Properties of Context-Free Languages

Reading: Chapter 7

Background Information for the Pumping Lemma for Context-Free Languages

- **Definition:** Let G = (V, T, P, S) be a CFG. If every production in P is one of the following two forms:

$$A \rightarrow BC$$
$$A \rightarrow a$$

where A, B and C are all in V and a is in T, then G is in Chomsky Normal Form (CNF).

- **Example:** (not quite!)

$$S \rightarrow AB \mid BA \mid aSb$$
$$A \rightarrow a$$
$$B \rightarrow b$$

- **Theorem:** Let L be a CFL. Then L – {ε} is a CFL.

- **Theorem:** Let L be a CFL not containing {ε}. Then there exists a CNF grammar G such that L = L(G).

- **Definition:** Let T be a tree. Then the <u>height</u> of T, denoted h(T), is defined as follows:
  - If T consists of a single vertex then h(T) = 0
  - If T consists of a root r and subtrees T_1, T_2, … T_k, then $h(T) = \max_i\{h(T_i)\} + 1$

- **Lemma:** Let G be a CFG in CNF. In addition, let w be a string of terminals where A=>*w and w has a derivation tree T. If T has height k>=1, then $|w| <= 2^{k-1}$.

- **Proof:** By induction on h(T) (exercise).

- **Corollary:** Let G be a CFG in CNF, and let w be a string in L(G). If $|w| >= 2^k$, where k>=0, then any derivation tree for w using G has height at least k+1.

- **Proof:** Follows from the lemma.

Pumping Lemma for Context-Free Languages

- **Lemma:**

  Let G = (V, T, P, S) be a CFG in CNF, and let $n = 2^{|V|}$. If z is a string in L(G) and $|z| >= n$, then there exist strings u, v, w, x and y in T* such that z=uvwxy and:

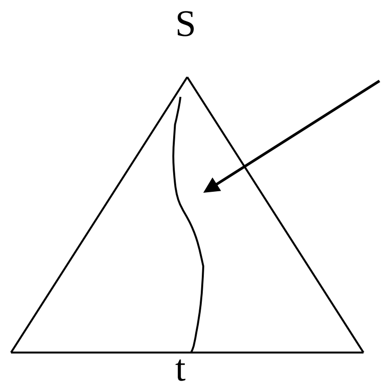
  – $|vx| >= 1$ (i.e., $|v| + |x| >= 1$)
  – $|vwx| <= n$
  – uv^iwx^iy is in L(G), for all $i >= 0$

- **Proof:**
  Let G = (V, T, P, S) be a CFG in CNF, let $n = 2^k$, where $k = |V|$, and let z be a string in L(G) where $|z| >= n$.

  Since $|z| >= n = 2^k$, it follows from the corollary that any derivation tree for z has height at least k+1.

  By definition such a tree contains a path of length at least k+1.

  Consider the longest such path in the tree:



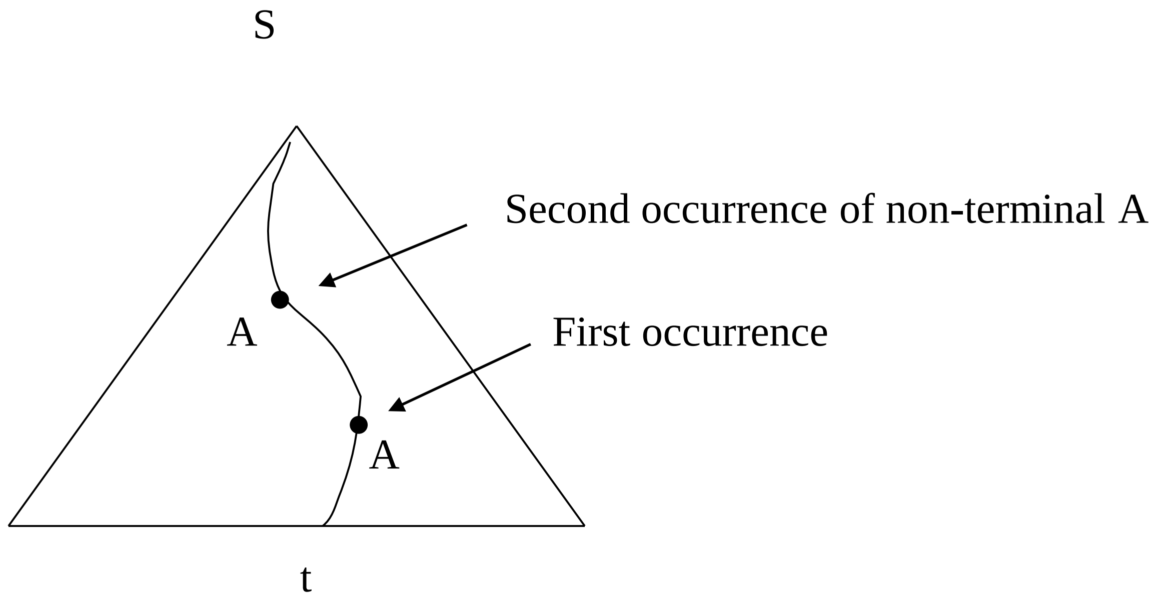
yield of T is z

Such a path has:

- Length >= k+1 (i.e., number of edges in the path is >= k+1)
- At least k+2 nodes
- 1 terminal
- At least k+1 non-terminals

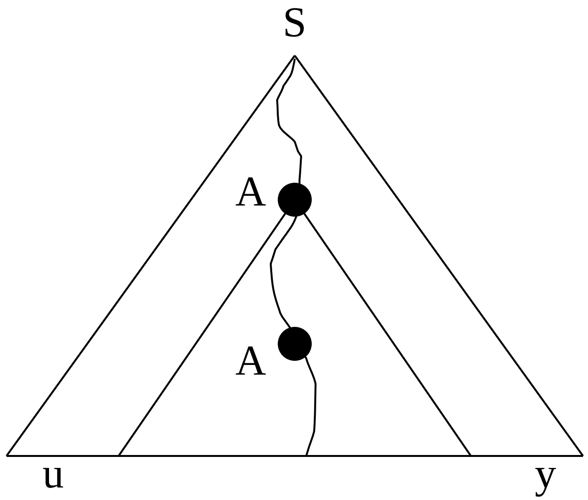
- Since there are only k non-terminals in the grammar, and since k+1 appear on this path, it follows that some non-terminal (perhaps many) appears at least twice on this path.

- Consider the first non-terminal that is repeated, when traversing the path from the leaf to the root.

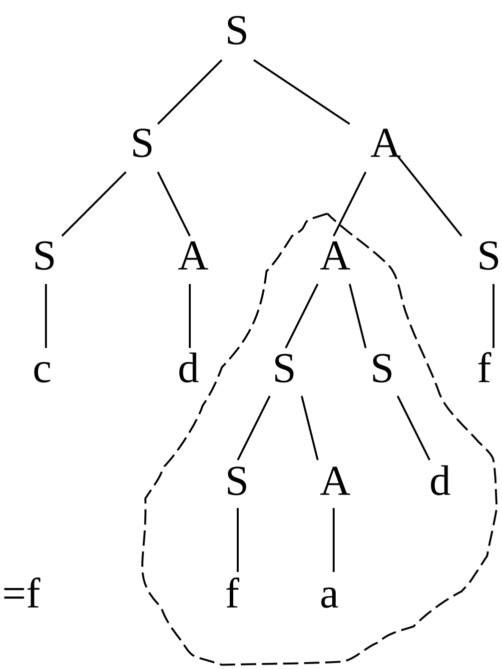


This path, and the non-terminal A will be used to break up the string z.

