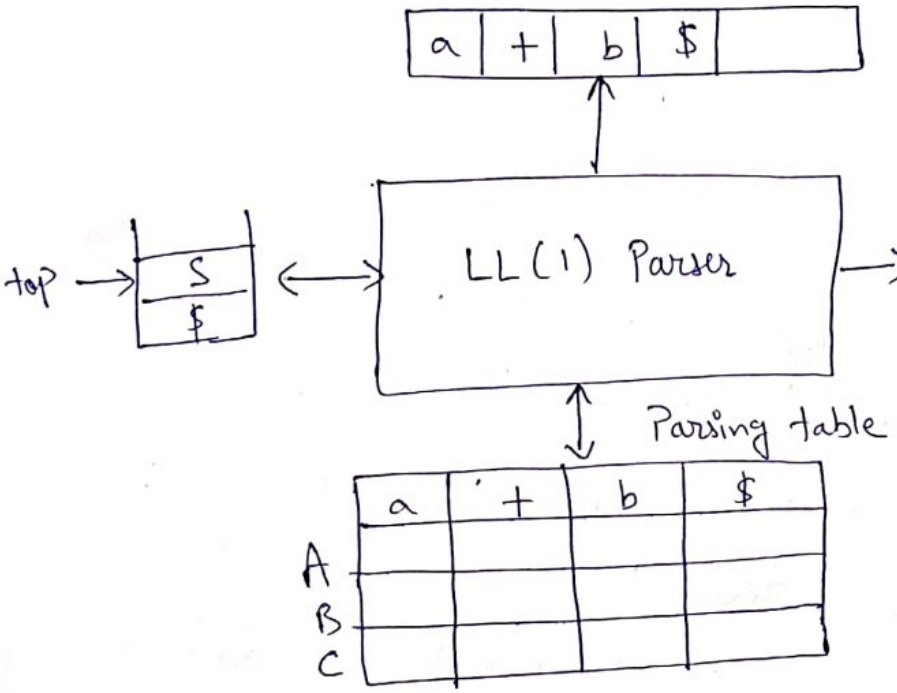
LL (1) → lookahead (uses only one i/p symbol to Predict Parsing Process)

- i/p is scanned from left to right.
- leftmost derivation

→ The data structures used by LL(1) are a) input buffer, b) stack, c) Parsing table. The LL(1) parser uses input buffer to store the input tokens. The stack is used to hold the left sentential form. The symbols in RHS of rule are pushed into the stack in reverse order i.e from right to left.

Scanned by CamScanner

Use of stack makes this parsing non-recursive. The table is basically a two dimensional array. The table can be represented as M[A, a] where 'A' is a non-terminal & 'a' is current input symbol..



Model for LL(1) Parser

→ The Parsing Program reads top of the stack & a current input symbol. with the help of these two symbols the Parsing action is determined. The Parser consults the table M[A, a] each time while taking the parsing actions. Hence this type of Parsing method is called "table drive Parsing".

* Construction of Predictive LL(1) Parsing:

~~Step 1:~~ ~~Comp~~

→ In LL(1) Parsing, Pre-Processing steps are as follows:

Scanned by CamScanner

1. Elimination of left Recursion.
2. left-factoring the grammar.
3. Computation of FIRST & FOLLOW.
4. Constructing Predictive Parsing Table using FIRST & FOLLOW.
5. Parse the input string with the help of Predictive Parsing table.

* FIRST( ) :

FIRST(α) is a set of terminal symbols that are first symbols appearing at RHS in derivation of 'α'.

→ Following are the rules used to compute FIRST functions:

- If 'X' is a terminal then First(X) is just 'X'.
- If there is a Production $X \rightarrow \epsilon$(epsilon) then add 'ϵ' to first(X).
- If there is a Production $X \rightarrow aB$ then add 'a' to first(X) where 'a' is a terminal & 'B' is non-terminal.
- If there is a Production $X \rightarrow Y_1Y_2 \cdots Y_k$ then first($Y_1Y_2 \cdots Y_k$) to first(X).
- First($Y_1Y_2 \cdots Y_k$) is either:
  - i) First(Y_1) (if First(Y_1) doesn't contain 'ϵ')
  - ii) OR (if First(Y_1) does contain 'ϵ') then First($Y_1Y_2 \cdots Y_k$) is everything in First(Y_1) <except for 'ϵ'> as well as everything in First($Y_2 \cdots Y_k$).

Scanned by CamScanner

iii) If First(Y_1) First(Y_2) ... First(Y_k) all contain 'ϵ' then add 'ϵ' to First($Y_1 Y_2 \cdots Y_k$) as well.

* Example :

(1). $E \rightarrow E+T \mid T$ (Left Recursion exists.)
$T \rightarrow T*F \mid F$
$F \rightarrow (E) \mid id$

sol : we need to remove left recursion from the grammar.
After removal of left recursion the grammar is :

$E \rightarrow TE'$
$E' \rightarrow +TE' \mid \epsilon$
$T \rightarrow FT'$
$T' \rightarrow *FT' \mid \epsilon$
$F \rightarrow (E) \mid id.$

(*) First(E) = $E \rightarrow T$ now go to 'T' RHS production & look.
$T \rightarrow FT'$ (again non-terminal)
so go to 'F' RHS side production
$F \rightarrow (E) \mid id$ (id → terminal.)
∴ First(E) = { (, id }

- First(E) = First(T) = First(F) = { (, id }
- First(E') = { +, ϵ }
- First(T') = { *, ϵ }

(2). $S \rightarrow aB \mid aC \mid Sd \mid Se$
$B \rightarrow bBc \mid f$
$C \rightarrow g.$

sol : First(S) = { a }
First(B) = { b, f }
First(C) = { g }

Scanned by CamScanner

* FOLLOW ( ) :

Let $S \rightarrow \alpha A a \beta$ where α & β may be terminals or non-terminals. FOLLOW(A) is defined as the set of terminal symbols that appear immediately to the right of 'A'.

→ Rules for computing FOLLOW are :

- For the start symbol 'S' place "$" in FOLLOW (S).
- If there is a production $A \rightarrow \alpha B \beta$ then everything in FIRST (β) without 'ϵ' is to be placed in FOLLOW(B).
- If there is a Production $A \rightarrow \alpha B \beta$ @L $A \rightarrow \alpha B$ and FIRST (β) = {ϵ} then FOLLOW (A) = FOLLOW (B) i.e everything in FOLLOW (A) is in FOLLOW (B).

* Examples :

(1)
$E \rightarrow TE'$
$E' \rightarrow +TE' \mid \epsilon$
$T \rightarrow FT'$
$T' \rightarrow *FT' \mid \epsilon$
$F \rightarrow (E) \mid id$.

Sol : • FOLLOW (E) = { $,) }

check all RHS side Productions where 'E' is there, we have
$F \rightarrow (E) \mid id$
↳ first(')') = ')'.
∴ FOLLOW(E) = { $,) }
since 'E' is start symbol place '$' in its follow.

- FOLLOW(E') = { $,) }
- FOLLOW (T) = { +, $,) } ⇒ $E \rightarrow TE'$
  first(E') = {+, ε} → to follow (E) = FOLLOW (T)
  FOLLOW(T) = {+, $,) }
- FOLLOW (T') = { +, $,) }
- FOLLOW (F) = { *, +, $,) }

Scanned by CamScanner

(2). $S \rightarrow aB \mid aC \mid Sd \mid Se$

$B \rightarrow bBc \mid f$

$C \rightarrow g$.

Sol: FOLLOW(S) = { $, d, e }

FOLLOW(B) = { c }

FOLLOW(C) = { d, e, $ }

* Construction of Predictive Parsing table:

The construction of Predictive Parsing table is an important activity in Predictive Parsing method. This algorithm requires FIRST & FOLLOW functions.

Input: Context-free grammar 'G'.

Output: Predictive Parsing table 'M'.

Algorithm: For the rule $A \rightarrow \alpha$ of grammar 'G':

- For each 'a' in FIRST(α) create entry M[A, a] = $A \rightarrow \alpha$ where 'a' is a terminal symbol.
- For 'ε' in FIRST(α) create entry M[A, b] = $A \rightarrow \alpha$ where 'b' is the symbols from FOLLOW(A).
- All the remaining entries in the table 'M' are marked as syntax error.

→ This algorithm can be applied to any grammar 'G' to produce a Parsing table 'M'. If the grammar 'G' is left recursive or ambiguous then 'M' will have atleast one multiple defined entry.

Scanned by CamScanner

* LL(1) grammars :- A grammar whose Parsing table has no multiple-defined entries is said to be LL(1). The LL(1) grammars are:

- Scanned from left-to-right.
- are Parsed by a leftmost derivation
- have one symbol look ahead.

* Problems :-

(1) Verify whether the following grammar is LL(1) or not.

$E \rightarrow E+T \mid T$

$T \rightarrow T*F \mid F$

$F \rightarrow (E) \mid id.$

sol :- we need to remove left recursion. After removal the grammar is:

$E \rightarrow TE'$

$E' \rightarrow +TE' \mid \epsilon$

$T \rightarrow FT'$

$T' \rightarrow *FT' \mid \epsilon$

$F \rightarrow (E) \mid id$

- First(E) = First(T) = First(F) = {(, id}
- First(E') = {+, ε}
- First(T') = {*, ε}

- Follow(E) = {$,)}
- Follow(E') = {$,)}
- Follow(T) = {+, $,)}
- Follow(T') = {+, $,)}
- Follow(F) = {*, +, $,)}

Scanned by CamScanner

* Predictive Parsing Table :

Non-terminals → Input | terminals → id, +, *, (,), $ | First Productions → (| Follow sets →), $

Input	id	+	*	(	)	$
E	$E \to TE'$			$E \to TE'$		
E'		$E' \to +TE'$			$E' \to \epsilon$	$E' \to \epsilon$.
T	$T \to FT'$			$T \to FT'$		
T'		$T' \to \epsilon$	$T' \to *FT'$		$T' \to \epsilon$	$T' \to \epsilon$
F	$F \to id$			$F \to (E)$		

Since there are no multiple entries in the above table, the given grammar is LL(1) grammar.

Input string : id + id * id $ is parsed using above table as below :

Stack	Input	Action
$ E' T → store in reverse order	id + id * id $	$E \to TE'$
$ E' T' F	id + id * id $	$E \to FT'$
$ E' T' id	id + id * id $	$F \to id$
$ E' T'	+ id * id $	
$ E'	+ id * id $	$T' \to \epsilon$
$ E' T +	+ id * id $	$E' \to +TE'$
$ E' T	id * id $	stack top & current pointer of input points to same character. Hence Pop it.

Scanned by CamScanner

stack	Input	Action
$ E' T' F	id * id $	T → F T'
$ E' T' id	id * id $	F → id
$ E' T'	* id $	
$ E' T' F *	* id $	T → * F T'
$ E' T' F	id $	
$ E' T' id	id $	F → id
$ E' T'	$	
$ E'	$	T → ϵ
$	$	E' → ϵ

∴ Thus input string gets parsed.

(2) check whether the following grammar is LL(1) grammar or not.

$$S \rightarrow iEtS \mid iEtSeS \mid a$$
$$E \rightarrow b$$

Sol:- After left factoring, the above grammar becomes as:

$$S \rightarrow iEtSS' \mid a$$
$$S' \rightarrow eS \mid \epsilon$$
$$E \rightarrow b$$

First & Follow for the above grammar:-

Scanned by CamScanner

- First (S) = { i, a }
- First (S') = { e, ϵ }
- First (E) = { b }
- Follow (S) = { $, e }
- Follow (S') = { $, e }
- Follow (E) = { t }

→ Predictive Parsing table can be constructed as:

	a	b	e	i	t	$
S	S → a			S → i E t S S'		
S'			S' → ϵ S' → e S	multiple entries.		S' → ϵ
E		E → b				

Since we got multiple entries in M[S', e], this grammar is not LL(1) grammar.

* Error Recovery in Predictive Parsing:

e/n error is detected during Predictive Parsing when the terminal on top of stack does not match the next input symbol or when non-terminal 'A' is on top of the stack, 'a' is the next input symbol, & the parsing table entry M[A, a] is empty.

Scanned by CamScanner

There are two types of error recovery techniques :-

(a) Panic-mode error recovery :- It is based on the idea of skipping symbols on the input until a token in a selected set of synchronizing tokens appears. Its effectiveness depends on the choice of synchronizing set. The sets should be chosen so that the parser recovers quickly from errors that are likely to occur in practice.

(b) Phase level Recovery :- It is implemented by filling in the blank entries in the predictive parsing table with pointers to error routines. These routines may change, insert or delete symbols on the input & issue appropriate error messages. They may also pop from the stack.

* Bottom-Up Parsing :-

A Bottom-up parse corresponds to the construction of a parse tree for an input string beginning at the leaves (the bottom) & working up towards the root (the top).

→ Input string is taken first & we try to reduce this string with the help of grammar & try to obtain the start symbol. The process of Parsing halts successfully as soon as we reach to start symbol.

Scanned by CamScanner

* Reduction :- Parser tries to identify RHS of production rule & replace it by corresponding LHS. This activity is called "Reduction".

→ The Primary task in bottom-up Parsing is to find the productions that can be used for reduction. The bottom up Parse tree construction process indicates that the tracing of derivations are to be done in reverse order.

* Handle Pruning :-

Handle is a string of substring that matches the right side of Production & we can reduce such string by a non-terminal on left hand side Production.

(OR)

Handle of right sentential form γ is a production $A \rightarrow \beta$ & a position of γ where the string β may be found & replaced by 'A' to produce the previous right sentential form in rightmost derivation of γ.

*Example :- $E \rightarrow E+E$
$E \rightarrow E*E$
$E \rightarrow (E)$
$E \rightarrow id$

Consider string id. + id. * id.

$E \xrightarrow{rm} E+\underline{E}$
$E \xrightarrow{rm} E+\underline{E*E}$
$E \xrightarrow{rm} E+E*\underline{id}$

Scanned by CamScanner

$E \xrightarrow{rm} E + \underline{id} * id$

$E \xrightarrow{rm} \underline{id} + id * id$

→ A right most derivation in reverse order can be obtained by "Handle Pruning".

Right Sentential Form	Handle ← Handle	Reduced Production.
$\underline{id} + id * id$	id	$E \rightarrow id$
$E + \underline{id} * id$	id	$E \rightarrow id$
$E + E * \underline{id}$	id	$E \rightarrow id$
$E + \underline{E * E}$	$E * E$	$E \rightarrow E * E$
$\underline{E + E}$	$E + E$	$E \rightarrow E + E$
E		

* Shift - Reduce Parsing :

Shift-reduce Parsing is a form of bottom-up Parsing in which a stack holds grammar symbols & an input buffer holds the rest of the string to be parsed.

→ The initial configuration of shift reduce Parser is as below:

STACK	INPUT
$	ω $

where stack is empty & 'ω' is a string.

Scanned by CamScanner

The shift-Reduce Parser can Perform the following operations:

(1) Shift : Moving of symbols from input buffer onto the stack.

(2) Reduce : If the handle appears on the top of the stack then reduction of it by appropriate rule is done. That means RHS of rule is popped off & LHS is pushed in. This action is called 'Reduce action'.

(3) Accept : Announce successful completion of Parsing if the stack contains start symbol only & input buffer is empty at the same time.

(4) Error : A situation in which parser cannot either shift or reduce the symbols. It cannot even. Perform the accept action. This is called as error.

* Example :

(1) E → E-E
E → E*E
E → id

Perform shift-Reduce parsing of the input string "id1 - id2 * id3".

Sol :

Stack	Input Buffer	Parsing Action
$	id1 - id2 * id3 $	shift
$id1	- id2 * id3 $	Reduce by E → id
$ E	- id2 * id3 $	shift
$ E -	id2 * id3 $	shift
$ E - id2	* id3 $	Reduce by E → id
$ E - E	* id3 $	shift

Scanned by CamScanner

$ E-E↑	id3 $	shift
$ E-E*id3	$	Reduce by E→id
$ E-E*E	$	Reduce by E→E*E
$ E-E	$	Reduce by E→E-E
$ E	$	Accept

Here we have followed two rules:

1. If the incoming operator has more priority than in stack operator then perform 'shift'.
2. If the stack operator has same or less priority than the priority of incoming operator then perform 'reduce'.

* Conflicts during shift-Reduce Parsing:

Shift-reduce Parsing cannot be used for all types of CFG's. For some CFG's, shift-reduce parser can reach a configuration in which the parser knowing the entire stack contents & the next input symbol, cannot decide whether to shift to reduce (a shift/reduce conflict) or cannot decide which of several reductions to make (a reduce/reduce conflict).

