Programming Languages & Translators

PARSING

Baishakhi Ray

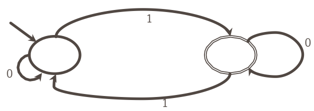
Fall 2018

These slides are motivated from Prof. Alex Aiken: Compilers (Stanford)

Languages and Automata

- Formal languages are very important in CS
  - Especially in programming languages
- Regular Languages
  - Weakest formal languages that are widely used
  - Many applications
- Many Languages are not regular

Automata that accepts odd numbers of 1



How many 1s it has accepted?

- Only solution is duplicate state

Automata does not have any memory

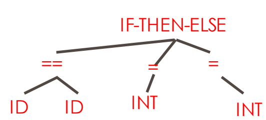
Intro to Parsing

- Regular Languages
  - Weakest formal languages that are widely used
  - Many applications
- Consider the language $\{(^i\ )^i \mid i \geq 0\}$
  - (), (( )), ((( )))
  - ((1 + 2) * 3)
- Nesting structures
  - if .. if.. else.. else..

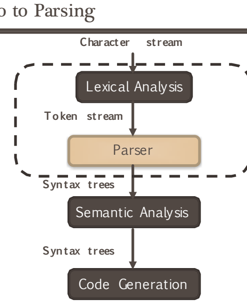
Regular languages cannot handle well

Intro to Parsing

- Input: if(x==y) 1 else 2;
- Parser Input (Lexical Input):
  KEY(IF) '(' ID(x) OP('==') ')' INT(1) KEY(ELSE) INT(2) ';'
- Parser Output



Intro to Parsing



- Nor every strings of tokens are valid
- Parser must distinguish between valid and invalid token strings.
- We need
  - A Language: to describe valid string
  - A method: to distinguish valid from invalid.

Context Free Grammar

- A CFG consists of
  - A set of terminal T
  - A set of non-terminal N
  - A start symbol S (S ϵ N)
  - A set of production rules
    - X -> $Y_1 \ldots Y_N$
    - X ϵ N
    - Y_i ϵ {N, T, ε}

- Ex: S -> (S) | ε
  - N = {S}
  - T = { (,) , ε}

Context Free Grammar

1. Begin with a string with only the start symbol S
2. Replace a non-terminal X with in the string by the RHS of some production rule: X->$Y_1 \ldots Y_n$
3. Repeat 2 again and again until there are no non-terminals

$X_1 \ldots X_i\ \underline{X}\ X_{i+1} \ldots X_n$ -> $X_1 \ldots X_i\ Y_1 \ldots Y_k\ X_{i+1} \ldots X_n$

For the production rule X -> $Y_1 \ldots Y_k$

$$\propto_0 \rightarrow \propto_1 \rightarrow \ldots \rightarrow \propto_n$$

$$\propto_0 \overset{*}{\rightarrow} \propto_n, n \geq 0$$

Context Free Grammar

- Let G be a CFG with start symbol S. Then the language L(G) of G is:

$$\{a_1 \ldots \ldots an | \forall_i\ ai \ \in T \wedge S \overset{*}{\rightarrow} a_1 a_2 \ldots \ldots an\}$$

Context Free Grammar

- There are no rules to replace terminals.
- Once generated, terminals are permanent
- Terminals ought to be tokens of programming languages
- Context-free grammars are a natural notation for this recursive structure

CFG: Simple Arithmetic expression

E → E + E
 | E * E
 | (E)
 | id

Languages can be generated: id, (id), (id + id) * id, …

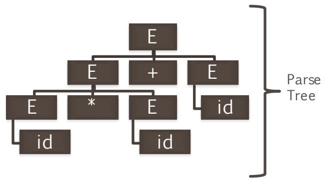
Derivation

- A derivation is a sequence of production
  - S -> … -> … ->

- A derivation can be drawn as a tree
  - Start symbol is tree's root
  - For a production X -> $Y_1 \ldots Y_n$, add children $Y_1 \ldots Y_n$ to node X

