Compiler Construction

Lecture 8: Syntax Analysis IV ($LR(k)$ Grammars)

Summer Semester 2016

Thomas Noll
Software Modeling and Verification Group
RWTH Aachen University

https://moves.rwth-aachen.de/teaching/ss-16/cc/

Software Modeling and Verification Chair | RWTH AACHEN UNIVERSITY

Eine Ladung Rückenwind*

***Rückenwind [ˈʀʏkn̩ˌvɪnt]:** Hilfe oder Beistand in einer schwierigen Situation oder
das Engagement der Förderer für ein erfolgreiches Studium nutzen.

Du studierst an der RWTH Aachen?
Dann bewirb Dich jetzt um ein Stipendium des RWTH Bildungsfonds!

Online-Bewerbung vom 01.06. bis 30.6.2016
www.rwth-aachen.de/bildungsfonds

Noch Fragen?
Wir helfen gerne weiter unter
bildungsfonds@rwth-aachen.de
oder Tel.: +49 241 80-90839

RWTH AACHEN UNIVERSITY

Shaping Business

Einladung zum itestra Studententag

- **Von den Bytes bis zum Topmanagement** – Wie itestra Unternehmen verändert
- **Shaping Business** – Gestaltung einer webbasierten Anwendung zur Optimierung der Vertriebswege mit AngularJS
- Ausklang des Tages bei einem gemeinsamen Essen

Freitag, **10.06.2016**, 15.00-19.00 Uhr

itestra GmbH, Hansaring 20, 50670 Köln

Die Fahrtkosten in Höhe einer Bahnfahrt 2. Klasse trägt die itestra GmbH.

Melden Sie sich bis zum **03.06.2016** an.
Seien Sie mit dabei!

Anne-Kristin Hauk – **hauk@itestra.de**

Softwarequalität– Praxis-Workshop

itestra
Software Productivity

Bringen Sie Informatik zur Wirkung!

Zuverlässig funktionierende Software, niedrige Kosten für Entwicklung und Betrieb oder schnelle Umsetzung neuer Anforderungen – Softwarequalität liefert einen wichtigen Beitrag zum Erfolg von Unternehmen. Dennoch existiert bis heute kein einheitliches Qualitätsverständnis für Software. Unterschiedlichste Verfahren und Kriterien werden eingesetzt.

Inhalte des Workshops:

- Wir stellen Ihnen im Workshop reale Beispiele aus der Praxis vor und zeigen, was neben Software-Tests und Code-Metriken wichtige Ansätze zur Qualitätssteigerung sind.
- Im Team erarbeiten Sie anschließend ein eigenes Software-Qualitätsmodell für Ihre Qualitätsziele und bewerten auf dieser Basis reale Softwaresysteme aus der Industrie. Unsere erfahrenen Kollegen unterstützen Sie dabei.

Dienstag, **28.06.2016**, 09.00–16.00 Uhr

Raum 9222, Geb. E3, 2. OG, Informatik-Zentrum, RWTH Aachen (Ahornstr. 55)

Melden Sie sich unter Angabe Ihres Semesters bis zum **19.06.2016** an.
Wir freuen uns auf Sie!

