Syntactic Analysis

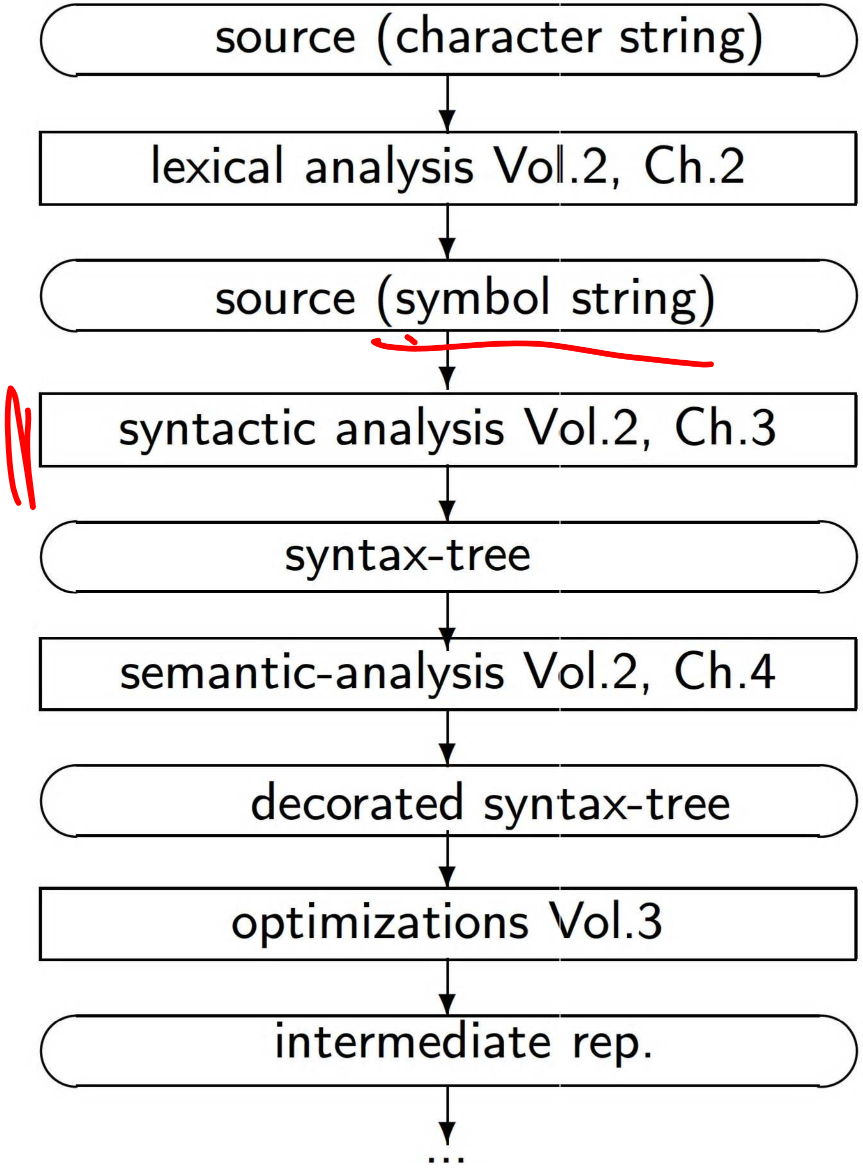
Reinhard Wilhelm
Universität des Saarlandes
wilhelm@cs.uni-sb.de
and
Mooly Sagiv
Tel Aviv University
sagiv@math.tau.ac.il

25. Oktober 2011

Subjects

- Introduction
  - The task of syntax analysis
  - Automatic generation
  - Error handling
- Context free grammars, derivations, and parse trees
- Grammar Flow Analysis
- Pushdown automata
- Top-down syntax analysis
- Bottom-up syntax analysis
- Bison — A parser generator

