

Principles of Programming Languages

<http://www.di.unipi.it/~andrea/Didattica/PLP-14/>

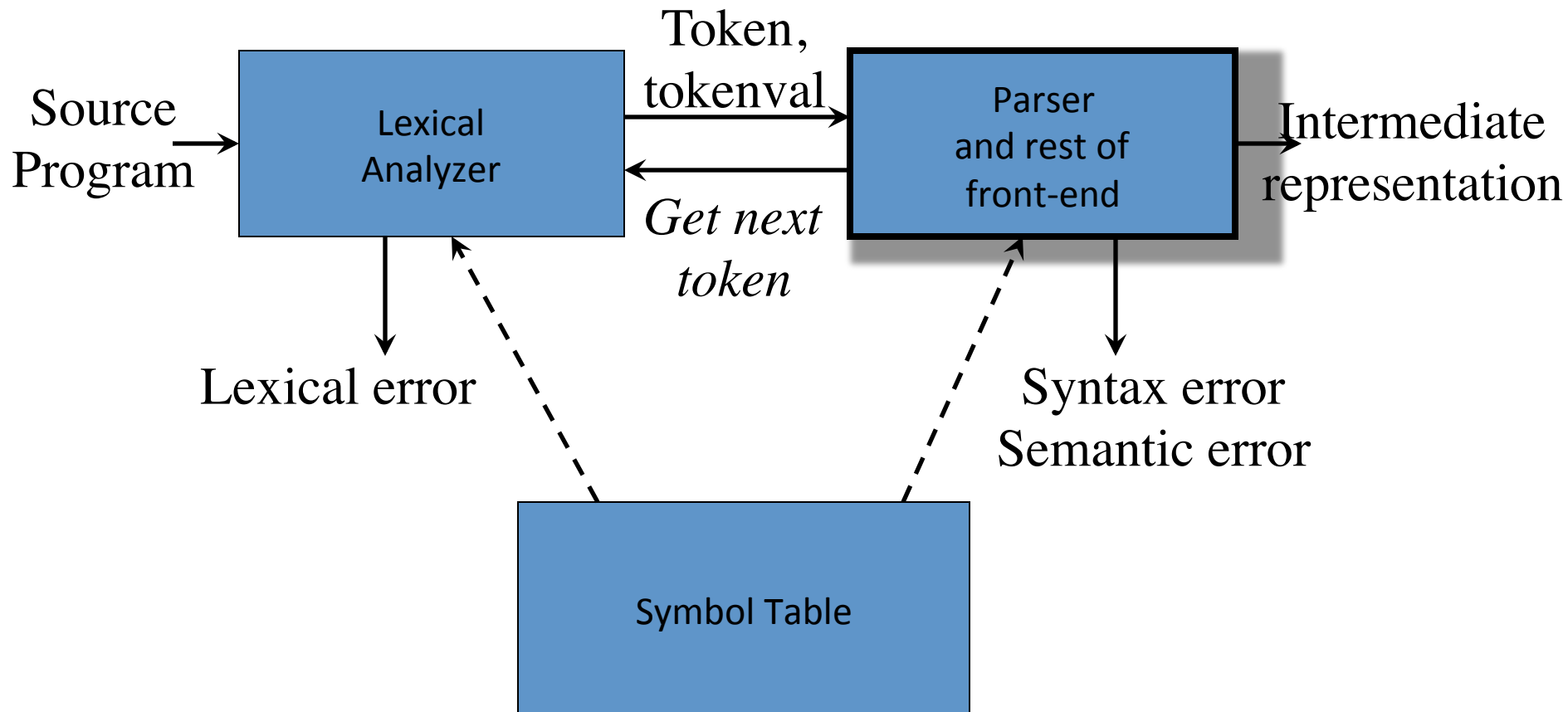
Prof. Andrea Corradini

Department of Computer Science, Pisa

Lesson 7

- Parsing: general concepts
- Top-down parsing, LL(1) grammars
- Descent recursive
- Predictive non-recursive
- Error Handling

Position of a Parser in the Compiler Model



The Parser

- A parser implements a C-F grammar as a recognizer of strings
- The role of the parser in a compiler is twofold:
 1. To check syntax (= string recognizer)
 - And to report syntax errors accurately
 2. To invoke semantic actions
 - For static semantics checking, e.g. type checking of expressions, functions, etc.
 - For syntax-directed translation of the source code to an intermediate representation
- Now we focus on 1.