Kontakt: Anne-Kristin Hauk (**hauk@itestra.de**) – www.itestra.com

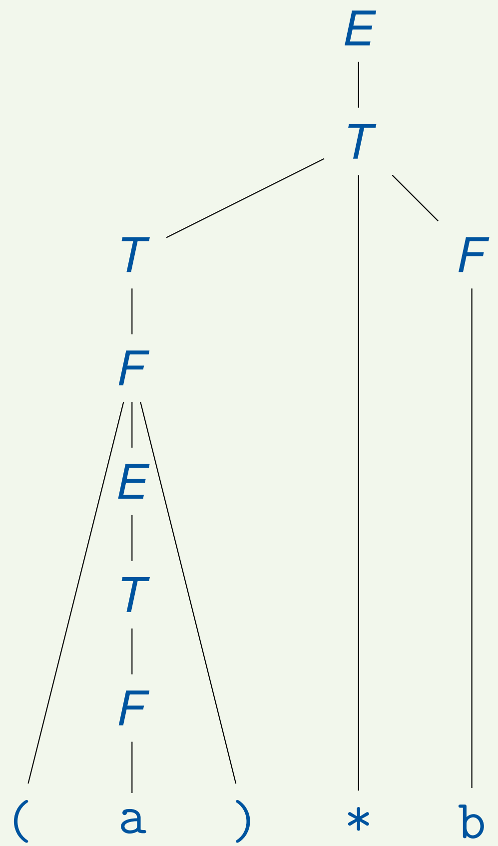
Recap: Top-Down Parsing

Example 8.1

Grammar for arithmetic expressions:

$$G_{AE}: \begin{array}{lll} E \rightarrow E{+}T \mid T & & (1,2) \\ T \rightarrow T{*}F \mid F & & (3,4) \\ F \rightarrow (E) \mid \mathtt{a} \mid \mathtt{b} & & (5,6,7) \end{array}$$

Leftmost analysis of `(a)*b`:
2 3 4 5 2 4 6 7



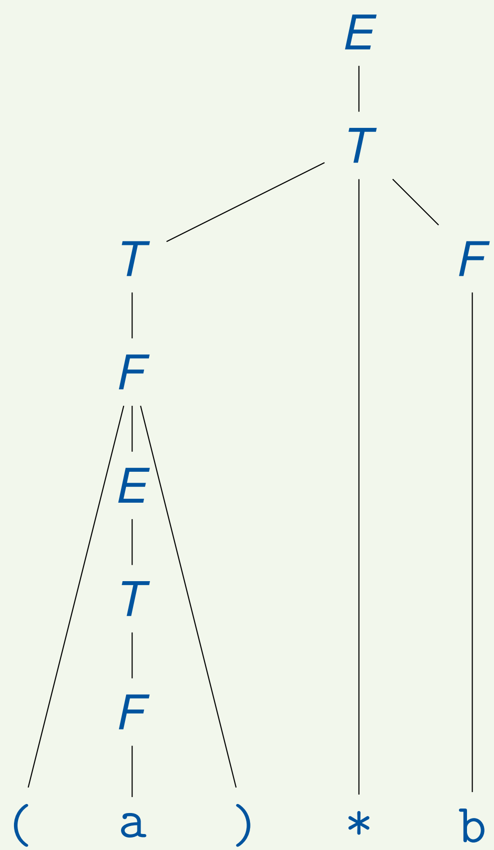
Bottom-Up Parsing I

Example 8.2

Grammar for
arithmetic expressions:

$$
\begin{aligned}
G_{AE}:\ & E \rightarrow E{+}T \mid T && (1,2)\\
& T \rightarrow T{*}F \mid F && (3,4)\\
& F \rightarrow (E) \mid \mathtt{a} \mid \mathtt{b} && (5,6,7)
\end{aligned}
$$

Reversed rightmost analysis
of `(a)*b`:
6 4 2 5 4 7 3 2



Bottom-Up Parsing II

Approach:

1. Given $G \in CFG_\Sigma$, construct a nondeterministic bottom-up parsing automaton (NBA) which accepts $L(G)$ and which additionally computes corresponding (reversed) rightmost analyses
   - input alphabet: Σ
   - pushdown alphabet: X
   - output alphabet: $[p]$ (where $p := |P|$)
   - state set: omitted
   - transitions:

     shift: shifting input symbols onto the pushdown

     reduce: replacing the right-hand side of a production by its left-hand side (= inverse expansion step)
2. Remove nondeterminism by allowing lookahead on the input:
   $G \in LR(k)$ iff $L(G)$ recognisable by deterministic bottom-up parsing automaton with lookahead of k symbols

Nondeterministic Bottom-Up Automaton I

Definition 8.3 (Nondeterministic bottom-up parsing automaton)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$. The nondeterministic bottom-up parsing automaton of G, $NBA(G)$, is defined by the following components.

