Check-in

Review: Predictive Parsing

Assume an LL(1) parser with… this selector table:

	(	)	{	}
S	(S)	)	{	}

this syntax stack:

S
)
)
eof

and this **(** lookahead token:

(

Draw the configuration of the parser after it processes the tokens ( )

Housekeeping
Administrivia

Projects

- P1 grades are in
- P2 (nominally) due Wednesday
- P3 out Friday

Trials

- Trial 1 due tonight

Housekeeping

Administrivia

Labs

- Based on the confusion about abstract classes, I’ve decided to shift the labs a bit

University of Kansas | Drew Davidson

EECS 665 COMPILER CONSTRUCTION

FIRST Sets

Last Time

Review – Predictive Parsing

Intro to Parsing

- Complexity

A New Type of Language – LL(k)

- Intro
- LL(1) parsing

<u>You Should Know</u>

- What parsing is
- What LL(1) languages are
- How an LL(1) parser operates

Parsing

Where we Left Off

Review – Predictive Parsing

The language might be LL(1) … even when the grammar is not!

Same language
Different grammars

Grammar 1

❶ *S* ::= **a b**

❷ | **a c**

Grammar 2

❶ *S* ::= **a** *X*

❷ *X* ::= **b**

❸ *X* ::= **c**

Predicted Parse Tree

S

Candidates

❶ *or* ❷

Input String

a

look

Predicted Parse Tree

S

Candidates

❶

Lookahead

a

look

Today’s Outline

Preview – FIRST Sets

Transforming Grammars

- Fixing LL(1) “near misses”

Building LL(1) Parsers

- What the selector table needs
- FIRST Sets

Parsing

LL(1) Grammar Limitations

Transforming Grammars – Fixing LL(1) Near Misses

Given a language, we can't always find an LL(1) grammar *even if one exists*

- Best we can do: simple transformations that remove "obvious" disqualifiers

Checking if a Grammar is LL(1)

Transforming Grammars – Fixing LL(1) Near Misses

If either of the following hold, the grammar is <u>not</u> LL(1):

- The grammar **is** left-recursive
- The grammar **isn't** left-factored

We can transform *some* grammars while preserving the recognized language

(Immediate) Left Recursion

Transforming Grammars – Fixing LL(1) Near Misses

- Recall, a grammar such that $X \overset{+}{\Rightarrow} X\ \alpha$ is left recursive
- A grammar is immediately left recursive if this can happen in one step:

$$A \rightarrow A\ \alpha \mid \beta$$

Immediate Left Recursion Removal

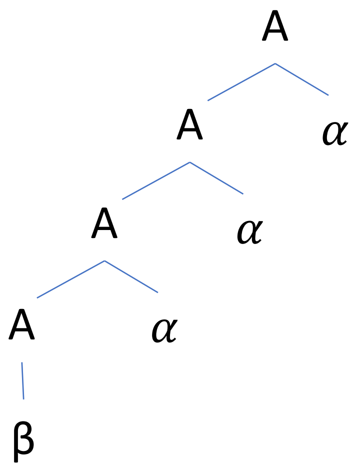
(Predictive) Parsing - LL(1) Transformations

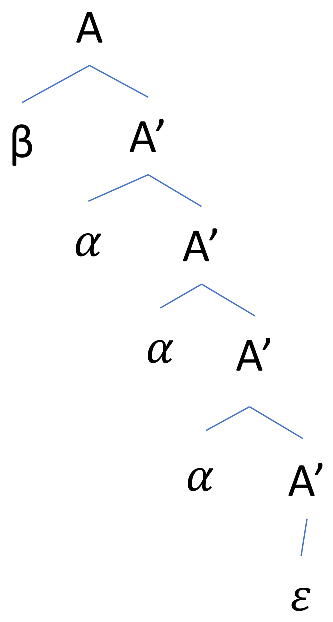
(for a single immediately left-recursive rule)

$$A \rightarrow A\ \alpha \mid \beta$$

Arbitrary Strings
(nonterminal or terminal)

$$A \rightarrow \beta\ A'$$
$$A' \rightarrow \alpha\ A' \mid \varepsilon$$





Immediate Left Recursion Removal

(Predictive) Parsing - LL(1) Transformations

$$A \to \overbrace{\underbrace{A}_{Exp}}^{Exp}\ \overbrace{\alpha}^{-\ Factor} \mid \underbrace{\beta}_{Factor}$$

$$\Rightarrow$$

$$A \to \beta\, A'$$
$$A' \to \alpha\, A' \mid \varepsilon$$

$$\overbrace{\boxed{Exp}}^{A}\ ::=\ \overbrace{\boxed{Exp}}^{A}\ \overbrace{\boxed{-\ Factor}}^{\alpha}$$
$$\mid\ \boxed{Factor} \sim \beta$$

Factor ::= **intlit** | **(** Exp **)**

$$\Rightarrow$$

Immediate Left Recursion Removal

(Predictive) Parsing - LL(1) Transformations

Exp - Factor

$$A \rightarrow \overbrace{A}^{Exp}\ \overbrace{\alpha}^{-\ Factor} \mid \underbrace{\beta}_{Factor}$$

