Parsing

A.k.a. Syntax Analysis

- Recognize *sentences* in a language.
- Discover the structure of a document/program.
- Construct (implicitly or explicitly) a tree (called as a parse tree) to represent the structure.
- The above tree is used later to guide the translation.

Grammars

The syntactic structure of a language is defined using *grammars*.

- Grammars (like regular expressions) specify a set of strings over an alphabet.
- Efficient *recognizers* (like DFA) can be constructed to efficiently determine whether a string is in the language.
- Language hierarchy:
  - Finite Languages (FL)

    Enumeration
  - Regular Languages (RL $\supset$ FL)

    Regular Expressions
  - Context-free Languages (CFL $\supset$ RL)

    Context-free Grammars

Regular Languages

Languages represented by regular expressions	$\equiv$	Languages recognized by finite automata

---

Examples:

- $\surd$ $\{a, b, c\}$
- $\surd$ $\{\epsilon, a, b, aa, ab, ba, bb, \ldots\}$
- $\surd$ $\{(ab)^n \mid n \geq 0\}$
- $\times$ $\{a^n b^n \mid n \geq 0\}$

Grammars

Notation where recursion is explicit. Examples

- $\{\epsilon, \mathrm{a}, \mathrm{b}, \mathrm{aa}, \mathrm{ab}, \mathrm{ba}, \mathrm{bb}, \ldots\}$:

$$\begin{array}{lcl} E & \longrightarrow & \mathrm{a} \\ E & \longrightarrow & \mathrm{b} \\ S & \longrightarrow & \epsilon \\ S & \longrightarrow & ES \end{array}$$

  Notational shorthand:

$$\begin{array}{lcl} E & \longrightarrow & \mathrm{a} \mid \mathrm{b} \\ S & \longrightarrow & \epsilon \mid ES \end{array}$$

- $\{\mathrm{a}^n\mathrm{b}^n \mid n \geq 0\}$:

$$\begin{array}{lcl} S & \longrightarrow & \epsilon \\ S & \longrightarrow & \mathrm{a}S\mathrm{b} \end{array}$$

- $\{w \mid \textit{no. of } \mathrm{a}\textit{'s in } w = \textit{no. of } \mathrm{b}\textit{'s in } w\}$

Context-free Grammars

- Terminal Symbols: Tokens
- Nonterminal Symbols: set of strings made up of tokens
- Productions: Rules for constructing the set of strings associated with non-terminal symbols.

  Example: $\mathit{Stmt} \longrightarrow \text{while}\ \mathit{Expr}\ \text{do}\ \mathit{Stmt}$

Start symbol: nonterminal symbol that represents the set of all strings in the language.

Example

$$
\begin{array}{lcl}
E & \longrightarrow & E + E \\
E & \longrightarrow & E - E \\
E & \longrightarrow & E * E \\
E & \longrightarrow & E \,/\, E \\
E & \longrightarrow & (\ E\ ) \\
E & \longrightarrow & \mathrm{id}
\end{array}
$$

$$\mathcal{L}(E) = \{\mathrm{id}, \mathrm{id} + \mathrm{id}, \mathrm{id} - \mathrm{id}, \ldots, \mathrm{id} + (\mathrm{id} * \mathrm{id}) - \mathrm{id}, \ldots\}$$

Context-free Grammars

Production: rule with *non-terminal* symbol on left hand side, and a (possibly empty) sequence of terminal or non-terminal symbols on the right-hand side.

Notations:

- Terminals: lower case letters, digits, punctuation
- Nonterminals: Upper case letters
- Arbitrary Terminals/Nonterminals: X, Y, Z
- Strings of Terminals: u, v, w
- Strings of Terminals/Nonterminals: α, β, γ
- Start Symbol: S

Context-Free Vs Other Types of Grammars

- Context-free grammar (CFG): Productions of the form $NT \longrightarrow [NT|T]*$
- Context-sensitive grammar (CSG): Productions of the form $[t|NT]*NT[t|NT]* \longrightarrow [t|NT]*$
- Unrestricted grammar: Productions of the form $[t|NT]* \longrightarrow [t|NT]*$

Examples of Non-Context-Free Languages

- Checking that variables are declared before use. If we simplify and abstract the problem, we see that it amounts to recognizing strings of the form wsw
- Checking whether the number of actual and formal parameters match. Abstracts to recognizing strings of the form $a^n b^m c^n d^m$
- In both cases, the rules are not enforced in grammar but deferred to type-checking phase
- Note: Strings of the form wsw^R and $a^n b^n c^m d^m$ can be described by a CFG

What types of Grammars Describe These Languages?

- Strings of 0's and 1's of form xx
- Strings of 0's and 1's in which 011 doesn't occur
- Strings of 0's and 1's in which each 0 is immediately followed by a 1
- Strings of 0's and 1's with ithe equal number of 0's and 1's.

Language Generated by Grammars, Equivalence of Grammars

- How to show that a grammar G generates a language $\mathcal{M}$? Show that
  - $\forall s \in \mathcal{M}$, show that $s \in \mathcal{L}(G)$
  - $\forall s \in \mathcal{L}(G)$, show that $s \in \mathcal{M}$
- How to establish that two grammars G_1 and G_2 are equivalent?

  Show that $\mathcal{L}(G_1) = \mathcal{L}(G_2)$

Grammar Examples

$$S \longrightarrow 0S1S \mid 1S0S \mid \epsilon$$

What is the language generated by this grammar?

Grammar Examples

$$S \longrightarrow 0A|1B|\epsilon$$

$$A \longrightarrow 0AA|1S$$

$$B \longrightarrow 1BB|0S$$

What is the language generated by this grammar?

The Two Sides of Grammars

Specify a set of strings in a language.

Recognize strings in a given language:

- Is a given string x in the language?

  Yes, if we can construct a *derivation* for x

- Example: Is $\mathrm{id} + \mathrm{id} \in \mathcal{L}(E)$?

$$\begin{aligned} \mathrm{id} + \mathrm{id} &\Longleftarrow E + \mathrm{id} \\ &\Longleftarrow E + E \\ &\Longleftarrow E \end{aligned}$$

Derivations

Grammar:
$$\begin{array}{lcl} E & \longrightarrow & E + E \\ E & \longrightarrow & \mathrm{id} \end{array}$$

E **derives** id + id:
$$\begin{array}{lcl} E & \Longrightarrow & E + E \\ & \Longrightarrow & E + \mathrm{id} \\ & \Longrightarrow & \mathrm{id} + \mathrm{id} \end{array}$$

- $\alpha A \beta \Longrightarrow \alpha \gamma \beta$ iff $A \longrightarrow \gamma$ is a production in the grammar.
- $\alpha \overset{*}{\Longrightarrow} \beta$ if α derives β in zero or more steps.

