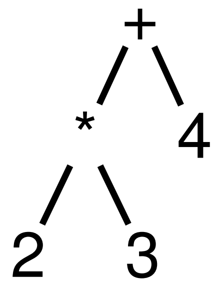
Parsing

Parsing

Objective: build an abstract syntax tree (AST) for the token sequence from the scanner.

`2 * 3 + 4` $\Rightarrow$



Goal: discard irrelevant information to make it easier for the next stage.

Parentheses and most other forms of punctuation removed.

Grammars

Most programming languages described using a *context-free grammar*.

Compared to regular languages, context-free languages add one important thing: recursion.

Recursion allows you to count, e.g., to match pairs of nested parentheses.

Which languages do humans speak? I'd say it's regular: I do not not not not not not not not not not understand this sentence.

Languages

Regular languages (t is a terminal):

$$A \to t_1 \dots t_n B$$

$$A \to t_1 \dots t_n$$

Context-free languages (P is terminal or a variable):

$$A \to P_1 \dots P_n$$

Context-sensitive languages:

$$\alpha_1 A \alpha_2 \to \alpha_1 B \alpha_2$$

"$B \to A$ only in the 'context' of $\alpha_1 \cdots \alpha_2$"

Issues

Ambiguous grammars

Precedence of operators

Left- versus right-recursive

Top-down vs. bottom-up parsers

Parse Tree vs. Abstract Syntax Tree

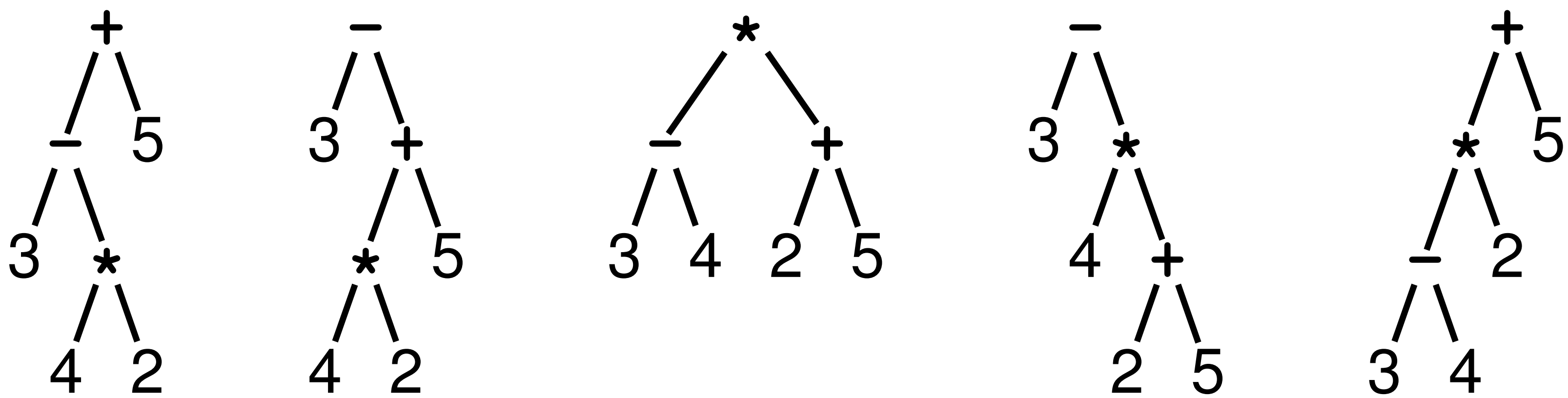
Ambiguous Grammars

A grammar can easily be ambiguous. Consider parsing

```
3 - 4 * 2 + 5
```

with the grammar

$$e \to e + e \mid e - e \mid e * e \mid e \,/\, e \mid N$$



Operator Precedence and Associativity

Usually resolve ambiguity in arithmetic expressions

Like you were taught in elementary school:

“My Dear Aunt Sally”

Mnemonic for multiplication and division before addition and subtraction.

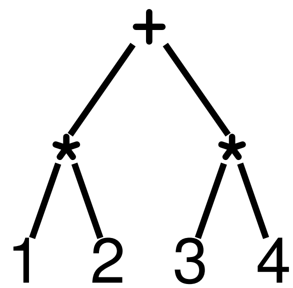
Operator Precedence

Defines how “sticky” an operator is.

`1 * 2 + 3 * 4`

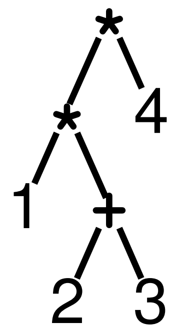
`*` at higher precedence than `+`:

`(1 * 2) + (3 * 4)`



`+` at higher precedence than `*`:

`1 * (2 + 3) * 4`



Associativity

Whether to evaluate left-to-right or right-to-left

Most operators are left-associative

`1 - 2 - 3 - 4`

```
           -                        -
          / \                      / \
         -   4                    1   -
        / \                          / \
       -   3                        2   -
      / \                              / \
     1   2                            3   4
```

((1 - 2) - 3) - 4 | 1 - (2 - (3 - 4))

left associative | right associative

Fixing Ambiguous Grammars

Original ANTLR grammar specification

```
expr
  : expr '+' expr
  | expr '-' expr
  | expr '*' expr
  | expr '/' expr
  | NUMBER
  ;
```

Ambiguous: no precedence or associativity.