Exp

Factor

Exp Exp'

$$A \rightarrow \beta\ A'$$

Exp' Exp'

$$A' \rightarrow \alpha\ A'$$

$$\mid\ \varepsilon$$

- Factor

A A α

$$\boxed{Exp} ::= \boxed{Exp}\ \boxed{-\ Factor}$$

$$\mid\ \boxed{Factor} \sim \beta$$

Factor ::= **intlit** | **(** Exp **)**

β A' A'

Exp ::= *Factor Exp'*

Exp' ::= *- Factor Exp'*

| ε

α

Factor ::= **intlit** | **(** Exp **)**

Immediate Left Recursion Removal

(Predictive) Parsing - LL(1) Transformations

(general rule)

Given Productions

$$
\begin{aligned}
A ::= &\ \beta_1 \\
| &\ \beta_2 \\
| &\ \beta_n \\
| &\ A\ \alpha_1 \\
| &\ A\ \alpha_2 \\
| &\ A\ \alpha_m
\end{aligned}
$$

Convert to

$$
\begin{aligned}
A ::= &\ \beta_1\ A' \\
| &\ \beta_2\ A' \\
| &\ \beta_n\ A' \\
A' ::= &\ \alpha_1\ A' \\
| &\ \alpha_2\ A' \\
| &\ \alpha_m\ A' \\
| &\ \varepsilon
\end{aligned}
$$

Left Factoring Grammar

(Predictive) Parsing - LL(1) Transformations

- If a nonterminal has (at least) two productions whose RHS has a common prefix, the grammar is **not** left factored

(and **not** an LL(1) grammar)

Question: What makes this grammar not left-factored?

Exp ::= **(** *Exp* **)**
| **{** *Exp* **}**
| **( )**
| **a b**
| **b b**

Left Factoring: Simple Rule

(Predictive) Parsing - LL(1) Transformations

Given Productions

$$A \to \alpha \beta_1 \mid \alpha \beta_2$$

Pull suffix into a new nonterminal

Convert to

$$A \to \alpha A'$$
$$A' \to \beta_1 \mid \beta_2$$

Add a new rule for suffixes

α β_1

X ::= $\boxed{\textbf{a b}}\,\boxed{\textbf{c d}}$

X ::= $\boxed{\textbf{a b}}\,\boxed{\textbf{e f}}$

β_2

X ::= **a b** *X'*

X' ::= **c d** | **e f**

Attempt LL(1) Conversion

(Predictive) Parsing - LL(1) Transformations

Remove immediate left-recursion

$$
\begin{aligned}
\overset{A}{\boxed{Exp}} \; ::= \;& \overset{\beta_1}{\boxed{(\ Exp\ )}} \\
\mid \;& \overset{A}{\boxed{Exp}}\ \overset{\alpha_1}{\boxed{Exp}} \\
\mid \;& \overset{\beta_2}{\boxed{(\ )}}
\end{aligned}
$$

$$
\begin{aligned}
\overset{A}{\boxed{Exp}} \; ::= \;& \overset{\beta_1}{\boxed{(\ Exp\ )}}\ \overset{A'}{\boxed{Exp'}} \\
\mid \;& \overset{\beta_2}{\boxed{(\ )}}\ \overset{A'}{\boxed{Exp'}} \\
\overset{A'}{\boxed{Exp'}} \; ::= \;& \overset{\alpha_1}{\boxed{Exp}}\ \overset{A'}{\boxed{Exp'}} \\
\mid \;& \overset{\text{new } \varepsilon}{\boxed{\varepsilon}}
\end{aligned}
$$

$$A \rightarrow A\ \alpha \mid \beta$$

becomes

$$
\begin{aligned}
A &\rightarrow \beta\ A' \\
A' &\rightarrow \alpha\ A' \\
&\mid\ \varepsilon
\end{aligned}
$$

Attempt LL(1) Conversion

(Predictive) Parsing - LL(1) Transformations

$$
\begin{aligned}
Exp &::= \mathbf{(}\ Exp\ \mathbf{)} \\
&\mid\ Exp \quad Exp \\
&\mid\ \mathbf{(}\ \mathbf{)}
\end{aligned}
$$

Remove immediate left-recursion ✔

$$
\begin{aligned}
\overset{A}{\boxed{Exp}} &::= \overset{\alpha}{\boxed{\mathbf{(}}}\ \overset{\beta_1}{\boxed{Exp\ \mathbf{)}\ \ Exp'}} \\
&\mid\ \overset{\alpha}{\boxed{\mathbf{(}}}\ \overset{\beta_2}{\boxed{\mathbf{)}\ \ Exp'}} \\
Exp' &::= Exp\ \ Exp' \\
&\mid\ \varepsilon
\end{aligned}
$$