* Example: Consider the following grammar where the productions are numbered as below:

E → E + T { Print '1' }

E → T { Print '2' }

Scanned by CamScanner

$T \rightarrow T * F$ { Print '3' }

$T \rightarrow F$ { Print '4' }

$E \rightarrow (E)$ { Print '5' }

$F \rightarrow id$ { Print '6' }

If a shift-reduce Parser writes the Production number immedi-tely after Performing any Production, what string will be Printed if the Parser input is id+id*id.

Sol:

Stack	Input	Operation
$	id+id*id$	shift
$id	+id*id$	Reduced by F→id Print '6'.
$F	+id*id$	Reduced by T→F Print '4'.
$T	+id*id$	Reduced by E→T Print '2'
$E	+id*id$	shift
$E+	id*id$	shift
$E+id	*id$	Reduced by F→id Print '6
$E+F	*id$	Reduced by T→F Print '4
$E+T	*id$	shift
$E+T*	id$	shift
$E+T*id	$	Reduced by F→id Print '6
$E+T*F	$	Reduced by T→T*F Print '3'
$E+T	$	Reduced by E→E+T Print '1'.
$E	$	

Scanned by CamScanner

∴ the final string obtained is 64264631.

* Operator Precedence Parsing :

A grammar 'G' is said to be operator Precedence if it Posses following Properties :

- No Production on the right side is 'ϵ' (epsilon).
- There should not be any Production rule Possessing two adjacent non-terminals at the right hand side.

* Example :

$$E \rightarrow EAE \mid (E) \mid -E \mid id$$
$$A \rightarrow + \mid - \mid * \mid / \mid \wedge$$

This grammar is not an operator Precedent grammar as in the Production rule $E \rightarrow EAE$.

→ It contains two consecutive non-terminals. Hence first we will convert it into equivalent operator Precedence grammar by removing 'A'.

$$E \rightarrow E+E \mid E-E \mid E*E \mid E/E \mid E \wedge E$$
$$E \rightarrow (E) \mid -E \mid id$$

→ In operator Precedence Parsing, we will first define Precedence relations $<\cdot$, $=$, and $\cdot>$ between pair of terminals.

	means
• $P <\cdot q$	P gives more Precedence than q.
• $P = q$	P has same " as q.
• $P \cdot> q$	P takes Precedence over q.

Scanned by CamScanner

Operator Precedence (Relation) table :

	id	+	*	$
id		.>	.>	.>
+	<.	.>	.>	.>
*	<.	.>	.>	.>
$	<.	<.	<.	

Consider the string : id + id * id .

We will insert $ symbols at the start & end of the input string. We will also insert Precedence operator by referring the Precedence relation table.

$$\$ <\cdot \text{ id } \cdot> + <\cdot \text{ id } \cdot> * <\text{ id } \cdot> \$$$

* Steps to Parse the given string :

- Scan the input from left to right until first .> is encountered.
- Scan backwards over = until <. is encountered.
- The handle is a string between <. and .>.

Parsing is done as follows :

$ <. id .> + <. id .> * <. id .> $	Handle id is obtained b/w <. .> Reduce this by E → id
E + <. id .> * <. id .> $	Handle id is obtained b/w <. .> Reduce this by E → id
E + E * <. id .> $	Handle id is obtained b/w <. .> Reduce this by E → id

Scanned by CamScanner

E + E * E	Remove all the non-terminals
+ *	Insert $ at the beginning & at the end. Also insert the Precedence operators.
$ <· + <· * ·> $	The * operator is surrounded by <· ·>. This indicates that * becomes handle. That means we have to reduce E*E operation first.
$ <· + ·> $	Now '+' becomes handle. hence we evaluate E+E
$ $	Parsing is done.

* Advantage : It is simple to implement.

* Disadvantages :

- The operator like minus has two different Precedence (unary & binary). Hence it is hard to handle such tokens.
- This kind of Parsing is applicable to only small class of grammars.

* Application :

SNOBOL language uses operator Precedence Parsing.

Scanned by CamScanner

UNIT-III : Syntax Analysis (Part-II)

* Introduction to LR Parsing :

LR Parsing is the most efficient method of bottom up parsing which can be used to parse the large class of context-free grammars. This method is also called LR(k) parsing.

L R (k) → no. of i/p symbols (where "k" is omitted if "k" is assumed to be 1)

- L → left to right scanning
- R → Rightmost derivation in reverse order.

→ LR Parsers are widely used for the following reasons :

(1) LR Parsers can be constructed to recognize most of the programming languages for which context-free grammar can be written.

(2) The class of grammar that can be parsed by LR Parser is a superset of class of grammars that can be parsed using Predictive Parsers.

(3) LR Parser works using non-backtracking shift reduce technique yet it is efficient one.

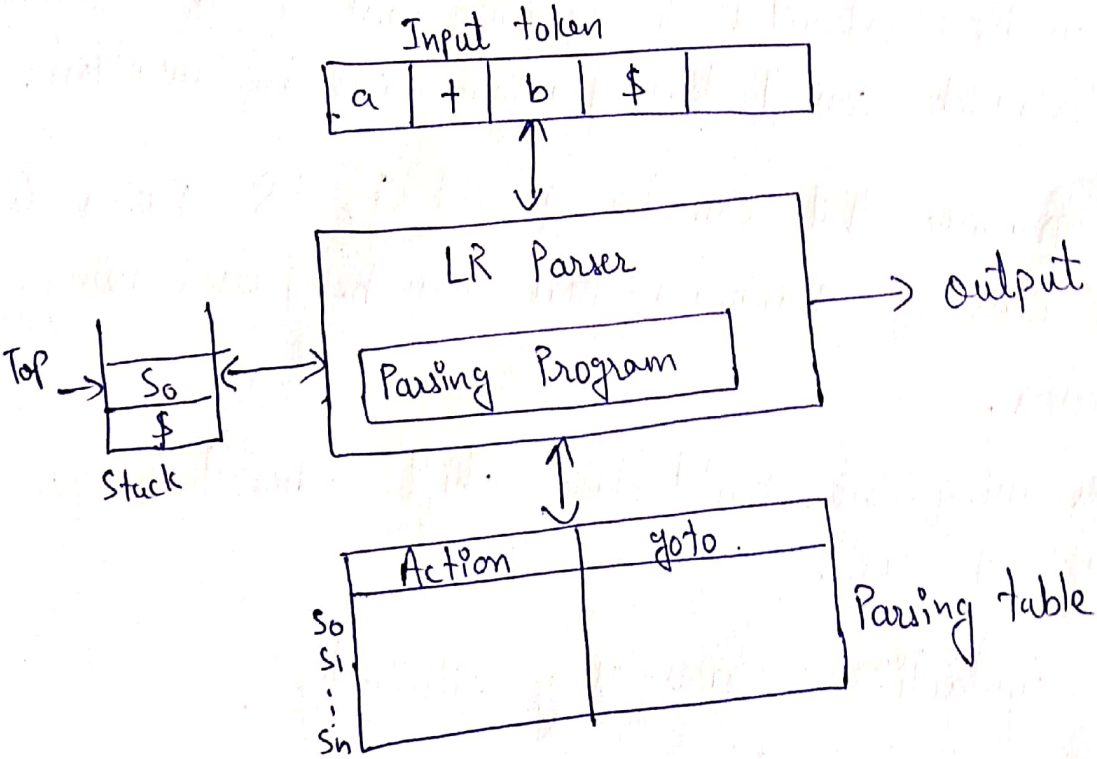
LR Parsers detect syntactical errors very efficiently.

* Structure of LR Parser.

LR Parser consists of input buffer for storing the input string, a stack for storing the grammar symbols, output & a Parsing table which has two parts : action & goto. There is one Parsing Program which reads the input symbol one at a time from

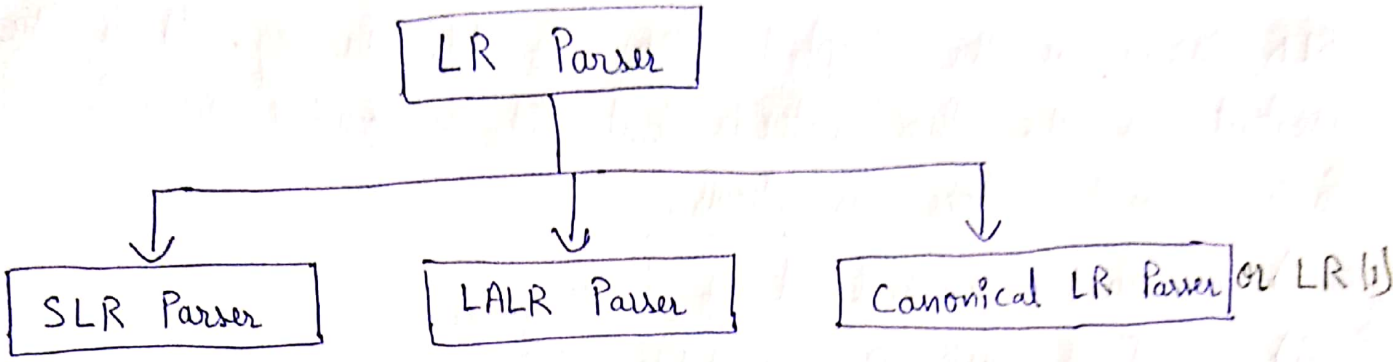
the input buffer. This Parsing Program works as follows:

- It initializes the stack with start symbol & invokes Scanner to get next token.
- It determines `s_j' the state currently on the top of the stack & a_i the current input symbol.
- It consults the Parsing table for the actions $[s_j, a_i]$ which can have one of the four values:
  - → s_j means shift state `j'.
  - → r_j " reduce by rule `j'.
  - → accept means successful Parsing is done
  - → error indicates syntactical error.

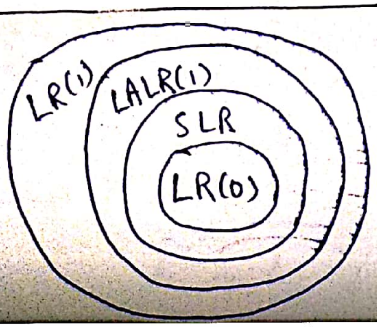


Structure Of LR Parser

*Types of LR Parsers:



SLR Parser	LALR Parser	Canonical LR Parser (CLR)
(1) SLR Parser is the smallest in size.	(1) LALR & SLR have the same size.	(1) LR Parser & CLR Parser is largest in size.
(2) It is an easiest method based on FOLLOW function.	(2) This method is applicable to wider class than SLR.	(2) This method is most Powerful than SLR & LALR.
(3) Error detection is not immediate in SLR.	(3) Error detection is not immediate in LALR.	(3) Immediate error detection is done by LR Parser.
(4) It requires less time & space complexity.	(4) The time & space complexity is more in LALR but efficient methods exist for constructing LALR Parsers directly.	(5) The time & space complexity is more for canonical LR Parser.



Powers: SLR(1) ≤ LALR(1) ≤ CLR(1) or LR(1).

Size: (SLR = LALR) < CLR.

* (a) SLR Parser (or) Simple LR

SLR Parser is the simplest form of LR Parsing. It is the weakest of the three methods but it is easiest to implement. Parsing can be done as follows:

i) Construct canonical set of items
ii) Construct SLR Parsing table
iii) Parsing of input string.

→ A grammar for which SLR Parser can be constructed is called SLR grammar.

* Augmented grammar: If a gramar 'G' is having start symbol 'S' then augmented grammar is a new grammar G' in which S' is a new start symbol such that S' → S. (Let S → ∈ be a Production) The Purpose of this grammar is to indicate the acceptance of input.

* Kernel items: It is collection of items S' → .S & all the items whose dots are not at the leftmost end of RHS of the rule.

* Non-Kernel items: The collection of all the items in which . are at the left end of RHS of the rule.

* Closure operation: For a CFG 'G', if 'I' is the set of items then the function closure(I) can be constructed using following rules:

- Consider 'I' is a set of canonical items & initially every item 'I' is added to closure(I).
- If rule A → α.Bβ is a rule in closure(I) & there is another rule for 'B' such as B → γ then

closure (I) : $A \rightarrow \alpha . B\beta$
$B \rightarrow .\gamma$

This rule has to be applied until no more items can be added to closure (I).

*goto operation:

If there is a production $A \rightarrow \alpha . B\beta$ then goto $(A \rightarrow \alpha . B\beta, B) = A \rightarrow \alpha B . \beta$. That means simply shifting of . one position ahead over the grammar symbol (may be terminal or non-terminal). The rule $A \rightarrow \alpha . B\beta$ is in 'I' then the same goto function can be written as goto (I, B).

(i) Construction of Canonical set of items:

- For the grammar 'G' initially add $S' \rightarrow .S$ in the set of item 'C'.
- For each set of items I_i in 'C' & for each grammar symbol X (may be terminal or non-terminal) add closure (I_i, X)

This process should be repeated by applying goto (I_i, x) for each 'x' in I_i such that goto (I_i, x) is not empty & not in 'C'. The set of items has to constructed until no more set of items can be added to 'C'.

(ii) Construction of SLR Parsing table

By considering basic parsing actions such as shift, reduce, accept & error we will fill up the action table. The goto table can be filled up using goto function.

Input : An Augmented grammar G'.

Output : SLR Parsing-table.

Algorithm :

- Initially construct set of items $C = \{I_0, I_1, I_2, \ldots I_n\}$ where 'C' is a collection of set of LR(0) items for the input grammar G'.
- The parsing actions are based on each Item I_i. The actions are as given below :
  - → If $A \to \alpha . a\beta$ is in I_i & goto $(I_i, a) = I_j$ then set of action $[i, a]$ as "shift j". Note that 'a' must be a terminal symbol.
  - → If there is a rule $A \to \alpha .$ in I_i then set actim $[i, a]$ to "reduce $A \to \alpha$" for all symbols 'a', where $a \in$ FOLLOW (A). Note that 'A' must not be an augmented grammar s'.
  - → If $S' \to S$ is in I_i then the entry in the action table action $[i, \$]$ = "accept".
- The goto part of the SLR table can be filled as : The goto transitions for state 'i' is considered for non-terminals only. If goto $(I_i, A) = I_j$ then goto $[I_i, A] = j$.
- All the entries not defined by rule 2 & 3 are considered to be "error".

iii) Parsing the Input using Parsing Table:

Input: The input string 'w' that is to be passed & Parsing table.

Output: Parse 'w' if $w \in L(G)$ using bottom up. If $w \notin L(G)$ then report syntactical error.

Algorithm:

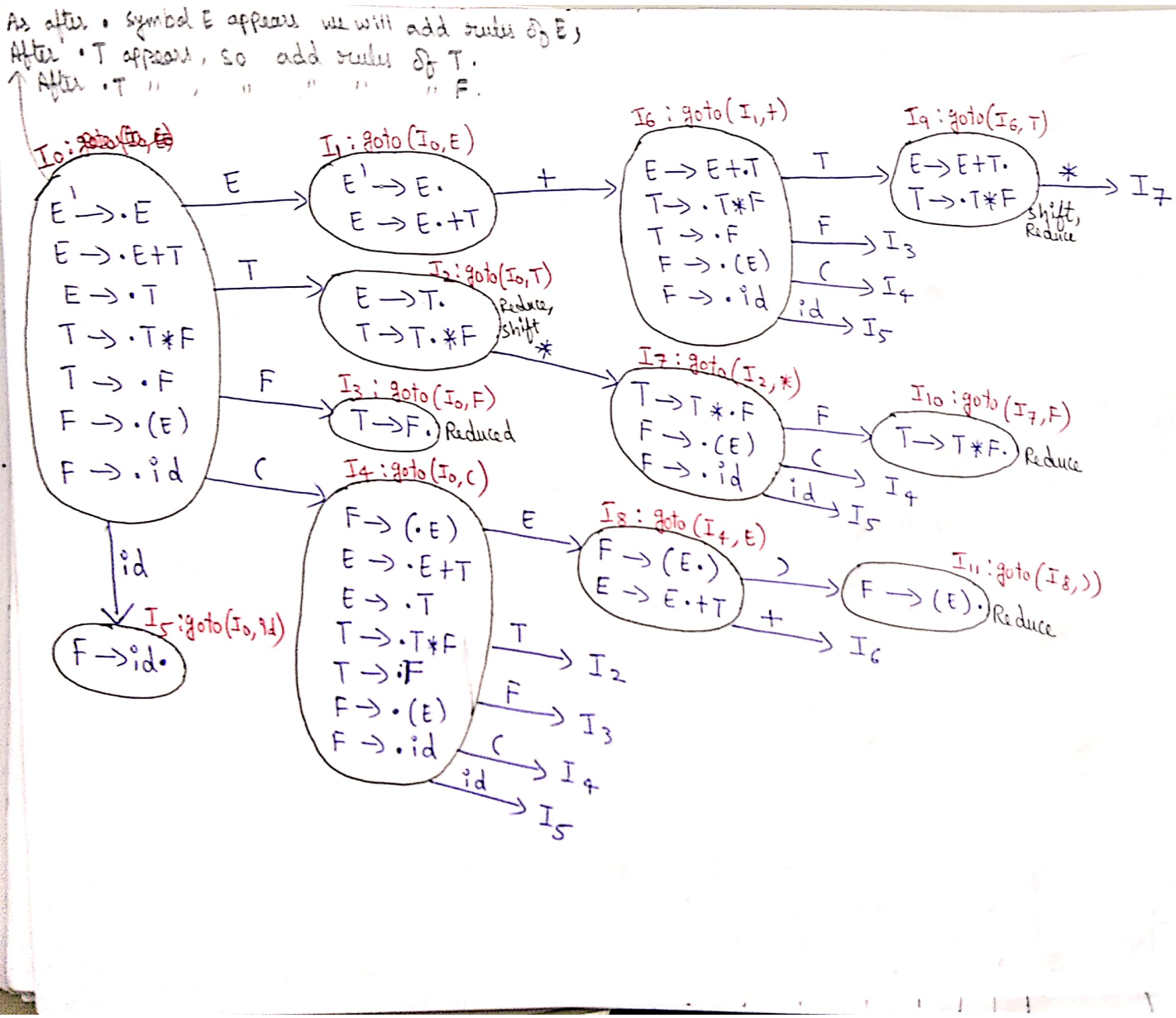
- Initially Push '0' as initial state onto the stack & place the input string with $ as end marker on the input tape.
- If 's' is the state on the top of the stack & 'a' is the symbol from input buffer pointed by a lookahead Pointer then
  - → If action [s,a] = shift 'j' then push 'a', then push 'j' onto the stack. Advance the input lookahead pointer.
  - → If action [s,a] = reduce $A \to \beta$ then pop $2*|\beta|$ symbols If 'i' is on the top of the stack then push 'A', then Push goto [i,A] on the top of the stack.
  - → If action [s,a] = accept then halt the Parsing Process. If indicates the successful Parsing.