  Example: $E \overset{*}{\Longrightarrow} \mathrm{id} + \mathrm{id}$
- Sentence: A sequence of terminal symbols w such that $S \overset{+}{\Longrightarrow} w$ (where S is the start symbol)
- Sentential Form: A sequence of terminal/nonterminal symbols α such that $S \overset{*}{\Longrightarrow} \alpha$

Derivations

- Rightmost derivation: Rightmost non-terminal is replaced first:

$$\begin{aligned} E &\Longrightarrow E + E \\ &\Longrightarrow E + \mathrm{id} \\ &\Longrightarrow \mathrm{id} + \mathrm{id} \end{aligned}$$

  Written as $E \overset{*}{\Longrightarrow}_{rm} \mathrm{id} + \mathrm{id}$

- Leftmost derivation: Leftmost non-terminal is replaced first:

$$\begin{aligned} E &\Longrightarrow E + E \\ &\Longrightarrow \mathrm{id} + E \\ &\Longrightarrow \mathrm{id} + \mathrm{id} \end{aligned}$$

  Written as $E \overset{*}{\Longrightarrow}_{lm} \mathrm{id} + \mathrm{id}$

Parse Trees

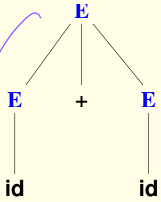
$E \rightarrow \mathrm{id}$

Graphical Representation of Derivations

Top-down

$$\begin{aligned} E &\Longrightarrow E + E \\ &\Longrightarrow \mathrm{id} + E \\ &\Longrightarrow \mathrm{id} + \mathrm{id} \end{aligned}$$

LL



$$\begin{aligned} E &\Longrightarrow E + E \\ &\Longrightarrow E + \mathrm{id} \\ &\Longrightarrow \mathrm{id} + \mathrm{id} \end{aligned}$$

Bottom up

LR

A **Parse Tree** succinctly captures the *structure* of a sentence.

Recursive descent → recursive, backtracks

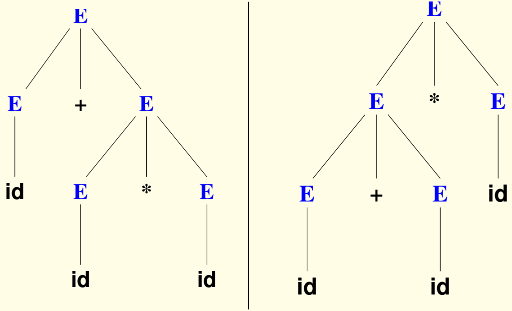
Predictive parsing LL

Antlr

Ambiguity

A Grammar is *ambiguous* if there are multiple parse trees for the same sentence.

Example: $\mathrm{id} + \mathrm{id} * \mathrm{id}$



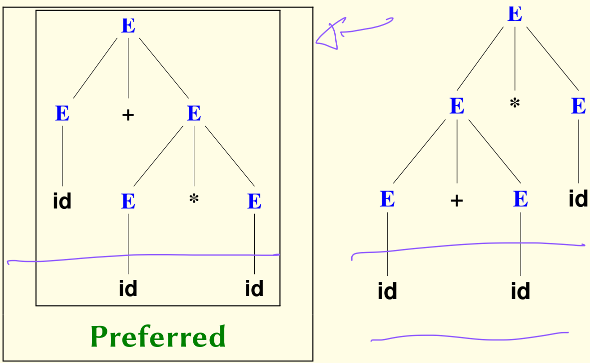
Disambiguition

Express Preference for one parse tree over others.

5ab

Example: id + id $*$ id

The usual precedence of $*$ over $+$ means:



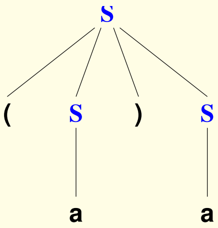
Preferred

Parsing

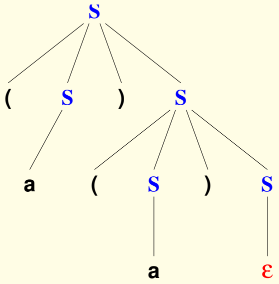
Construct a parse tree for a given string.

$$\begin{aligned} S &\longrightarrow (S)S \\ S &\longrightarrow a \\ S &\longrightarrow \epsilon \end{aligned}$$

(a)a



(a)(a)



A Procedure for Parsing

Grammar: $S \longrightarrow a$

→ Top-down

→ Bottom-up

```
procedure parse_S() {
    switch (input_token) {
        case TOKEN_a:
            consume(TOKEN_a);
            return;
        default:
            /* Parse Error */
    }
}
```

Predictive Parsing

Grammar:

$$\begin{aligned} S &\longrightarrow a \\ S &\longrightarrow \epsilon \end{aligned}$$

```
procedure parse_S() {
    switch (input_token) {
        case TOKEN_a:   /* Production 1 */
            consume(TOKEN_a);
            return;
        case TOKEN_EOF:   /* Production 2 */
            return;
        default:
            /* Parse Error */
    }
}
```

Predictive Parsing (contd.)

Grammar:

$$S \longrightarrow (S)S$$
$$S \longrightarrow a$$
$$S \longrightarrow \epsilon$$

```
procedure parse_S() {
    switch (input_token) {
        case TOKEN_OPEN_PAREN:   /* Production 1 */
            consume(TOKEN_OPEN_PAREN);
            parse_S();
            consume(TOKEN_CLOSE_PAREN);
            parse_S();
            return;
```

$S \to A \mid B$

parse_S()

parse_A()

parse_B()

LL(1)

LL(k)

LR(1)

Predictive Parsing (contd.)