- Input alphabet: Σ
- Pushdown alphabet: X
- Output alphabet: $[p]$
- Configurations: $\Sigma^* \times X^* \times [p]^*$ (top of pushdown to the right)
- Transitions for $w \in \Sigma^*$, $\alpha \in X^*$, and $z \in [p]^*$:
  shifting steps: $(aw, \alpha, z) \vdash (w, \alpha a, z)$ if $a \in \Sigma$
  reduction steps: $(w, \alpha\beta, z) \vdash (w, \alpha A, zi)$ if $\pi_i = A \to \beta$
- Initial configuration for $w \in \Sigma^*$: $(w, \varepsilon, \varepsilon)$
- Final configurations: $\{\varepsilon\} \times \{S\} \times [p]^*$

Nondeterministic Bottom-Up Automaton II

Example 8.4

Grammar for arithmetic expressions (cf. Example 8.2):

$$G_{AE}: E \to E{+}T \mid T \quad (1,2)$$
$$T \to T{*}F \mid F \quad (3,4)$$
$$F \to (E) \mid a \mid b \quad (5,6,7)$$

Bottom-up parsing of (a)*b:

$$
\begin{array}{ll}
 & ((a)*b, \varepsilon, \varepsilon) \\
\vdash & (a)*b, (, \varepsilon) \\
\vdash & ()*b, (a, \varepsilon) \\
\vdash & ()*b, (F, 6) \\
\vdash & ()*b, (T, 64) \\
\vdash & ()*b, (E, 642) \\
\vdash & (*b, (E), 642) \\
\vdash & (*b, F, 6425) \\
\vdash & (*b, T, 64254) \\
\vdash & (b, T*, 64254) \\
\vdash & (\varepsilon, T*b, 64254) \\
\vdash & (\varepsilon, T*F, 642547) \\
\vdash & (\varepsilon, T, 6425473) \\
\vdash & (\varepsilon, E, 64254732)
\end{array}
$$

Correctness of $NBA(G)$

Theorem 8.5 (Correctness of $NBA(G)$)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ and $NBA(G)$ as before. Then, for every $w \in \Sigma^*$ and $z \in [p]^*$,

$$(w, \varepsilon, \varepsilon) \vdash^* (\varepsilon, S, z) \quad \textit{iff} \quad \overleftarrow{z} \textit{ is a rightmost analysis of } w$$

Proof.

similar to the top-down case (Theorem 6.1) □

Nondeterministic Bottom-Up Parsing

Nondeterminisn in $NBA(G)$

Observation: $NBA(G)$ is generally nondeterministic

- Shift or reduce? Example:

$$(bw, \alpha a, z) \vdash \begin{cases} (w, \alpha ab, z) \\ (bw, \alpha A, zi) \end{cases} \text{if } \pi_i = A \to a$$

- If reduce: which "handle" β? Example:

$$(w, \alpha ab, z) \vdash \begin{cases} (w, \alpha A, zi) \\ (w, \alpha aB, zj) \end{cases} \text{if } \pi_i = A \to ab \text{ and } \pi_j = B \to b$$

- If reduce β: which left-hand side A? Example:

$$(w, \alpha a, z) \vdash \begin{cases} (w, \alpha A, zi) \\ (w, \alpha B, zj) \end{cases} \text{if } \pi_i = A \to a \text{ and } \pi_j = B \to a$$

- When to terminate parsing? Example:

$$\underbrace{(\varepsilon, S, z)}_{\text{final}} \vdash (\varepsilon, A, zi) \text{ if } \pi_i = A \to S$$

Resolving Termination Nondeterminism I

General assumption to avoid nondeterminism of last type:
every grammar is start separated

Definition 8.6 (Start separation)

A grammar $G = \langle N, \Sigma, P, S\rangle \in CFG_\Sigma$ is called start separated if S only occurs in productions of the form $S \to A$ where $A \neq S$.

Remarks:

- Start separation always possible by adding $S' \to S$ with new start symbol S'
- From now on consider only reduced grammars of this form (and let $\pi_0 := S' \to S$)

Resolving Termination Nondeterminism II

Start separation removes "When to terminate parsing?" nondeterminism:

Lemma 8.7

If $G \in CFG_\Sigma$ is start separated, then no successor of a final configuration (ε, S', z) in $NBA(G)$ is again a final configuration.
(Thus parsing should be stopped in the first final configuration.)

