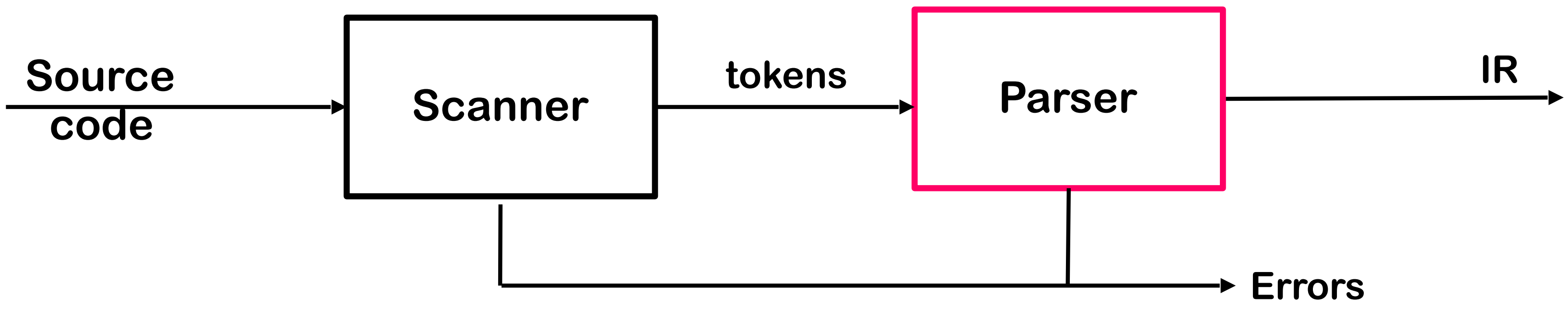
Introduction to Parsing

The Front End



Parser

- Checks the stream of <u>words</u> and their <u>parts of speech</u> (produced by the scanner) for grammatical correctness
- Determines if the input is syntactically well formed
- Guides checking at deeper levels than syntax
- Builds an IR representation of the code

The Study of Parsing

The process of discovering a derivation for some sentence

- Need a mathematical model of syntax — a grammar G
- Need an algorithm for testing membership in L(G)

Roadmap for our study of parsing

1. Context-free grammars and derivations
2. Top-down parsing
   - Generated LL(1) parsers & hand-coded recursive descent parsers
3. Bottom-up parsing
   - Generated LR(1) parsers

Why Not Use Regular Languages & DFAs?

Not all languages are regular (RL's $\subset$ CFL's $\subset$ CSL's)

You cannot construct DFA's to recognize these languages

- L = { $p^k q^k$ } (parenthesis languages)
- L = { $wcw^r \mid w \in \Sigma^*$ }

Neither of these is a regular language

To recognize these features requires an arbitrary amount of context (left or right ...)

But, this issue is somewhat subtle. You can construct DFA's for

- Strings with alternating 0's and 1's
  $(\varepsilon \mid 1) (01)^* (\varepsilon \mid 0)$
- Strings with an even number of 0's and 1's

RE's can count bounded sets and bounded differences

$\Rightarrow$ Cannot add parenthesis, brackets, begin-end pairs, ...

A More Useful Grammar Than Sheep Noise

To explore the uses of CFGs, we need a more complex grammar

0	Expr	→	Expr Op Expr
1		|	num
2		|	id
3	Op	→	+
4		|	-
5		|	*
6		|	/

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

- Such a sequence of rewrites is called a derivation
- Process of discovering a derivation is called parsing

We denote this derivation: Expr $\Rightarrow$ id – num * id

Derivations

The point of parsing is to construct a derivation

- At each step, we choose a nonterminal to replace
- Different choices can lead to different derivations

Two derivations are of interest

- Leftmost derivation — replace leftmost NT at each step
- Rightmost derivation — replace rightmost NT at each step

These are the two systematic derivations

(We don't care about randomly-ordered derivations!)