Grammar:

$$\begin{array}{lcl} S & \longrightarrow & (S)S \\ S & \longrightarrow & a \\ S & \longrightarrow & \epsilon \end{array}$$

```
case TOKEN_a:   /* Production 2 */
    consume(TOKEN_a);
    return;
case TOKEN_CLOSE_PAREN:
case TOKEN_EOF:   /* Production 3 */
    return;
default:
    /* Parse Error */
```

Predictive Parsing: Restrictions

Grammar cannot be left-recursive

Example: $E \longrightarrow E + E \mid a$

```
procedure parse_E() {
    switch (input_token) {
        case TOKEN_a:   /* Production 1 */
            parse_E();
            consume(TOKEN_PLUS);
            parse_E();
            return;
        case TOKEN_a:   /* Production 2 */
            consume(TOKEN_a);
            return;
    }
}
```

Removing Left Recursion

$$\begin{aligned} A &\longrightarrow A\ a \\ A &\longrightarrow b \end{aligned}$$

$$\mathcal{L}(A) = \{b, ba, baa, baaa, baaaa, \ldots\}$$

$$A \longrightarrow bA'$$

$$\begin{aligned} A' &\longrightarrow aA' \\ A' &\longrightarrow \epsilon \end{aligned}$$

Removing Left Recursion

More generally,

$$\begin{aligned} A &\longrightarrow A\alpha_1 | \cdots | A\alpha_m \\ A &\longrightarrow \beta_1 | \cdots | \beta_n \end{aligned}$$

Can be transformed into

$$\begin{aligned} A &\longrightarrow \beta_1 A' | \cdots | \beta_n A' \\ A' &\longrightarrow \alpha_1 A' | \cdots | \alpha_m A' | \epsilon \end{aligned}$$

Removing Left Recursion: An Example

$$\begin{aligned} E &\longrightarrow E + E \\ E &\longrightarrow \mathrm{id} \end{aligned}$$

$$\Downarrow$$

$$E \longrightarrow \mathrm{id}\ E'$$

$$\begin{aligned} E' &\longrightarrow +\ E\ E' \\ E' &\longrightarrow \epsilon \end{aligned}$$

Predictive Parsing: Restrictions

May not be able to choose a *unique* production

$$\begin{aligned} S &\longrightarrow a\ B\ d \\ B &\longrightarrow b \\ B &\longrightarrow bc \end{aligned}$$

Left-factoring can help:

$$\begin{aligned} S &\longrightarrow a\ B\ d \\ B &\longrightarrow bC \\ C &\longrightarrow c|\epsilon \end{aligned}$$

Predictive Parsing: Restrictions

In general, though, we may need a backtracking parser:

Recursive Descent Parsing

$$\begin{array}{lcl} S & \longrightarrow & a\ B\ d \\ B & \longrightarrow & b \\ B & \longrightarrow & bc \end{array}$$

Recursive Descent Parsing

Grammar:

$$\begin{array}{lcl} S & \longrightarrow & a\ B\ d \\ B & \longrightarrow & b \\ B & \longrightarrow & bc \end{array}$$

```
procedure parse_B() {
    switch (input_token) {
        case TOKEN_b:   /* Production 2 */
            consume(TOKEN_b);
            return;
        case TOKEN_b:   /* Production 3 */
            consume(TOKEN_b);
            consume(TOKEN_c);
            return;
}}
```

Non-recursive Parsing

Instead of recursion,

use an explicit *stack* along with the parsing table.

Data objects:

- **Parsing Table**: $M(A, a)$, a two-dimensional array, dimensions indexed by nonterminal symbols (A) and terminal symbols (a).
- A **Stack** of terminal/nonterminal symbols
- **Input stream** of tokens

The above data structures manipulated using a *table-driven parsing program.*

Table-driven Parsing

Grammar:

$$\begin{array}{lcl} A & \longrightarrow & a \\ B & \longrightarrow & b \end{array} \qquad \begin{array}{lcl} S & \longrightarrow & A\ S\ B \\ S & \longrightarrow & \epsilon \end{array}$$

Parsing Table:

NONTERMINAL	INPUT SYMBOL		
	a	b	EOF
S	$S \longrightarrow A\ S\ B$	$S \longrightarrow \epsilon$	$S \longrightarrow \epsilon$
A	$A \longrightarrow a$		
B		$B \longrightarrow b$	

Table-driven Parsing Algorithm

```
stack initialized to EOF.
while (stack is not empty) {
    X = top(stack);
    if (X is a terminal symbol)
        consume(X);
    else /* X is a nonterminal */
        if ($M[X, input\_token] = X \longrightarrow Y_1, Y_2, \ldots, Y_k$) {
            pop(stack);
            for $i = k$ downto 1 do
                push(stack, $Y_i$);
        }
        else /* Syntax Error */
}
```

FIRST and FOLLOW

Grammar: $S \longrightarrow (S)S \mid a \mid \epsilon$

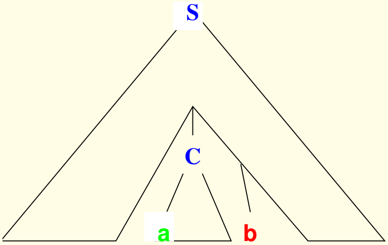
- FIRST(X) = First character of any string that can be derived from X

  $\text{FIRST}(S) = \{(, a, \epsilon\}.$

- FOLLOW(A) = First character that, in any derivation of a string in the language, appears immediately after A.

  $\text{FOLLOW}(S) = \{), \text{EOF}\}$

FIRST and FOLLOW (contd.)



$$a \in \text{FIRST}(C)$$
$$b \in \text{FOLLOW}(C)$$

FIRST and FOLLOW

FIRST(X): First terminal in some α such that $X \overset{*}{\Longrightarrow} \alpha$.

FOLLOW(A): First terminal in some β such that $S \overset{*}{\Longrightarrow} \alpha A \beta$.

Grammar:

$$\begin{array}{lll lll} A & \longrightarrow & a & S & \longrightarrow & A\ S\ B \\ B & \longrightarrow & b & S & \longrightarrow & \epsilon \end{array}$$

$$\begin{array}{lcl lcl} \textit{First}(S) & = & \{\,\mathrm{a}, \epsilon\,\} & \textit{Follow}(S) & = & \{\,\mathrm{b}, \mathrm{EOF}\,\} \\ \textit{First}(A) & = & \{\,\mathrm{a}\,\} & \textit{Follow}(A) & = & \{\,\mathrm{a}, \mathrm{b}\,\} \\ \textit{First}(B) & = & \{\,\mathrm{b}\,\} & \textit{Follow}(B) & = & \{\,\mathrm{b}, \mathrm{EOF}\,\} \end{array}$$

Definition of FIRST

Grammar:

$$\begin{array}{lllllll} A & \longrightarrow & a & \quad & S & \longrightarrow & A\ S\ B \\ B & \longrightarrow & b & \quad & S & \longrightarrow & \epsilon \end{array}$$

