CSE 396 Lecture Tue 4/2 Spr 2019

Top Hat 0832

Theorem: Given any CFG $G_0 = (V_0, \Sigma, R_0, S_0)$ we can build a grammar G' in Chomsky NF s.t. $L(G') = L(G_0) \setminus \{\epsilon\}$.

Proof: skipped — part will be covered in Ch 4.

CFG Pumping Lemma: Let $L \subseteq \Sigma^*$ be any CFL. Then:

there exists an $N > 0$ such that [text]

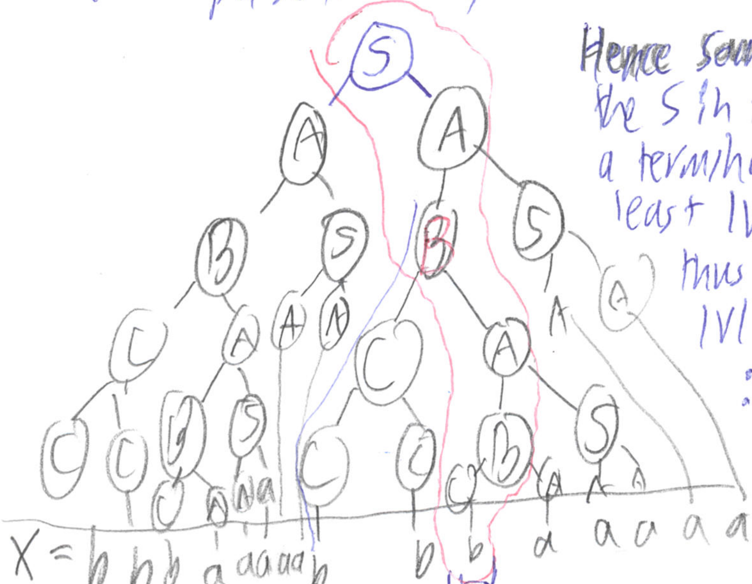
forall $x \in L$ [w] with $|x| > N$ [$w =: uvxyz$]

there exists a breakdown $x =: yuvwz$ s.t. $|uvw| \leq N$ and $|uw| > 0$ i.e. $uw \neq \epsilon$ [$|vxy| \leq N$ and $|vy| > 0$ — v and y are not both ϵ]

[$s_i = uv^i x y^i z$ is in L. $[uxz \in L$ too$]$]

and forall $i \geq 0$, $x^{(i)} = yu^i v w^i z$ is in L. [$yvz \in L$ as case $i = 0$] "pumping down".

Proof: By the Theorem, there is a CFG G [$G = (V, \Sigma, R, S)$] in Chomsky NF s.t. $L(G) = L \setminus \{\epsilon\}$. Take $N = 2^{|V|}$. Let any $x \in L(G)$ $|x| > N$, be given. By $x \in L(G)$, we can take a parse tree T for x. Since G is in ChNF, T is a binary tree.