- Grammar
  - E -> E + E | E * E | (E) | id
- String
  - id * id + id
- Derivation

E -> E + E
  -> E * E + E
  -> id * E + E
  -> id * id + E
  -> id * id + id



Parse Tree

- A parse tree has
  - Terminals at the leaves
  - Non-terminals at the interior nodes
- An in-order traversal of the leaves is the original input
- The parse tree shows the association of operations, the input string does not

Parse Tree

- Left-most derivation
  - At each step, replace the left-most non-terminal

E -> E + E
  -> E * E + E
  -> id * E + E
  -> id * id + E
  -> id * id + id

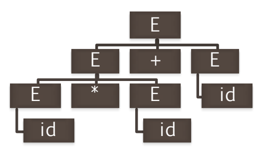
- Right-most derivation
  - At each step, replace the right-most non-terminal

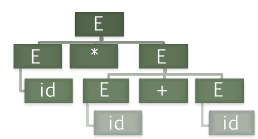
E -> E + E
  -> E + id
  -> E * E + id
  -> E * id + id
  -> id * id + id

Note that, right-most and left-most derivations have the same parse tree

Ambiguity

- Grammar
  - E -> E + E | E * E | (E) | id
- String
  - id * id + id





Ambiguity

- A grammar is ambiguous if it has more than one parse tree for a string
  - There are more than one right-most or left-most derivation for some string
- Ambiguity is bad
  - Leaves meaning for some programs ill-defined

Error Handling

- Purpose of the compiler is
  - To detect non-valid programs
  - To translate the valid ones
- Many kinds of possible errors (e.g., in C)

Error Kind	Example	Detected by
Lexical	... $ ...	Lexer
Syntax	... x*%...	Parser
Semantic	... int x; y = x(3);...	Type Checker
Correctness	your program	tester/user

Error Handling

- Error Handler should
  - Recover errors accurately and quickly
  - Recover from an error quickly
  - Not slow down compilation of valid code
- Types of Error Handling
  - Panic mode
  - Error productions
  - Automatic local or global correction

Panic Mode Error Handling

- Panic mode is simplest and most popular method
- When an error is detected
  - Discard tokens until one with a clear role is found
  - Continue from there
- Typically looks for “synchronizing” tokens
  - Typically the statement of expression terminators

Panic Mode Error Handling

- Example:
  - (1 + + 2) + 3
- Panic-mode recovery:
  - Skip ahead to the next integer and then continue
- Bison: use the special terminal error to describe how much input to skip
  - E -> $\underbrace{\text{int | E + E | (E)}}_{\text{Normal mode}}$ | $\underbrace{\text{error int | (error)}}_{\text{Error mode}}$

Error Productions

- Specify known common mistakes in the grammar
- Example:
  - Write 5x instead of 5 * x
  - Add production rule E -> .. | E E
- Disadvantages
  - complicates the grammar

Error Corrections

- Idea: find a correct “nearby” program
  - Try token insertions and deletions (goal: minimize edit distance)
  - Exhaustive search
- Disadvantages
  - Hard to implement
  - Slows down parsing of correct programs
  - “Nearby” is not necessarily “the intended” program

Error Corrections

- Past
  - Slow recompilation cycle (even once a day)
  - Find as many errors in once cycle as possible
- Disadvantages
  - Quick recompilation cycle
  - Users tend to correct one error/cycle
  - Complex error recovery is less compelling

Abstract Syntax Trees

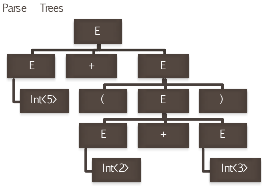
- A parser traces the derivation of a sequence of tokens
- But the rest of the compiler needs a structural representation of the program
- Abstract Syntax Trees
  - Like parse trees but ignore some details
  - Abbreviated as AST