“Standard” Structure



source (character string)

lexical analysis Vol.2, Ch.2 — finite automata

source (symbol string)

syntactic analysis Vol.2, Ch.3 — pushdown automata

syntax-tree

semantic-analysis Vol.2, Ch.4 — attribute grammar evaluators

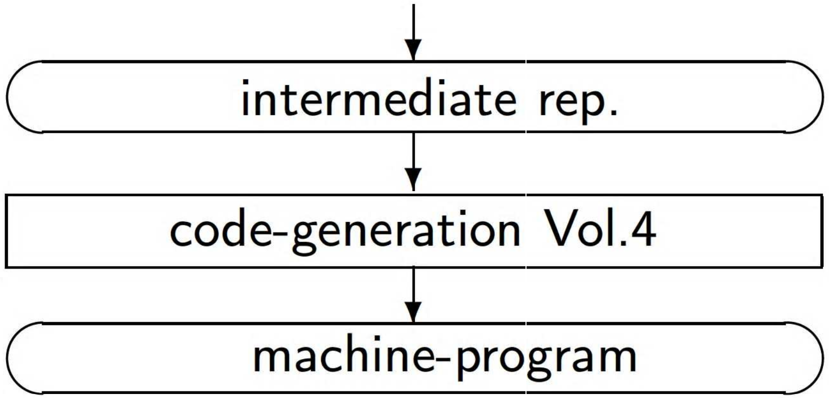
decorated syntax-tree

optimizations Vol.3 — abstract interpretation + transformations

intermediate rep.

…

“Standard” Structure cont’d



tree automata + dynamic programming + $\cdots$

Syntax Analysis (Parsing)

- Functionality

Input	Sequence of symbols (tokens)
Output	Parse tree

- Report syntax errors, e,g., unbalanced parentheses
- Create "'pretty-printed" version of the program (sometimes)
- In many cases the tree need not be generated (one-pass compilers)

Note: Input is considered as a word over a new (finite) alphabet, i.e. the set of all symbol classes.

Handling Syntax Errors

- Report and locate the error (symptom)
- Diagnose the error ✓
- Correct the error
- Recover from the error in order to discover more errors (without reporting too many follow up errors)

Example

$$a := a * \cancel{(}b + c) * \underset{\uparrow}{d};$$

The Valid Prefix Property

- For every word u that the parser identifies as a legal prefix, there exists a word w such that uw is a valid program — u has a continuation w
- Property of a parsing method
- All the parsing methods treated, i.e. LL-parsing and LR-parsing, have the valid prefix property.

Error Diagnosis Data

- Line number (may be far from the actual error)
- The current symbol
- The symbols expected in the current parser state
- Parser configuration

Error Recovery

- Becomes less important in interactive environments
- Example heuristics:
  - Search for a “significant” symbol and ignore the string up to this symbol (*panic mode*)
  - Try to “replace” symbols for common errors
  - Refrain from reporting more than 3 subsequent errors
- Globally optimal solutions — For every illegal input w, find a legal input w' with a “minimal distance” from w

Example Context Free Grammar (Section)

Stat	→	If_Stat | While_Stat | Repeat_Stat | Proc_Call | Assignment

If_Stat	→	**if** Cond **then** Stat_Seq **else** Stat_Seq **fi** |
		if Cond **then** Stat_Seq **fi**

While_Stat	→	**while** Cond **do** Stat_Seq **od**
Repeat_Stat	→	**repeat** Stat_Seq **until** Cond
Proc_Call	→	Name **(** Expr_Seq **)**
Assignment	→	Name := Expr
Stat_Seq	→	Stat | Stat_Seq; Stat
Expr_Seq	→	Expr | Expr_Seq, Expr

expand →

reduce ←

Context-Free-Grammar Definition

A context-free-grammar is a quadruple $G = (V_N, V_T, P, S)$ where:

- V_N — finite set of nonterminals
- V_T — finite set of terminals
- $P \subseteq V_N \times (V_N \cup V_T)^*$ — finite set of production rules
- $S \in V_n$ — the start nonterminal

~~$X \to$~~

(X, α)

$X \to \alpha$

Examples

$$
\begin{array}{l}
G_0 = (\{\underline{E, T, F}\}, \{+, *, (,), \mathbf{id}\}, \\
\quad \{ \begin{array}{lll} E & \rightarrow & E + T \mid T \\ T & \rightarrow & T * F \mid F \quad E) \\ F & \rightarrow & (E) \mid \mathbf{id}\}, \end{array} \\
G_1 = (\{\underline{E}\}, \{+, *, (,), \mathbf{id}\}, \underline{\{E \rightarrow E + E \mid E * E \mid (E) \mid \mathbf{id}\}}, E)
\end{array}
$$

Derivations

A context-free-grammar $G = (V_N, V_T, P, S)$

- $\varphi \Longrightarrow \psi$

  if there exist $\varphi_1, \varphi_2 \in (V_N \cup V_T)^*$, $A \in V_N$

  - $\varphi \equiv \varphi_1 \, A \, \varphi_2$
  - $\boxed{A \rightarrow \alpha \in P}$
  - $\psi \equiv \varphi_1 \, \alpha \, \varphi_2$