Proof.

- To (ε, S', z), only reductions by ε-productions can be applied:

$$(\varepsilon, S', z) \vdash (\varepsilon, S'A, zi) \quad \text{if } \pi_i = A \to \varepsilon$$

- Thereafter, only reductions by productions of the form $A_0 \to A_1 \dots A_n$ ($n \geq 0$) applicable
- Every resulting configuration is of the (non-final) form

$$(\varepsilon, S'B_1 \dots B_k, z) \quad \text{where } k \geq 1$$

□

$LR(k)$ Grammars I

Goal: resolve remaining nondeterminism of $NBA(G)$ by supporting lookahead of $k \in \mathbb{N}$ symbols on the input

$\Longrightarrow$ $LR(k)$: reading of input from left to right with k-lookahead, computing a rightmost analysis

Definition 8.8 ($LR(k)$ grammar)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ be start separated and $k \in \mathbb{N}$. Then G has the $LR(k)$ property (notation: $G \in LR(k)$) if for all rightmost derivations of the form

$$S \begin{cases} \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta w \\ \Rightarrow_r^* \gamma B x \Rightarrow_r \alpha\beta y \end{cases}$$

such that $\mathrm{first}_k(w) = \mathrm{first}_k(y)$, it follows that $\alpha = \gamma$, $A = B$, and $x = y$.

$LR(k)$ Grammars II

Remarks:

- If $G \in LR(k)$, then the reduction of $\alpha\beta w$ to $\alpha A w$ is already determined by $\mathrm{first}_k(w)$.
- Therefore $NBA(G)$ in configuration $(w, \alpha\beta, z)$ can decide to reduce and how to reduce.
- Computation of $NBA(G)$ for $S \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta w$:

$$(w'w, \varepsilon, \varepsilon) \vdash^* (w, \alpha\beta, z) \overset{\text{red } i}{\vdash} (w, \alpha A, zi) \vdash \ldots$$

  where $\pi_i = A \to \beta$
- Computation of $NBA(G)$ for $S \Rightarrow_r^* \gamma B x \Rightarrow_r \alpha\beta y$:
  - with direct reduction ($y = x$, $\alpha\beta = \gamma\delta$, $\pi_j = B \to \delta$):

$$(y'y, \varepsilon, \varepsilon) \vdash^* (y, \alpha\beta, z') = (x, \gamma\delta, z') \overset{\text{red } j}{\vdash} (x, \gamma B, z'j) \vdash \ldots$$

  - with previous shifts ($y = x'x$, $\alpha\beta x' = \gamma\delta$, $\pi_j = B \to \delta$):

$$(y'y, \varepsilon, \varepsilon) \vdash^* (y, \alpha\beta, z') = (x'x, \alpha\beta, z') \overset{\text{shift}^*}{\vdash} (x, \alpha\beta x', z') = (x, \gamma\delta, z') \overset{\text{red } j}{\vdash} (x, \gamma B, z'j) \vdash \ldots$$

$LR(0)$ Grammars

The case $k = 0$ is relevant (in contrast to $LL(0)$): here the decision is just based on the contents of the pushdown, without any lookahead.

Corollary 8.9 ($LR(0)$ grammar)

$G \in CFG_\Sigma$ has the $LR(0)$ property if for all rightmost derivations of the form

$$S \begin{cases} \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta w \\ \Rightarrow_r^* \gamma B x \Rightarrow_r \alpha\beta y \end{cases}$$

it follows that $\alpha = \gamma$, $A = B$, and $x = y$.

