Chapter 12: Context-Free Grammar

I. Theory of Automata

→ II. Theory of Formal Languages

III. Theory of Turing Machines …

- programming languages
- compiling a program: an operation that generates an equivalent program in machine or assembler language.
- 2 phases:
  1. parsing ←
  2. translation to machine language

Example: AE (Arithmetic Expressions)

- Rule 1: Any number is in AE
- Rule 2: If x and y are in AE, then so are:

  (x) – (x) (x+y) (x–y) (x*y)

A different way for defining the set AE is to use a set of substitutions rules similar to the grammatical rules:

Substitution rules that define the AE's:

$S \rightarrow EA$
$EA \rightarrow (EA + EA)$
$EA \rightarrow (EA–EA)$
$EA \rightarrow (EA*EA)$
$EA \rightarrow (EA)$
$EA \rightarrow –(EA)$
$EA \rightarrow NUMBER$

NUMBERS??

NUMBER → FIRST-DIGIT
FIRST-DIGIT → FIRST-DIGIT OTHER-DIGIT
FIRST-DIGIT → 1 2 3 4 5 6 7 8 9
OTHER-DIGIT → 0 1 2 3 4 5 6 7 8 9

$S \Rightarrow EA \Rightarrow (EA*EA) \Rightarrow ((EA+EA)*EA) \Rightarrow ((EA+EA)*(EA+EA))$

$\ldots \Rightarrow ((3+4)*(6+7))$

EA:

$S \rightarrow EA$

$EA \rightarrow (EA + EA)$

$EA \rightarrow (EA - EA)$

$EA \rightarrow (EA*EA)$

$EA \rightarrow (EA)$

$EA \rightarrow -(EA)$

$EA \rightarrow NUMBER$

How to generate the number 1066?

NUMBER ⇒ FIRST-DIGIT
⇒ FIRST-DIGIT OTHER-DIGIT
⇒ FIRST-DIGIT OTHER-DIGIT OTHER-DIGIT
⇒ FIRST-DIGIT OTHER-DIGIT OTHER-DIGIT OTHER-DIGIT
⇒ 1 0 6 6

NUMBERS

NUMBER → FIRST-DIGIT
FIRST-DIGIT → FIRST-DIGIT OTHER-DIGIT
FIRST-DIGIT → 1 2 3 4 5 6 7 8 9
OTHER-DIGIT → 0 1 2 3 4 5 6 7 8 9

<u>Definition:</u> A context free grammar (CFG) is:

1. an alphabet S of letters, called terminals.
2. a set of symbols, called nonterminals or variables. One symbol S is called the start symbol.
3. a finite set of productions of the form:

$$A \to a$$

where A is a nonterminal and a is a finite sequence (word) of nonterminals and terminals.

- Examples:

Terminals: (,), +, -, *, numbers

Nonterminals: S, EA

EA:

S → EA

EA → (EA + EA)

EA → (EA–EA)

EA → (EA*EA)

EA → (EA)

EA → –(EA)

EA → NUMBER

NUMBERS

NUMBER	→ FIRST-DIGIT
FIRST-DIGIT	→ FIRST-DIGIT OTHER-DIGIT
FIRST-DIGIT	→ 1 2 3 4 5 6 7 8 9
OTHER-DIGIT	→ 0 1 2 3 4 5 6 7 8 9

Terminals : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Nonterminals : S, FIRST-DIGIT, OTHER-DIGIT

Definition: A sequence of applications of productions starting with the start symbol and ending in a sequence of terminals is called a derivation.

Definition: The language generated by a CFG is the set of all sequences of terminals produced by derivations. We also say language defined by, language derived from, or language produced by the CFG.

Definition: A language generated by a CFG is called a context-free language.

Examples: $S \rightarrow aS$ $S \rightarrow \Lambda$

$$
\begin{aligned}
S &\Rightarrow aS \\
&\Rightarrow aaS \\
&\Rightarrow aaaS \\
&\Rightarrow aaaaS \\
&\Rightarrow aaaaaS \\
&\Rightarrow aaaaaaS \\
&\Rightarrow aaaaaa\Lambda = aaaaaa
\end{aligned}
$$

Generated Language:

$\{\Lambda, a, aa, aaa, \ldots\}$ = language($\mathbf{a^*}$).