Left-factored

$$
\begin{aligned}
\overset{A}{\boxed{Exp}} &::= \overset{\alpha}{\boxed{\mathbf{(}}}\ \overset{A'}{\boxed{Exp''}} \\
\overset{A'}{\boxed{Exp''}} &::= \overset{\beta_1}{\boxed{Exp\ \mathbf{)}\ Exp'}} \\
&\mid\ \overset{\beta_2}{\boxed{\mathbf{)}\ Exp'}} \\
\mathrm{Exp}' &::= Exp\ Exp' \\
&\mid\ \varepsilon
\end{aligned}
$$

$$A \rightarrow \alpha\ \beta_1 \mid \alpha\ \beta_2$$

becomes

$$A \rightarrow \alpha\ A'$$
$$A' \rightarrow \beta_1 \mid \beta_2$$

Attempt LL(1) Conversion

(Predictive) Parsing - LL(1) Transformations

$$
\begin{aligned}
Exp &::= \textbf{(}\ Exp\ \textbf{)} \\
&\mid\ Exp\ \ Exp \\
&\mid\ \textbf{(}\ \textbf{)}
\end{aligned}
$$

Remove immediate left-recursion

$$
\begin{aligned}
Exp &::= \textbf{(}\ Exp\ \textbf{)}\ \ Exp' \\
&\mid\ \textbf{(}\ \textbf{)}\ \ Exp' \\
Exp' &::= Exp\ \ Exp' \\
&\mid\ \varepsilon
\end{aligned}
$$

Left-factored

$$
\begin{aligned}
Exp &::= \textbf{(}\ \ Exp'' \\
Exp'' &::= Exp\ \textbf{)}\ Exp' \\
&\mid\ \textbf{)}\ Exp' \\
\mathrm{Exp}' &::= Exp\ Exp' \\
&\mid\ \ \varepsilon
\end{aligned}
$$

Current Status

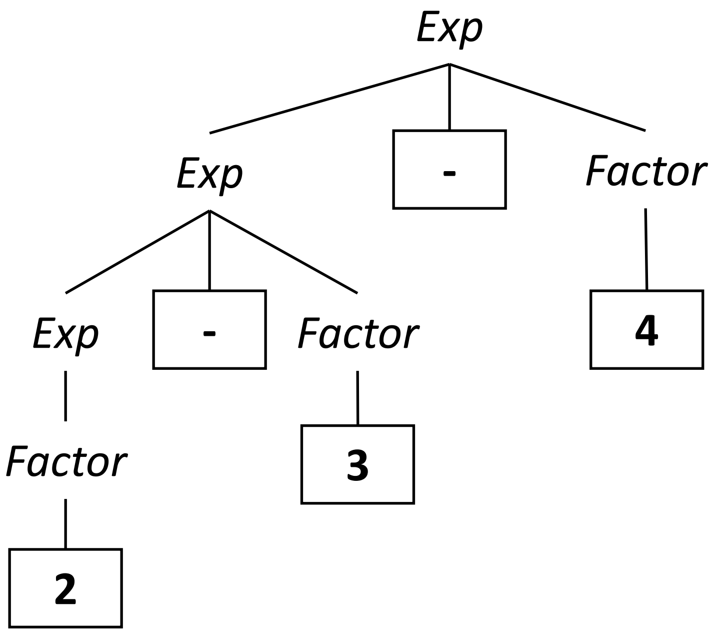
(Predictive) Parsing - LL(1) Transformations

- We’ve removed 2 disqualifiers from LL(1)
  - Left-recursive grammar
  - **Not** Left-Factored grammar

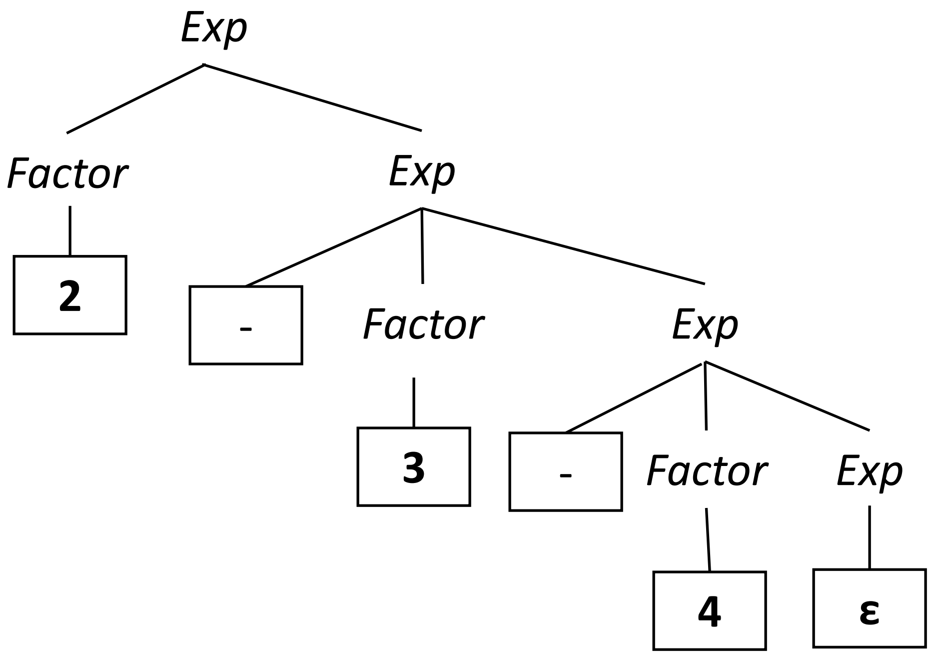
Let’s Check on the Parse Tree

LL(1) Grammar Transformations

$$
\begin{aligned}
\mathit{Exp} &\rightarrow \mathit{Exp}\ \mathbf{-}\ \mathit{Factor} \\
&\mid\ \mathit{Factor} \\
\text{Factor} &\rightarrow \textbf{intlit} \mid \textbf{(}\ \text{Exp}\ \textbf{)}
\end{aligned}
$$