* Problems:

(1) Construct the SLR parsing table for the given grammar. Also Parse the Input id*id+id. $E \to E+T \mid T$, $T \to T*F \mid F$, $F \to (E) \mid id$.

Sol: Let us number the Production rules in the grammar.

1. $E \to E+T$
2. $E \to T$
3. $T \to T*F$
4. $T \to F$
5. $F \to (E)$
6. $F \to id$



(5)

- First(E) = First(T) = First(F) = {(, id} ✓ (here first & Follow do not do left recursion, left factoring)
- ~~First(E')~~ = ~~{+, ε}~~
- ~~First(T')~~ = ~~{*, ε}~~
- Follow(E) = {+, $,)} ✓
- ~~Follow(E')~~ = ~~{$,)}~~
- Follow(T) = {+, *, $,)} ✓
- ~~Follow(T')~~ = ~~{+, $,)}~~
- Follow(F) = {*, +, $,)} ✓

* SLR Parsing Table:

state	Action						goto		
	id	+	*	(	)	$	E	T	F
0	S_5			S_4			1	2	3
1		S_6				Accept			
2		r_2	S_7		r_2	r_2			
3		r_4	r_4		r_4	r_4			
4	S_5			S_4			8	2	3
5		r_6	r_6		r_6	r_6			
6	S_5			S_4				9	3
7	S_5			S_4					10
8		S_6			S_{11}				
9		r_1	S_7		r_1	r_1			
10		r_3	r_3		r_3	r_3			
11		r_5	r_5		r_5	r_5			

r_1 : E → E+T
r_2 : E → T
r_3 : T → T*F
r_4 : T → F
r_5 : F → (E)
r_6 : F → id

→ For Item 2, E→T. is reduced. check the list. It is numbered r_2. Now Follow(E) = ($,), +) in $,) columns write r_2.

→ If it is shift operation just write S_{no}.
If: Item 0; on id it is going to I_5. So write "S_5".

* Parsing input "id * id + id".

Stack	Input	Action	output
$0	id * id + id $	shift s_5	whenever we do Reduce we need to pop out 2 symbols.
$0 id5	* id + id $	Reduce by F → id	
$0 F_3	* id + id $	Reduce by T → F	
$0 T_2	* id + id $	shift s_7	→ when you are shifting no need to pop.
$0 T_2 * 7	id + id $	shift s_5	
$0 T_2 * 7 id5	+ id $	Reduce by F → id	
$0 T_2 * 7 F10	+ id $	Reduce by T → T * F	∴ since there are no conflicts in SLR table. The given grammar is SLR(1).
$0 T_2	+ id $	Reduce by E → T	
$0 E1	+ id $	shift s_6	
$0 E1 + 6	id $	shift s_5	
$0 E1 + 6 id5	$	Reduce by F → id	
$0 E1 + 6 F3	$	Reduce by T → F	
$0 E1 + 6 T9	$	Reduce by E → E + T	
$0 E1	$	accept	

(2) Show that the following grammar is not SLR(1).

S → Aa | bAc | dc | bda

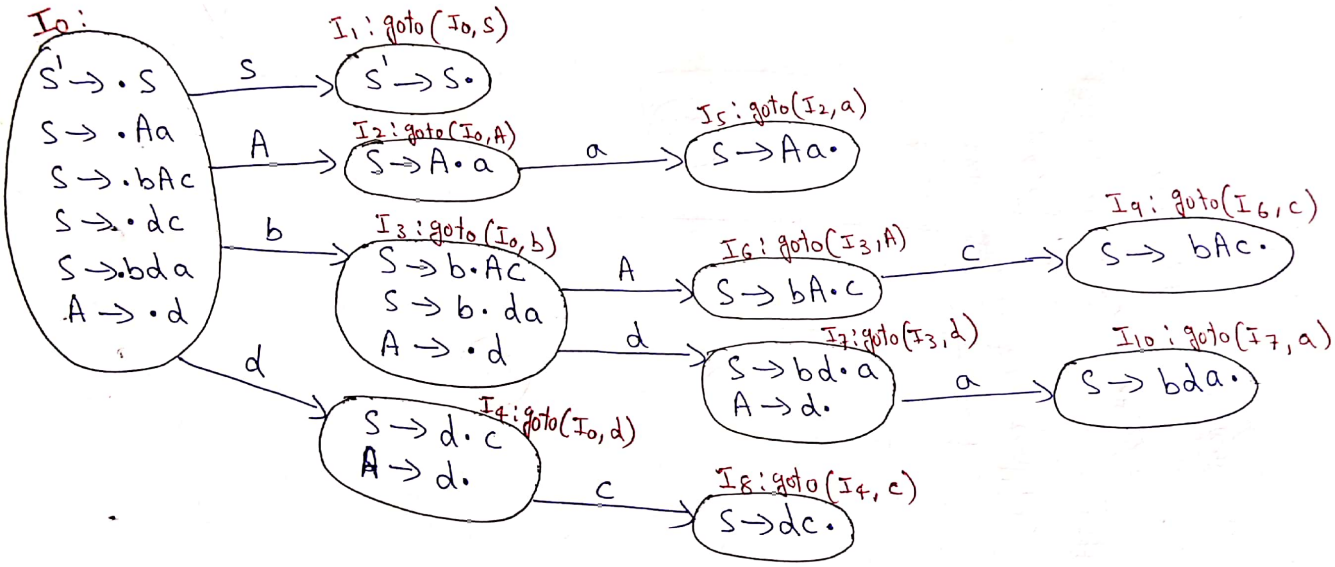
A → d.

Sol:- Let us number the Production rules in the grammar.

1. S → Aa
2. S → bAc
3. S → dc
4. S → bda
5. A → d

CD Unit-3

(6)



- Follow (S) = {$}
- Follow (A) = {a, c}

* SLR(1) Parsing Table :

	Action					Goto	
state	a	b	c	d	$	S	A
0		S_3		S_4		1	2
1					Accept		
2	S_5						
3				S_7			6
4	r_5		S_8 / r_5				
5					r_1		
6			S_9				
7	S_{10} / r_5		r_5				
8					r_3		
9					r_2		
10					r_4		

Since the above table has multiple entries in Action [7, a] and Action [4, c], this means shift/reduce conflict will occur while Parsing the input string.

∴ The given grammar is not SLR(1). //

* More Powerful Parsers:

(a) Canonical LR : (b) LR(1) Parser

The canonical set of items is the parsing technique in which a lookahead symbol is generated while constructing set of items. Hence the collection of set of items is referred as LR(1). The value 1 in the bracket indicates that there is one lookahead symbol in the set of items.

→ Steps to conitruct canonical LR:

- Construction of canonical set of items along with the lookahead.
- Building canonical LR Parsing table
- Parsing the input string using canonical LR Parsing table

* Construction of canonical set of items along with lookahead:

(1) For the grammar 'G' initially add $S' \rightarrow \cdot S$ in the set of item 'C'.

(2) For each set of items I_i in 'C' & for each grammar Symbol X (may be terminal or non-terminal) add closure(I_i, X). This process should be repeated by applying goto(I_i, X) for each X in I_i such that goto(I_i, X) is not empty & not in 'C'. The set of items has to be constructed until no more set of items can be added to 'C'.

(3) The closure function can be computed as follows:
For each item $A \rightarrow \alpha \cdot X\beta a$ and rule $X \rightarrow \gamma$ & $b \in FIRST(\beta a)$ such that $X \rightarrow \cdot\gamma$ & 'b' is not in I

then add $X \to .\gamma, b$ to I.

(4) Similarly the goto function can be computed as: For each item $[A \to \alpha . X\beta, a]$ is in I & rule $[A \to \alpha X . \beta, a]$ is not in I ~~then add~~ ~~$X \to .\gamma, b$~~ goto items then add $[A \to \alpha X . \beta, a]$ to goto items.

This process is repeated until no more set of items can be added to the collection 'C'.

* Construction of canonical LR Parsing table:

(1) Initially construct set of items $C = \{I_0, I_1, I_2, \ldots I_n\}$ where 'C' is a collection of set of LR(1) items for the input grammar G'.

(2) The parsing actions are based on each item I_i. The actions are as given below:

- If $[A \to \alpha . a\beta, b]$ is in I_i and goto $(I_i, a) = I_j$ then Create an entry in the action table action $[I_i, a]$ = shift j.
- If there is a Production $[A \to \alpha ., a]$ in I_i then in the action table action $[I_i, a]$ = reduce by $A \to \alpha$. Here 'A' should not be S'.
- If there is a Production $S' \to S., \$$ in I_i then action $[i, \$]$ = accept.

(3) The goto Part of the LR table can be filled as: The goto transitions for state 'i' is considered for non-terminals only. If goto $(I_i, A) = I_j$ then goto $[I_i, A] = j$.

(4) ~~All the entries not defined by rule 2 and 3 are considered to be "error"~~ All entries not defined by rule 2 & 3 are considered to be "error"

* Problems :

(1). Construct LR(1) or CLR set of items for the below grammar

$S \rightarrow CC$

$C \rightarrow aC \mid d$

Sol : Add $S' \rightarrow .S, \$$ as the first rule in I_0. The grammar is :

$S' \rightarrow .S, \$$

$S \rightarrow CC$

$C \rightarrow aC \mid d$

Add dot to every Production : $(A \rightarrow \alpha . X\beta, a)$, $A \in S'$, $\alpha \in \epsilon$, $X \in S$, $a \in \$$

I_0

$S' \rightarrow .S, \$$

$S \rightarrow .CC, \$$

$C \rightarrow .aC, a/d$

$C \rightarrow .d, a/d$

$X \rightarrow \gamma, b$ then add $X \rightarrow .\gamma, b$

$b \in First(\beta a)$

$b \in First(\epsilon \$)$

$b \in First(\$)$

$b = \{\$\}$

→ If there is a Production $X \rightarrow \gamma, b$ then add $X \rightarrow .\gamma, b$

$C \rightarrow .aC$

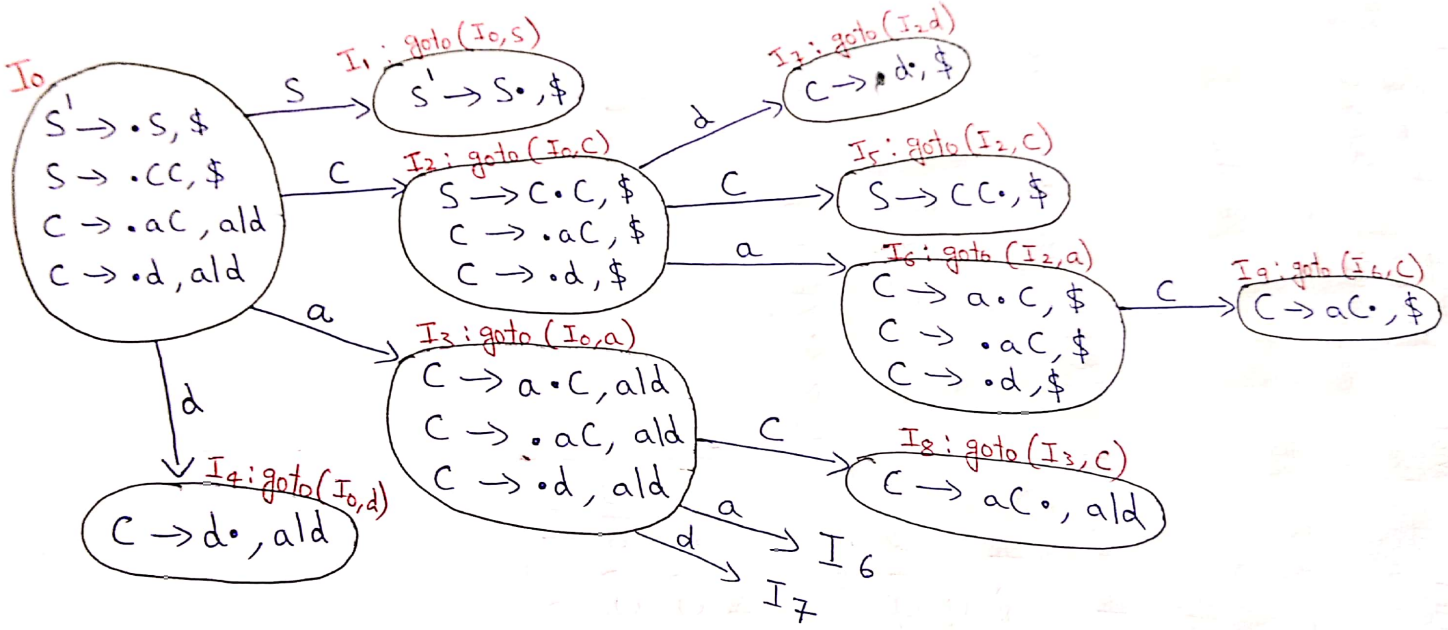
$C \rightarrow .d$

$b \in First(\beta a)$

$b \in First(C\$)$

$b \in First(C)$ as $First(C) = \{a, d\}$

$b = \{a, d\}$



* LR(1) Parsing table :-

	action			goto	
	a	d	$	S	C
0	S_3	S_4		1	2
1			Accept		
2	S_6	S_7			5
3	S_3	S_4			8
4	r_3	r_3			
5			r_1		
6	S_6	S_7			
7			r_3		
8	r_2	r_2			
9			r_2		

r_1 : S → CC

r_2 : C → aC

r_3 : C → d

I_0 :- on S it is shifted to I_1
C " " " " I_2
a " " " " I_3
d " " " " I_4

} write S no in appropriate reduce.

I_4 :- reduced check numbering : "r_3".

C → d·, (a/d) ⇒ write in a/d "r_3".

* Parsing the input using LR(1) Parsing table

Let us Parse for the i/p string "aadd" by using above Parsing table.

stack	Input buffer	action table	goto table	Passing action
$0	aadd$	action [0,a] = S_3		
$0a3	add $	action [3,a] = S_3		shift
$0a3a3	dd$	action [3,d] = S_4		shift
$0a3a3d4	d$	action [4,d] = r_3	[3,C] = 8	Reduce by C→d
$0a3a3C8	d$	action [8,d] = r_2	[3,C] = 8	Reduce by C→aC
$0a3C8	d$	action [8,d] = r_2	[0,C] = 2	Reduce by C→aC
$0C2	d$	action [2,d] = S7		shift
$0C2d7	$	action [7,$] = r_3	[2,C] = 5	Reduce by C→d
$0C2C5	$	action [5,$] = r_1	[0,S] = 1	Reduce by S→CC
$0S1	$	accept		

Thus the given input string is successfully Parsed using LR Parser or canonical LR Parser.

(2) show that the following grammar is LR(1).