$S \to SS$ $\quad$ $S \to a$ $\quad$ $S \to \Lambda$

$$
\begin{aligned}
S &\Rightarrow SS \\
&\Rightarrow SSS \\
&\Rightarrow SaS \\
&\Rightarrow SaSS \\
&\Rightarrow \Lambda aSS \\
&\Rightarrow \Lambda aaS \\
&\Rightarrow \Lambda aa\Lambda = aa
\end{aligned}
$$

Generated Language:
$\{\Lambda, a, aa, aaa, \ldots\}$ = language(**a***).

(An infinite number of derivations for the word aa.)

<u>In general:</u> variables: Upper case letters
terminals: Lower case letters

The empty word:

is it a nonterminal? $\Lambda \rightarrow \ldots$

no

a terminal? $\Lambda aa\Lambda = aa$

not exactly, because it is erased.

$N \rightarrow \Lambda$ N can simply be deleted.

- $S \to aS \quad S \to bS \quad S \to a \quad S \to b$

  $S \Rightarrow aS \Rightarrow abS \Rightarrow abbS \Rightarrow abba$

- $S \to X \quad S \to Y \quad X \to \Lambda \quad Y \to aY$

  $Y \to bY \quad Y \to a \quad Y \to b$

- $S \to aS \quad S \to bS \quad S \to a \quad S \to b \quad S \to \Lambda$

  $S \Rightarrow aS \Rightarrow abS \Rightarrow abbS \Rightarrow abbaS \Rightarrow abba$

- $S \to aS \quad S \to bS \quad S \to \Lambda$

$S \to XaaX \quad X \to aX \quad X \to bX \quad X \to \Lambda$

$S \Rightarrow XaaX \Rightarrow aXaaX \Rightarrow abXaaX \Rightarrow abXaabX \Rightarrow abaab$

How many derivations are possible for the word baabaab?

$S \to XY$

$X \to aX$

$X \to bX$

$X \to a$

$Y \to Ya$

$Y \to Yb$

$Y \to a$

Abbreviation:

$S \to XY$

$X \to aX \mid bX \mid a$

$Y \to Ya \mid Yb \mid a$

$S \rightarrow SS$ / ES / SE / Λ /DSD

$E \rightarrow aa \mid bb$

$D \rightarrow ab \mid ba$

EVEN-EVEN=language([**aa+bb+(ab + ba)(aa+bb)*(ab+ba)]***)

$S \rightarrow aSb$ / Λ

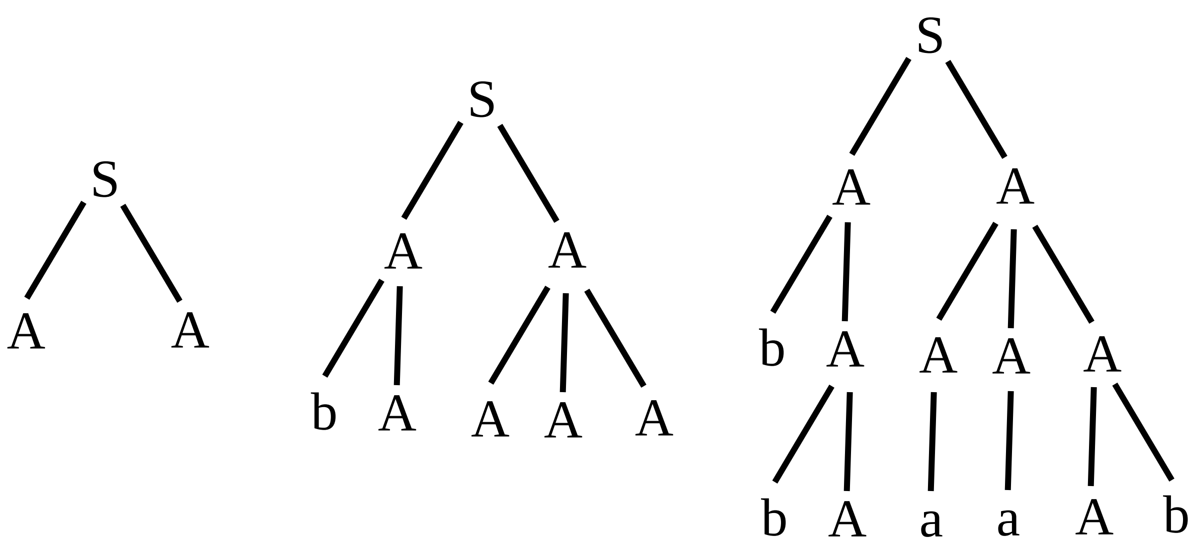
$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaaaSbbbb$$
$$\Rightarrow aaaaaSbbbbb \Rightarrow aaaaabbbbb$$