Abstract Syntax Trees

- Grammar
  - E -> int | (E) | E + E
- String
  - 5 + (2 + 3)
- After lexical analysis
  - Int<5> '+' '(' Int<2> '+' Int<3> ')'

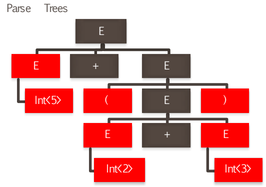
Abstract Syntax Trees: 5 + (2 + 3)

Parse Trees



Abstract Syntax Trees: 5 + (2 + 3)

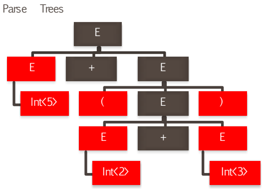
Parse Trees



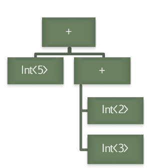
- Have too much information
  - Parentheses
  - Single-successor nodes

Abstract Syntax Trees: 5 + (2 + 3)

Parse Trees



AST



- Have too much information
  - Parentheses
  - Single-successor nodes

- ASTs capture the nesting structure
- But abstracts from the concrete syntax
  - More compact and easier to use

Disadvantages of ASTs

- AST has many similar forms
  - E.g., for, while, repeat...until
  - E.g., if, ?:, switch
- Expressions in AST may be complex, nested
  - (x * y) + (z > 5 ? 12 * z : z + 20)
- Want simpler representation for analysis
  - ...at least, for dataflow analysis

30

Parsing algorithm: Recursive Descent Parsing

- The parse tree is constructed
  - From the top
  - From left to right
- Terminals are seen in order of appearance in the token stream

Parsing algorithm: Recursive Descent Parsing

- Grammar:
  - E -> T | T + E
  - T -> int | int * T | (E)
- Token Stream: (int<5>)
- Start with top level non-terminal E
  - Try the rules for E in order

Recursive Descent Parsing Example

E -> **T** | T + E

T -> **int** | int * T | (E)

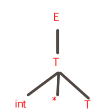
mismatch: int does not match arrowhead (
backtrack

(int<5>)

Recursive Descent Parsing Example

E -> **T** | T + E

T -> int | **int * T** | (E)



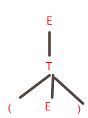
backtrack

(int<5>)

Recursive Descent Parsing Example

E -> **T** | T + E

T -> int | int * T | **(E)**



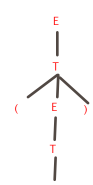
Match! Advance input

(int<5>)

Recursive Descent Parsing Example

E -> **T** | T + E

T -> **int** | int * T | (E)



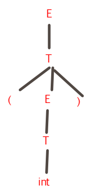
Match! Advance input

(int<5>)

Recursive Descent Parsing Example

E -> T | T + E

T -> int | int * T | (E)



Match! Advance input

(int<5>)

A Recursive Descent Parser. Preliminaries

- Let TOKEN be the type of tokens
  - Special tokens INT, OPEN, CLOSE, PLUS, TIMES
- Let the global **next** point to the next token

A (Limited) Recursive Descent Parser

- Define boolean functions that check the token string for a match of
  - A given token terminal
    `bool term (TOKEN tok) { return *next++ == tok; }`
  - The n^{th} production of S:
    `bool Sn() { … }`
  - Try all productions of S:
    `bool S() { … }`

A (Limited) Recursive Descent Parser

- For production E → T
  `bool E1() { return T(); }`
- For production E → T + E
  `bool E2() { return T() && term(PLUS) && E(); }`
- For all productions of E (with backtracking)

```
bool E() {
  TOKEN *save = next;
  return (next = save, E1( )) || (next = save, E2( ));
}
```

A (Limited) Recursive Descent Parser (4)

