Context free grammar $G = (N, \Sigma, P, S)$

Parse tree is a tree satisfying the following conditions.

1. Each interior node is labelled with an element of N
2. Each leaf node is labelled with Σ or ϵ.
3. if an interior node is labelled A and its children are labelled $B_1, \dots B_k$.

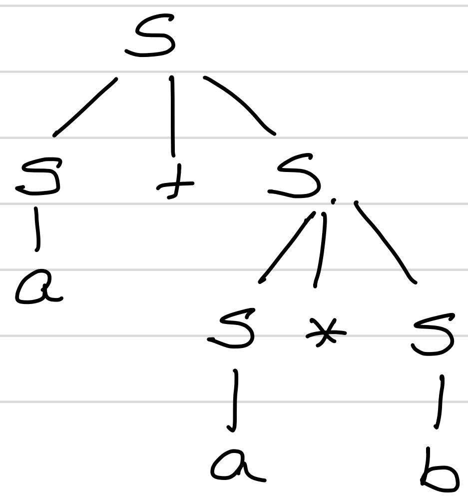
Then $A \rightarrow B_1 B_2, \dots B_k \in P$.

$S \rightarrow S+S \mid S*S \mid (S) \mid I$

$I \rightarrow a \mid b$

String $a + a*b$

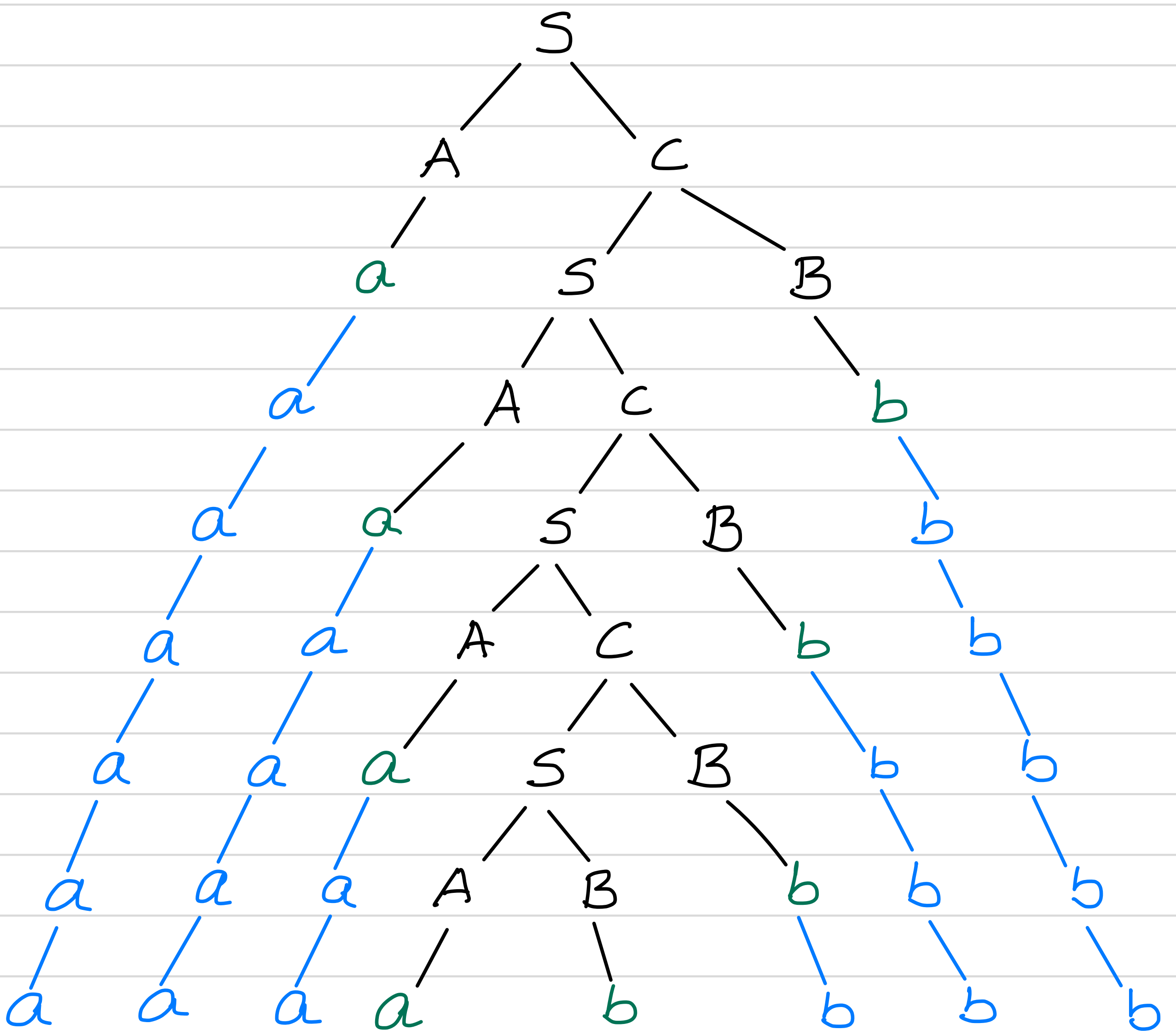
$S \rightarrow S+S \rightarrow S+S*S \xrightarrow{*} a+a*b$.