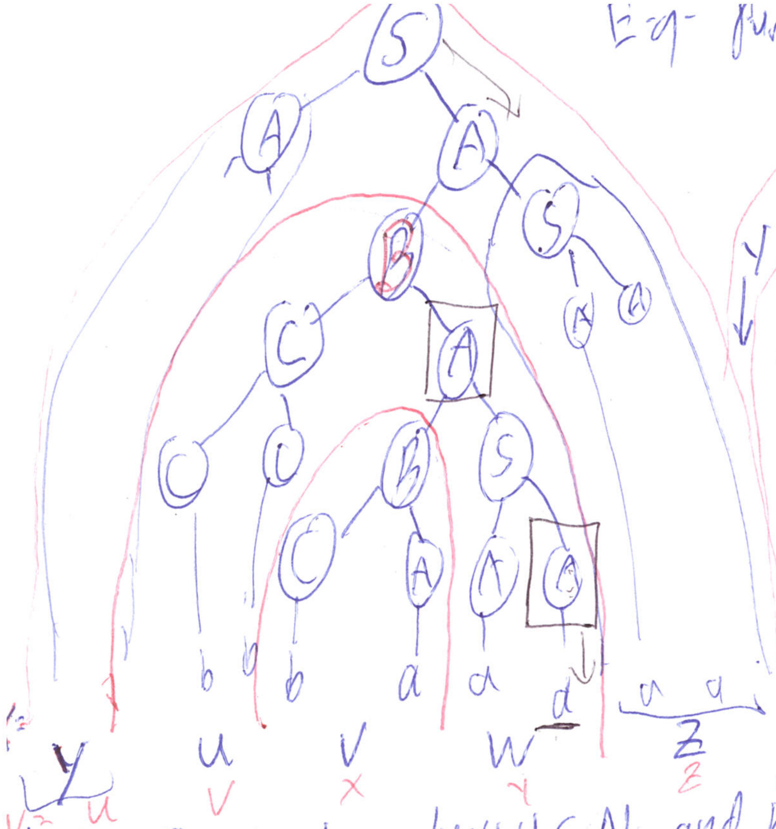


Hence some path from the S in the root to a terminal has at least $|V|+1$ edges, thus more than $|V|+1$ variables. $\therefore$ By PHP, some variable appears twice.

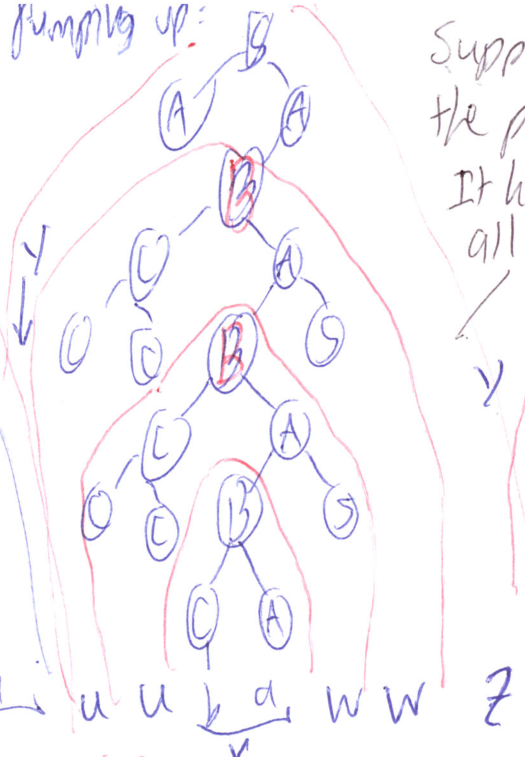
Example: $S \to AA \mid BC$
$|V| = 4$ so $N = 16$.
$A \to BS \mid a$
$B \to CA \mid AS$
$C \to CC \mid b$

Key principle: Any binary tree of height h in edges has at most 2^h leaves. Hence if T has $n > N = 2^h$ leaves, then it has height $\geq h+1$ in edges.

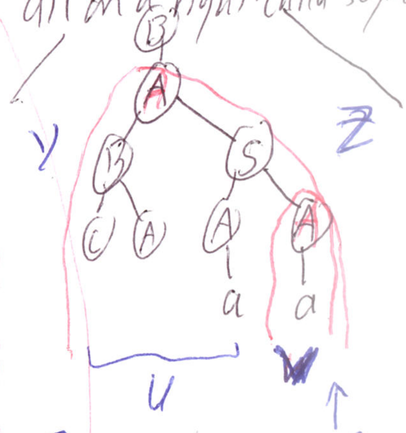
Take the lowest such occurrence. Then the higher occurrence of the variable (here, the upper B in red), has height $\leq |V|+1$ in nodes, which is $\leq |V|$ in edges. Hence the yield of the subtree of that variable has length at most $2^{|V|} = N$. Look at T now



E.g. pumping up:



Suppose we chose (B) the path to the a in blac[k]. It has two A's that are all on a right-child segme[nt]