Assigning Precedence Levels

Split into multiple rules, one per level

```
expr : expr '+' expr
     | expr '-' expr
     | term ;

term : term '*' term
     | term '/' term
     | atom ;

atom : NUMBER ;
```

Still ambiguous: associativity not defined

Assigning Associativity

Make one side or the other the next level of precedence

```
expr : expr '+' term
     | expr '-' term
     | term ;

term : term '*' atom
     | term '/' atom
     | atom ;

atom : NUMBER ;
```

Parsing Context-Free Grammars

There are $O(n^3)$ algorithms for parsing arbitrary CFGs, but most compilers demand $O(n)$ algorithms.

Fortunately, the LL and LR subclasses of CFGs have $O(n)$ parsing algorithms. People use these in practice.

The CYK algorithm (Cocke-Younger-Kasami)

Inputs: a string $w = a_1 a_2 \dots a_n$ and a grammar in Chomsky Normal Form (only productions $X \to a$ and $X \to YZ$)

Construct an $n \times n$ table T, where $T[i, j]$ is the set of nonterminals that generate the substring $a_i a_{i+1} \dots a_j$. Using dynamic programming, each table entry can be filled in $O(n)$ time. The algorithm runs in $O(n^3)$ time.

Question: how to determine that $a_1 a_2 \dots a_n$ is in the language?

Parsing LL(k) Grammars

LL: Left-to-right, Left-most derivation

k: number of tokens to look ahead

Parsed by top-down, predictive, recursive parsers

Basic idea: look at the next token to predict which production to use

ANTLR builds recursive LL(k) parsers

Almost a direct translation from the grammar.

Implementing a Top-Down Parser

```
stmt : 'if' expr 'then' expr
     | 'while' expr 'do' expr
     | expr ':=' expr ;
expr : NUMBER | '(' expr ')' ;
```

```
stmt() {
  switch (next-token) {
  case IF:
      match(IF); expr(); match(THEN); expr();          break;
  case WHILE:
      match(WHILE); expr(); match(DO); expr();         break;
  case NUMBER or LPAREN:
      expr(); match(COLEQ); expr();                    break;
}}
```

Writing LL(k) Grammars

Cannot have left-recursion

```
expr : expr '+' term | term ;
```

becomes

```
AST expr() {
  switch (next-token) {
  case NUMBER : expr(); /* Infinite Recursion */
```

Writing LL(1) Grammars

Cannot have common prefixes

```
expr : ID '(' expr ')'
     | ID '=' expr
```

becomes

```
expr() {
  switch (next-token) {
  case ID:
    match(ID); match(LPAR); expr(); match(RPAR);  break;
  case ID:
    match(ID); match(EQUALS); expr();             break;
```

Eliminating Common Prefixes

Consolidate common prefixes:

```
expr
  : expr '+' term
  | expr '-' term
  | term
  ;
```

becomes

```
expr
  : expr ('+' term | '-' term )
  | term
  ;
```

Eliminating Left Recursion

Understand the recursion and add tail rules

```
expr
  : expr ('+' term | '-' term )
  | term
  ;
```

becomes

```
expr : term exprt ;
exprt : '+' term exprt
      | '-' term exprt
      | /* nothing */
      ;
```

Using ANTLR's EBNF

ANTLR makes this easier since it supports `*` and `+`:

```
expr : expr '+' term
     | expr '-' term
     | term ;
```

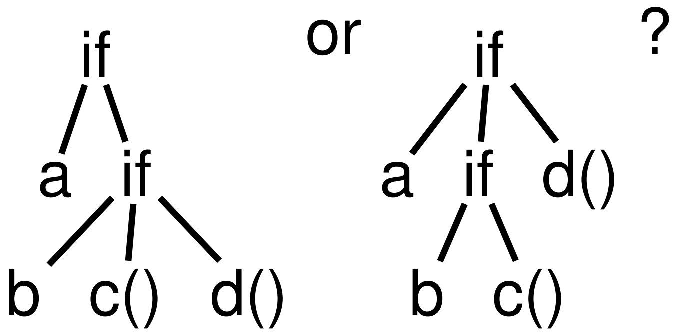
becomes

```
expr : term ('+' term | '-' term)* ;
```

The Dangling Else Problem

Who owns the *else*?

```
if (a) if (b) c(); else d();
```



Grammars are usually ambiguous; manuals give disambiguating rules such as C’s:

> As usual the “else” is resolved by connecting an else with the last encountered elseless if.

The Dangling Else Problem

```
stmt : "if" expr "then" stmt iftail
     | other-statements ;

iftail
   : "else" stmt
   | /* nothing */
   ;
```

Problem comes when matching “iftail.”

Normally, an empty choice is taken if the next token is in the “follow set” of the rule. But since “else” can follow an iftail, the decision is ambiguous.

The Dangling Else Problem

ANTLR can resolve this problem by making certain rules “greedy.” If a conditional is marked as greedy, it will take that option even if the “nothing” option would also match:

```
stmt
  : "if" expr "then" stmt
     ( options {greedy = true;}
     : "else" stmt
     )?
  | other-statements
  ;
```

The Dangling Else Problem

Some languages resolve this problem by insisting on nesting everything.

E.g., Algol 68:

```
if a < b then a else b fi;
```

“fi” is “if” spelled backwards. The language also uses do–od and case–esac.