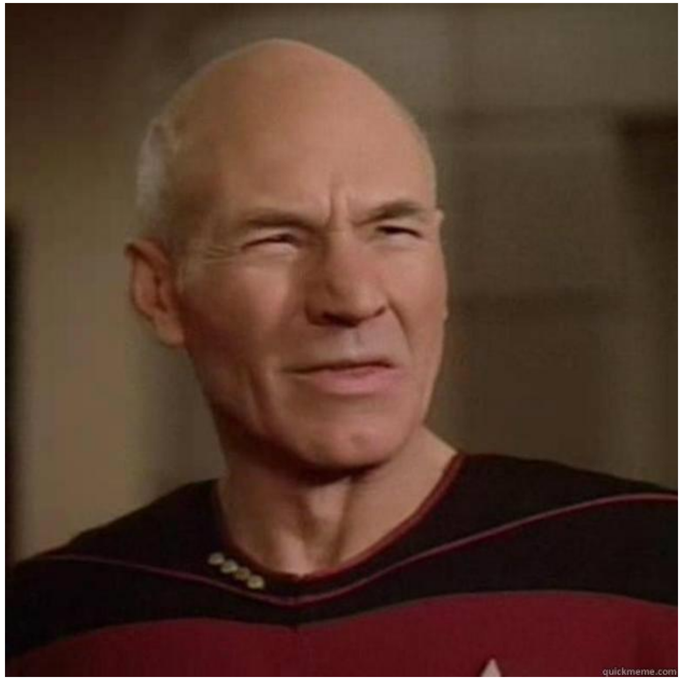


$$
\begin{aligned}
\mathit{Exp} &\rightarrow \mathit{Factor}\ \mathit{Exp'} \\
\mathit{Exp'} &\rightarrow -\ \mathit{Factor}\ \mathit{Exp'} \\
&\mid\ \varepsilon \\
\text{Factor} &\rightarrow \textbf{intlit} \mid \textbf{(}\ \text{Exp}\ \textbf{)}
\end{aligned}
$$

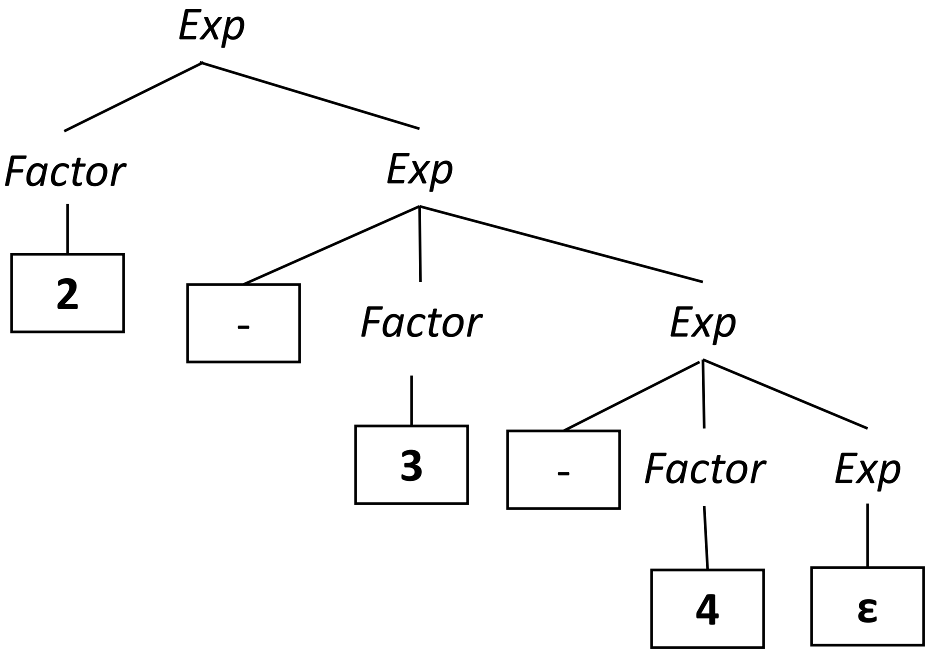


Let's Check on the Parse Tree

LL(1) Grammar Transformations



Exp → *Factor Exp'*
Exp' → - *Factor Exp'*
| ε
Factor → **intlit** | **(** Exp **)**



Nevermind, We’ll Fix Parse Trees Later

LL(1) Grammar Transformations

¯_(ツ)_/¯

Today's Outline

Lecture 9 – FIRST sets

Transforming Grammars

- Fixing LL(1) "near misses"

Building LL(1) Parsers

- Understanding LL(1) Selector Tables
- FIRST Sets

Parsing

Recall the LL(1) Parser's Operation

Building LL(1)Selector Table

LL(1)

- Processes **L**eft-to-right
- **L**eftmost derivation
- **1** token of lookahead