The example on the preceding slide was a leftmost derivation

- Of course, there is also a rightmost derivation
- Interestingly, it turns out to be different

Derivations

The point of parsing is to construct a derivation

A derivation consists of a series of rewrite steps

$$S \Rightarrow \gamma_0 \Rightarrow \gamma_1 \Rightarrow \gamma_2 \Rightarrow \ldots \Rightarrow \gamma_{n-1} \Rightarrow \gamma_n \Rightarrow \text{sentence}$$

- Each γ_i is a sentential form
  - If γ contains only terminal symbols, γ is a sentence in L(G)
  - If γ contains 1 or more non-terminals, γ is a sentential form
- To get γ_i from γ_{i-1}, expand some NT $A \in \gamma_{i-1}$ by using $A \rightarrow \beta$
  - Replace the occurrence of $A \in \gamma_{i-1}$ with β to get γ_i
  - In a leftmost derivation, it would be the first NT $A \in \gamma_{i-1}$

A left-sentential form occurs in a leftmost derivation

A right-sentential form occurs in a rightmost derivation

The Two Derivations for x – 2 * y

Leftmost derivation

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

Rightmost derivation

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	Expr Op <id,y>
5	Expr * <id,y>
0	Expr Op Expr * <id,y>
1	Expr Op <num,2> * <id,y>
4	Expr - <num,2> * <id,y>
2	<id,x> - <num,2> * <id,y>

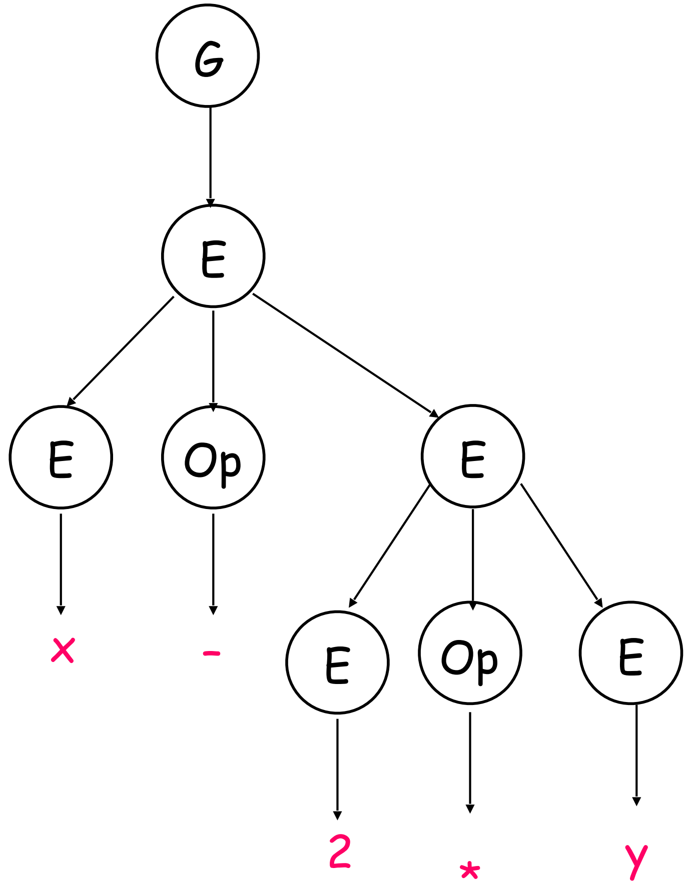
In both cases, Expr $\Rightarrow$ id – num * id

- The two derivations produce different parse trees
- The parse trees imply different evaluation orders!