Statement separators/terminators

C uses `;` as a statement terminator.

```
if (a<b) printf("a less");
else {
  printf("b"); printf(" less");
}
```

Pascal uses `;` as a statement separator.

```
if a < b then writeln('a less')
else begin
  write('a'); writeln(' less')
end
```

Pascal later made a final `;` optional.

Table-driven Top-Down Parsing

Nomenclature

$a, b, c, \ldots$ represent terminal symbols

$A, B, C, \ldots$ represent nonterminal symbols

S represents the initial nonterminal symbol

$X, Y, Z, \ldots$ represent terminal or non-terminal symbols

$u, v, w, \ldots$ represent words of terminal symbols

$\alpha, \beta, \gamma, \ldots$ represent words of terminal and non-terminal symbols

$\Rightarrow_G$ represents a one-step derivation with grammar G

$\stackrel{*}{\Rightarrow}_G$ represents zero or more steps in a derivation

First and follow

$$\textsc{First}(\alpha) \equiv$$

$$\{a : \alpha \stackrel{*}{\Rightarrow} a\beta\} \cup$$

$$(\texttt{\textbf{if}}\ \ \alpha \stackrel{*}{\Rightarrow} \varepsilon\ \ \texttt{\textbf{then}}\ \ \{\varepsilon\}\ \ \texttt{\textbf{else}}\ \ \emptyset)$$

$$\textsc{Follow}(A) \equiv$$

$$\{a : S \stackrel{+}{\Rightarrow} \alpha A a\beta\} \cup$$

$$(\texttt{\textbf{if}}\ \ S \stackrel{*}{\Rightarrow} \alpha A\ \ \texttt{\textbf{then}}\ \ \{\varepsilon\}\ \ \texttt{\textbf{else}}\ \ \emptyset)$$

Calculating first and follow

FIRST:

For all terminals a: $\textsc{First}(a) = \{a\}$;

For all nonterminals X: $\textsc{First}(X) = \emptyset$;

For all productions $X \to \varepsilon$, add ε to $\textsc{First}(X)$;

repeat

For all productions $X \to Y_1 Y_2 \ldots Y_k$ ⟨outer loop⟩

For i in $1 \ldots k$ ⟨inner loop⟩

add $(\textsc{First}(Y_i) \setminus \{\varepsilon\})$ to $\textsc{First}(X)$;

if $\varepsilon \notin \textsc{First}(Y_i)$ **continue outer loop**;

add ε to $\textsc{First}(X)$;

until no further progress

Calculating first and follow

FOLLOW:

FOLLOW$(S) = \{\varepsilon\}$, where S is the start symbol

For all other symbols X, FOLLOW$(X) = \emptyset$

`repeat`

For all productions $A \rightarrow \alpha B \beta$

add (FIRST$(\beta) \setminus \{\varepsilon\}$) to FOLLOW$(B)$

For all productions $A \rightarrow \alpha B$

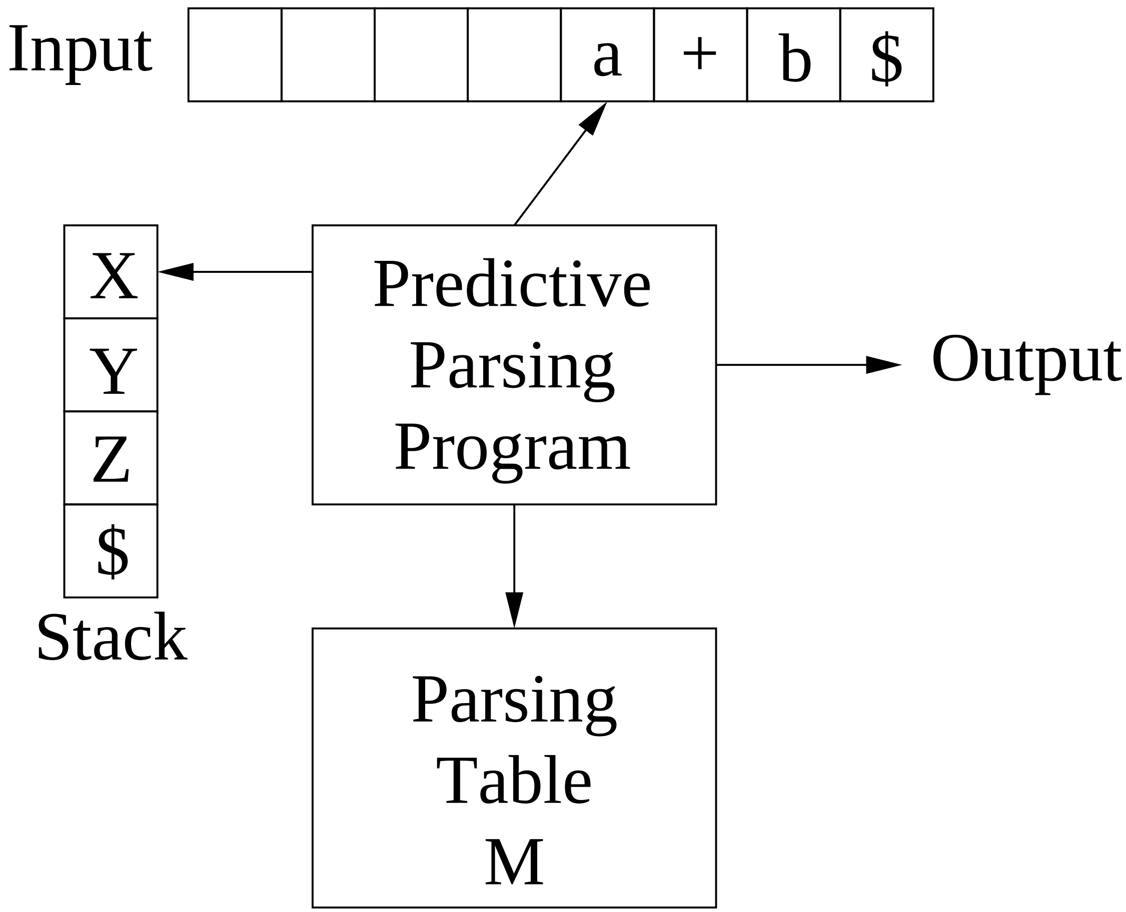
or $A \rightarrow \alpha B \beta$, where $\varepsilon \in$ FIRST(β)