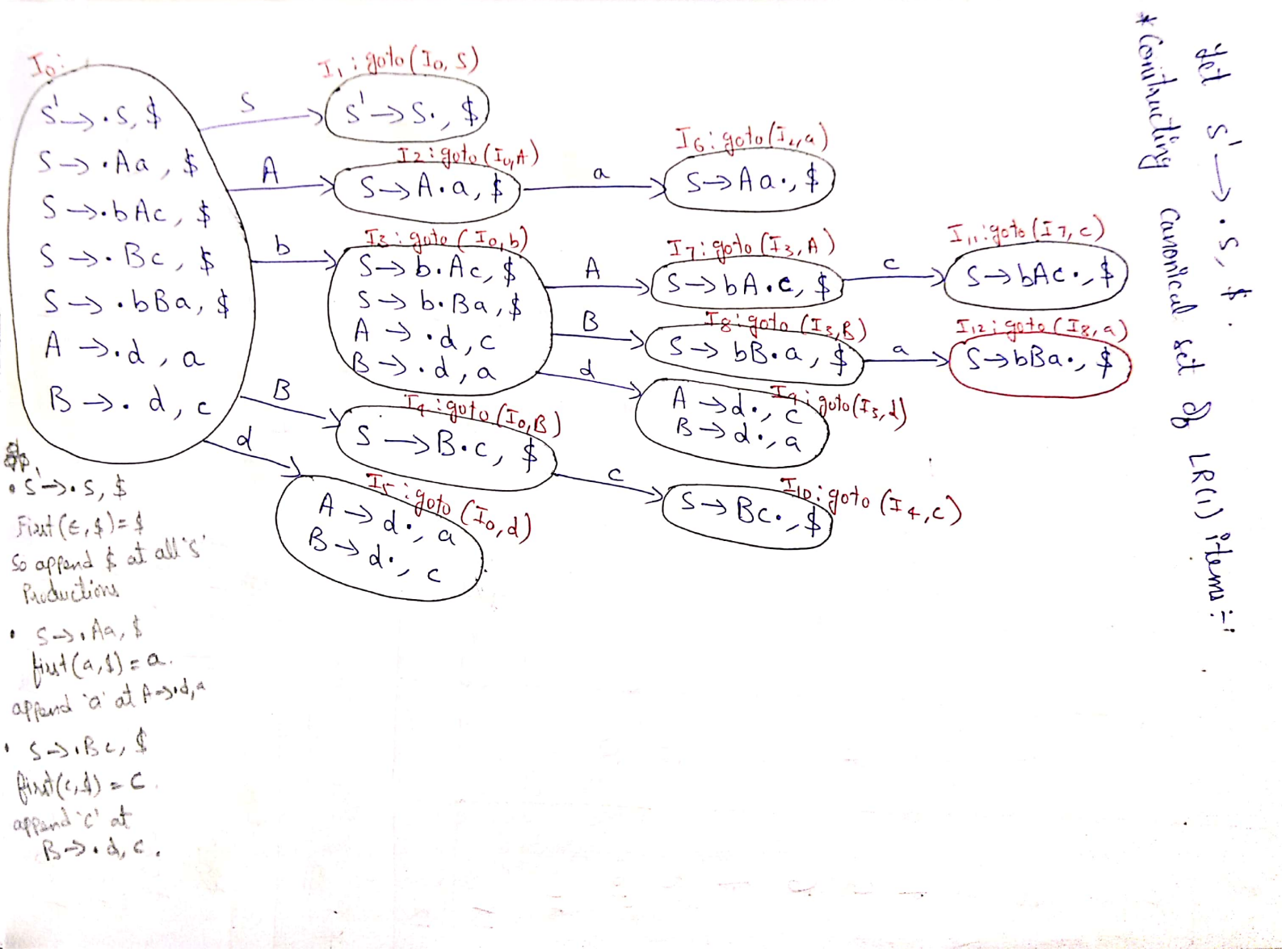
S → Aa | bAc | Bc | bBa

A → d

B → d.

Sol: We will number out the Production rules in given grammar.

1. S → Aa
2. S → bAc
3. S → Bc
4. S → bBa
5. A → d
6. B → d



* Parsing table :

State	Action					goto		
	a	b	c	d	$	S	A	B
0		S_3		S_5		1	2	4
1					Accpt			
2	S_6							
3				S_9			7	8
4			S_{10}					
5	r_5		r_6					
6					r_1			
7			S_{11}					
8	S_{12}							
9	r_6		r_5					
10					r_3			
11					r_2			
12					r_4			

* Parsing the string "bda" using above Parsing table of LR(1) :

stack	Input Buffer	Action
$0	bda$	shift 3
$0b3	da$	shift 9
$0b3d9	a$	Reduce by B → d
$0b3B8	a$	shift 12

Stack	Input Buffer	Action
$ 0b3B8a12	$	Reduce by S → bBa
$ 0S1	$	Accept

The above string 'bda' is accepted. Since in the Parsing table we do not have multiple entries in the same cell, the above grammar is LR(1) or CLR(1).

* LALR :-

The algorithm for construction of LALR Parsing table is as follows:-

step 1 :- Construct the LR(1) set of items.

step 2 :- Merge the two states I_i and I_j if the first component (i.e the production rules with dots) are matching & create a new state replacing one of the older state such as $I_{ij} = I_i \cup I_j$.

step 3 :- The Parsing actions are based on each item I_i. The actions are as given below:

a) If $[A \to \alpha . a\beta, b]$ is in I_i & goto $(I_i, a) = I_j$ then create an entry in the action table action $[I, a]$ = shift 'j'.

b) If there is a production $[A \to \alpha \cdot, a]$ in I_i then in the action table action $[I_i, a]$ = reduce by $A \to \alpha$. Here 'A' should not be S'.

c) If there is a production $S' \to S \cdot, \$$ in I_i then action $[i, \$]$ = accept.

step4 : The goto part of the LR table can be filled as: The goto transitions for state 'i' is considered for non-terminals only. If goto $(I_i, A) = I_j$ then goto $[I_i, A] = j$.

step5 : If the parsing action conflict then the algorithm fails to produce LALR Parser & grammar is not LALR(1). All the entries not defined by rule 2 & 4 are considered to be "error".

* **Problems** :

(1) Construct Parsing table for LALR(1) Parser for the below given grammar.

$S \to CC$

$C \to aC$

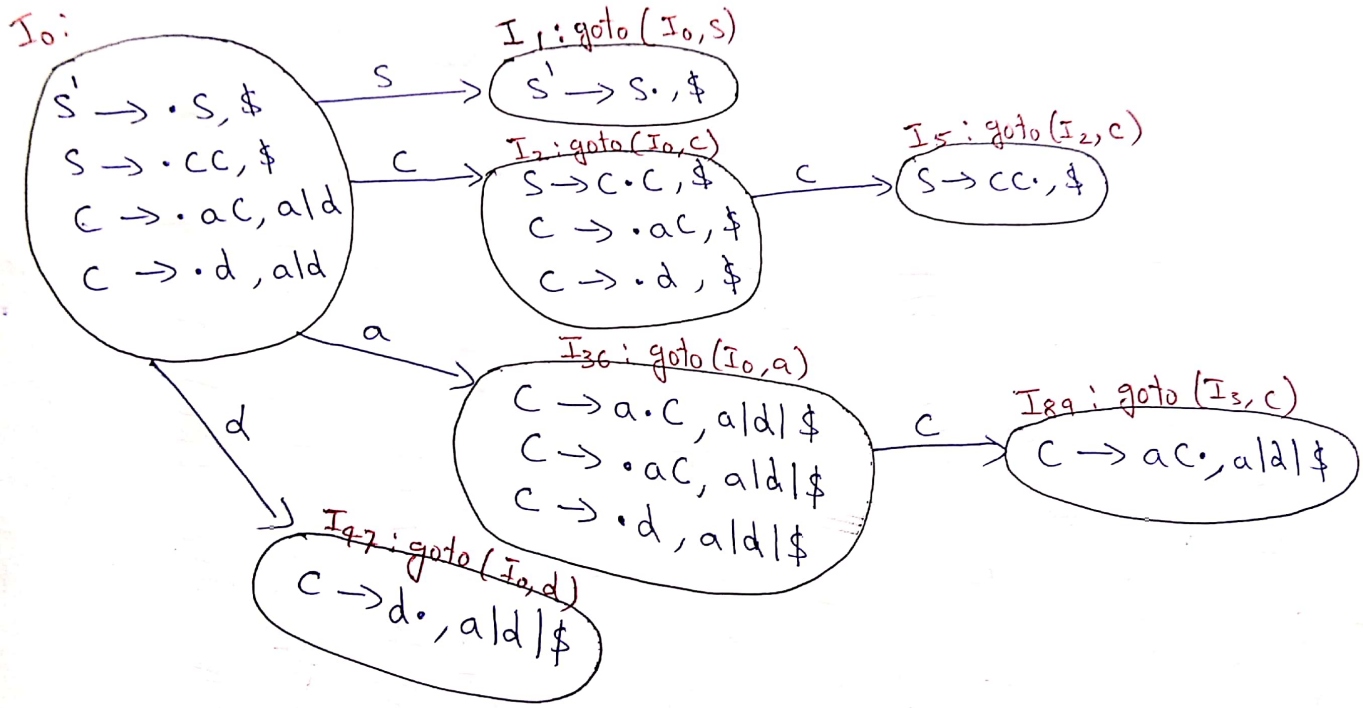
$C \to d$.

sol : First the set LR(1) items can be constructed as follows with merged states.

I_1. $S \to CC$

r_2. $C \to aC$

r_3. $C \to d$.



Now consider state I_0 there is a match with the rule $[A \to \alpha . a\beta, b]$ and goto $(I_i, a) = I_j$

$C \to .aC, a/d/\$$ and if the goto is applied on 'a' then we get the state I_{36}. Hence,

in I_0:

$C \to .d$, a/d

$[A \to \alpha . a\beta, b]$

where $A = C$, $\alpha = \epsilon$, $a = d$, $\beta = \epsilon$, $b = a/d$

goto $(I_0, d) = I_{47}$.

hence action $[0, d]$ = shift 47.

→ For state I_{47} : $C \to d.$, $a/d/\$$

$A \to \alpha .$, a

where $A = C$, $\alpha = d$, $a = a/d/\$$.

action $[47, a]$ = reduce by $C \to d$ i.e rule 3

action $[47, d]$ = " " " " "

action $[47, \$]$ = " " " " "

→ $S' \to S.$, $\$$ in I_1, so we will create action $[1, \$]$ = accept.

The goto table can be filled by using the goto functions.

Instance goto $(I_0, S) = I_1$, so goto $[0, S] = 1$. continuing in this manner we can fill the LR(1) Parsing table as follows:-

State	Action			goto	
	a	d	$	S	C
0	S_{36}	S_{47}		1	2
1			Accept		
2	S_{36}	S_{47}			5
36	S_{36}	S_{47}			89
47	r_3	r_3	r_3		
5			r_1		
89	r_2	r_2	r_2		

* Parsing string "adad" using LALR Parser :

Stack	Input buffer	Action table	goto table	Parsing action
$0	aadd$	action [0,a] = S36		
$0a36	add$	action [36,a] = S36		shift
$0a36a36	dd$	action [36,d] = S47		shift
$0a36a36d47	d$	action [47,d] = r36	[36,C] = 89	Reduce by C→d
$0a36a36C89	d$	action [89,d] = r2	[36,C] = 89	Reduce by C→aC
$0a36C89	d$	action [89,d] = r2	[0,C] = 2	Reduce by C→ac
$0C2	d$	action [2,d] = S47		shift
$0C2d47	$	action [47,$] = r36	[2,C] = 5	Reduce by C→d
$0C2C5	$	action [5,$] = r1	[0,S] = 1	Reduce by S→CC
$0S1	$	accept		

(2) Show that the following grammar is not LALR(1).

S → Aa | bAc | Bc | bBa

A → d

B → d

Sol:-

Let us number the production rules of the grammar:

γ_1 : S → Aa

γ_2 : S → bAc

γ_3 : S → Bc

γ_4 : S → bBa

γ_5 : A → d

γ_6 : B → d

LALR(1) set of items are as follow

State	Action					GOTO		
	a	b	c	d	$	S	A	B
0		S_3		S_5		1	2	4
1					Accept			
2	S_6							
3				S_9			7	8
4			S_{10}					
59	γ_5 γ_6		γ_5 γ_6					
6					γ_1			
7			S_{11}					
8	S_{12}							
10					γ_3			
11					γ_2			
12					γ_4			

The Parsing table shows multiple entries in Action [59, a] & Action [59, c]. This is called reduce/reduce conflict. Because of this conflict we cannot parse input.

∴ The given grammar is not LALR(1).

* Handling Ambiguous Grammar :

If all the grammar is ambiguous then it creates the conflicts & we cannot parse the input string with such ambiguous grammar. But for some languages in which arithmetic

expressions are given ambiguous grammar are most compact & provide more natural specification as compared to equivalent unambiguous grammar.

→ while using ambiguous grammar for parsing the input string we will use all the disambiguating rules so that each time only one parse tree will be generated for that specific input. Thus ambiguous grammar can be used in controlled manner for parsing the input.

* Using Precedence & Associativity to resolve Parsing Action Conflicts

Consider an ambiguous grammar:

$E \rightarrow E+E$

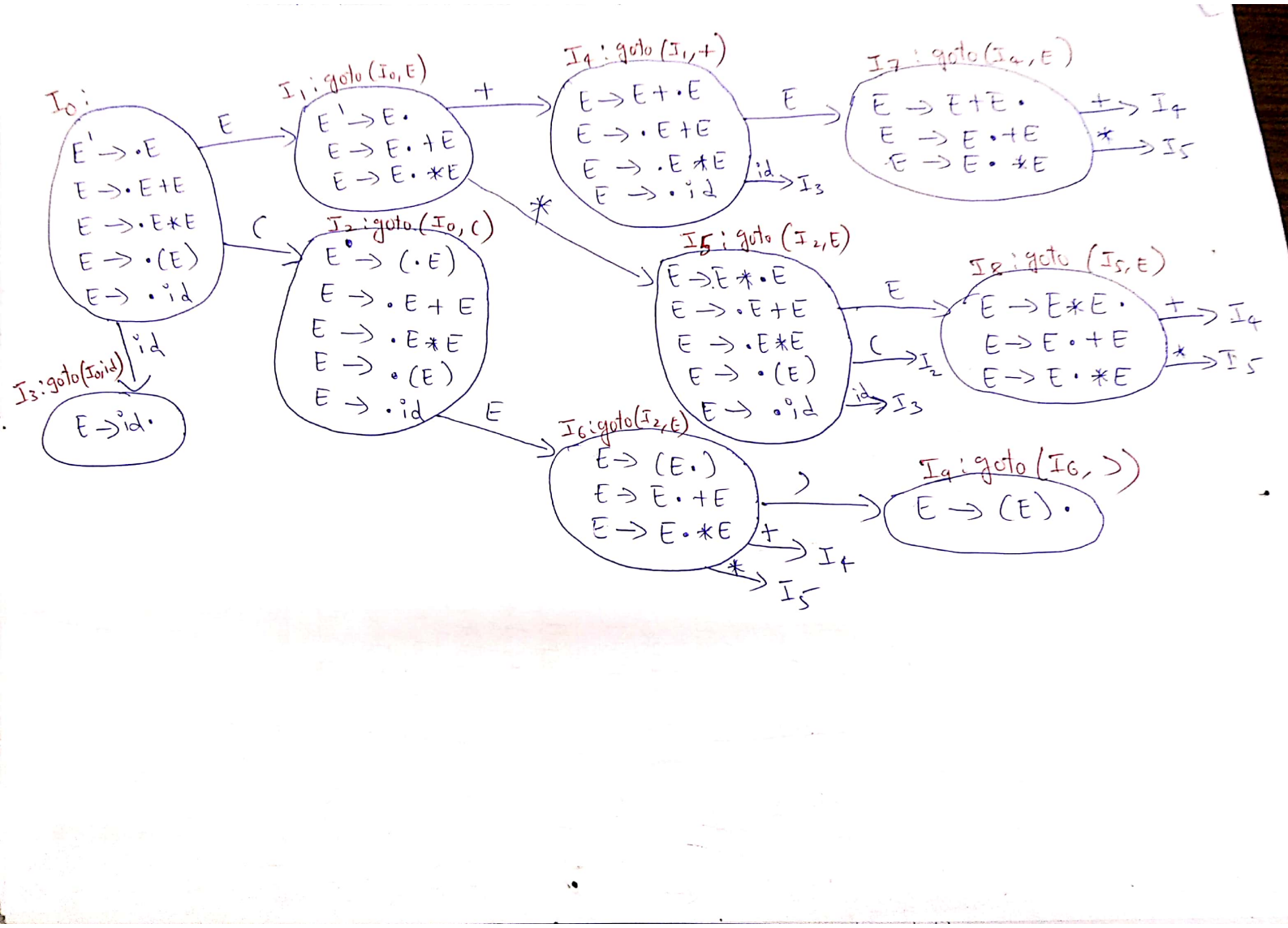
$E \rightarrow E*E$

$E \rightarrow (E)$

$E \rightarrow id$

Now we will build the set of LR(0) items for this grammar.