$FIRST(\alpha)$ is the smallest set such that

$\alpha =$	Property of $FIRST(\alpha)$
a, a terminal	$a \in FIRST(\alpha)$
A, a nonterminal	$A \longrightarrow \epsilon \in G \Longrightarrow \epsilon \in FIRST(\alpha)$ $A \longrightarrow \beta \in G, \beta \neq \epsilon \Longrightarrow FIRST(\beta) \subseteq FIRST(\alpha)$
$X_1 X_2 \cdots X_k$, a string of terminals and non-terminals	$FIRST(X_1) - \{\epsilon\} \subseteq FIRST(\alpha)$ $FIRST(X_i) \subseteq FIRST(\alpha)$ if $\forall j < i \quad \epsilon \in FIRST(X_j)$ $\epsilon \in FIRST(\alpha)$ if $\forall j < k \quad \epsilon \in FIRST(X_j)$

Definition of FOLLOW

Grammar:

$$\begin{array}{lll@{\qquad}lll} A & \longrightarrow & a & S & \longrightarrow & A\ S\ B \\ B & \longrightarrow & b & S & \longrightarrow & \epsilon \end{array}$$

$FOLLOW(A)$ is the smallest set such that

A	Property of $FOLLOW(A)$
$= S$, the start symbol	EOF $\in FOLLOW(S)$ Book notation: $\$ \in FOLLOW(S)$
$B \longrightarrow \alpha A \beta \in G$	$FIRST(\beta) - \{\epsilon\} \subseteq FOLLOW(A)$
$B \longrightarrow \alpha A$, or $B \longrightarrow \alpha A \beta$, $\epsilon \in FIRST(\beta)$	$FOLLOW(B) \subseteq FOLLOW(A)$

A Procedure to Construct Parsing Tables

```
procedure $table\_construct(G)$ {
    for each $A \longrightarrow \alpha \in G$ {
        for each $a \in FIRST(\alpha)$ such that $a \neq \epsilon$
            add $A \longrightarrow \alpha$ to $M[A, a]$;
        if $\epsilon \in FIRST(\alpha)$
            for each $b \in FOLLOW(A)$
                add $A \longrightarrow \alpha$ to $M[A, b]$;
}}
```

LL(1) Grammars

Grammars for which the parsing table constructed earlier has no multiple entries.

$$\begin{array}{lcl} E & \longrightarrow & \mathrm{id}\ E' \\ E' & \longrightarrow & +\ E\ E' \\ E' & \longrightarrow & \epsilon \end{array}$$

NONTERMINAL	INPUT SYMBOL		
	id	+	EOF
E	$E \longrightarrow \mathrm{id}\ E'$		
E'		$E' \longrightarrow +\ E\ E'$	$E' \longrightarrow \epsilon$

Parsing with LL(1) Grammars

Nonterminal	Input Symbol		
	id	+	EOF
E	$E \longrightarrow \text{id}\ E'$		
E'		$E' \longrightarrow +\ E\ E'$	$E' \longrightarrow \epsilon$

$\$E$	$\text{id} + \text{id}\$$	E	$\Longrightarrow$	$\text{id}E'$
$\$E'\text{id}$	$\text{id} + \text{id}\$$			
$\$E'$	$+ \text{id}\$$		$\Longrightarrow$	$\text{id}{+}EE'$
$\$E'E{+}$	$+ \text{id}\$$			
$\$E'E$	$\text{id}\$$		$\Longrightarrow$	$\text{id}{+}\text{id}E'E'$
$\$E'E'\text{id}$	$\text{id}\$$			
$\$E'E'$	$\$$		$\Longrightarrow$	$\text{id}{+}\text{id}E'$
$\$E'$	$\$$		$\Longrightarrow$	$\text{id}{+}\text{id}$
$\$$	$\$$			

LL(1) Derivations

Left to Right Scan of input

Leftmost Derivation

(1) look ahead 1 token at each step

Alternative characterization of LL(1) Grammars:

Whenever $A \longrightarrow \alpha \mid \beta \in G$

1. $\mathit{FIRST}(\alpha) \cap \mathit{FIRST}(\beta) = \{\ \}$, and
2. if $\alpha \overset{*}{\Longrightarrow} \epsilon$ then $\mathit{FIRST}(\beta) \cap \mathit{FOLLOW}(A) = \{\ \}$.

Corollary: No Ambiguous Grammar is LL(1).

Leftmost and Rightmost Derivations

$$
\begin{array}{lcl}
\hline
E & \longrightarrow & E+T \\
E & \longrightarrow & T \\
T & \longrightarrow & \mathrm{id} \\
\hline
\end{array}
$$

Derivations for id + id:

$$
\begin{array}{lcl|lcl}
E & \Longrightarrow & E+T & E & \Longrightarrow & E+T \\
 & \Longrightarrow & T+T & & \Longrightarrow & E+\mathrm{id} \\
 & \Longrightarrow & \mathrm{id}+T & & \Longrightarrow & T+\mathrm{id} \\
 & \Longrightarrow & \mathrm{id}+\mathrm{id} & & \Longrightarrow & \mathrm{id}+\mathrm{id} \\
\end{array}
$$

LEFTMOST | RIGHTMOST

Bottom-up Parsing

Given a stream of tokens w, *reduce* it to the start symbol.

$$\begin{array}{lcl} E & \longrightarrow & E+T \\ E & \longrightarrow & T \\ T & \longrightarrow & \mathrm{id} \end{array}$$

Parse input stream: id + id:

$$\begin{array}{c} \mathrm{id} + \mathrm{id} \\ T + \mathrm{id} \\ E + \mathrm{id} \\ E + T \\ E \end{array}$$

Reduction ≡ Derivation^{-1}.

Handles

$A + \boxed{A + id} + id$

$A \to \boxed{A + id}$

Informally, a “handle” of a sentential form is a substring that matches the right side of a production, and

whose reduction to the non-terminal on the left hand side of the production represents one step along the reverse rightmost derivation.

Handles

A structure that furnishes a means to perform reductions.

$$\begin{array}{ccc} E & \longrightarrow & E+T \\ E & \longrightarrow & T \\ T & \longrightarrow & \text{id} \end{array}$$

Parse input stream: id + id:

$$\begin{array}{c} \boxed{\text{id}} + \text{id} \\ \boxed{T} + \text{id} \\ E + \boxed{\text{id}} \\ \boxed{E + T} \\ E \end{array}$$

Handles

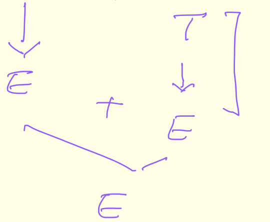
Handles are substrings of sentential forms:

1. A substring that matches the right hand side of a production
2. Reduction using that rule can lead to the start symbol

$$\begin{aligned} E &\implies \boxed{E + T} \\ &\implies E + \boxed{\text{id}} \\ &\implies \boxed{T} + \text{id} \\ &\implies \boxed{\text{id}} + \text{id} \end{aligned}$$

$T \to T * T$

id + id * id



Handle Pruning: replace handle by corresponding LHS.

