Session 24

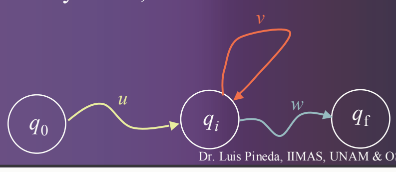
Pumping Lemma for CFG

How can we tell whether a language is a *CFL*?

- First answer: define a CFG or design a PDA for such a language
- But, what if we have a language described by some other means:
  - $L = \{a^i b^i c^i \mid i \geq 1\}$
  - Is this language a CFL?
- Use the pumping lemma for CFL
- Antecedents:
  - Chomsky Normal Form (1959)
  - Due to Bar-Hillel, Perles and Shamir (1961)
  - The pumping lemma for RE is a simplification of the corresponding lemma for CFL

The Pumping Lemma for *RL*

- Suppose L is a regular language recognized by a FA with n states; then, for any $x \in L$ with $|x| \geq n$, $x = uvw$ for some strings satisfying:
  - $|uv| \leq n$
  - $|v| > 0$
  - For any $m \geq 0$, $uv^mw \in L$



The "loop" in strings of *CFLs*

- In long enough derivations, variables have to repeat:

  $S \Rightarrow^* vAz \Rightarrow^* vwAyz \Rightarrow^* vwxyz$

  where $v, w, x, y, z \in \Sigma^*$
- The context before and after a variable in the right-side of a production (e.g. w and y in $A \rightarrow wAy$) is pumped up with the repetition of the variable in a derivation:

  $S \Rightarrow^* vAz \Rightarrow^* vwAyz \Rightarrow^* vw^2Ay^2z \Rightarrow^* vw^3Ay^3z \Rightarrow^* \ldots$

  since x can be derived from each A,

  $vAz \Rightarrow^* vxz \in L$

  $vwAyz \Rightarrow^* vwxyz \in L$

  $vw^2Ay^2z \Rightarrow^* vw^2xy^2z \in L$

  $vw^3Ay^3z \Rightarrow^* vw^3xy^3z \in L$

  …

The Pumping Lemma for *CFL*

- Let $G = (V, \Sigma, S, P)$, be a CFG in CNF with a total of p variables. Any string u in $L(G)$ with $|u| \geq 2^{p+1}$ can be written as $u = vwxyz$, for some strings v, w, x, y, and z satisfying:
  - $|wy| > 0$
  - $|wxy| \leq 2^{p+1}$
  - For any $m \geq 0$, $vw^mxy^mz \in L$

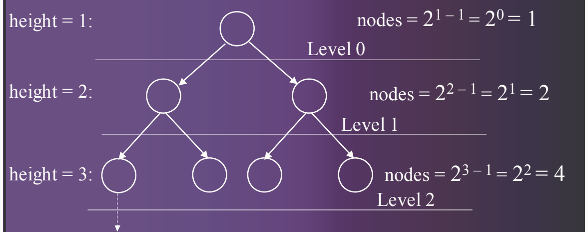
The Pumping Lemma for *CFL*

- The conditions do not come from Mars:

  (*i*) $|u| \geq 2^{p+1}$ (*ii*) $|wy| > 0$ (*iii*) $|wxy| \leq 2^{p+1}$
  - for a parameter p (i.e. the number of distinct variables in V)
- A sketch of the story:
  - Syntactic structures produce by Grammars in CNF are binary trees all the way down until nodes dominating terminal symbols, which have only one descendant
  - A binary tree of height h has a yield of size $\leq 2^{h-1}$, so a binary tree having more than 2^{h-1} leafs has a height greater than h
  - If the grammar has p variables, a derivation of any string of size equal or greater than 2^{p+1} has a path whose height is greater than $p + 2$ and some variable must appear at least twice!
  - (*i*) and (*ii*) are constraints on substrings generated by binary trees with paths of height long enough (i.e. constraint (*i*)) to have a variable at least twice!

