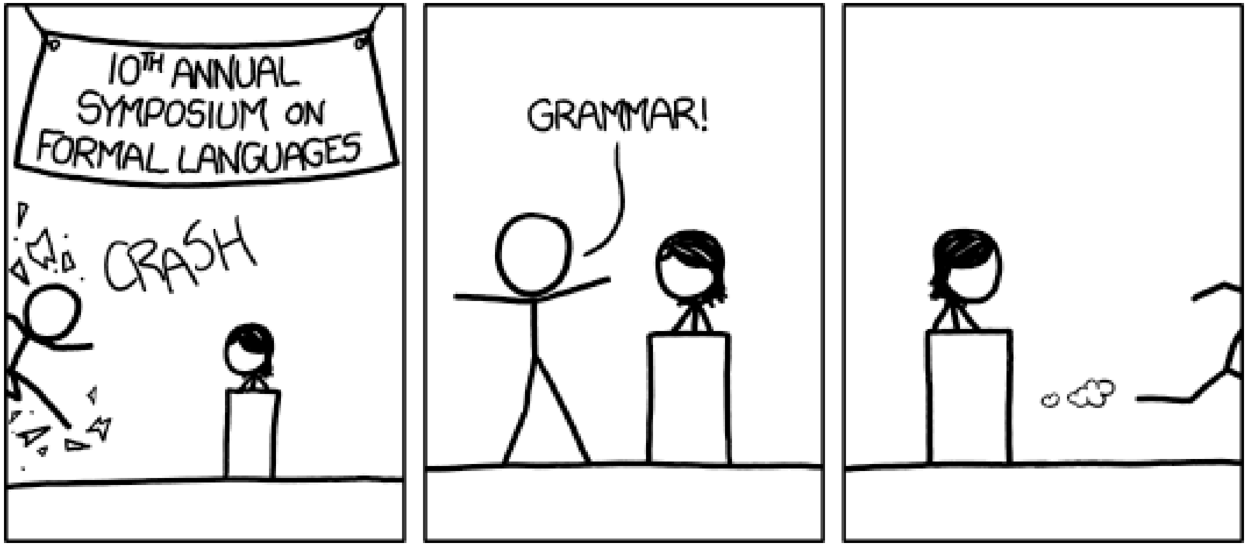
CSE 311: Foundations of Computing

Lecture 21: Context-Free Grammars



[Audience looks around]
"What is going on? There must be some context we're missing"

Recap: Regular expressions are "patterns"

ε matches the **empty string**

a matches the one character string *a*

A $\cup$ **B** matches all strings that either **A** matches or **B** matches (or both)

AB matches all strings that have a first part that **A** matches followed by a second part that **B** matches

A* matches all strings that are concatenations of any number of strings (even 0) that **A** matches, (equivalently, $A^* = \varepsilon \cup A \cup AA \cup AAA \cup \ldots$)

Example: $(00 \cup 01 \cup 10 \cup 11)^*$ corresponds to set of even length binary strings Σ, Σ^*

Fact: Now every language can be described by a regular expression (i.e. palindromes; proof later in this course)

Example Context-Free Grammars

$\Sigma = \{0, 1\}$

$S \to A \mid B$

$A \to B \mid \varepsilon$

$B \to 01$

Context-Free Grammars

- A Context-Free Grammar (CFG) is given by a finite set of substitution rules involving
  - A finite set **V** of *variables* that can be replaced
  - Alphabet **Σ** of *terminal symbols* that can't be replaced
  - One variable, usually **S**, is called the *start symbol*
  - The empty string
- The rules involving a variable **A** are written as

$$\mathbf{A} \rightarrow w_1 \mid w_2 \mid \cdots \mid w_k$$

where each w_i is a string of variables and terminals – that is $w_i \in (\mathbf{V} \cup \boldsymbol{\Sigma})^*$

How CFGs generate strings

- Begin with start symbol **S**

- If there is some variable **A** in the current string you can replace it by one of the w's in the rules for **A**
  - $\mathbf{A} \rightarrow w_1 \mid w_2 \mid \cdots \mid w_k$
  - Write this as $x\mathbf{A}y \Rightarrow xwy$
  - Repeat until no variables left

