Bottom-up Syntax Analysis

Mooly Sagiv
Tel Aviv University
sagiv@math.tau.ac.il
and
Reinhard Wilhelm
Universität des Saarlandes
wilhelm@cs.uni-sb.de

June 2, 1997

Subjects

- Functionality and Method
- Example Parsers
- Derivation of a Parser
- Conflicts
- $LR(k)$–Grammars
- $LR(1)$–Parser Generation
- Bison

Bottom-Up Syntax Analysis

input A stream of symbols (tokens)

output A syntax tree or error

method **until** input consumed or error **do**

- shift next symbol or reduce by some production
- decide what to do by looking one symbol ahead

Properties

- Constructs the syntax tree in a bottom-up manner
- Finds the right most derivation (in reversed order)
- Reports error as soon as the already read part of the input is not a prefix of a program (valid prefix property)

Parsing $aabb$ in the grammar G_{ab}

$$S \rightarrow aSb|\epsilon$$

Stack	Input	Action	Dead ends
$\$$	$aabb\#$	**shift**	**reduce** $S \rightarrow \epsilon$
$\$a$	$abb\#$	**shift**	**reduce** $S \rightarrow \epsilon$
$\$aa$	$bb\#$	**reduce** $S \rightarrow \epsilon$	**shift**
$\$aaS$	$bb\#$	**shift**	**reduce** $S \rightarrow \epsilon$
$\$aaSb$	$b\#$	**reduce** $S \rightarrow aSb$	**shift**, **reduce** $S \rightarrow \epsilon$
$\$aS$	$b\#$	**shift**	**reduce** $S \rightarrow \epsilon$
$\$aSb$	$\#$	**reduce** $S \rightarrow aSb$	**reduce** $S \rightarrow \epsilon$
$\$S$	$\#$	**accept**	**reduce** $S \rightarrow \epsilon$

Issues:

- Shift vs. Reduce
- Reduce $A \rightarrow \beta$, Reduce $B \rightarrow \alpha\beta$

Parsing aa in the grammar
$S \to AB, S \to A, A \to a, B \to a$

Stack	**Input**	**Action**	**Dead ends**
$\$$	$aa\#$	**shift**	
$\$a$	$a\#$	**reduce** $A \to a$	**reduce** $B \to a$, **shift**
$\$A$	$a\#$	**shift**	**reduce** $S \to A$
$\$Aa$	$\#$	**reduce** $B \to a$	**reduce** $A \to a$
$\$AB$	$\#$	**reduce** $S \to AB$	
$\$S$	$\#$	**accept**	

Issues:

- Shift vs. Reduce
- Reduce $A \to \beta$, Reduce $B \to \alpha\beta$

The Handle

- The bottom up–Parser is a shift–reduce–parser, each step is

  a shift: consuming the next input symbol,
  a reduction: reducing a suffix of the stack contents by some production.

- the problem is the localization of the "handle", the next right side to reduce.

  reducing too early: dead end,
  reducing too late: burying the handle.

From P_G to LR–Parsers for G

- P_G has non-deterministic choice of expansions,
- LL–parsers eliminate non–determinism by looking ahead at expansions,
- LR–parsers follow all possibilities in parallel (corresponds to the subset–construction in **NFA** $\rightarrow$ **DFA**).

Derivation

1. Characteristic finite automaton of P_G, a description of P_G
2. Make deterministic
3. Interpret as control of a push down automaton
4. Check for “inedaquate” states

Characteristic Finite Automaton of P_G

NFA $char(P_G) = (Q_c, V_c, \Delta_c, q_c, F_c)$ — the **characteristic finite automaton** of P_G :

- $Q_c = It_G$ — states: the items of G
- $V_c = V_T \cup V_N$ — input alphabet: the sets of terminal and non-terminal symbols
- $q_c = [S' \rightarrow .S]$ — start state
- $F_c = \{[X \rightarrow \alpha.] \mid X \rightarrow \alpha \in P\}$ — final states: the complete items
- $\Delta_c =$
  $\{([X \rightarrow \alpha.Y\beta], Y, [X \rightarrow \alpha Y.\beta]) \mid X \rightarrow \alpha Y\beta \in P$ and $Y \in V_N \cup V_T\} \cup$
  $\{([X \rightarrow \alpha.Y\beta], \varepsilon, [Y \rightarrow .\gamma]) \mid X \rightarrow \alpha Y\beta \in P$ and $Y \rightarrow \gamma \in P\}$

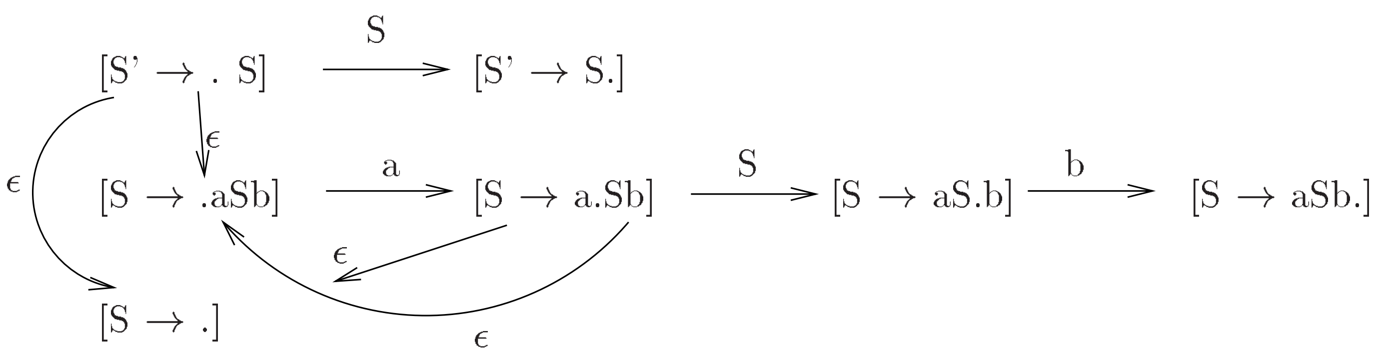
Characteristic NFA for G_{ab}

$$S \rightarrow aSb|\epsilon$$

$P_{G_{ab}}$

Stack	Input	New State
$[S' \rightarrow .S]$	ϵ	$[S \rightarrow .aSb]$
$[S' \rightarrow .S]$	ϵ	$[S \rightarrow .]$
$[S \rightarrow .aSb]$	a	$[S \rightarrow a.Sb]$
$[S \rightarrow a.Sb]$	ϵ	$[S \rightarrow .aSb]$
$[S \rightarrow a.Sb]$	ϵ	$[S \rightarrow .]$
$[S \rightarrow aS.b]$	b	$[S \rightarrow aSb.]$
$[S \rightarrow a.Sb][S \rightarrow .]$	ϵ	$[S \rightarrow aS.b]$
$[S \rightarrow a.Sb][S \rightarrow aSb.]$	ϵ	$[S \rightarrow aS.b]$
$[S' \rightarrow .S][S \rightarrow aSb.]$	ϵ	$[S' \rightarrow S.]$
$[S' \rightarrow .S][S \rightarrow .]$	ϵ	$[S' \rightarrow S.]$