add FOLLOW(A) to FOLLOW(B)

`until` no further progress

Note: FIRST(β) is calculated in a similar way as the inner loop of FIRST.

Table-driven predictive parser



Parsing table

For each production $A \to \alpha$ of the grammar do:

1. For each terminal $a \in \textsc{First}(\alpha)$, add $A \to \alpha$ to $M[A, a]$.

2. If $\varepsilon \in \textsc{First}(\alpha)$, then for each terminal b in $\textsc{Follow}(A)$, add $A \to \alpha$ to $M[A, b]$. If $\varepsilon \in \textsc{First}(\alpha)$ and $\$ \in \textsc{Follow}(A)$, add $A \to \alpha$ to $M[A, \$]$ as well.

All empty entries must be set to **error**.

Parsing table

Production	First	Follow
$E \to TE'$	$\{(, \textbf{id}\}$	$\{), \$\}$
$E' \to +TE' \mid \varepsilon$	$\{+, \varepsilon\}$	$\{), \$\}$
$T \to FT'$	$\{(, \textbf{id}\}$	$\{+,), \$\}$
$T' \to *FT' \mid \varepsilon$	$\{*, \varepsilon\}$	$\{+,), \$\}$
$F \to (E) \mid \textbf{id}$	$\{(, \textbf{id}\}$	$\{+, *,), \$\}$

	id	$+$	$*$	(	) , \$
E	$E \to TE'$			$E \to TE'$	
E'		$E' \to +TE'$			$E' \to \varepsilon$
T	$T \to FT'$			$T \to FT'$	
T'		$T' \to \varepsilon$	$T' \to *FT'$		$T' \to \varepsilon$
F	$F \to \textbf{id}$			$F \to (E)$	

LL(1) grammars

A grammar is LL(1) if

- For every pair of rules $A \to \alpha_1$ and $A \to \alpha_2$, then $\textsc{First}(\alpha_1) \cap \textsc{First}(\alpha_2) = \emptyset$.
- If $A \to \alpha$ and A is nullable, then $\textsc{First}(\alpha) \cap \textsc{Follow}(A) = \emptyset$.

If a grammar is LL(1), then no conflicts appear in the parsing table.

Predicting parsing algorithm

ip is the pointer to the input (initially pointing at the first token);
The stack initially contains $S\$$, with S on the top;
Set X to the top of the stack symbol;

while $(X \neq \$)$ **do**
    a = INPUT[*ip*];
    if $(X = a)$ pop the stack and advance *ip*;
    else if (X is a terminal) error();
    else if ($M[X, a] =$ **error**) error();
    else if $(M[X, a] = X \to Y_1 Y_2 \cdots Y_k)$
        Output the production $X \to Y_1 Y_2 \cdots Y_k$;
        Pop the stack;
        Push $Y_k, Y_{k-1}, \ldots, Y_1$ onto the stack, with Y_1 on top;
    Set X to the top of the stack symbol;

Exercise: Execute the parsing algorithm with the input $z * (x + y)\$$.

Generation of top-down parsers
(ANTLR style)

A simple example

Grammar

```
instruction_list: ( instruction ) *

instruction: IDENT ":=" expr |
             IF expr THEN instruction_list

expr: (IDENT | NUM) (PLUS (IDENT | NUM))*
```

Parser

```
void instruction_list () {
  while (token==IDENT || token==IF) {
    instruction();
  }
}
```

A simple example

```
instruction: IDENT ":=" expr |
             IF expr THEN instruction_list
```

```
void instruction () {
  if (token==IDENT) {
    nexttoken();
    if (token== COLON_EQUAL) {
      nexttoken(); expr();
    } else SYNTAXERROR();
  } else if (token==IF) {
    nexttoken(); expr();
    if (token==THEN) {
      nextoken(); instruction_list();
    } else SYNTAXERROR();
  } else SYNTAXERROR();
}
```

A simple example

```
expr: (IDENT | NUM) (PLUS (IDENT | NUM))*
```

```
void expr() {
  if (token==IDENT || token==NUM) {
    nexttoken();
    while (token==PLUS) {
      nexttoken();
      if (token==IDENT || token==NUM){
        nexttoken();
      } else SYNTAXERROR();
    }
  } else SYNTAXERROR();
}
```

ANTLR

- ANTLR uses the EBFN notation
- NULLABLE, FIRST and FOLLOW must be extended to the EBNF notation.
- A recursive top-down parser is generated.