- The set of strings the CFG generates are all strings produced in this way that have no variables

Example Context-Free Grammars

$\Sigma = \{0, 1\}$

$S \to A \mid B$

$A \to B \mid \varepsilon$

$B \to 01$

$S \Rightarrow A \Rightarrow \varepsilon$

$S \Rightarrow B \Rightarrow 01$

$S \Rightarrow A \Rightarrow B \Rightarrow 01$

Example Context-Free Grammars

Example: $\mathbf{S} \rightarrow 0\mathbf{S}0 \mid 1\mathbf{S}1 \mid 0 \mid 1 \mid \varepsilon$

$$S \Rightarrow 0S0 \Rightarrow 01S10 \Rightarrow 01110$$

Example Context-Free Grammars

Example: $\mathbf{S} \to 0\mathbf{S}0 \mid 1\mathbf{S}1 \mid 0 \mid 1 \mid \varepsilon$

The set of all binary palindromes

Example Context-Free Grammars

Example: $\mathbf{S} \rightarrow 0\mathbf{S}0 \mid 1\mathbf{S}1 \mid 0 \mid 1 \mid \varepsilon$

The set of all binary palindromes

Example: $\mathbf{S} \rightarrow 0\mathbf{S} \mid \mathbf{S}1 \mid \varepsilon$

$S \Rightarrow 0S \Rightarrow 0S1 \Rightarrow 00S1 \Rightarrow 001$

Example Context-Free Grammars

Example: $\mathbf{S} \rightarrow 0\mathbf{S}0 \mid 1\mathbf{S}1 \mid 0 \mid 1 \mid \varepsilon$

The set of all binary palindromes

Example: $\mathbf{S} \rightarrow 0\mathbf{S} \mid \mathbf{S}1 \mid \varepsilon$

0*1*

Example Context-Free Grammars

Grammar for $\{0^n 1^n : n \geq 0\}$

(all strings with same # of 0's and 1's with all 0's before 1's)

$$S \to 0S1 \mid \varepsilon$$

$$S \Rightarrow 0S1 \Rightarrow 00S11 \Rightarrow 000S111$$
$$\Rightarrow 000111$$

Example Context-Free Grammars

Grammar for $\{0^n 1^n : n \geq 0\}$

(all strings with same # of 0's and 1's with all 0's before 1's)

$$\mathbf{S} \rightarrow 0\mathbf{S}1 \mid \varepsilon$$

Example Context-Free Grammars

Grammar for $\{0^n 1^n : n \geq 0\}$

(all strings with same # of 0's and 1's with all 0's before 1's)

$$S \rightarrow 0S1 \mid \varepsilon$$

$S \Rightarrow SS \Rightarrow (S)S$

$((S))S$

$()S$

$()$

Example: $S \rightarrow (S) \mid SS \mid \varepsilon$

$S \rightarrow () \mid SS \mid \varepsilon$

$A \rightarrow (A) \mid AA$

$()$, $()()$, $()()()$

$(())$

Example Context-Free Grammars

Grammar for $\{0^n 1^n : n \geq 0\}$

(all strings with same # of 0's and 1's with all 0's before 1's)

$$\mathbf{S} \rightarrow 0\mathbf{S}1 \mid \varepsilon$$

Example: $\mathbf{S} \rightarrow (\mathbf{S}) \mid \mathbf{SS} \mid \varepsilon$

The set of all strings of matched parentheses

Simple Arithmetic Expressions

$$E \rightarrow E{+}E \mid E{*}E \mid (E) \mid x \mid y \mid z \mid 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$

Generate $(2*x) + y$

Simple Arithmetic Expressions

$$E \rightarrow E{+}E \mid E*E \mid (E) \mid x \mid y \mid z \mid 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$

Generate $(2*x) + y$

$$E \Rightarrow E{+}E \Rightarrow (E){+}E \Rightarrow (E*E){+}E \Rightarrow (2*E){+}E \Rightarrow (2*x){+}E \Rightarrow (2*x){+}y$$

Lecture 21 Activity

You will be assigned to breakout rooms. Please:

- Introduce yourself
- Choose someone to share their screen, showing this PDF
- Consider the CFG (start symbol is E)

