Lecture 14

CMSC 350: COMPILER DESIGN
Creating an abstract representation of program syntax.

PARSING
Ambiguity

• Consider this grammar:

\[ S \rightarrow S + S \mid (S) \mid \text{number} \]

• Claim: it accepts the \textit{same} set of strings as the previous one.
• What’s the difference?
• Consider these \textit{two} leftmost derivations:

\[
\begin{align*}
S & \rightarrow S + S \rightarrow 1 + S \rightarrow 1 + S + S \rightarrow 1 + 2 + S \rightarrow 1 + 2 + 3 \\
S & \rightarrow S + S \rightarrow S + S + S \rightarrow 1 + S + S \rightarrow 1 + 2 + S \rightarrow 1 + 2 + 3
\end{align*}
\]

• One derivation gives left associativity, the other gives right associativity to ‘+’
  – Which is which?

AST 1

\[
\begin{array}{c}
1 \\
+ \\
3 \\
+ \\
1 \\
2
\end{array}
\]

AST 2

\[
\begin{array}{c}
2 \\
+ \\
3
\end{array}
\]
Why do we care about ambiguity?

• The ‘+’ operation is associative, so it doesn’t matter which tree we pick. Mathematically, \( x + (y + z) = (x + y) + z \)
  – But, some operations aren’t associative. Examples?
  – Some operations are only left (or right) associative. Examples?

• Moreover, if there are multiple operations, ambiguity in the grammar leads to ambiguity in their *precedence*.

• Consider:

\[
S \rightarrow S + S \mid S * S \mid (S) \mid \text{number}
\]

• Input: 1 + 2 * 3
  – One parse = \((1 + 2) * 3 = 9\)
  – The other = \(1 + (2 * 3) = 7\)
Eliminating Ambiguity

- We can often eliminate ambiguity by adding nonterminals and allowing recursion only on the left (or right).
- Higher-precedence operators go farther from the start symbol.
- Example:

  \[
  S \rightarrow S + S \mid S * S \mid (S) \mid \text{number}
  \]

- To disambiguate:
  - Decide (following math) to make ‘*’ higher precedence than ‘+’
  - Make ‘+’ left associative
  - Make ‘*’ right associative
- Note:
  - \( S_2 \) corresponds to ‘atomic’ expressions

  \[
  \begin{align*}
  S_0 & \rightarrow S_0 + S_1 \mid S_1 \\
  S_1 & \rightarrow S_2 * S_1 \mid S_2 \\
  S_2 & \rightarrow \text{number} \mid (S_0)
  \end{align*}
  \]
• Context-free grammars allow concise specifications of programming languages.
  – An unambiguous CFG specifies how to parse: convert a token stream to a (parse tree)
  – Ambiguity can (often) be removed by encoding precedence and associativity in the grammar.

• Even with an unambiguous CFG, there may be more than one derivation
  – Though all derivations correspond to the same abstract syntax tree.

• Still to come: finding a derivation
  – But first: happy
Searching for derivations.

LL & LR PARSING
CFGs Mathematically

• A Context-free Grammar (CFG) consists of
  – A set of *terminals* (e.g., a token or $\varepsilon$)
  – A set of *nonterminals* (e.g., S and other syntactic variables)
  – A designated nonterminal called the *start symbol*
  – A set of productions: LHS $\rightarrow$ RHS
    • LHS is a nonterminal
    • RHS is a *string* of terminals and nonterminals

• Example: The balanced parentheses language:

  $S \rightarrow (S)S$

  $S \rightarrow \varepsilon$

• How many terminals? How many nonterminals? Productions?
Consider finding left-most derivations

- Look at only one input symbol at a time.

<table>
<thead>
<tr>
<th>Partly-derived String</th>
<th>Look-ahead</th>
<th>Parsed/Unparsed Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow E + S$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (S) + S$</td>
<td>1</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (E + S) + S$</td>
<td>1</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + S) + S$</td>
<td>2</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + E + S) + S$</td>
<td>2</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + S) + S$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + E) + S$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + (S)) + S$</td>
<td>3</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + (E + S)) + S$</td>
<td>3</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow ...$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$S \rightarrow E + S \mid E$
$E \rightarrow \text{number} \mid (S)$
There is a problem

- We want to decide which production to apply based on the look-ahead symbol.
- But, there is a choice:

  (1)  \[ S \rightarrow E \rightarrow (S) \rightarrow (E) \rightarrow (1) \]

  vs.

