

Compiler Construction

Lecture 6: Syntax Analysis II ($LL(k)$ Grammars)

Thomas Noll

Lehrstuhl für Informatik 2
(Software Modeling and Verification)

RWTH Aachen University

noll@cs.rwth-aachen.de

<http://www-i2.informatik.rwth-aachen.de/i2/cc12/>

Summer Semester 2012

- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead
- 4 $LL(k)$ Grammars
- 5 Follow Sets
- 6 $LL(1)$ Grammars

Approach:

- 1 Given $G \in CFG_{\Sigma}$, construct a **nondeterministic pushdown automaton** (PDA) which accepts $L(G)$ and which additionally computes corresponding leftmost derivations (similar to the proof of " $L(CFG_{\Sigma}) \subseteq L(PDA_{\Sigma})$ ")
 - input alphabet: Σ
 - pushdown alphabet: X
 - output alphabet: $[p]$
 - state set: not required
- 2 **Remove nondeterminism** by allowing **lookahead** on the input:
 $G \in LL(k)$ iff $L(G)$ recognizable by deterministic PDA with lookahead of k symbols

The Nondeterministic Top-Down Automaton I

Definition (Nondeterministic top-down parsing automaton)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_{\Sigma}$. The **nondeterministic top-down parsing automaton** of G , $NTA(G)$, 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 left)
- **Transitions** for $w \in \Sigma^*$, $\alpha \in X^*$, and $z \in [p]^*$:
expansion steps: if $\pi_i = A \rightarrow \beta$, then $(w, A\alpha, z) \vdash (w, \beta\alpha, zi)$
matching steps: for every $a \in \Sigma$, $(aw, a\alpha, z) \vdash (w, \alpha, z)$
- **Initial configuration** for $w \in \Sigma^*$: (w, S, ε)
- **Final configurations:** $\{\varepsilon\} \times \{\varepsilon\} \times [p]^*$

Remark: $NTA(G)$ is nondeterministic iff G contains $A \rightarrow \beta \mid \gamma$

- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead
- 4 $LL(k)$ Grammars
- 5 Follow Sets
- 6 $LL(1)$ Grammars

Correctness of $\text{NTA}(G)$

Theorem 6.1 (Correctness of $\text{NTA}(G)$)

Let $G = \langle N, \Sigma, P, S \rangle \in \text{CFG}_\Sigma$ and $\text{NTA}(G)$ as before. Then, for every $w \in \Sigma^*$ and $z \in [p]^*$,

$(w, S, \varepsilon) \vdash^* (\varepsilon, \varepsilon, z)$ iff z is a leftmost analysis of w

Proof.

$\Rightarrow$ (soundness): see Exercise 2

$\Leftarrow$ (completeness): on the board



- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead**
- 4 $LL(k)$ Grammars
- 5 Follow Sets
- 6 $LL(1)$ Grammars

Adding Lookahead

Goal: resolve nondeterminism of $\text{NTA}(G)$ by supporting lookahead of $k \in \mathbb{N}$ symbols on the input
 $\implies$ determination of expanding A -production by next k symbols

Definition 6.2 (first_k set)

Let $G = \langle N, \Sigma, P, S \rangle \in \text{CFG}_\Sigma$, $\alpha \in X^*$, and $k \in \mathbb{N}$. Then the first_k set of α , $\text{first}_k(\alpha) \subseteq \Sigma^*$, is given by

$$\text{first}_k(\alpha) := \{v \in \Sigma^k \mid \text{ex. } w \in \Sigma^* \text{ such that } \alpha \Rightarrow^* vw\} \cup \{v \in \Sigma^{<k} \mid \alpha \Rightarrow^* v\}$$

Remark: $\text{first}_k(\alpha)$ is effectively computable

Example 6.3 (first_k set)

Let $G : S \rightarrow aSb \mid \varepsilon$.

- ❶ $\text{first}_1(ab) = \{a\} = \text{first}_2(a)$
- ❷ $\text{first}_3(S) = \{\varepsilon, ab, aab, aaa\}$
- ❸ $\text{first}_3(Sa) = \{a, aba, aab, aaa\}$

- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead
- 4 $LL(k)$ Grammars
- 5 Follow Sets
- 6 $LL(1)$ Grammars

$LL(k)$: reading of input from **L**eft to right with k -lookahead, computing a **L**eftmost analysis

Definition 6.4 ($LL(k)$ grammar)

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

$$S \Rightarrow_i^* wA\alpha \begin{cases} \Rightarrow_l w\beta\alpha \Rightarrow_i^* wx \\ \Rightarrow_l w\gamma\alpha \Rightarrow_i^* wy \end{cases}$$

such that $\beta \neq \gamma$, it follows that $\text{first}_k(x) \neq \text{first}_k(y)$ (i.e., different productions must not yield the same lookahead).

Remarks:

- If $G \in LL(k)$, then the leftmost derivation step for $wA\alpha$ in

$$S \Rightarrow_i^* wA\alpha \begin{cases} \Rightarrow_i w\beta\alpha \Rightarrow_i^* wx \\ \Rightarrow_i w\gamma\alpha \Rightarrow_i^* wy \end{cases}$$

is **determined by the next k symbols** following w .