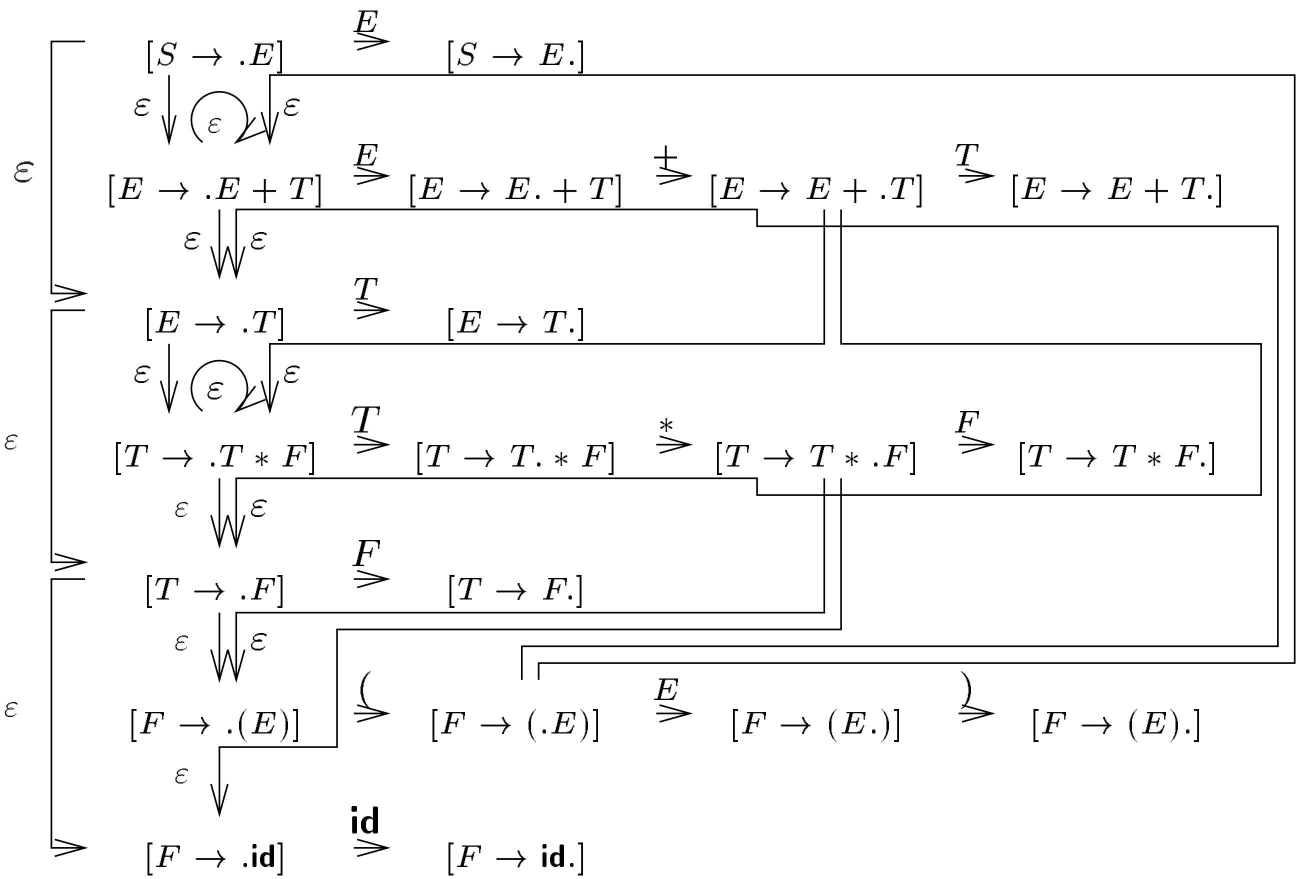
$char(P_{G_{ab}})$



G_0, the grammar for arithmetic expressions

$$
\begin{array}{lcl}
S & \rightarrow & E \\
E & \rightarrow & E + T \mid T \\
T & \rightarrow & T * F \mid F \\
F & \rightarrow & (E) \mid \textbf{id}
\end{array}
$$

Characteristic NFA of G_0



Interpreting $char(P_G)$

State of $char(P_G)$ is the *current* state of P_G, i.i. the state on top of P_G's stack. Adding actions to the transitions and states of $char(P_G)$ to describe P_G:

ε–**transitions:** push new state of $char(P_G)$ onto stack of P_G: new current state.

reading transitions: reading transitions of P_G: replace current state of P_G by the shifted one.

final state: Actions in P_G:

- pop final state $[X \rightarrow \alpha.]$ from the stack,
- do a transition from the new topmost state ander X,
- push the new state onto the stack.

The Handle

- The bottom up–Parser is a shift–reduce–parser, each step is

  a shift: consuming the next input symbol,
  making a transition ander it from the current state,
  pushing the new state onto the stack.
  a reduction: reducing a suffix of the stack contents by some production,
  making a transition ander the left side non–terminal from the new current state,
  pushing the new state.

- the problem is the localization of the "handle", the next right side to reduce.

  reducing too early: dead end,
  reducing too late: burying the handle.

Handles and Viable Prefixes

Some Abbreviations:
RMD – rightmost derivation
RSF – right sentential form

$S' \underset{rm}{\overset{*}{\Longrightarrow}} \beta X u \underset{rm}{\Longrightarrow} \beta\alpha u$ – a RMD of cfg G.

- α is the **handle** of $\beta\alpha u$.
  The part of a RSF next to be reduced.

- Each prefix of $\beta\alpha$ is a **viable prefix**.
  A prefix of a RSF stretching at most up to the handle,
  i.e. reductions if possible then only at the end.

Examples in G_0

	handle	viable prefix	Reason
$E+F$	F	$E,\ E+,\ E+F$	$S \underset{rm}{\Longrightarrow} E \underset{rm}{\Longrightarrow} E+T \underset{rm}{\Longrightarrow} E+F$
$T * \textbf{id}$	**id**	$T,\ T*,\ T * \textbf{id}$	$S \underset{rm}{\overset{3}{\Longrightarrow}} T * F \underset{rm}{\Longrightarrow} T * \textbf{id}$
$F + \textbf{id}$	F	F	$S \underset{rm}{\overset{4}{\Longrightarrow}} T * \textbf{id} \underset{rm}{\Longrightarrow} F * \textbf{id}$

Valid Items

$[X \to \alpha.\beta]$ is **valid** for the viable prefix $\gamma\alpha$, if there exists a RMD $S' \overset{*}{\underset{rm}{\Longrightarrow}} \gamma X w \underset{rm}{\Longrightarrow} \gamma\alpha\beta w$.

An item valid for a viable prefix gives one interpretation of the parsing situation.

Some viable prefixes of G_0

Viable Prefix	Valid Items	Reason	γ	w	X	α	β
$E+$	$[E \to E + .T]$	$S \underset{rm}{\Longrightarrow} E \underset{rm}{\Longrightarrow} E + T$	ε	ε	E	$E+$	T
	$[T \to .F]$	$S \overset{*}{\underset{rm}{\Longrightarrow}} E + T \underset{rm}{\Longrightarrow} E + F$	$E+$	ε	T	ε	F
	$[F \to .\mathbf{id}]$	$S \overset{*}{\underset{rm}{\Longrightarrow}} E + F \underset{rm}{\Longrightarrow} E + \mathbf{id}$	$E+$	ε	F	ε	$\mathbf{id}$
$(E + ($	$[F \to (.E)]$	$S \overset{*}{\underset{rm}{\Longrightarrow}} (E + F) \underset{rm}{\Longrightarrow} (E + (E))$	$(E+$	$)$	F	$($	$E)$

Valid Items and Parsing Situations

Given some input string $xuvw$.

The RMD

$$S' \underset{rm}{\overset{*}{\Longrightarrow}} \gamma X w \underset{rm}{\Longrightarrow} \gamma\alpha\beta w \underset{rm}{\overset{*}{\Longrightarrow}} \gamma\alpha v w \underset{rm}{\overset{*}{\Longrightarrow}} \gamma u v w \underset{rm}{\overset{*}{\Longrightarrow}} x u v w$$

describes the following sequence of partial derivations:

1. $\gamma \underset{rm}{\overset{*}{\Longrightarrow}} x$

2. $\alpha \underset{rm}{\overset{*}{\Longrightarrow}} u$

3. $\beta \underset{rm}{\overset{*}{\Longrightarrow}} v$

4. $X \underset{rm}{\Longrightarrow} \beta$

5. $S' \underset{rm}{\overset{*}{\Longrightarrow}} \gamma X w$

executed by the bottom up–parser in this order.