Error Handling

- A good compiler should assist in identifying and locating errors
 - *Lexical errors*: compiler can easily recover and continue (e.g. misspelled identifiers)
 - *Syntax errors*: can almost always recover (e.g. missing ';' or '{', misplaced **case**)
 - *Static semantic errors*: can sometimes recover (e.g. type mismatches, variable used before declaration)
 - *Dynamic semantic errors*: hard or impossible to detect at compile time, runtime checks are required (e.g. null pointer, division by zero, invalid array access)
 - *Logical errors*: hard or impossible to detect (e.g. `if (b = true) ...`)

Viable-Prefix Property

- The *viable-prefix property* of parsers allows early detection of syntax errors
 - Enjoyed by LL(1), LR(1) parsers
 - Goal: detection of an error *as soon as possible* without further consuming unnecessary input
 - How: detect an error as soon as the prefix of the input does not match a prefix of any string in the language

Prefix {
...
for (;)
...
Error is detected here ↓

Prefix {
...
DO 10 I = 1 ; 0
...
Error is detected here ↓

Error Recovery Strategies

- *Panic mode*
 - Discard input until a token in a set of designated “synchronizing tokens” is found
- *Phrase-level recovery*
 - Perform local correction on the input to repair the error
- *Error productions*
 - Augment grammar with productions for erroneous constructs
- *Global correction*
 - Choose a minimal sequence of changes to obtain a global least-cost correction

Grammars (Recap)

- A grammar is a 4-tuple $G = (N, T, P, S)$ where
 - T is a finite set of tokens (*terminal* symbols)
 - N is a finite set of *nonterminals*
 - P is a finite set of *productions* of the form
$$\alpha \rightarrow \beta$$
where $\alpha \in (NUT)^* N (NUT)^*$ and $\beta \in (NUT)^*$
 - $S \in N$ is a designated *start symbol*

Notational Conventions Used

- Terminals
 $a, b, c, \dots \in T$
specific terminals: **0**, **1**, **id**, **+**
- Nonterminals
 $A, B, C, \dots \in N$
specific nonterminals: *expr*, *term*, *stmt*
- Grammar symbols
 $X, Y, Z \in (NUT)$
- Strings of terminals
 $u, v, w, x, y, z \in T^*$
- Strings of grammar symbols
 $\alpha, \beta, \gamma \in (NUT)^*$

Derivations (Recap)

- The *one-step derivation* is defined by
$$\gamma \alpha \delta \Rightarrow \gamma \beta \delta$$
where $\alpha \rightarrow \beta$ is a production in the grammar
- In addition, we define
 - \Rightarrow is *leftmost* \Rightarrow_{lm} if γ does not contain a nonterminal
 - \Rightarrow is *rightmost* \Rightarrow_{rm} if δ does not contain a nonterminal
 - Transitive closure \Rightarrow^* (zero or more steps)
 - Positive closure \Rightarrow^+ (one or more steps)
- The *language generated by G* is defined by
$$L(G) = \{w \in T^* \mid S \Rightarrow^+ w\}$$

Derivation (Example)

