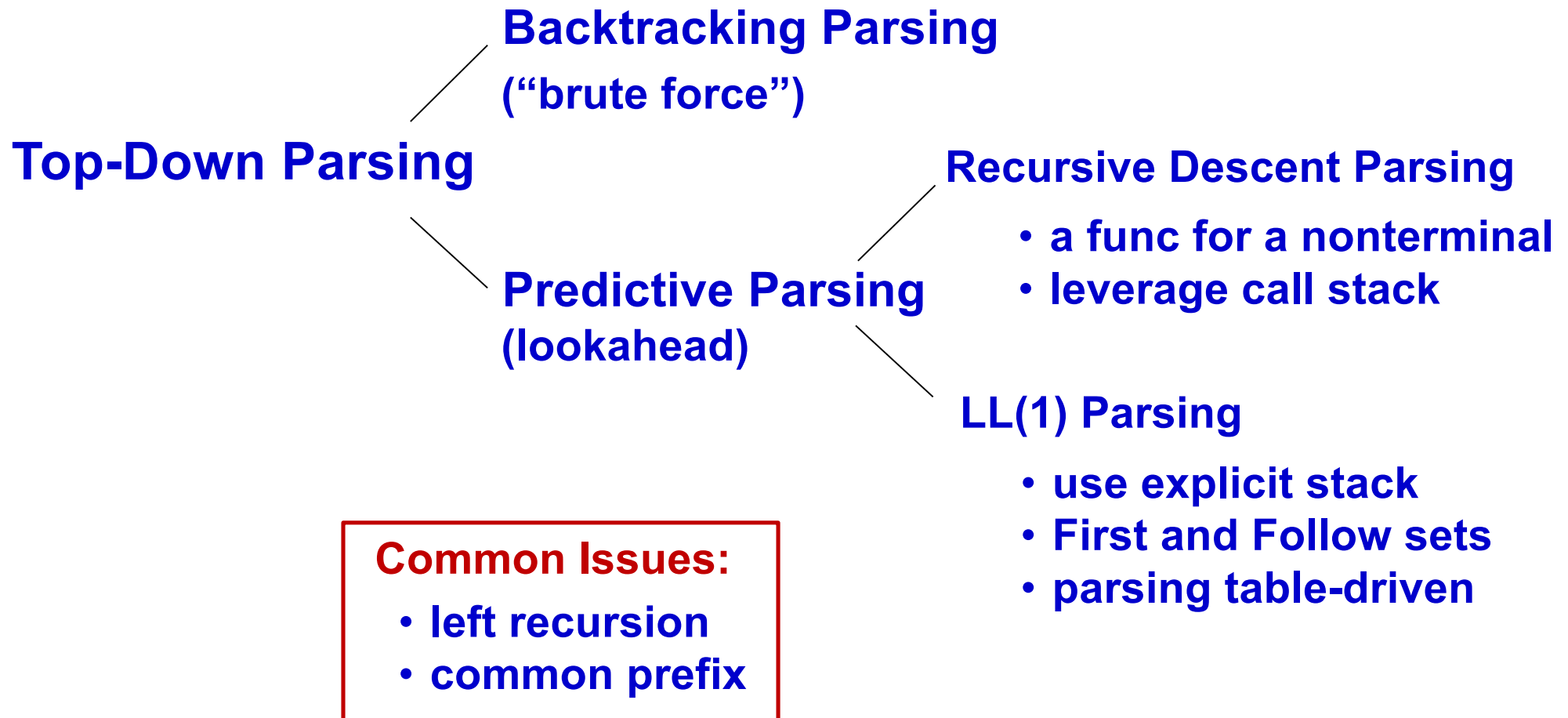


Summary: Top-Down Parsing



Example



$S \rightarrow A \mid B C$

$A \rightarrow a A \mid \epsilon$

$B \rightarrow b B \mid \epsilon$

$C \rightarrow c C \mid d C \mid \epsilon$

	FIRST	FOLLOW
S	a, b, c, d, ϵ	\$
A	a, ϵ	\$
B	b, ϵ	c, d, \$
C	c, d, ϵ	\$

	a	b	c	d	\$
S	$S \rightarrow A$	$S \rightarrow B C$	$S \rightarrow B C$	$S \rightarrow B C$	$S \rightarrow A$ $S \rightarrow B C$
A	$A \rightarrow a A$				$A \rightarrow \epsilon$
B		$B \rightarrow b B$	$B \rightarrow \epsilon$	$B \rightarrow \epsilon$	$B \rightarrow \epsilon$
C			$C \rightarrow c C$	$C \rightarrow d C$	$C \rightarrow \epsilon$

Bottom-up Parsing

Basic Idea :

- Scan the input string from left to right.
- Try to construct a parse tree starting at the bottom (i.e., the leaves) and working towards the root.

Shift-reduce parsing :

Basic Idea : Apply a sequence of “*reductions*” to transform the input string to the start symbol of the grammar.

reduction: replace a substring matching the RHS of a production by the LHS.

Example : Consider the grammar

$$S \longrightarrow aABe$$

$$A \longrightarrow Abc$$

$$A \longrightarrow b$$

$$B \longrightarrow d$$

Input: **abbcde**
 \rightsquigarrow **aAbcde**
 \rightsquigarrow **aAde**
 \rightsquigarrow **aABe**
 \rightsquigarrow **S**

Handles

Intuition : A *handle* of a string s is a substring α s.t.:

1. α matches the RHS of a production $A \longrightarrow \alpha$; and
2. replacing α by the LHS A represents a step in the reverse of a *rightmost derivation* of s .

Example : Consider the grammar

$$\begin{aligned} S &\rightarrow aABe \\ A &\rightarrow Abc \mid b \\ B &\rightarrow d \end{aligned}$$

The rightmost derivation for the input **abbcede** is:

$$\begin{aligned} \underline{S} &\Rightarrow \underline{a}A\underline{B}e \Rightarrow \underline{a}A\underline{d}e \Rightarrow \underline{a}A\underline{bc}de \\ &\Rightarrow \underline{a}bcde. \end{aligned}$$

The string **aAbcde** can be reduced in two ways:

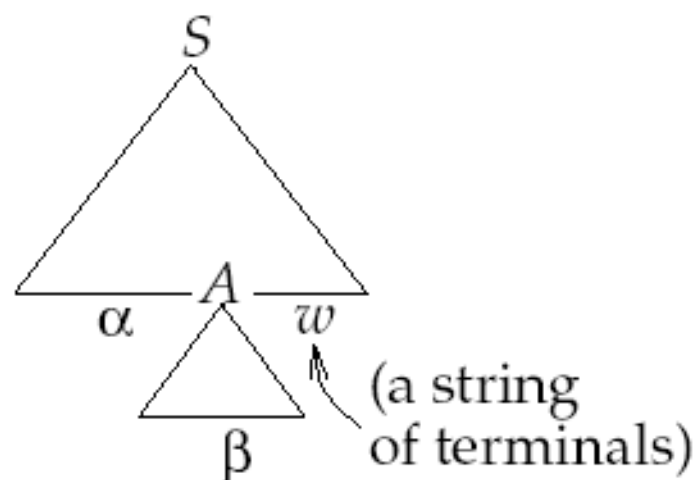
1. **aAbcde** \rightsquigarrow **aAde**; and
2. **aAbcde** \rightsquigarrow **aAbcBe**.

But (2) is not in a rightmost derivation, so **Abc** is the only handle.