G: $S \to AC \mid AB$, $A \to a$, $B \to b$, $C \to SB$.

CNF for $\{a^n b^n \mid n \geq 1\}$

Consider the derivation of $a^4 b^4$.



Observation. For a grammar in Chomsky normal form, any parse tree for a long string should have a long path.

Any long path should have at least two occurences of some nonterminal symbol.

For a grammar in CNF - the number of symbols can atmost double going down a level in the parse tree - RHS of each production contains atmost 2 symbols.

We have 1 symbol at level 0
atmost 2 symbols at level 1.
⋮
2^i symbols at level i.

To have 2^n symbols at the bottom level, the tree must be of depth* atleast n. Thus the parse tree has at least $n+1$ levels.

* Depth - number of edges in the longest path from the root to a leaf node.

Pumping Lemma for CFLs.

if $A \subseteq \Sigma^*$ is a CFL then there exist $k \geq 0$ such that for every $z \in A$ of length at least k can be split into five substrings $z = u v w x y$ such that $vx \neq \epsilon$, $|vwx| \leq k$ and for all $i \geq 0$, $u v^i w x^i y \in A$.

Proof.

Let G be a grammar for A in CNF.

Take $k = 2^{n+1}$, where n is the number of nonterminals of G.

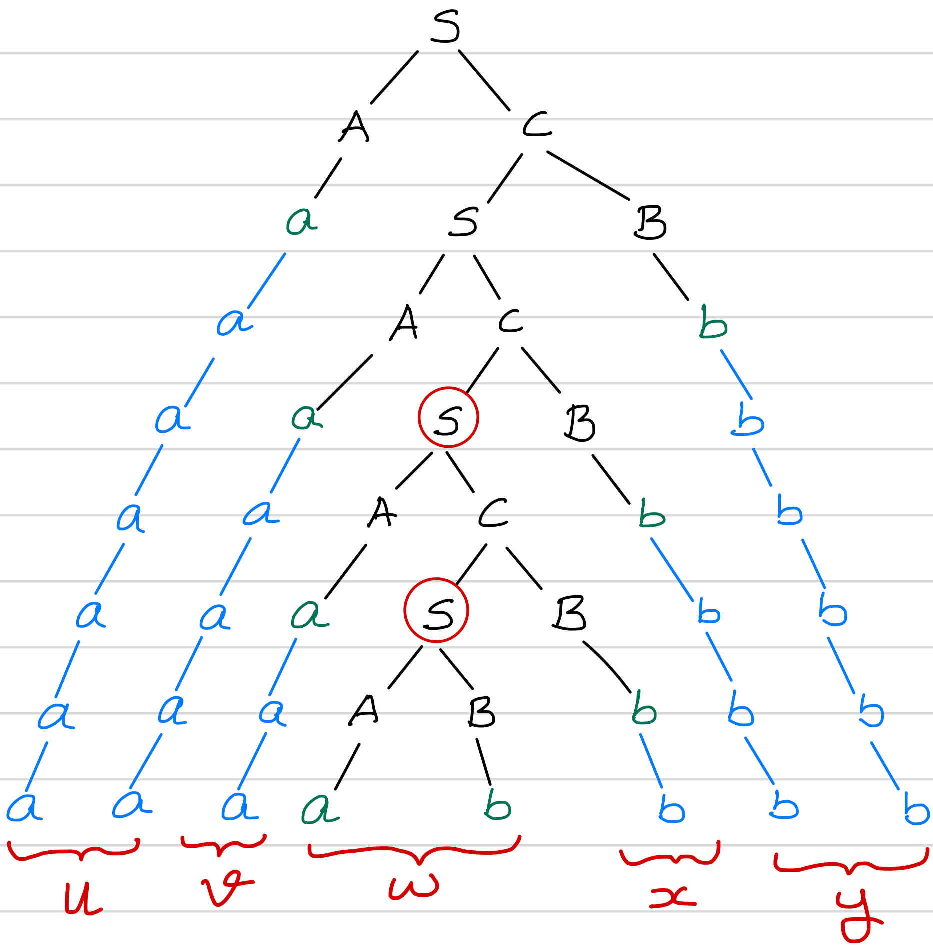
Suppose $z \in A$ and $|z| \geq k$.

Any parse tree in G for z must be of depth at least $n+1$ (i.e. there are $n+2$ levels).

Consider the longest path in the tree (it is of length atleast $n+1$).

The longest path contains atleast $n+1$ occurence of nonterminals.

This implies: Some nonterminal occurs more than once.