Some facts about binary trees

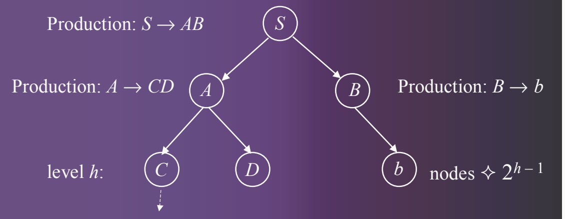
- Height of a path: number of nodes in a path
- In a complete binary tree number of the nodes in level h is 2^{h-1}



- The derivation of a string u (or yield) with more than 2^{h-1} symbols has a height greater than h
- Nodes in level $l = 2^l$

Structures in CNF

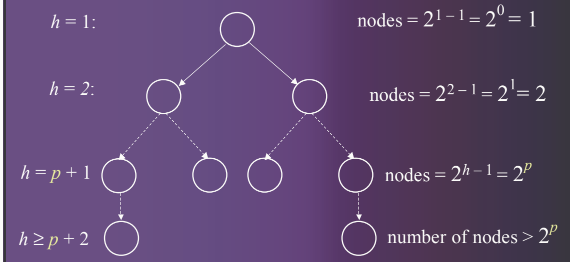
- Height of a tree: the height of the largest path
- In a structure produce by a CFG in CNF there may be fewer but no more than 2^{h-1} nodes in the level h



- A string (or yield) with more than 2^{h-1} symbols has a height greater than h

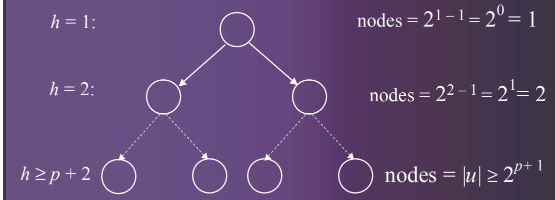
Some facts about binary trees

- If number of leafs $> 2^p$ then the height is at least $p + 2$



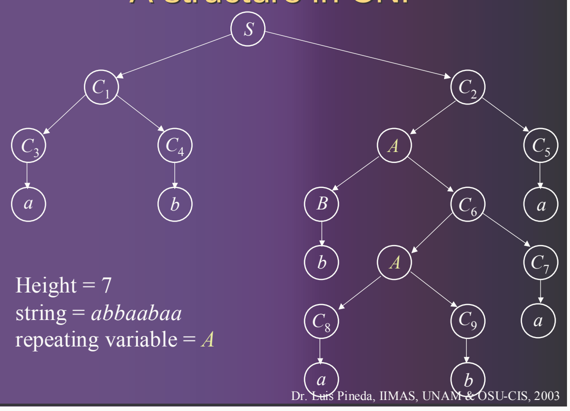
size of strings and derivation length

- If $|u| \geq 2^{p+1}$ then the height is at least $p + 2$



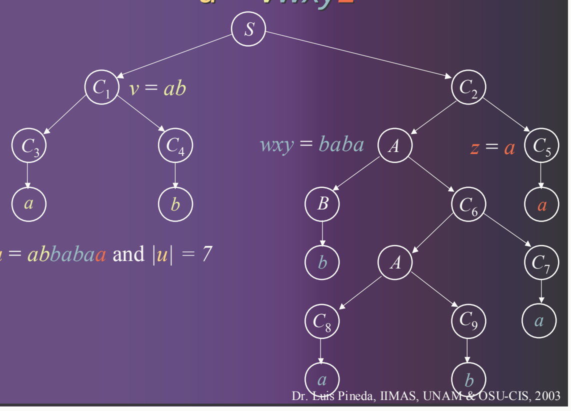
- If there are p different variables in the grammar, a string u such that $|u| \geq 2^{p+1}$ has a syntactic tree whose height is at least $p + 2$