  (1) + 2  \[ S \rightarrow E + S \rightarrow (S) + S \rightarrow (E) + S \rightarrow (1) + S \rightarrow (1) + E \rightarrow (1) + 2 \]

- Given the look-ahead symbol: ‘(‘ it isn’t clear whether to pick \[ S \rightarrow E \] or \[ S \rightarrow E + S \] first.
LL(1) GRAMMARS
Grammar is the problem

- Not all grammars can be parsed “top-down” with only a single lookahead symbol.
- **Top-down**: starting from the start symbol (root of the parse tree) and going down

- LL(1) means
  - Left-to-right scanning
  - Left-most derivation,
  - 1 lookahead symbol

- This language isn’t “LL(1)”
- Is it LL(k) for some k?

- What can we do?
Making a grammar LL(1)

- **Problem:** We can’t decide which S production to apply until we see the symbol after the first expression.
- **Solution:** “Left-factor” the grammar. There is a common E prefix for each choice, so add a new non-terminal $S'$ at the decision point:

\[
S \rightarrow E + S \mid E \\
E \rightarrow \text{number} \mid (S) \\
S' \rightarrow \epsilon \\
S' \rightarrow + S \\
E \rightarrow \text{number} \mid (S)
\]

- Also need to eliminate left-recursion somehow. Why?
- Consider:

\[
S \rightarrow S + E \mid E \\
E \rightarrow \text{number} \mid (S)
\]
### LL(1) Parse of the input string

- Look at only one input symbol at a time.

<table>
<thead>
<tr>
<th>Partly-derived String</th>
<th>Look-ahead</th>
<th>Parsed/Unparsed Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow E S'$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (S) S'$</td>
<td>1</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (E S') S'$</td>
<td>1</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 S') S'$</td>
<td>+</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + S) S'$</td>
<td>2</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + E S') S'$</td>
<td>2</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 S') S'$</td>
<td>+</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + S) S'$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + E S') S'$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$\rightarrow (1 + 2 + (S)S') S'$</td>
<td>3</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
</tbody>
</table>
Predictive Parsing

- Given an LL(1) grammar:
  - For a given nonterminal, the lookahead symbol uniquely determines the production to apply.
  - Top-down parsing = predictive parsing
  - Driven by a predictive parsing table:
    nonterminal * input token → production

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>+</th>
<th>(</th>
<th>)</th>
<th>$ (EOF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>→ S$</td>
<td></td>
<td>→ S$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>→ E S’</td>
<td>→ E S’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S’</td>
<td>→ + S</td>
<td>→ ε</td>
<td>→ ε</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>→ num.</td>
<td>→ ( S )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Note: it is convenient to add a special end-of-file token $ and a start symbol T (top-level) that requires $.
How do we construct the parse table?

• Consider a given production: \(A \rightarrow \gamma\)
• Construct the set of all input tokens that may appear \textit{first} in strings that can be derived from \(\gamma\)
  – Add the production \(\rightarrow \gamma\) to the entry \((A, \text{token})\) for each such token.
• If \(\gamma\) can derive \(\varepsilon\) (the empty string), then we construct the set of all input tokens that may \textit{follow} the nonterminal \(A\) in the grammar.
  – Add the production \(\rightarrow \gamma\) to the entry \((A, \text{token})\) for each such token.

• Note: if there are two different productions for a given entry, the grammar is not LL(1)
Example

- First(T) = First(S)
- First(S) = First(E)
- First(S') = { + }
- First(E) = { number, ‘(‘ }

- Follow(S') = Follow(S)
- Follow(S) = { $, ‘)’ } ∪ Follow(S')

Note: we want the least solution to this system of set equations... a fixpoint computation.