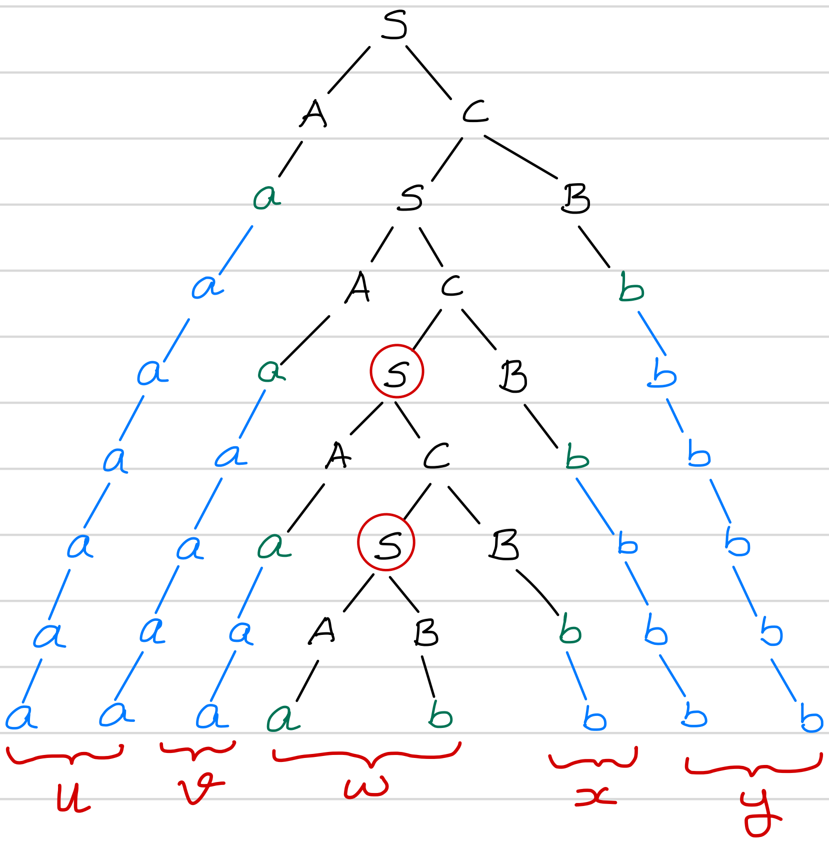


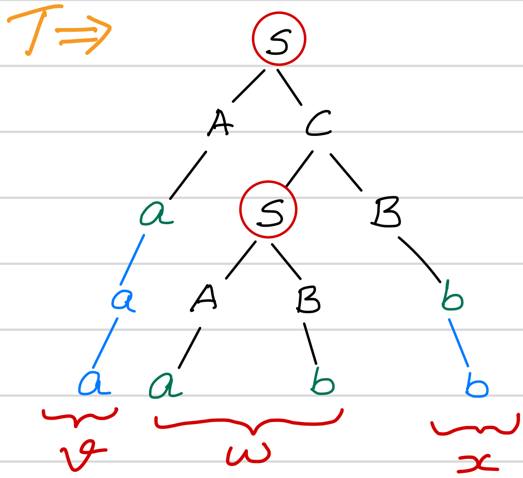
Take the first pair of occurences of the same nonterminal along the path - traversing bottom to top.

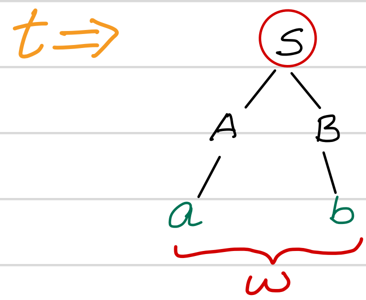