The valid item $[X \to \alpha.\beta]$ for the viable prefix $\gamma\alpha$ describes the situation after partial derivation 2.

Theorems

$char(P_G) = (Q_c, V_c, \Delta_c, q_c, F_c)$

Theorem. *For each viable prefix there is at least one valid item.*

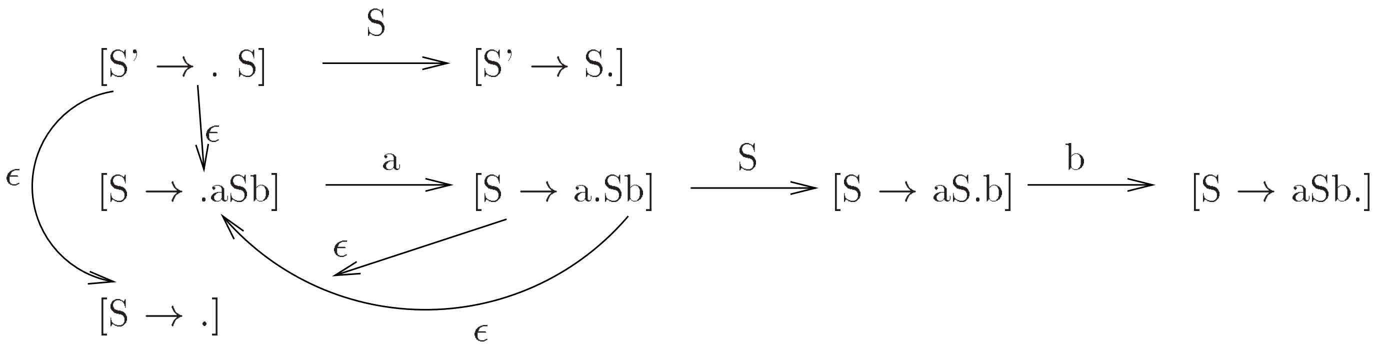
Theorem. *Let* $\gamma \in (V_T \cup V_N)^*$ *and* $q \in Q_c$.
$(q_c, \gamma) \underset{char(P_G)}{\overset{*}{\vdash}} (q, \varepsilon)$ *iff* γ *is a viable prefix and* q *is a valid item for* γ.

Theorem. *The language of viable prefixes of a cfg is regular.*

Making $char(P_G)$ deterministic

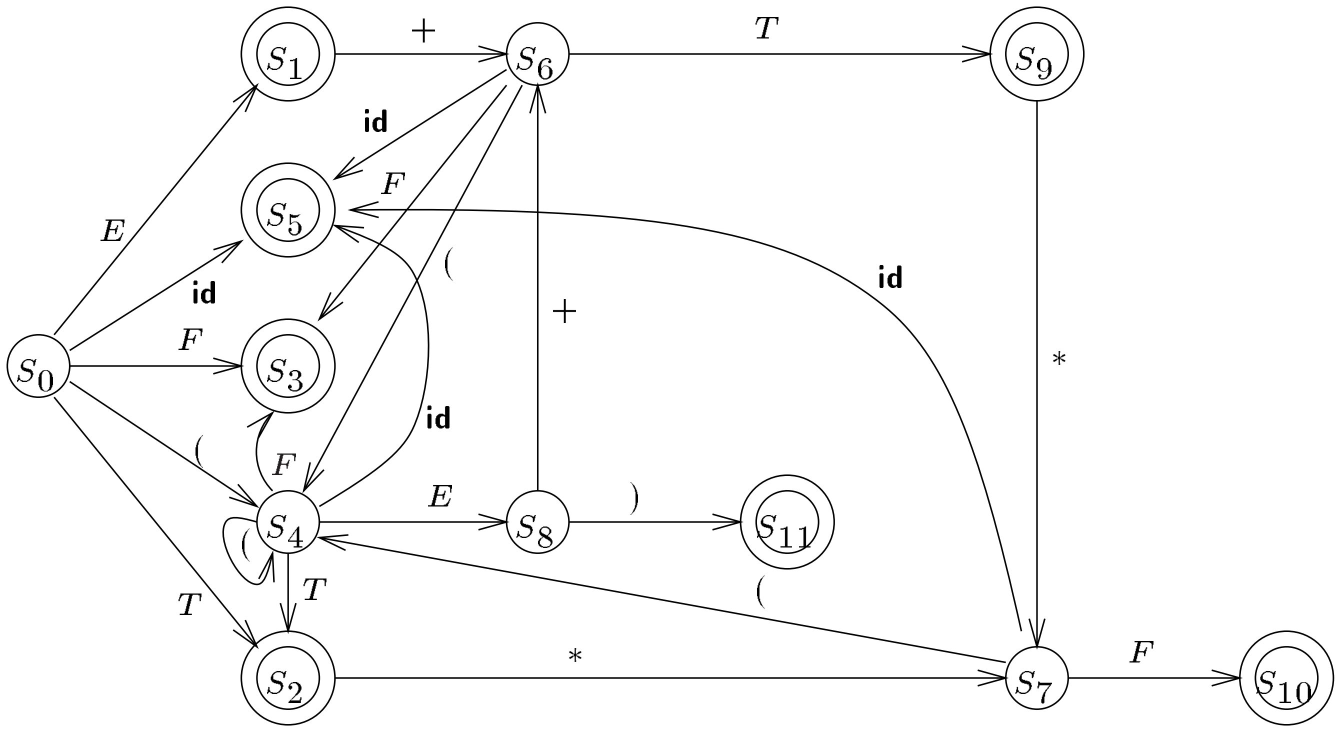
Apply **NFA** $\rightarrow$ **DFA** to $char(P_G)$: Result LR-DFA(G).

Example: $char(P_{G_{ab}})$



LR-DFA(G_{ab}):

LR-DFA(G_0)



The States as Sets of Items

$$
\begin{array}{llll}
S_0 & = \{ & [S & \rightarrow .E], \\
& & [E & \rightarrow .E+T], \\
& & [E & \rightarrow .T], \\
& & [T & \rightarrow .T*F], \\
& & [T & \rightarrow .F], \\
& & [F & \rightarrow .(E)], \\
& & [F & \rightarrow .\textbf{id}]\} \\
S_1 & = \{ & [S & \rightarrow E.], \\
& & [E & \rightarrow E.+T]\} \\
S_2 & = \{ & [E & \rightarrow T.], \\
& & [T & \rightarrow T.*F]\} \\
S_3 & = \{ & [T & \rightarrow F.]\} \\
S_4 & = \{ & [F & \rightarrow (.E)], \\
& & [E & \rightarrow .E+T], \\
& & [E & \rightarrow .T], \\
& & [T & \rightarrow .T*F] \\
& & [T & \rightarrow .F] \\
& & [F & \rightarrow .(E)] \\
& & [F & \rightarrow .\textbf{id}]\}
\end{array}
$$

$$
\begin{array}{llll}
S_5 & = \{ & [F & \rightarrow \textbf{id}.]\} \\
S_6 & = \{ & [E & \rightarrow E+.T], \\
& & [T & \rightarrow .T*F], \\
& & [T & \rightarrow .F], \\
& & [F & \rightarrow .(E)], \\
& & [F & \rightarrow .\textbf{id}]\} \\
S_7 & = \{ & [T & \rightarrow T*.F], \\
& & [F & \rightarrow .(E)], \\
& & [F & \rightarrow .\textbf{id}]\} \\
S_8 & = \{ & [F & \rightarrow (E.)], \\
& & [E & \rightarrow E.+T]\} \\
S_9 & = \{ & [E & \rightarrow E+T.], \\
& & [T & \rightarrow T.*F]\} \\
S_{10} & = \{ & [T & \rightarrow T*F.]\} \\
S_{11} & = \{ & [F & \rightarrow (E).]\}
\end{array}
$$

