Pumping Lemma for CFL's

A pigeonhole argument also can be used to prove a pumping lemma for CFL's, but it is more involved than for RL's.

Pumping Lemma for CFL's: *(Sipser, 2.34)* Suppose L is a context-free language. Then there exists a number p such that, if s is any string in L of length at least p, then s may be written $s = uvxyz$ so that the following all hold:

1. $uv^ixy^iz \in L$ for all $i >= 0$
2. $|vy| > 0$ (*i.e.* either v or y is not ϵ)
3. $|vxy| \leq p$.

Proof Idea:

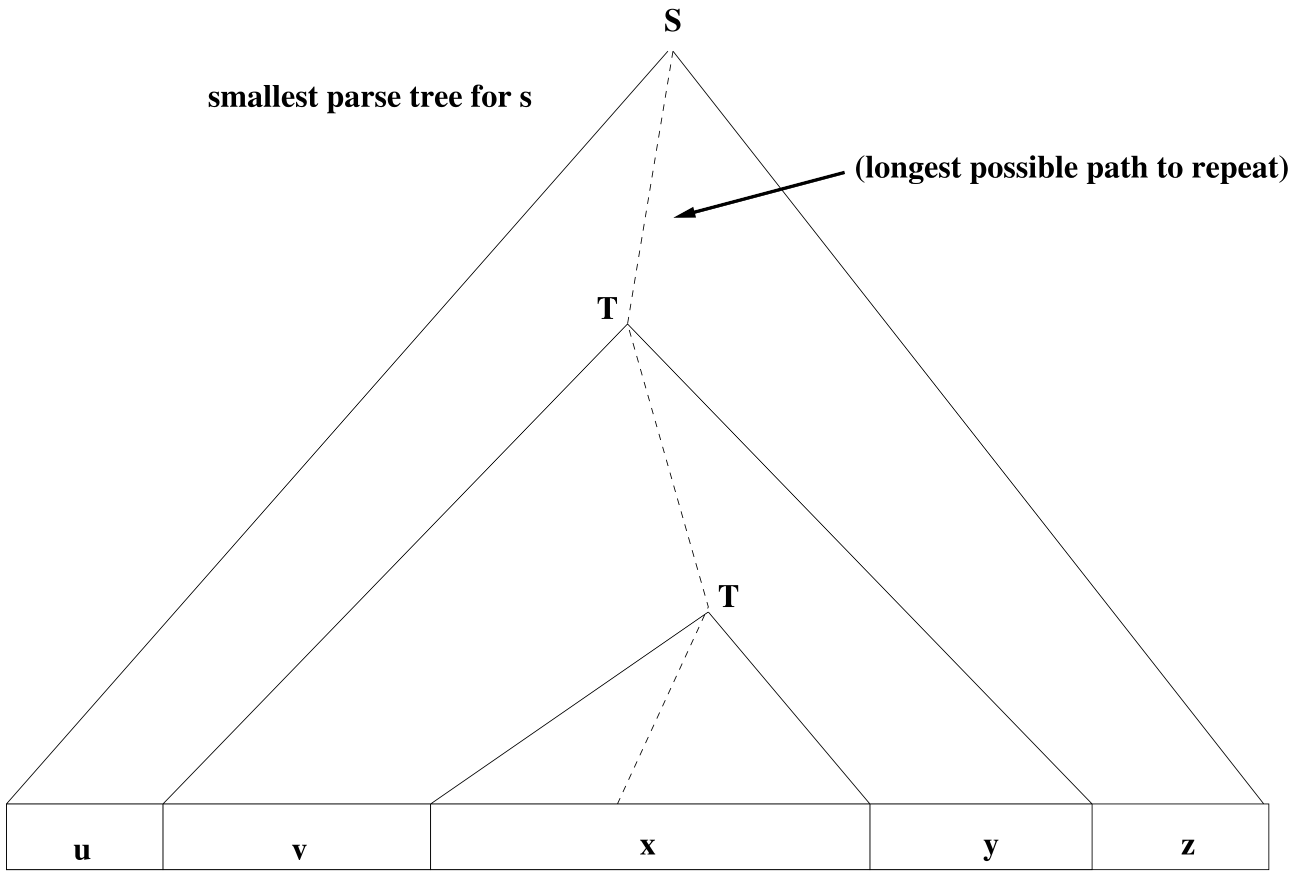
1. Parse trees for long strings must be tall, because the branching is bounded by the maximum number of symbols on the RHS of a rule.
2. A tall parse tree must have a long path from the root. On a sufficiently long path, there must be a repeated variable (by pigeonhole).
3. We can "chop out" or "replicate" the portion of a parse tree between two occurrences of the same variable on a path from the root.

Details

Suppose L is a CFL and let $G = (V, \Sigma, R, S)$ be a CFG with $L = L(G)$.

- A parse tree of height h can have at most b^h leaves, where b is the maximum number of symbols on the RHS of a rule of G.
- Therefore, any parse tree for a string s with $|s| > b^h$ must have height $> h$.
- Let $p = b^{|V|+1}$. For any string $s \in L$, if $|s| \geq p$, then any parse tree for s must have height $\geq |V| + 1$.

- Given a string s with $|s| \geq p$, consider a "smallest" parse tree for s (*e.g.* in terms of number of nodes).
- The chosen parse tree must have height $\geq |V| + 1$, so there is a path from the root to a leaf of height $\geq |V| + 1$. Since there are only $|V|$ variables, some variable repeats along that path (pigeonhole!).
- Consider a repeat that is "lowest" (*i.e* the path from the root to the first instance of the repeated variable is as long as possible).


S
smallest parse tree for s
(longest possible path to repeat)
T
T
u
v
x
y
z

- Let u, v, x, y, z be as in the figure.
- We have $|vy| > 0$; otherwise both v and y are ϵ and we could obtain a smaller parse tree by "chopping out" the portion between the repeated variables.
- We have $|vxy| \leq p$ because the "lowest" repeat must be found no more than height $|V|+1$ above the leaves. This means that $|vxy| \leq p = b^{|V|+1}|$.
- We have $uv^ixy^iz \in L$ for all $i \geq 0$, because we can "chop out" and then "replicate" the portion of the tree between the repeated variables as many times as we like.

Application

Prop: The language $L = \{ww \mid w \in \{0,1\}^*\}$ is not context-free.

Proof: Suppose L were CF. Let G be a CFG with $L = L(G)$. Let p be the pumping length associated with G. Consider the following string in L:

$$s = 0^p 1^p 0^p 1^p$$

Applying the PL, $s = uvxyz$ where $|vy| > 0$ and $|vxy| \le p$.

Consider the possible overlaps of vxy with s:

```
        0...0...01...1...10...0...01...1...1
(1)             vxy
(3)                      vxy
(2)                                vxy
```

1. If vxy is entirely in the first half of s, then $uvvxyyz$ has a 1 in the first position of the second half, so cannot be in L.

2. If vxy is entirely in the second half of s, then $uvvxyyz$ has a 0 in the last position of the first half, so cannot be in L.

3. If vxy overlaps the middle, then $uxz = uv^0xy^0z$ cannot be in L, because we have taken away some 1s from the first half and some 0s from the second half.

Intuitive Interpretation

For strings in a CFL, the portions of a string that are "far apart" can have only a very limited connection to each other.