Suppose X is the nonterminal with two occurences.
Split $z = uvwxy$ such that.

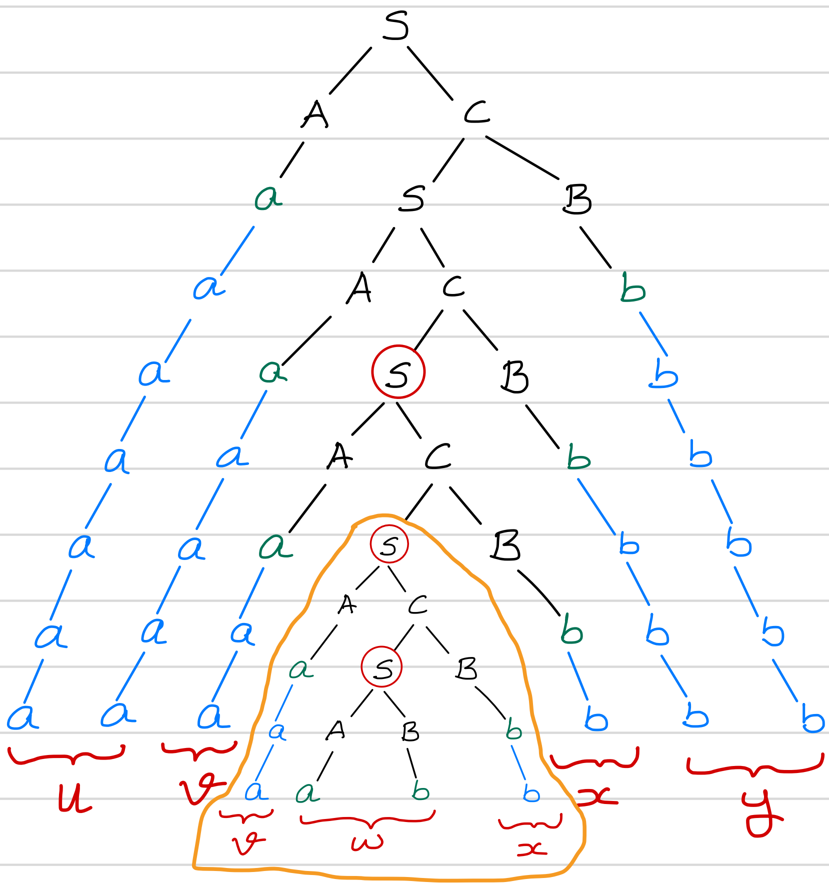
w - string (of terminals) generated by lower occurence of X
vwx - string generated by the upper occurence of X

