Pumping Lemma for CFL

Let L be a context-free language. Then there exists an integer $p \geq 1$, called the pumping length, *such that the following holds: Every string s in L, with $|s| \geq p$, can be written as $s = uvxyz$, such that*

1. $|vy| \geq 1$ *(i.e., v and y are not both empty),*
2. $|vxy| \leq p$, *and*
3. $uv^ixy^iz \in L$, *for all $i \geq 0$.*

P r o o f :

The proof of the pumping lemma will use the following result about parse trees:
Let L be a context-free language and let ∑ be the alphabet of L,then there exists a context-free grammar in Chomsky normal form, G = (V, ∑,P,S),such that L = L(G).

Define r to be the number of variables of G and define $p = 2^r$. We will prove that the value of p can be used as the pumping length. Consider an arbitrary string s in L such that $|s| \geq p$, and let T be a parse tree for s. Let ℓ be the height of T. Then, by Lemma 3.8.2, we have

$$|s| \leq 2^{\ell-1}.$$

On the other hand, we have

$$|s| \geq p = 2^r.$$

By combining these inequalities, we see that $2^r \leq 2^{\ell-1}$, which can be rewritten as

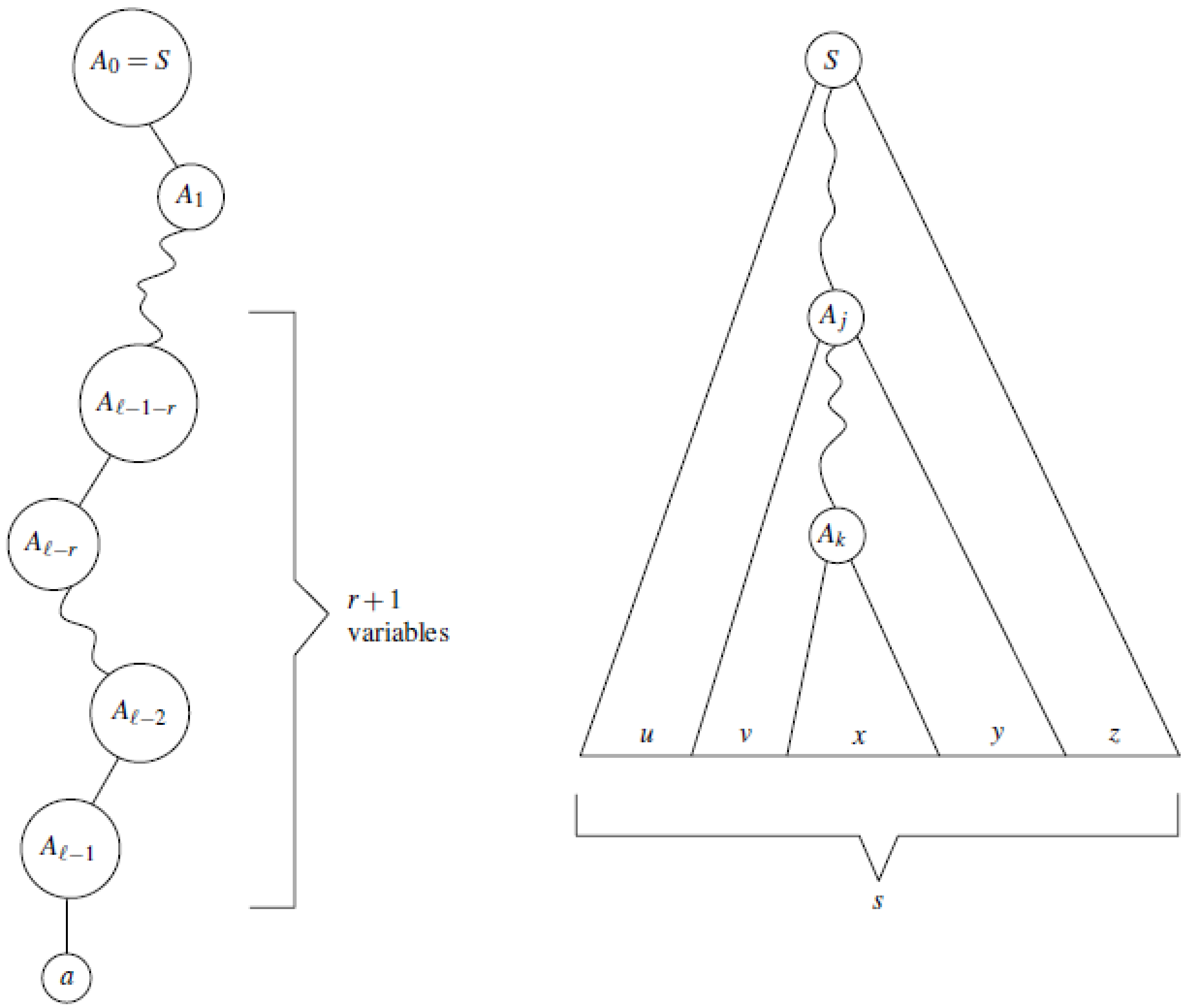
$$\ell \geq r + 1.$$

Consider the nodes on a longest root-to-leaf path in T. Since this path consists of ℓ edges, it consists of $\ell + 1$ nodes. The first ℓ of these nodes store variables, which we denote by $A_0, A_1, \ldots, A_{\ell-1}$ (where $A_0 = S$), and the last node (which is a leaf) stores a terminal, which we denote by a.

