Top Hat #6570

**Def^n** A CFG $G = (V, \Sigma, R, S)$ is in Chomsky normal form (ChNF) if every rule is in $V \to \Sigma$ or $V \to VV$.

That is, every rule has the form $A \to c$, $c \in \Sigma$, or $A \to BB$ or $A \to AA$ or $A \to BC$, $A, B, C \in V$.

(Rest of page discusses last two steps to ChNF, which meaningless) ADDED:

**Theorem**: For every CFG $G_0$ we can build a CFG $G_4$ in ChNF such that $L(G_4) = L(G_0) \setminus \{\epsilon\}$.

[If $\epsilon \in L(G_0)$, the text allows adding a new $S_4'$: $S_4' \to \epsilon \mid$ Right-hand sides of rules for $S_4$ in $G_4$. Then the grammar $G_4'$ is in "Relaxed ChNF" and gives $L(G_4') = L(G_0)$ exactly. OK by me

The last two steps are "silly" IMHO:

We will get $G_2$ in which each rule is $A \to c$ or $A \to \vec{X}$ with $|\vec{X}| \geq 2$.

3. If a terminal $c$ occurs in $\vec{X}$, add an "alias variable" $P_c$ and the rule $P_c \to c$. Then replace every occurrence of $c$ in $\vec{X}$ by $P_c$. Doing this for all rules makes $L(G_3) = L(G_2)$, all rules in $V \times \Sigma$ or $V \times V^*$.
4. Given a rule $A \to \vec{X}$ where $\vec{X}$ consists of 3 or more variables, break $\vec{X}$ down in steps of 2 like so: $A \to BCDE$ becomes $A \to BQ_1$, $Q_1 \to CQ_2$, $Q_2 \to DE$, where these are the only rules for the new variables $Q_1, Q_2$ in $G_4$

**Def$^n$:** A variable $A$ is **nullable** if $A \Rightarrow^* \varepsilon$.

The first steps on the way to ChNF are:

1. Identify the subset $\mathrm{NULLABLE} \subseteq V$ of nullable vars.
2. Alter the rules so that $\varepsilon$-rules $B \to \varepsilon$ are no longer needed. * Text blends these steps *.

**Algm for 1.** Initialize $N = \{A \in V : A \to \varepsilon \text{ is a rule}\}$

```
bool changed = true
  while (changed) {
     changed = false;
     foreach B in V \ N:
        if (B has a rule B -> X with X ∈ N*):
           N := N ∪ {B}
           changed = true;
  }
output N  // = final NULLABLE.
```

Step 2 explained in other notes.

Example: $S \to (S)S \mid \varepsilon$

There are 2 occurrences of $S$ in the first rule, so we need $2^2 = 4$ combos:

$G_1 = \begin{cases} \text{① The original rule } S \to (S)S \\ \text{② } S \to ()S \quad \text{deleting first occurrence} \\ \text{③ } S \to (S) \quad \text{deleting 2nd occurrence} \\ \text{④ } S \to () \quad \text{deleting both.} \end{cases}$

Theorem: $G_1$ derives all the nonempty balanced parentheses strings

Second Step: "Bypass" unit rules $A \to B$.

(i) Build the directed graph $H$ of edges $(A, B)$ for all unit rules in $G$.

We want to identify cases where $A \to B \to C$.

(ii) Take $H^*$ to be the transitive closure of $H$

(iii) including "Gen-1" $(A, B)$ cases
For each $(A, C) \in H^*$, and all rules $C \to \vec{X}$, add $A \to \vec{X}$ as a rule. Sound.

(iv) delete all unit rules (incl any new ones)

Comprehensive because added $A \to \vec{X}$ rules cover all situations that had unit steps.

The resulting CFG $G_2$ has $L(G_2) = L(G_1) = L(G) \setminus \{\varepsilon\}$ and all rules have the form $A \to c$ or $A \to X$ with $X \in VVV^*$ ie. $|X| \geq 2$.

Final "silly" steps making RHS have length = 2 were already discussed orally. $\boxtimes$

# Unit Rules Example:

$S \to AB \mid cA \mid \varepsilon$   $A \to SS \mid Sa$   $B \to AS \mid c$

NULLABLE = {A, B, S}   Grammar $G_1$ is:

$G_1$: $S \to AB \mid A \mid B \mid cA \mid c$   $A \to SS \mid S \mid Sa \mid a$

$B \to AS \mid A \mid S \mid c$

Graph H


Transitive closure adds (A, B).

$G_2$: $S \to AB \mid cA \mid c \mid SS \mid Sa \mid a \mid AS \mid$ already have c

A and B get exactly the same here. "YUCK!"

We still need aliases $X_c \to c$  $X_a \to a$ here for ChNF. Yuck$^2$!