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>+</th>
<th>(</th>
<th></th>
<th>$ (EOF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>$S$</td>
<td></td>
<td></td>
<td>$S$</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>E S'</td>
<td></td>
<td></td>
<td>E S'</td>
<td></td>
</tr>
<tr>
<td>S'</td>
<td>+ S</td>
<td></td>
<td></td>
<td>$\epsilon$</td>
<td>$\epsilon$</td>
</tr>
<tr>
<td>E</td>
<td>num.</td>
<td></td>
<td>(</td>
<td>S</td>
<td></td>
</tr>
</tbody>
</table>
Converting the table to code

- Define \( n \) mutually recursive function, one for each nonterminal \( A \): \texttt{parse} \_\_A
- Each function “peeks” at the lookahead token and then follows the production rule in the corresponding entry.
  - Consume terminal tokens from the input stream
  - Call \texttt{parse} \_\_X to create sub-tree for nonterminal \( X \)
  - This function builds the ast tree itself and returns it.
<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>+</th>
<th>(</th>
<th>)</th>
<th>$ \text{ (EOF)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>$\rightarrow S$</td>
<td></td>
<td></td>
<td></td>
<td>$\rightarrow S$</td>
</tr>
<tr>
<td>S</td>
<td>$\rightarrow E\ S'$</td>
<td></td>
<td></td>
<td></td>
<td>$\rightarrow E\ S'$</td>
</tr>
<tr>
<td>S'</td>
<td></td>
<td>$\rightarrow +\ S$</td>
<td></td>
<td>$\rightarrow \varepsilon$</td>
<td>$\rightarrow \varepsilon$</td>
</tr>
<tr>
<td>E</td>
<td>$\rightarrow \text{num.}$</td>
<td></td>
<td></td>
<td>$\rightarrow (\ S\ )$</td>
<td></td>
</tr>
</tbody>
</table>
LL(1) Summary

• Top-down parsing that finds the leftmost derivation.
• Language Grammar ⇒ LL(1) grammar ⇒ prediction table ⇒ recursive-descent parser

• Problems:
  – Grammar must be LL(1)
  – Can extend to LL(k) (it just makes the table bigger)
  – Grammar cannot be left recursive (parser functions will loop!)

• Is there a better way?
LR GRAMMARS
Bottom-up Parsing (LR Parsers)

• LR(k) parser:
  – Left-to-right scanning
  – Rightmost derivation
  – k lookahead symbols

• LR grammars are more expressive than LL
  – Can handle left-recursive (and right recursive) grammars; virtually all programming languages
  – Easier to express programming language syntax (no left factoring)

• Technique: “Shift-Reduce” parsers
  – Work bottom up instead of top down
  – Construct right-most derivation of a program in the grammar
  – Used by many parser generators (e.g. yacc, CUP, ocamlyacc, merlin, etc.)
  – Better error detection/recovery
Top-down vs. Bottom up

- Consider the left-recursive grammar:

  \[
  S \rightarrow S + E \mid E \\
  E \rightarrow \text{number} \mid (S) 
  \]

- \((1 + 2 + (3 + 4)) + 5\)

- What part of the tree must we know after scanning just \((1 + 2)\)

- In top-down, must be able to guess which productions to use...

Top-down

Bottom-up

Note: ‘(‘ has been scanned but not consumed. Processing it is still pending.
### Progress of Bottom-up Parsing

<table>
<thead>
<tr>
<th>Reductions</th>
<th>Scanned</th>
<th>Input Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1 + 2 + (3 + 4)) + 5 \leftarrow$</td>
<td>(</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
<tr>
<td>$(E + 2 + (3 + 4)) + 5 \leftarrow$</td>
<td>(1</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + 2 + (3 + 4)) + 5 \leftarrow$</td>
<td>(1 + 2</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + E + (3 + 4)) + 5 \leftarrow$</td>
<td>(1 + 2</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + (3 + 4)) + 5 \leftarrow$</td>
<td>(1 + 2</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + (E + 4)) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + (S + 4)) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + (S + E)) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + (S)) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S + E) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$(S) + 5 \leftarrow$</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$E + 5 \leftarrow$</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$S + 5 \leftarrow$</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
<td>$1 + 2 + (3 + 4) + 5$</td>
</tr>
<tr>
<td>$S + E \leftarrow$</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
<td>$(1 + 2 + (3 + 4)) + 5$</td>
</tr>
</tbody>
</table>

**Rightmost derivation**

$S \rightarrow S + E \mid E$

$E \rightarrow \text{number} \mid (S)$
Shift/Reduce Parsing

• Parser state:
  – Stack of terminals and nonterminals.
  – Unconsumed input is a string of terminals
  – Current derivation step is stack + input

• Parsing is a sequence of shift and reduce operations:
  • Shift: move look-ahead token to the stack
  • Reduce: Replace symbols $\gamma$ at top of stack with nonterminal $X$ such that $X \rightarrow \gamma$ is a production. (pop $\gamma$, push $X$)