Grammar $G = (\{E\}, \{+, *, (,), -, \text{id}\}, P, E)$ with productions

$$P = E \rightarrow E + E$$

$$E \rightarrow E * E$$

$$E \rightarrow (E)$$

$$E \rightarrow - E$$

$$E \rightarrow \text{id}$$

Example derivations:

$$E \Rightarrow - E \Rightarrow - \text{id}$$

$$E \Rightarrow_{rm} E + E \Rightarrow_{rm} E + \text{id} \Rightarrow_{rm} \text{id} + \text{id}$$

$$E \Rightarrow^* E$$

$$E \Rightarrow^* \text{id} + \text{id}$$

$$E \Rightarrow^+ \text{id} * \text{id} + \text{id}$$

Chomsky Hierarchy: Language Classification

- A grammar G is said to be
 - *Regular* if it is *right linear* where each production is of the form
$$A \rightarrow w B \quad \text{or} \quad A \rightarrow w$$
or *left linear* where each production is of the form
$$A \rightarrow B w \quad \text{or} \quad A \rightarrow w$$
 - *Context free* if each production is of the form
$$A \rightarrow \alpha$$
where $A \in N$ and $\alpha \in (NUT)^*$
 - *Context sensitive* if each production is of the form
$$\alpha A \beta \rightarrow \alpha \gamma \beta$$
where $A \in N$, $\alpha, \gamma, \beta \in (NUT)^*$, $|\gamma| > 0$
 - *Unrestricted*

Chomsky Hierarchy

$\mathcal{L}(\text{regular}) \subset \mathcal{L}(\text{context free}) \subset \mathcal{L}(\text{context sensitive}) \subset \mathcal{L}(\text{unrestricted})$

Where $\mathcal{L}(T) = \{ L(G) \mid G \text{ is of type } T \}$

That is: the set of all languages
generated by grammars G of type T

Examples:

Every *finite language* is regular! (construct a FSA for strings in $L(G)$)

$L_1 = \{ \mathbf{a}^n \mathbf{b}^n \mid n \geq 1 \}$ is context free

$L_2 = \{ \mathbf{a}^n \mathbf{b}^n \mathbf{c}^n \mid n \geq 1 \}$ is context sensitive

Parsing

- *Universal* (any C-F grammar)
 - Cocke-Younger-Kasimi
 - Earley
- *Top-down* (C-F grammar with restrictions)
 - Recursive descent (predictive parsing)
 - LL (Left-to-right, Leftmost derivation) methods
- *Bottom-up* (C-F grammar with restrictions)
 - Operator precedence parsing
 - LR (Left-to-right, Rightmost derivation) methods
 - SLR, canonical LR, LALR

Top-Down Parsing

- LL methods (Left-to-right, Leftmost derivation) and recursive-descent parsing

Grammar:

$E \rightarrow T + T$

$T \rightarrow (E)$

$T \rightarrow - E$

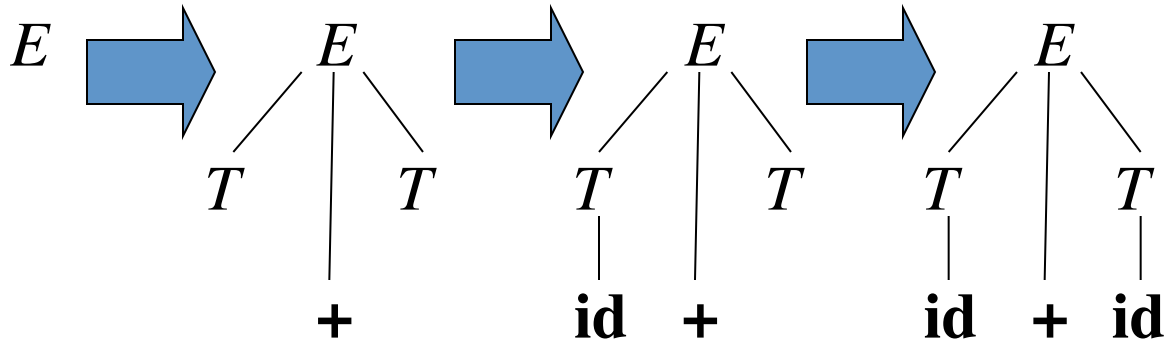
$T \rightarrow \mathbf{id}$

Leftmost derivation:

$E \Rightarrow_{lm} T + T$

$\Rightarrow_{lm} \mathbf{id} + T$

$\Rightarrow_{lm} \mathbf{id} + \mathbf{id}$



Left Recursion (Recap)

- Productions of the form
$$A \rightarrow A \alpha \mid \beta \mid \gamma$$
are left recursive
- Left recursion can be indirect
- When one of the productions in a grammar is (indirectly) left recursive, then a recursive descent or predictive parser loops forever on certain inputs

Immediate Left-Recursion Elimination

Rewrite every left-recursive production

$$A \rightarrow A \alpha$$

$$| \beta$$

$$| \gamma$$

$$| A \delta$$

into a right-recursive production:

$$A \rightarrow \beta A_R$$

$$| \gamma A_R$$

$$A_R \rightarrow \alpha A_R$$

$$| \delta A_R$$

$$| \varepsilon$$

A General Systematic Left Recursion Elimination Method

Input: Grammar G with no cycles or ε -productions

Arrange the nonterminals in some order A_1, A_2, \dots, A_n

for $i = 1, \dots, n$ **do**

for $j = 1, \dots, i-1$ **do**

 replace each

$$A_i \rightarrow A_j \gamma$$

 with

$$A_i \rightarrow \delta_1 \gamma \mid \delta_2 \gamma \mid \dots \mid \delta_k \gamma$$

 where

$$A_j \rightarrow \delta_1 \mid \delta_2 \mid \dots \mid \delta_k$$

enddo

 eliminate the *immediate left recursion* in A_i

enddo

Example Left Recursion Elim.

$$\left. \begin{array}{l} A \rightarrow B C | a \\ B \rightarrow C A | A b \\ C \rightarrow A B | C C | a \end{array} \right\} \text{Choose arrangement: } A, B, C$$

$i = 1$: nothing to do

$$\begin{aligned} i = 2, j = 1: \quad & B \rightarrow C A | \underline{A} b \\ & \Rightarrow B \rightarrow C A | \underline{B} \underline{C} b | \underline{a} b \\ & \Rightarrow_{(imm)} B \rightarrow C A B_R | a b B_R \\ & \quad B_R \rightarrow C b B_R | \varepsilon \end{aligned}$$

$$\begin{aligned} i = 3, j = 1: \quad & C \rightarrow \underline{A} B | C C | a \\ & \Rightarrow C \rightarrow \underline{B} \underline{C} B | \underline{a} B | C C | a \end{aligned}$$

$$\begin{aligned} i = 3, j = 2: \quad & C \rightarrow \underline{B} C B | a B | C C | a \\ & \Rightarrow C \rightarrow \underline{C} \underline{A} B_R C B | \underline{a} \underline{b} B_R C B | a B | C C | a \\ & \Rightarrow_{(imm)} C \rightarrow a b B_R C B C_R | a B C_R | a C_R \\ & \quad C_R \rightarrow A B_R C B C_R | C C_R | \varepsilon \end{aligned}$$

Left Factoring

- When a nonterminal has two or more productions whose right-hand sides start with the same grammar symbols, the grammar is not LL(1) and cannot be used for predictive parsing

- Replace productions

$$A \rightarrow \alpha \beta_1 \mid \alpha \beta_2 \mid \dots \mid \alpha \beta_n \mid \gamma$$

with

$$A \rightarrow \alpha A_R \mid \gamma$$

$$A_R \rightarrow \beta_1 \mid \beta_2 \mid \dots \mid \beta_n$$

Predictive Parsing

- Eliminate left recursion from grammar
- Left factor the grammar
- Compute FIRST and FOLLOW
- Two variants:
 - Recursive (recursive-descent parsing)
 - Non-recursive (table-driven parsing)

FIRST (Revisited)