Handles : cont'd

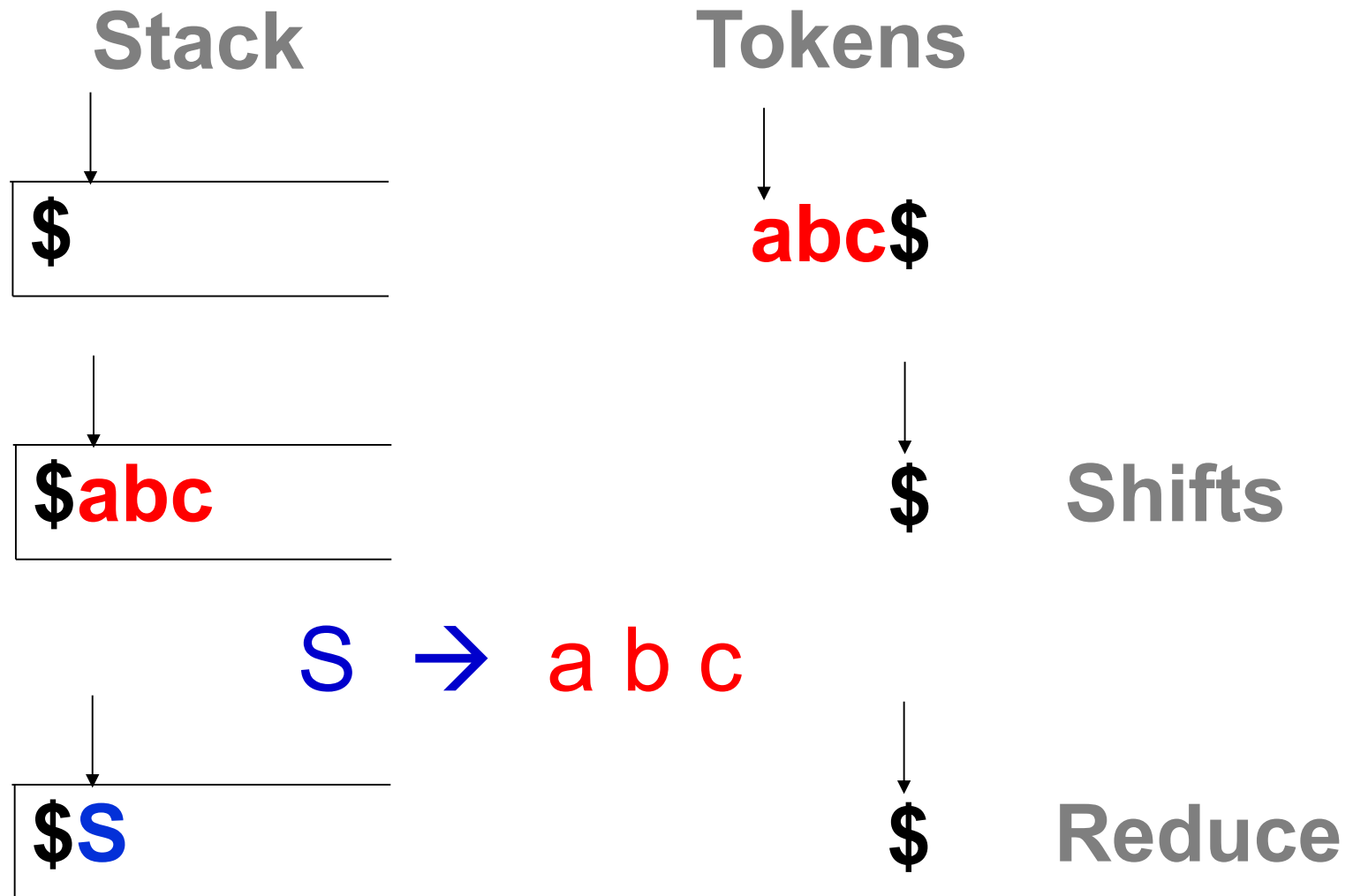
Definition : A *handle* of a right-sentential form γ is

1. a production $A \rightarrow \beta$, and
2. a position in γ where β may be found and replaced by A to produce the *previous* sentential form in a rightmost derivation of γ .



The handle $A \rightarrow \beta$ in $\alpha\beta w$

Shift Reduce Parsing



Stack Implementation of Shift-Reduce Parsing:

Data Structures :

- *the stack*, its bottom marked by \$, initially empty.
- *the input string*, its right end marked by \$, initially w .

Action :

repeat

1. *Shift* zero or more input symbols onto the stack, until a handle β is on the top of the stack.
2. *Reduce* β to the LHS of the appropriate production.

until ready to accept.

Acceptance : When the stack contains the start symbol and the input is empty.

Example : Consider the grammar

$S \rightarrow aABe$

$A \rightarrow Abc$

$A \rightarrow b$

$B \rightarrow d$

Input: **abbcde**

\rightsquigarrow **aAbcde**

\rightsquigarrow **aAde**

\rightsquigarrow **aABe**

\rightsquigarrow **S**

	c		e	
	b	d	B	
b	A	A	A	
a	a	a	a	S
\$	\$	\$	\$	\$

Example :

Grammar: $S \rightarrow aABe$
 $A \rightarrow Abc \mid b$
 $B \rightarrow d$

Input string : **abbcde**

<i>Stack (\rightarrow)</i>	<i>Input</i>	<i>Action</i>
\$	abbcde\$	shift
\$a	bbcde\$	shift
\$ab	bcde\$	reduce by $A \rightarrow b$
\$aA	bcde\$	shift
\$aAb	cde\$	shift
\$aAbc	de\$	reduce by $A \rightarrow Abc$
\$aA	de\$	shift
\$aAd	e\$	reduce by $B \rightarrow d$
\$aAB	e\$	shift
\$aABe	\$	reduce by $S \rightarrow aABe$
\$S	\$	accept

Conflicts during Shift-Reduce Parsing :

1. Can't decide whether to shift or to reduce (*"shift-reduce conflict"*).

Example : "dangling else":

$$\begin{aligned} Stmt \longrightarrow & \text{if } Expr \text{ then } Stmt \mid \\ & \text{if } Expr \text{ then } Stmt \text{ else } Stmt \mid \dots \end{aligned}$$

2. Can't decide which of several possible reductions to make (*"reduce-reduce conflict"*).

Example :

$$\begin{aligned} Stmt \longrightarrow & \text{id } (params) \mid Expr := Expr \mid \dots \\ Expr \longrightarrow & \text{id } (params) \end{aligned}$$

Given the input A(I, J) the parser doesn't know whether it's a procedure call or an array reference.

LR Parsing

- Bottom-up.
- LR(k) parser:
 - Scans the input L-to-R.
 - Produces a Rightmost derivation.
 - Uses k-symbol lookahead.

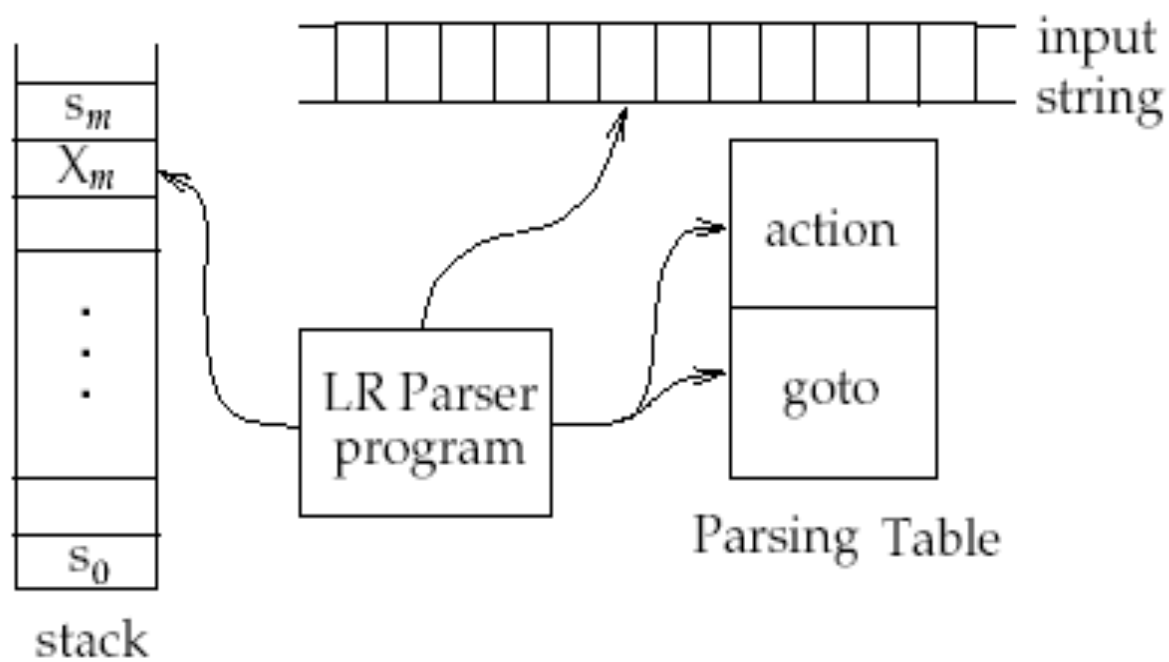
Advantages :

- Very general and flexible.
- Efficiently implemented.
- Parses a large class of grammars.

Disadvantages :

- Difficult to implement by hand for typical programming language grammars.
(Use tools such as **yacc** or **bison**.)

Schematic of an LR Parser :



- The driver program is the same for all LR parsers (SLR(1), LALR(1), LR(1), ...) : only the parsing table changes.

- The stack holds strings of the form

$$s_0 X_1 s_1 X_2 s_2 \cdots X_m s_m$$

where s_m is on top, the s_i are "states", and X_i are grammar symbols.

- The configuration of an LR parser is given by a pair $\langle \text{stack contents, unexpanded input} \rangle$.

A configuration $\langle s_0 X_1 s_1 \cdots X_m s_m, a_i a_{i+1} \cdots a_n \rangle$ represents the right-sentential form

$$X_1 \cdots X_m a_i a_{i+1} \cdots a_n$$

The sequence of symbols $X_1 \cdots X_m$ on the parser stack is called a viable prefix of the right sentential form.

LR Parse Tables

- The parsing table consists of two parts: a parsing **action** function, and a **goto** function.
- For a given configuration of the parser, the next move is determined by the parse table entry

$\text{action}[s_m, a_i]$.

where s_m is the topmost state on the stack, and a_i is the next input symbol.