Theorems

$char(P_G) = (Q_c, V_c, \Delta_c, q_c, F_c)$
$LR-DFA(G) = (Q_d, V_N \cup V_T, \Delta, q_d, F_d)$

Theorem. *Let γ be a viable prefix, and let p(γ) $\in$ Q_d be the uniquely determined state, into which LR-DFA(G) transfers out of the initial state by reading γ, i.e., $(q_d, \gamma) \overset{*}{\vdash}_{LR-DFA(G)} (p(\gamma), \varepsilon)$.*
Then

(a) $p(\varepsilon) = q_d$

(b) $p(\gamma) = \{q \in Q_c \mid (q_c, \gamma) \overset{*}{\vdash}_{char(P_G)} (q, \varepsilon)\}$

(c) $p(\gamma) = \{i \in It_G \mid i \text{ valid for } \gamma\}$

(d) *Let Γ the (in general infinite) set of all viable prefixes of G. The mapping $p : \Gamma \rightarrow Q_d$ defines a finite partition on Γ.*

(e) $L($*LR-DFA*$(G))$ *is the set of viable prefixes of G, which end in a handle.*

Examples from G_0

$\gamma = E + F$ is a viable prefix of G_0.
With the state $p(\gamma) = S_3$ are also associated:
$F,\ (F,\ ((F,\ (((F, \ldots$
$T * (F,\ T * ((F,\ T * (((F, \ldots$
$E + F,\ E + (F,\ E + ((F, \ldots$

Regard S_6 in LR-DFA(G_0).
It consists of all valid items for the viable prefix $E+$,
namely the items $[E \to E + .T], [T \to .T * F]$,
$[T \to .F], [F \to .\mathbf{id}], [F \to .(E)]$.

Reason:
$E+$ is prefix of the RSF $E + T$;

$$
\begin{array}{llllllll}
S \underset{rm}{\Longrightarrow} E \underset{rm}{\Longrightarrow} & E + T & \underset{rm}{\Longrightarrow} & E + F & \underset{rm}{\Longrightarrow} & E + \mathbf{id} \\
& \uparrow & & \uparrow & & \uparrow \\
\text{Therefore} & [E \to E + .T] & & [T \to .F] & & [F \to .\mathbf{id}]
\end{array}
$$

are valid.

What the LR-DFA(G) describes

LR-DFA(G) interpreted as a PDA
$P_0(G) = (\Gamma, V_T, \Delta, q_0, \{q_f\})$

Γ, (stack alphabet): the set Q_d of states of LR-DFA(G).

$q_0 = q_d$ (initial state): in the stack of $P_0(G)$ initially.

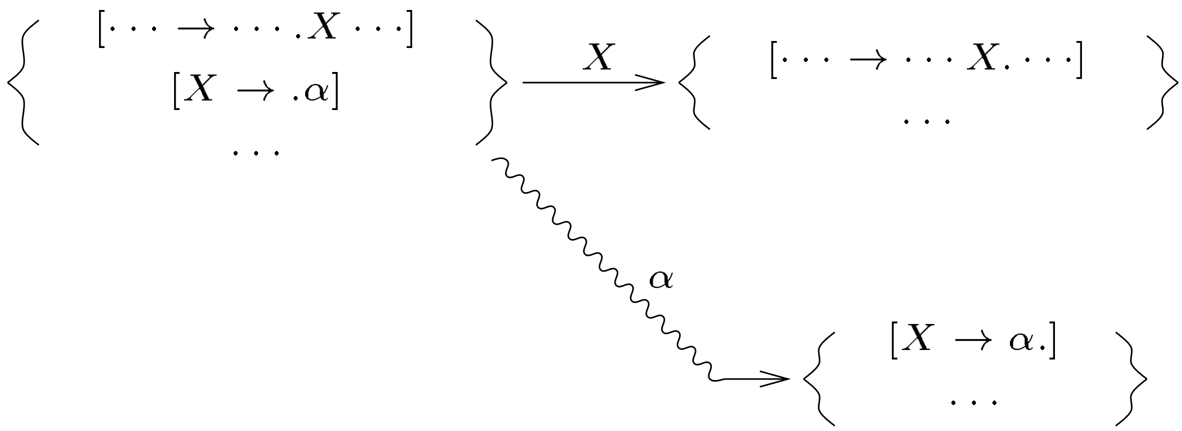
q_f (the one final state): the final state of LR-DFA(G), which contains $[S' \rightarrow S.]$.

$\Delta \subseteq \Gamma^* \times (V_T \cup \{\varepsilon\}) \times \Gamma^*$ (transition relation): Defined as follows:

read: $(q, a, q\delta_d(q,a)) \in \Delta$, if $\delta_d(q,a)$ defined.
Read next input symbol a and push successor state of q ander a (item $[X \rightarrow \cdots .a \cdots] \in q$).

reduce: $(qq_1 \ldots q_n, \varepsilon, q\delta_d(q,X)) \in \Delta$, if $[X \rightarrow \alpha.] \in q_n$, $|\alpha| = n$.
Remove $|\alpha|$ entries from the stack.
Push the successor of the new topmost state ander X onto the stack.

Reduction in PDA $P_0(G)$



Some observations and recollections:

- also works for reductions of ϵ,
- each state has a unique entry symbol,
- the stack contents uniquely determine a viable prefix,
- current state (topmost) is the state associated with this viable prefix,
- current state consists of all items valid for this viable prefix.

Non-determinism in $P_0(G)$

$P_0(G)$ is non-deterministic if either

Shift–reduce conflict: There are read as well as reduce transitions out of one state, or

Reduce–reduce conflict: There are more than one reduce transitions from one state.

States with a shift–reduce conflict have at least one read item $[X \rightarrow \alpha.a\beta]$ and at least one complete item $[Y \rightarrow \gamma.]$.

States with a reduce–reduce conflict have at least two complete items $[Y \rightarrow \alpha.]$, $[Z \rightarrow \beta.]$.

A state with a conflict is **inadequate**.

Some Inadequate States

LR-DFA(G_0) has three inadequate states, S_1, S_2 and S_9.

S_1: Can reduce E to S (complete item $[S \rightarrow E.]$) or read "$+$" (shift–item $[E \rightarrow E. + T]$);

S_2: Can reduce T to E (complete item $[E \rightarrow T.]$) or read "$*$" (shift-item $[T \rightarrow T. * F]$);

S_9: Can reduce E+T to E (complete item $[E \rightarrow E + T.]$) or read "$*$" (shift–item $[T \rightarrow T. * F]$).