- $\text{FIRST}(\alpha) = \{ \text{the set of terminals that begin all strings derived from } \alpha \}$
- $\text{FIRST}(a) = \{a\}$ if $a \in T$
- $\text{FIRST}(\varepsilon) = \{\varepsilon\}$
- $\text{FIRST}(A) = \bigcup_{A \rightarrow \alpha} \text{FIRST}(\alpha)$ for $A \rightarrow \alpha \in P$
- $\text{FIRST}(X_1 X_2 \dots X_k) =$
 - if** for all $j = 1, \dots, i-1 : \varepsilon \in \text{FIRST}(X_j)$ **then**
add non- ε in $\text{FIRST}(X_i)$ to $\text{FIRST}(X_1 X_2 \dots X_k)$
 - if** for all $j = 1, \dots, k : \varepsilon \in \text{FIRST}(X_j)$ **then**
add ε to $\text{FIRST}(X_1 X_2 \dots X_k)$

FOLLOW

- $\text{FOLLOW}(A) = \{ \text{the set of terminals that can immediately follow nonterminal } A \}$
- $\text{FOLLOW}(A) =$
 - for** all $(B \rightarrow \alpha A \beta) \in P$ **do**
add $\text{FIRST}(\beta) \setminus \{\epsilon\}$ to $\text{FOLLOW}(A)$
 - for** all $(B \rightarrow \alpha A \beta) \in P$ and $\epsilon \in \text{FIRST}(\beta)$ **do**
add $\text{FOLLOW}(B)$ to $\text{FOLLOW}(A)$
 - for** all $(B \rightarrow \alpha A) \in P$ **do**
add $\text{FOLLOW}(B)$ to $\text{FOLLOW}(A)$
 - if** A is the start symbol S **then**
add $\$$ to $\text{FOLLOW}(A)$

LL(1) Grammar

- A grammar G is LL(1) if it is not left recursive and for each collection of productions

$$A \rightarrow \alpha_1 \mid \alpha_2 \mid \dots \mid \alpha_n$$

for nonterminal A the following holds:

1. $\text{FIRST}(\alpha_i) \cap \text{FIRST}(\alpha_j) = \emptyset$ for all $i \neq j$
2. if $\alpha_j \Rightarrow^* \varepsilon$ then
 - 2.a. $\alpha_i \not\Rightarrow^* \varepsilon$ for all $i \neq j$
 - 2.b. $\text{FIRST}(\alpha_j) \cap \text{FOLLOW}(A) = \emptyset$
for all $i \neq j$

Non-LL(1) Examples

<i>Grammar</i>	<i>Not LL(1) because:</i>
$S \rightarrow S a \mid a$	Left recursive
$S \rightarrow a S \mid a$	$\text{FIRST}(a S) \cap \text{FIRST}(a) \neq \emptyset$
$S \rightarrow a R \mid \varepsilon$ $R \rightarrow S \mid \varepsilon$	For R : $S \Rightarrow^* \varepsilon$ and $\varepsilon \Rightarrow^* \varepsilon$
$S \rightarrow a R a$ $R \rightarrow S \mid \varepsilon$	For R : $\text{FIRST}(S) \cap \text{FOLLOW}(R) \neq \emptyset$

Recursive-Descent Parsing (Recap)

- Grammar must be LL(1)
- Every nonterminal has one (recursive) procedure responsible for parsing the nonterminal's syntactic category of input tokens
- When a nonterminal has multiple productions, each production is implemented in a branch of a selection statement based on input look-ahead information