Since $\ell - 1 - r \geq 0$, the sequence

$$A_{\ell-1-r}, A_{\ell-r}, \ldots, A_{\ell-1}$$

of variables is well-defined. Observe that this sequence consists of $r + 1$ variables. Since the number of variables in the grammar G is equal to r, the pigeonhole principle implies that there is a variable that occurs at least twice in this sequence. In other words, there are indices j and k, such that $\ell - 1 - r \leq j < k \leq \ell - 1$ and $A_j = A_k$. Refer to the figure below for an illustration.



Recall that T is a parse tree for the string s. Therefore, the terminals stored at the leaves of T, in the order from left to right, form s. As indicated in the figure above, the nodes storing the variables A_j and A_k partition s into five substrings u, v, x, y, and z, such that $s = uvxyz$.

It remains to prove that the three properties stated in the pumping lemma hold. We start with the third property, i.e., we prove that

$$uv^ixy^iz \in L, \text{ for all } i \geq 0.$$

In the grammar G, we have

$$S \stackrel{*}{\Rightarrow} uA_jz. \tag{3.3}$$

Since $A_j \stackrel{*}{\Rightarrow} vA_ky$ and $A_k = A_j$, we have

$$A_j \stackrel{*}{\Rightarrow} vA_jy. \tag{3.4}$$

Finally, since $A_k \stackrel{*}{\Rightarrow} x$ and $A_k = A_j$, we have

$$A_j \stackrel{*}{\Rightarrow} x. \tag{3.5}$$

From (3.3) and (3.5), it follows that

$$S \stackrel{*}{\Rightarrow} uA_jz \stackrel{*}{\Rightarrow} uxz,$$

which implies that the string uxz is in the language L. Similarly, it follows from (3.3), (3.4), and (3.5) that

$$S \stackrel{*}{\Rightarrow} uA_jz \stackrel{*}{\Rightarrow} uvA_jyz \stackrel{*}{\Rightarrow} uvvA_jyyz \stackrel{*}{\Rightarrow} uvvxyyz.$$

Hence, the string uv^2xy^2z is in the language L. In general, for each $i \geq 0$, the string uv^ixy^iz is in the language L, because

$$S \stackrel{*}{\Rightarrow} uA_jz \stackrel{*}{\Rightarrow} uv^iA_jy^iz \stackrel{*}{\Rightarrow} uv^ixy^iz.$$

This proves that the third property in the pumping lemma holds.

Next we show that the second property holds. That is, we prove that $|vxy| \leq p$. Consider the subtree rooted at the node storing the variable A_j. The path from the node storing A_j to the leaf storing the terminal a is a longest path in this subtree. (Convince yourself that this is true.) Moreover, this path consists of $\ell - j$ edges. Since $A_j \stackrel{*}{\Rightarrow} vxy$, this subtree is a parse tree for the string vxy (where A_j is used as the start variable). Therefore, by Lemma 3.8.2, we can conclude that $|vxy| \leq 2^{\ell-j-1}$. We know that $\ell - 1 - r \leq j$, which is equivalent to $\ell - j - 1 \leq r$. It follows that

$$|vxy| \leq 2^{\ell-j-1} \leq 2^r = p.$$

Finally, we show that the first property in the pumping lemma holds. That is, we prove that $|vy| \geq 1$. Recall that

$$A_j \stackrel{*}{\Rightarrow} vA_k y.$$

Let the first rule used in this derivation be $A_j \to BC$. (Since the variables A_j and A_k, even though they are equal, are stored at different nodes of the parse tree, and since the grammar G is in Chomsky normal form, this first rule exists.) Then

$$A_j \Rightarrow BC \stackrel{*}{\Rightarrow} vA_k y.$$

Observe that the string BC has length two. Moreover, by applying rules of a grammar in Chomsky normal form, strings cannot become shorter. (Here, we use the fact that the start variable does not occur on the right-hand side of any rule.) Therefore, we have $|vA_k y| \geq 2$. But this implies that $|vy| \geq 1$. This completes the proof of the pumping lemma.

Lemma 3.8.2 *Let G be a context-free grammar in Chomsky normal form, let s be a non-empty string in $L(G)$, and let T be a parse tree for s. Let ℓ be the height of T, i.e., ℓ is the number of edges on a longest root-to-leaf path in T. Then*

$$|s| \leq 2^{\ell-1}.$$