Predictive Parser: "guess & check"

- Starts at the root, *guesses* how to unfold a nonterminal (derivation step)
- *Checks* that terminals match prediction

Recall the LL(1) Parser's Operation

Building LL(1)Selector Table

Example LL(1) Grammar:

S ::= **a** *X*

X ::= **b a** | **c**

Example Input:

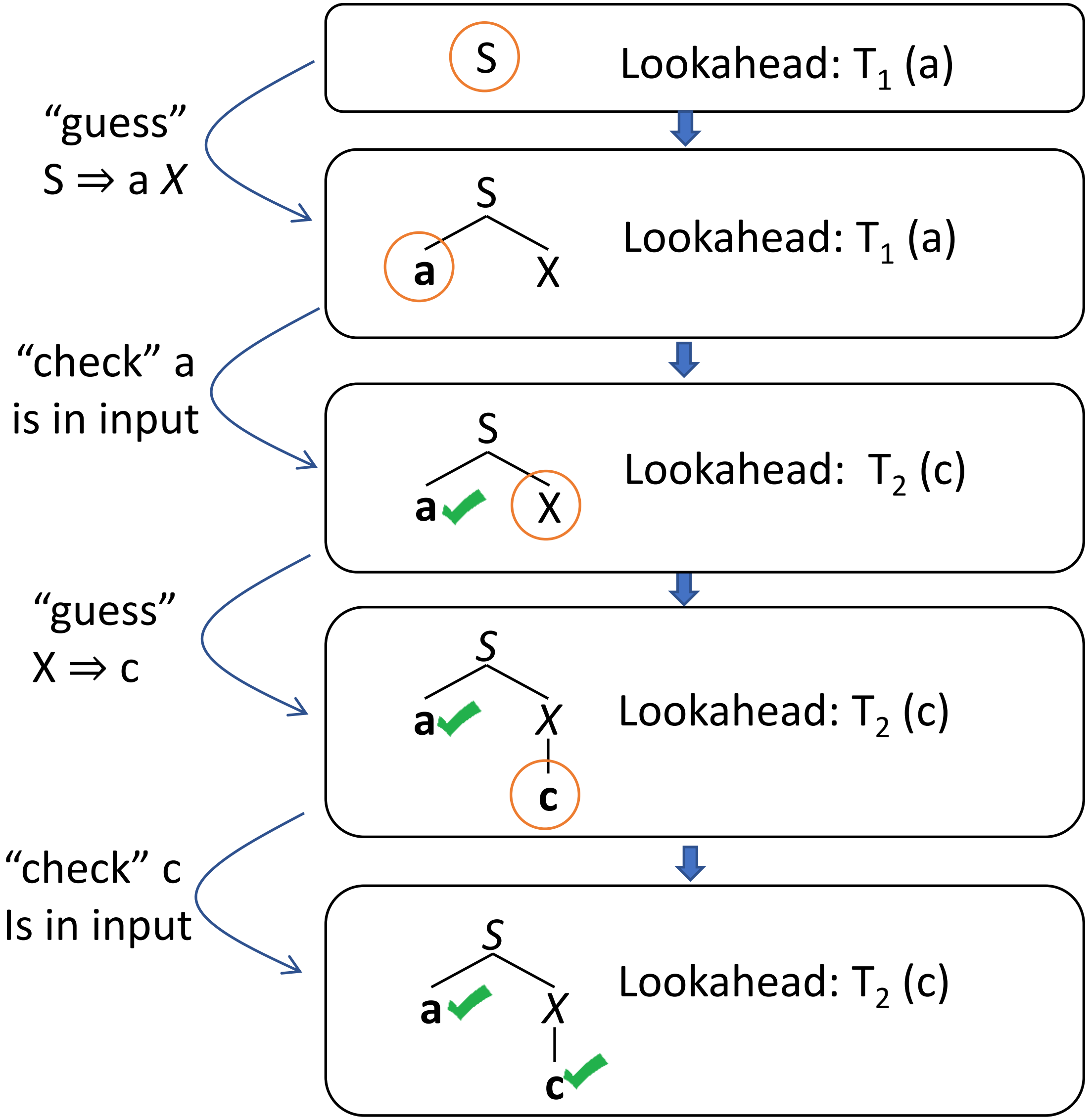
a c

$\uparrow$ $\uparrow$

T_1 T_2

In practice, table-driven parser uses a stack to match this tree

Parser state



How does the Parser Guess?

Building Parser Tables

The intuition is a bit tricky

- We need to get into the mindset of the parser

Pretend your consciousness has been transported inside an LL(1) parser

Become the Parser

Building Parser Tables

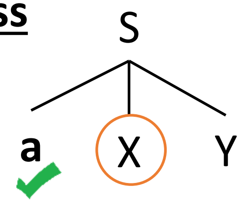
You need to unfold a nonterminal *X* with lookahead token **t**

Assume there's an *X* production *X ::=* $\pi_1\ \pi_2$ (where π_1 and π_2 are some kind of symbol)

How do we know to guess this production?