- An action table entry can be of four types:
 1. **shift** s , where s is a state.
 2. **reduce** by a grammar production $A \rightarrow \beta$.
 3. **accept**
 4. **error**

LR Parsing : cont'd

Suppose the parser configuration is

$$\langle s_0 X_1 s_1 \cdots X_m s_m, a_i \cdots a_n \$ \rangle.$$

- if $\text{action}[s_m, a_i] = \text{shift } s$ then the parser executes a shift move. The new configuration is

$$\langle s_0 X_1 s_1 \cdots X_m s_m \underbrace{a_i s}_{\text{pushed}}, a_{i+1} \cdots a_n \$ \rangle.$$

- if $\text{action}[s_m, a_i] = \text{reduce } A \rightarrow \beta$ then the parser does a reduce move. The new configuration is

$$\langle s_0 X_1 s_1 \cdots X_{m-r} s_{m-r} \underbrace{A s}_{\text{new}} \quad a_i \cdots a_n \$ \rangle.$$

where

- $r = \text{length of } \beta$; and
 - $s = \text{goto}[s_{m-r}, A]$.
- if $\text{action}[s_m, a_i] = \text{accept}$ then parsing is done.
 - if $\text{action}[s_m, a_i] = \text{error}$ the parser calls an error recovery routine.

5.2. Finite Automata to recognize Viable Prefixes

Definition : An *LR(0) item* of a grammar G is a production of G with a dot '.' added at some position in the RHS.

Example : The production $A \rightarrow aAb$ gives the items

$$A \rightarrow \cdot aAb$$

$$A \rightarrow a \cdot Ab$$

$$A \rightarrow aA \cdot b$$

$$A \rightarrow aAb \cdot$$

Intuition : An item $A \rightarrow \alpha \cdot \beta$ denotes:

- we have seen a string derivable from α ; and
- we hope to see a string derivable from β .

Overall Goal : Given a grammar with start symbol S ,

- Construct an augmented grammar by adding a new start symbol S' and production $S' \rightarrow S$;
- Starting with the item $S' \rightarrow \cdot S$, recognize the viable prefix $S' \rightarrow S \cdot$.

Viable Prefix DFA

1. closure :

Definition : If I is a set of items for a grammar G , then $closure(I)$ is the set of items constructed as follows:

repeat

1. add every item in I to $closure(I)$;
2. if $A \rightarrow \alpha \cdot B \beta$ is in $closure(I)$ and $B \rightarrow \gamma$ is a production of G , then add $B \rightarrow \cdot \gamma$ to $closure(I)$.

until no new item can be added to $closure(I)$.

Intuition : If $A \rightarrow \alpha \cdot B \beta$ is in $closure(I)$ then we hope to see a string derivable from B in the input. So if $B \rightarrow \gamma$ is a production of G , then we should hope to see a string derivable from γ in the input. Hence, $B \rightarrow \cdot \gamma$ is in $closure(I)$.

Viable Prefix DFA – cont'd:

2. goto :

Definition : If I is a set of items for a grammar G and X a grammar symbol, then $goto(I, X)$ is the set of items

$$closure(\{A \rightarrow \alpha X \cdot \beta \mid A \rightarrow \alpha \cdot X \beta \in I\}).$$

Intuition :

- A set of items I corresponds to a state.
- If $A \rightarrow \alpha \cdot X \beta \in I$ then
 - we've seen a string derivable from α ; and
 - we hope to see a string derivable from $X\beta$;

- now suppose we see a string derivable from X : the resulting state should be one in which:
 - we've seen a string derivable from αX ; and
 - we hope to see a string derivable from β ;
- The item corresponding to this is $A \dashrightarrow \alpha X \cdot \beta$.

Constructing the Viable Prefix DFA for LR(0) Items

- Given a grammar G with start symbol S , construct the *augmented grammar* by adding a special production

$$S' \longrightarrow S$$

where S' does not appear in G .

- Algorithm for constructing the canonical collection of LR(0) items for an augmented grammar G' :

begin

$C := \{closure(\{S' \longrightarrow \cdot S\})\};$

repeat

for each set of items $I \in C$ **do**

for each grammar symbol X **do**

if $goto(I, X) \neq \emptyset$ **then**

 add $goto(I, X)$ to C ;

fi

until no new set of items can be added to C ;

return C ;