Derivations and Parse Trees

Leftmost derivation

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

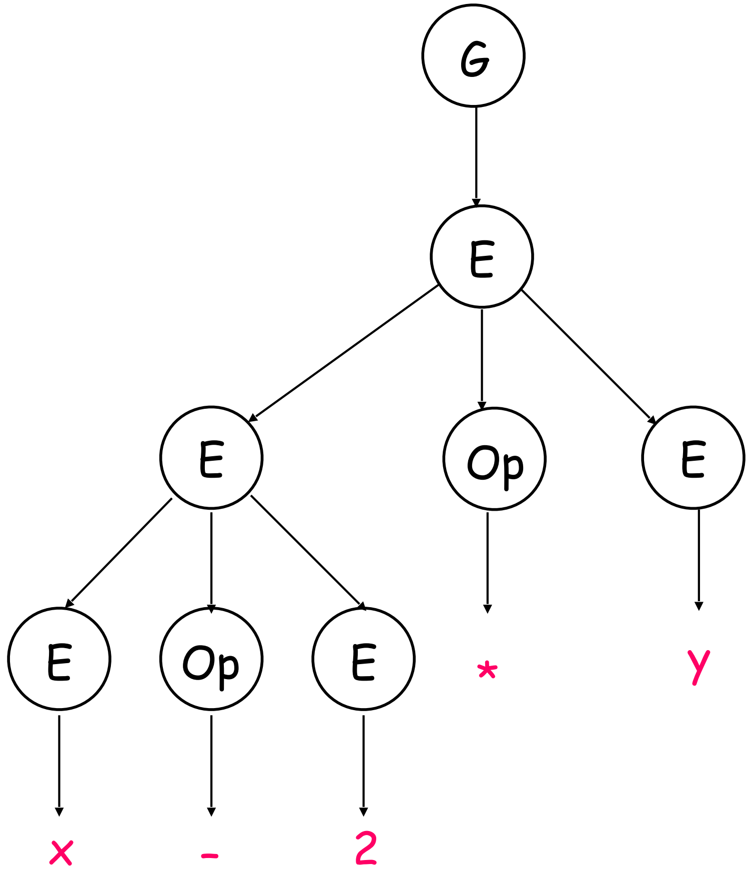


This evaluates as x - (2 * y)

Derivations and Parse Trees

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	Expr Op <id,y>
5	Expr * <id,y>
0	Expr Op Expr * <id,y>
1	Expr Op <num,2> * <id,y>
4	Expr - <num,2> * <id,y>
2	<id,x> - <num,2> * <id,y>

This evaluates as (x – 2) * y



This ambiguity is NOT good

Derivations and Precedence

These two derivations point out a problem with the grammar:

It has no notion of precedence, or implied order of evaluation

To add precedence

- Create a nonterminal for each level of precedence
- Isolate the corresponding part of the grammar
- Force the parser to recognize high precedence subexpressions first

For algebraic expressions

- Parentheses first (level 1)
- Multiplication and division, next (level 2)
- Subtraction and addition, last (level 3)

Derivations and Precedence

Adding the standard algebraic precedence produces:

	0	Goal	→	Expr
level 3	1	Expr	→	Expr + Term
	2		|	Expr - Term
level 2	3		|	Term
	4	Term	→	Term * Factor
level 1	5		|	Term / Factor
	6		|	Factor
	7	Factor	→	(Expr)
	8		|	<u>number</u>
	9		|	<u>id</u>

This grammar is slightly larger

- Takes more rewriting to reach some of the terminal symbols
- Encodes expected precedence
- Produces same parse tree under leftmost & rightmost derivations
- Correctness trumps the speed of the parser

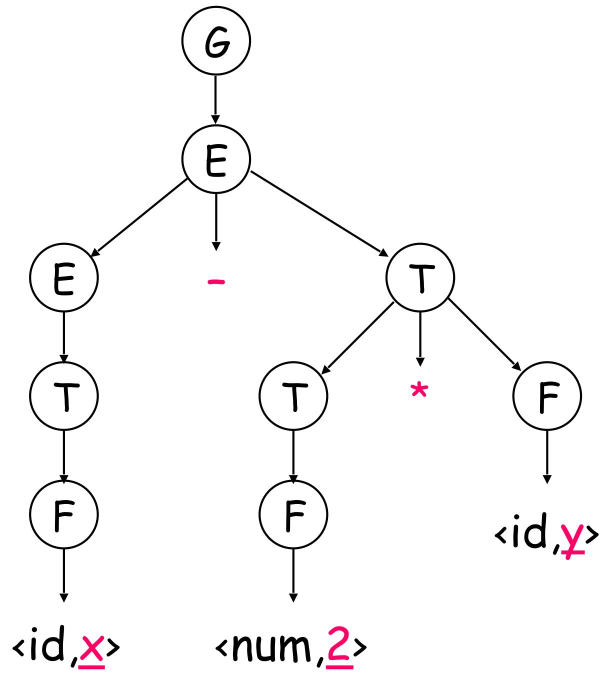
Let's see how it parses x - 2 * y

Cannot handle precedence in an RE for expressions

Introduced parentheses, too (beyond power of an RE)

Derivations and Precedence

Rule	Sentential Form
—	Goal
0	Expr
2	Expr - Term
4	Expr - Term * Factor
9	Expr - Term * <id,y>
6	Expr - Factor * <id,y>
8	Expr - <num,2> * <id,y>
3	Term - <num,2> * <id,y>
6	Factor - <num,2> * <id,y>
9	<id,x> - <num,2> * <id,y>



Its parse tree

It derives x – (2 * y), along with an appropriate parse tree.

Both the leftmost and rightmost derivations give the parse tree, because the grammar directly and explicitly encodes the desired precedence.

Ambiguous Grammars

Let's leap back to our original expression grammar.

It had other problems.