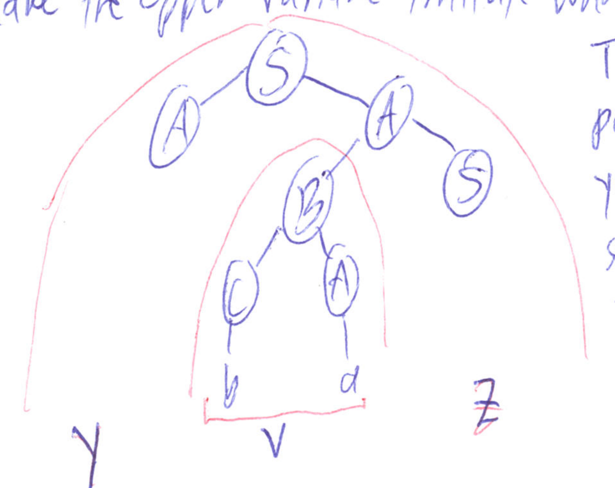


Then $w = \varepsilon$.

We have $|uvw| \le N$, and because one B is below the other, $uw \neq \varepsilon$, i.e. at least one of u or w is not empty.

We can do two kinds of adjustments to T to derive new strings.

- Make the upper variable imitate what the ~~lower~~ lower one did.



Thus we have a parse tree T' yielding yvz, so $x^{(0)} = yu^0vw^0z$ is in $L(G) = L$.

- Make the lower variable imitate the upper multiple times until finally doing what it did.

Thus we have a parse tree T'' for $x^{(2)} = yuuvwwz$ so $x^{(2)} \in L(G) = L$.

Repeating gives $x^{(i)} = yu^ivw^iz \in L$ for all i.

But then u cannot be ε, so $x^{(i)} = yu^ivz$ are all different strings as $i = 0, 1, 2, \ldots$

In the symmetric all left-children case, we get $u = \varepsilon$, $w \neq \varepsilon$.

Hence in every case $|uvw| \le N$, $uw \neq \varepsilon$, and $x^{(i)} \in L$ for all i, which completes the proof. ☒

This immediately implies the contrapositive form.

Original: If L is a CFL then blah →

Contra: If ¬ blah... then L is not a CFL!

Contrapositive: Let L be any language. Suppose that

for all $N > 0$

there exists $x \in L$ with $|x| > N$ such that

for all breakdowns $x =: yuvwz$ subject to $|uvw| \le N$ and $uw \ne \varepsilon$,

there exists $i \ge 0$ such that $x^{(i)} = yu^ivw^iz$ is not in L.

Then L is not a CFL. This form gives a proof script.

Example: $L = \{a^nb^nc^n : n \ge 1\}$.

Let any $N > 0$ be given

Take $x = \underline{a^Nb^Nc^N}$. Then $x \in L$. (clearly)

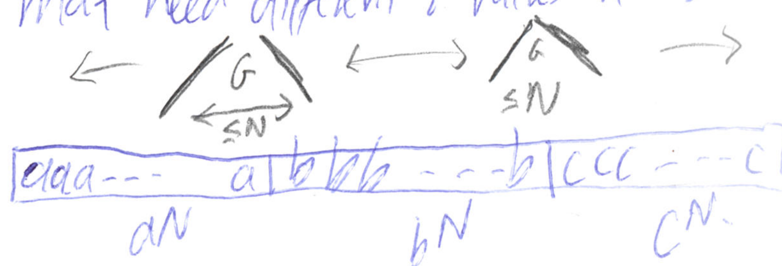
Let any breakdown $x =: yuvwz$ with $|uvw| \le N$, $uw \ne \varepsilon$, be given.

Take $i = \underline{0}$.

(Do lots of analysis and often cases about where the breakdown is, which may need different i values in cases.)

In each case we get $x^{(i)} \notin L$, so L is not a CFL by the CFL PL.

$x =$



By $|uvw| \le N$, we cannot have that u and w collectively include both an 'a' and a 'c'. But by $uw \ne \varepsilon$, uw must include at least one a, b, or c. Hence $x^{(0)} = yvz$ subtracts at least one letter but cannot balance the counts of all three letters. Hence $x^{(i)} \notin L$ with $i = 0$, so L is not a CFL