Direct Construction of the LR-DFA(G)

Algorithm LR-DFA:
Input: cfg $G = (V'_N, V_T, P', S')$
Output: LR-DFA$(G) = (Q_d, V_N \cup V_T, q_d, \delta_d, F_d)$
Method: The states and the transitions of the LR-DFA(G) are constructed using the following three functions *Start*, *Closure* and *Succ*
F_d – all states containing at least one complete item.

var q, q': **set of item**;
Q_q: **set of set of item**;
δ_d: **set of item** $\times (V_N \cup V_T) \rightarrow$ **set of item**;

function *Start:* **set of item**; **return**$(\{[S' \rightarrow .S]\})$;

function $Closure(s$: **set of item**$)$: **set of item**;
(∗ ε-Succ states of algorithm **NFA → DFA** ∗)
begin $q := s$;
  while *exists* $[X \rightarrow \alpha.Y\beta]$ **in** q **and** $Y \rightarrow \gamma$ **in** P
    and $[Y \rightarrow .\gamma]$ **not in** q **do**
    add $[Y \rightarrow .\gamma]$ to q
  od;
  return(q)
end ;

```
function  Succ(s : set of item, Y : V_N ∪ V_T) : set of item;
      (* read–transitions of P_G *)
   return({[X → αY.β] | [X → α.Yβ] ∈ s});

begin
   Q_d := {Closure(Start)}; (* start state *)
   δ_d := ∅;
   foreach  q in  Q_d and  X in  V_N ∪ V_T do
      let  q' = Closure(Succ(q, X)) in
         if  q' ≠ ∅
         then
            if  q' not in Q_d
            then  Q_d := Q_d ∪ {q'}
            fi;
            δ_d := δ_d ∪ {q --X--> q'}
         fi
      tel
   od
end .
```

LR(k)–Grammars

G – LR(k)–Grammar iff in each RMD
$S' = \alpha_0 \underset{rm}{\Longrightarrow} \alpha_1 \underset{rm}{\Longrightarrow} \alpha_2 \cdots \underset{rm}{\Longrightarrow} \alpha_m = v$
and in each RSF α_i

- the handle can be localized, and
- the production to be applied can be determined

by regarding the prefix of α_i including the handle and at most k symbols behind the handle.

I.e., the splitting of α_i in $\gamma\beta w$ and the production $X \rightarrow \beta$, such that $\alpha_{i-1} = \gamma X w$, is uniquely determined by $\gamma\beta$ and $k : w$.

Definition: A cfg G is an **LR(k)-Grammar**, iff
$S' \overset{*}{\underset{rm}{\Longrightarrow}} \alpha X w \underset{rm}{\Longrightarrow} \alpha\beta w$ and
$S' \overset{*}{\underset{rm}{\Longrightarrow}} \gamma Y x \underset{rm}{\Longrightarrow} \alpha\beta y$ and
$k : w = k : y$ implies that
$\alpha = \gamma$ and $X = Y$ and $x = y$.

Example1

Cfg G_{nLL} with the productions

$$
\begin{array}{lcl}
S & \rightarrow & A \mid B \\
A & \rightarrow & aAb \mid 0 \\
B & \rightarrow & aBbb \mid 1
\end{array}
$$

- $L(G) = \{a^n 0 b^n \mid n \geq 0\} \cup \{a^n 1 b^{2n} \mid n \geq 0\}$.
- G_{nLL} is not LL(k) for arbitrary k.
- G_{nLL} is LR(0)-grammar.
- The RSFs of G_{nLL} (<u>handle</u>)
  - $S, \underline{A}, \underline{B}$,
  - $a^n \underline{aBbb} b^{2n}, a^n \underline{aAb} b^n$,
  - $a^n a \underline{0} b b^n, a^n a \underline{1} bb b^{2n}$.

- Only $a^n aAbb^n$ and $a^n aBbbb^{2n}$ allow 2 different reductions.
  - reduce $a^n aAbb^n$ to $a^n Ab^n$: part of a RMD $S \overset{*}{\underset{rm}{\Longrightarrow}} a^n Ab^n \underset{rm}{\Longrightarrow} a^n aAbb^n$,
  - reduce $a^n aAbb^n$ to $a^n aSbb^n$: not part of any RMD.

- The prefix a^n of $a^n Ab^n$ uniquely determines, whether
  - A is the handle $(n = 0)$, or
  - whether aAb is the handle $(n > 0)$.

- The RSFs $a^n Bb^{2n}$ are treated analogously.

Example2

Cfg G_1 with
$S \rightarrow aAc$
$A \rightarrow Abb \mid b$

- $L(G_1) = \{ab^{2n+1}c \mid n \geq 0\}$
- G_1 is LR(0)–grammar.

  RSF $aAbbb^{2n}c$**:** only legal reduction is to $aAb^{2n}c$,
  uniquely determined by the prefix $aAbb$.
  RSF $abb^{2n}c$**:** b is the handle,
  uniquely determined by the prefix ab.

Example3

Cfg G_2 with

$$\begin{array}{lcl} S & \to & aAc \\ A & \to & bbA \mid b. \end{array}$$

- $L(G_2) = L(G_1)$
- G_2 is LR(1)–grammar.
- Critical RSF $ab^n w$.
  - $1 : w = b$ implies, handle in w;
  - $1 : w = c$ implies, last b in b^n is handle.

Example4

Cfg G_3 with

$$\begin{array}{lcl} S & \rightarrow & aAc \\ A & \rightarrow & bAb \mid b. \end{array}$$

- $L(G_3) = L(G_1)$,
- G_3 is not LR(k)–grammar for arbitrary k.

Chose an arbitrary k.
Regard two RMDs
$S \underset{rm}{\overset{*}{\Longrightarrow}} ab^nAb^nc \underset{rm}{\Longrightarrow} ab^nbb^nc$
$S \underset{rm}{\overset{*}{\Longrightarrow}} ab^{n+1}Ab^{n+1}c \underset{rm}{\Longrightarrow} ab^{n+1}bb^{n+1}c$ where $n \geq k$
Choose $\alpha = ab^n, \beta = b, \gamma = ab^{n+1}, w = b^nc, y = b^{n+2}c$.
It holds $k : w = k : y = b^k$.
$\alpha \neq \gamma$ implies that G_3 is not an LR(k)–grammar.

Adding Lookahead

Lookahead will be used to resolve conflicts.

- $[X \rightarrow \alpha_1.\alpha_2, L]$ – **LR(k)–Item**, if $X \rightarrow \alpha_1\alpha_2 \in P$ and $L \subseteq V_{T\#}^{\leq k}$.

- $[X \rightarrow \alpha_1.\alpha_2]$ – **core** of $[X \rightarrow \alpha_1.\alpha_2, L]$,

- L – the **lookahead set** of $[X \rightarrow \alpha_1.\alpha_2, L]$.

- $[X \rightarrow \alpha_1.\alpha_2, L]$ is **valid** for a viable prefix $\alpha\alpha_1$, if for all $u \in L$ there is a RMD $S'\# \underset{rm}{\overset{*}{\Longrightarrow}} \alpha X w \underset{rm}{\Longrightarrow} \alpha\alpha_1\alpha_2 w$ with $u = k : w$.

The context–free items can be regarded as LR(0)-Items if $[X \rightarrow \alpha_1.\alpha_2, \{\varepsilon\}]$ is identified with $[X \rightarrow \alpha_1.\alpha_2]$.

Example from G_0

(1)	$[E \;\rightarrow\; E + .T, \{), +\}]$	is a valid LR(1)–item for $(E+$
(2)	$[E \;\rightarrow\; T., \{*\}]$	is not a valid LR(1)-item for any viable prefix

Reason:

(1) $S' \underset{rm}{\overset{*}{\Longrightarrow}} (E) \underset{rm}{\Longrightarrow} (E + T) \underset{rm}{\overset{*}{\Longrightarrow}} (E + T + \textbf{id})$ where
$\alpha = (,\; \alpha_1 = E+,\; \alpha_2 = T,\; u = +,\; w = +\textbf{id})$
(2) The string $E*$ can occur in no RMD.