end

Example

Original Grammar

$$E \rightarrow E + T \mid T$$

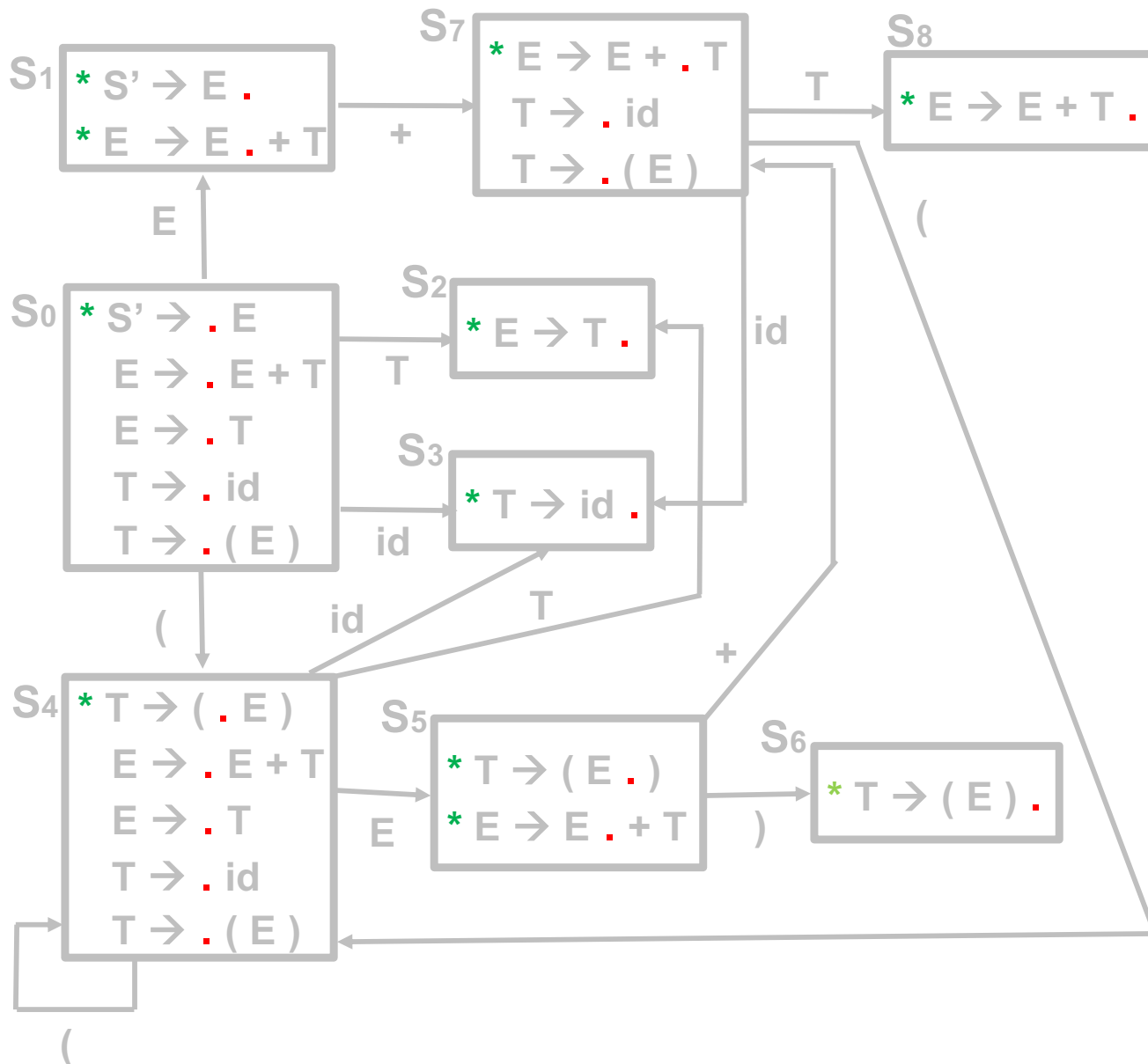
$$T \rightarrow \text{id} \mid (E)$$

Augmented Grammar

$$S' \rightarrow E$$

$$E \rightarrow E + T \mid T$$

$$T \rightarrow \text{id} \mid (E)$$



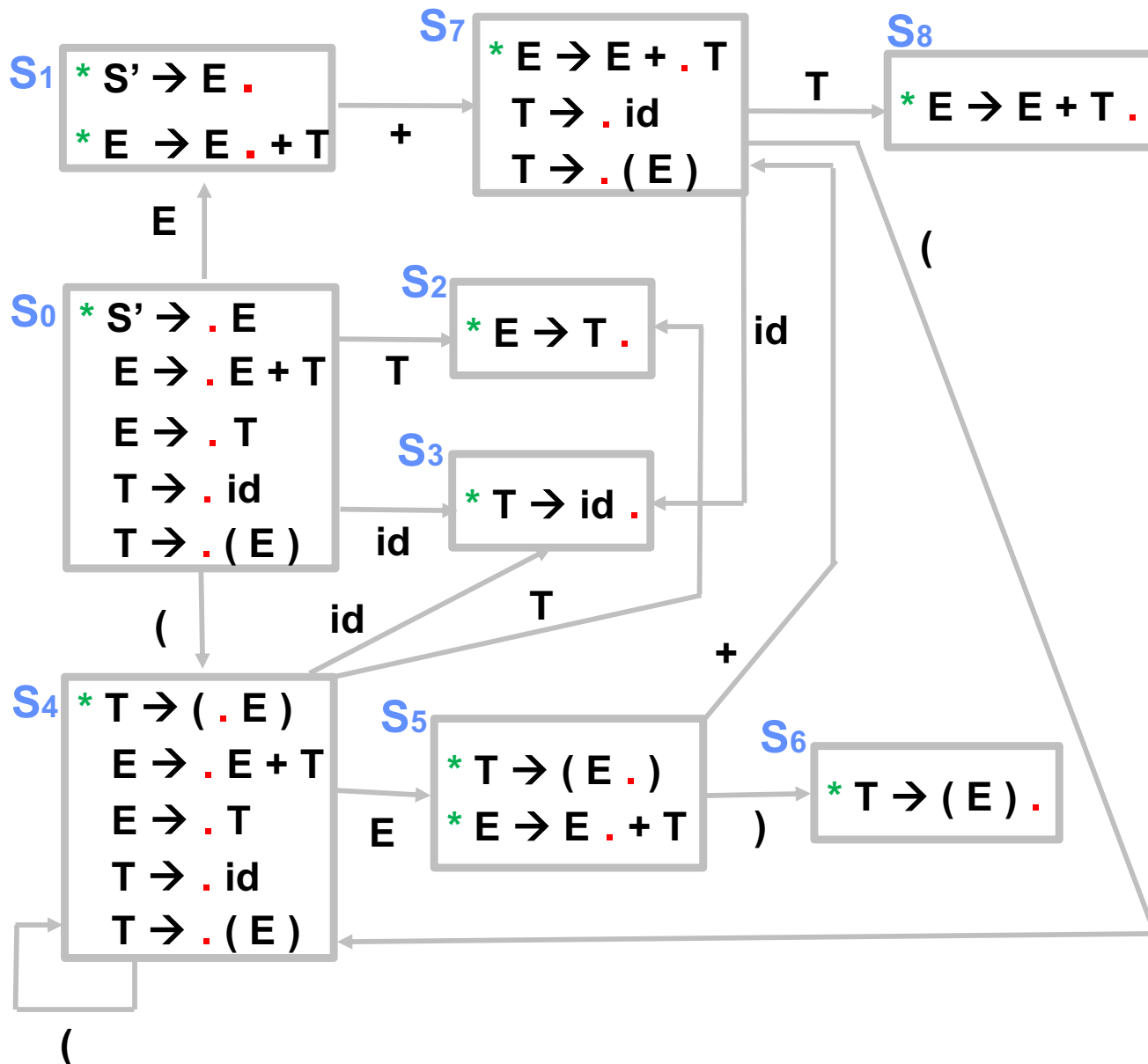
Augmented Grammar

$S' \rightarrow E$
 $E \rightarrow E + T \mid T$
 $T \rightarrow id \mid (E)$

Kernel items are
Marked with *

Rest of the items
added by closure

• Tells where we are
in the production



Augmented Grammar

$S' \rightarrow E$
 $E \rightarrow E + T \mid T$
 $T \rightarrow id \mid (E)$

Kernel items are
Marked with *

Rest of the items
added by closure

• Tells where we are
in the production

5.3. Constructing an SLR(1) Parse Table

1. Given a grammar G , construct the augmented grammar G' by adding the production $S' \rightarrow S$.
2. Construct $C = \{I_0, \dots, I_n\}$, the set of states of the viable prefix DFA for G' .
3. State i is constructed from I_i , with parsing action determined as follows:
 - (a) $A \rightarrow \alpha \cdot a \beta \in I_i$, a a terminal, $goto(I_i, a) = I_j$:
set $action[i, a] = \underline{shift\ j}$.
 - (b) $A \rightarrow \alpha \cdot \in I_i$, $A \neq S'$: for each $a \in FOLLOW(A)$,
set $action[i, a] = \underline{reduce\ A \rightarrow \alpha}$.
 - (c) $S' \rightarrow S \cdot \in I_i$: set $action[i, \$] = \underline{accept}$.

4. goto transitions are constructed as follows: for each nonterminal A , if $goto(I_i, A) = I_j$ then $goto[i, A] = j$.

5. All entries not defined by the above steps are made error.

If there are any multiply defined entries, then G is not SLR.

6. Initial state of the parser: that constructed from $I_0 \sim S' \rightarrow *S$.

	ACTION					GOTO		
	+	Id	()	\$	E	T	S
S0		S,S3	S,S4			S1	S2	
S1	S,S7				accept			
S2	R,#3			R,#3	R,#3			
S3	R,#4			R,#4	R,#4			
S4		S,S3	S,S4			S5	S2	
S5	S,S7			S,S6				
S6	R,#5			R,#5	R,#5			
S7		S,S3	S,S4				S8	
S8	R,#2			R,#2	R,#2			

#1	$S' \rightarrow E$
#2	$E \rightarrow E + T$
#3	$E \rightarrow T$
#4	$T \rightarrow id$
#5	$T \rightarrow (E)$

Follow(S') \rightarrow { \$ }
Follow(E) \rightarrow { +,), \$ }
Follow(T) \rightarrow { +,), \$ }

S - SHIFT R - REDUCE

S# - Next State

#n - Production Rule Number

The LR Parsing Algorithm

begin

set ip to point to the first symbol of the input w \$;

while TRUE **do**

let s be the state on top of the stack,
 a the symbol pointed at by ip ;

if $\text{action}[s, a] = \text{shift } s'$ **then**
 push a then s' on top of the stack;
 advance ip to the next input symbol;

else if $\text{action}[s, a] = \text{reduce } A \rightarrow \beta$ **then**
 pop $2 * |\beta|$ symbols off the stack;
 let s' be the state now on top of the stack;
 push A then $\text{goto}[A, s']$ on top of the stack;

else if $\text{action}[s, a] = \text{accept}$ **then return**;

else $\text{error}()$;
 fi

od

end

Stack	Input	Action
\$ S0	id + id \$	shift S3
\$ S0 id S3	+ id \$	red. $T \rightarrow id$ GOTO[S0,T]=S2
\$ S0 T S2	+ id \$	red $E \rightarrow T$ GOTO[S0,E]=S1
\$ S0 E S1	+ id \$	shift S7
\$ S0 E S1 + S7	id \$	shift S3
\$ S0 E S1 + S7 id S3	\$	red $T \rightarrow id$ GOTO[S7,T]=S8
\$ S0 E S1 + S7 T S8	\$	red $E \rightarrow E+T$ GOTO[S0,E]=S1
\$ S0 E S1	\$	accept

Limitations of SLR Parsing

Cannot handle many “reasonable” grammars, e.g.:

$$\begin{aligned} S &\longrightarrow R \mid L=R \\ L &\longrightarrow * R \mid \text{id} \\ R &\longrightarrow L \end{aligned}$$

The SLR parse table contains a state

$$I = \{S \longrightarrow L \cdot = R, R \longrightarrow L \cdot\}$$

which causes a shift/reduce conflict on ‘=’, since ‘=’ is in FOLLOW(L).