0	Expr	→	Expr Op Expr
1		|	number
2		|	id
3	Op	→	+
4		|	-
5		|	*
6		|	/

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

- This grammar allows multiple leftmost derivations for x - 2 * y
- Hard to automate derivation if > 1 choice
- The grammar is ambiguous

we have alternatives here

Two Leftmost Derivations for x - 2 * y

The Difference:

- Different productions chosen on the second step

Original choice

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
1	<id,x> - <num,2> * <id,y>

New choice

Rule	Sentential Form
—	Expr
0	Expr Op Expr
0	Expr Op Expr Op Expr
2	<id,x> Op Expr Op Expr
4	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

- Both derivations succeed in producing x - 2 * y

Two Leftmost Derivations for x – 2 * y

The Difference:

- Different productions chosen on the second step

Rule	Sentential Form
—	Expr
0	Expr Op Expr
2	<id,x> Op Expr
4	<id,x> - Expr
0	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

Rule	Sentential Form
—	Expr
0	Expr Op Expr
0	Expr Op Expr Op Expr
2	<id,x> Op Expr Op Expr
4	<id,x> - Expr Op Expr
1	<id,x> - <num,2> Op Expr
5	<id,x> - <num,2> * Expr
2	<id,x> - <num,2> * <id,y>

New choice

Different choice is possible, we are in the same situation

Ambiguous Grammars

Definitions

- If a grammar has more than one leftmost derivation for a single sentential form, the grammar is ambiguous
- If a grammar has more than one rightmost derivation for a single sentential form, the grammar is ambiguous
- The leftmost and rightmost derivations for a sentential form may differ, even in an unambiguous grammar
  - However, they must have the same parse tree!

Classic example — the if-then-else problem

Stmt → if Expr then Stmt
| if Expr then Stmt else Stmt
| … other stmts …

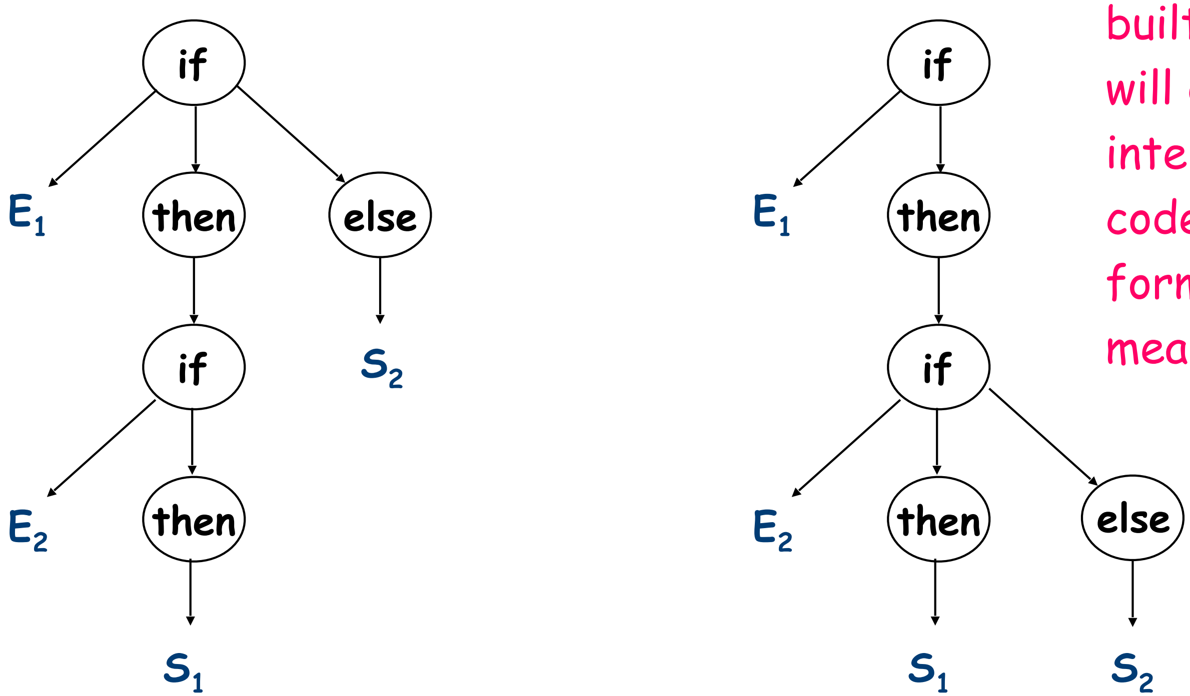
This ambiguity is inherent in the grammar

Ambiguity

$$\begin{aligned} Stmt \rightarrow\ & \underline{if}\ Expr\ \underline{then}\ Stmt \\ |\ & \underline{if}\ Expr\ \underline{then}\ Stmt\ \underline{else}\ Stmt \\ |\ & \ldots\ other\ stmts\ \ldots \end{aligned}$$

This sentential form has two derivations

$\underline{if}\ Expr_1\ \underline{then}\ \underline{if}\ Expr_2\ \underline{then}\ Stmt_1\ \underline{else}\ Stmt_2$



production 2, then production 1

production 1, then production 2

Part of the problem is that the structure built by the parser will determine the interpretation of the code, and these two forms have different meanings!