- Follow(E) = {+, *,), $}

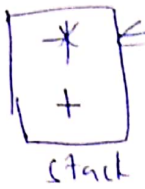


state	action						(goto) E
	id	+	*	(	)	$	E
0	S_3			S_2			1
1		S_4	S_5			Accept	
2	S_3			S_2			6
3		r_4	r_4		r_4	r_4	
4	S_3			S_2			7
5	S_3			S_2			8
6		S_4	S_5		S_9		
7		S_4 or r_1	S_5 or r_1		r_1	r_1	
8		S_4 or r_2	S_5 or r_2		r_2	r_2	
9			r_3		r_3	r_3	

The shift/reduce conflicts occur at state 7 & 8. Let us consider id + id * id.

Stack	Input	Action with conflict resolution
$0	id+id*id$	shift
$0id3	+id*id$	Reduce by E→id
$0E1	+id*id$	shift
$0E1+4	id*id$	shift
$0E1+4id3	*id$	Reduce by E→id
$0E1+4E7	*id$	The conflict can be resolved by shift 5
$0E1+4E7*5	id$	shift
$0E1+4E7*5id3	$	Reduce by E→id
$0E1+4E7*5E8	$	Reduce by E→E*E
$0E1+4E7	$	Reduce by E→E+E
$0E1	$	Accept

eg: * has precedence over '+' we have to perform multiplication operation first. And for that it is necessary to push * on the top of the stack. The stack position will be:



By this we can perform E*E first & then E+E. The Parsing conflict can be resolved by assigning shift operation. Hence action[7,*] = S_5.

state	Action						goto
	id	+	*	(	)	$	E
0	S_3			S_2			1
1		S_4	S_5			Accept	
2	S_3			S_2			
3		r_4	r_4		r_4	r_4	
4	S_3			S_2			
5	S_3			S_2			
6		S_4	S_5		S_9		
7		r_1	S_5		r_1	r_1	
8		r_2	r_2		r_2	r_2	
9			r_3		r_3	r_3	

* Using Dangling Else Ambiguity :

Consider the grammar :

$S \rightarrow iSeS \mid iS \mid a$

where iSeS = if expression then statement else statement

iS = if " "

a = all other productions.

Let the grammar be:

$S' \rightarrow S$

$S \rightarrow iSeS \mid iS \mid a$

Let us construct LR(0) set of items as:

First(S) = {i, a}

Follow(S) = {e, $}

State	Action				GOTO
	i	e	a	$	S
0	S_2		S_3		1
1				Accept	
2	S_2		S_3		4
3		r_3		r_3	
4		S_5 or r_2		r_2	
5	S_2		S_3		6
6		r_1		r_2	

Action [5,e) has shift/reduce conflict.

Consider the input "iiaea$" for Procesing.

Stack	Input	Action with conflict resolution
$0	iiaea$	shift
$0i2	iaea$	shift
$0i2i2	aea$	shift
$0i2i2a3	ea$	Reduce S → a
$0i2i2S4	ea$	From the conflict we have chosen Reduce S → iS
$0i2S4	ea$	Reduce S → iS
$0S1	a$	Error!!

That means the choice of r_2 in action [4,e] is not valid. Hence we will try it by choosing the shift action.

Stack	Input	Action with conflict resolution
$0	iiaea$	shift
$0i2	iaea$	shift
$0i2i2	aea$	shift
$0i2i2a3	ea$	Reduce S→a
$0i2i2 S4	ea$	From the conflict we have chosen Reduce S→iS
$0i2i2 S4e5	a$	shift
$0i2i2S4e5a3	$	Reduce S→a
$0i2i2S4e5 S6.	$	Reduce S→ iSeS
$0i2S4	$	Reduce S→iS
$0S1	$	Accept.

Logically also we should choose the shift operation as by shifting the else we can associate it with previous "if expression then statement". Therefore shift / reduce conflict is resolved in favour of shift.

→ After resolving the conflict we get Parsing table for dangling else problem as:

state	Action				goto
	i	e	a	$	S
0	S2		S3		1
1				Accept	
2	S2		S3		4
3		r3		r3	
4		S5		r2	
5	S2		S3		6
6		r1		r2	

* Error Recovery in LR Parser :

The LR parser is a table driven parsing method in which the blank entries are treated as errors. When we compile any Program we first get the syntactical errors. These errors are usually denoted by user friendly error messages.

*Example :

$E \to E+E$
$E \to E*E$
$E \to (E)$
$E \to id$.

(we have already constructed LR(0) items for this grammar in handling Ambiguous grammar topic).

state	Action						goto
	id	+	*	(	)	$	E
0	S_3			S_2			1
1		S_4	S_5			Acept	
2	S_3			S_2			6
3		r_4	r_4		r_4	r_4	
4	S_3			S_2			7
5	S_3			S_2			8
6		S_4	S_5		S_9		
7		r_1	S_5		r_1	r_1	
8		r_2	r_2		r_2	r_2	
9			r_3		r_3	r_3	

Following are the rules to fill the table:

(1) If there are entries for r_j in some particular state then fill up all the blank entries r_j only.

(2) If there are shift entries only in some particular state then do not replace blank entries. Keep them as it is.

Using the error detection & recovery Procedure we have replaced some blank (error) entries by particular reduction rules. This change means we are postponing the error detection until one or more reductions are done & error will be introduced before any shift move take place.

Consider the set of items generated for obtaining the error messages.

$I_0 : E' \rightarrow \cdot E$

It means that there is no symbol before the dot. And being the initial state the stack is empty. In such a case if + @1 * @2 $ comes in the input string then we say that operand is missing for these operator.

Stack	Input	Error could be
$	+	Missing operand
$	*	missing operand
$	$	Missing operand
$	)	Unbalanced right Parenthesis

I_1 : goto (I_0, E)
$E' \rightarrow E \cdot$

If we ultimately reduce $E \rightarrow id$ then id = missing operator, id (= missing operator, id) = unbalanced right Parenthesis.

Stack	Input	Error could be
$ ... id	id	Missing operator
$ ... id	(	" "
$ ... id	)	unbalanced right Parenthesis

I_6 : goto (I_2, E)
$E \rightarrow (E \cdot)$

If we eventually reduce $E \rightarrow id$ then the rule becomes $E \rightarrow (id \cdot)$. That means after id we expect ')' & if $ comes then the error will be missing right Parenthesis.

Stack	Input	Error could be
$ .... Cid	id	Missing operator
$ .... Cid	(	" "
$ ... Cid	$	Missing right Parenthesis

From all these situations, some error messages are:

1) E1 : These errors are in states I_0, I_2, I_4 & I_5. This indica-tes that the operand should appear before operator. Hence the error message will be "missing operand".

2) E2 : This error is in ')' column & from states I_0, I_1, I_2 I_4 & I_5 indicating unbalancing in right Parenthesis. Hence error message will be "unbalanced right Parenthesis".

3) E3 : The operator is expected in this case of error as it is from state I_1 or I_6. The error message will be "missing operator".

• E_4 : This error occurs at state 6 in the) column. The state 6 expects ')' Parenthesis at the end of expression. Hence the error message will be "missing right Parenthesis"

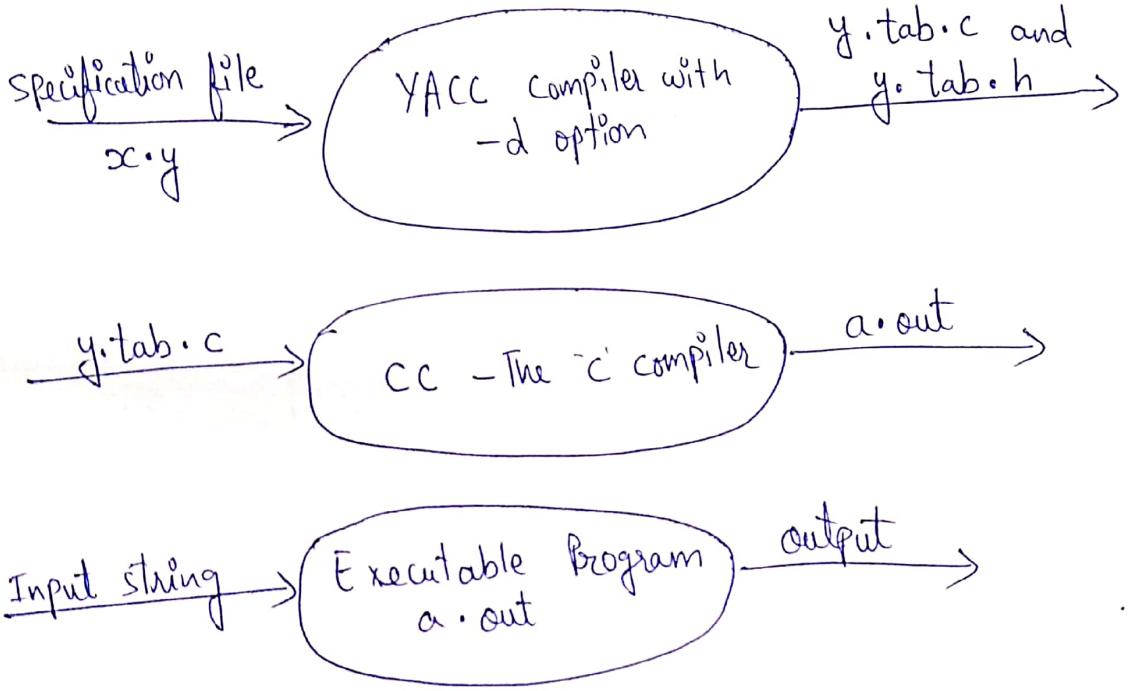
→ Modified table with included error messages :-

state	Action						GOTO
	id	+	*	(	)	$	E
0	S_3	E_1	E_1	S_2	E_2	E_1	1
1	E_3	S_4	S_5	E_3	E_2	Accept	
2	S_3	E_1	E_1	S_2	E_2	E_1	6
3	r_4	r_4	r_4	r_4	r_4	r_4	
4	S_3	E_1	E_1	S_2	E_2	E_1	7
5	S_3	E_1	E_1	S_2	E_2	E_1	8
6	E_3	S_4	S_5	E_3	S_9	E_4	
7	r_1	r_1	S_5	r_1	r_1	r_1	
8	r_2	r_2	r_2	r_2	r_2	r_2	
9	r_3	r_3	r_3	r_3	r_3	r_3	

While Parsing the input, LR Parser will refer the Parsing table if it detects the error entry then respective error message will be reported. This is how we get syntax errors when we compile our Program.

* YACC – Automatic Parser Generator :

YACC is an automatic tool for generating the Parser Program. YACC stands for "Yet Another Compiler Compiler", which is basically the utility available from Unix. Basically YACC is LALR Parser generator. The YACC can report conflicts or ambiguities in the form of error messages. LEX & YACC work together to analyse the program syntactically.



YACC : Parser generator model

YACC specification file need to be written first, for example "x.y" This file is given to the YACC compiler by Unix command as:

```
$ yacc x.y
```

Then it will generate a Parser Program using your YACC specification file. This Parser Program has a standard name as "y.tab.c". This is basically Parser Program in 'C'

generated automatically.

```
$ yacc -d xc.y
```

By -d option two files will get generated one is "y.tab.c" & other is "y.tab.h". The header file y.tab.h will store all the tokens & we need not have to create "y.tab.h" explicitly. The generated y.tab.c program will then be compiled by 'C' compiler & generates the executable a.out file. We can test YACC program with the help of some valid & invalid strings.

* YACC specification file :-

YACC specification file consists of three parts: (a) declaration section, (b) translation rule section & (c) supporting 'C' functions.

```
%{
   /* declaration section */        (a)
%}
%%
   /* Translation rule section */   (b)
%%
/* Required 'C' functions */        (c)
```

YACC specification file

(a) **Declaration section** :- In this section ordinary 'C' declarations can be written. We can also declare grammar tokens in this section. The declaration of tokens should be after %{ %}.

(b) Translation Rule Section: It consists of all the production rules of context free grammar with corresponding actions.

If:

Rule 1	action 1
Rule 2	action 2
⋮	
Rule n	Action n

→ If there are more than one alternatives to a single rule then those alternatives should be seperated by "|" character. The actions are typical 'C' statements.

If CFG is:

LHS → alternative 1 | alternative 2 | .... | alternative n

then:

```
LHS : alternative 1    {action 1}
    |      "      2    {  "   2}
    ⋮
    | alternative n     {action n}
```

(c) 'C' functions section: This section consists of one main function in which the routine yyparse() will be called. And it also consists of required 'C' functions.

*Program : Write a YACC Program that will take arithmetic Expression as input & Produce the corresponding Postfix expression as output.

Sol:

```
/* LEX Program to Convert Infix expression to Postfix form */
%{
  #include <stdlib.h>
  #include <stdio.h>
  #include "y.tab.h" /* contain token definitions */
%}
%%
 /* Valid digit */
 [0-9]+ {
   /* Convert from character form to integer form */
   yylval.no = atoi(yytext);
   /* Return the appropriate token */
   return(DIGIT);
 }
/* Valid Identifier / Variable */
[a-z|A-Z][a-z|A-Z|0-9|_]* {
      strcpy(yylval.str, yytext);
      return(ID);
    }
/* Operators */
```

```
"+"   { return (PLUS); }
"-"   { return (MINUS); }
"*"   { return (MUL); }
"/"   { return (DIV); }
"^"   { return (EXPO); }
"("   { return (OPEN); }
")"   { return (CLOSE); }
"\n"  { return 0; }
/* Ignore white spaces */
 [\t] ;
/* Ignore remaining characters */
 . ;
%%
```

* YACC Program :

```
/* YACC Program to convert Infix expression to Postfix form */
%{
   #include <stdio.h>
%}
/* The identifier can be a digit (number) or a Variable (str) */
%union
  { int no;
    char str[10];
  }
```

```
/* Declaring the tokens */
%token <no> DIGIT
%token <str> ID
/* + and - are left associative & have the least Precedence */
%left PLUS MINUS
/* then comes * & / which are also left associative */
%left MUL DIV
/* And lastly exponent operator ^ which has higher Precedence
& is right associative */
%right EXPO
/* Brackets which has the highest Precedence */
%left OPEN CLOSE
%%
STMT : EXPR          { Printf ("\n"); }
EXPR : EXPR PLUS EXPR    { Printf ("+"); }
     | EXPR MINUS EXPR   { Printf ("-"); }
     | EXPR MUL EXPR     { Printf ("*"); }
     | EXPR DIV EXPR     { Printf ("/"); }
     | EXPR EXPO EXPR    { Printf ("^"); }
     | OPEN EXPR CLOSE
     | DIGIT     { Printf ("%d", yyval.no); }
     | ID        { Printf ("%s", yyval.str); }
     ;
%%
```

```
/* user sub-routine section */
int main(void)
{
    Printf("\n");
    yyparse();
    Printf("\n");
    return 0;
}
```

* output :-

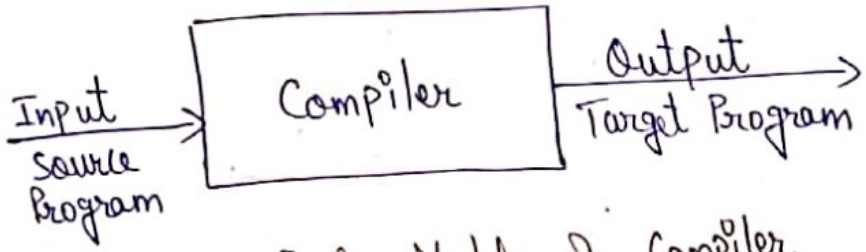
```
$ lex postfix.l
$ yacc postfix.y
$ gcc lex.yy.c y.tab.c -ll -ly
$ ./a.out
(2+3)*(3/4+4) - 3^2
23+34/4+*32^-
```

COMPILER DESIGN

UNIT I : Overview of Compiler & Lexical Analysis

*** Introduction :** Compiler is a program which takes one language (source Program) as input and translates it into an equivalent another language (target Program).