Shift-Reduce Parsing

Bottom-up parsing.

- **Shift:** Construct leftmost handle on top of stack
- **Reduce:** Identify handle and replace by corresponding RHS
- **Accept:** Continue until string is reduced to start symbol and input token stream is empty
- **Error:** Signal parse error if no handle is found.

Implementing Shift-Reduce Parsers

and nonterminals

- Stack to hold grammar symbols (corresponding to tokens seen thus far).

terminals

- Input stream of yet-to-be-seen tokens.
- Handles appear on top of stack.
- Stack is initially empty (denoted by \$).
- Parse is successful if stack contains only the start symbol when the input stream ends.

Shift-Reduce Parsing: An Example

$$
\begin{array}{lcl}
S & \longrightarrow & aABe \\
A & \longrightarrow & Abc \mid b \\
B & \longrightarrow & d
\end{array}
$$

To parse: *a b b c d e*

Stack	Input
$a\ \boxed{b}$	$\cancel{a}\ \cancel{b}\ b\ c\ d\ e\ \$$
$a\ \boxed{A\ b\ c}$	$\cancel{b}\ \cancel{c}\ d\ e\ \$$
$a\ A\ \boxed{d}$	$\cancel{d}\ e\ \$$
$\boxed{a\ A\ B\ e}$	$\cancel{e}\ \$$
S	$\$$

Shift-Reduce Parsing: An Example

$$\begin{array}{lcl} E & \longrightarrow & E+T \\ E & \longrightarrow & T \\ T & \longrightarrow & \mathrm{id} \end{array}$$

$A \to b$

$B \to b$

Stack	Input Stream	Action
\$	id + id \$	shift
\$ id	+ id \$	reduce by $T \longrightarrow \mathrm{id}$
\$ T	+ id \$	reduce by $E \longrightarrow T$
\$ E	+ id \$	shift
\$ E +	id \$	shift
\$ E + id	\$	reduce by $T \longrightarrow \mathrm{id}$
\$ E + T	\$	reduce by $E \longrightarrow E+T$
\$ E	\$	**ACCEPT**

More on Handles

$\boxed{id + id} + id$

Handle: Let $S \Longrightarrow^{*}_{rm} \alpha A w \Longrightarrow_{rm} \alpha\beta w$.

Then $A \longrightarrow \beta$ is a handle for $\alpha\beta w$ at the position imeediately following α.

Notes:

- For unambiguous grammars, every right-sentential form has a unique handle.
- In shift-reduce parsing, handles always appear on top of stack, i.e., $\alpha\beta$ is in the stack (with β at top), and w is unread input.

Identification of Handles and Relationship to Conflicts

Case 1: With $\underline{\alpha\beta}$ on the stack, don't know if we have a handle on top of the stack, or we need to shift some more input to get $\underline{\beta x}$ which is a handle.

- Shift-reduce conflict
- Example: if-then-else

Case 2: With $\alpha\underline{\beta_1\beta_2}$ on the stack, don't know if $\underline{A \longrightarrow \beta_2}$ is the handle, or $\underline{B \longrightarrow \beta_1\beta_2}$ is the handle

- Reduce-reduce conflict
- Example: $E \longrightarrow E - E \mid - E \mid id$

Viable Prefix

- Prefix of a right-sentential form that does not continue beyond the rightmost handle.
- With $\alpha\beta w$ example of the previous slides, a viable prefix is something of the form $\alpha\beta_1$ where $\beta = \beta_1\beta_2$

LR Parsing

- Stack contents as $s_0X_1s_1X_2\cdots X_ms_m$
- Its actions are driven by two tables, *action* and *goto*

Parser Configuration: $(\underbrace{s_0X_1s_1X_2\cdots X_ms_m}_{\text{stack}},\quad \underbrace{a_ia_{i+1}\cdots a_n\$}_{\text{unconsumed input}}\quad)$

$action[s_m, a_i]$ can be:

- shift s: new config is $(s_0X_1s_1X_2\cdots X_ms_ma_is,\ a_{i+1}\cdots a_n\$)$
- reduce $A \longrightarrow \beta$: Let $|\beta| = r$, $goto[s_{m-r}, A] = s$: new config is $(s_0X_1s_1X_2\cdots X_{m-r}s_{m-r}As,\ a_ia_{i+1}\cdots a_n\$)$
- error: perform recovery actions
- accept: Done parsing

$A \rightarrow aBC$

$C \rightarrow de$

Knuth

r states

r non terminals

LR Parsing

- *action* and *goto* depend only on the state at the top of the stack, not on all of the stack contents
  - The s_i states compactly summarize the "relevant" stack content that is at the top of the stack.
- You can think of *goto* as the action taken by the parser on "consuming" (and shifting) nonterminals
  - similar to the shift action in the *action* table, except that the transition is on a nonterminal rather than a terminal
- The *action* and *goto* tables define the transitions of an FSA that accepts RHS of productions!

Example of LR Parsing Table and its Use

- See Text book Algorithm 4.7: (follows directly from description of LR parsing actions 2 slides earlier)
- See expression grammar (Example 4.33), its associated parsing table in Fig 4.31, and the use of the table to parse $id * id + id$ (Fig 4.32)

LR Versus LL Parsing

Intuitively:

- LL parser needs to guess the production based on the first symbol (or first few symbols) on the RHS of a production
- LR parser needs to guess the production *after* seeing all of the RHS

Both types of parsers can use next k input symbols as look-ahead symbols (LL(k) and LR(k) parsers)

- Implication: $LL(k) \subset LR(k)$

How to Construct LR Parsing Table?