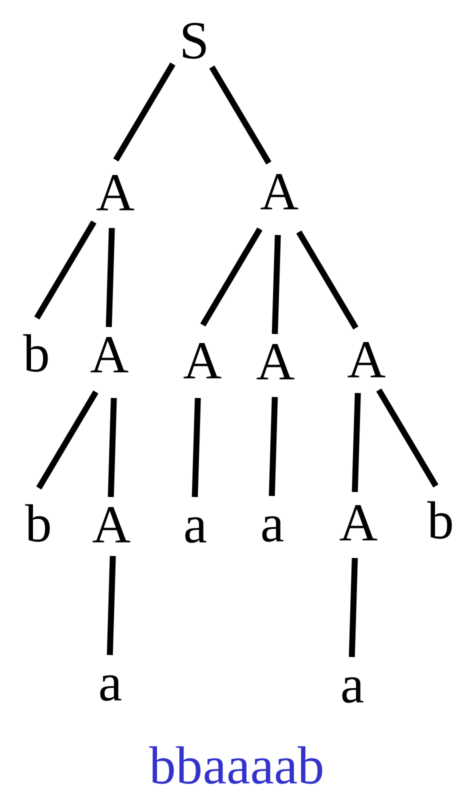
$S \rightarrow aSa$ / bSb / Λ

$S \rightarrow AA$

$A \rightarrow AAA \mid bA \mid Ab \mid a$

Parse Trees for the word bbaaaab





Parse trees are also called syntax trees,generation trees, production trees, or derivation trees.

<u>Remark:</u> In a parse tree every internal nodes is labelled with a variable (nonterminal) and every leaf is labelled with a terminal.

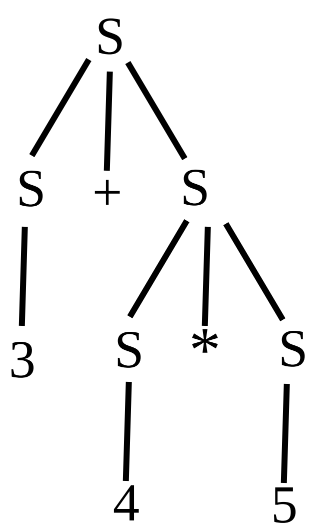
Example:

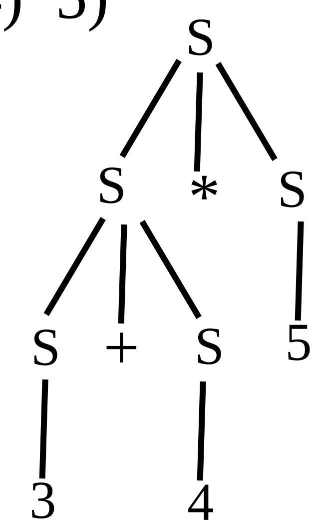
S → (S+S) | (S*S) | NUMBER

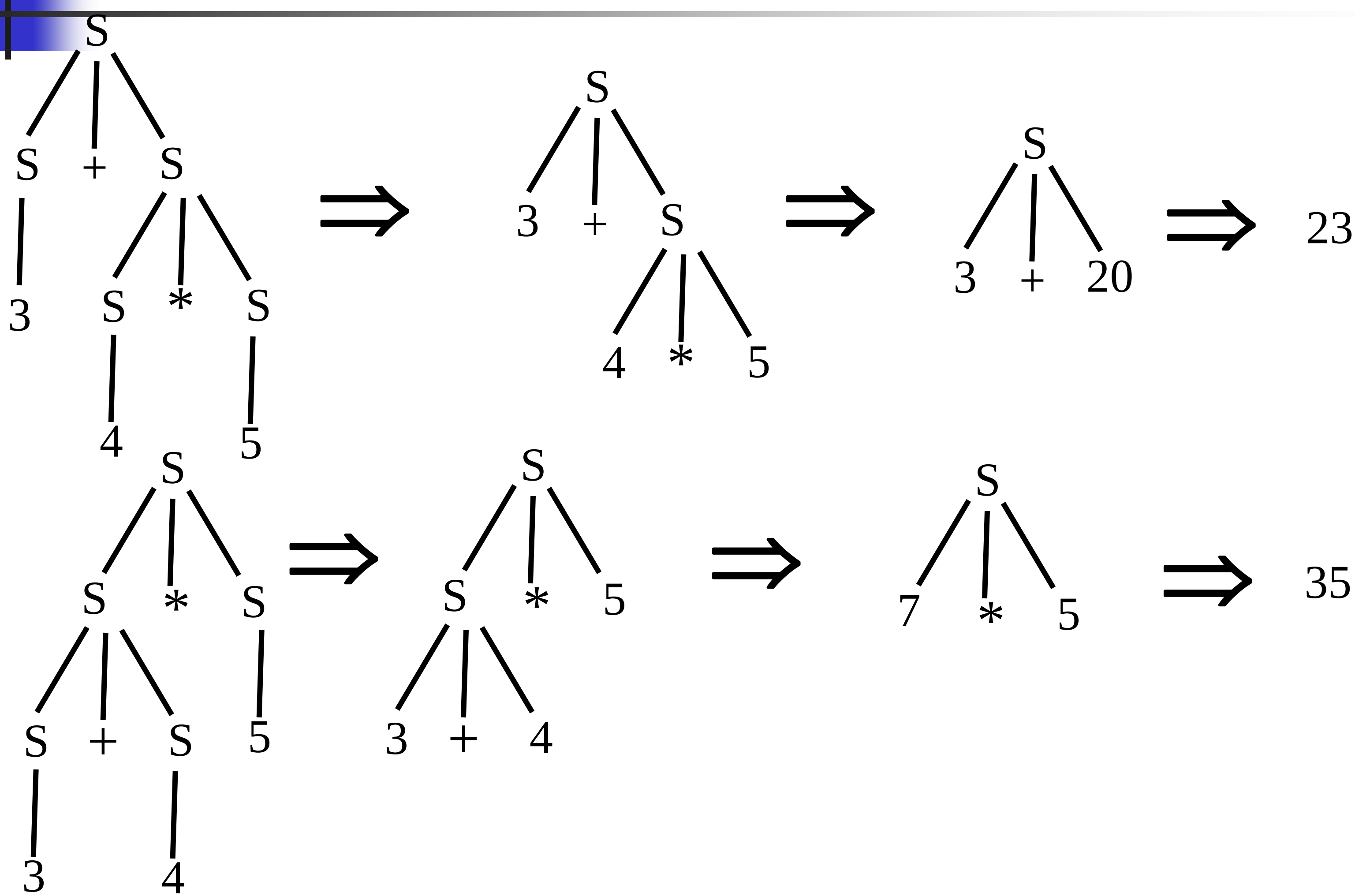
NUMBER → …

$$S \Rightarrow (S+S) \Rightarrow (S+(S*S)) \Rightarrow \ldots \Rightarrow (3+(4*5))$$

