SNS COLLEGE OF ENGINEERING
Kurumbapalayam (Po), Coimbatore - 641 107
AN AUTONOMOUS INSTITUTION
Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai

19CS502 - AUTOMATA THEORY AND COMPILER DESIGN

UNIT-1

FORMAL LANGUAGE AND REGULAR EXPRESSIONS.

Formal language and Regular Expression: Languages, Definition Languages regular expressions, Finite Automata DFA, NFA. conversion of regular expression to NFA NFA to DFA. Applications of finite Automata to lexical analysis, lex tools.

Automata Theory :

→ Automata is the study of abstract machines and the computation problems that can be solved using these machines. The abstract machine is called automata.

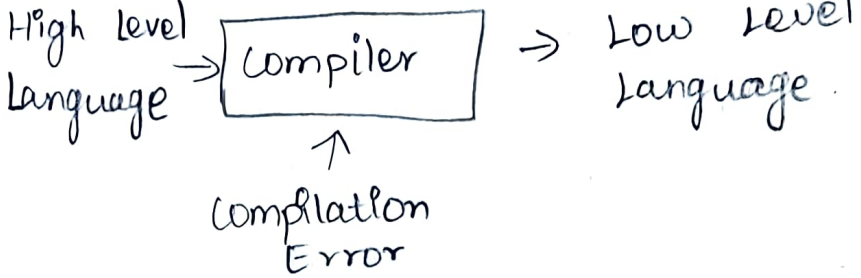
→ This automation consists of states and transitions The State is represented by circles, and the Transition is represented by arrows.