Let T - subtree rooted at upper occurence of X.
t - subtree rooted at the lower occurence of X



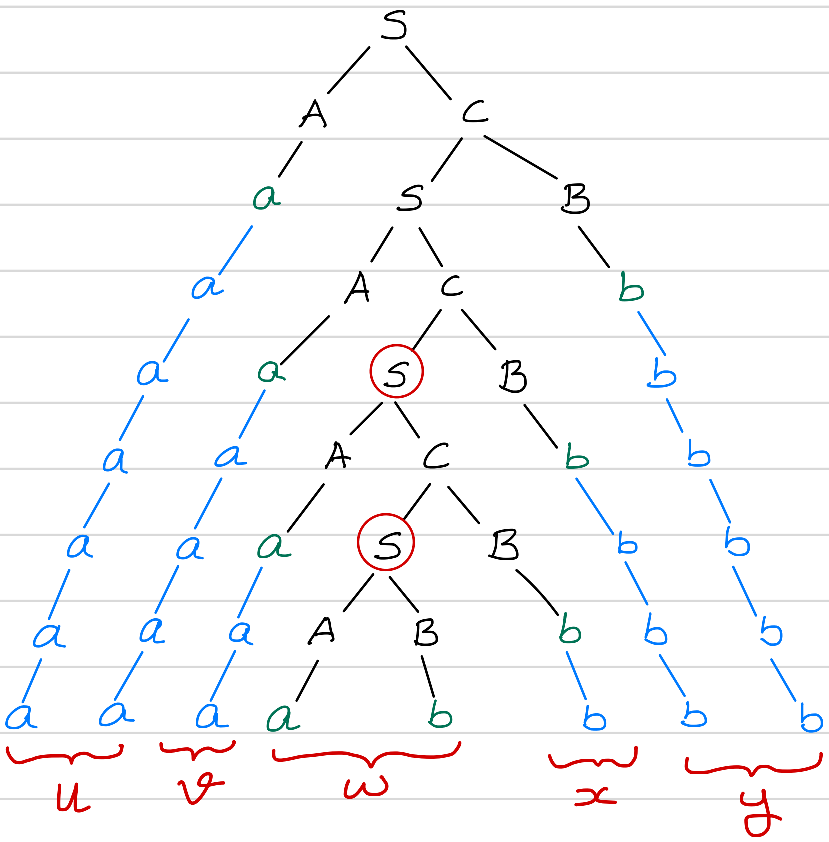




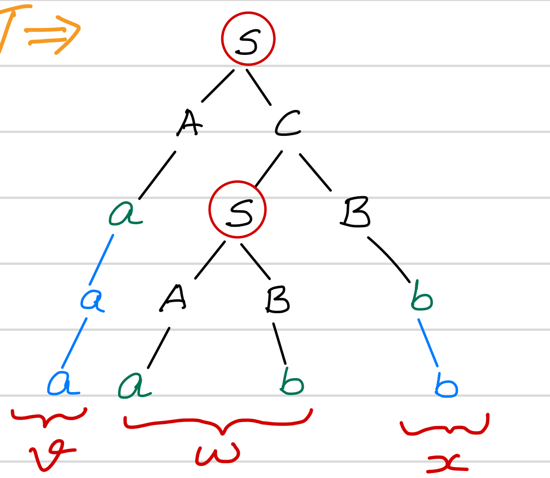


Consider the string uv^2wx^2y

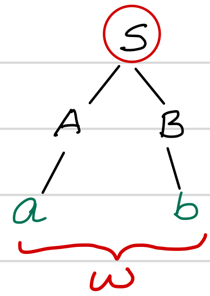
We can replace t with as many copies of T to get a parse tree for uv^iwx^iy for all $i \geq 1$.