Ambiguity

The grammar forces the structure to match the desired meaning.

Removing the ambiguity

- Must rewrite the grammar to avoid generating the problem
- Match each else to innermost unmatched if (common sense rule)

0	Stmt	→	if Expr then Stmt
1		|	if Expr then WithElse else Stmt
2		|	Other Statements
3	WithElse	→	if Expr then WithElse else WithElse
4		|	Other Statements

With this grammar, example has only one rightmost derivation

Intuition: once into WithElse, we cannot generate an unmatched else

… a final if without an else can only come through rule 2 …

Ambiguity

<u>if</u> $Expr_1$ <u>then</u> <u>if</u> $Expr_2$ <u>then</u> $Stmt_1$ <u>else</u> $Stmt_2$

Rule	Sentential Form
—	Stmt
0	<u>if</u> Expr <u>then</u> Stmt
1	<u>if</u> Expr <u>then</u> <u>if</u> Expr <u>then</u> WithElse <u>else</u> Stmt
2	<u>if</u> Expr <u>then</u> <u>if</u> Expr <u>then</u> WithElse <u>else</u> S_2
4	<u>if</u> Expr <u>then</u> <u>if</u> Expr <u>then</u> S_1 <u>else</u> S_2
?	<u>if</u> Expr <u>then</u> <u>if</u> E_2 <u>then</u> S_1 <u>else</u> S_2
?	<u>if</u> E_1 <u>then</u> <u>if</u> E_2 <u>then</u> S_1 <u>else</u> S_2

Other productions to derive Exprs

This grammar has only one rightmost derivation for the example

Deeper Ambiguity

Ambiguity usually refers to confusion in the CFG

Overloading can create deeper ambiguity

    a = f(17)

In many Algol-like languages, f could be either a function or a subscripted variable

Disambiguating this one requires context

- Need values of declarations
- Really an issue of type, not context-free syntax
- Requires an extra-grammatical solution (not in CFG)
- Must handle these with a different mechanism
  - Step outside grammar rather than use a more complex grammar

Ambiguity - the Final Word

Ambiguity arises from two distinct sources

- Confusion in the context-free syntax (<u>if</u>-<u>then</u>-<u>else</u>)
- Confusion that requires context to resolve (overloading)

Resolving ambiguity

- To remove context-free ambiguity, rewrite the grammar
- To handle context-sensitive ambiguity takes cooperation
  - Knowledge of declarations, types, …
  - Accept a superset of L(G) & check it by other means (Context Sensitive analysis)
  - This is a language design problem

Sometimes, the compiler writer accepts an ambiguous grammar

- Parsing techniques that "do the right thing"
- i.e., always select the same derivation

Parsing Techniques

Top-down parsers (LL(1), recursive descent)

- Start at the root of the parse tree and grow toward leaves
- Pick a production & try to match the input
- Bad “pick” $\Rightarrow$ may need to backtrack
- Some grammars are backtrack-free (predictive parsing)

Bottom-up parsers (LR(1), operator precedence)

- Start at the leaves and grow toward root
- As input is consumed, encode possibilities in an internal state
- Start in a state valid for legal first tokens
- Bottom-up parsers handle a large class of grammars

Top-down Parsing

A top-down parser starts with the root of the parse tree

The root node is labeled with the goal symbol of the grammar

Top-down parsing algorithm:

> Construct the root node of the parse tree
>
> Repeat until lower fringe of the parse tree matches the input string
>
> 1. At a node labeled A, select a production with A on its lhs and, for each symbol on its rhs, construct the appropriate child
> 2. When a terminal symbol is added to the border and it doesn't match the border, backtrack
> 3. Find the next node to be expanded (label $\in$ NT)

The key is picking the right production in step 1

- That choice should be guided by the input string

Remember the expression grammar?