LR–Parser

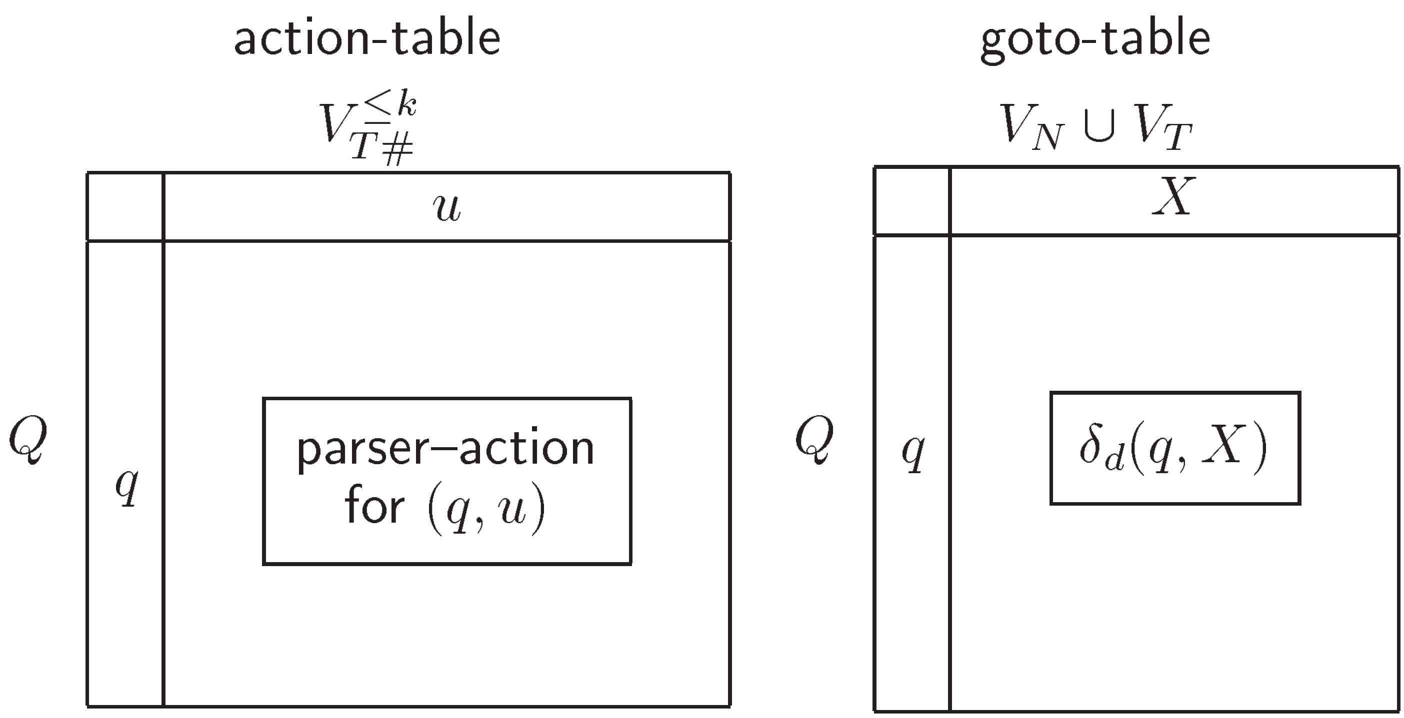
Take their decisions (to shift or to reduce) by consulting

- the viable prefix γ in the stack, actually the by γ uniquely determined state (on top of the stack),
- the next k symbols of the remaining input.
- Recorded in an **action**–table.
- The entries in this table are:

shift:	read next input symbol;
reduce $(X \rightarrow \alpha)$:	reduce by production $X \rightarrow \alpha$;
error:	report error
accept:	report successful termination.

A **goto**–table records the transition function of the LR–DFA(G).

The action– and the goto–table



Parser table for $S \rightarrow aSb|\epsilon$

Action–table

state	sets of items	symbols		
		a	b	$\#$
0	$\left\{\begin{array}{l}[S' \rightarrow .S],\\ [S \rightarrow .aSb],\\ [S \rightarrow .]\}\end{array}\right\}$	s		$r(S \rightarrow \epsilon)$
1	$\left\{\begin{array}{l}[S \rightarrow a.Sb],\\ [S \rightarrow .aSb],\\ [S \rightarrow .]\}\end{array}\right\}$	s	$r(S \rightarrow \epsilon)$	
2	$\{[S \rightarrow aS.b]\}$		s	
3	$\{[S \rightarrow aSb.]\}$		$r(S \rightarrow aSb)$	$r(S \rightarrow aSb)$
4	$\{[S' \rightarrow S.]\}$			$accept$

Goto–table

state	**symbol**			
	a	b	$\#$	S
0	1			4
1	1			2
2		3		
3				
4				

Parsing $aabb$

Stack	**Input**	**Action**
$\$\,0$	$aabb\#$	**shift** 1
$\$\,0\,1$	$abb\#$	**shift** 1
$\$\,0\,1\,1$	$bb\#$	**reduce** $S \to \epsilon$
$\$\,0\,1\,1\,2$	$bb\#$	**shift** 3
$\$\,0\,1\,1\,2\,3$	$b\#$	**reduce** $S \to aSb$
$\$\,0\,1\,2$	$b\#$	**shift** 3
$\$\,0\,1\,2\,3$	$\#$	**reduce** $S \to aSb$
$\$\,0\,4$	$\#$	**accept**

Compressed Representation

- Integrate the terminal columns of the goto–table into the action–table.
- Combine **shift** entry for q and a with $\delta_d(q, a)$.
- Interpret action$[q, a] =$ **shift** p as read a and push p.

Compressed Parser table for $S \rightarrow aSb|\epsilon$

st.	sets of items	symbols			goto
		a	b	$\#$	S
0	$\left\{ \begin{array}{l} [S' \rightarrow .S], \\ [S \rightarrow .aSb], \\ [S \rightarrow .]\} \end{array} \right\}$	$s1$		$rS \rightarrow \epsilon$	4
1	$\left\{ \begin{array}{l} [S \rightarrow a.Sb], \\ [S \rightarrow .aSb], \\ [S \rightarrow .]\} \end{array} \right\}$	$s1$	$rS \rightarrow \epsilon$		2
2	$\{[S \rightarrow aS.b]\}$		$s3$		
3	$\{[S \rightarrow aSb.]\}$		$rS \rightarrow aSb$	$rS \rightarrow aSb$	
4	$\{[S' \rightarrow S.]\}$			$accept$	

Compressed Parser table for $S \to AB, S \to A, A \to a, B \to a$

s	sets of items	symbols		goto		
		a	$\#$	A	B	S
0	$\left\{\begin{array}{l}[S' \to .S],\\ [S \to .AB],\\ [S \to .A],\\ [A \to .a]\end{array}\right\}$	$s1$		2		5
1	$\{[A \to a.]\}$	$rA \to a$	$rA \to a$			
2	$\left\{\begin{array}{l}[S \to A.B],\\ [S \to A.],\\ [B \to .a]\end{array}\right\}$	$s3$	$rS \to A$		4	
3	$\{[B \to a.]\}$		$rB \to a$			
4	$\{[S \to AB.]\}$		$rS \to AB$			
5	$\{[S' \to S.]\}$		a			

Parsing aa

Stack	Input	Action
$\$ 0$	$aa\#$	**shift** 1
$\$ 0\ 1$	$a\#$	**reduce** $A \to a$
$\$ 0\ 2$	$a\#$	**shift** 3
$\$ 0\ 2\ 3$	$\#$	**reduce** $B \to a$
$\$ 0\ 2\ 4$	$\#$	**reduce** $S \to AB$
$\$ 0\ 5$	$\#$	**accept**

Algorithm LR(1)–PARSER

```
type state = set of item;
var  lookahead: symbol;
     (* the next not yet consumed input symbol *)
     S : stack of state;
proc  scan;
     (* reads the next symbol into lookahead *)
proc  acc;
     (* report successful parse; halt *)
proc  err(message: string);
     (* report error; halt *)
scan; push(S, q_d);
forever do
   case  action[top(S), lookahead] of
     shift: begin  push(S, goto[top(S), lookahead]);
                   scan
            end ;
     reduce (X → α) : begin
                          pop^|α|(S); push(S, goto[top(S), X]);
                          output("X → α")
                      end ;
     accept:  acc;
     error:   err("...");
   end case
od
```

Construction of LR(1)–Parsers

Classes of LR–Parsers:

canonical LR(1): analyze languages of LR(1)–grammars,

SLR(1): use $FOLLOW_1$ to resolve conflicts,
size is size of LR(0)–parser,

LALR(1): refine lookahead sets compared to $FOLLOW_1$,
size is size of LR(0)–parser.
BISON is an LALR(1)–parser generator.

