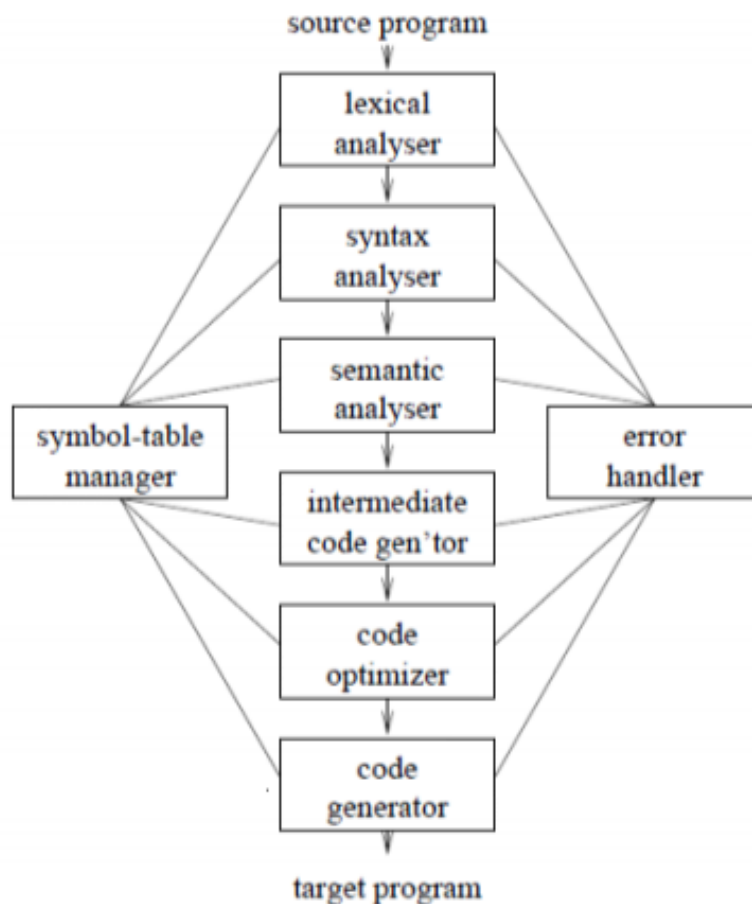


Lecturer (3)

1.3 The Phases of a Compiler

The process of compilation is split up into six phases, each of which interacts with a symbol table manager and an error handler. This is called the analysis/synthesis model of compilation. There are many variants on this model, but the essential elements are the same.



1.3.1 Lexical Analysis

A lexical analyser or scanner is a program that groups sequences of characters into lexemes, and outputs (to the syntax analyser) a sequence of tokens. Here:

- (a) Tokens are symbolic names for the entities that make up the text of the program; e.g. if for the keyword if, and id for any identifier. These make up the output of the lexical analyser.
- (b) A pattern is a rule that specifies when a sequence of characters from the input constitutes a token; e.g the sequence i, f for the token if, and any sequence of alphanumeric starting with a letter for the token id.

- (c) A lexeme is a sequence of characters from the input that match a pattern (and hence constitute an instance of a token); for example `if` matches the pattern for `i f`, and `foo123bar` matches the pattern for `i d`.

For example, the following code might result in the table given below.

<i>Lexeme</i>	<i>Token</i>	<i>Pattern</i>
program	program	p, r, o, g, r, a, m newlines, spaces, tabs
foo	id (foo)	letter followed by seq. of alphanumerics
(leftpar	a left parenthesis
input	input	i, n, p, u, t
,	comma	a comma
output	output	o, u, t, p, u, t
)	rightpar	a right parenthesis
;	semicolon	a semi-colon
var	var	v, a, r
x	id (x)	letter followed by seq. of alphanumerics
:	colon	a colon
integer	integer	i, n, t, e, g, e, r
;	semicolon	a semi-colon
begin	begin	b, e, g, i, n newlines, spaces, tabs
readln	readln	r, e, a, d, l, n
(leftpar	a left parenthesis
x	id (x)	letter followed by seq. of alphanumerics
)	rightpar	a right parenthesis
;	semicolon	a semi-colon
writeln	writeln	w, r, i, t, e, l, n
(leftpar	a left parenthesis
'value read ='	literal ('value read =')	seq. of chars enclosed in quotes
,	comma	a comma
x	id (x)	letter followed by seq. of alphanumerics
)	rightpar	a right parenthesis
		newlines, spaces, tabs
end	end	e, n, d
.	fullstop	a fullstop

It is the sequence of tokens in the middle column that are passed as output to the syntax analyser. This token sequence represents almost all the important information from the input program required by the syntax analyser. Whitespace (newlines, spaces and tabs), although often important in separating lexemes, is usually not returned as a token. Also, when outputting an id or literal token, the lexical analyser must also return the value of the matched lexeme (shown in parentheses) or else this information would be lost.