Case 1: π_1 subtree may start with **t**

Parse in Progress



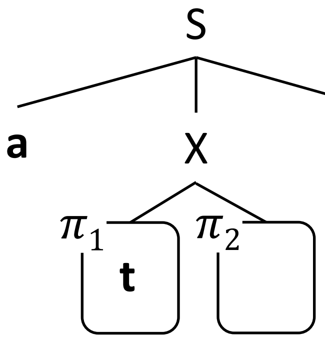
Lookahead: T_2 (**t**)

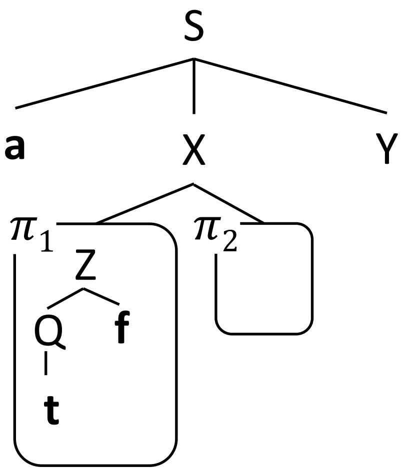
Grammar Fragment

…

X ::= $\pi_1\ \pi_2$

…





Become the Parser

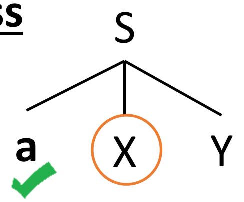
Building Parser Tables

You need to unfold a nonterminal *X* with lookahead token **t**

Assume there's an *X* production *X* ::= $\pi_1\ \pi_2$ (where π_1 and π_2 are some kind of symbol)

How do we know to guess this production?

Parse in Progress



Lookahead: T_2 (**t**)

Grammar Fragment

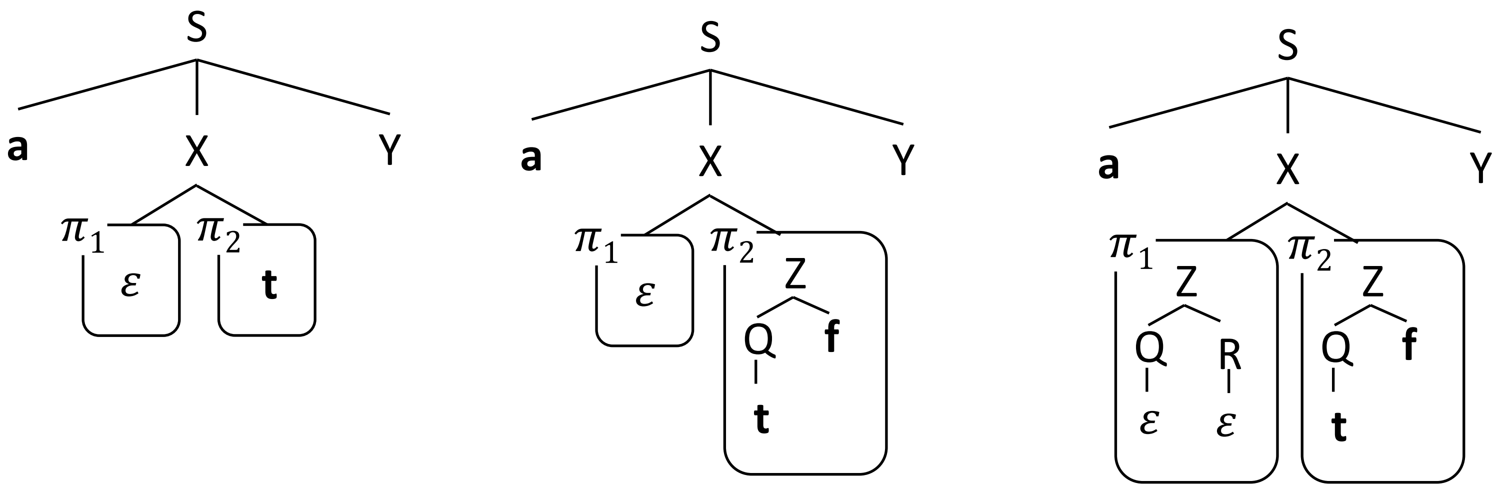
…

X ::= $\pi_1\ \pi_2$

…

Case 1: π_1 subtree may start with **t**

Case 2: π_1 subtree may be empty and π_2 starts with **t**



Become the Parser

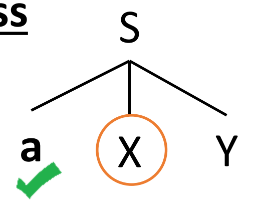
Building Parser Tables

You need to unfold a nonterminal *X* with lookahead token **t**

Assume there's an *X* production *X ::=* $\pi_1\ \pi_2$ (where π_1 and π_2 are some kind of symbol)

How do we know to guess this production?