LR(1)–Conflicts

Set of LR(1)-items I has a

shift-reduce-conflict:
if exists at least one item $[X \rightarrow \alpha.a\beta, L_1] \in I$
and at least one item $[Y \rightarrow \gamma., L_2] \in I$,
and if $a \in L_2$.

reduce-reduce-conflict:
if it contains at least two items $[X \rightarrow \alpha., L_1]$
and $[Y \rightarrow \beta., L_2]$ where $L_1 \cap L_2 \neq \emptyset$.

A state with a conflict is called **inadequate**.

Construction of an LR(1)–Action Table

Input: set of LR(1)–states Q without inadequate states
Output: action-table
Method:

foreach $q \in Q$ **do**
    foreach $LR(1)\text{–}item\ [K, L] \in q$ **do**
        if $K = [S' \rightarrow S.]$ **and** $L = \{\#\}$
        then $action[q, \#] := accept$
        elseif $K = [X \rightarrow \alpha.]$
        then **foreach** $a \in L$ **do**
                $action[q, a] := reduce(X \rightarrow \alpha)$
                od
        elseif $K = [X \rightarrow \alpha.a\beta]$
        then $action[q, a] := shift$
        fi
    od
od;
foreach $q \in Q$ **and** $a \in V_T$ *such that* action$[q, a]$ *is undef.* **do**
    action$[q, a] := error$
od;

Computing Canonical LR(1)–States

Input: cfg G
Output: char. NFA of a canonical LR(1)–Parser for G.
Method: The states and transitions are constructed
using the functions *Start*, *Closure* and *Succ*.

var q, q' : **set of item**;
var Q : **set of set of item**;
var δ : **set of item** $\times$ $(V_N \cup V_T) \rightarrow$ **set of item**;
function *Start:* **set of item**;
  return$(\{[S' \rightarrow .S, \{\#\}]\})$;
function $Closure(q$: **set of item**$)$: **set of item**;
begin
  foreach $[X \rightarrow \alpha.Y\beta, L]$ **in** q **and** $Y \rightarrow \gamma$ **in** P **do**
    if *exist.* $[Y \rightarrow .\gamma, L']$ **in** q
    then *replace* $[Y \rightarrow .\gamma, L']$ *by* $[Y \rightarrow .\gamma, L' \cup \varepsilon\text{-}ffi(\beta L)]$
    else $q := q \cup \{[Y \rightarrow .\gamma, \varepsilon\text{-}ffi(\beta L)]\}$
    fi
  od;
  return(q)
end ;
function $Succ(q$: **set of item**, $Y : V_N \cup V_T)$: **set of item**;
  return$(\{[X \rightarrow \alpha Y.\beta, L] \mid [X \rightarrow \alpha.Y\beta, L] \in q\})$;

```
begin
  Q := {Closure(Start)};   δ := ∅;
  foreach q in Q and  X in V_N ∪ V_T do
    let  q' = Closure(Succ(q, X)) in
      if  q' ≠ ∅
      then
        if  q' not in Q
        then  Q := Q ∪ {q'}
        fi;
        δ := δ ∪ {q --X--> q'}
      fi
    tel
  od
end .
```

- The test "q' not in Q" uses an equality test on LR(1)–items. $[K_1, L_1] = [K_2, L_2]$ iff $K_1 = K_2$ and $L_1 = L_2$.
- The canonical LR(1)–parser generator splits LR(0)–states.
- LALR(1)–parsers could be generated by
  - using the equality' test $[K_1, L_1] = [K_2, L_2]$ iff $K_1 = K_2$.
  - and replacing an existing state q'' by a state, in which equal' items $[K_1, L_1] \in q'$ and $[K_2, L_2] \in q''$ are merged to new items $[K_1, L_1 \cup L_2]$.

Example from G_0

$$
\begin{aligned}
S'_0 = \;& Closure(\text{Start}) \\
= \{ & [S \rightarrow .E, \{\#\}] \\
& [E \rightarrow .E + T, \{\#, +\}], \\
& [E \rightarrow .T, \{\#, +\}], \\
& [T \rightarrow .T * F, \{\#, +, *\}], \\
& [T \rightarrow .F, \{\#, +, *\}], \\
& [F \rightarrow .(E), \{\#, +, *\}], \\
& [F \rightarrow .\mathbf{id}, \{\#, +, *\}] \;\; \}
\end{aligned}
$$

$$
\begin{aligned}
S'_1 = \;& Closure(Succ(S'_0, E)) \\
= \{ & [S \rightarrow E., \{\#\}], \\
& [E \rightarrow E. + T, \{\#, +\}] \;\; \}
\end{aligned}
$$

$$
\begin{aligned}
S'_2 = \;& Closure(Succ(S'_1, T)) \\
= \{ & [E \rightarrow T., \{\#, +\}], \\
& [T \rightarrow T. * F, \{\#, +, *\}] \;\; \}
\end{aligned}
$$

$$
\begin{aligned}
S'_6 = \;& Closure(Succ(S'_1, +)) \\
= \{ & [E \rightarrow E + .T, \{\#, +\}], \\
& [T \rightarrow .T * F, \{\#, +, *\}], \\
& [T \rightarrow .F, \{\#, +, *\}], \\
& [F \rightarrow .(E), \{\#, +, *\}], \\
& [F \rightarrow .\mathbf{id}, \{\#, +, *\}] \;\; \}
\end{aligned}
$$

$$
\begin{aligned}
S'_9 = \;& Closure(Succ(S'_6, T)) \\
= \{ & [E \rightarrow E + T., \{\#, +\}], \\
& [T \rightarrow T. * F, \{\#, +, *\}] \;\; \}
\end{aligned}
$$

Inadequate LR(0)–states S_1, S_2 und S_9 are adequate after adding lookahead sets.

S'_1 shifts under "$+$", reduces under "$\#$".
S'_2 shifts under "$*$", reduces under "$\#$" and "$+$",
S'_9 shifts under "$*$", reduces under "$\#$" and "$+$".

Non–canonical LR–Parsers

SLR(1)– and LALR(1)–Parsers are constructed by

1. building an LR(0)–parser,
2. testing for inadequate LR(0)–states,
3. extending complete items by lookahead sets,
4. testing for inadequate LR(1)–states.

The lookahead set for item $[X \rightarrow \alpha.\beta]$ in q is denoted $LA(q, [X \rightarrow \alpha.\beta])$
The function $LA : Q_d \times It_G \rightarrow 2^{V_T \cup \{\#\}}$ is differently defined for SLR(1) (LA_S) und LALR(1) (LA_L).

SLR(1)– and LALR(1)–Parsers have the size of the LR(0)–parser, i.e., no states are split.

Constructing SLR(1)–Parsers

- Add $LA_S(q, [X \rightarrow \alpha.]) = FOLLOW_1(X)$ to all complete items;
- Check for inadequate SLR(1)–states.
- Cfg G is **SLR(1)** if it has no inadequate SLR(1)–states.

Example from G_0:
Extend the complete items in the inadequate states S_1, S_2 and S_9 by $FOLLOW_1$ as their lookahead sets.

$$
\begin{array}{llll}
S_1'' = \{ & [S \rightarrow E., \{\#\}], & & \text{conflict removed,} \\
 & [E \rightarrow E.+T]\} & & \text{"} + \text{" is not in } \{\#\} \\
 & & & \\
S_2'' = \{ & [E \rightarrow T., \{\#, +,)\}], & & \text{conflict removed,} \\
 & [T \rightarrow T.*F]\ \} & & \text{"} * \text{" is not in } \{\#, +,)\} \\
 & & & \\
S_9'' = \{ & [E \rightarrow E+T., \{\#, +,)\}], & & \text{conflict removed,} \\
 & [T \rightarrow T.*F]\ \} & & \text{"} * \text{" is not in } \{\#, +,)\}
\end{array}
$$

G_0 is an SLR(1)–grammar.