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>(1 + 2 + (3 + 4)) + 5</td>
<td>shift (</td>
</tr>
<tr>
<td>(1</td>
<td>1 + 2 + (3 + 4)) + 5</td>
<td>shift 1</td>
</tr>
<tr>
<td>(E</td>
<td>+ 2 + (3 + 4)) + 5</td>
<td>reduce: E $\rightarrow$ number</td>
</tr>
<tr>
<td>(S</td>
<td>+ 2 + (3 + 4)) + 5</td>
<td>reduce: S $\rightarrow$ E</td>
</tr>
<tr>
<td>(S +</td>
<td>2 + (3 + 4)) + 5</td>
<td>shift +</td>
</tr>
<tr>
<td>(S + 2</td>
<td></td>
<td>shift 2</td>
</tr>
<tr>
<td>(S + 2</td>
<td>+ (3 + 4)) + 5</td>
<td>reduce: E $\rightarrow$ number</td>
</tr>
</tbody>
</table>
Simple LR parsing with no look ahead.

**LR(0) GRAMMARS**
LR Parser States

• Goal: know what set of reductions are legal at any given point.
• Idea: Summarize all possible stack prefixes $\alpha$ as a finite parser state.
  – Parser state is computed by a DFA that reads the stack $\sigma$.
  – Accept states of the DFA correspond to unique reductions that apply.

• Example: LR(0) parsing
  – Left-to-right scanning, Right-most derivation, zero look-ahead tokens
  – Too weak to handle many language grammars (e.g. the “sum” grammar)
  – But, helpful for understanding how the shift-reduce parser works.
Example LR(0) Grammar: Tuples

• Example grammar for non-empty tuples and identifiers:

\[
S \rightarrow ( \ L \ ) \mid \text{id} \\
L \rightarrow S \mid L , S
\]

• Example strings:
  - x
  - (x, y)
  - (((x))))
  - (x, (y, z), w)
  - (x, (y, (z, w)))

Parse tree for:

(x, (y, z), w)
Shift/Reduce Parsing

• Parser state:
  – Stack of terminals and nonterminals.
  – Unconsumed input is a string of terminals
  – Current derivation step is stack + input
• Parsing is a sequence of *shift* and *reduce* operations:
• **Shift**: move look-ahead token to the stack: e.g.

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x, (y, z), w)</td>
<td>shift (</td>
<td></td>
</tr>
<tr>
<td>(x, (y, z), w)</td>
<td>shift x</td>
<td></td>
</tr>
</tbody>
</table>

• **Reduce**: Replace symbols γ at top of stack with nonterminal X such that X ⟷ γ is a production. (pop γ, push X): e.g.

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x, (y, z), w)</td>
<td>reduce S ⟷ id</td>
<td></td>
</tr>
<tr>
<td>(S, (y, z), w)</td>
<td>reduce L ⟷ S</td>
<td></td>
</tr>
</tbody>
</table>
### Example Run

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x, (y, z), w)</td>
<td>shift (</td>
<td></td>
</tr>
<tr>
<td>(x, (y, z), w)</td>
<td>shift x</td>
<td></td>
</tr>
<tr>
<td>(x, (y, z), w)</td>
<td>reduce S (\rightarrow) id</td>
<td></td>
</tr>
<tr>
<td>(S, (y, z), w)</td>
<td>reduce L (\rightarrow) S</td>
<td></td>
</tr>
<tr>
<td>(L, (y, z), w)</td>
<td>shift ,</td>
<td></td>
</tr>
<tr>
<td>(L, (y, z), w)</td>
<td>reduce S (\rightarrow) id</td>
<td></td>
</tr>
<tr>
<td>(L, (S, z), w)</td>
<td>reduce L (\rightarrow) S</td>
<td></td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>shift ,</td>
<td></td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>shift z</td>
<td></td>
</tr>
<tr>
<td>(L, (L, z), w)</td>
<td>reduce S (\rightarrow) id</td>
<td></td>
</tr>
<tr>
<td>(L, (L, S), w)</td>
<td>reduce L (\rightarrow) L, S</td>
<td></td>
</tr>
<tr>
<td>(L, L), w)</td>
<td>shift )</td>
<td></td>
</tr>
<tr>
<td>(L, L), w)</td>
<td>reduce S (\rightarrow) ( L )</td>
<td></td>
</tr>
<tr>
<td>(L, S), w)</td>
<td>reduce L (\rightarrow) L, S</td>
<td></td>
</tr>
<tr>
<td>(L, w)</td>
<td>shift ,</td>
<td></td>
</tr>
<tr>
<td>(L, w)</td>
<td>shift w</td>
<td></td>
</tr>
<tr>
<td>(L, w)</td>
<td>reduce S (\rightarrow) id</td>
<td></td>
</tr>
<tr>
<td>(L, S)</td>
<td>reduce L (\rightarrow) L, S</td>
<td></td>
</tr>
<tr>
<td>(L)</td>
<td>shift )</td>
<td></td>
</tr>
<tr>
<td>(L)</td>
<td>reduce S (\rightarrow) ( L )</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>reduce S</td>
<td></td>
</tr>
</tbody>
</table>

