Derivation Trees and Regular Languages

We consider special derivations and ways to represent derivations. Then we connect CFGs and regular languages.
Strings $\alpha$ yields string $\beta$, written $\alpha \rightarrow^* \beta$, if it is possible to get from $\alpha$ to $\beta$ using the productions. A derivation of $\beta$ is the sequence of steps that gets to $\beta$. A leftmost derivation is where at each stage one replaces the leftmost variable. A rightmost derivation is defined similarly.
In a *derivation tree*, the root is the start variable, all internal nodes are labeled with variables, while all leaves are labeled with terminals. The children of an internal node are labeled from left to right with the right-hand side of the production used.
Example

Recall the CFG for equal 0’s and 1’s:

\[ S \rightarrow 0S1S \mid 1S0S \mid \varepsilon \]

The derivation for 011100 given earlier was left-most:

\[
S \Rightarrow 0S1S \Rightarrow 01S \Rightarrow 011S0S \Rightarrow 0111S0S0S \\
\Rightarrow 01110S0S \Rightarrow 011100S \Rightarrow 011100
\]
Here is derivation tree for 011100 in the above grammar
A grammar is **unambiguous** if there is a unique leftmost derivation for each string in the language. Equivalently, for each string there is a unique derivation tree.

For example, our grammar for equality is ambiguous:

\[ S \rightarrow 0S1S \mid 1S0S \mid \varepsilon \]

(The string 0101 has two derivation trees.)
Grammars and Compilers

Grammars are used in compilers. A compiler checks that the input file is valid by essentially finding the derivation tree. This derivation tree is then used in finding code for the file or expression. If grammar is unambiguous, the derivation tree is always unique.
The language of arithmetical expressions using only multiplication and addition can be described as the following CFG with start variable $E$:

\[
E \rightarrow E + T \mid T \\
T \rightarrow T \times F \mid F \\
F \rightarrow (E) \mid \text{number}
\]

(Think of Expression, Factor and Term.)

This generates expressions such as $1 + (3 + 2) \times 5$ and $1 + 2 \times 3$. 
The derivation tree for $1+2\times3$:

Note that derivation tree automatically knows that multiplication takes precedence over addition.
**Theorem.** *Every regular language is generated by a context-free grammar.*

One way to prove this is to provide an algorithm to convert an RE to a CFG.

This can be achieved recursively. For example, if overall language is union of two pieces, one can write \( S \rightarrow A \mid B \); and if the concatenation of two pieces, one can write \( S \rightarrow CD \). If overall language is the star of a piece, say generated by \( E \), then one can write \( S \rightarrow ES \mid \varepsilon \).
Consider RE $(11+00)^*11$. At top level, it is concatenation of two pieces; the first piece is $or$ of two parts. So:

\[
S \rightarrow TU \\
U \rightarrow 11 \\
T \rightarrow TV \mid \varepsilon \\
V \rightarrow 00 \mid 11
\]
Each string in the language has a leftmost derivation and a derivation tree. If these are unique for all strings, then the grammar is called unambiguous.