A Non–SLR(1)–Grammar

$$\begin{array}{lcl} S' & \to & S \\ S & \to & L = R \mid R \\ L & \to & *R \mid \textbf{id} \\ R & \to & L \end{array}$$

Slightly abstracted form of the C–assignment.

States of the LR–DFA as sets of items.

$$\begin{array}{llll} S_0 & = \{ & [S' \to .S], & \\ & & [S \to .L = R], & \\ & & [S \to .R], & \\ & & [L \to . * R], & \\ & & [L \to .\textbf{id}], & \\ & & [R \to .L] & \} \end{array}$$

$$S_1 = \{ \ [S' \to S.] \ \}$$

$$\begin{array}{llll} S_2 & = \{ & [S \to L. = R], & \\ & & [R \to L.] & \} \end{array}$$

$$S_3 = \{ \ [S \to R.] \ \}$$

$$\begin{array}{llll} S_4 & = \{ & [L \to * .R], & \\ & & [R \to .L], & \\ & & [L \to . * R], & \\ & & [L \to .\textbf{id}] & \} \end{array}$$

$$S_5 = \{ \ [L \to \textbf{id}.] \ \}$$

$$\begin{array}{llll} S_6 & = \{ & [S \to L = .R], & \\ & & [R \to .L], & \\ & & [L \to . * R], & \\ & & [L \to .\textbf{id}] & \} \end{array}$$

$$S_7 = \{ \ [L \to * R.] \ \}$$

$$S_8 = \{ \ [R \to L.] \ \}$$

$$S_9 = \{ \ [S \to L = R.] \ \}$$

S_2 is the only inadequate LR(0)–state.
Extend $[R \ \rightarrow \ L.] \in S_2$ by $FOLLOW_1(R) = \{\#, =\}$
does not remove the shift-reduce-conflict,
since the symbol to shift, "$=$", is in the lookahead set.

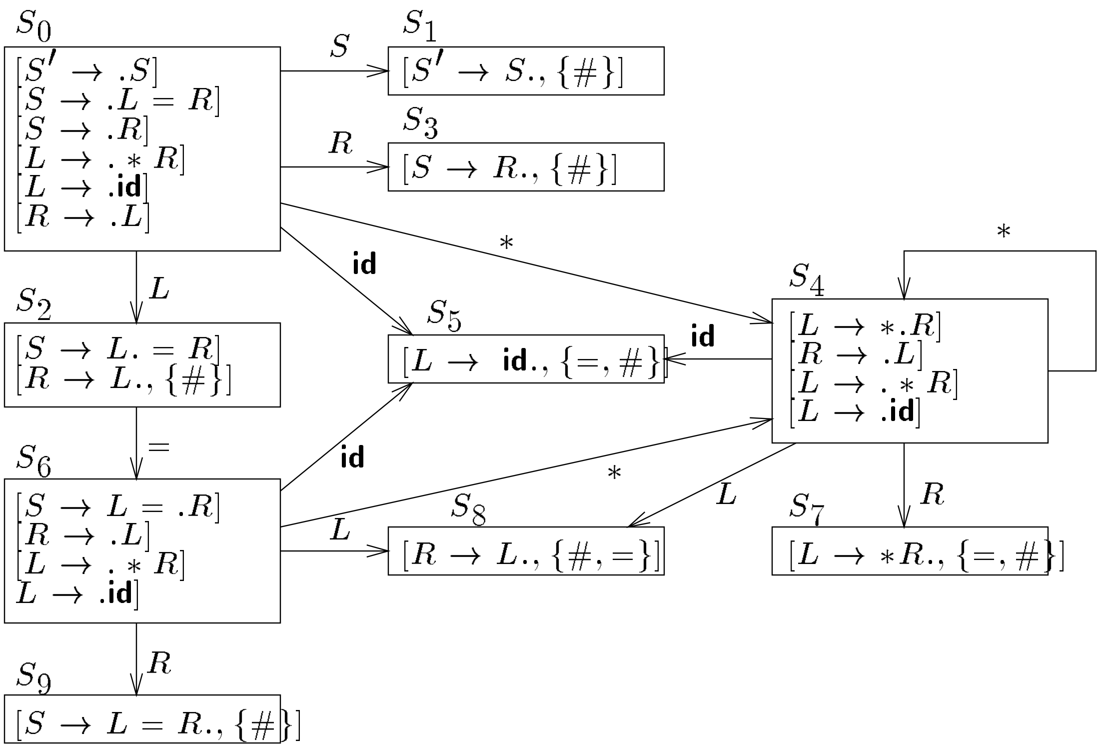
LALR(1)–Parsers

SLR(1): $LA_S(q, [X \rightarrow \alpha.]) =$
$\{a \in V_T \cup \{\#\} \mid S'\# \overset{*}{\Longrightarrow} \beta X a \gamma\} = FOLLOW_1(X)$

LALR(1): $LA_L(q, [X \rightarrow \alpha.]) =$
$\{a \in V_T \cup \{\#\} \mid S'\# \underset{rm}{\overset{*}{\Longrightarrow}} \beta X a w$ and $\delta_d^*(q_d, \beta\alpha) = q\}$
Lookahead set $LA_L(q, [X \rightarrow \alpha.])$ depends on the state q.

- Add $LA_L(q, [X \rightarrow \alpha.])$ to all complete items;
- Check for inadequate LALR(1)–states.
- Cfg G is **LALR(1)** if it has no inadequate LALR(1)–states.
- Definition is not constructive.
- Construction by modifying the LR(1)–Parser Generator, merging items with identical cores.

Non–SLR–Example



Grammar is LALR(1)–grammar.

Interesting Non $LR(1)$ Grammars

- Common "derived" prefix

$$\begin{array}{rcl} A & \rightarrow & B_1 ab \\ A & \rightarrow & B_2 ac \\ B_1 & \rightarrow & \epsilon \\ B_2 & \rightarrow & \epsilon \end{array}$$

- Optional non-terminals

$$\begin{array}{rcl} St & \rightarrow & OptLab\, St' \\ OptLab & \rightarrow & id : \\ OPtlab & \rightarrow & \epsilon \\ St' & \rightarrow & id := Exp \end{array}$$

- Ambiguous:
  - Ambiguous arithmetic expressions
  - Dangling-else

Bison Specification

Definitions: start-non-terminal+tokens+associativity
%%
Productions
%%
C-Routines

Bison Example

```
%{
int line_number = 1 ;
int error_occ = 0 ;
void yyerror(char *);
#include <stdio.h>
%}
%start exp
%left '+'
%left '*'
%right UMINUS
%token INTCONST
%%
exp:   exp '+' exp { $$ = $1 + $3 ;}
   |   exp '*' exp { $$ = $1 * $3 ;}
   |   '-' exp  %prec UMINUS { $$ = - $2 ; }
   |  '(' exp ')' { $$ = $2 ; }
   | INTCONST
   ;
%%
void yyerror(char *message)
{ fprintf(stderr, "%s near line %ld. \n", message, line_number);
  error_occ=1;
}
```

Flex for the Example

```
%{
#include <math.h>
#include "calc.tab.h"
extern int line_number;
%}
Digit [0-9]
%%
{Digit}+                  {yylval = atoi(yytext) ;
                           return(INTCONST); }
\n    {line_number++ ; }
[\t ]+                    ;
.                         {return(*yytext); }
%%
```