

For many simple context-free grammars (CFGs), visual examination suffices to convince one that the CFG generates what it is supposed to. But once one deals with more complex CFGs, formal proof becomes necessary. For instance, here is one possible CFG for the language  $L = \{x \in \{0, 1\}^* : N_0(x) = N_1(x)\}$ , where  $N_a(x)$  denotes the number of appearances of the character  $a$  in the string  $x$ :

$$S \longrightarrow 0S1S \mid 1S0S \mid \varepsilon \quad (1)$$

We shall formally prove that this CFG “works.” Before doing so, we recall the formal definition of what it means for a CFG to generate a string.

**Definition:** Let  $G = (V, \Sigma, R, S)$  be a CFG. For strings  $s, s' \in (V \cup \Sigma)^*$ , we write  $s \Rightarrow s'$  if  $\exists \alpha, \beta, \gamma \in (V \cup \Sigma)^*$  and  $v \in V$  such that

- $s = \alpha v \gamma$
- $s' = \alpha \beta \gamma$
- $v \rightarrow \beta$  is a rule in  $R$ .

We write  $s \overset{*}{\Rightarrow} s'$  if  $\exists s_1, s_2, \dots, s_n \in (V \cup \Sigma)^*$  such that  $s \Rightarrow s_1 \Rightarrow s_2 \Rightarrow \dots \Rightarrow s_n \Rightarrow s'$ . We say that  $G$  generates  $x \in \Sigma^*$  if  $S \overset{*}{\Rightarrow} x$ . Finally, we define  $\mathcal{L}(G) := \{x \in \Sigma^* : S \overset{*}{\Rightarrow} x\}$ .

**Theorem:** Let  $G$  denote CFG (1) above. Then  $\mathcal{L}(G) = L$ .

**Proof:** We are trying to prove equality between two sets. As usual, the proof breaks down into two major steps.

*Step 1:  $\mathcal{L}(G) \subseteq L$ :* The application of any of the three rules in  $G$  always produces an equal number of 0s and 1s: either one of each is produced ( $S \rightarrow 0S1S$  and  $S \rightarrow 1S0S$ ) or else none of them is produced ( $S \rightarrow \varepsilon$ ). Therefore any string generated by  $G$  must have an equal number of 0s and 1s. In other words,  $\mathcal{L}(G) \subseteq L$ .

*Step 2:  $L \subseteq \mathcal{L}(G)$ :* The argument for this is much more involved, so we give a detailed proof using mathematical induction. We shall prove the statement “ $\forall x \in L (x \in \mathcal{L}(G))$ ” by induction on  $|x|$ .

The base case is when  $|x| = 0$ . This is trivial: we must have  $x = \varepsilon$  and we indeed have  $\varepsilon \in \mathcal{L}(G)$ .

For the induction step, suppose  $x = a_1 a_2 \dots a_k \in L$  with  $k > 0$  and each  $a_i \in \{0, 1\}$ . We shall assume that  $a_1 = 0$  (the proof is essentially the same for the other case,  $a_1 = 1$ ). Let us define

$$\begin{aligned} d_i &:= N_0(a_1 a_2 \dots a_i) - N_1(a_1 a_2 \dots a_i), \quad \text{for } 0 \leq i \leq k, \\ t &:= \min\{i > 0 : d_i = 0\}. \end{aligned}$$

Putting these definitions in words,  $d_i$  is the difference between the numbers of 0s and 1s in the length- $i$  prefix of  $x$  (it is a *signed* difference, so it could be positive, zero or negative). Since  $a_1 = 0$ , we have  $d_1 = 1$ . We also have  $d_k = N_0(x) - N_1(x) = 0$ , because  $x \in L$ . Thus, there must exist a *smallest positive index*  $i$  such that  $d_i = 0$ : this value of  $i$  is defined to be  $t$ .

Since  $t$  was chosen to be the smallest possible and  $d_1 > 0$ , we must have  $d_{t-1} > 0$ . Since  $d_{t-1} > 0$  and  $d_t = 0$ , we must have  $a_t = 1$ . Therefore, we have

$$x = 0a_2a_3 \dots a_{t-1}1a_{t+1} \dots a_k.$$

Let  $y := a_2a_3 \dots a_{t-1}$  and  $z := a_{t+1} \dots a_k$ . Since  $d_t = 0$ , we have  $N_0(a_1 \dots a_t) = N_1(a_1 \dots a_t)$ , i.e.,  $N_0(0y1) = N_1(0y1)$ , which in turn implies  $N_0(y) = N_1(y)$ , so that  $y \in L$ . It is also clear that  $N_0(z) = N_1(z)$ , so that  $z \in L$ .

It is time to apply the induction hypothesis. Since  $|y| < k$  and  $|z| < k$  and  $y, z \in L$ , we have  $y, z \in \mathcal{L}(G)$ . By definition, this means  $S \overset{*}{\Rightarrow} y$  and  $S \overset{*}{\Rightarrow} z$ . Now, consulting the rules of  $G$ , we see that

$$S \Rightarrow 0S1S \overset{*}{\Rightarrow} 0y1S \overset{*}{\Rightarrow} 0y1z = x,$$

which shows that  $x \in \mathcal{L}(G)$ . This completes the induction step and the proof of the theorem. QED.