Key idea: Construct an FSA to recognize RHS of productions

- States of FSA remember which parts of RHS have been seen already.
- We use “ · ” to separate seen and unseen parts of RHS

LR(0) item: A production with “ · ” somewhere on the RHS. Intuitively,

- ▷ grammar symbols before the “ · ” are on stack;
- ▷ grammar symbols after the “ · ” represent symbols in the input stream.

$$I_0: \quad \begin{array}{l} E' \longrightarrow \cdot E \\ E \longrightarrow \cdot E+T \\ E \longrightarrow \cdot T \\ T \longrightarrow \cdot \mathrm{id} \end{array}$$

$E' \to E\cdot$

$E \to E \cdot + T$

$E \to E + \cdot T$

$E \to E + T \cdot$

$E \to T \cdot$

$T \to id \cdot$

$E' \to E$

$S' \to \cdot S$

How to Construct LR Parsing Table?

- If there is no way to distinguish between two different productions at some point during parsing, then the same state should represent both.
  - *Closure* operation: If a state s includes LR(0) item $A \longrightarrow \alpha \cdot B\beta$, and there is a production $B \longrightarrow \gamma$, then s should include $B \longrightarrow \cdot \gamma$
  - *goto* operation: For a set I of items, $goto[I, X]$ is the closure of all items $A \longrightarrow \alpha X \cdot \beta$ for each $A \longrightarrow \alpha \cdot X\beta$ in I

Item set: A set of items that is closed under the *closure* operation, corresponds to a <u>state</u> of the parser.

Constructing Simple LR (SLR) Parsing Tables

Step 1: Construct LR(0) items (Item set construction) $\rightarrow$ states

Step 2: Construct a DFA for recognizing items

Step 3: Define *action* and *goto* based on the DFA

Item Set Construction

1. Augment the grammar with a rule $S' \longrightarrow S$, and make S' the new start symbol
2. Start with initial set I_0 corresponding to the item $S' \longrightarrow \cdot S$
3. apply *closure* operation on I_0.
4. For each item set I and grammar symbol X, add $goto[I, X]$ to the set of items
5. Repeat previous step until no new item sets are generated.

Item Set Construction

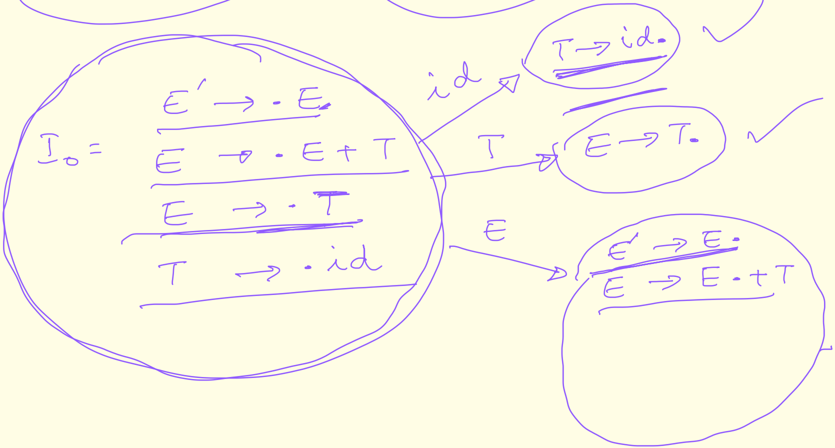
$E' \longrightarrow E \qquad E \longrightarrow E + T \mid T \qquad T \longrightarrow T * F \mid F \qquad F \longrightarrow (E) \mid id$

$I_0 : E' \longrightarrow \cdot E$

$I_1 : E' \longrightarrow E \cdot$

$I_2 : E \longrightarrow T \cdot$

$I_3 : T \longrightarrow F \cdot$



Item Set Construction (Contd.)

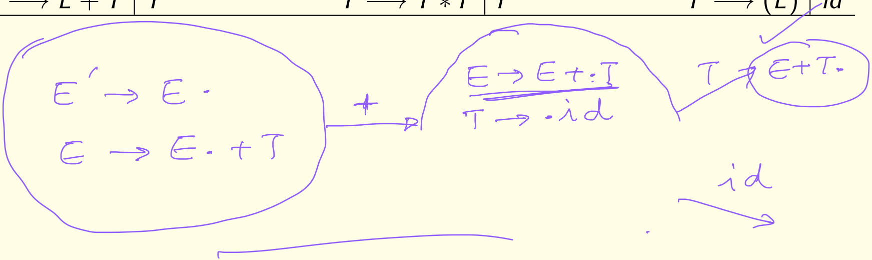
$E' \longrightarrow E \qquad E \longrightarrow E + T \mid T \qquad T \longrightarrow T * F \mid F \qquad F \longrightarrow (E) \mid id$

$I_4 : F \longrightarrow (\cdot E)$

$I_5 : F \longrightarrow id \cdot$

$I_6 : E \longrightarrow E + \ \cdot T$

$I_7 : T \longrightarrow T * \ \cdot F$



Item Set Construction (Contd.)

$E' \longrightarrow E \qquad E \longrightarrow E + T \mid T \qquad T \longrightarrow T * F \mid F \qquad F \longrightarrow (E) \mid id$

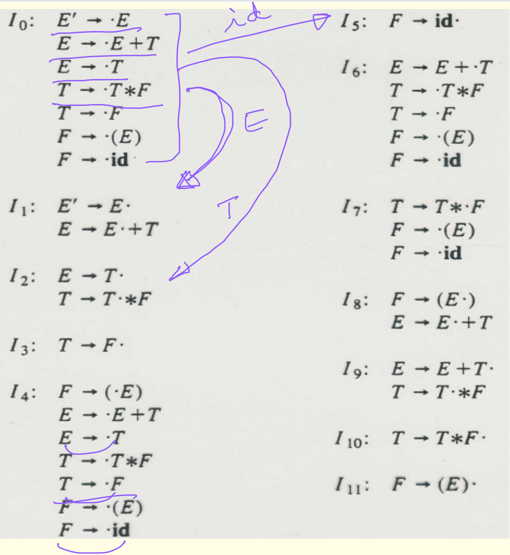
$I_8 : F \longrightarrow (E \cdot)$

$I_9 : E \longrightarrow E + T \cdot$

$I_{10} : T \longrightarrow T * F \cdot$

$I_{11} : F \longrightarrow (E) \cdot$

Item Sets for the Example



I_0: $E' \to \cdot E$
$E \to \cdot E + T$
$E \to \cdot T$
$T \to \cdot T * F$
$T \to \cdot F$
$F \to \cdot (E)$
$F \to \cdot \mathbf{id}$