$\varphi_1 A \varphi_2 \Longrightarrow \varphi_1 \alpha \varphi_2$

- $\varphi \overset{*}{\Longrightarrow} \psi$ reflexive transitive closure

$\gamma_1, \ldots, \gamma_n$

$\gamma_i \Rightarrow \gamma_{i+1}$

- The language defined by G

$$L(G) = \{w \in V_T^* \mid S \overset{*}{\Longrightarrow} w\}$$

$\varphi = \gamma_1, \psi = \gamma_n$

Reduced and Extended Context Free Grammars

A nonterminal A is

- reachable: There exist φ_1, φ_2 such that $S \overset{*}{\Longrightarrow} \varphi_1 A \varphi_2$
- productive: There exists $w \in V_T^*$, $A \overset{*}{\Longrightarrow} w$

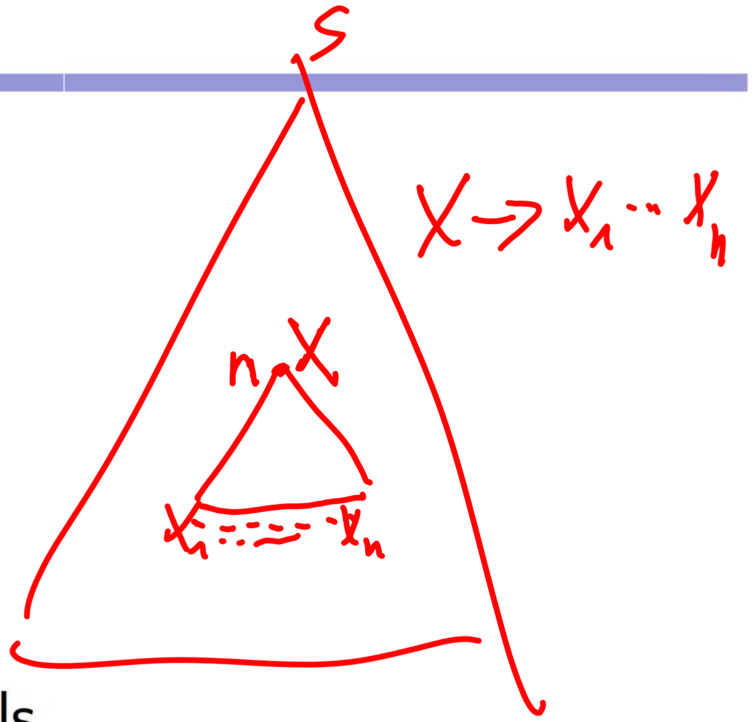
Removal of unreachable and non-productive nonterminals and the productions they occur in doesn't change the defined language.

A grammar is reduced if it has neither unreachable nor non-productive nonterminals.

A grammar is extended if a new startsymbol S' and a new production $S' \rightarrow S$ are added to the grammar.