Nullable

- $\textsc{Nullable}(\varepsilon) = \textit{true}$.
- $\textsc{Nullable}(E*) = \textit{true}$.
- If $\textsc{Nullable}(E)$, then $\textsc{Nullable}(E+) = \textit{true}$.
- If $V \to E \in G$ and $\textsc{Nullable}(E)$, then $\textsc{Nullable}(V) = \textit{true}$.
- If $\textsc{Nullable}(E_1)$ and $\textsc{Nullable}(E_2)$, then $\textsc{Nullable}(E_1 E_2) = \textit{true}$.
- If $\textsc{Nullable}(E_1)$ or $\textsc{Nullable}(E_2)$, then $\textsc{Nullable}(E_1 | E_2) = \textit{true}$.
- Nothing else is $\textsc{Nullable}$.

First

- If c is a terminal, $\textsc{First}(c) = \{c\}$.
- $\textsc{First}(E*)$ and $\textsc{First}(E+)$ contain $\textsc{First}(E)$.
- If $V \to E \in G$, then $\textsc{First}(V)$ contains $\textsc{First}(E)$.
- $\textsc{First}(E_1 E_2)$ contains $\textsc{First}(E_1)$.
- If $\textsc{Nullable}(E_1)$, then $\textsc{First}(E_1 E_2)$ contains $\textsc{First}(E_2)$.
- $\textsc{First}(E_1 | E_2)$ contains $\textsc{First}(E_1)$ and $\textsc{First}(E_2)$.
- Nothing else belongs to FIRST.

Follow

- If $E_1 E_2$ is and expression of the grammar, then $\text{F}\textsc{ollow}(E_2)$ contains $\text{F}\textsc{ollow}(E_1 E_2)$, $\text{F}\textsc{ollow}(E_1)$ contains $\text{F}\textsc{irst}(E_2)$ and, if $\text{N}\textsc{ullable}(E_2)$ then $\text{F}\textsc{ollow}(E_1)$ contains $\text{F}\textsc{ollow}(E_1 E_2)$.
- If $E_1 | E_2$ is and expression of the grammar, then $\text{F}\textsc{ollow}(E_1)$ and $\text{F}\textsc{ollow}(E_2)$ contain $\text{F}\textsc{ollow}(E_1 | E_2)$.
- If $E*$ or $E+$ are expressions of the grammar, then their FOLLOW is contained in $\text{F}\textsc{ollow}(E)$.
- If $V \rightarrow E \in G$, then $\text{F}\textsc{ollow}(E)$ contains $\text{F}\textsc{ollow}(V)$.
- Nothing else belongs to FOLLOW.

Generating an LL(1) recursive-descent predictive parser

For every rule $A \to E$, a function is generated:

void A() $\{$

    Parse$(E,$ ***Follow***$(A))$

$\}$

where **Parse**(E, F) is the code generated to recognize the EBNF expression E followed by the tokens in F.

Token is a variable that represents the current token.

match(T**)** $\equiv$

    if (Token == T**) nexttoken();**

    else SyntaxError();

Parse(E,F)

```
Parse(E_1 | E_2 | ... | E_n, F) ≡
    if (Token ∈ First(E_1)) Parse(E_1, F);
    else if (Token ∈ First(E_2)) Parse(E_2, F);
    ...
    else if (Token ∈ First(E_n)) Parse(E_n, F);
    else if (no E_i is nullable) SyntaxError();
    else if (Token ∉ F) SyntaxError();
```

If the BNF version of the grammar is LL(1) then there are no conflicts between the different branches (including the case of a nullable E_i).

Parse(E,F)

Parse($E_1 E_2 \ldots E_n$**,** F**)** $\equiv$

    Parse(E_1**,** ***First***$(E_2 \ldots E_n \cdot F)$**);**

    Parse(E_2**,** ***First***$(E_3 \ldots E_n \cdot F)$**);**

    $\ldots$

    Parse(E_n**,** F**);**

where ***First***$(E \cdot F)$ **is**

- ***First***(E) **if** E **is not nullable**
- ***First***$(E) \cup F$**, otherwise**

Parse(E,F)

Parse($E*$ **,** F **)** $\equiv$

  while (Token $\in$ ***First*** (E) **) Parse(** E **,** F **);**

Parse($E+$ **,** F **)** $\equiv$

  do Parse(E **,** F **); while (Token** $\in$ ***First*** (E, F) **);**

Parse($E?$ **,** F **)** $\equiv$

  if (Token $\in$ ***First*** (E) **) Parse(** E **,** F **);**

Parse(a **,** F **)** $\equiv$ **Match(** a **);** // a is a terminal symbol

Parse(A **,** F **)** $\equiv$ A **();** // A is a non-terminal symbol

LL(1) example

Grammar:

$$E \to T\ ('\mathbf{+}'\ T \mid '\mathbf{-}'\ T)*$$

$$T \to F\ ('\mathbf{*}'\ F \mid '\mathbf{/}'\ F)*$$

$$F \to \mathbf{ID} \mid \mathbf{NUM} \mid '\mathbf{(}'\ E\ '\mathbf{)}'$$

```
void E() {
    T();
    while (Token == '+' or Token == '-') {
        if (Token == '+') { Match('+'); T(); }
        else if (Token == '-') { Match('-'); T(); }
        else SyntaxError();        // redundant
    }
}
```

LL(1) example

```
void T() {
    F();
    while (Token == '*' or Token == '/') {
        if (Token == '*') { Match('*'); F(); }
        else if (Token == '/') { Match('/'); F(); }
        else SyntaxError();        // redundant
    }
}

void F() {
    if (Token == ID) Match(ID);
    else if (Token == NUM) Match(NUM);
    else if (Token == '(') {
        Match('('); E(); Match(')');
    } else SyntaxError();
}
```