Using FIRST and FOLLOW in a Recursive-Descent Parser

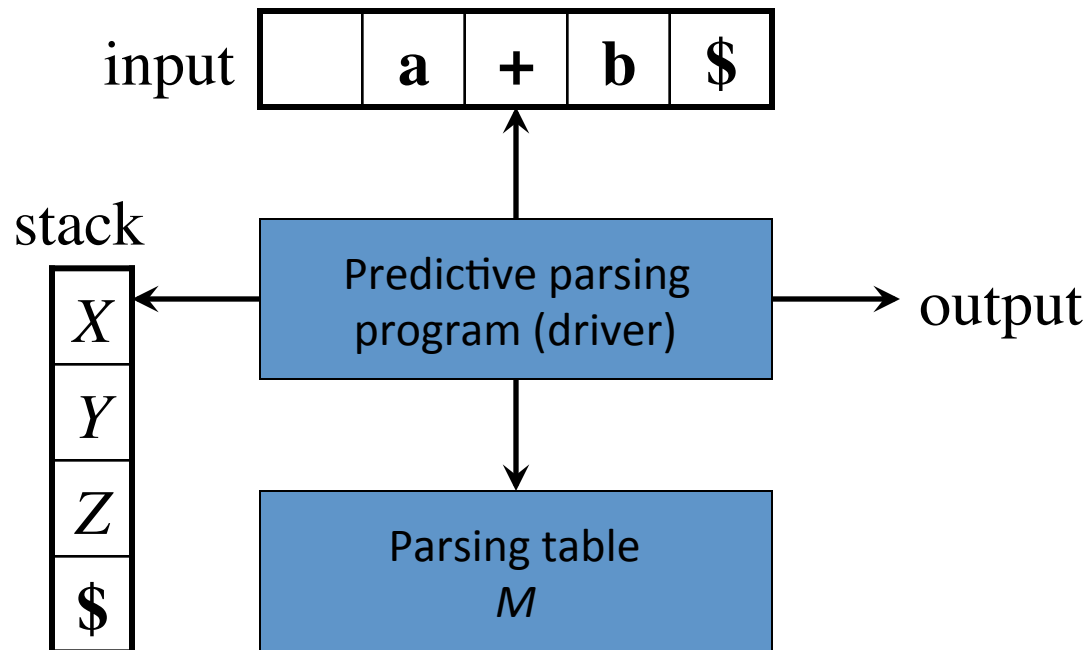
$expr \rightarrow term\ rest$
 $rest \rightarrow +\ term\ rest$
 $| -\ term\ rest$
 $| \epsilon$
 $term \rightarrow id$

```
procedure rest();  
begin  
  if lookahead in FIRST(+ term rest) then  
    match( '+' ); term(); rest()  
  else if lookahead in FIRST(- term rest) then  
    match( '-' ); term(); rest()  
  else if lookahead in FOLLOW(rest) then  
    return  
  else error()  
end;
```

where $FIRST(+\ term\ rest) = \{ + \}$
 $FIRST(-\ term\ rest) = \{ - \}$
 $FOLLOW(rest) = \{ \$ \}$

Non-Recursive Predictive Parsing: Table-Driven Parsing

- Given an LL(1) grammar $G = (N, T, P, S)$ construct a table $M[A, a]$ for $A \in N, a \in T$ and use a *driver program* with a *stack*

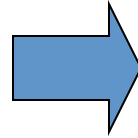


Constructing an LL(1) Predictive Parsing Table

```
for each production  $A \rightarrow \alpha$  do  
    for each  $a \in \text{FIRST}(\alpha)$  do  
        add production  $A \rightarrow \alpha$  to  $M[A,a]$   
    enddo  
    if  $\epsilon \in \text{FIRST}(\alpha)$  then  
        for each  $b \in \text{FOLLOW}(A)$  do  
            add  $A \rightarrow \alpha$  to  $M[A,b]$   
        enddo  
    endif  
enddo  
Mark each undefined entry in  $M$  error
```

Example Table

$E \rightarrow T E_R$
 $E_R \rightarrow + T E_R \mid \epsilon$
 $T \rightarrow F T_R$
 $T_R \rightarrow * F T_R \mid \epsilon$
 $F \rightarrow (E) \mid \mathbf{id}$