  $E \rightarrow E{+}E \mid E{*}E \mid (E) \mid x \mid y \mid z \mid 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$
- Example to generate (2*x)+y

  $E \Rightarrow E{+}E \Rightarrow (E){+}E \Rightarrow (E{*}E){+}E \Rightarrow (2{*}E){+}E \Rightarrow (2{*}x){+}E \Rightarrow (2{*}x){+}y$

- Find two different ways to generate 2 + 3*4

Fill out the poll everywhere for Activity Credit!
Go to pollev.com/philipmg and login with your UW identity

Lecture 21 Activity

You will be assigned to breakout rooms. Please:

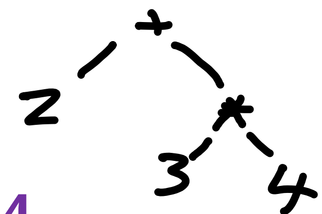
- Introduce yourself
- Choose someone to share their screen, showing this PDF
- Consider the CFG (start symbol is E)

  $E \rightarrow E+E \mid E*E \mid (E) \mid x \mid y \mid z \mid 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$

- Example to generate (2*x)+y

  $E \Rightarrow E+E \Rightarrow (E)+E \Rightarrow (E*E)+E \Rightarrow (2*E)+E \Rightarrow (2*x)+E \Rightarrow (2*x)+y$

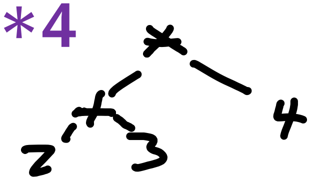
- Find two different ways to generate 2 + 3*4



$\Rightarrow E+E\cdot E$

$E \Rightarrow E+E \Rightarrow 2+E \Rightarrow 2+E*E \Rightarrow 2+3*E \Rightarrow \underline{2+3*4}$

$E \Rightarrow E*E \Rightarrow E+E*E \Rightarrow 2+E*E \Rightarrow 2+3*E \Rightarrow 2+3*4$

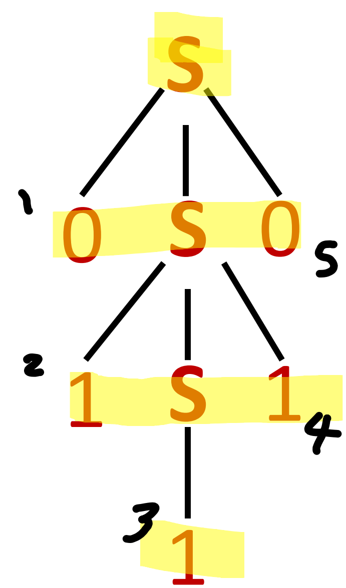


Parse Trees

Suppose that grammar **G** generates a string **x**

- A *parse tree* of **x** for **G** has
  - Root labeled **S** (start symbol of **G**)
  - The children of any node labeled **A** are labeled by symbols of **w** left-to-right for some rule $\mathbf{A} \to \mathbf{w}$
  - The symbols of **x** label the leaves ordered left-to-right

$$S \to 0S0 \mid 1S1 \mid 0 \mid 1 \mid \varepsilon$$



Parse tree of 01110

CFGs and recursively-defined sets of strings

- A CFG with the start symbol **S** as its only variable recursively defines the set of strings of terminals that **S** can generate

- A CFG with more than one variable is a simultaneous recursive definition of the sets of strings generated by *each* of its variables
  - Sometimes necessary to use more than one

building precedence in simple arithmetic expressions

- **E** – expression   (start symbol)
- **T** – term   **F** – factor   **I** – identifier   **N** - number

$$\mathbf{E} \rightarrow \mathbf{T} \mid \mathbf{E}+\mathbf{T}$$

$$\mathbf{T} \rightarrow \mathbf{F} \mid \mathbf{F}*\mathbf{T}$$

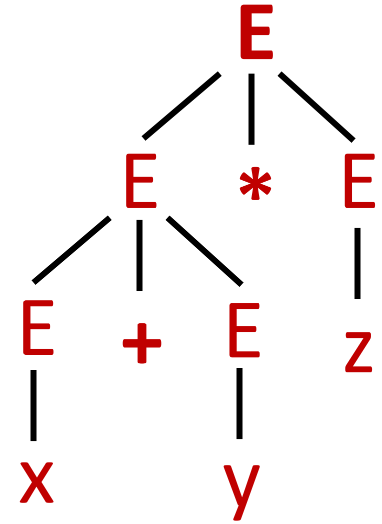
$$\mathbf{F} \rightarrow (\mathbf{E}) \mid \mathbf{I} \mid \mathbf{N}$$