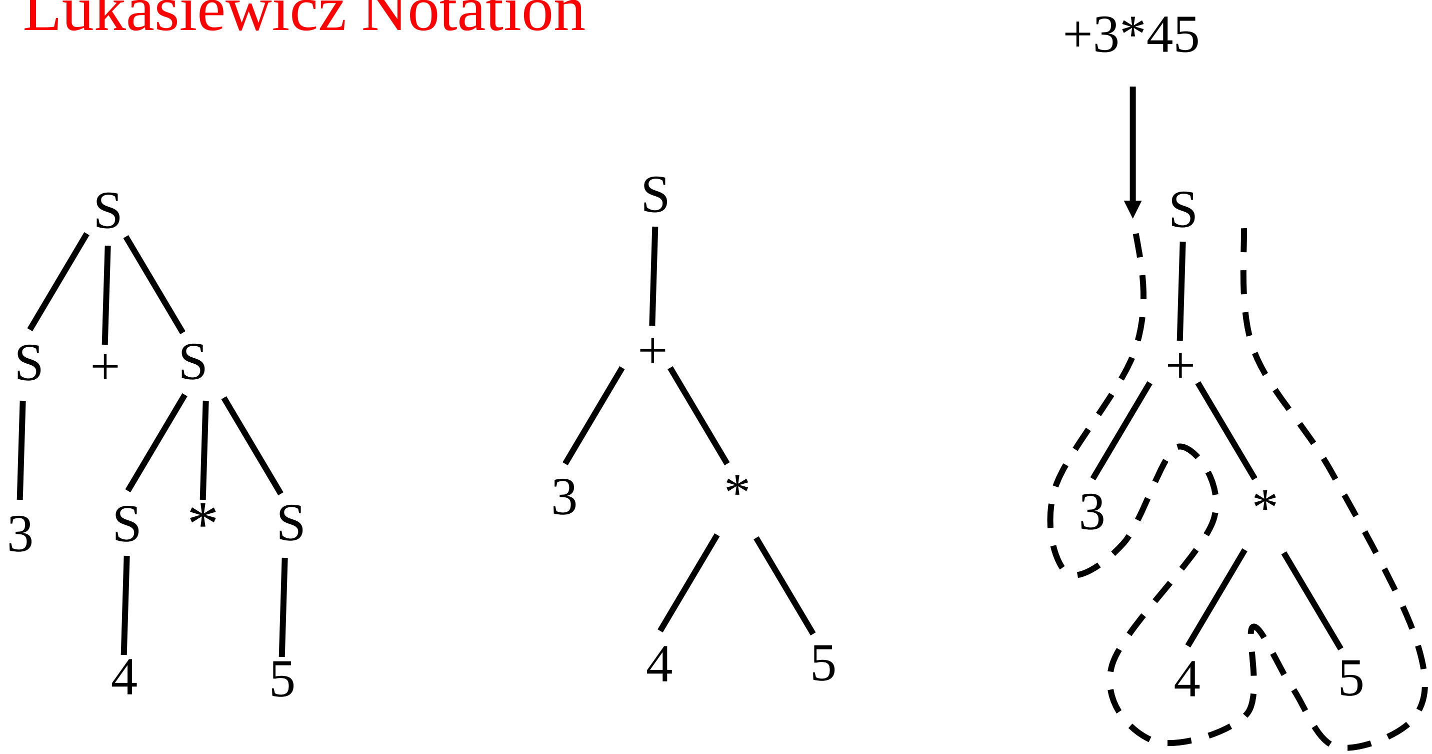
$$S \Rightarrow (S*S) \Rightarrow ((S+S)*S) \Rightarrow \ldots \Rightarrow ((3+4)*5)$$