- **Generic Description:**



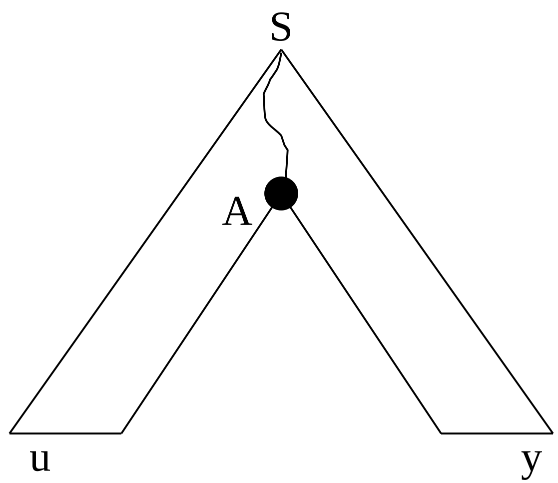
- **Example:**



S -> SA
A -> SS | AS
S -> c | f | d
A -> d | a

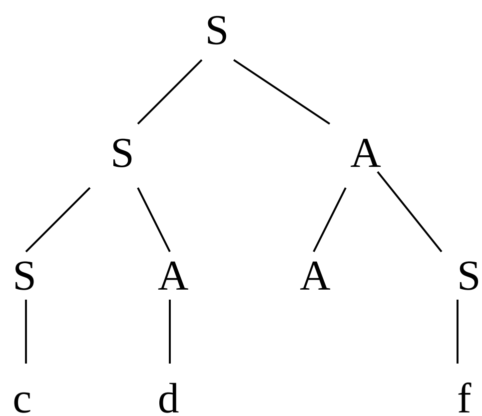
In this case u = cd and y =f

- **Cut out the subtree rooted at A:**



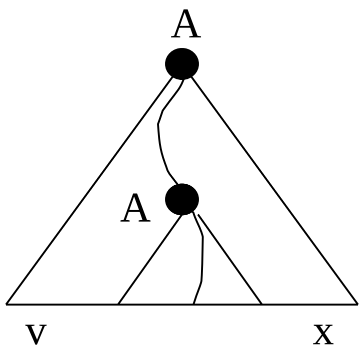
S =>* uAy (1)

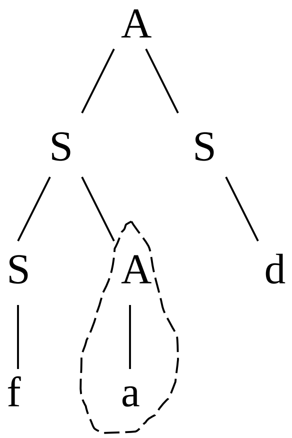
- **Example:**



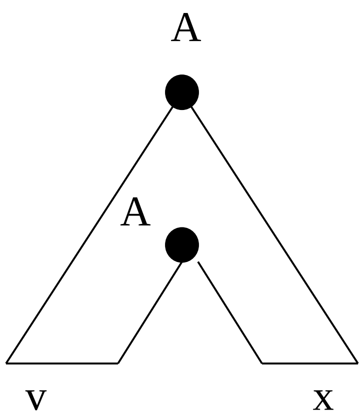
S =>* cdAf

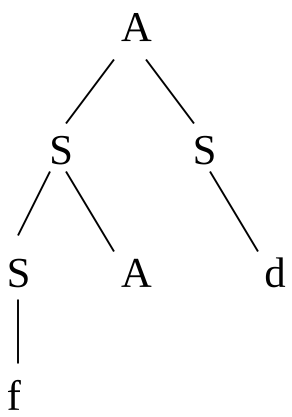
- **Consider the subtree rooted at A:**





- **Cut out the subtree rooted at the first occurrence of A:**

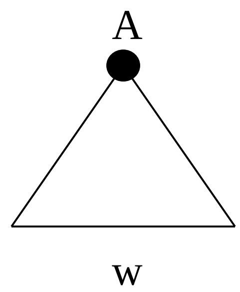


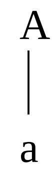


A =>* vAx (2)

A =>* fAd

- **Consider the smallest subtree rooted at A:**





$A \Rightarrow^* w$ (3)

$A \Rightarrow^* a$

- **Collectively (1), (2) and (3) give us:**

$$
\begin{aligned}
S &\Rightarrow^* uAy && (1)\\
&\Rightarrow^* uvAxy && (2)\\
&\Rightarrow^* uvwxy && (3)\\
&\Rightarrow^* z && \text{since } z=uvwxy
\end{aligned}
$$

- **In addition, (2) also tells us:**

$$
\begin{array}{ll}
S \Rightarrow^* uAy & (1) \\
\quad \Rightarrow^* uvAxy & (2) \\
\quad \Rightarrow^* uv^2Ax^2y & (2) \\
\quad \Rightarrow^* uv^2wx^2y & (3)
\end{array}
$$

- **More generally:**

$$S \Rightarrow^* uv^iwx^iy \qquad \text{for all } i>=1$$

- **And also:**

$$
\begin{array}{ll}
S \Rightarrow^* uAy & (1) \\
\quad \Rightarrow^* uwy & (3)
\end{array}
$$

- **Hence:**

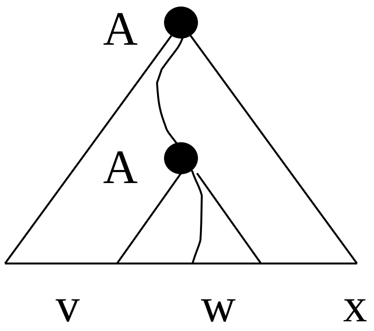
$$S \Rightarrow^* uv^iwx^iy \qquad \text{for all } i>=0$$

- **Consider the statement of the Pumping Lemma:**

–*What is n?*

$n = 2^k$, where k is the number of non-terminals in the grammar.