$$\mathbf{I} \rightarrow x \mid y \mid z$$

$$\mathbf{N} \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$

$x + y \cdot z$

No longer allows:



building precedence in simple arithmetic expressions

- **E** – expression (start symbol)
- **T** – term **F** – factor **I** – identifier **N** - number

$$\mathbf{E} \rightarrow \mathbf{T} \mid \mathbf{E}+\mathbf{T}$$

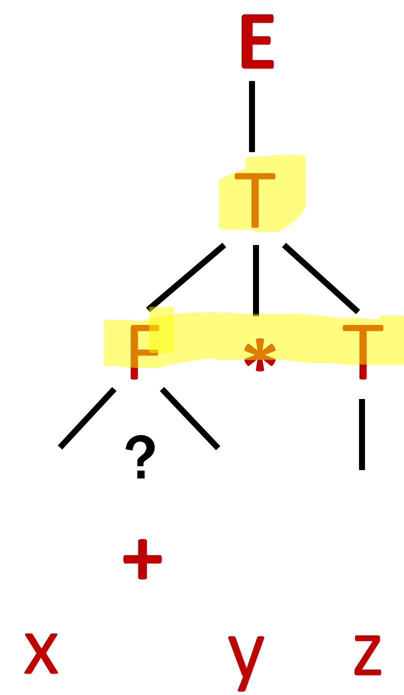
$$\mathbf{T} \rightarrow \mathbf{F} \mid \mathbf{F}*\mathbf{T}$$

$$\mathbf{F} \rightarrow (\mathbf{E}) \mid \mathbf{I} \mid \mathbf{N}$$

$$\mathbf{I} \rightarrow x \mid y \mid z$$

$$\mathbf{N} \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$

$(x+y)\cdot z$



building precedence in simple arithmetic expressions

- **E** – expression (start symbol)
- **T** – term **F** – factor **I** – identifier **N** - number

$$\mathbf{E} \rightarrow \mathbf{T} \mid \mathbf{E}+\mathbf{T}$$

$$\mathbf{T} \rightarrow \mathbf{F} \mid \mathbf{F}*\mathbf{T}$$

$$\mathbf{F} \rightarrow (\mathbf{E}) \mid \mathbf{I} \mid \mathbf{N}$$

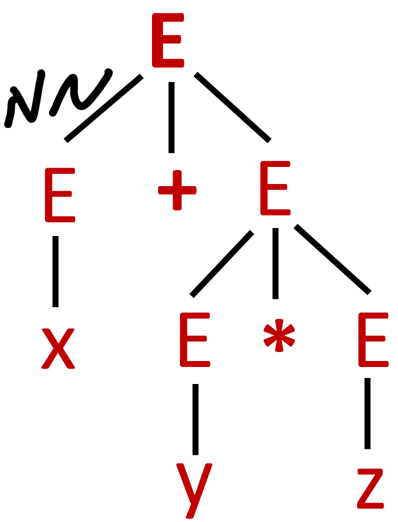
$$\mathbf{I} \rightarrow x \mid y \mid z$$

$$\mathbf{N} \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9 \mid NN$$

22 + x

005 + x

Still allows:



building precedence in simple arithmetic expressions

- **E** – expression (start symbol)
- **T** – term **F** – factor **I** – identifier **N** - number