Problem : For an input

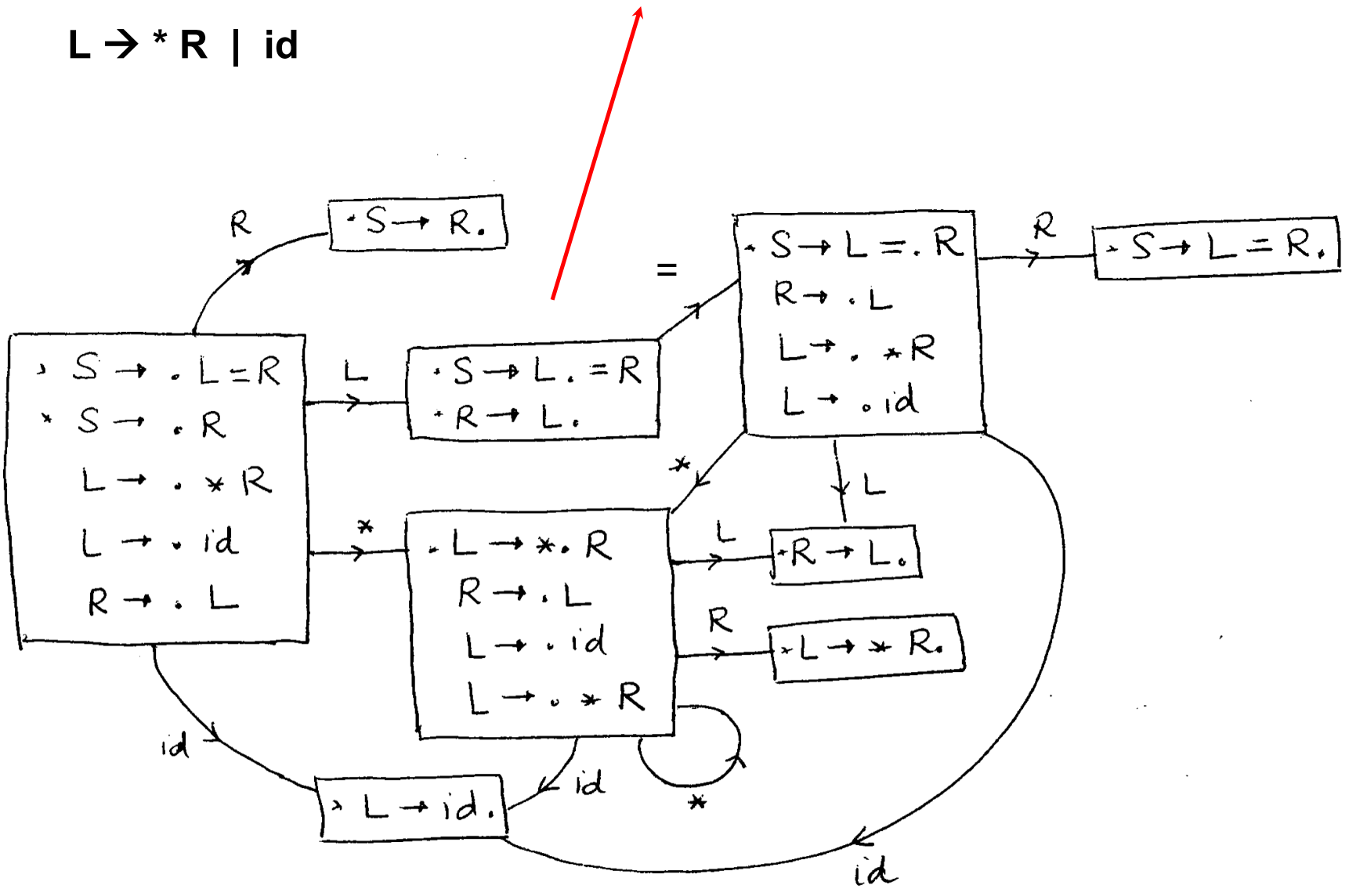
$$*id = id$$

we want to remember enough “left context” after seeing $*$ to make the right shift/reduce decision. SLR cannot do this adequately.

$S \rightarrow L = R \mid R$
 $R \rightarrow L$
 $L \rightarrow * R \mid id$

$S \rightarrow L . = R$
 $R \rightarrow L .$

Shift on =
 Since = is in Follow(R)
 so Reduce on =



5.4. LR(1) Parsing

Idea : Extend SLR parsing to incorporate lookahead.

LR(1) Item :

- Of the form $[A \rightarrow \alpha \cdot \beta, a]$, where **a** is a terminal or is the endmarker \$.
- The lookahead has no effect on items of the form $[A \rightarrow \alpha \cdot \beta, a]$, where $\beta \neq \epsilon$.
- For items of the form $[A \rightarrow \alpha \cdot, a]$, reduce only if the next symbol is **a**.

Note: For an item of the form $[A \rightarrow \alpha \cdot \beta, a]$, $a \in \text{FOLLOW}(A)$. But there may be $b \in \text{FOLLOW}(A)$ for which there is no item $[A \rightarrow \alpha \cdot \beta, b]$.

LR(1) Parsing: closure and goto Functions

1. closure(I) :

begin

$S := I;$

repeat

for(each item $[A \rightarrow \alpha \cdot B\beta, a] \in I,$

 each production $B \rightarrow \gamma,$

 each terminal $b \in \text{FIRST}(\beta a)$) **do**

 add $[B \rightarrow \cdot \gamma, b]$ to $S;$

until no new item can be added to $S;$

return $S;$

end

2. goto(I, X) :

begin

let $J = \{[A \rightarrow \alpha X \cdot \beta, a] \mid [A \rightarrow \alpha \cdot X\beta, a] \in I\};$

return $\text{closure}(J);$

end

Constructing the Viable Prefix DFA for LR(1) Items

- Given : An augmented grammar G' .

- Algorithm :

begin

$C := \{closure(\{[S' \rightarrow \cdot S, \$])\};$

repeat

for each set of items $I \in C$ **do**

for each grammar symbol X **do**

if $goto(I, X) \neq \emptyset$ **then**

 add $goto(I, X)$ to C ;

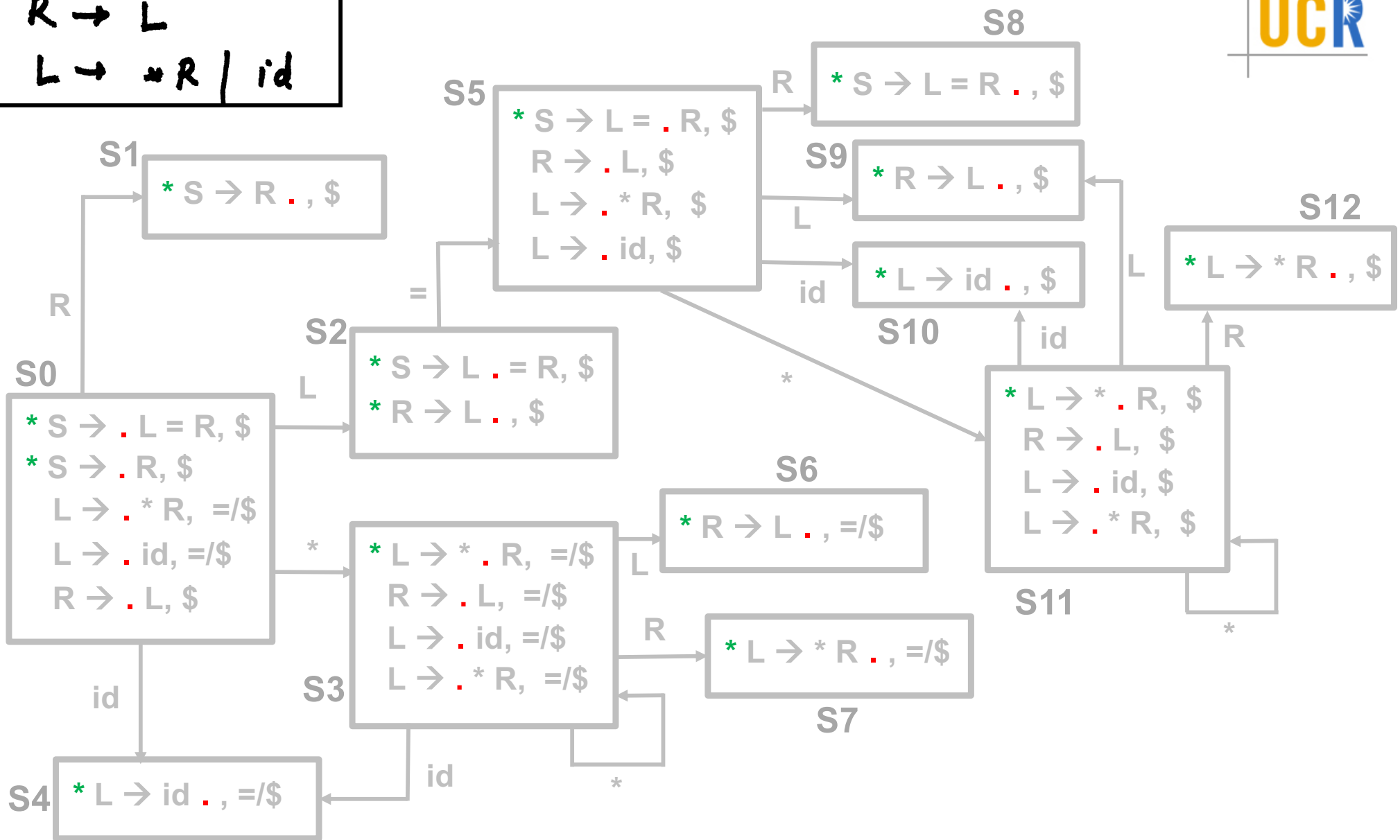
until no new set of items can be added to C ;

return C ;

end

- Note : The set of items construction is essentially the same as for the SLR(1) case.

$S \rightarrow L=R \mid R$
 $R \rightarrow L$
 $L \rightarrow *R \mid id$



	ACTION				GOTO		
	*	=	id	\$	S	R	L
S0	S,S3		S,S4			S1	S2
S1				accept			
S2		S,S5		R,R→L			
S3	S,S3		S,S4			S7	S6
S4		R,L→id		R,L→id			
S5	S,S11		S,S10			S8	S9
S6		R,R→L		R,R→L			
S7		R,L→*R		R,L→*R			
S8				accept			
S9				R,R→L			
S10				R,L→id			
S11	S,S11		S,S10			S12	S9
S12				R,L→*R			

S - SHIFT R - REDUCE S# - Next State #n - Production Number

\$ S0	* id = id \$
\$ S0 * S3	id = id \$
\$ S0 * S3 id S4	= id \$
\$ S0 * S3 L S6	= id \$
\$ S0 * S3 R S7	= id \$
\$ S0 L S2	= id \$
\$ S0 L S2 = S5	id \$
\$ S0 L S2 = S5 id S10	\$
\$ S0 L S2 = S5 L S9	\$
\$ S0 L S2 = S5 R S8	\$

accept

Constructing an LR(1) Parse Table

1. Given a grammar G , construct the augmented grammar G' by adding the production $S' \rightarrow S$.
2. Construct $C = \{I_0, \dots, I_n\}$, the viable prefix DFA for G' .
3. State i is constructed from I_i , with parsing action determined as follows:
 - (a) $[A \rightarrow \alpha \cdot a \beta, b] \in I_i$, a a terminal, $goto(I_i, a) = I_j$: set $action[i, a] = \underline{shift\ j}$.
 - (b) $[A \rightarrow \alpha \cdot, a] \in I_i, A \neq S'$: set $action[i, a] = \underline{reduce\ A \rightarrow \alpha}$.
 - (c) $[S' \rightarrow S \cdot, \$] \in I_i$: set $action[i, \$] = \underline{accept}$.