Parse in Progress



Lookahead: T_2 (**t**)

Grammar Fragment

…

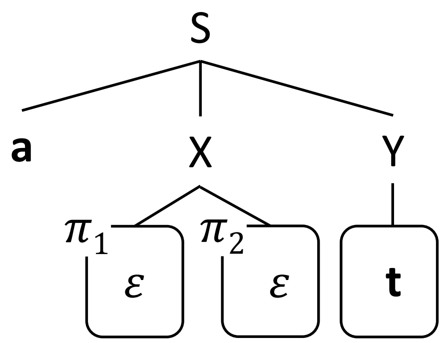
X ::= $\pi_1\ \pi_2$

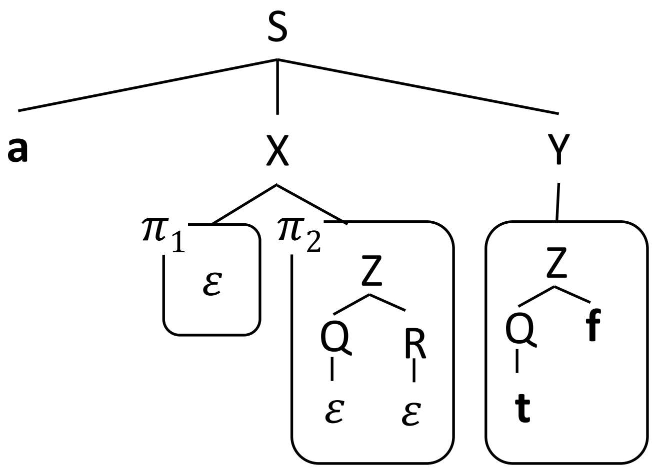
…

Case 1: π_1 subtree may start with **t**

Case 2: π_1 subtree may be empty and π_2 starts with **t**

Case 3: both π_1 and π_2 may be empty and the sibling may start with **t**





Become the Parser

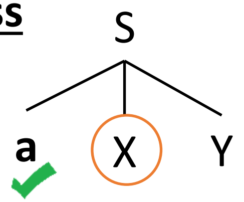
Building Parser Tables

You need to unfold a nonterminal *X* with lookahead token **t**

Assume there's an *X* production $X ::= \pi_1\ \pi_2$
(where π_1 and π_2 are some kind of symbol)

How do we know to guess this production?

Parse in Progress



Lookahead: T_2 (**t**)

Grammar Fragment

…
$X ::= \pi_1\ \pi_2$
…

Case 1: π_1 subtree may start with **t**

Case 2: π_1 subtree may be empty and π_2 starts with **t**

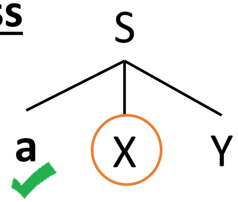
Case 3: both π_1 and π_2 may be empty and the sibling may start with **t**

How can we account for these cases when building the parser?

Become the Parser

Building Parser Tables

Parse in Progress



Lookahead: T_2 (**t**)

You need to unfold a nonterminal *X* with lookahead token **t**

Assume there's an *X* production *X ::=* $\pi_1\ \pi_2$ (where π_1 and π_2 are some kind of symbol)

How do we know to guess this production?

Grammar Fragment

…

X ::= $\pi_1\ \pi_2$

…

Case 1: π_1 subtree may start with **t**

Case 2: π_1 subtree may be empty and π_2 starts with **t**

FIRST Sets

Case 3: both π_1 and π_2 may be empty and the sibling may start with **t**

FOLLOW Sets

Two sets are sufficient to capture these cases and to build the selector table

Today's Outline

Lecture 9 – FIRST sets

Transforming Grammars

✔ Fixing LL(1) "near misses"

Building LL(1) Parsers

✔ Reverse-Engineering Selector Tables

- FIRST Sets

Parsing

An Informal Definition

Building LL(1) Selector Table: FIRST sets, single symbol

FIRST(α) = The set of terminals that begin strings derivable from α, and also, if α can derive ε, then ε is in FIRST(α).

A Formal Definition

Building LL(1) Selector Table: FIRST sets, single symbol

FIRST(α) = The set of terminals that begin strings derivable from α, and also, if α can derive ε, then ε is in FIRST(X).

Formally, FIRST(α) =

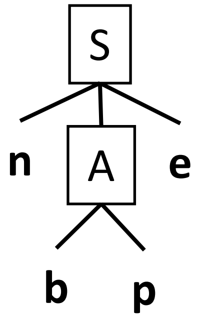
$$\left\{ \hat{\alpha} \,\middle|\, \left(\hat{\alpha} \in \Sigma \wedge \alpha \overset{*}{\Rightarrow} \hat{\alpha}\beta\right) \vee \left(\hat{\alpha} = \varepsilon \wedge \alpha \overset{*}{\Rightarrow} \varepsilon\right) \right\}$$

A Parse Tree Perspective

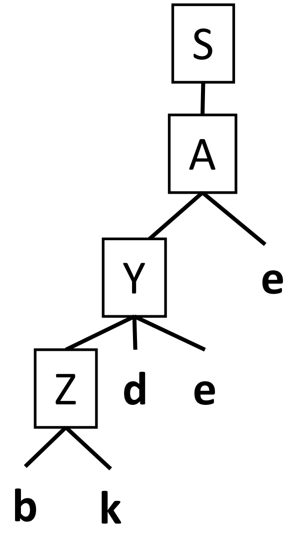
Building LL(1) Selector Table: FIRST sets, single symbol

FIRST(α) = The set of terminals that begin strings derivable from α, and also, if α can derive ε, then ε is in FIRST(X).