Bottom-up Parsing

Rightmost Derivation

$$
\begin{aligned}
1: &\quad e \rightarrow t + e \\
2: &\quad e \rightarrow t \\
3: &\quad t \rightarrow \textbf{Id} * t \\
4: &\quad t \rightarrow \textbf{Id}
\end{aligned}
$$

A rightmost derivation for **Id** $*$ **Id** $+$ **Id**:

$$
\begin{gathered}
\boxed{e} \\
t + \boxed{e} \\
t + \boxed{t} \\
\boxed{t} + \textbf{Id} \\
\textbf{Id} * \boxed{t} + \textbf{Id} \\
\textbf{Id} * \textbf{Id} + \textbf{Id}
\end{gathered}
$$

Basic idea of bottom-up parsing: construct this rightmost derivation backward.

Handles

$$1: \quad e \rightarrow t + e$$
$$2: \quad e \rightarrow t$$
$$3: \quad t \rightarrow \textbf{Id} * t$$
$$4: \quad t \rightarrow \textbf{Id}$$

$\textbf{Id} * \textbf{Id} + \textbf{Id}$

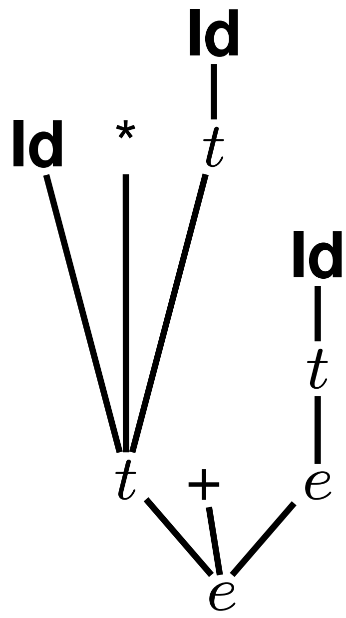
$\textbf{Id} * t + \textbf{Id}$

$t + \textbf{Id}$

$t + t$

$t + e$

e



This is a reverse rightmost derivation for $\textbf{Id} * \textbf{Id} + \textbf{Id}$.

Each highlighted section is a handle.

Taken in order, the handles build the tree from the leaves to the root.

Shift-reduce Parsing

$1: \quad e \rightarrow t + e$

$2: \quad e \rightarrow t$

$3: \quad t \rightarrow \textbf{Id} * t$

$4: \quad t \rightarrow \textbf{Id}$

stack	input	action
	$\textbf{Id} * \textbf{Id} + \textbf{Id}$	shift
$\textbf{Id}$	$* \textbf{Id} + \textbf{Id}$	shift
$\textbf{Id}*$	$\textbf{Id} + \textbf{Id}$	shift
$\textbf{Id} * \boxed{\textbf{Id}}$	$+ \textbf{Id}$	reduce (4)
$\boxed{\textbf{Id} * t}$	$+ \textbf{Id}$	reduce (3)
t	$+ \textbf{Id}$	shift
$t+$	$\textbf{Id}$	shift
$t + \boxed{\textbf{Id}}$		reduce (4)
$t + \boxed{t}$		reduce (2)
$\boxed{t + e}$		reduce (1)
e		accept

Scan input left-to-right, looking for handles.

An oracle tells what to do

LR Parsing

$1: \quad e \rightarrow t + e$

$2: \quad e \rightarrow t$

$3: \quad t \rightarrow \textbf{Id} * t$

$4: \quad t \rightarrow \textbf{Id}$

stack	input	action
$\boxed{0}$	**Id** * **Id** + **Id** $	shift, goto 1

	action				goto	
	Id	$+$	$*$	$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				acc		

1. Look at state on top of stack
2. and the next input token
3. to find the next action
4. In this case, shift the token onto the stack and go to state 1.

LR Parsing

$1: \quad e \rightarrow t + e$

$2: \quad e \rightarrow t$

$3: \quad t \rightarrow \textbf{Id} * t$

$4: \quad t \rightarrow \textbf{Id}$

	action				goto	
	Id	$+$	$*$	$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				acc		

stack	input	action
$\boxed{0}$	**Id** * **Id** + **Id** $	shift, goto 1
$\boxed{0}\ \boxed{\overset{\textbf{Id}}{1}}$	* **Id** + **Id** $	shift, goto 3
$\boxed{0}\ \boxed{\overset{\textbf{Id}}{1}}\ \boxed{\overset{*}{3}}$	**Id** + **Id** $	shift, goto 1
$\boxed{0}\ \boxed{\overset{\textbf{Id}}{1}}\ \boxed{\overset{*}{3}}\ \boxed{\overset{\textbf{Id}}{1}}$	+ **Id** $	reduce w/ 4

Action is reduce with rule 4 ($t \rightarrow$ **Id**). The right side is removed from the stack to reveal state 3. The goto table in state 3 tells us to go to state 5 when we reduce a t:

stack	input	action
$\boxed{0}\ \boxed{\overset{\textbf{Id}}{1}}\ \boxed{\overset{*}{3}}\ \boxed{\overset{t}{5}}$	+ **Id** $	