$A \rightarrow \alpha$	FIRST(α)	FOLLOW(A)
$E \rightarrow T E_R$	(id	\$)
$E_R \rightarrow + T E_R$	+	\$)
$E_R \rightarrow \epsilon$	ϵ	
$T \rightarrow F T_R$	(id	+ \$)
$T_R \rightarrow * F T_R$	*	+ \$)
$T_R \rightarrow \epsilon$	ϵ	
$F \rightarrow (E)$	(* + \$)
$F \rightarrow \mathbf{id}$	id	* + \$)



	id	+	*	()	\$
E	$E \rightarrow T E_R$			$E \rightarrow T E_R$		
E_R		$E_R \rightarrow + T E_R$			$E_R \rightarrow \epsilon$	$E_R \rightarrow \epsilon$
T	$T \rightarrow F T_R$			$T \rightarrow F T_R$		
T_R		$T_R \rightarrow \epsilon$	$T_R \rightarrow * F T_R$		$T_R \rightarrow \epsilon$	$T_R \rightarrow \epsilon$
F	$F \rightarrow \mathbf{id}$			$F \rightarrow (E)$		

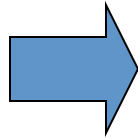
LL(1) Grammars are Unambiguous

Ambiguous grammar

$$S \rightarrow i E t S S_R \mid a$$

$$S_R \rightarrow e S \mid \epsilon$$

$$E \rightarrow b$$



$A \rightarrow \alpha$	FIRST(α)	FOLLOW(A)
$S \rightarrow i E t S S_R$	i	e \$
$S \rightarrow a$	a	
$S_R \rightarrow e S$	e	e \$
$S_R \rightarrow \epsilon$	ϵ	
$E \rightarrow b$	b	t

Error: duplicate table entry

	a	b	e	i	t	\$
S	$S \rightarrow a$			$S \rightarrow i E t S S_R$		
S_R			$S_R \rightarrow \epsilon$ $S_R \rightarrow e S$			$S_R \rightarrow \epsilon$
E		$E \rightarrow b$				

Predictive Parsing Program (Driver)

```
push($)  
push(S)  
a := lookahead  
repeat  
     $X := \text{pop}()$   
    if  $X$  is a terminal or  $X = \$$  then  
        match(X) // moves to next token and a := lookahead  
    else if  $M[X,a] = X \rightarrow Y_1 Y_2 \dots Y_k$  then  
        push( $Y_k, Y_{k-1}, \dots, Y_2, Y_1$ ) // such that  $Y_1$  is on top  
        ... invoke actions and/or produce IR output ...  
    else    error()  
    endif  
until  $X = \$$ 
```

Example Table-Driven Parsing

Stack	Input	Production applied
$\$E$	$\underline{\text{id}}+\text{id}*\text{id}\$$	$E \rightarrow T E_R$
$\$E_R T$	$\underline{\text{id}}+\text{id}*\text{id}\$$	$T \rightarrow F T_R$
$\$E_R T_R F$	$\underline{\text{id}}+\text{id}*\text{id}\$$	$F \rightarrow \text{id}$
$\$E_R T_R \underline{\text{id}}$	$\underline{\text{id}}+\text{id}*\text{id}\$$	
$\$E_R T_R$	$\underline{+}\text{id}*\text{id}\$$	$T_R \rightarrow \epsilon$
$\$E_R$	$\underline{+}\text{id}*\text{id}\$$	$E_R \rightarrow + T E_R$
$\$E_R T \underline{+}$	$\underline{+}\text{id}*\text{id}\$$	
$\$E_R T$	$\underline{\text{id}}*\text{id}\$$	$T \rightarrow F T_R$
$\$E_R T_R F$	$\underline{\text{id}}*\text{id}\$$	$F \rightarrow \text{id}$
$\$E_R T_R \underline{\text{id}}$	$\underline{\text{id}}*\text{id}\$$	
$\$E_R T_R$	$\underline{*}\text{id}\$$	$T_R \rightarrow * F T_R$
$\$E_R T_R F \underline{*}$	$\underline{*}\text{id}\$$	
$\$E_R T_R F$	$\underline{\text{id}}\$$	$F \rightarrow \text{id}$
$\$E_R T_R \underline{\text{id}}$	$\underline{\text{id}}\$$	
$\$E_R T_R$	$\underline{\$}$	$T_R \rightarrow \epsilon$
$\$E_R$	$\underline{\$}$	$E_R \rightarrow \epsilon$
$\underline{\$}$	$\underline{\$}$	