\[
S \rightarrow ( L ) \mid \text{id} \\
L \rightarrow S \mid L, S
\]
Action Selection Problem

• Given a stack $\sigma$ and a look-ahead symbol $b$, should the parser:
  – Shift $b$ onto the stack (new stack is $\sigma b$)
  – Reduce a production $X \rightarrow \gamma$, assuming that $\sigma = \alpha \gamma$ (new stack is $\alpha X$)?

• Sometimes the parser can reduce but shouldn’t
  – For example, $X \rightarrow \varepsilon$ can always be reduced
• Sometimes the stack can be reduced in different ways

• Main idea: decide what to do based on a prefix $\alpha$ of the stack plus the look-ahead symbol.
  – The prefix $\alpha$ is different for different possible reductions since in productions $X \rightarrow \gamma$ and $Y \rightarrow \beta$, $\gamma$ and $\beta$ might have different lengths.

• Main goal: know what set of reductions are legal at any point.
  – How do we keep track?
LR(0) States

• An LR(0) state is a set of items keeping track of progress on possible upcoming reductions.
• An LR(0) item is a production from the language with an extra separator “.” somewhere in the right-hand-side

\[
\begin{align*}
S & \rightarrow ( L ) \mid \text{id} \\
L & \rightarrow S \mid L , S
\end{align*}
\]

• Example items: \( S \rightarrow .( L ) \) or \( S \rightarrow ( . L ) \) or \( L \rightarrow S \).
• Intuition:
  – Stuff before the ‘.’ is already on the stack  
    (beginnings of possible \( \gamma \)’s to be reduced)
  – Stuff after the ‘.’ is what might be seen next
  – The prefixes \( \alpha \) are represented by the state itself
Constructing the DFA: Start state & Closure

• First step: Add a new production
  \[ S' \rightarrow S\$ \] to the grammar

• Start state of the DFA = empty stack, so it contains the item:
  \[ S' \rightarrow .S\$ \]

• Closure of a state:
  – Adds items for all productions whose LHS nonterminal occurs in an item in the state just after the ‘.’
  – The added items have the ‘.’ located at the beginning (no symbols for those items have been added to the stack yet)
  – Note that newly added items may cause yet more items to be added to the state... keep iterating until a fixed point is reached.

  **Example:** \[ \text{CLOSURE}\{S' \rightarrow .S\$\} = \{S' \rightarrow .S\$, S \rightarrow .(L), S \rightarrow .id\} \]

• Resulting “closed state” contains the set of all possible productions that might be reduced next.
Example: Constructing the DFA

• First, we construct a state with the initial item $S' \rightarrow .S$
Example: Constructing the DFA

- Next, we take the closure of that state:
  \[ \text{CLOSURE}\{\{S' \rightarrow .S$\}\} = \{S' \rightarrow .S$, S \rightarrow .( L ), S \rightarrow .id\} \]

- In the set of items, the nonterminal S appears after the ‘.’
- So we add items for each S production in the grammar
Example: Constructing the DFA

- Next we add the transitions:
- First, we see what terminals and nonterminals can appear after the ‘.’ in the source state.
  - Outgoing edges have those label.
- The target state (initially) includes all items from the source state that have the edge-label symbol after the ‘.’, but we advance the ‘.’ (to simulate shifting the item onto the stack)
Example: Constructing the DFA

- Finally, for each new state, we take the closure.
- Note that we have to perform two iterations to compute CLOSURE(\{S \mapsto ( . L )\})
  - First iteration adds \(L \mapsto .S\) and \(L \mapsto .L, S\)
  - Second iteration adds \(S \mapsto .(L)\) and \(S \mapsto .id\)
Done!
Using the DFA

• Run the parser stack through the DFA.
• The resulting state tells us which productions might be reduced next.
  – If not in a reduce state, then shift the next symbol and transition according to DFA.
  – If in a reduce state, $X \rightarrow \gamma$ with stack $\alpha\gamma$, pop $\gamma$ and push $X$.