○ → States

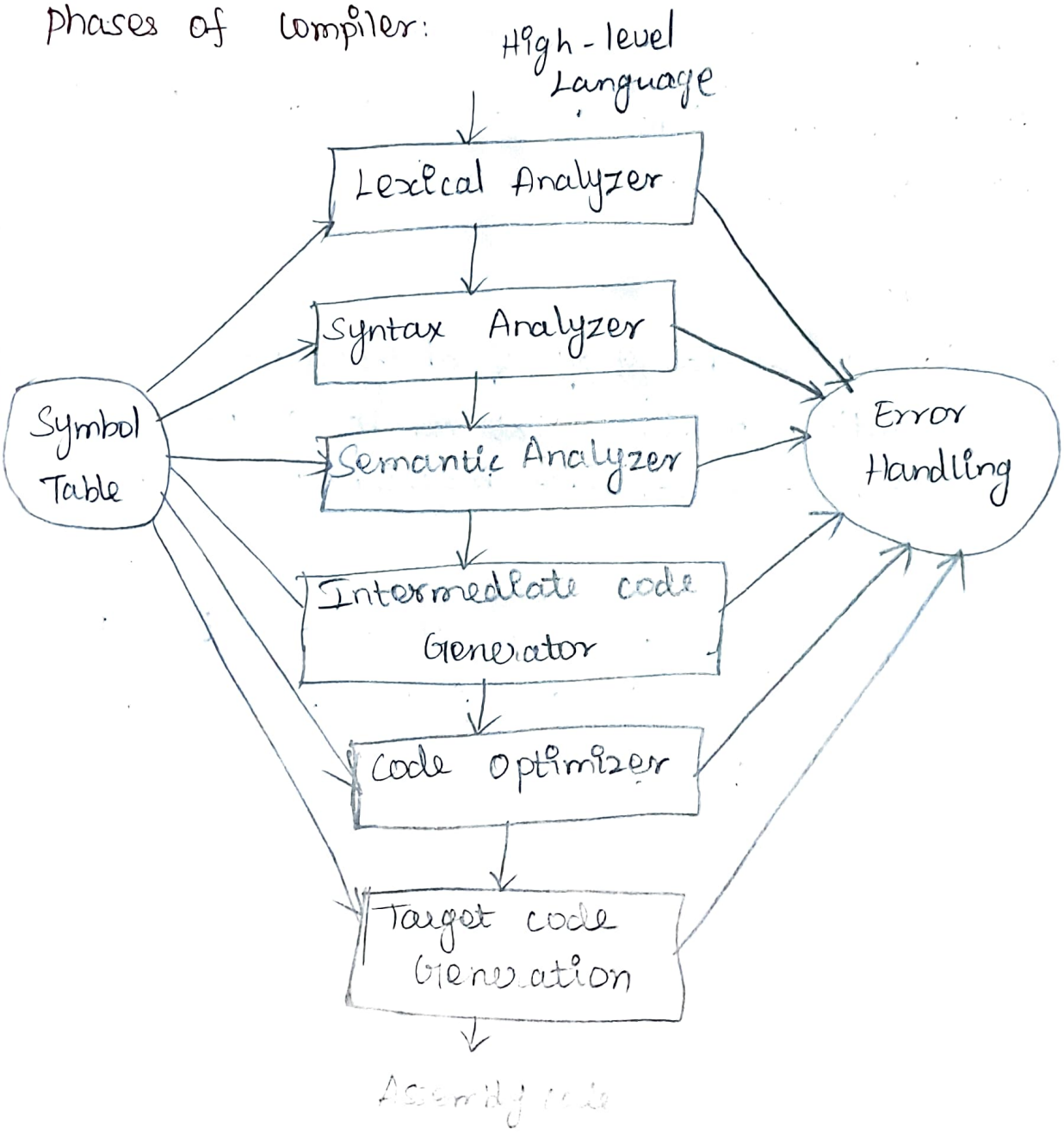
⟶ → Transition

Compiler Design:

→ The compiler is s/w that converts a program written in a high-level language (source language) to a low-level language (object / Target / Machine language 0's & 1's).

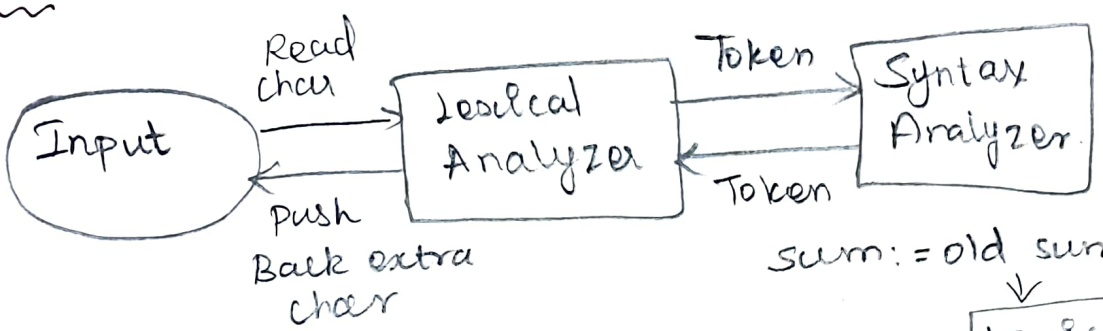


Phases of compiler:



SNS COLLEGE OF ENGINEERING
Kurumbapalayam (Po), Coimbatore – 641 107
AN AUTONOMOUS INSTITUTION
Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai

→ Its the first phase of compiler also known as a Scanner.
→ It converts High level i/p program into a sequence of Tokens.



sum := old sum + Rate * 50
↓
Lexical Analyzer

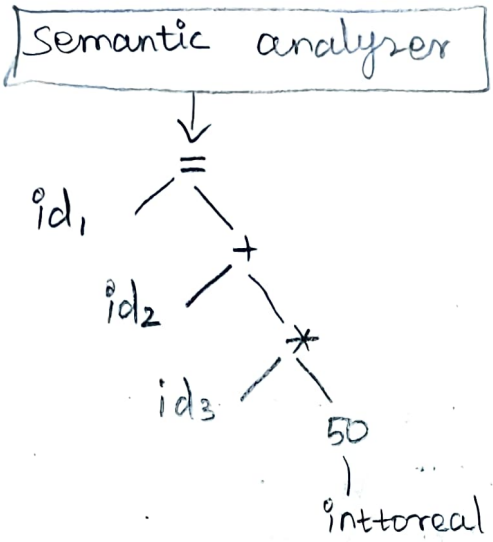
Ex:
→ Type token (id, number, real, ...)
→ Punctuation token (IF, void, return. .....)
→ Alphabetic token (keywords).

Syntax Analysis:
→ Its the second phase of compilation process.
→ It takes tokens as input and generates a parse tree as output.
→ The parser checks that the expression made by the tokens is syntactically correct or not.

id1 = id2 + id3 * id4.