We will call this version "the classic expression grammar"

0	Goal	→	Expr
1	Expr	→	Expr + Term
2		|	Expr - Term
3		|	Term
4	Term	→	Term * Factor
5		|	Term / Factor
6		|	Factor
7	Factor	→	(Expr)
8		|	number
9		|	id

And the input x – 2 * y

Example

Let's try x - 2 * y :

Rule	Sentential Form	Input
—	Goal	↑x - 2 * y

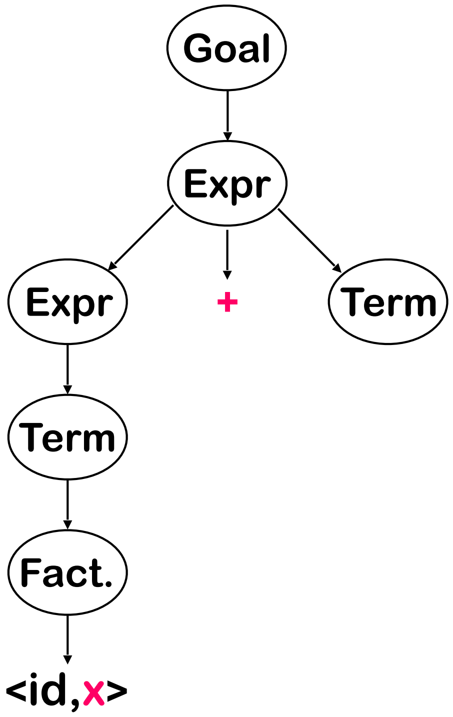
↑ is the position in the input buffer

Example

↑ is the position in the input buffer

Let's try x – 2 * y :

Rule	Sentential Form	Input
—	Goal	↑x - 2 * y
0	Expr	↑x - 2 * y
1	Expr +Term	↑x - 2 * y
3	Term +Term	↑x - 2 * y
6	Factor +Term	↑x - 2 * y
9	<id,x> +Term	↑x - 2 * y
→	<id,x> +Term	x ↑- 2 * y



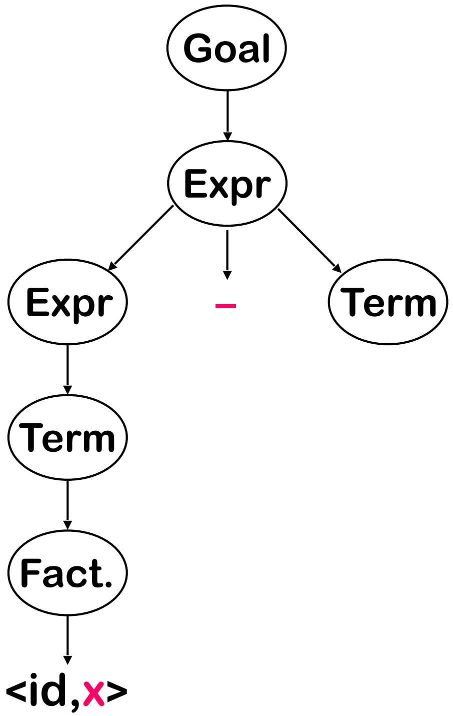
This worked well, except that "–" doesn't match "+"

The parser must backtrack to here

Example

Continuing with x – 2 * y :

Rule	Sentential Form	Input
—	Goal	↑x - 2 * y
0	Expr	↑x - 2 * y
2	Expr -Term	↑x - 2 * y
3	Term -Term	↑x - 2 * y
6	Factor -Term	↑x - 2 * y
9	<id,x> - Term	↑x - 2 * y
→	<id,x> -Term	x ↑- 2 * y
→	<id,x> -Term	x - ↑2 * y



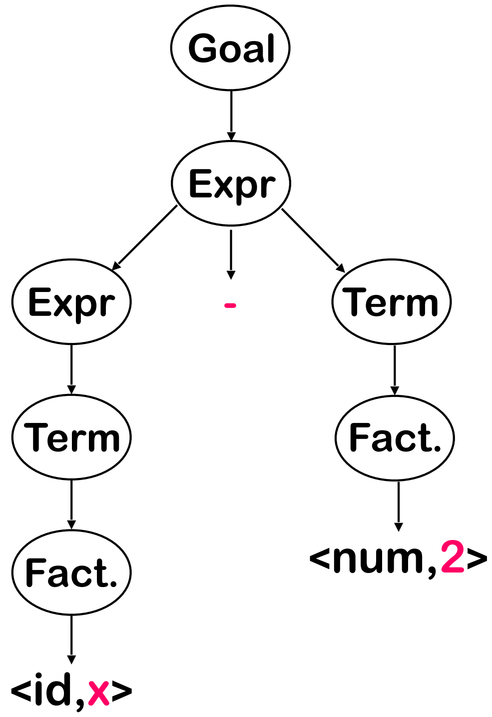
Now, "-" and "-" match

Now we can expand Term to match "2"