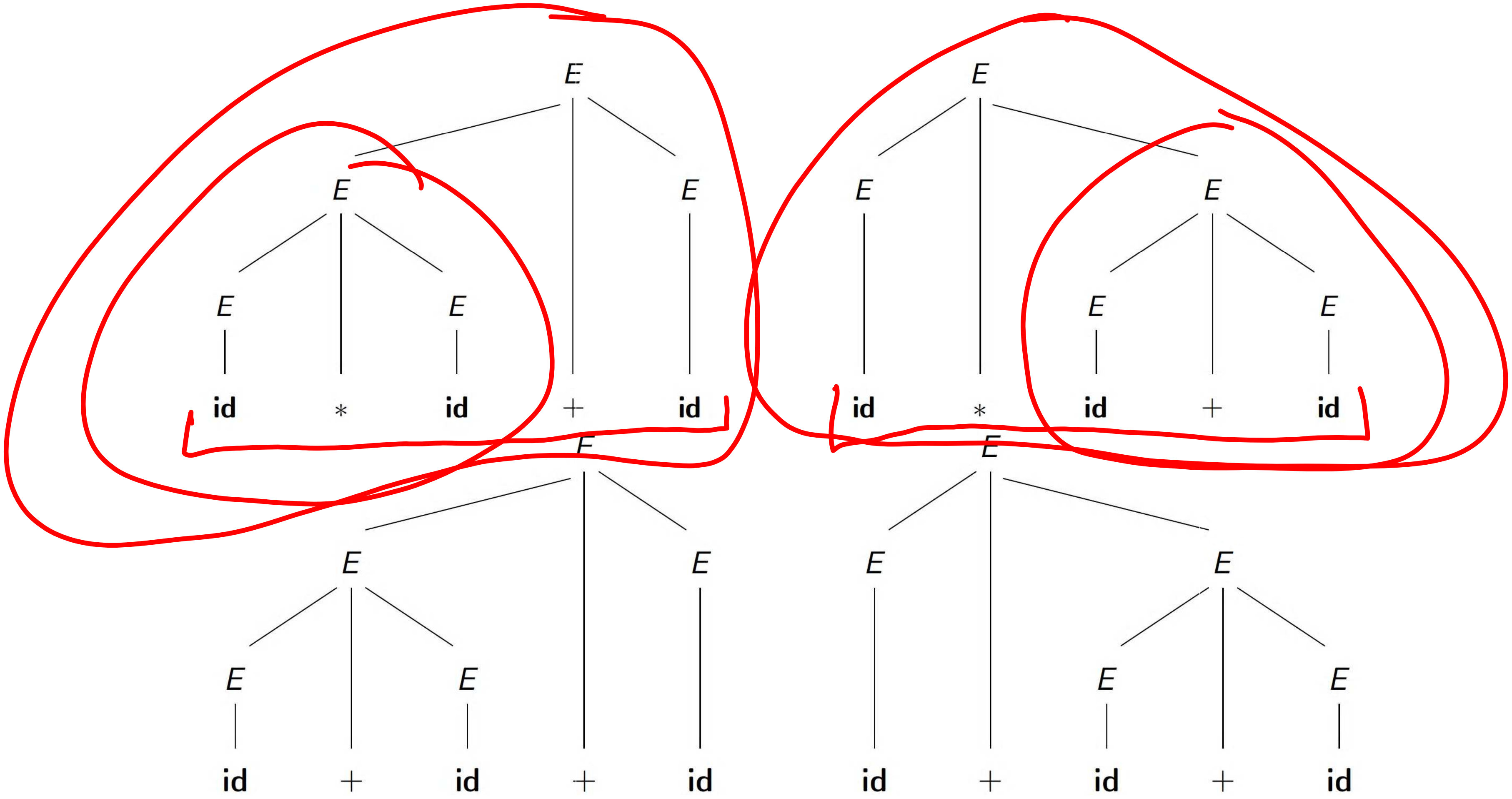
From now on, we only consider reduced and extended grammars.

Syntax-Tree (Parse-Tree)



- An ordered tree.
- Root is labeled with S.
- Internal nodes are labeled by nonterminals.
- Leaves are labeled by terminals or by ε.
- For internal nodes n: Is n labeled by N and are its children $n.1, \ldots, n.n_p$ labeled by $N_1, \ldots, N_{n_p}$, then $N \to N_1, \ldots, N_{n_p} \in P$.

Examples



Leftmost (Rightmost) Derivations

Given a context-free-grammar $G = (V_N, V_T, P, S)$

- $\varphi \underset{lm}{\Longrightarrow} \psi$ if there exist $\varphi_1 \in V_T^*$, $\varphi_2 \in (V_N \cup V_T)^*$, and $A \in V_N$
  - $\varphi \equiv \varphi_1\, A\, \varphi_2$
  - $A \rightarrow \alpha \in P$
  - $\psi \equiv \varphi_1\, \alpha\, \varphi_2$

  replace leftmost nonterminal
- $\varphi \underset{rm}{\Longrightarrow} \psi$ if there exist $\varphi_2 \in V_T^*$, $\varphi_1 \in (V_N \cup V_T)^*$, and $A \in V_N$
  - $\varphi \equiv \varphi_1\, A\, \varphi_2$
  - $A \rightarrow \alpha \in P$
  - $\psi \equiv \varphi_1\, \alpha\, \varphi_2$

  replace rightmost nonterminal
- $\varphi \overset{*}{\underset{lm}{\Longrightarrow}} \psi$, $\varphi \overset{*}{\underset{rm}{\Longrightarrow}} \psi$ are defined as usual

Ambiguous Grammar

A grammar that has (equivalently)

- two leftmost derivations for the same string,
- two rightmost derivations for the same string,
- two syntax trees for the same string.

context-free languages

non-determ., unambig.   C++

deterministically analyzable
context-free language