Goal: derive a finite information from the pushdown which suffices to resolve the nondeterminism (similar to abstraction of right context in LL parsing by fo-sets)

$LR(0)$ Grammars

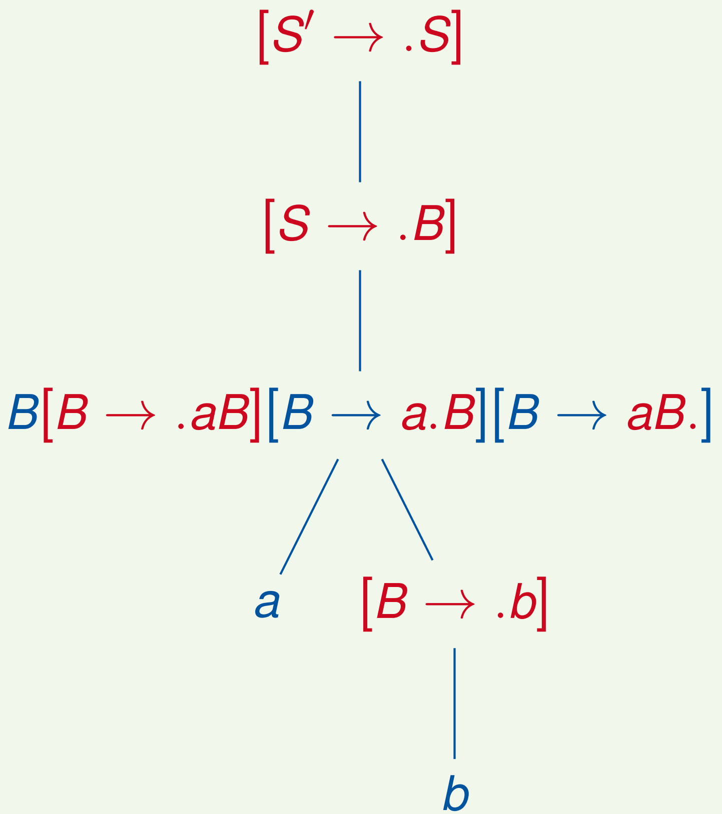
$LR(0)$ Items and Sets I

Example 8.10

$$G: \quad S' \to S \qquad (0)$$
$$S \to B \mid C \qquad (1,2)$$
$$B \to aB \mid b \qquad (3,4)$$
$$C \to aC \mid c \qquad (5,6)$$

$NBA(G)$:

$$(ab, \varepsilon, \varepsilon)$$
$$\vdash (b, a, \varepsilon)$$
$$\vdash (\varepsilon, ab, \varepsilon)$$
$$\vdash (\varepsilon, aB, 4)$$
$$\vdash (\varepsilon, B, 43)$$
$$\vdash (\varepsilon, S, 431)$$
$$\vdash (\varepsilon, S', 4310)$$



$LR(0)$ Items and Sets II

Definition 8.11 ($LR(0)$ items and sets)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ be start separated by $S' \to S$ and $S' \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta_1\beta_2 w$ (i.e., $A \to \beta_1\beta_2 \in P$).

- $[A \to \beta_1 \cdot \beta_2]$ is called an $LR(0)$ item for $\alpha\beta_1$.
- Given $\gamma \in X^*$, $LR(0)(\gamma)$ denotes the set of all $LR(0)$ items for γ, called the $LR(0)$ set (or: $LR(0)$ information) of γ.
- $LR(0)(G) := \{LR(0)(\gamma) \mid \gamma \in X^*\}$.

Corollary 8.12

1. *For every $\gamma \in X^*$, $LR(0)(\gamma)$ is finite.*
2. *$LR(0)(G)$ is finite.*
3. *The item $[A \to \beta\cdot] \in LR(0)(\gamma)$ indicates the possible reduction $(w, \alpha\beta, z) \vdash (w, \alpha A, zi)$ where $\pi_i = A \to \beta$ and $\gamma = \alpha\beta$.*
4. *The item $[A \to \beta_1 \cdot Y\beta_2] \in LR(0)(\gamma)$ indicates an incomplete handle β_1 (to be completed by shift operations or ε-reductions).*

$LR(0)$ Conflicts

Definition 8.13 ($LR(0)$ conflicts)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ and $I \in LR(0)(G)$.