A structure in CNF

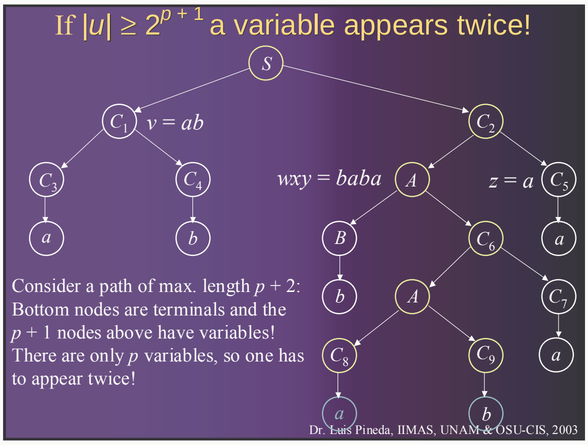


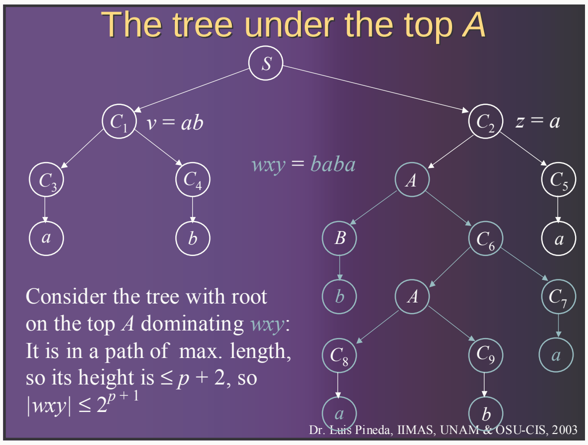
Height = 7
string = *abbaabaa*
repeating variable = *A*