→ Compilers basically act as translators. The basic model of compiler can be represented as follows :



Basic Model of Compiler

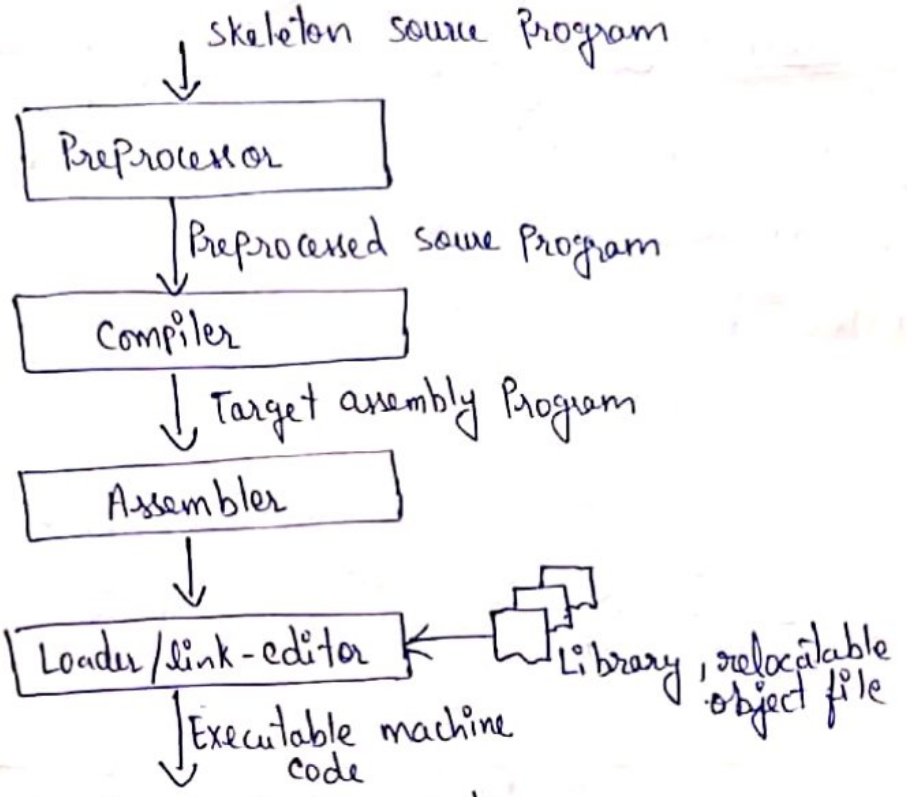
→ The compiler takes a source Program as higher level languages such as C, Pascal, Fortran & Converts it into low level language or a machine level language such as assembly language.

*** Properties of a Compiler :**

- The compiler itself must be bug-free.
- It must generate correct machine-code
- The generated machine code must run fast.
- The compiler must be portable.
- It must give good diagnostics & error messages.
- It must have consistent optimization.

Scanned by CamScanner

→ To create an executable form of your source Program only a compiler Program is not sufficient. we may require several other Programs to create an executable target Program



(or) Target Machine code.

Process of Execution of Program.

* Structure of a Compiler :

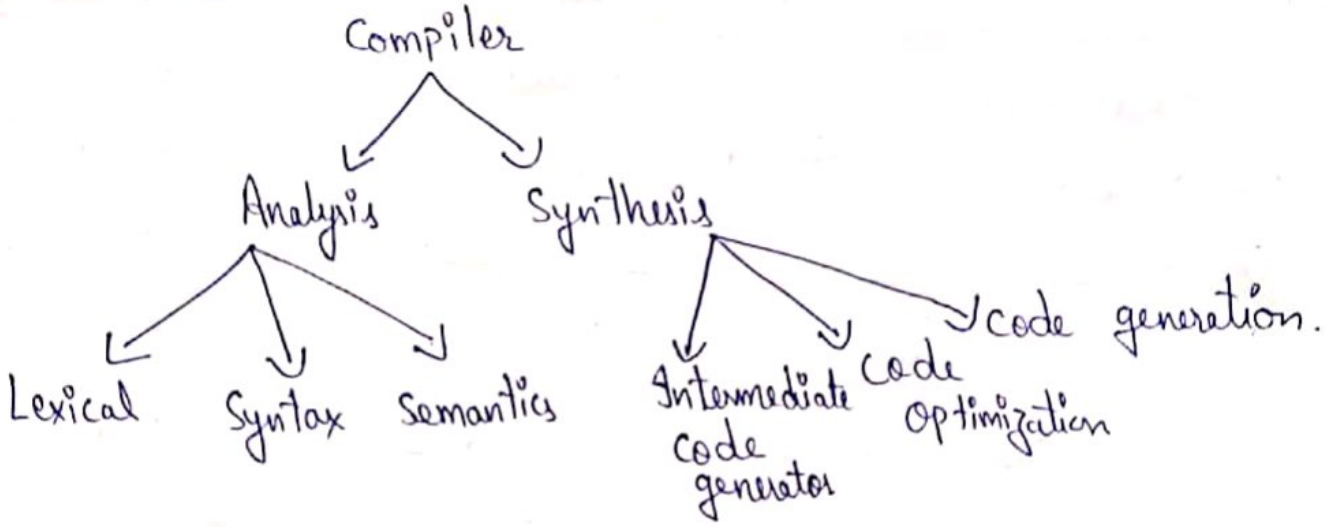
The Process of Compilation is carried out in two Parts namely: "Analysis & Synthesis".

→ Analysis Part is the front end of the compiler.

→ Synthesis Part is the back end of the compiler.

→ Compilation Process operates as a sequence of phases, each of which transforms one representation of the source program to another.

Scanned by CamScanner



Phases of compiler

(a) Lexical Analysis :

The lexical Analysis is also called "Scanning". The complete source code is scanned & the source program is broken up into group of strings called "lexemes". For each lexeme, the lexical analyzer produces as output a "token" of the form : (token-name, attribute-value).

sequence of characters having a collective meaning

→ The blank characters which are used in the Programming statement are eliminated during this phase.

Eg : c := a+10 ;

these are mapped to the tokens like this.

c, a ⇒ identifiers
:= ⇒ assignment
+ ⇒ operator
10 ⇒ constant.

tokens

<id, 1> <:=> <id, 2> <+> <10> ;

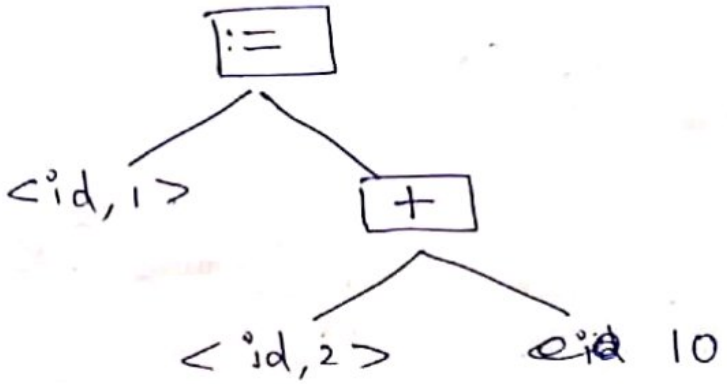
→ we use regular expressions to recognize tokens.

Scanned by CamScanner

(b) **Syntax Analysis** :- The syntax analysis is also called "Parsing". In this phase the tokens generated by the lexical analyzer are grouped together to form a hierarchical tree-like structure, called as Parse tree (or) Syntax tree. We use Context-free grammar (CFG) to recognize syntax or grammar.

Syntax tree for <id,1> <:=> <id,2> <+> ~~<id,3>~~ 10 is :-

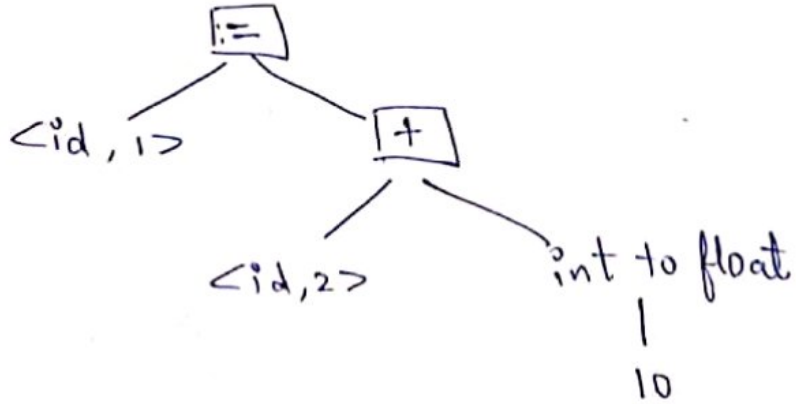
2 Parsing techniques in this phase
- Top-down
- Bottom-up



Parse tree (or) syntax tree.

(c) **Semantic Analysis** :-

Semantic Analysis determines the meaning of the source string. It also does "type checking" where the compiler checks that each operator has matching operands. For example, matching of Parenthesis in the expression, matching of else stmts, checking the scope of operation etc.



Scanned by CamScanner

(d) Intermediate code generation :

The intermediate code is a kind of code which is easy to generate & this code can be easily converted to target code. Intermediate code is of many forms such as three address code, quadruple, triple, Posix etc. For the above example let us convert it into three address form (consists of instructions each of which has at most three operands).

Eg :
$t_1 = \text{int to float}(10)$
$t_2 = id2 + t_1$
$id1 = t_2$

(e) Code optimization :

The code optimization phase attempts to improve the intermediate code. This is necessary to have a faster executing code or less consumption of memory. By optimizing the code the overall running time of the target program can be improved.

Eg :
$t_1 = id2 + 10.0$ // code is optimized here. unnecessary variables are removed.
$id1 = t_1$

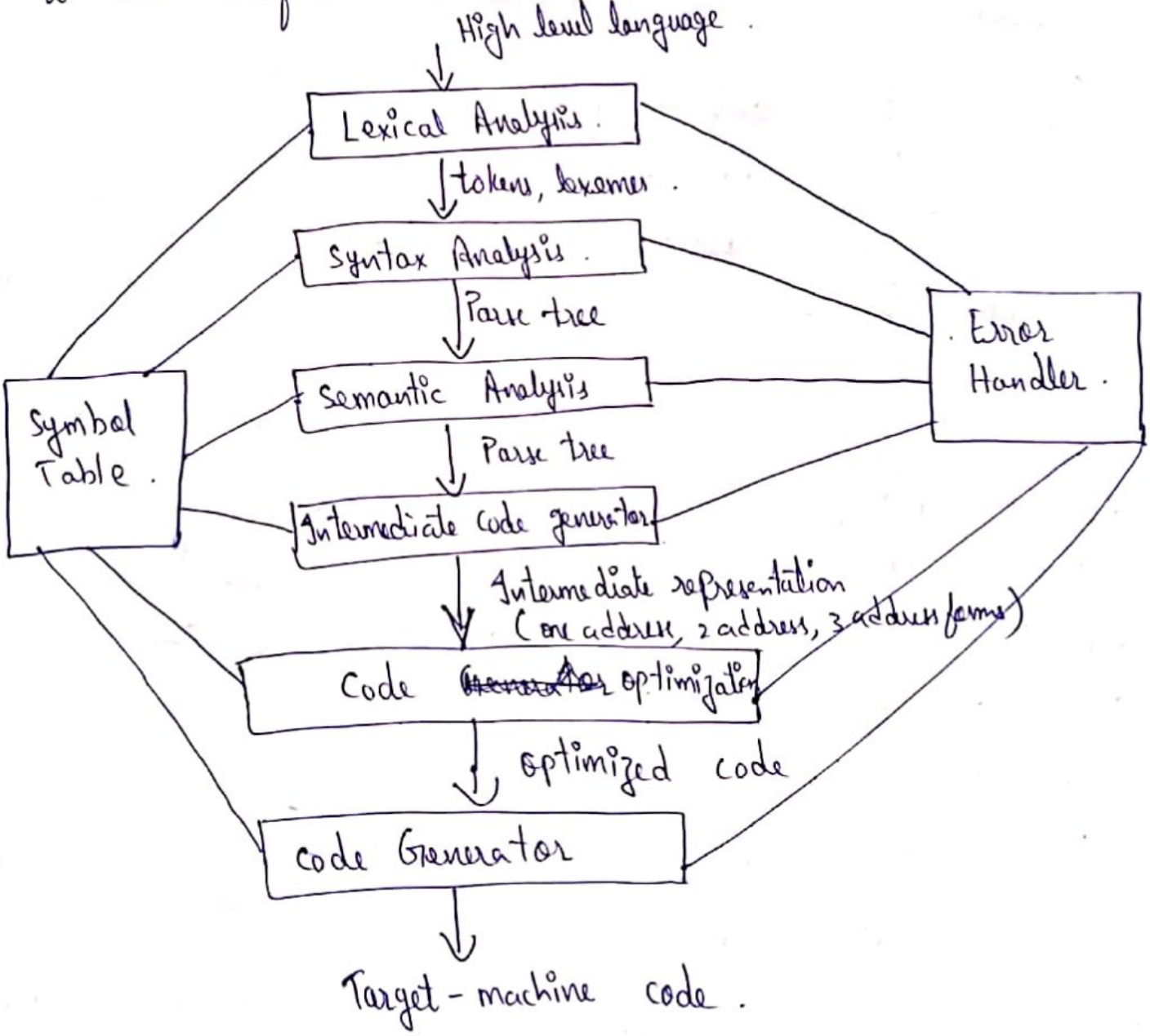
(c) Code Generation : In code generation phase the target code gets generated. The intermediate code instructions are translated into sequence of machine instructions.

Scanned by CamScanner

Eg: Mov, id2, R1
ADDF R1, #10.0, R1

```
Mov   id2, R1
ADDF  #10.0, R1
Mov   R1, id3
```

we are moving value of id2 to 'R1' register. Then we add id2 & 10 & result is stored in 'R1'. Then value of R1 is assigned to id3 identifier.



Phases Of Compiler

Scanned by CamScanner

* Symbol Table

Symbol table is a data structure containing a record for each variable name, with fields for the attributes of the name. Symbol table also stores information about the subroutines used in the program.

→ It allows us to find the record for each identifier quickly & to store or retrieve data from that record efficiently.

* Error detection & Handling :

In compilation, each phase detects errors. These errors must be reported to error handler, whose task is to handle the errors so that the compilation can proceed. Errors are reported in the form of 'messages'.

→ Large no. of errors can be detected in syntax analysis phase. Such errors are called as "syntax errors".

* Pass : when several phases are grouped together, we call it as Pass.

* The Science of Building a Computer :

A compiler must accept all source programs that conform to the specification of the language. Any transformation performed by the compiler while translating a source program must preserve the meaning of the program being compiled. Compiler writers thus have influence over not just the

Scanned by CamScanner

Compilers they create, but all the Programs that their compilers compile.

a) Modeling in Compiler Design & Implementation

The study of compilers is mainly a study of how we design the right mathematical models & choose the right algorithms, while balancing the need for generality & Power against simplicity & efficiency.

→ Some of most fundamental models are finite-state machines & regular expressions (for describing the lexical units of Programs), context-free grammars (to describe syntactic structure).

b) The Science of code optimization:

The term 'optimization' in compiler design refers to the attempts that a compiler makes to produce code that is more efficient than the obvious code.

→ Compiler optimizations must have the following design objectives:

- The optimization must be correct, i.e Preserve the meaning of the compiled Program.
- The optimization must improve the Performance of many Programs.
- The compilation time must be kept reasonable
- The Engineering effort required must be manageable.

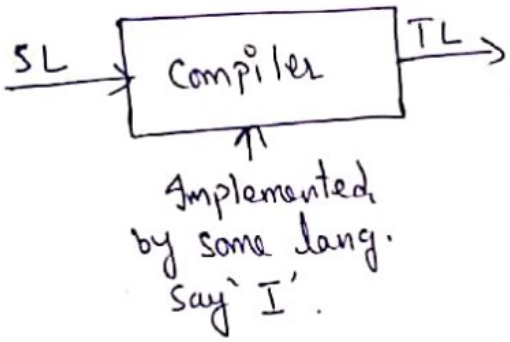
Scanned by CamScanner

* Bootstrapping :

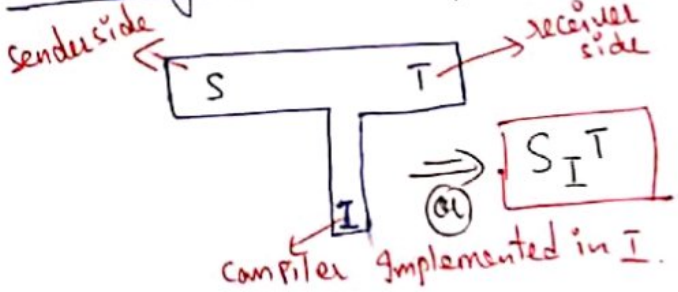
A Process by which simple language is used to translate more complicated Program, which in turn may handle an even more complicated Program.

→ There are three types of languages.