- I has a shift/reduce conflict if there exist $A \to \alpha_1 a \alpha_2, B \to \beta \in P$ such that

$$[A \to \alpha_1 \cdot a\alpha_2], [B \to \beta \cdot] \in I.$$

- I has a reduce/reduce conflict if there exist $A \to \alpha, B \to \beta \in P$ with $A \neq B$ or $\alpha \neq \beta$ such that

$$[A \to \alpha \cdot], [B \to \beta \cdot] \in I.$$

Lemma 8.14

$G \in LR(0)$ iff no $I \in LR(0)(G)$ contains conflicting items.

Proof.

omitted □

Computing $LR(0)$ Sets I

Theorem 8.15 (Computing $LR(0)$ sets)

Let $G = \langle N, \Sigma, P, S\rangle \in CFG_\Sigma$ be start separated by $S' \to S$ and reduced.

1. $LR(0)(\varepsilon)$ *is the least set such that*
   - $[S' \to \cdot S] \in LR(0)(\varepsilon)$ *and*
   - *if* $[A \to \cdot B\gamma] \in LR(0)(\varepsilon)$ *and* $B \to \beta \in P$,
     then $[B \to \cdot\beta] \in LR(0)(\varepsilon)$.
2. $LR(0)(\alpha Y)$ *(*$\alpha \in X^*, Y \in X$*) is the least set such that*
   - *if* $[A \to \gamma_1 \cdot Y\gamma_2] \in LR(0)(\alpha)$,
     then $[A \to \gamma_1 Y \cdot \gamma_2] \in LR(0)(\alpha Y)$ *and*
   - *if* $[A \to \gamma_1 \cdot B\gamma_2] \in LR(0)(\alpha Y)$ *and* $B \to \beta \in P$,
     then $[B \to \cdot\beta] \in LR(0)(\alpha Y)$.

Computing $LR(0)$ Sets II

Example 8.16 (cf. Example 8.10)

$G: \; S' \to S \quad S \to B \mid C$
$\quad B \to aB \mid b \quad C \to aC \mid c$

$[S' \to \cdot S] \in$

$LR(0)(\varepsilon)$ $\dfrac{[A \to \cdot B\gamma] \in LR(0)(\varepsilon), B \to \beta \in P}{\Longrightarrow [B \to \cdot\beta] \in LR(0)(\varepsilon)}$ $\dfrac{[A \to \gamma_1 \cdot Y\gamma_2] \in LR(0)(\alpha)}{\Longrightarrow [A \to \gamma_1 Y \cdot \gamma_2] \in LR(0)(\alpha Y)}$ $\dfrac{[A \to \gamma_1 \cdot B\gamma_2] \in}{\Longrightarrow [B \to \cdot\beta] \in}$

$I_0 := LR(0)(\varepsilon): \; [S' \to \cdot S] \; [S \to \cdot B] \; [S \to \cdot C] \; [B \to \cdot aB] \; [B \to \cdot b] \; [C \to \cdot aC] \; [C \to \cdot c]$
$I_1 := LR(0)(S): \; [S' \to S\cdot]$
$I_2 := LR(0)(B): \; [S \to B\cdot]$
$I_3 := LR(0)(C): \; [S \to C\cdot]$
$I_4 := LR(0)(a): \; [B \to a \cdot B] \; [C \to a \cdot C] \; [B \to \cdot aB] \; [B \to \cdot b] \; [C \to \cdot aC] \; [C \to \cdot c]$
$I_5 := LR(0)(b): \; [B \to b\cdot]$
$I_6 := LR(0)(c): \; [C \to c\cdot]$
$I_7 := LR(0)(aB): \; [B \to aB\cdot]$
$I_8 := LR(0)(aC): \; [C \to aC\cdot]$

($LR(0)(aa) = LR(0)(a) = I_4$, $LR(0)(ab) = LR(0)(b) = I_5$, $LR(0)(ac) = LR(0)(c) = I_6$, ...,
$I_9 := LR(0)(\gamma) = \emptyset$ in all remaining cases, e.g., for $\gamma = bB$)

No conflicts $\Longrightarrow G \in LR(0)$ (but $G \notin LL(1)$)