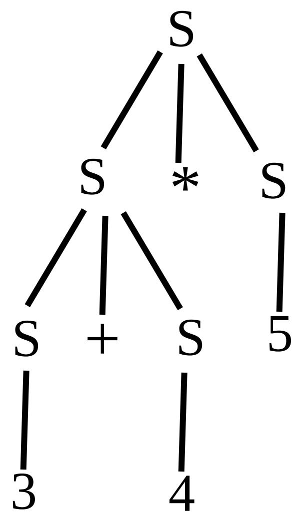


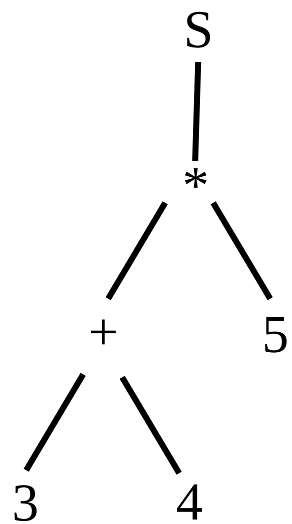


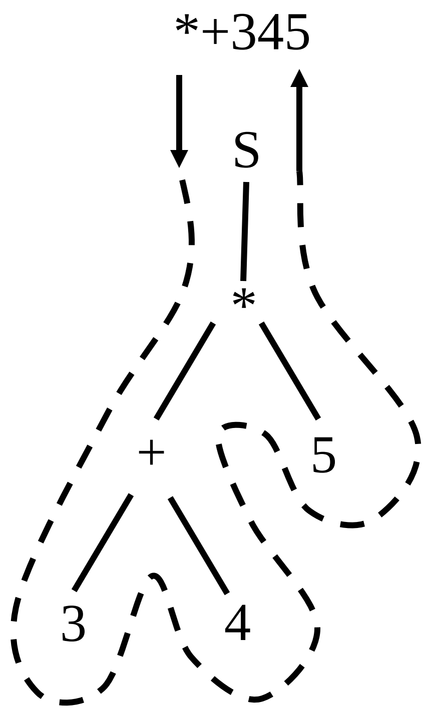

S
S + S
3 S * S
4 5
3 + S
4 * 5
3 + 20
23
S * S
S + S 5
3 4
S * 5
3 + 4
7 * 5
35

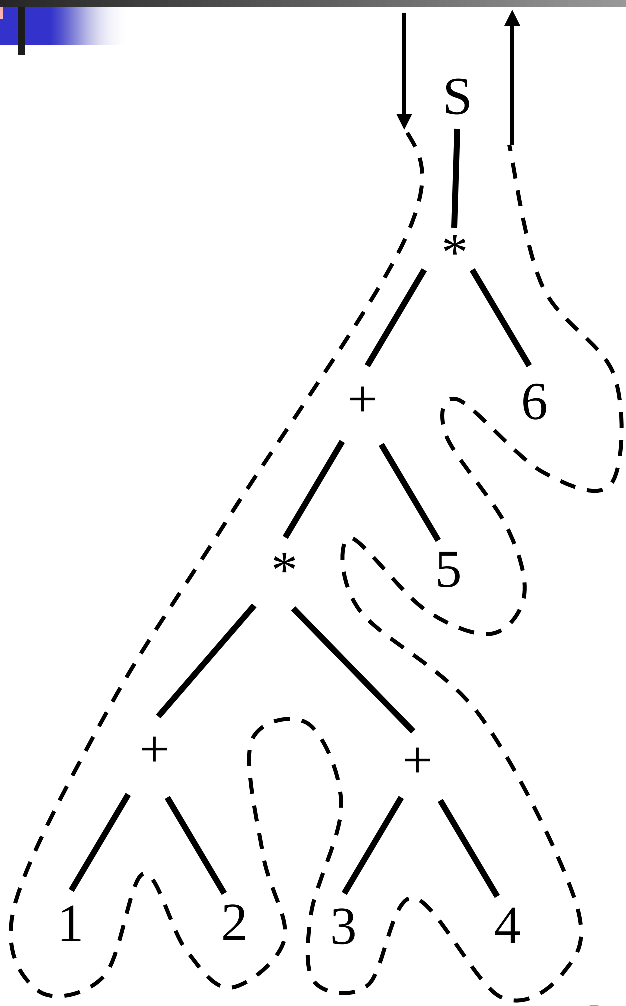
Lukasiewicz Notation




S
S
*
S
S
+
S
5
3
4


S
*
+
5
3
4


*+345
S
*
+
5
3
4

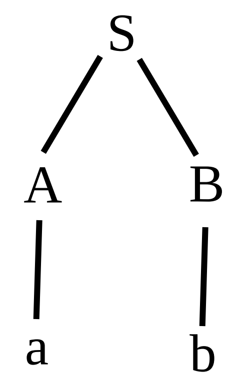


* + * + 1 2 + 3 4 5 6	+ 1 2
⇓	
* + * 3 + 3 4 5 6	+ 3 4
⇓	
* + * 3 7 5 6	* 3 7
⇓	
* + 21 5 6	+ 21 5
⇓	
* 26 6	* 26 6
⇓	
156	

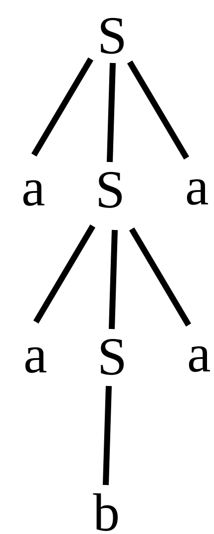
Example: $S \rightarrow AB \quad A \rightarrow a \quad B \rightarrow b$

$$S \Rightarrow AB \Rightarrow aB \Rightarrow ab$$

$$S \Rightarrow AB \Rightarrow Ab \Rightarrow ab$$



- *A CFG is ambiguous if there is at least one word in the language that has at least two derivation trees. It is called unambiguous otherwise.*