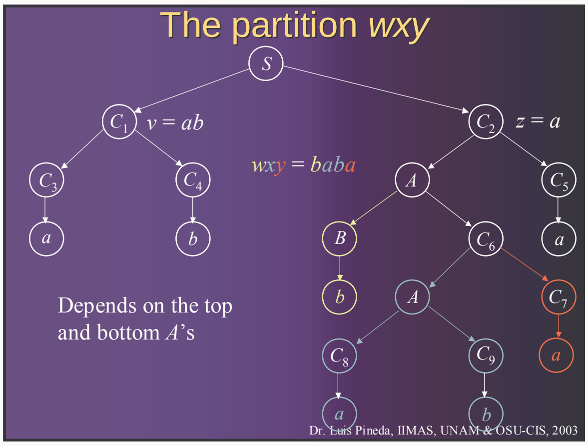
$u = vwxyz$

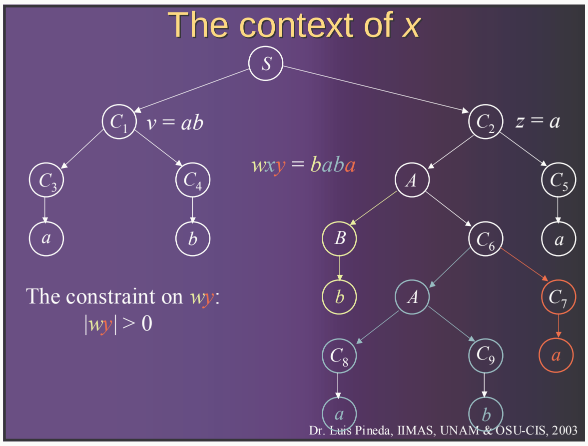


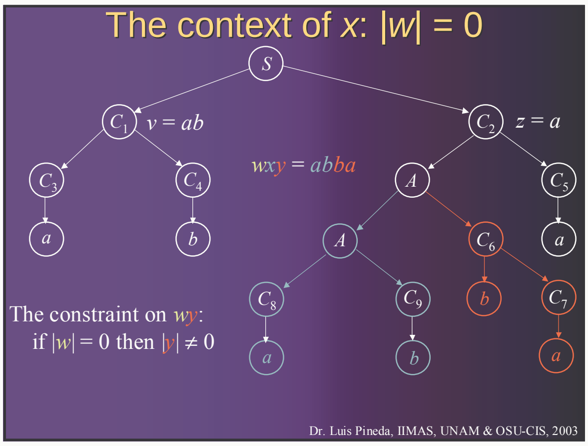
$u = abbabaa$ and $|u| = 7$

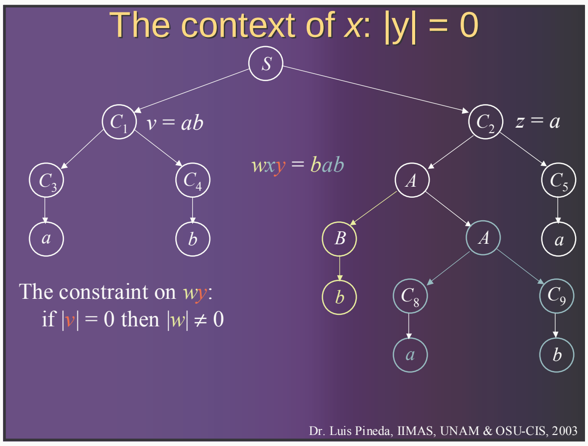

If |u| ≥ 2^(p + 1) a variable appears twice!
v = ab
wxy = baba
z = a
Consider a path of max. length p + 2:
Bottom nodes are terminals and the
p + 1 nodes above have variables!
There are only p variables, so one has
to appear twice!
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003


The tree under the top A
v = ab
z = a
wxy = baba
Consider the tree with root
on the top A dominating wxy:
It is in a path of max. length,
so its height is ≤ p + 2, so
|wxy| ≤ 2^(p + 1)
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003


The partition wxy
v = ab
z = a
wxy = baba
Depends on the top
and bottom A's
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003


The context of x
v = ab
z = a
wxy = baba
The constraint on wy:
|wy| > 0
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003


The context of x: |w| = 0
v = ab
z = a
wxy = abba
The constraint on wy:
if |w| = 0 then |y| ≠ 0
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003


The context of x: |y| = 0
v = ab
z = a
wxy = bab
The constraint on wy:
if |y| = 0 then |w| ≠ 0
Dr. Luis Pineda, IIMAS, UNAM & OSU-CIS, 2003

The Pumping Lemma for *CFL*

- Let L a *CFL*. There is an integer n so that for any u satisfying $|u| \geq n$ there are strings $v, w, x, y,$ and z satisfying:
  - $u = vwxyz$
  - $|wy| > 0$
  - $|wxy| \leq n$
  - For any $m \geq 0$, $vw^mxy^mz \in L$
- Proof:
  - Find a CFG in CNF that generates $L - \{\Lambda\}$.
  - Let p be the number of variables in this grammar and $n = 2^{p+1}$

Using the pumping lemma for *CFL*

- If we have a language described by some other means:
  - $L = \{a^ib^ic^i \in \Sigma^* \mid i \geq 1\}$
  - Is this a *CFL*?
- Strategy:
  - Assume that the pumping lemma for *CFL* holds
  - If a contradiction follows from this assumption the language is not context free!

$L = \{a^ib^ic^i \in \Sigma^* \mid i \geq 1\}$

- Let n be the constant and $u = a^nb^nc^n$
  - $|u| = 3n$ (This is ok: $n = 2^{p+1}$)
- Partition u into $vwxyz$ such that $|wxy| \leq n$ and $|wy| > 0$; since $|wxy| \leq n$ this substring has at most two distinct types of symbols:
- Choose $m = 0$ in vw^mxy^mz
  - Since $|wy| > 0$, either $|w| > 0$ or $|y| > 0$ (or both!)
  - The segments of two symbols containing w and y have less symbols than the segment including the symbol which is not in wxy
  - L is not a CFL!