- Functions for non-terminal T

```
bool T1() { return term(INT); }
bool T2() { return term(INT) && term(TIMES) && T(); }
bool T3() { return term(OPEN) && E() && term(CLOSE); }

bool T() {
    TOKEN *save = next;
    return (next = save, T1())
        || (next = save, T2())
        || (next = save, T3());
}
```

Recursive Descent Parsing

- To start the parser
  - Initialize next to point to first token
  - Invoke E() · Notice how this simulates the example parse ·

Example

Grammar:

E → T |T + E
T → int | int * T | (E)

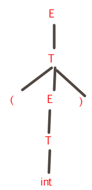
Input: (int)

Code:

```
bool term(TOKEN tok) { return *next++ == tok; }

bool E1() { return T(); }
bool E2() { return T() && term(PLUS) && E(); }
bool E() {TOKEN *save = next;
        return (next = save, E1()) || (next = save, E2()); }

bool T1() { return term(INT); }
bool T2() { return term(INT) && term(TIMES) && T(); }
bool T3() { return term(OPEN) && E() && term(CLOSE); }
bool T() { TOKEN *save = next;
    return (next = save, T1())
        || (next = save, T2())
        || (next = save, T3()); }
```



When Recursive Descent Does Not Work

Grammar:

E → T |T + E
T → int | int * T | (E)

Input: int * int

Code:

```
bool term(TOKEN tok) { return *next++ == tok; }

bool E1() { return T(); }
bool E2() { return T() && term(PLUS) && E(); }
bool E() {TOKEN *save = next;
        return (next = save, E1()) || (next = save, E2()); }

bool T1() { return term(INT); }
bool T2() { return term(INT) && term(TIMES) && T(); }
bool T3() { return term(OPEN) && E() && term(CLOSE); }
bool T() { TOKEN *save = next;
    return (next = save, T1())
        || (next = save, T2())
        || (next = save, T3()); }
```

Recursive Descent Parsing: Limitation

- If production for non-terminal X succeeds
  - Cannot backtrack to try different production for X later
- General recursive descent algorithms support such full backtracking
  - Can implement any grammar
- Presented RDA is not general
  - But easy to implement
- Sufficient for grammars where for any non-terminal at most one production can succeed
- The grammar can be rewritten to work with the presented algorithm
  - By left factoring

Left Factoring

A -> $\alpha\beta_1$ | $\alpha\beta_2$

- The input begins with a nonempty string derived from α, we do not know whether to expand A to $\alpha\beta_1$ or $\alpha\beta_2$.
- We can defer the decision by expanding A to αA'.
- Then, after seeing the input derived from α, we expand A' to β_1 or β_2 (left-factored)
- The original productions become:

A -> $\alpha A'$, A' -> β_1 | β_2

When Recursive Descent Does Not Work

- Consider a production S → S a
  bool S1() { return S() && term(a); }
  bool S() { return S1(); }
- S() goes into an infinite loop
- A left-recursive grammar has a non-terminal S

  S →+ Sα for some α
- Recursive descent does not work for left recursive grammar

Elimination of Left Recursion

- Consider the left-recursive grammar

  S → S α | β
- S generates all strings starting with a β and followed by a number of α
- Can rewrite using right-recursion

  S → β S'

  S' → α S' | ε

More Elimination of Left-Recursion

- In general
  $S \rightarrow S\ \alpha_1 \mid \ldots \mid S\ \alpha_n \mid \beta_1 \mid \ldots \mid \beta_m$
- All strings derived from S start with one of $\beta_1,\ldots,\beta_m$ and continue with several instances of $\alpha_1,\ldots,\alpha_n$
- Rewrite as
  $S \rightarrow \beta_1\ S' \mid \ldots \mid \beta_m\ S'$
  $S' \rightarrow \alpha_1\ S' \mid \ldots \mid \alpha_n\ S' \mid \varepsilon$

General Left Recursion

- The grammar
  $S \rightarrow A\ \alpha \mid \delta$
  $A \rightarrow S\ \beta$
  is also left-recursive because
  $S \rightarrow^{+} S\ \beta\ \alpha$
- This left-recursion can also be eliminated