4. goto transitions are constructed as follows: for each nonterminal A , if $goto(I_i, A) = I_j$ then $goto[i, A] = j$.
5. All entries not defined by the above steps are made error.
If there are any multiply defined entries, then G is not LR(1).
6. Initial state of the parser: that constructed from $I_0 \sim [S' \rightarrow \cdot S, \$]$.

LR(1) vs. SLR(1) :

- LR(1) more powerful, can handle a strictly larger class of grammars than SLR(1).
- The parse tables for LR(1) become very large — may be impractical for realistic grammars.
- A compromise between parsing power and table size that is commonly used is seen in LALR parsers.

An LALR parser can be thought of as an LR(1) parser, some of whose states have been merged into a single state. This can be done in many (but not all) cases without causing problems.

The parsers generated by tools such as yacc and bison are LALR.

5.4.3. LALR(1) Parsing

Observation : Every SLR grammar is an LR(1) grammar, but the LR(1) parser usually has many more states than the SLR parser.

Many of these states differ only on the lookahead token. But the lookahead token does not play any role except on reductions.

Definition : The core of a set of LR(1) items I is

$$\text{core}(I) = \{J \mid [J, \mathbf{a}] \in I \text{ for some } \mathbf{a}\}$$

I.e., $\text{core}(I)$ is the set of first components of I .

Example : Suppose

$$I = \{[A \rightarrow c\cdot, \mathbf{a}], \\ [A \rightarrow c\cdot, \mathbf{b}], \\ [B \rightarrow c\cdot, \mathbf{c}]\}$$

Then,

$$\text{core}(I) = \{A \rightarrow c\cdot, B \rightarrow c\cdot\}$$

Merging sets of LR(1) Items

- If sets of items with the same core are merged, the parser behaves essentially as before.

However, some redundant reductions might be done before an error is detected.

- $core(goto(I, X))$ depends only on $core(I)$, so $goto$'s of merged sets may themselves be merged.

- Suppose we take a set C_0 of sets of LR(1) items for a given grammar, and merge those sets of items that have the same core to get a set C_1 of sets of LR(1) items.

LR(1) parse table construction using C_1 will not introduce any new shift/reduce conflicts compared to C_0 .

However, this can introduce new reduce/reduce conflicts.

Example of reduce/reduce conflicts due to merging :

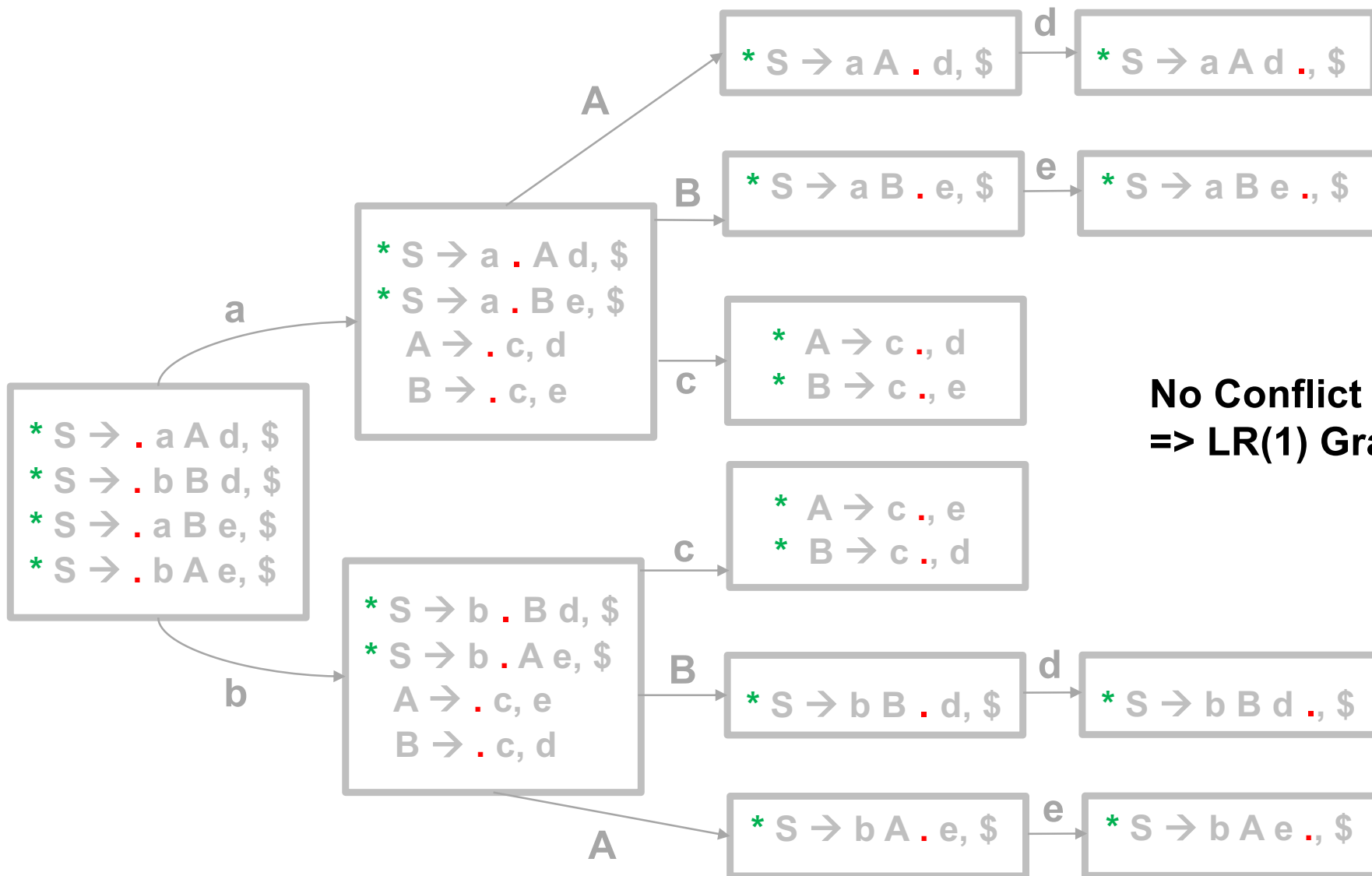
Consider the grammar given by

$$S \rightarrow aAd \mid bBd \mid aBe \mid bAe$$
$$A \rightarrow c$$
$$B \rightarrow c$$