• Optimization: No need to re-run the DFA from beginning every step
  – Store the state with each symbol on the stack: e.g. $1(3_3L_5)_6$
  – On a reduction $X \rightarrow \gamma$, pop stack to reveal the state too:
    e.g. From stack $1(3_3L_5)_6$ reduce $S \rightarrow (L)$ to reach stack $1_3$
  – Next, push the reduction symbol: e.g. to reach stack $1_3S$
  – Then take just one step in the DFA to find next state: $1_3S_7$
Implementing the Parsing Table

Represent the DFA as a table of shape:
\[
\text{state} \times (\text{terminals} + \text{nonterminals})
\]

- Entries for the “action table” specify two kinds of actions:
  - Shift and goto state \( n \)
  - Reduce using reduction \( X \leftarrow \gamma \)
    - First pop \( \gamma \) off the stack to reveal the state
    - Look up \( X \) in the “goto table” and goto that state

<table>
<thead>
<tr>
<th>State</th>
<th>Terminal Symbols</th>
<th>Nonterminal Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Action table</strong></td>
<td><strong>Goto table</strong></td>
</tr>
</tbody>
</table>
### Example Parse Table

<table>
<thead>
<tr>
<th></th>
<th>(</th>
<th>)</th>
<th>id</th>
<th>,</th>
<th>$</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td></td>
<td>g4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>S→id</td>
<td>S→id</td>
<td>S→id</td>
<td>S→id</td>
<td>S→id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td></td>
<td>g7</td>
<td>g5</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DONE</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>s6</td>
<td></td>
<td>s8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>S → (L)</td>
<td>S → (L)</td>
<td>S → (L)</td>
<td>S → (L)</td>
<td>S → (L)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>L → S</td>
<td>L → S</td>
<td>L → S</td>
<td>L → S</td>
<td>L → S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>s3</td>
<td>s2</td>
<td></td>
<td></td>
<td></td>
<td>g9</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>L → L,S</td>
<td>L → L,S</td>
<td>L → L,S</td>
<td>L → L,S</td>
<td>L → L,S</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sx  = shift and goto state x
gx  = goto state x
## Example

- Parse the token stream: \( (x, (y, z), w) \$

<table>
<thead>
<tr>
<th>Stack</th>
<th>Stream</th>
<th>Action (according to table)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varepsilon_1 )</td>
<td>( (x, (y, z), w) $</td>
<td>s3</td>
</tr>
<tr>
<td>( \varepsilon_1(3) )</td>
<td>( x, (y, z), w) $</td>
<td>s2</td>
</tr>
<tr>
<td>( \varepsilon_1(3x_2) )</td>
<td>( , (y, z), w) $</td>
<td>Reduce: ( S \rightarrow \text{id} )</td>
</tr>
<tr>
<td>( \varepsilon_1(3S) )</td>
<td>( , (y, z), w) $</td>
<td>g7 (from state 3 follow ( S ))</td>
</tr>
<tr>
<td>( \varepsilon_1(3S_7) )</td>
<td>( , (y, z), w) $</td>
<td>Reduce: ( L \rightarrow S )</td>
</tr>
<tr>
<td>( \varepsilon_1(3L) )</td>
<td>( , (y, z), w) $</td>
<td>g5 (from state 3 follow ( L ))</td>
</tr>
<tr>
<td>( \varepsilon_1(3L_5) )</td>
<td>( , (y, z), w) $</td>
<td>s8</td>
</tr>
<tr>
<td>( \varepsilon_1(3L_{5,8}) )</td>
<td>( (y, z), w) $</td>
<td>s3</td>
</tr>
<tr>
<td>( \varepsilon_1(3L_{5,8}(3) )</td>
<td>( y, z), w) $</td>
<td>s2</td>
</tr>
</tbody>
</table>
LR(0) Limitations

- An LR(0) machine only works if states with reduce actions have a single reduce action.
  - In such states, the machine always reduces (ignoring lookahead)
- With more complex grammars, the DFA construction will yield states with shift/reduce and reduce/reduce conflicts:

  OK shift/reduce reduce/reduce

  \[
  S \rightarrow ( L ). \\
  S \rightarrow ( L ). \\
  L \rightarrow . L , S \\
  S \rightarrow L , S . \\
  S \rightarrow , S .
  \]

- Such conflicts can often be resolved by using a look-ahead symbol: LR(1)
Examples

- Consider the left associative and right associative “sum” grammars:

  left
  
  \[
  S \rightarrow S + E \mid E \\
  E \rightarrow \text{number} \mid (S)
  \]

  right
  
  \[
  S \rightarrow E + S \mid E \\
  E \rightarrow \text{number} \mid (S)
  \]

- One is LR(0) the other isn’t… which is which and why?
- What kind of conflict do you get? Shift/reduce or Reduce/reduce?
- Ambiguities in associativity/precedence usually lead to shift/reduce conflicts.
LR(1) Parsing

- Algorithm is similar to LR(0) DFA construction:
  - LR(1) state = set of LR(1) items
  - An LR(1) item is an LR(0) item + a set of look-ahead symbols:
    \[ A \rightarrow \alpha.\beta, \ L \]

- LR(1) closure is a little more complex:
- Form the set of items just as for LR(0) algorithm.
- Whenever a new item \( C \rightarrow .\gamma \) is added because \( A \rightarrow \beta.C\delta, \ L \) is already in the set, we need to compute its look-ahead set \( M \):
  1. The look-ahead set \( M \) includes FIRST(\( \delta \))
     (the set of terminals that may start strings derived from \( \delta \))
  2. If \( \delta \) can derive \( \varepsilon \) (it is nullable), then the look-ahead \( M \) also contains \( L \)
Example Closure

\[
\begin{align*}
S' & \rightarrow S$ \\
S & \rightarrow E + S \mid E \\
E & \rightarrow \text{number} \mid (S)
\end{align*}
\]

• Start item: \( S' \rightarrow .S$ , \{\}\n• Since \( S \) is to the right of a ‘.’, add:
  \[
  \begin{align*}
  S & \rightarrow .E + S \quad \{\$\} \quad \text{Note:}\ \{\$\} \text{ is } \text{FIRST}($) \\
  S & \rightarrow .E \quad \{\$\}
  \end{align*}
  \]
• Need to keep closing, since \( E \) appears to the right of a ‘.’ in ‘.E + S’:
  \[
  \begin{align*}
  E & \rightarrow .\text{number} \quad \{+\} \quad \text{Note:}\ + \text{ added for reason 1} \\
  E & \rightarrow .(S) \quad \{+\}
  \end{align*}
  \]
• Because \( E \) also appears to the right of ‘.’ in ‘.E’ we get:
  \[
  \begin{align*}
  E & \rightarrow .\text{number} \quad \{$\} \quad \text{Note:}\ \$ \text{ added for reason 2} \\
  E & \rightarrow .(S) \quad \{$\}
  \end{align*}
  \]
• All items are distinct, so we’re done
Using the DFA

The behavior is determined if:
- There is no overlap among the look-ahead sets for each reduce item, and
- None of the look-ahead symbols appear to the right of a ‘.’

Choice between shift and reduce is resolved.

Fragment of the Action & Goto tables
LR variants

- LR(1) gives maximal power out of a 1 look-ahead symbol parsing table
  - DFA + stack is a push-down automaton
- In practice, LR(1) tables are big.
  - Modern implementations (e.g. menhir) directly generate code

- LALR(1) = “Look-ahead LR”
  - Merge any two LR(1) states whose items are identical except for the look-ahead sets:

```
S' \rightarrow .S$ {}
S \rightarrow .E + S ${}$
S \rightarrow .E ${}$
E \rightarrow .num {+}
E \rightarrow .( S ) {+}
E \rightarrow .num ${}$
E \rightarrow .( S ) ${}$
```

- Such merging can lead to nondeterminism (e.g. reduce/reduce conflicts), but
- Results in a much smaller parse table and works well in practice
- This is the usual technology for automatic parser generators: yacc, ocamlyacc, happy

- GLR = “Generalized LR” parsing
  - Efficiently compute the set of all parses for a given input
  - Later passes should disambiguate based on other context
Classification of Grammars

LR(1)
LALR(1)
LL(1)
SLR
LR(0)