LR Parsing

$1: \quad e \rightarrow t + e$

$2: \quad e \rightarrow t$

$3: \quad t \rightarrow \textbf{Id} * t$

$4: \quad t \rightarrow \textbf{Id}$

	action				goto	
	Id	$+$	$*$	$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				acc		

stack	input	action
0	**Id** * **Id** + **Id** $	shift, goto 1
0 **Id** 1	* **Id** + **Id** $	shift, goto 3
0 **Id** 1 * 3	**Id** + **Id** $	shift, goto 1
0 **Id** 1 * 3 **Id** 1	+ **Id** $	reduce w/ 4
0 **Id** 1 * 3 t 5	+ **Id** $	reduce w/ 3
0 t 2	+ **Id** $	shift, goto 4
0 t 2 + 4	**Id** $	shift, goto 1
0 t 2 + 4 **Id** 1	$	reduce w/ 4
0 t 2 + 4 t 2	$	reduce w/ 2
0 t 2 + 4 e 6	$	reduce w/ 1
0 e 7	$	accept

Constructing the SLR Parse Table

The states are places we could be in a reverse-rightmost derivation. Let's represent such a place with a dot.

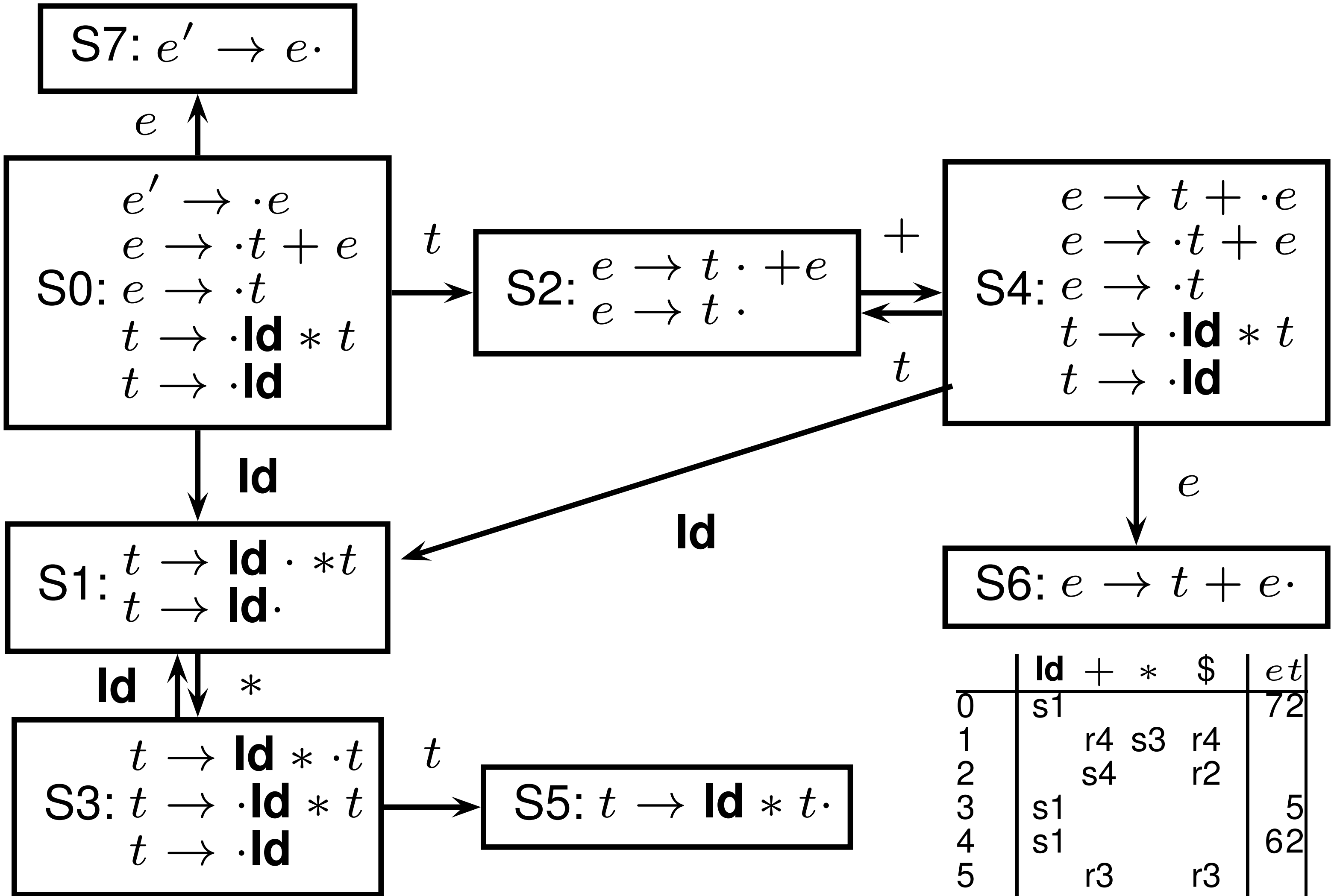
$$1: \quad e \rightarrow t + e$$
$$2: \quad e \rightarrow t$$
$$3: \quad t \rightarrow \textbf{Id} * t$$
$$4: \quad t \rightarrow \textbf{Id}$$

Say we were at the beginning ($\cdot e$). This corresponds to

$$\begin{aligned} e' &\rightarrow \cdot e \\ e &\rightarrow \cdot t + e \\ e &\rightarrow \cdot t \\ t &\rightarrow \cdot \textbf{Id} * t \\ t &\rightarrow \cdot \textbf{Id} \end{aligned}$$

The first is a placeholder. The second are the two possibilities when we're just before e. The last two are the two possibilities when we're just before t.