$S \rightarrow aAd \mid bBd \mid aBe \mid bAe$

$A \rightarrow c$

$B \rightarrow c$



$S \rightarrow aAd \mid bBd \mid aBe \mid bAe$

$A \rightarrow c$

$B \rightarrow c$



$A \rightarrow c \cdot, d$
 $B \rightarrow c \cdot, e$

$A \rightarrow c \cdot, e$
 $B \rightarrow c \cdot, d$

Merge

$A \rightarrow c \cdot, d/e$
 $B \rightarrow c \cdot, d/e$

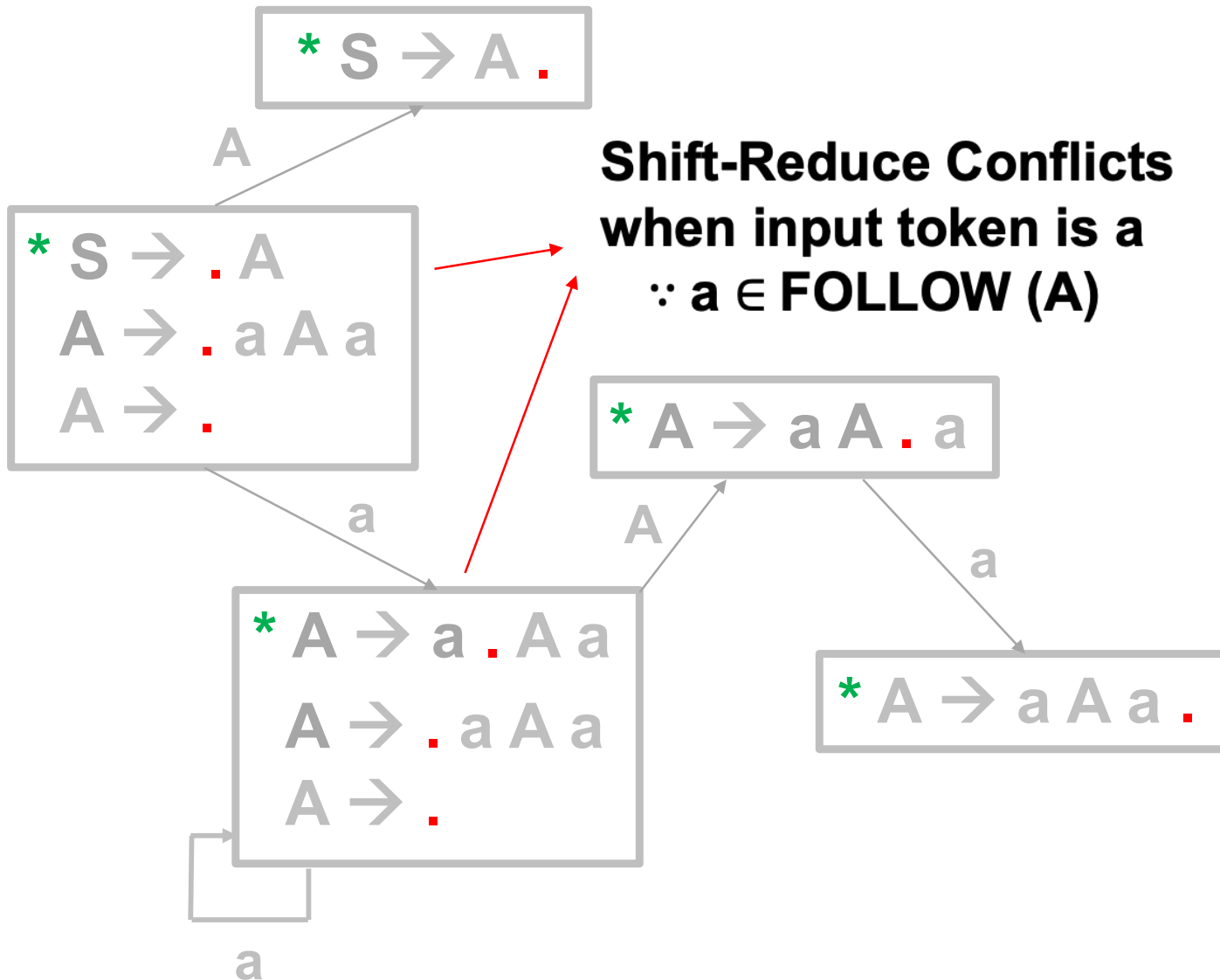
Contains

reduce-reduce conflict

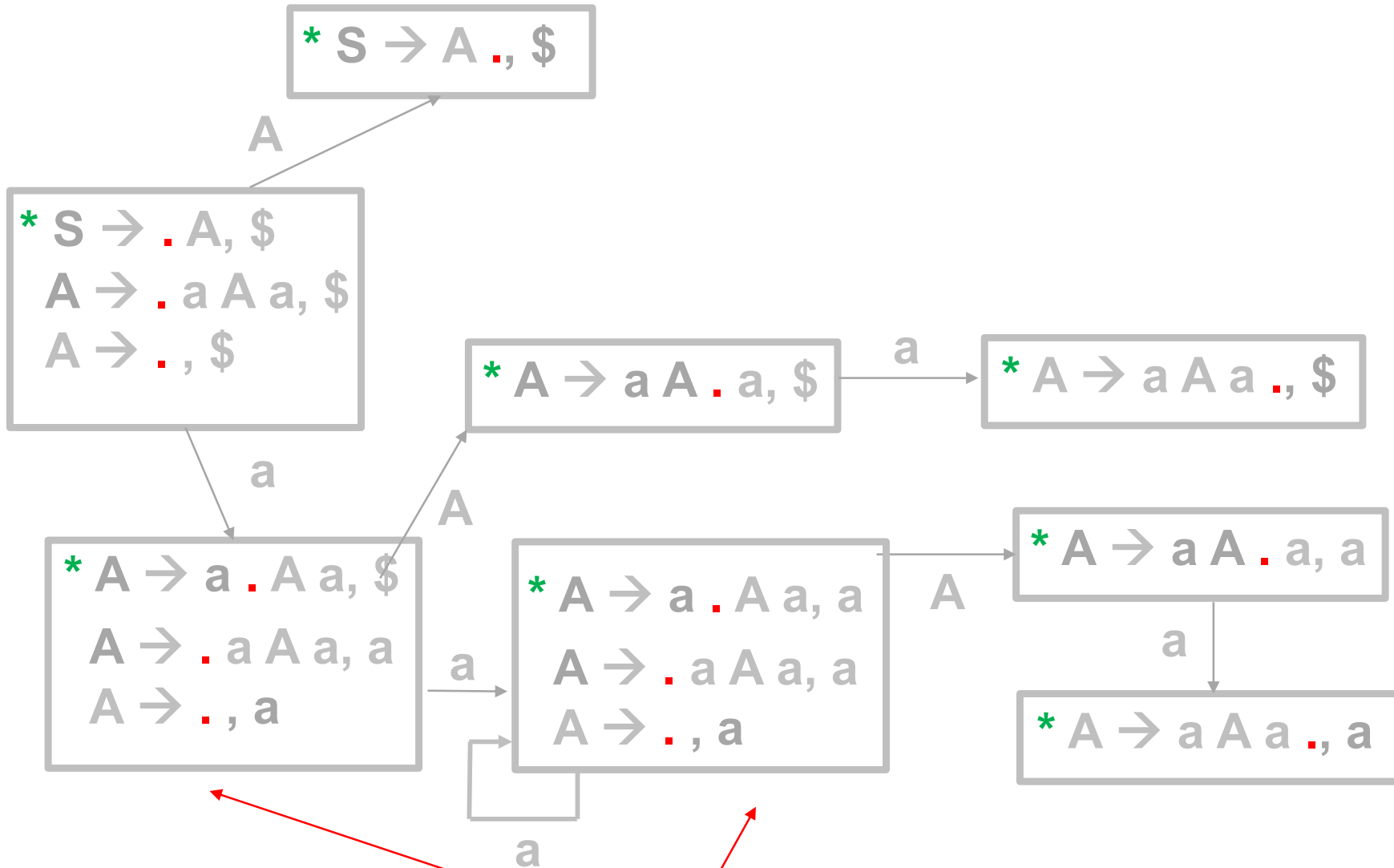
→ not LALR(1)