I_1: $E' \to E \cdot$
$E \to E \cdot + T$

I_2: $E \to T \cdot$
$T \to T \cdot * F$

I_3: $T \to F \cdot$

I_4: $F \to (\cdot E)$
$E \to \cdot E + T$
$E \to \cdot T$
$T \to \cdot T * F$
$T \to \cdot F$
$F \to \cdot (E)$
$F \to \cdot \mathbf{id}$

I_5: $F \to \mathbf{id} \cdot$

I_6: $E \to E + \cdot T$
$T \to \cdot T * F$
$T \to \cdot F$
$F \to \cdot (E)$
$F \to \cdot \mathbf{id}$

I_7: $T \to T * \cdot F$
$F \to \cdot (E)$
$F \to \cdot \mathbf{id}$

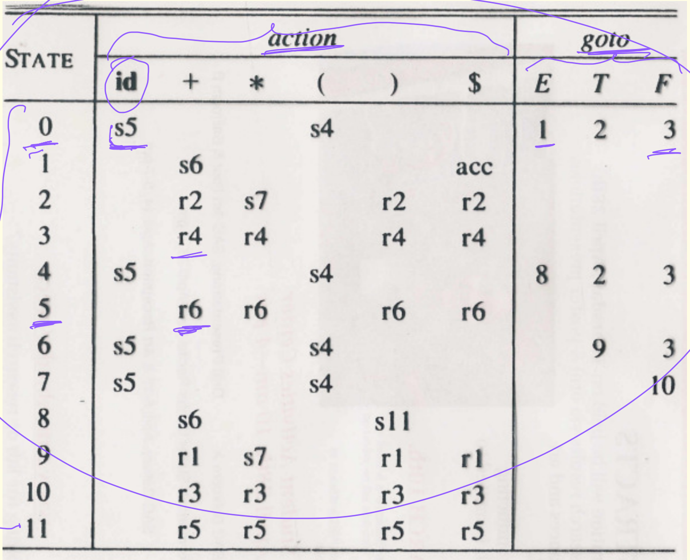
I_8: $F \to (E \cdot)$
$E \to E \cdot + T$

I_9: $E \to E + T \cdot$
$T \to T \cdot * F$

I_{10}: $T \to T * F \cdot$

I_{11}: $F \to (E) \cdot$

SLR(1) Parse Table for the Example Grammar



STATE	action						goto		
	id	+	*	(	)	$	E	T	F
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

id + id * id

~~0 id 5~~

1

0 F 3

F

Defining *action* and *goto* tables

- Let $I_0, I_1, \ldots, I_n$ be the item sets constructed before

$\alpha\, a \cdot B$

$A \to \alpha\cdot$

$A \to \alpha \cdot a B$

LR(0)

- Define *action* as follows
  - If $A \longrightarrow \alpha \cdot a\beta$ is in I_i and there is a DFA transition to I_j from I_i on symbol a then $action[i, a] =$ "shift j"
  - If $A \longrightarrow \alpha\,\cdot$ is in I_i then $action[i, \overset{b}{\cancel{a}}] =$ "reduce $A \longrightarrow \alpha$" for every $\overset{b}{\cancel{a}} \in FOLLOW(A)$
  - If $S' \longrightarrow S\,\cdot$ is in I_i then $action[I_i, \$] =$ "accept"
- If any conflicts arise in the above procedure, then the grammar is *not* SLR(1).
- *goto* transition for LR parsing defined directly from the DFA transitions.
- All undefined entries in the table are filled with "error"

$A \to \alpha\cdot$

$B \to \beta\cdot$

FOLLOW(A) $\cap$

$\subseteq$ $\in$ FOLLOW(B)

Defining *action* and *goto* tables

$\alpha\, a \cdot \beta$

$A \to \alpha\cdot$

$A \to \alpha \cdot a\beta$

- Let $I_0, I_1, \ldots, I_n$ be the item sets constructed before
- Define *action* as follows

  LR(0)

  - If $A \longrightarrow \alpha \cdot a\beta$ is in I_i and there is a DFA transition to I_j from I_i on symbol a then $action[i, a] =$ "shift j"
  - If $A \longrightarrow \alpha\cdot$ is in I_i then $action[i, \overset{b}{\cancel{a}}] =$ "reduce $A \longrightarrow \alpha$" for every $\overset{b}{\cancel{a}} \in FOLLOW(A)$
  - If $S' \longrightarrow S\cdot$ is in I_i then $action[I_i, \$] =$ "accept"
- If any conflicts arise in the above procedure, then the grammar is *not* SLR(1).
- *goto* transition for LR parsing defined directly from the DFA transitions.
- All undefined entries in the table are filled with "error"

$A \to \alpha\cdot$

$B \to \beta\cdot$

FOLLOW(A) ∩

c ∈ FOLLOW(B)

Deficiencies of SLR Parsing

SLR = LR(0) item sets + LR(1)
1 - look ahead for reduction

SLR(1) treats all occurrences of a RHS on stack as identical.

Only a few of these reductions may lead to a successful parse.

$S \to ab \mid ba$

Example:

$$\begin{array}{lll} S \longrightarrow AaAb & & A \longrightarrow \epsilon \\ S \longrightarrow BbBa & & B \longrightarrow \epsilon \end{array}$$

FOLLOW(A) = {a, b}

FOLLOW(B) = {a, b}

$I_0 = \{[S' \to \cdot S], [S \to \cdot AaAb], [S \to \cdot BibBa], [A \to \cdot], [B \to \cdot]\}$.

Since $FOLLOW(A) = FOLLOW(B)$, we have reduce/reduce conflict in state 0.

LR(1) Item Sets

Construct LR(1) items of the form $A \longrightarrow \alpha \cdot \beta, a$, which means:

> *The production $A \longrightarrow \alpha\beta$ can be applied when the next token on input stream is a.*

S	$\longrightarrow AaAb$	A	$\longrightarrow \epsilon$
S	$\longrightarrow BbBa$	B	$\longrightarrow \epsilon$

An example LR(1) item set:

$$I_0 = \{[S' \to \cdot S, \$], [S \to \cdot AaAb, \$], [S \to \cdot BbBa, \$],$$
$$[A \to \cdot, a], [B \to \cdot, b]\}.$$

$[S' \to \cdot S, \$]$

$S \to \cdot AaAb, \$]$

$[A \to \cdot \epsilon, a]$

$[S \to \cdot BbBa, \$]$

$[B \to \cdot, b]$

LR(1) and LALR(1) Parsing

LR(1) item sets that are identical except for the look ahead

↪ merged

LR(1) parsing: Parse tables built using LR(1) item sets.