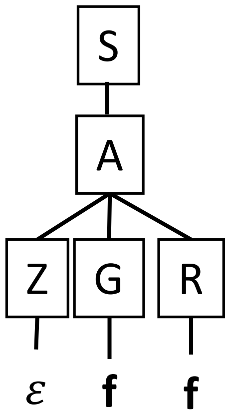
What does the parse tree say about FIRST(*A*)?



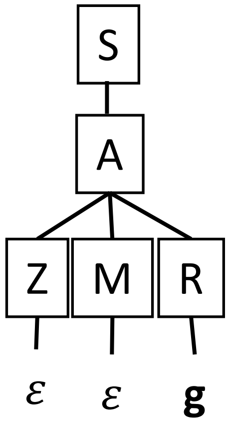
FIRST(A) includes { **b** }



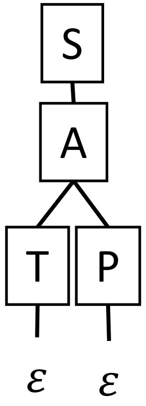
Again, FIRST(A) includes { **b** }



FIRST(A) Includes { **f** }



FIRST(A) includes { **g** }



FIRST(A) includes { ε }

If these were the only possible parse trees, then FIRST(*A*) = { **b**, f, **g,** ε }

A Parse Tree Perspective

Building LL(1) Selector Table: FIRST sets, single symbol

FIRST(α) = The set of terminals that begin strings derivable from α, and also, if α can derive ε, then ε is in FIRST(X).

This isn't how you build FIRST sets

- Looking at parse trees is illustrative for concepts only
- We need to derive FIRST sets directly from the grammar

Building FIRST Sets: Methodology

Building Parser Tables

First sets exist for any arbitrary string of symbols α

- Defined in terms of FIRST sets for a single symbol
  - FIRST of an alphabet terminal
  - FIRST for ε
  - FIRST for a nonterminal
- Use single-symbol FIRST to construct symbol-string FIRSTS

Rules for Single Symbols

Building Parser Tables

FIRST(X) = The set of terminals that begin strings derivable from X, and also, if X can derive ε, then ε is in FIRST(X).

<u>Building FIRST for terminals</u>

FIRST(**t**) = { **t** } for **t** in Σ

FIRST(ε) = { ε }

<u>Building FIRST(*X*) for nonterminal *X*</u>

For each X ::= $\alpha_1\ \alpha_2\ \dots\ \alpha_n$

C_1: add FIRST(α_1) - ε

C_2: If ε could "prefix" FIRST(α_k), add FIRST(α_k)- ε

C_3: If ε is in every FIRST set $\alpha_1 \dots \alpha_n$, add ε

Rules for Single Symbols

Building LL(1) Parsers

Building FIRST(*X*) for nonterminal *X*

For each X ::= $\alpha_1\ \alpha_2 \ldots \alpha_n$

C_1: add FIRST(α_1) - ε

C_2: If ε could "prefix" FIRST(α_k), add FIRST(α_k)- ε

C_3: If ε is in every FIRST set $\alpha_1 \ldots \alpha_n$, add ε

Rules for Single Symbols

Building LL(1) Parsers

<u>Building FIRST(*X*) for nonterminal *X*</u>

For each X ::= $\alpha_1\ \alpha_2 \ldots \alpha_n$

C_1: add FIRST(α_1) - ε

C_2: If ε could "prefix" FIRST(α_k), add FIRST(α_k)- ε

C_3: If ε is in every FIRST set $\alpha_1 \ldots \alpha_n$, add ε

Say there's a production

X ::= *Y Z R T*

and we know

FIRST(*Y*) = { ε, **a** }

FIRST(*Z*) = { ε, **b, m** }

FIRST(*R*) = { **c** }

FIRST(*T*) = { **d** }

By C_2 clause FIRST(*X*) includes **b, m** and **c**

b,**m** because FIRST of every symbol before the 2nd includes ε)

Z in this case

c because FIRST of every symbol before the 3rd includes ε)

R in this case

FIRST(*X*) does not add **d** in this clause because not every FIRST set before the T includes ε

Building FIRST Sets for Symbol Strings

Building LL(1) Parsers

<u>Building FIRST(</u>α<u>)</u>

Let α be composed of symbols $\alpha_1\ \alpha_2 \ldots \alpha_n$

C_1: add FIRST(α_1) - ε

C_2: If $\alpha_1 \ldots \alpha_{k-1}$ is nullable, add FIRST(α_k)- ε

C_3: If $\alpha_1 \ldots \alpha_n$ is nullable, add ε

Base Cases:

α_i is is a terminal **t**. Add **t**

α_i is is a nonterminal *X*. Add every leaf symbol that could begin an *X* subtree

(this gets a bit complicated due to dependencies)

Summary: Explored the LL(1) Mindset

FIRST Sets

LL(1) "Parseability" Qualification

- Knowing the leftmost terminal of a parse (sub)tree is enough to pick the next derivation step

Elusive Conditions

- Two different rules could start with the same terminal (not left factored)
- The same rule(s) could be applied repeatedly (left recursive)

Began choosing matching productions to input

- What terminal could the production be the start of (FIRST)?