Summary of Recursive Descent

- Simple and general parsing strategy
  - Left-recursion must be eliminated first
  - … but that can be done automatically
- Unpopular because of backtracking
  - Thought to be too inefficient
- In practice, backtracking is eliminated by restricting the grammar

Predictive Parsers

- Like recursive-descent but parser can "predict" which production to use
  - By looking at the next few tokens
  - No backtracking
- Predictive parsers accept LL(k) grammars
  - L means "left-to-right" scan of input
  - L means "leftmost derivation"
  - k means "predict based on k tokens of lookahead"
  - In practice, LL(1) is used

LL(1) vs. Recursive Descent

- In recursive-descent
  - At each step, many choices of production to use
  - Backtracking used to undo bad choices
- In LL(1)
  - At each step, only one choice of production
  - That is
    - When a non-terminal A is leftmost in a derivation
    - The next input symbol is t
    - There is a unique production $A \rightarrow \alpha$ to use
      - Or no production to use (an error state)
- LL(1) is a recursive descent variant without backtracking

Predictive Parsing and Left Factoring

- Recall the grammar
  $E \rightarrow T + E \mid T$
  $T \rightarrow \text{int} \mid \text{int} * T \mid (E)$
- Hard to predict because
  - For T two productions start with int
  - For E it is not clear how to predict
- We need to left-factor the grammar

Left-Factoring Example

- Grammar
  E → T + E | T
  T → int | int * T | (E)
- Factor out common prefixes of productions
  E → T X
  X → + E | ε
  T → (E) | int Y
  Y → * T | ε

LL(1) Parsing Table Example

- Left-factored grammar
  E → T X
  X → + E | ε
  T → (E) | int Y
  Y → * T | ε
- The LL(1) parsing table:

		next input tokens					
Left-most non-terminals		int	*	+	(	)	$
	E	TX			TX		
	X			+E		ε	ε
	T	int Y			(E)		
	Y		*T	ε		ε	ε

LL(1) Parsing Table Example (Cont.)

- Consider the [E, int] entry
  - "When current non-terminal is E and next input is int, use production E → T X"
  - This can generate an int in the first position
- Consider the [Y,+] entry
  - "When current non-terminal is Y and current token is +, get rid of Y"
  - Y can be followed by + only if Y → ε

LL(1) Parsing Tables. Errors

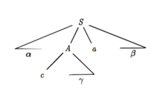
- Blank entries indicate error situations
- Consider the [E,*] entry
  - "There is no way to derive a string starting with * from non-terminal E"

Using Parsing Tables

- Method similar to recursive descent, except
  - For the leftmost non-terminal S
  - We look at the next input token a
  - And choose the production shown at [S,a]
- A stack records frontier of parse tree
  - Non-terminals that have yet to be expanded
  - Terminals that have yet to match against the input
  - Top of stack = leftmost pending terminal or non-terminal
- Reject on reaching error state
- Accept on end of input & empty stack

First & Follow

- During top down parsing, FIRST and FOLLOW allow us to choose which production to apply, based on the next input symbol.
- FIRST(α), α is any string of grammar symbols
  - A set of **terminals** that begin strings derived from α.
  - If $\alpha \xrightarrow{*} \epsilon$, then ϵ is in FIRST(α).
  - if $\alpha \xrightarrow{*} cY$, the c is in FIRST($\alpha$).



- FOLLOW(A), A is a **nonterminal**
  - the set of **terminals** that can appear immediately to the right of A
  - A set of terminals "a" such that $S \xrightarrow{*} \alpha A a \beta$ for some α and β.