Example

Trying to match the "2" in x – 2 * y :

Rule	Sentential Form	Input
→	<id,x> - Term	x - ↑2 * y
6	<id,x> - Factor	x - ↑2 * y
8	<id,x> - <num,2>	x - ↑2 * y
→	<id,x> - <num,2>	x - 2 ↑* y



Where are we?

- "2" matches "2"
- We have more input, but no NTs left to expand
- The expansion terminated too soon

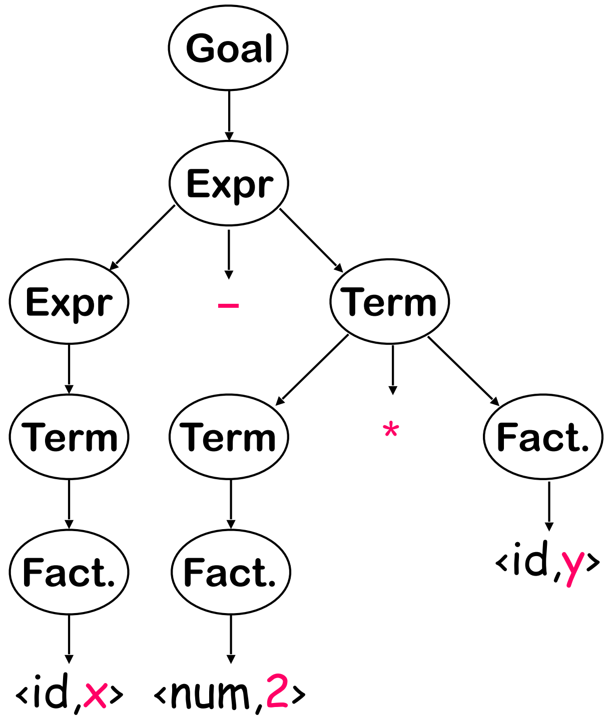
⇒ Need to backtrack

Example

The Point: The parser must make the right choice when it expands a NT. Wrong choices lead to wasted effort.

Trying again with “2” in x – 2 * y :

Rule	Sentential Form	Input
→	<id,x> - Term	x - ↑2 * y
4	<id,x> - Term * Factor	x - ↑2 * y
6	<id,x> - Factor * Factor	x - ↑2 * y
8	<id,x> - <num,2> * Factor	x - ↑2 * y
→	<id,x> - <num,2> * Factor	x - 2 ↑* y
→	<id,x> - <num,2> * Factor	x - 2 * ↑y
9	<id,x> - <num,2> * <id,y>	x - 2 * ↑y
→	<id,x> - <num,2> * <id,y>	x - 2 * y↑



This time, we matched & consumed all the input

⇒Success!

Another possible parse

Other choices for expansion are possible

Rule	Sentential Form	Input
—	Goal	↑x - 2 * y
0	Expr	↑x - 2 * y
1	Expr +Term	↑x - 2 * y
1	Expr + Term +Term	↑x - 2 * y
1	Expr + Term +Term + Term	↑x - 2 * y
1	And so on	↑x - 2 * y

Consumes no input!

This expansion doesn't terminate

- Wrong choice of expansion leads to non-termination
- Non-termination is a bad property for a parser to have
- Parser must make the right choice

Left Recursion

Top-down parsers cannot handle left-recursive grammars

Formally,

A grammar is left recursive if $\exists\, A \in NT$ such that $\exists$ a derivation $A \Rightarrow^{+} A\alpha$, for some string $\alpha \in (NT \cup T)^{+}$

Our classic expression grammar is left recursive

- This can lead to non-termination in a top-down parser
- In a top-down parser, any recursion must be right recursion
- We would like to convert the left recursion to right recursion

Non-termination is <u>always</u> a bad property in a compiler

0	Goal	¯ Expr
1	Expr	¯ Expr + Term
2		| Expr - Term
3		| Term
4	Term	¯ Term * Factor
5		| Term / Factor
6		| Factor
7	Factor	¯ (Expr)
8		| <u>number</u>
9		| <u>id</u>

Eliminating Left Recursion

To remove left recursion, we can transform the grammar

Consider a grammar fragment of the form

$$\begin{aligned} Fee &\rightarrow Fee\ \alpha \\ &\mid\ \beta \end{aligned}$$

where neither α nor β start with Fee

We can rewrite this fragment as

$$\begin{aligned} Fee &\rightarrow \beta\ Fie \\ Fie &\rightarrow \alpha\ Fie \\ &\mid\ \varepsilon \end{aligned}$$

where Fie is a new non-terminal

The new grammar defines the same language as the old grammar, using only right recursion.