Panic Mode Recovery

Add synchronizing actions to undefined entries based on FOLLOW

$\text{FOLLOW}(E) = \{) \$ \}$
 $\text{FOLLOW}(E_R) = \{) \$ \}$
 $\text{FOLLOW}(T) = \{ +) \$ \}$
 $\text{FOLLOW}(T_R) = \{ +) \$ \}$
 $\text{FOLLOW}(F) = \{ + *) \$ \}$

Pro: Can be automated
 Cons: Error messages are needed

	id	+	*	()	\$
E	$E \rightarrow T E_R$			$E \rightarrow T E_R$	<i>synch</i>	<i>synch</i>
E_R		$E_R \rightarrow + T E_R$			$E_R \rightarrow \epsilon$	$E_R \rightarrow \epsilon$
T	$T \rightarrow F T_R$	<i>synch</i>		$T \rightarrow F T_R$	<i>synch</i>	<i>synch</i>
T_R		$T_R \rightarrow \epsilon$	$T_R \rightarrow * F T_R$		$T_R \rightarrow \epsilon$	$T_R \rightarrow \epsilon$
F	$F \rightarrow \text{id}$	<i>synch</i>	<i>synch</i>	$F \rightarrow (E)$	<i>synch</i>	<i>synch</i>

synch: the driver pops current nonterminal A and skips input till *synch* token or skips input until one of $\text{FIRST}(A)$ is found

Phrase-Level Recovery

Change input stream by inserting missing tokens

For example: **id id** is changed into **id * id**

Pro: Can be fully automated

Cons: Recovery not always intuitive

Can then continue here

	id	+	*	()	\$
E	$E \rightarrow T E_R$			$E \rightarrow T E_R$	<i>synch</i>	<i>synch</i>
E_R		$E_R \rightarrow + T E_R$			$E_R \rightarrow \epsilon$	$E_R \rightarrow \epsilon$
T	$T \rightarrow F T_R$	<i>synch</i>		$T \rightarrow F T_R$	<i>synch</i>	<i>synch</i>
T_R	<i>insert *</i>	$T_R \rightarrow \epsilon$	$T_R \rightarrow * F T_R$		$T_R \rightarrow \epsilon$	$T_R \rightarrow \epsilon$
F	$F \rightarrow \mathbf{id}$	<i>synch</i>	<i>synch</i>	$F \rightarrow (E)$	<i>synch</i>	<i>synch</i>

*insert **: driver inserts missing * and retries the production

Error Productions

$$\begin{aligned}
 E &\rightarrow T E_R \\
 E_R &\rightarrow + T E_R \mid \varepsilon \\
 T &\rightarrow F T_R \\
 T_R &\rightarrow * F T_R \mid \varepsilon \\
 F &\rightarrow (E) \mid \mathbf{id}
 \end{aligned}$$

Add “*error production*”:

$$T_R \rightarrow F T_R$$

to ignore missing *, e.g.: **id id**

Pro: Powerful recovery method

Cons: Manual addition of productions

	id	+	*	()	\$
E	$E \rightarrow T E_R$			$E \rightarrow T E_R$	<i>synch</i>	<i>synch</i>
E_R		$E_R \rightarrow + T E_R$			$E_R \rightarrow \varepsilon$	$E_R \rightarrow \varepsilon$
T	$T \rightarrow F T_R$	<i>synch</i>		$T \rightarrow F T_R$	<i>synch</i>	<i>synch</i>
T_R	$T_R \rightarrow F T_R$	$T_R \rightarrow \varepsilon$	$T_R \rightarrow * F T_R$		$T_R \rightarrow \varepsilon$	$T_R \rightarrow \varepsilon$
F	$F \rightarrow \mathbf{id}$	<i>synch</i>	<i>synch</i>	$F \rightarrow (E)$	<i>synch</i>	<i>synch</i>