- Corresponding **computations of NTA(G)**:

$$\begin{array}{lcl} (wx, S, \varepsilon) \vdash^* (x, A\alpha, z) & \stackrel{(*)}{\vdash} & (x, \beta\alpha, zi) \vdash^* (\varepsilon, \varepsilon, ziz') \\ (wy, S, \varepsilon) \vdash^* (y, A\alpha, z) & \stackrel{(*)}{\vdash} & (y, \gamma\alpha, zj) \vdash^* (\varepsilon, \varepsilon, zjz'') \end{array}$$

where $\pi_i = A \rightarrow \beta$ and $\pi_j = A \rightarrow \gamma$

- **Deterministic decision** in $(*)$ possible if $\text{first}_k(x) \neq \text{first}_k(y)$
- **Problem:** how to **determine the A -production** from the lookahead (potentially infinitely many derivations $\beta\alpha \Rightarrow_i^* x / \gamma\alpha \Rightarrow_i^* y$)?

Lemma 6.5 (Characterization of $LL(k)$)

$G \in LL(k)$ iff for all leftmost derivations of the form

$$S \Rightarrow_l^* wA\alpha \begin{cases} \Rightarrow_l w\beta\alpha \\ \Rightarrow_l w\gamma\alpha \end{cases}$$

such that $\beta \neq \gamma$, it follows that $\text{first}_k(\beta\alpha) \cap \text{first}_k(\gamma\alpha) = \emptyset$.

Proof.

omitted □

Remarks:

- If $G \in LL(k)$, then the A -production is **determined by the lookahead sets** $\text{first}_k(\beta\alpha)$ (for every $A \rightarrow \beta \in P$).
- **Problem:** still **infinitely many right contexts** α to be considered (if β [or γ] “too short”, i.e., $\text{first}_k(\beta\alpha) \neq \text{first}_k(\beta)$).
- **Idea:** α derives to **“everything that follows A ”**

- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead
- 4 $LL(k)$ Grammars
- 5 Follow Sets**
- 6 $LL(1)$ Grammars

Goal: determine all possible lookaheads from production alone
(by combining all possible right contexts)

Definition 6.6 (follow_k set)

Let $G = \langle N, \Sigma, P, S \rangle \in \text{CFG}_\Sigma$, $A \in N$, and $k \in \mathbb{N}$. Then the follow_k set of A , $\text{follow}_k(A) \subseteq \Sigma^*$, is given by

$\text{follow}_k(A) := \{v \in \text{first}_k(\alpha) \mid \text{ex. } w \in \Sigma^*, \alpha \in X^* \text{ such that } S \Rightarrow_l^* wA\alpha\}.$

- 1 Repetition: Nondeterministic Top-Down Parsing
- 2 Correctness of $\text{NTA}(G)$
- 3 Adding Lookahead
- 4 $LL(k)$ Grammars
- 5 Follow Sets
- 6 $LL(1)$ Grammars

Motivation:

- $k = 1$ sufficient to resolve nondeterminism in “most” practical applications
- Implementation of $LL(k)$ parsers for $k > 1$ rather involved (cf. **ANTLR** [ANother Tool for Language Recognition; formerly PCCTS] at <http://www.antlr.org/>)

Abbreviations: $\text{fi} := \text{first}_1$, $\text{fo} := \text{follow}_1$, $\Sigma_\epsilon := \Sigma \cup \{\epsilon\}$

Corollary 6.7

- ① For every $\alpha \in X^*$,

$$\text{fi}(\alpha) = \{a \in \Sigma \mid \text{ex. } w \in \Sigma^* : \alpha \Rightarrow^* aw\} \cup \{\epsilon \mid \alpha \Rightarrow^* \epsilon\} \subseteq \Sigma_\epsilon$$

- ② For every $A \in N$,

$$\text{fo}(A) = \{x \in \text{fi}(\alpha) \mid \text{ex. } w \in \Sigma^*, \alpha \in X^* : S \Rightarrow_j^* wA\alpha\} \subseteq \Sigma_\epsilon.$$

Definition 6.8 (Lookahead set)

Given $\pi = A \rightarrow \beta \in P$,

$$\text{la}(\pi) := \text{fi}(\beta \cdot \text{fo}(A)) \subseteq \Sigma_\varepsilon$$

is called the **lookahead set** of π (where $\text{fi}(\Gamma) := \bigcup_{\gamma \in \Gamma} \text{fi}(\gamma)$).

Corollary 6.9

❶ For all $a \in \Sigma$,

$$a \in \text{la}(A \rightarrow \beta) \text{ iff } a \in \text{fi}(\beta) \text{ or } (\beta \Rightarrow^* \varepsilon \text{ and } a \in \text{fo}(A))$$

❷ $\varepsilon \in \text{la}(A \rightarrow \beta)$ iff $\beta \Rightarrow^* \varepsilon$ and $\varepsilon \in \text{fo}(A)$