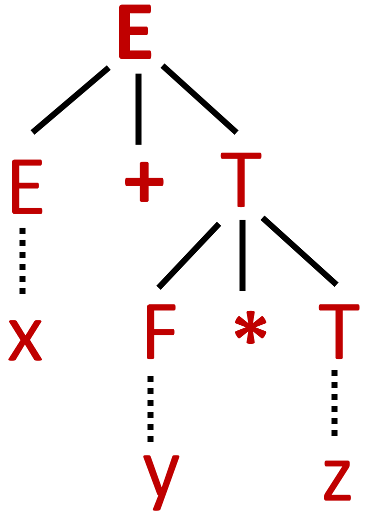
$$\mathbf{E} \rightarrow \mathbf{T} \mid \mathbf{E}+\mathbf{T}$$

$$\mathbf{T} \rightarrow \mathbf{F} \mid \mathbf{F}*\mathbf{T}$$

$$\mathbf{F} \rightarrow (\mathbf{E}) \mid \mathbf{I} \mid \mathbf{N}$$

$$\mathbf{I} \rightarrow x \mid y \mid z$$

$$\mathbf{N} \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$$



CFGs and regular expressions

Theorem: For any set of strings (language) A described by a regular expression, there is a CFG that recognizes A.

Proof idea: Structural induction based on the recursive definition of regular expressions…

Regular Expressions over Σ

- **Basis:**
  - $\varnothing$, ε are regular expressions
  - ***a*** is a regular expression for any $a \in \Sigma$
- **Recursive step:**
  - If **A** and **B** are regular expressions then so are:

    (A $\cup$ B)

    (AB)

    A*

CFGs are more general than REs

- CFG to match RE $\boldsymbol{\varepsilon}$

  S $\rightarrow \varepsilon$

- CFG to match RE ***a*** (for any $a \in \Sigma$)

  S $\rightarrow$ a

CFGs are more general than REs

- CFG to match RE **$\boldsymbol{\varepsilon}$**

  S $\rightarrow \varepsilon$

- CFG to match RE ***a*** (for any $a \in \Sigma$)

  S $\rightarrow$ a

CFGs are more general than REs

Suppose CFG with start symbol $\mathbf{S_1}$ matches RE **A**
CFG with start symbol $\mathbf{S_2}$ matches RE **B**

- CFG to match RE **A** $\cup$ **B**

$$\mathbf{S} \rightarrow \mathbf{S_1} \mid \mathbf{S_2}$$

- CFG to match RE **AB**

$$\mathbf{S} \rightarrow \mathbf{S}_1\, \mathbf{S}_2$$

CFGs are more general than REs

Suppose CFG with start symbol $\mathbf{S_1}$ matches RE $\mathbf{A}$

- CFG to match RE $\mathbf{A}^*$ (= $\varepsilon \cup \mathbf{A} \cup \mathbf{AA} \cup \mathbf{AAA} \cup \ldots$)

$$\mathbf{S} \rightarrow \mathbf{S_1}\,\mathbf{S} \mid \varepsilon$$

Backus-Naur Form (The same thing…)

BNF (Backus-Naur Form) grammars

- Originally used to define programming languages
- Variables denoted by long names in angle brackets, e.g.

  <identifier>, <if-then-else-statement>, <assignment-statement>, <condition>

  $::=$ used instead of $\rightarrow$

BNF for C

```
statement:
  ((identifier | "case" constant-expression | "default") ":")*
  (expression? ";" |
   block |
   "if" "(" expression ")" statement |
   "if" "(" expression ")" statement "else" statement |
   "switch" "(" expression ")" statement |
   "while" "(" expression ")" statement |
   "do" statement "while" "(" expression ")" ";" |
   "for" "(" expression? ";" expression? ";" expression? ")" statement |
   "goto" identifier ";" |
   "continue" ";" |
   "break" ";" |
   "return" expression? ";"
  )

block: "{" declaration* statement* "}"

expression:
  assignment-expression%

assignment-expression: (
    unary-expression (
      "=" | "*=" | "/=" | "%=" | "+=" | "-=" | "<<=" | ">>=" | "&=" |
      "^=" | "|="
    )
  )* conditional-expression

conditional-expression:
  logical-OR-expression ( "?" expression ":" conditional-expression )?
```

Parse Trees

Back to middle school:

<sentence>::=<noun phrase><verb phrase>

<noun phrase>::==<article><adjective><noun>

<verb phrase>::=<verb><adverb>|<verb><object>

<object>::=<noun phrase>

Parse:

The yellow duck squeaked loudly

The red truck hit a parked car