–*Why is |v| + |x| >=1?*



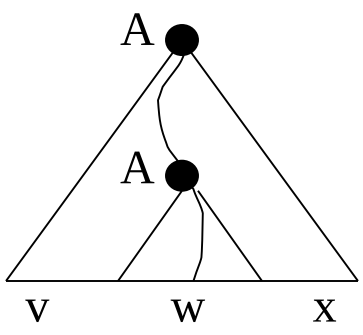
Since the height of this subtree is >=2, the first production is A->V_1V_2. Since no non-terminal derives the empty string (in CNF), either V_1 or V_2 must derive a non-empty v or x. More specifically, if w is generated by V_1, then x contains at least one symbol, and if w is generated by V_2, then v contains at least one symbol.

–*Why is |vwx| <= n?*

Observations:

- The repeated variable was the first repeated variable on the path from the bottom, and therefore (by the pigeon-hole principle) the path from the leaf to the second occurrence of the non-terminal has length *at most* k+1.

- Since the path was the largest in the entire tree, this path is the longest in the subtree rooted at the second occurrence of the non-terminal. Therefore the subtree has height <=k+1. From the lemma, the yield of the subtree has length $<=2^k=n$.•



- **Examples of showing languages are not context-free:**
  - http://my.fit.edu/~pbernhar/Teaching/FormalLanguages/nonContextFree1.pdf
  - http://my.fit.edu/~pbernhar/Teaching/FormalLanguages/nonContextFree2.pdf

Closure Properties for Context-Free Languages

- **Theorem:** The CFLs are closed with respect to the union, concatenation and Kleene star operations.

- **Proof:** (details left as an exercise) Let L_1 and L_2 be CFLs. By definition there exist CFGs G_1 and G_2 such that $L_1 = L(G_1)$ and $L_2 = L(G_2)$.

  – For union, show how to construct a grammar G_3 such that $L(G_3) = L(G_1) \cup L(G_2)$.
  – For concatenation, show how to construct a grammar G_3 such that $L(G_3) = L(G_1)L(G_2)$.
  – For Kleene star, show how to construct a grammar G_3 such that $L(G_3) = L(G_1)^*$.•

- **Theorem:** The CFLs are not closed with respect to intersection.

- **Proof:** (counter example)
  Let

  $$L_1 = \{a^i b^i c^j \mid i,j >= 0\}$$

  and

  $$L_2 = \{a^i b^j c^j \mid i,j >= 0\}$$

  Note that both of the above languages are CFLs.

  If the CFLs were closed with respect to intersection then

  $$L_1 \cap L_2$$

  would have to be a CFL.

  But this is equal to:

  $$\{a^i b^i c^i \mid i >= 0\}$$

  which is not a CFL.•

- **Lemma:** Let L_1 and L_2 be subsets of Σ^*. Then $\overline{L_1 \cup L_2} = \overline{L_1} \cap \overline{L_2}$.

- **Theorem**: The CFLs are not closed with respect to complementation.

- **Proof:** (by contradiction) Suppose that the CFLs were closed with respect to complementation, and let L_1 and L_2 be CFLs. Then:

  $\overline{L_1}$ would be a CFL

  $\overline{L_2}$ would be a CFL

  $\overline{L_1} \cup \overline{L_2}$ would be a CFL

  $\overline{\overline{L_1} \cup \overline{L_2}}$ would be a CFL

  But by the lemma:

  $\overline{\overline{L_1} \cup \overline{L_2}} = \overline{\overline{L_1}} \cap \overline{\overline{L_2}} = L_1 \cap L_2$ a contradiction.•

- **Theorem:** Let L be a CFL and let R be a regular language. Then $L \cap R$ is a CFL.

- **Proof:** (exercise – sort of)•

- **Question:** Is $L \cap R$ regular?

- **Answer:** Not always. Let L = $\{a^i b^i \mid i >= 0\}$ and R = $\{a^i b^j \mid i,j >= 0\}$, then $L \cap R = L$ which is not regular.