LALR(1) parsing: *Look Ahead* LR(1)

Merge LR(1) item sets; then build parsing table.

Typically, LALR(1) parsing tables are much smaller than LR(1) parsing table.

$$\begin{pmatrix}[A \to \cdot, a] \\ [B \to \cdot, b]\end{pmatrix} \xleftrightarrow{\text{merged}} \begin{pmatrix}[A \to \cdot, b] \\ [B \to \cdot, b]\end{pmatrix}$$

$$\times \quad ([A \to \cdot, a]) \quad I_0$$

$$\times \quad ([B \to \cdot, a]) \quad I_1$$

$$\begin{pmatrix}[A \to \cdot b] \\ [B \to \cdot b]\end{pmatrix} \quad I_2$$

YACC

Yet Another Compiler Compiler:

LALR(1) parser generator.

- Grammar rules are written in a specification (`.y`) file, analogous to the regular definitions in a `lex` specification file.
- Yacc translates the specifications into a parsing function `yyparse()`.

$$\texttt{spec.y} \xrightarrow{\texttt{yacc}} \texttt{spec.tab.c}$$

- `yyparse()` calls `yylex()` whenever input tokens need to be consumed.
- `bison`: GNU variant of `yacc`.

Using Yacc

```
%{
  ... C headers (#include)
%}
... Yacc declarations:
       %token ...
       %union{...}
       precedences
%%
... Grammar rules with actions:
Expr:  Expr TOK_PLUS Expr
    |  Expr TOK_MINUS Expr
    ;
%%
... C support functions
```

YACC

Yet **A**nother **C**ompiler **C**ompiler:

LALR(1) parser generator.

- Grammar rules are written in a specification (`.y`) file, analogous to the regular definitions in a `lex` specification file.
- Yacc translates the specifications into a parsing function `yyparse()`.

$$\texttt{spec.y} \xrightarrow{\texttt{yacc}} \texttt{spec.tab.c}$$

- `yyparse()` calls `yylex()` whenever input tokens need to be consumed.
- `bison`: GNU variant of `yacc`.

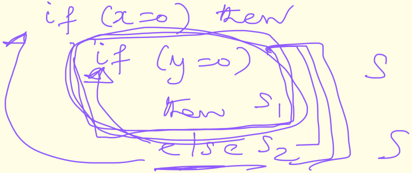
Using Yacc

```
%{
  ... C headers (#include)
%}
... Yacc declarations:
       %token ...
       %union{...}
       precedences
%%
... Grammar rules with actions:
Expr:  Expr TOK_PLUS Expr
    |  Expr TOK_MINUS Expr
    ;
%%
... C support functions
```

Conflicts and Resolution

ifStmt → if Expr then S
if Expr then S else S

- Operator precedence works well for resolving conflicts that involve operators
  - But use it with care – only when they make sense, not for the sole purpose of removing conflict reports
- Shift-reduce conflicts: Bison favors shift
  - Except for the dangling-else problem, this strategy does not ever seem to work, so don't rely on it.



Reduce-Reduce Conflicts

$S \to \varepsilon \mid \cancel{M} \mid S\ W$

$\cancel{M \to \varepsilon} \mid W$

```
sequence: /* empty */
            { printf ("empty sequence\n"); }
          | maybeword
          | sequence word
            { printf ("added word %s\n", $2); };
maybeword: /* empty */
             { printf ("empty maybeword\n"); }
          | word
            { printf ("single word %s\n", $1); };
```

$S \to M \to \varepsilon$

$S \to \varepsilon$

In general, grammar needs to be rewritten to eliminate conflicts.

Sample Bison File: Postfix Calculator

```
input:    /* empty */
        | input line
;
line:     '\n'
        | exp '\n'      { printf ("\t%.10g\n", $1); }
;
exp:      NUM           { $$ = $1;           }
        | exp exp '+'   { $$ = $1 + $2;      }
        | exp exp '-'   { $$ = $1 - $2;      }
        | exp exp '*'   { $$ = $1 * $2;      }
        | exp exp '/'   { $$ = $1 / $2;      }
         /* Exponentiation */
        | exp exp '^'   { $$ = pow ($1, $2); }
         /* Unary minus    */
        | exp 'n'       { $$ = -$1;          };
%%
```

Infix Calculator

```
%{
#define YYSTYPE double
#include <math.h>
#include <stdio.h>
int yylex (void);
void yyerror (char const *);
%}
/* Bison Declarations */
%token NUM
%left '-' '+'
%left '*' '/'
%left NEG      /* negation--unary minus */
%right '^'     /* exponentiation */
```

lower precedence

higher

$5+5*3$

$(5+5)*3$ $\quad$ $5\,5+3*$

$5+(5*3)$ $\quad$ $5\,5\,3*+$

$(-5)*5$

$5-(5*5)$

Infix Calculator (Continued)

```
%% /* The grammar follows.  */
input:    /* empty */
        | input line
;
line:     '\n'
        | exp '\n'  { printf ("\t%.10g\n", $1); }
;
exp:      NUM                { $$ = $1;         }
        | exp '+' exp        { $$ = $1 + $3;    }
        | exp '-' exp        { $$ = $1 - $3;    }
        | exp '*' exp        { $$ = $1 * $3;    }
        | exp '/' exp        { $$ = $1 / $3;    }
        | '-' exp  %prec NEG { $$ = -$2;        }
        | exp '^' exp        { $$ = pow ($1, $3); }
        | '(' exp ')'        { $$ = $2;         }
;
%%
```

$I \to \varepsilon \mid I\ L$

$\underbrace{L\ L\ L\ L}$

0 or more

$1

$2

Error Recovery

```
line:      '\n'
         | exp '\n'   { printf ("\t%.10g\n", $1); }
         | error '\n' { yyerrok;                  };
```

5 + * 5 \n

$

~~exp +~~ error

- Pop stack contents to expose a state where an `error` token is acceptable
- Shift error token onto the stack
- Discard input until reaching a token that can follow this error token

Error recovery strategies are never perfect — some times they lead to cascading errors, unless carefully designed.

Left Versus Right Recursion

LL

5, 5, 5, 5, 5

```
expseq1:   exp | expseq1 ',' exp;
```

is a left-recursive definition of a sequence of exp's, whereas

```
expseq1:   exp | exp ',' expseq1;
```

is a right-recursive definition

- Left-recursive definitions are a no-no for LL parsing, but yes-yes for LR parsing
- Right-recursive definition is bad for LR parsing as it needs to shift ithe entire list on stack before any reduction — increases stack usage