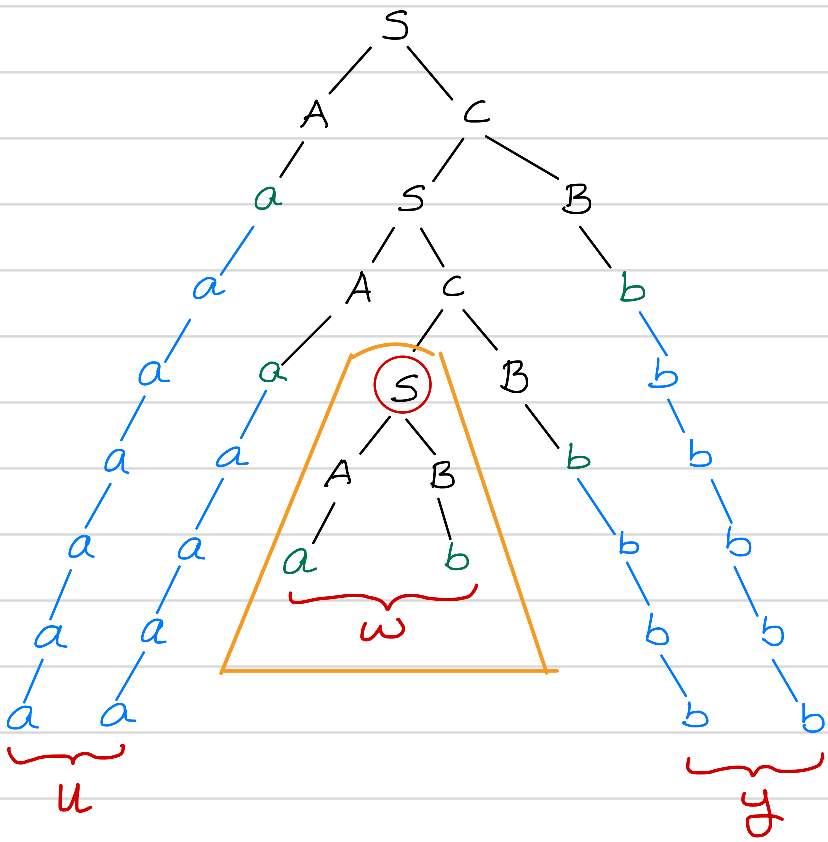


$T \Rightarrow$



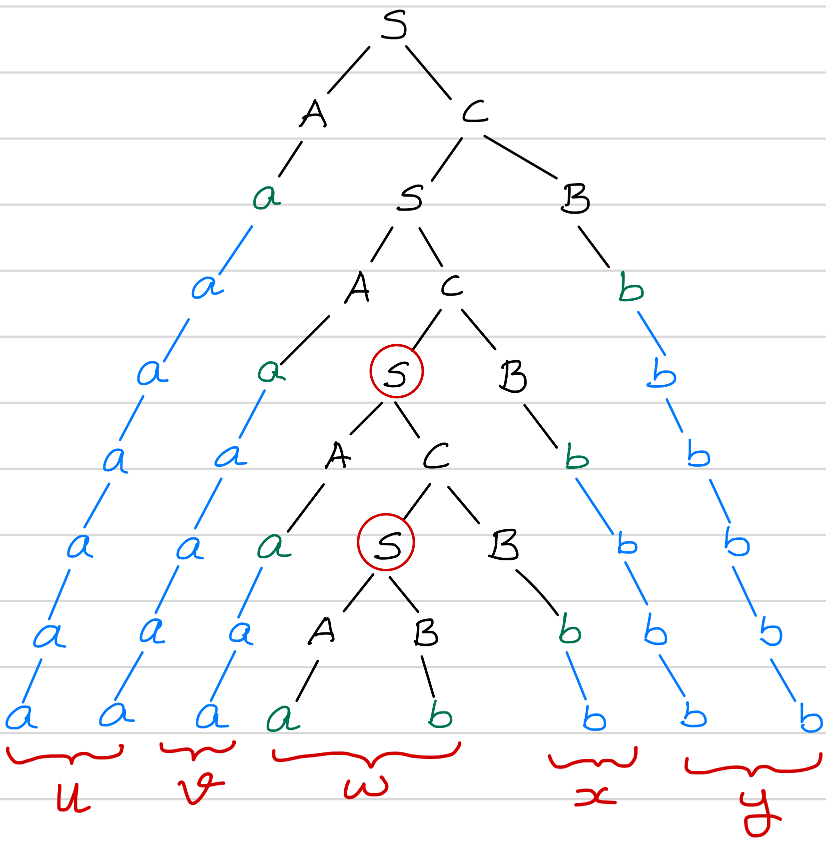
$t \Rightarrow$

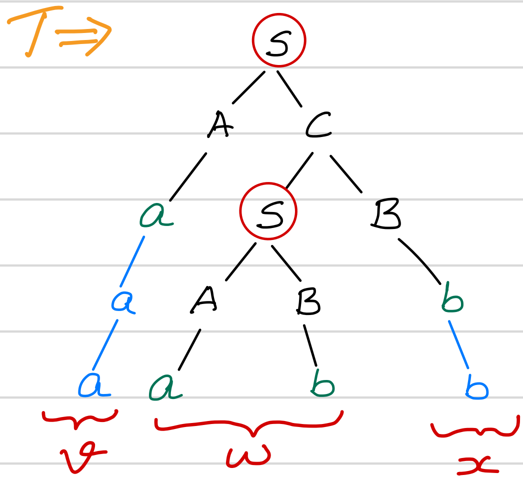


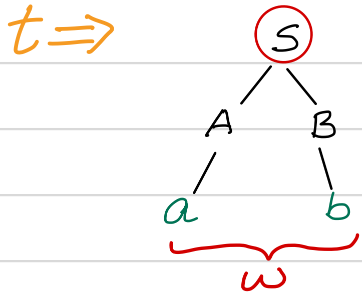


Consider the string uv^2wx^2y

We can cut T and replace it with t to get a parse tree for $uv^0wx^0y = uwy$.







Note.

1. $vx \neq \epsilon$. – v and x are not both ϵ.

2. $|vwx| \leq k$ – Since we chose the first repeated occurence of a nonterminal from the bottom.

   Depth of the subtree under the upper occurence of the repeated nonterminal X is atmost $n+1$

   $\therefore$ it can have atmost $2^{n+1} = k$ terminals

$\square$

To show that a set is not a CFL - use pumping lemma in the contrapositive form.

For all $k \geq 0$, $\exists z \in A$ s.t $|z| \geq k$ and for all split of z into substrings $z = uvwxy$ with $vx \neq \epsilon$ and $|vwx| \leq k$, there exists an $i \geq 0$ s.t $uv^i w x^i y \notin A$

$A = \{a^n b^n a^n \mid n \geq 0\}$ is not context free.

Proof. Given k, let $z = a^k b^k a^k$. We have $z \in A$, $|z| = 3k$

Now consider any split $z = uvwxy$ with $vx \neq \epsilon$ and $|vwx| \leq k$. Let $i = 2$. Consider the string uv^2wx^2y.

Case 1. v or x contains atleast one "a" and at least one "b". then uv^2wx^2y is not of the form $a^* b^* a^*$

Case 2. v and x contains only a's. Then uv^2wx^2y has more a's than b's.

Case 3. v and x contains only b's. Then the number of b's is greater than the number of a's.

Case 4. One of v or x contains only a's and the other only b's. Then uv^2wx^2y is not of the form $a^m b^m a^m$.

Thus in all cases, the resulting string $uv^2wx^2y \notin A$. By pumping lemma, A is not a CFL.