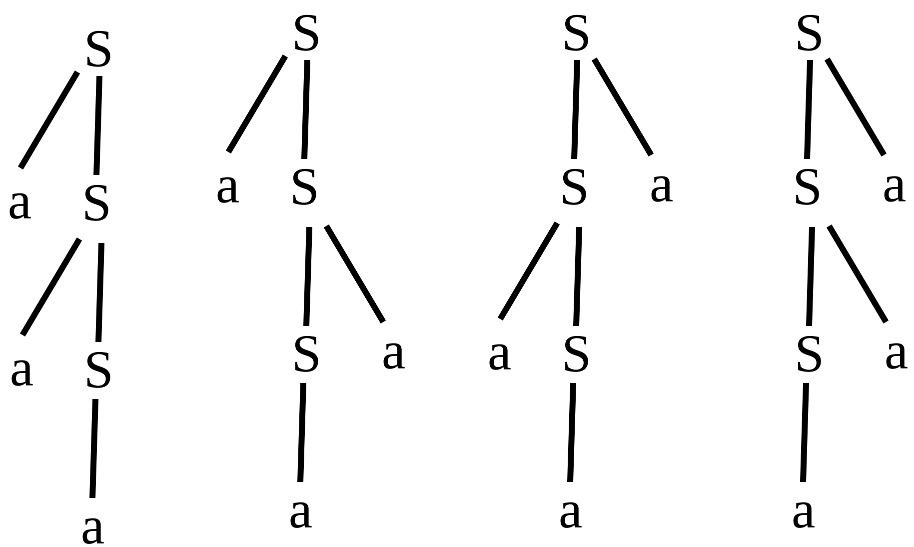
- Example: $S \rightarrow aSa \quad S \rightarrow bSb$

  $S \rightarrow a \quad S \rightarrow b \quad S \rightarrow \Lambda$

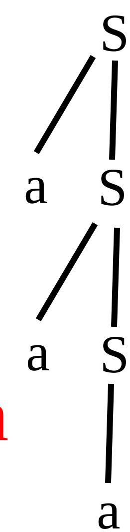
$$S \Rightarrow aSa \Rightarrow aaSaa \Rightarrow aabaa$$

Language(**a+**)