Added a reference to the empty string

Eliminating Left Recursion

The expression grammar contains two cases of left recursion

Expr	→	Expr + Term	Term	→	Term * Factor
	|	Expr - Term		|	Term * Factor
	|	Term		|	Factor

Applying the transformation yields

Expr	→	Term Expr′	Term	→	Factor Term′
Expr′	→	+ Term Expr′	Term′	→	* Factor Term′
	|	- Term Expr′		|	/ Factor Term′
	|	ε		|	ε

These fragments use only right recursion

Right recursion often means right associativity. In this case, the grammar does not display any particular associative bias.

Eliminating Left Recursion

Substituting them back into the grammar yields

0	Goal	$\rightarrow$	Expr
1	Expr	$\rightarrow$	Term Expr'
2	Expr'	$\rightarrow$	+ Term Expr'
3		|	- Term Expr'
4		|	ε
5	Term	$\rightarrow$	Factor Term'
6	Term'	$\rightarrow$	* Factor Term'
7		|	/ Factor Term'
8		|	ε
9	Factor	$\rightarrow$	<u>(</u> Expr <u>)</u>
10		|	<u>number</u>
11		|	<u>id</u>

- This grammar is correct, if somewhat non-intuitive.
- It is left associative, as was the original
  - ⇒ The naïve transformation yields a right recursive grammar, which changes the implicit associativity
- A top-down parser will terminate using it.
- A top-down parser may need to backtrack with it.

Eliminating Left Recursion

The transformation eliminates immediate left recursion

What about more general, indirect left recursion ?

The general algorithm:

arrange the NTs into some order $A_1, A_2, \ldots, A_n$

for i $\leftarrow$ 1 to n

for s $\leftarrow$ 1 to i - 1

replace each production $A_i \rightarrow A_s\gamma$ with $A_i \rightarrow \delta_1\gamma \mid \delta_2\gamma \mid \ldots \mid \delta_k\gamma$,

where $A_s \rightarrow \delta_1 \mid \delta_2 \mid \ldots \mid \delta_k$ are all the current productions for A_s

eliminate any immediate left recursion on A_i

using the direct transformation

This assumes that the initial grammar has no cycles ($A_i \Rightarrow^+ A_i$),

and no epsilon productions

Eliminating Left Recursion

How does this algorithm work?

1. Impose arbitrary order on the non-terminals
2. Outer loop cycles through NT in order
3. Inner loop ensures that a production expanding A_i has no non-terminal A_s in its rhs, for $s < i$
4. Last step in outer loop converts any direct recursion on A_i to right recursion using the transformation showed earlier
5. New non-terminals are added at the end of the order & have no left recursion

At the start of the i^{th} outer loop iteration

For all $k < i$, no production that expands A_k contains a non-terminal A_s in its rhs, for $s < k$

Example

- Order of symbols: G, E, T

1. A_i = G

G → E

E → E + T

E → T

T → E * T

T → <u>id</u>

2. A_i = E

G → E

E → T E'

E' → + T E'

E' → ε

T → E * T

T → <u>id</u>

3. A_i = T, A_s = E

G → E

E → T E'

E' → + T E'

E' → ε

T → T E' * T

T → <u>id</u>

4. A_i = T

G → E

E → T E'

E' → + T E'

E' → ε

T → <u>id</u> T'

T' → E ' * T T'

T' → ε

Go to Algorithm

Picking the “Right” Production

If it picks the wrong production, a top-down parser may backtrack

Alternative is to look ahead in input & use context to pick correctly

How much lookahead is needed?

- In general, an arbitrarily large amount
- Use the Cocke-Younger, Kasami algorithm or Earley’s algorithm

Fortunately,

- Large subclasses of CFGs can be parsed with limited lookahead
- Most programming language constructs fall in those subclasses

Among the interesting subclasses are LL(1) and LR(1) grammars

We will focus, for now, on LL(1) grammars & predictive parsing