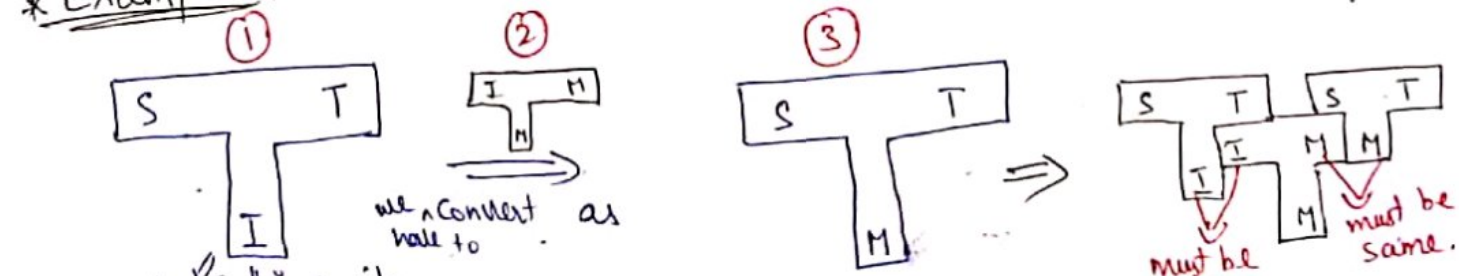
- Source language(SL) ; i.e the application program
- Target language(TL) in which machine code is written.
- Implementation language in which a compiler is written.



⇒ This can be represented by "T-diagram" as follows;



* Example :



Say we have this compiler

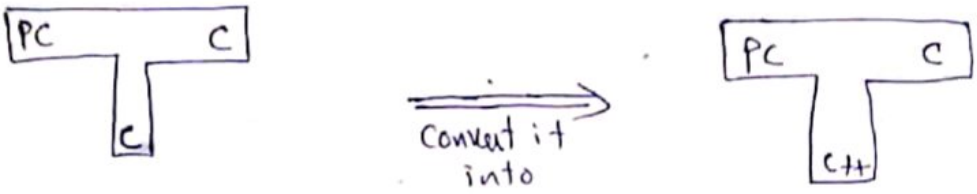
→ The Process of converting 'I' into 'M' lang. machine is Boot-strapping.

Scanned by CamScanner

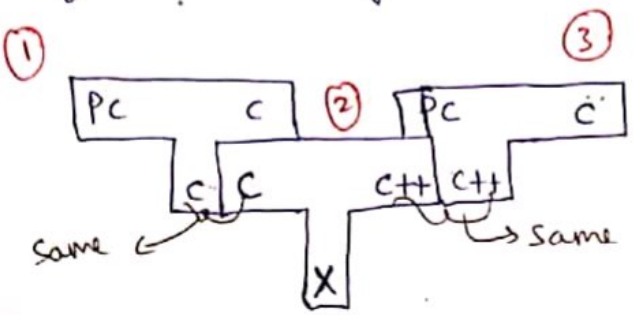
Example :

We have Pascal Translator written in → 'c' lang.
i/p is Pascal code(Pc), o/p – 'c'. create pascal translator in c++.

sol : Given :



with help of Bootstrapping :



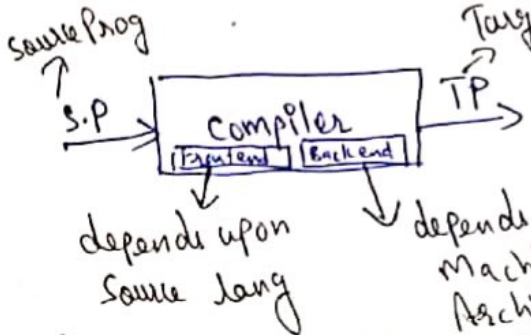
X ⇒ can be anything like java or any other lang

(or) $P_c c + C_x c++ \Rightarrow P_{c++} c$

* Cross Compiler :

Cross compiler is a compiler which is capable of creating executable code for a platform other than the one on which compiler is running.

eg : A compiler which runs on windows 7 but this generates a code which can run on Android smartphone. Then this compiler is known as Cross-compiler.



⇒ we need to change only the backend to implement Cross compiler.

→ using cross compilation, platform independency can be achieved.

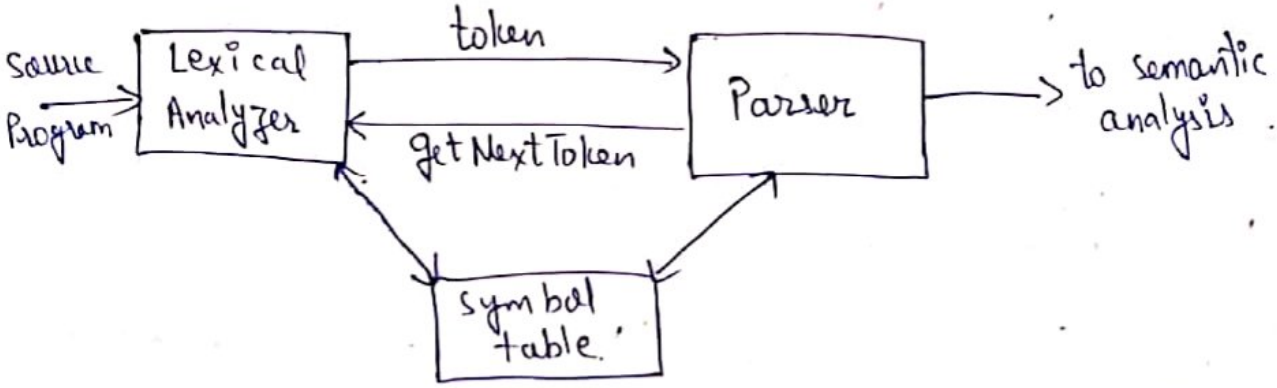
Scanned by CamScanner

LEXICAL ANALYSIS

* Role of Lexical Analyzer :-

Lexical Analyzer reads the input source program from left to right one character at a time & generates the sequence of tokens. It is the first phase of a compiler.

→ Each token is a single logical cohesive unit such as identifier, keywords, operators & punctuation marks. Then the Parser to determine the syntax of the source program can use these tokens.



Interactions between the lexical analyzer & the Parser

→ As the lexical analyzer scans the source program to recognize the tokens, it is also called as "Scanner".

* Functions of lexical Analyzer :-

- It Produces stream of tokens
- It eliminates whitespaces (blank, newline, tab) & comments.
- It generates symbol table which stores the information about identifiers, constants encountered in the i/p.
- It keeps track of line nos.
- It reports the error encountered while generating the tokens.

Scanned by CamScanner

* Tokens : A token is a pair consisting of a token name & an optional attribute value. It describes the class or category of input string. for example identifiers, keywords, constants are called tokens.

* Pattern : A pattern is a description of the form that the lexemes of a token may take. It is a set of rules that describe the token.

* Lexemes : Sequence of characters in the source program that are matched with the pattern of the token. For example int, i, num, choice;

Example : if (a < b)

if	(	a	<	b	)

⇒ tokens which are generated.

here "if", "(", "a", "<", "b", ")" ⇒ are all lexemes.

→ 'if' is a keyword
→ "(" is openening Parenthesis
→ "a" is identifier → collection of letter.
→ "<" is an operater
→ "b" is identifier
→ ")" is closing Parenthesis.

Scanned by CamScanner

* Attributes for tokens :

Eg :

Token		Lexemes
number	— 0, 1, 2, ...	0, 6.2, 1, ...

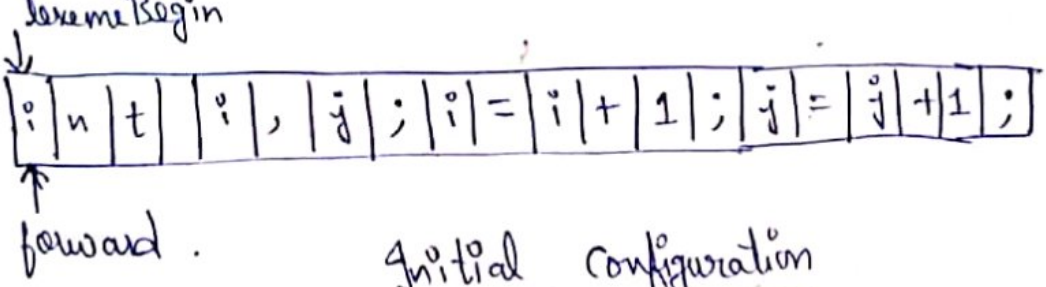
when more than one lexeme can match a pattern, the lexical analyzer must provide the subsequent compiler phases additional information about the particular lexeme that matched.

Eg : Pattern for "token: Number" matches both '0' and '1'.

→ Lexical analyzer returns to the parser not only a token name but an attribute value that describes the lexeme represented by the token. Token name influences parsing decisions, while the attribute value influences translation of tokens after the ~~the~~ Parse.

* Input Buffering :

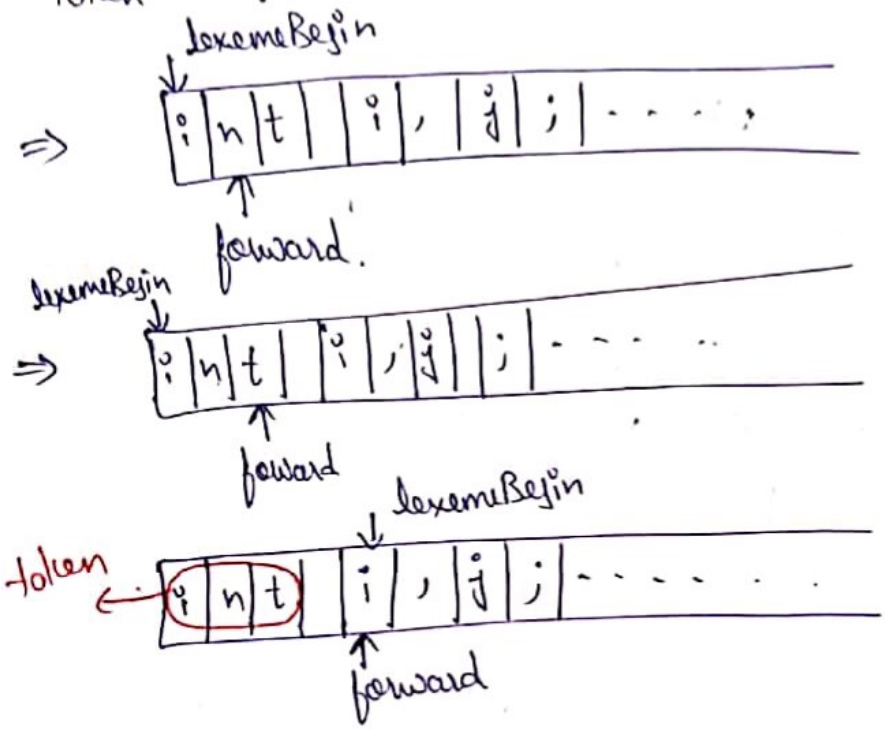
The lexical analyzer scans the input string from left to right one character at a time. It uses two pointers: begin Pointer "lexemeBegin" and foward Pointer "foward" to keep track of the portion of the input scanned. Initially both the pointers point to the first character of the input string as:



Initial Configuration

Scanned by CamScanner

The 'lexemeBegin' pointer remains at the beginning of the sting to be read & the 'forward' pointer moves ahead to search for end of lexeme. As soon as the blank space is encountered, it indicates end of lexeme. & then both 'lexemeBegin' & 'forward' is set at next token 'i'.



→ Reading i/p's from secondary storage is costly. so, a block of data is first read into a buffer & then scanned by lexical analyzer.

Input Buffering

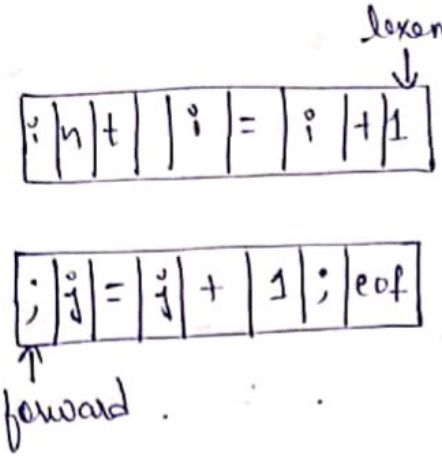
→ Before one buffer scheme was used but now we are using two buffer scheme.

* Two Buffer Scheme

→ to store i/p string.

Two buffers each of size 'N' (usually size of disk block) are used ~~A~~ Each buffer can read 'N' characters. In this method, first buffer & second buffer are scanned alternately. when the end of current buffer is reached the other buffer is filled.

Scanned by CamScanner



Two Buffer scheme

* Sentinel :
Both buffers have a special character "eof" at the end. when 'lexemeBegin' Pointer encounters 'eof' at Buffer 1, then it starts filling up second buffer. Similarly when 'eof' is encountered at Buffer2, it starts filling Buffer1 & so on.

→ This 'eof' character introduced at the end is called "Sentinel" which is used to identify the end of buffer.

* Specifications of tokens :

To specify tokens 'regular expressions' are used. when a pattern is matched by some regular expression then token can be recognized.

* Alphabet : Alphabet is a finite, non-empty set of symbols. It is represented by the symbol 'Σ'.

Eg : $\Sigma = \{0, 1\}$, $\Sigma = \{a, b\}$

$\Sigma = \{a, b, c, \cdots\}$, $\Sigma = \{e_0, e_1, \cdots e_9\}$

Scanned by CamScanner

* String : String is a finite sequence of symbols choosen from some alphabet.

Eg : $\Sigma = \{0, 1\}$.

0011 is a string of Σ , 102 is not a string of Σ.

1010 " " " " " , 234 " " " " " .

→ length of string is denoted by $|s|$..

→ Empty string can be denoted by 'ϵ'.

* Language : Set of strings which belong to some alphabet 'Σ' is called as a Language (L).

Eg : $\Sigma = \{0, 1\}$.

$L = \{00, 01, 10, 11\}$

$L = \{0^n 1^n \mid n \geq 1\}$.

* Operations on Language :

There are various operations which can be performed on a language as follows : below. Let 'L', M' be two languages.

Operation	definition & Notation.
• Union of L & M	$L \cup M = \{ s \mid s \text{ is in } L \text{ or } s \text{ is in } M \}$
• Concatenation of L & M	$LM = \{ st \mid s \text{ is in } L \text{ \& } t \text{ is in } M \}$
• Kleen closure of 'L'	$L^* = \bigcup_{i=0}^{\infty} L^i$
• Positive closure of 'L'	$L^+ = \bigcup_{i=1}^{\infty} L^i$

Definitions of operations on languages

Scanned by CamScanner

Eg :- $L = \{A, B, C, \ldots Z, a, b, c, \ldots\}$, $D = \{0, 1, 2, \ldots 9\}$ are two languages.

- LUD is a set of letters & digits
- LD is a set of strings consisting of letters followed by digits.
- L^5 is a set of strings having length of 5 each.
- L^* is " " " " all strings including 'ϵ'.
- L^+ " " " " except 'ϵ'.

* **Regular Expressions** :- Regular Expression is a notation to represent certain sets of strings in an algebraic fashion. This notation involves a combination of strings of symbols from some alphabet 'Σ', the symbol of null string 'ϵ', star operator (*) & plus operator (+).

* **Examples** :-

(1) write a Regular Expression (RE) for a language containing the strings of length two over $\Sigma = \{0, 1\}$.

Sol :- RE = (0+1)(0+1).

(2) write a RE for lang. containing strings which end with 'abb' over $\Sigma = \{a, b\}$

Sol :- $(a+b)^*abb$.

(3) write a RE for a recognizing identifier.

Sol :- For denoting identifier, we will consider a set of letters

Scanned by CamScanner

& digits because identifier is a combination of letters or letter & digits but having first character as letter always. Hence RE can be denoted as:

$$RE = letter\ (letter + digit)^*$$

where letter = (A, B, ... Z, a, b, ... z) & digit = (0, 1, 2, ... 9).

* Regular Definitions:

If 'Σ' is an alphabet of basic symbols, then a regular definition is a sequence of definitions of the form:

$$d_1 \longrightarrow r_1$$
$$d_2 \longrightarrow r_2$$
$$\vdots \qquad \vdots$$
$$d_n \longrightarrow r_n$$

where,

- Each 'di' is a new symbol, not in 'Σ' & not the same as any other of the d's &
- Each 'ri' is a regular expression over the alphabet $\Sigma \cup \{d_1, d_2, \ldots, d_{i-1}\}$.

→ Some of the notations used for writing the regular expressions are as follows:

(1) One or more instance : To represent one or more instances "+" sign is used. Eg : r^+.

(2) Zero or more instances : To represent zero or more instances "*" sign is used. Eg : r^*.

(3) Character classes : A class of symbols can be denoted by [ ]. Eg : [012] means 0 or 1 or 2.

Scanned by CamScanner

* Examples :

(1) write Regular definition for 'C' language identifiers.

sol : 'C' identifiers are strings of letters, digits & underscores.

Regular definition is :

letter_ → [A-z a-z]
digit → [0-9]
id → letter_ (letter_ | digit)*

(2) write Regular definition for Unsigned numbers.

sol : unsigned nos (integer or floating point are strings like :
5080, 0.01234, 6.33458, 0.89E-4 etc.