SAMPLE PROBLEMS

$S \rightarrow A$
 $A \rightarrow a A a \mid \epsilon$

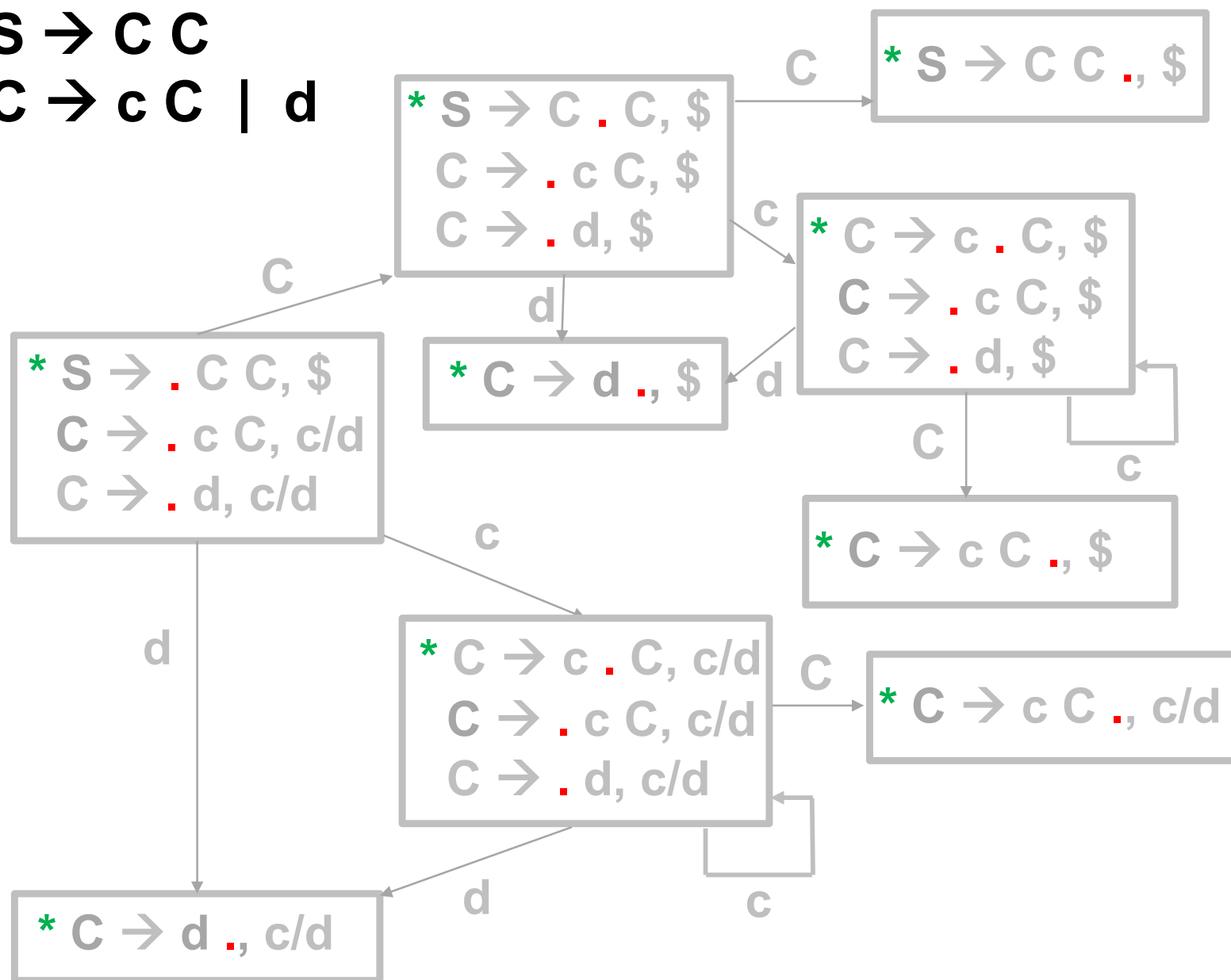


$S \rightarrow A$
 $A \rightarrow a A a \mid \epsilon$

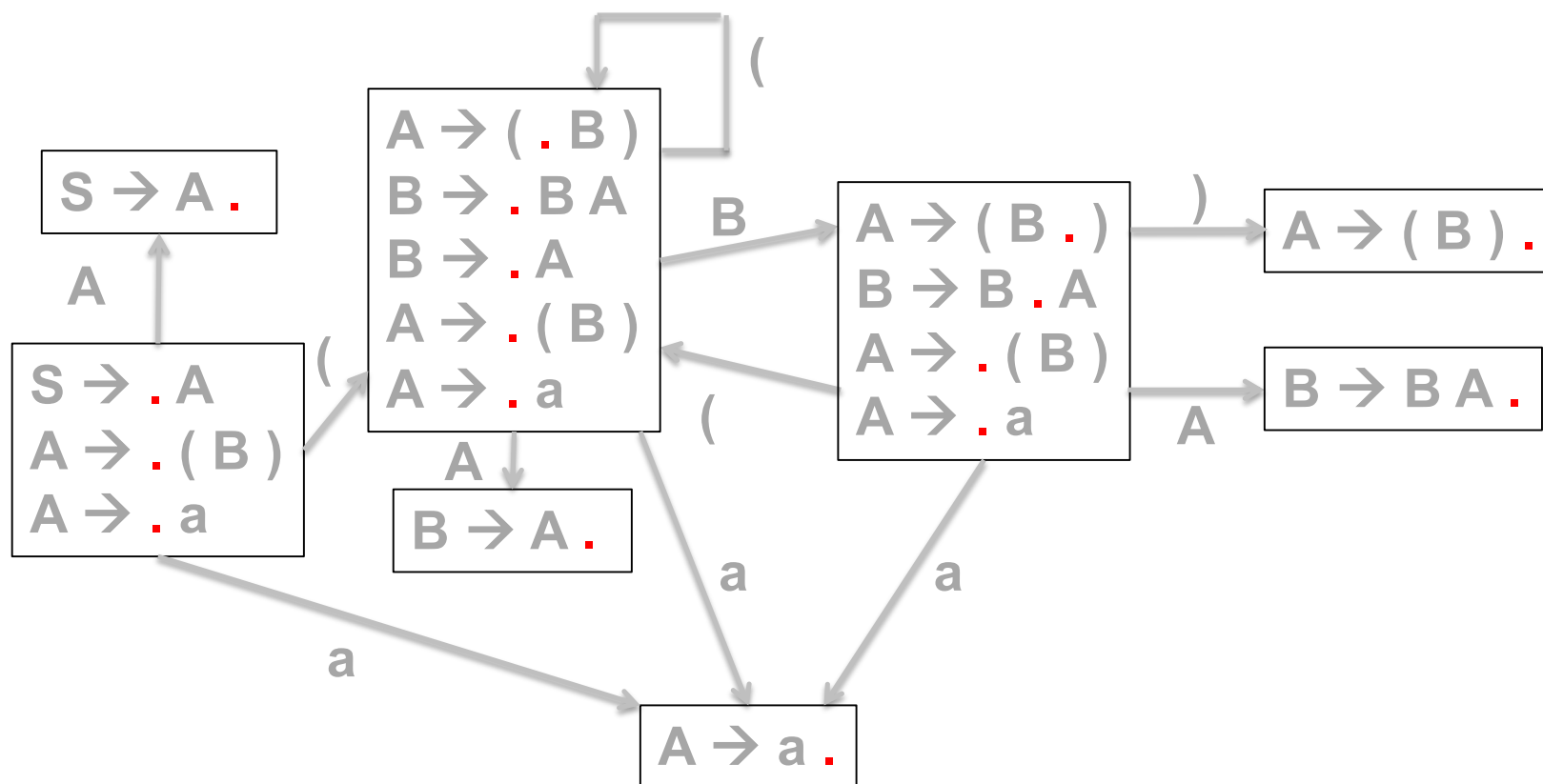


Shift-Reduce Conflicts
when input token is **a**

$S \rightarrow C C$
 $C \rightarrow c C \mid d$

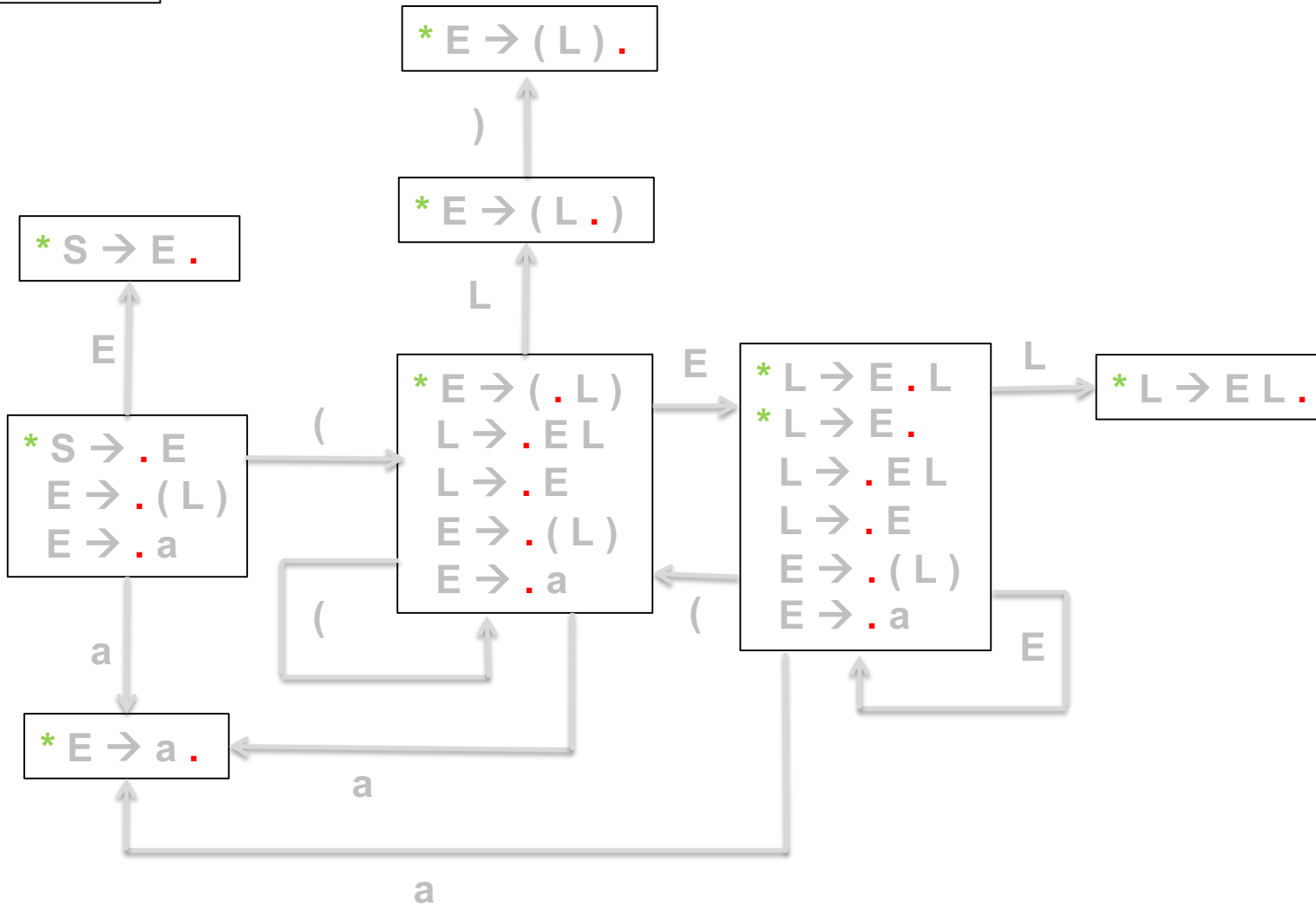


$S \rightarrow A$
 $A \rightarrow (B) \mid a$
 $B \rightarrow BA \mid A$



No Conflicts!

$S \rightarrow E$
 $E \rightarrow (L) | a$
 $L \rightarrow EL | E$



$S \rightarrow E$
 $E \rightarrow (L) | a$
 $L \rightarrow EL | E$