Example: $S \rightarrow aS \mid Sa \mid a$



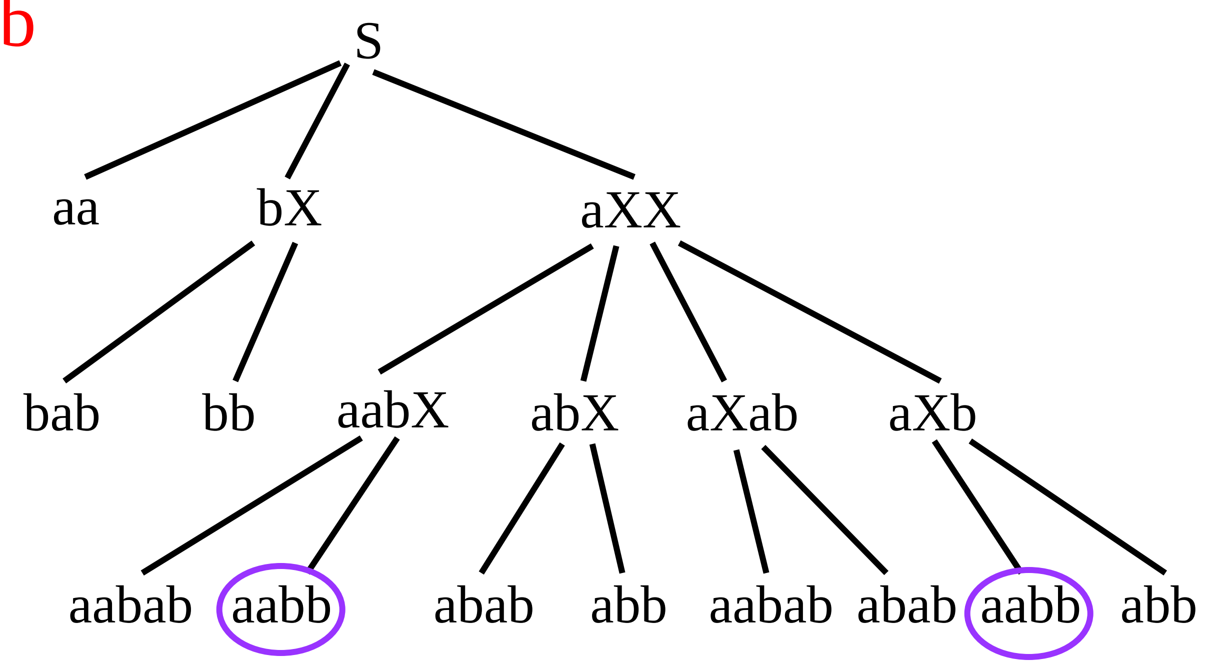
Example: $S \rightarrow aS \mid a$



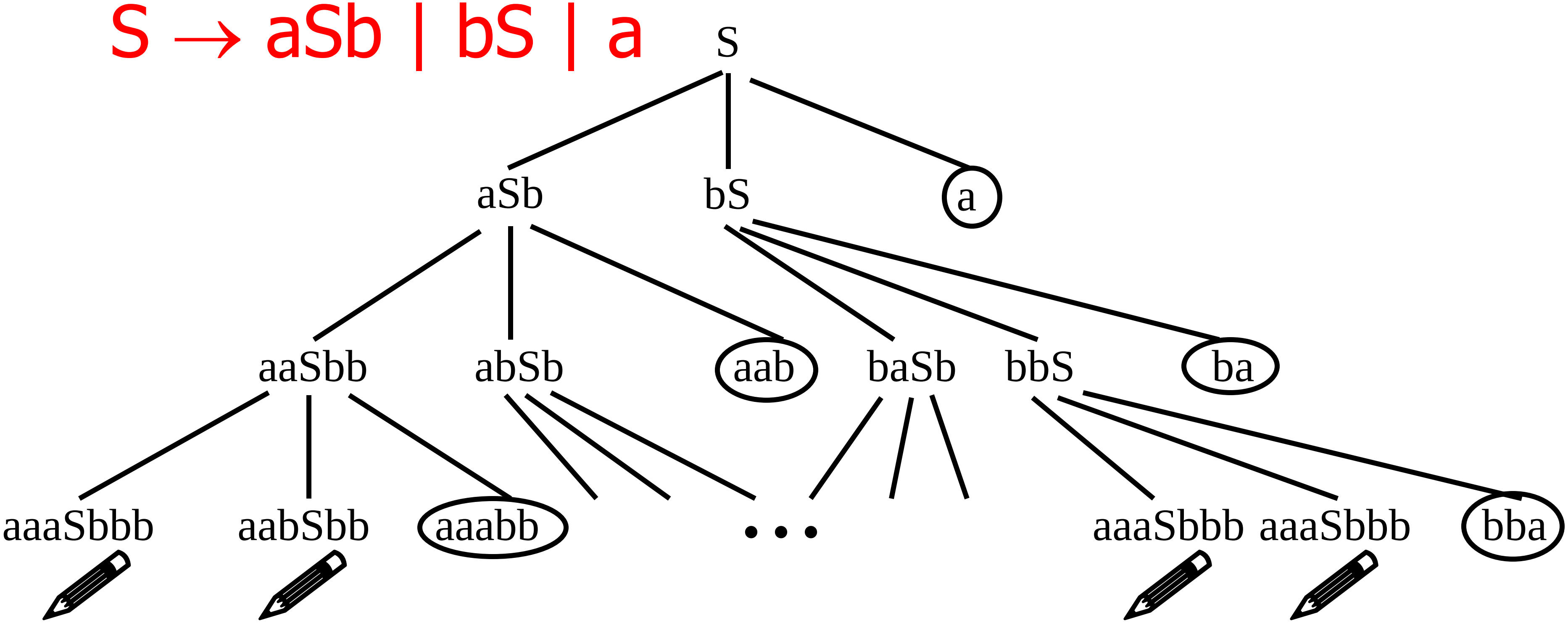
Total Language Trees

$S \rightarrow aa \mid bX \mid aXX$

$X \rightarrow ab \mid b$



S → aSb | bS | a


S
aSb
bS
a
aaSbb
abSb
aab
baSb
bbS
ba
aaaSbbb
aabSbb
aaabb
• • •
aaaSbbb
aaaSbbb
bba

$S \rightarrow X \mid b$

$X \rightarrow aX$