$L = \{a^ib^ic^i \in \Sigma^* \mid i \geq 1\}$

Choose $m = 0$ in vw^mxy^mz:

- Case 1: wxy is in the a's block:
  - $w = a^i$, $y = a^j$ and $a^{n-i-j}b^nc^n \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 2: wxy is in the a's and b's block:
  - $a^ib^jc^n \notin L$ as $i < n$ or $j < n$ (maybe both) and $i + j < 2n$
- Case 3: wxy in the b's block:
  - $w = b^i$, $y = b^j$ and $a^nb^{n-i-j}c^n \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 4: wxy in the b's anc c's blocks:
  - $a^nb^ic^j \notin L$ as $i < n$ or $j < n$ (maybe both) and $i + j < 2n$
- Case 5: wxy in the c's block:
  - $w = c^i$, $y = c^j$ and $a^nb^nc^{n-i-j} \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$

$L = \{a^ib^ic^i \in \Sigma^* \mid i \geq 1\}$

- The abstraction:
  - The segments of two symbols containing w and y have less symbols than the segment including the symbol which is not in wxy
- Case 1: wxy is in the a's block:
  - $w = a^i$, $y = a^j$ and $a^{n-i-j}b^nc^n \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 2: wxy is in the a's and b's block:
  - $a^ib^jc^n \notin L$ as $i < n$ or $j < n$ (maybe both) and $i + j < 2n$
- Case 3: wxy in the b's block:
  - $w = b^i$, $y = b^j$ and $a^nb^{n-i-j}c^n \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 2 includes case 1 and case 3!
  - The segment $|a^ib^j| < 2n$

$L = \{a^ib^ic^i \in \Sigma^* \mid i \geq 1\}$

- The abstraction:
  - The segments of two symbols containing w and y have less symbols than the segment including the symbol which is not in wxy
- Case 3 (again): wxy in the b's block:
  - $w = b^i$, $y = b^j$ and $a^nb^{n-i-j}c^n \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 4: wxy in the b's anc c's blocks:
  - $a^nb^ic^j \notin L$ as $i < n$ or $j < n$ (maybe both) and $i + j < 2n$
- Case 5: wxy in the c's block:
  - $w = c^i$, $y = c^j$ and $a^nb^nc^{n-i-j} \notin L$ as $i > 0$ or $j > 0$ and $n - i - j < n$
- Case 4 includes case 3 and 5!
  - The segment $|b^ic^j| < 2n$

$L = \{a^i b^i c^i \in \Sigma^* \mid i \geq 1\}$

- The abstraction:
  - The segments of two symbols containing w and y have less symbols than the segment including the symbol which is not in wxy
- The abstraction: let $m = 0$
  - p: The segment $|a^i b^j| < 2n$ and $a^i b^j c^n \notin L$
  - q: The segment $|b^i c^j| < 2n$ and $a^n b^i c^j \notin L$
  - L is not a CFL as p or q

$L = \{x \in \{a, b\}^* \mid n_a(x) < n_b(x)$ and $n_a(x) < n_c(x)$

- Let n be the constant and $u = a^n b^{n+1} c^{n+1}$
  - $|u| = 3n + 2$ (This is ok: $n = 2^{p+1}$)
- Partition u into $vwxyz$ such that $|wxy| \leq n$ and $|wy| > 0$
  - Again wxy has at most two kinds of symbols
- Case 1:
  - w or y have at least one a
  - choose $m = 2$ and $a^i b^j c^{n+1} \notin L$ as $i \geq n + 1$ so $a^i \geq c^{n+1}$
- Case 2:
  - w or y have no a's
  - chose $m = 0$ and $a^n b^i c^j \notin L$ as $i < n + 1$ or $j < n + 1$ (maybe both) and $n_a(u) \geq n_b(u)$ or $n_a(u) \geq n_c(x)$

Non-CFL