Constructing Parsing Tables: The Intuition

- Consider non-terminal A, production A → α, & token t
- T[A,t] = α in two cases:
- If α →* t β
  - α can derive a t in the first position
  - We say that t ∈ First(α)
- If A → α and α →* ε and S →* β A t δ
  - Useful if stack has A, input is t, and A cannot derive t
  - In this case only option is to get rid of A (by deriving ε)
  - We say t ∈ Follow(A)

Computing First Sets

- Definition

  First(X) = { t | X →* tα} ∪ {ε | X →* ε} , X can be single terminal, single non-terminal, or string including both
- Algorithm sketch:

1. First(t) = { t } , t is terminal
2. ε ∈ First(X)
   - if X → ε
   - if X → $A_1 \ldots A_n$ and ε ∈ First(A_i) for $1 \le i \le n$
3. First(α) ⊆ First(X) if X → $A_1 \ldots A_n$ α
   - ε ∈ First(A_i) for $1 \le i \le n$

First Sets. Example

- grammar

  E → T X
  X → + E | ε
  T → (E) | int Y
  Y → * T | ε
- First sets

  First(( ) = { (}
  First( )) = {) }
  First(int) = { int }
  First(+) = { + }
  First(*) = { * }

  First(E) ⊇ = First(T) = {int, (}
  First(X) = {+, ε }
  First(Y) = {*, ε }

Computing Follow Sets

- Definition:

  Follow(X) = { t | S →* β X t δ }
- Intuition:
  - If X → A B then First(B) ⊆ Follow(A) and Follow(X) ⊆ Follow(B)
  - If B →* ε then Follow(X) ⊆ Follow(A)
  - If S is the start symbol then $ ∈ Follow(S)

Computing Follow Sets (Cont.)

Algorithm sketch:

1. $ ∈ Follow(S)
2. First(β) - {ε} ⊆ Follow(X)
   - For each production A → α X β
3. Follow(A) ⊆ Follow(X)
   - For each production A → α X β where ε ∈ First(β)

Follow Sets. Example

- Recall the grammar

  E → T X　　　X → + E | ε
  T → (E) | int Y　　　Y → * T | ε
- Follow sets

```
Follow( + ) = { int, ( }
Follow( ( ) = { int, ( }          Follow( E ) = {), $}
Follow( * ) = { int, ( }          Follow( T ) = {+, ) , $}
Follow( ) ) = {+, ) , $}          Follow( Y ) = {+, ) , $}
Follow( int) = {*, +, ) , $}.     Follow( X ) = {$, ) }
```

Constructing LL(1) Parsing Tables

- Construct a parsing table T for CFG G
- For each production A → α in G do:
  - For each terminal t ∈ First(α) do
    - T[A, t] = α
  - If ε ∈ First(α), for each t ∈ Follow(A) do
    - T[A, t] = α
  - If ε ∈ First(α) and $ ∈ Follow(A) do
    - T[A, $] = α

LL(1) Parsing Table Example

- Left-factored grammar

  E → T X
  X → + E | ε
  T → (E) | int Y
  Y → * T | ε

Rules:
For each production A → α in G do:
For each terminal t ∈ First(α) do
T[A, t] = α
If ε ∈ First(α), for each t ∈ Follow(A) do
T[A, t] = α
If ε ∈ First(α) and $ ∈ Follow(A) do
T[A, $] = α

- The LL(1) parsing table:

Left-most non-terminals	next input tokens: int	*	+	(	)	$
E	TX			TX		
X			+E		ε	ε
T	int Y			(E)		
Y		*T	ε		ε	ε

Notes on LL(1) Parsing Tables

- If any entry is multiply defined then G is not LL(1) [Eg: S->Sa|b]
  - If G is ambiguous
  - If G is left recursive
  - If G is not left-factored
  - other: e.g., LL(2)
- Most programming language CFGs are not LL(1)
  - too weak
  - However they build on these basic ideas

Bottom-Up Parsing

- Bottom-up parsing is more general than (deterministic) top-down parsing
  - just as efficient
  - Builds on ideas in top-down parsing
- Bottom-up parsers don't need left-factored grammars
- Revert to the "natural" grammar for our example:

  E → T + E | T
  T → int * T | int | (E)
- Consider the string: int * int + int

Bottom-Up Parsing

- Revert to the "natural" grammar for our example:

  E → T + E | T
  T → int * T | int | (E)
- Consider the string: int * int + int
- Bottom-up parsing reduces a string to the start symbol by inverting productions:

```
int * int + int        T → int
int * T + int          T → int * T
T + int                T → int
T + T                  E → T
T + E                  E → T + E
E
```

Observation

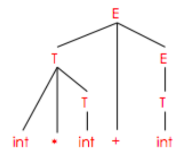
- Read the productions in reverse (from bottom to top)
- This is a rightmost derivation!

```
int * int + int        T → int
int * T + int          T → int * T
T + int                T → int
T + T                  E → T
T + E                  E → T + E
E
```

Bottom-Up Parsing

- A bottom-up parser traces a rightmost derivation in reverse

```
int * int + int          T → int
int * T  + int           T → int * T
T + int                  T → int
T + T                    E → T
T + E                    E → T + E
E
```



A trivial Bottom-Up Parsing Algorithm

```
Let I = input string
  repeat
    pick a non-empty substring β of I
        where X→ β is a production
    if no such β, backtrack
    replace one β by X in I
  until I = "S" (the start symbol) or all possibilities are exhausted
```

Where Do Reductions Happen?

- Right-most derivation has an interesting consequence:
  - Let $\alpha\beta\omega$ be a step of a bottom-up parse
  - Assume the next reduction is by $X \rightarrow \beta$
  - Then ω is a string of terminals
- Why? Because $\alpha X\omega \rightarrow \alpha\beta\omega$ is a step in a rightmost derivation

Notation

- Idea: Split string into two substrings
  - Right substring is as yet unexamined by parsing (a string of terminals)
  - Left substring has terminals and non-terminals
- The dividing point is marked by a |
  - The | is not part of the string
- Initially, all input is unexamined | $x_1x_2 \ldots x_n$

Shift-Reduce Parsing

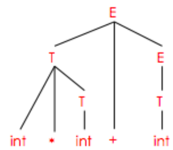
- Bottom-up parsing uses only two kinds of actions:
  - Shift
  - Reduce
- Shift: Move | one place to the right
  - Shifts a terminal to the left string ABC|xyz ⇒ ABCx|yz
- Reduce: Apply an inverse production at the right end of the left string
  - If A → xy is a production, then Cbxy|ijk ⇒ CbA|ijk

The Example with Reductions Only

```
int * int | + int        reduce T → int
int * T | + int          reduce T → int * T
T + int |                reduce T → int
T + T |                  reduce E → T
T + E |                  reduce E → T + E
E |
```

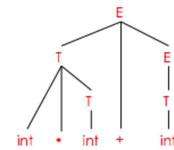
The Example with Shift-Reduce Parsing

|int * int + int	shift	
int | * int + int	shift	
int * | int + int	shift	
int * int | + int	reduce	T → int
int * T | + int	reduce	T → int * T
T | + int	shift	
T + | int	shift	
T + int |	reduce	T → int
T + T |	reduce	E → T
T + E |	reduce	E → T + E
E |		



A shift-reduce parsing

```
|int * int + int
int | * int + int
int * | int + int
int * int | + int
int * T | + int
T | + int
T + | int
T + int |
T + T |
T + E |
E |
```



The Stack

- Left string can be implemented by a stack
  - Top of the stack is the |
- Shift pushes a terminal on the stack
- Reduce pops 0 or more symbols off of the stack (production rhs) and pushes a nonterminal on the stack (production lhs)

Conflicts

- In a given state, more than one action (shift or reduce) may lead to a valid parse
- If it is legal to shift or reduce, there is a shift-reduce conflict
- If it is legal to reduce by two different productions, there is a reduce-reduce conflict.