Proofs of equivalence in CFGs

K. Subramani
LDCSEE,
West Virginia University,
Morgantown, WV
{ksmani@csee.wvu.edu}

1 Definitions

Definition 1.1 *A Context-Free Grammar (CFG) is a 4-tuple $G = \langle V, T, P, S \rangle$, with V representing a set of non-terminals, T representing a set of terminals, P representing a set of productions or rules and S denoting a special symbol in V called the start symbol.*

Definition 1.2 *The set of all terminal strings that can be derived from the start symbol (say S) of a grammar (say G) is called the language of that grammar and is denoted by $L(G)$.*

2 Example

Consider the CFG $G = \langle V, T, P, S \rangle$, with $V = \{S\}$, $T = \{a, b\}$ and P represented by:

$$S \quad \to \quad a \cdot S \mid b \cdot S \mid \epsilon$$

Show that $L(G)$ consists of **all** and **only** those strings in $\{a, b\}^*$.

Solution: It is important to note that $\{a, b\}^*$ is a regular expression that represents the set of all possible strings over the binary alphabet $\{a, b\}$. Thus, we are being asked to prove that every string derived from S is in $\{a, b\}^*$ **and** that every string in $\{a, b\}^*$ can be derived from S. We prove the hypothesis in two parts.

Lemma 2.1 *Let x denote an arbitrary terminal string derived from S; x must belong to $\{a, b\}^*$.*

Proof: We use induction on the number of derivations used to derive x from S.
BASIS: Assume that x was derived from S in one step. In this case, x must be ϵ, since $S \to \epsilon$ is the only derivation that results in a terminal string in one step.
Inasmuch as $\epsilon \in \{a, b\}^*$, the basis is proven.
We assume that whenever x is derived from S in exactly k steps, $x \in \{a, b\}^*$. Now consider the case in which x is derived from S in $(k + 1)$ steps. The first step of the derivation is either $S \to a \cdot S$ or $S \to b \cdot S$. (Why?) Without loss of generality, assume that $S \to a \cdot S$ was the first step in the derivation. It follows that x can be broken up as $a \cdot y$, where y is derived from the S in $a \cdot S$. But the derivation of y takes exactly k steps and hence $y \in \{a, b\}^*$. It follows that x which is the concatenation of a with a string $y \in \{a, b\}^*$ is also in $\{a, b\}^*$.
By applying the first principle of mathematical induction, we can conclude that all strings derived from S belong to $\{a, b\}^*$. $\Box$

Lemma 2.2 *Let x denote an arbitrary string in $\{a, b\}^*$; there exists a derivation for x from S.*

Proof: We use induction on the length of x, i.e., $|x|$.
BASIS: $|x| = 0$. In this case, x must be ϵ; since $S \to \epsilon$ is a production in G, x can be derived from S. Thus, the basis is proven.

Assume that whenever $|x| = k$, there is a derivation of x from S.

Now, consider a string x of length $(k+1)$. Without loss of generality, we assume that $x = a \cdot y$, where $y \in \{a, b\}^*$. (The proof can be reworded if $x = b \cdot y$!) Since $|y|$ is k, as per the inductive hypothesis, there must exist a derivation of y from S, i.e., $S \Rightarrow y$. But then x can be derived from S as follows: $S \Rightarrow a \cdot S \cdots a \cdot y \Rightarrow x$.

Using the first principle of mathematical induction, we conclude that every string in $\{a, b\}^*$ can be derived from S. $\square$

$\square$