Regular definition is :

digit → [0-9]
digits → digit+
number → digits (. digits)? (E [+-]? digits)?

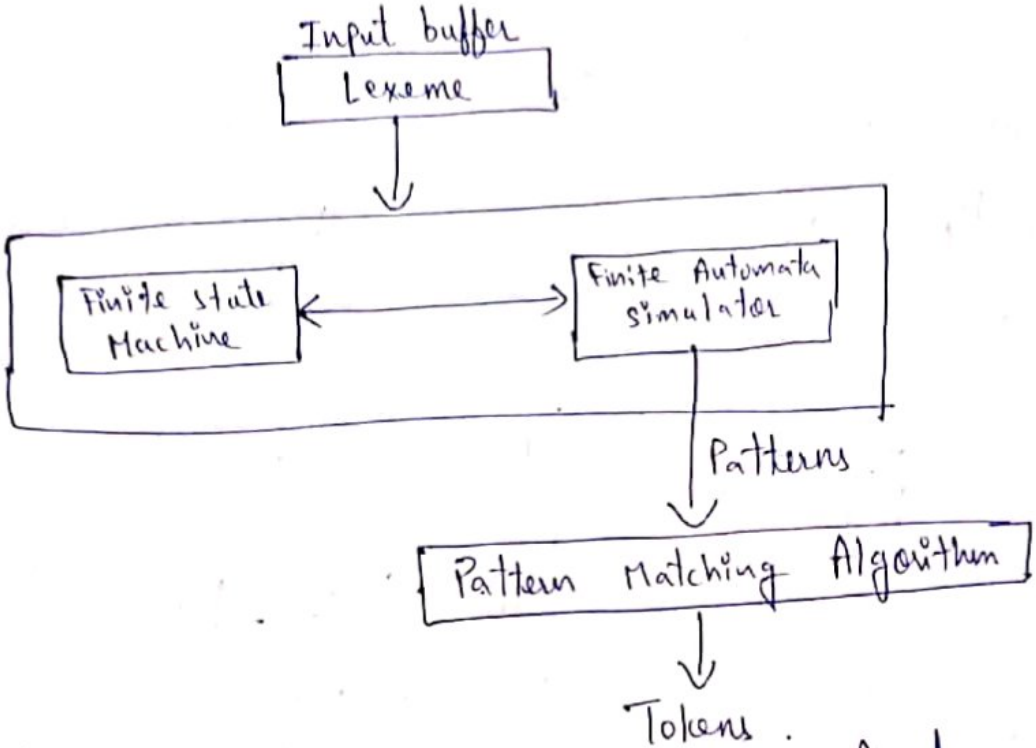
(? → zero or one instance)

* Recognition of Tokens :

The token is usually represented as: | Token type | Token value/attribute |

→ The token type tells us the category of token & token value gives us the information regarding token. During lexical analysis, Process, the symbol table is maintained. The token value can be a pointer to symbol table in case of identifiers & constants.

Scanned by CamScanner



Block diagram of Lexical Analyzer

* Example :

digit → [0-9]
digits → digit+
number → digits (. digits)? (E [+-]? digits)?
letter → [A-Z a-z]
id → letter (letter | digit)*
if → if
then → then
else → else
relational operators ↷ relop → < | > | <= | >= | = | <>

Sample Regular definition

Scanned by CamScanner

Lexemes	Token Name	Attribute Value
Any ws	—	—
if	if	—
then	then	—
else	else	—
Any id	id	Pointer to table entry
Any number	number	Pointer to table entry.
<	relop	LT
<=	relop	LE
=	relop	EQ
<>	relop	NE
>	relop	GT
>=	relop	GE

Tokens, their patterns & attribute values

* Transition Diagrams:

In the middle of lexical analyzer Process we convert patterns into Transition diagrams. Transition diagrams have a collection of nodes or circles called states. Each state represents a condition that could occur during the Process of scanning the input looking for a lexeme that matches one of several patterns.

→ Edges are directed from one state to another. Each edge is labeled by a symbol or set of symbols.

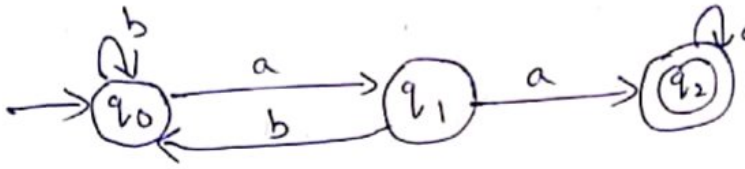
Scanned by CamScanner

- $\rightarrow (q_0)$ ⇒ initial state
- $\textcircled{\textcircled{q}}$ ⇒ final state

* Examples :-

(1) Design a transition diagram for the language consisting of all the strings containing at least one pair of consecutive a's over $\Sigma = \{a, b\}$.

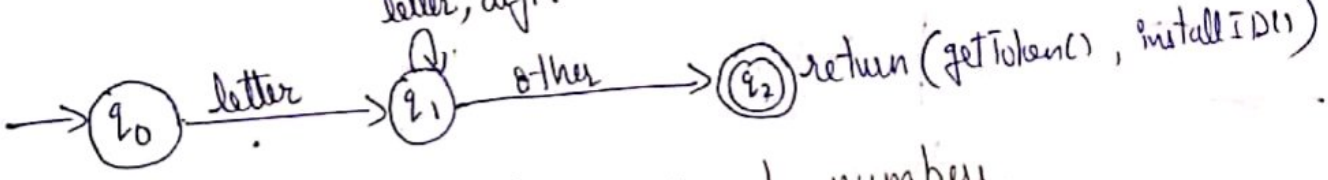
Sol :-



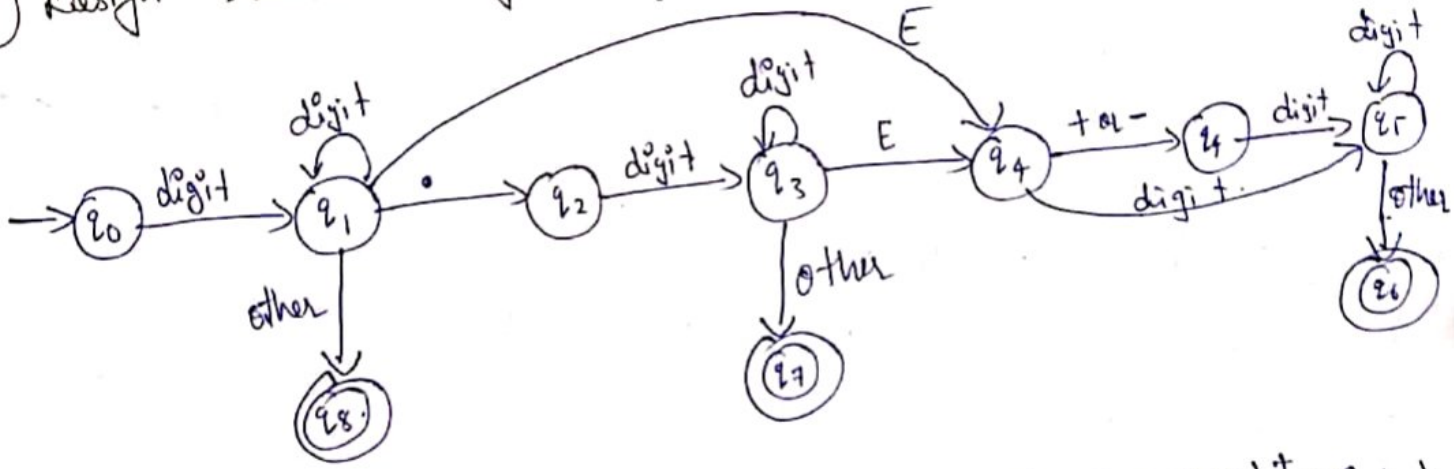
	a	b
$\rightarrow q_0$	q_1	q_0
q_1	q_2	q_0
q_2	q_2	q_2

(2) Design transition diagram for identifiers.

Sol :- RE for identifier :- r.e = letter (letter | digit)*



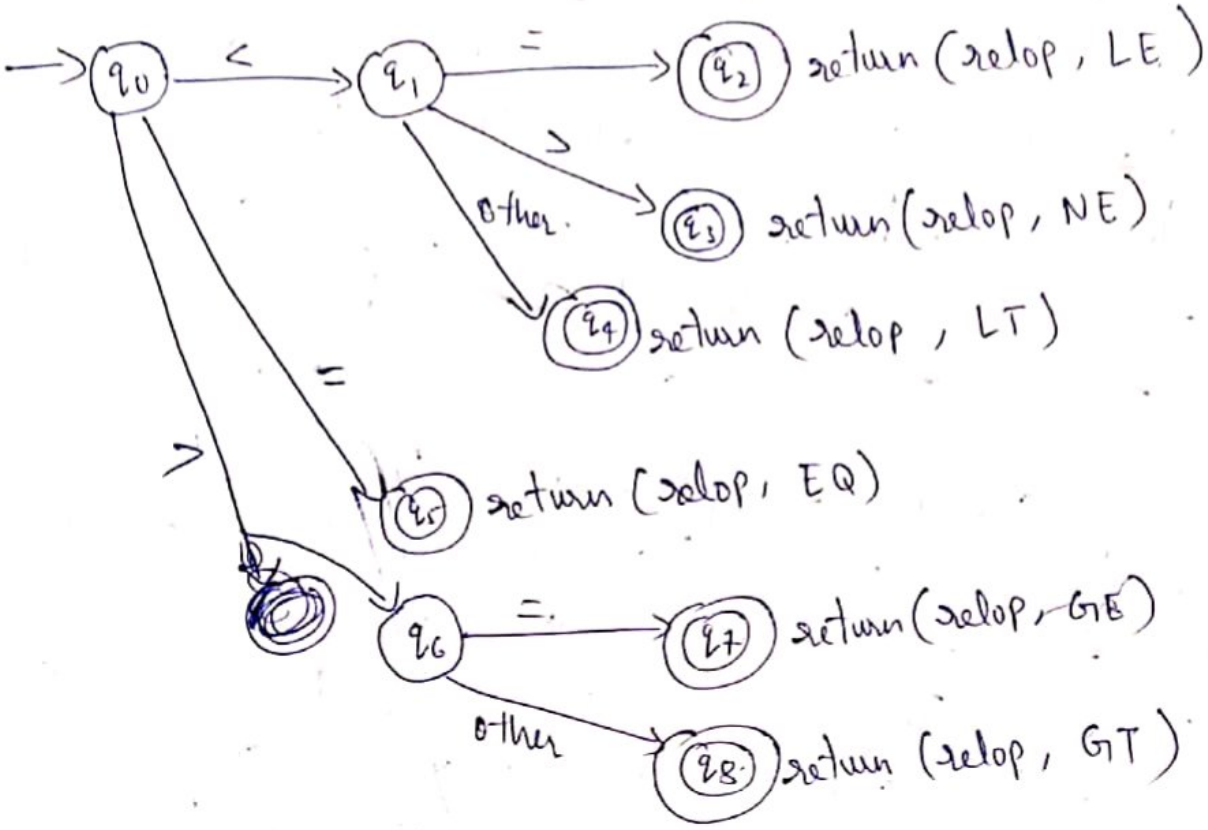
(3) Design transition diagram for unsigned numbers.



re = digit$^+$ | (digit)$^+$ (.) (digit)$^+$ | (digit)$^+$ (.) (digit)$^+$ E (+|-) (digit$^+$).

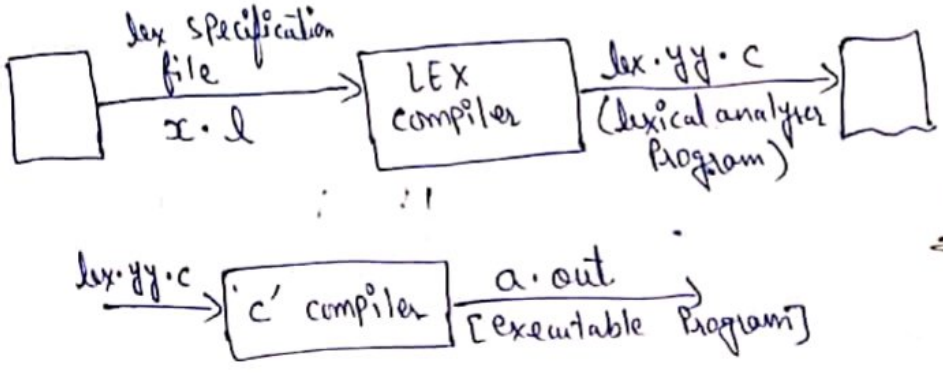
Scanned by CamScanner

④ Design transition diagram for relational operators (relop).



* LEX — Lexical Analyzer Generator: (or FLEX)

Regular expressions are used in recognizing the tokens. LEX is a tool called or FLEX (recent implementation) allows one to specify a lexical analyzer by specifying regular expressions to describe patterns for tokens.



Generation of lexical analyser using LEX

Scanned by CamScanner

→ Specification file can be created using a file with extension ".l" (dot 'l'). Say for example "x.l". This x.l is then given to LEX compiler to Produce lex.yy.c. This lex.yy.c is a 'c' Program which is actually a lexical analyzer Program we know that the specification file stores the regular expressions for the tokens. The lex.yy.c file consists of the tabular representation of the transition diagrams constructed for the regular expressions of specification file say x.l.

→ The lexemes can be recognized with help of tabular transitions diagrams & some standard routines. In the specification file of LEX, actions are associated with each regular expression. These actions are simply pieces of 'c' code. These pieces of 'c' code are directly carried over to the lex.yy.c. Finally the 'c' compiler compiles this generated lex.yy.c & Produces an object Program a.out. When some input stream is given to a.out then sequence of tokens get generated.

→ LEX Program consists of three parts:
a) Declaration section, b) Rule section & c) Procedure section.

```
%{
   Declaration section
%}
%%
Rule section
%%
Auxiliary Procedure section
```

LEX Program format

Scanned by CamScanner

(a) Declaration section : In this section, declaration of variables, constants can be done. Some regular definitions can also be written in this section. The regular definitions are basically components of regular expressions appearing in the rule section.

(b) Rule section: It consists of regular expressions with associated actions. These translation rules can be given as :

R_1	{action 1}
R_2	{action 2}
⋮	
R_n	{action n}

where each R_i is a regular expression & each action; is a program fragment describing what action is to be taken for corresponding regular expression. These actions can be specified by piece of 'C' code.

(c) Auxiliary Procedure section: In this section, all the requi -red procedures are defined. Sometimes these procedures are required by the actions in the rule section.

→ The lexical analyzer or scanner works in co-ordination with Parser. when activated by the parser, the lexical analyzer begins reading its remaining input, character by character at a time. when the string is matched with one of the regular expressions R_i then corresponding action; will

Scanned by CamScanner

get executed & this action; returns the control to the Parser.

* Example LEX Programming :

```
// Program name : x.l
%{
%}  // Declaration section
%%  // Rule section
"Rama"  |
"seeta"  |
"Geeta"  |
"Neeta"  |  Printf ("In Noun");
"sings"  |
"dances"  |
"eats"      Printf ("In Verb");
%%
main()  // Auxiliary Procedure section.
{
   yylex();
}
int yywrap()
{
   return 1;
}
```

→ This is a simple program that recognizes noun & verb from the string.

Scanned by CamScanner

* Execution :

```
$ lex x.l ↵
$ cc lex.yy.c ↵
$ ./a.out ↵
```

o/p : ip1 : Rama eats ↵
o/p : Noun
Verb.

ip2 : seeta sings ↵
o/p : Noun
Verb

* LEX Actions :

(1) BEGIN : It indicates the start state. The lexical analyzer starts at state '0'.

(2) ECHO : It emits the input as it is.

(3) yylex() : As soon as call to yylex() is encountered, scanner starts scanning the source Program.

(4) yywrap() : The function yywrap() is called when scanner encounters end of file. If yywrap() returns '0'(zero) then scanner continues scanning. when " " '1'(one) that means end of file is encountered.

(5) yyin : It is the standard file that stores input source Program.

Scanned by CamScanner

*FLEX :

Flex is a tool for generating scanners. A scanner, sometimes called a 'tokenizer' is a program which recognizes lexical patterns in text. The flex program reads user-specified input files or its standard input if no file names are given for a description of a scanner to generate. The description is in the form of pairs of regular expressions & 'C' code called rules. Flex generates a 'C' source file named "lex.yy.c" which defines the function yylex(). The file lex.yy.c can be compiled & linked to produce an executable.

→ when the executable is run, it analyzes its i/p for occurances of text matching with the regular expressions for each rule. whenever it finds a match, it executes the corresponding 'C' code.

Scanned by CamScanner