Constructing the SLR Parsing Table



	Id	$+$	$*$	\$	e	t
0	s1				7	2
1		r4	s3	r4		
2		s4		r2		
3	s1					5
4	s1				6	2
5		r3		r3		
6				r1		
7				acc		

The Punchline

This is a tricky, but mechanical procedure. The parser generators YACC, Bison, Cup, and others (but not ANTLR) use a modified version of this technique to generate fast bottom-up parsers.

You need to understand it to comprehend error messages:

Shift/reduce conflicts are caused by a state like

$t \to \textbf{Id} \cdot *t$

$t \to \textbf{Id} * t \cdot$

Reduce/reduce conflicts are caused by a state like

$t \to \textbf{Id} * t \cdot$

$e \to t + e \cdot$

Exercises (grammars)

1. Write unambiguous grammars for the following languages:
   - The set of all strings of a's and b's that are palindromes.
   - Strings that match the pattern $a * b*$ and have more a's than b's.
   - Strings with balanced parenthesis and square braces. Example:

     `( [ [ ] (  ( ) [ ( ) ] [ ] ) ] )`
   - The set of all strings of a's and b's such that every a is immediatelly followed by at least one b.
   - The set of all strings of a's and b's with an equal number of a's and b's.
   - The set of all strings of a's and b's with an different number of a's and b's.
   - Blocks of statements in Pascal or MH, where the semicolons (`';'`) separate the statements:

     `( statement;  ( statement ;  statement  ) ; statement )`
   - Blocks of statemens in C, where the semicolons (`';'`) follow each statement:

     `{ statement;  { statement;  statement; }  statement; }`
2. Specify the previous grammars in ANTLR notation, modifying the grammar when necessary.

Exercises (parsing)

1. Calculate NULLABLE, FIRST and FOLLOW of the non-terminal symbols in the following grammar:

$$\begin{array}{lcl} A & \rightarrow & B \mid a \\ B & \rightarrow & b \mid \varepsilon \\ C & \rightarrow & c \mid A\,B\,C \end{array}$$

2. Consider the following grammar:

$$S \rightarrow S\ S\ + \mid S\ S\ * \mid a$$

and the string $a\,a + a\,*$.

- Give a leftmost derivation for the string.
- Give a rightmost derivation for the string.
- Give a parse tree for the string.
- Is the grammar ambiguous or unambiguous? Justify your answer.
- Describe the language generated by this grammar.

3. Consider the following grammar:

$$
\begin{array}{lcl}
S & \rightarrow & c\,A\,B\,c \\
A & \rightarrow & a\,A\,a \mid c \\
B & \rightarrow & b\,B\,b \mid c
\end{array}
$$

- Calculate FIRST and FOLLOW for the non-terminal symbols.
- Construct the LL(1) parsing table and check whether it is an LL(1) grammar.

4. Calculate NULLABLE, FIRST and FOLLOW for the following grammar:

$$
\begin{array}{lcl}
S & \rightarrow & u\,B\,D\,z \\
B & \rightarrow & B\,v \mid w \\
D & \rightarrow & E\,F \\
E & \rightarrow & y \mid \varepsilon \\
F & \rightarrow & x \mid \varepsilon
\end{array}
$$

Construct the LL(1) parsing table and give evidence that this grammar is not LL(1). Modify the grammar as little as possible to make an LL(1) grammar that accepts the same language.

5. Design a table-driven top-down parser for the following grammar:

$$
\begin{array}{lcl}
S & \rightarrow & E\ \$ \\
E & \rightarrow & T + E \mid T \\
T & \rightarrow & \mathbf{num} * T \mid \mathbf{num}
\end{array}
$$

6. Design a recursive-descent parser (ANTLR style) for the following grammar:

$$\begin{array}{rcl} E & \rightarrow & T\ ('\texttt{+}'\ T)\ * \\ T & \rightarrow & F\ ('\texttt{*}'\ F)\ * \\ F & \rightarrow & '\texttt{(}'\ E\ '\texttt{)}'\ \mid \texttt{id} \end{array}$$

7. Consider the following grammar:

$$\begin{array}{rcl} G & \rightarrow & S\ \$ \\ S & \rightarrow & A\,M \\ M & \rightarrow & S \mid \varepsilon \\ A & \rightarrow & a\,E \mid b\,A\,A \\ E & \rightarrow & a\,B \mid b\,A \mid \varepsilon \\ B & \rightarrow & b\,E \mid a\,B\,B \end{array}$$

(a) Describe the language generated by the grammar.

(b) Give a parse treee for the string *abaa*\$.

(c) Is it an LL(1) grammar ? Build the parsing table and identify the conflicts.

8. Design an SLR(1) parser for the following grammar:

$$\begin{array}{lcl} S' & \rightarrow & S\ \$ \\ S & \rightarrow & V\ \texttt{;}\ S \mid \varepsilon \\ V & \rightarrow & \texttt{int id} \end{array}$$

9. Design an SLR(1) and LL(1) parsers for the following grammar:

$$\begin{array}{lcl} P & \rightarrow & E\ \$ \\ E & \rightarrow & \texttt{atom} \mid {}'\ E \mid \texttt{(}E\,E_S\texttt{)} \\ E_S & \rightarrow & E\,E_S \mid \varepsilon \end{array}$$

Give the leftmost derivation and the parse tree for the string **`(cdr '(a b c))$`**.