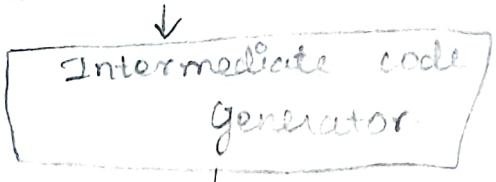
Semantics Analysis:

→ Its third phase of compilation. It checks whether the parse tree follows the rules of language. Semantic analyzer keeps track of identifier, their types & expressions. The o/p of semantic analysis phase is the annotated tree syntax.



Intermediate code generation:

→ Compiler generates the source code into the intermediate code. Intermediate code is generated between the high-level language and the machine language. It should be generated in such a way that you can easily translate it into the target machine code.



temp1 := inttoreal (50)
temp2 := id3 * temp1
temp3 := id2 * temp2
id1 := temp3

SNS COLLEGE OF ENGINEERING
Kurumbapalayam (Po), Coimbatore – 641 107
AN AUTONOMOUS INSTITUTION
Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai

Code optimization:

→ Its the optional phase. Its used to improve the intermediate code so that the output of the program could run faster and take less space. It removes the unnecessary lines of the code and arranges the sequence of statements in order to speed up the program execution.

code optimization

↓

```
temp1 := id3 * 50
id1 := id2 + temp1
```

Code Generation:

→ Its the final stage of compilation process. It takes the optimized intermediate code as i/p & maps it to the target machine language. It translates the intermediate code into the machine code of the specified computer.

code generation

↓

```
MOVF id3, R2
MULF #50, R2
MOVF id2, R2
ADDF R2, R1
MOVF R1, id1
```

Formal language

Should to know
-> Alphabet string, Lang
-> Mathematical Indu
-> Finite automata.

equiv NFA & DFA
NFA with ε moves.

Alphabets

-> An alphabets is a finite, non-empty set of symbols we use the symbol Σ for an alphabet.

eg:

$\Sigma = \{0, 1\}$, $\Sigma = \{a, b, \ldots z\}$.

String:

-> A string is a finite sequence of symbols from some alphabet.

eg: "xyz" is a string over an alphabet $\Sigma = \{a, b, \ldots z\}$

The empty string or null string is denoted by ϵ.

Length of string:

-> The length of a string is the no. of symbols in that string. If "w" is a string then its length is denoted by $|w|$

eg:

$w = abcd$, then length of w is $|w| = 4$.

$n = 100$ is a string, then $|n| = 3$

ϵ is the empty string & has length zero.

The set of strings of length k ($k \geq 1$)

Let Σ be an alphabet and $\Sigma = \{a, b\}$, then all strings of length k ($k \geq 1$) is denoted by Σ^k. where.

$\Sigma^k = \{w : w \text{ is a string of length } k, k \geq 1\}$

eg:

$\Sigma = \{a, b\}$ then

$\Sigma^1 = \{a, b\}$

$\Sigma^2 = \{aa, ab, ba, bb\}$

$\Sigma^3 = \{aaa, bbb, aba, baa, bab, bba, aab, abb\}$

SNS COLLEGE OF ENGINEERING
Kurumbapalayam (Po), Coimbatore - 641 107
AN AUTONOMOUS INSTITUTION
Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai

$|\Sigma^1| = 2 = 2^1$ (No. of strings of length one)

$|\Sigma^2| = 4 = 2^2$ (" " " " " two).

Eg2:

$S = \{0, 1, 2\}$ then $S^2 = \{00, 01, 02, 11, 10, 12, 22, 21, 20\}$.

$$|S^2| = 3^2 = 9.$$

Regular Expression (RE)

1. ε is a regular expression, $L(\varepsilon) = \{\varepsilon\}$
2. If a is a symbol in Σ then a is a regular expression
   $L(a) = \{a\}$
3. Suppose r & s are regular expression denoting the language $L(r)$ and $L(s)$
   - → $(r) | (s)$ is a RE denoting the language $L(r) \cup L(s)$
   - → $(r)(s)$ is a RE denoting the language $L(r).L(s)$.
   - → (r^*) is a RE denoting the Language $(L(r))^*$.
   - → (r) is a RE denoting the language $L(r)$

Identifier: