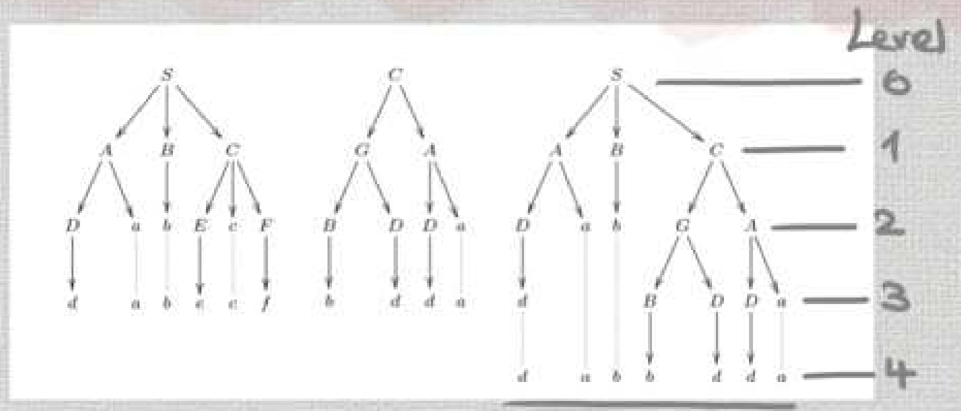
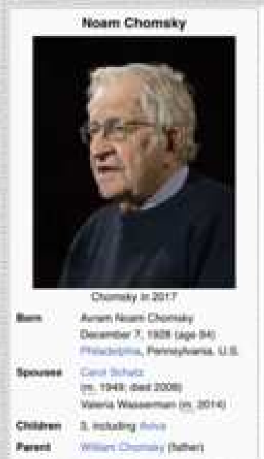
XI Automata & Formal Languages

Eleventh Lecture — 7 November 2024

Trees

grafting





Chomsky Normal Form CNF

$A \longrightarrow BC$ BINARY RULES

$A \longrightarrow a$ TERMINAL RULES

L 3.3 If G in CNF, $w \in \mathcal{L}(G)$, then every G-derivation of w has length $2|w|-1$.

L 3.4 If G in CNF, $\mathbb{T}$ a G-parse tree of height $h+1$, then

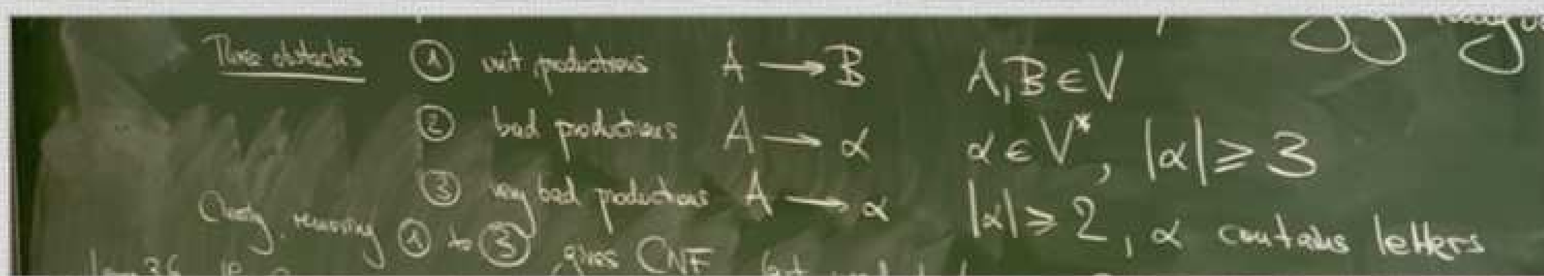
$$|\sigma_{\mathbb{T}}| \leq 2^h.$$

T 3.5 (Chomsky) If G context-free, then there is G' in CNF s.t.

$$\mathcal{L}(G) = \mathcal{L}(G').$$

PROOF OF T 3.5

What happened so far?



Three obstacles:

(1) unit productions $A \longrightarrow B$

(2) bad productions $A \longrightarrow \alpha$
$\alpha \in V^*, |\alpha| \geq 3$

(3) very bad productions $A \longrightarrow \alpha$
$|\alpha| \geq 2$, α contains letters

L 3.6 Can remove very bad productions

<u>Remark</u> May add unit and bad productions.

G unit-closed if for $A \to B$, $B \longrightarrow \alpha \in P$, also $A \to \alpha \in P$.

L 3.7 Can form the unit closure

<u>Remark</u> May add unit and bad productions.
Only adds very bad productions if the original grammar had some.

L 3.8 If G context-free and unit closed, then removing unit productions doesn't change the language.

Lemma 3.9 If $A \longrightarrow \alpha = A_0 \dots A_n$ is a bad production $(n \geq 2)$

define $P_{A,\alpha} := \{ A \to A_0 X_0,\ X_0 \longrightarrow A_1 X_1,\ \dots,\ X_{n-3} \longrightarrow A_{n-2} X_{n-2},\ X_{n-2} \longrightarrow A_{n-1} A_n \}$

Introduce new variables $X_0, \dots, X_{n-2}$:

if $V' := V \cup \{X_0, \dots, X_{n-2}\}$

and $P' := (P \cup P_{A,\alpha}) \setminus \{A \longrightarrow \alpha\}$

and $G' = (\Sigma, V', S, P')$, then $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof Clear. $\mathcal{L}(G) \subseteq \mathcal{L}(G')$.

In order to see $\mathcal{L}(G') \subseteq \mathcal{L}(G)$. If $S \xrightarrow{G'} w$ and the derivation has no X-variables, then $S \xrightarrow{G} w$. If some X occurs, all of the new rules were used in precisely the natural order. Reorder and replace the string of new rules with $A \longrightarrow \alpha$ to get a G-derivation. q.e.d

Remark The construction in L 3.9 produces no very bad, bad, or unit rules

PROOF OF CHOMSKY'S THM:

Start from G context-free.

Apply L 3.6 to remove very bad rules.

Apply L 3.7 to produce unit closure (note: no new very bad rules).

Apply L 3.8 to remove unit rules

For each bad rule remaining, apply L 3.9 to remove it.

q.e.d

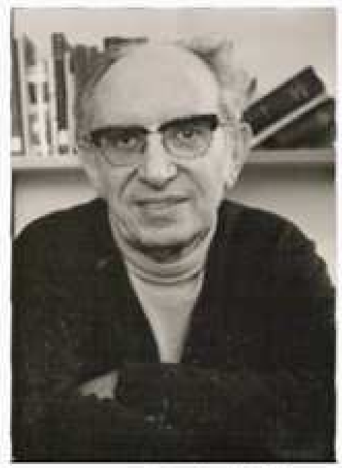
Remark: Number of variables goes up considerably

3.3 The pumping lemma for context-free languages

Definition 3.10. Let $L \subseteq \mathbb{W}$ be a language. We say that *L satisfies the (context-free) pumping lemma with pumping number n* if for every word $w \in L$ such that $|w| \geq n$ there are words u, v, x, y, z such that $w = xuyvz$, $|uv| > 0$, $|uyv| \leq n$ and for all $k \in \mathbb{N}$, we have that $xu^k yv^k z \in L$. We say that L *satisfies the (context-free) pumping lemma* if there is some n such that it satisfies the (context-free) pumping lemma with pumping number n.

The first proof of the context-free pumping lemma is usually attributed to Yehoshua Bar-Hillel (1915–1975); the statement is therefore also known as the *Bar-Hillel Lemma.*[12]

Yehoshua Bar-Hillel



Born	September 8, 1915 Vienna, Austria-Hungary
Died	September 25, 1975 (aged 60) Jerusalem, Israel

$w = xuyvz$

$|uv| > 0$

$|uyv| \leq n$

$\forall k \quad xu^k y v^k z \in L.$

Proposition 3.11. Every language that satisfies the (regular) pumping lemma satisfies the (context-free) pumping lemma.

Proof. If $w = xuz$ with $|u| > 0$ and $|xu| \leq n$, then let $y := \varepsilon$ and $v := \varepsilon$. Clearly, $|uv| \geq |u| > 0$ and $|uyv| = |u\varepsilon\varepsilon| = |u| \leq |xu| \leq n$ and $xu^k yv^k z = xu^k \varepsilon\varepsilon^k z = xu^k z$. Q.E.D.

L satisfies CFPL w PN n if

$\forall w \in L$ $|w| \geq n$ then there are x, u, y, v, z with

1. $w = xuyvz$
2. $|uv| > 0$
3. $|uyv| \leq n$

s.t. $\forall k$ $xu^k y v^k z \in L$.

RPL

$w = xuz$

$|u| > 0$

$|xu| \leq n$

$xu^k z \in L$

Clearly, if L satisfies RPL, then CFPL.

[Just let $y := \varepsilon$, $v := \varepsilon$.]

<u>Theorem</u> Every context-free language satisfies the CFPL with some pumping number n.

Proof. WLOG by Chomsky, G is in CNF. Consider $m := |V|$ and define $n := 2^m + 1$. Claim: n is the PN of G.

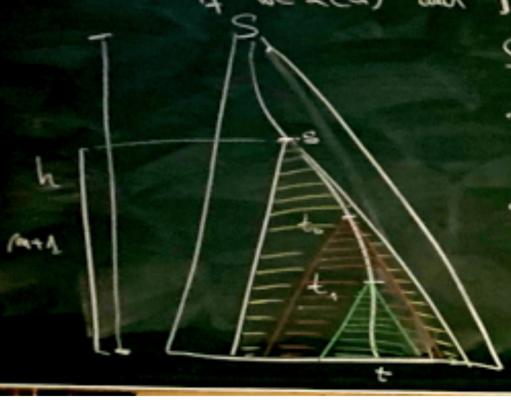
If $w \in \mathcal{L}(G)$ and Π is a G-parse tree for w, then the height h of Π is at least $m+1$.

So, there is a terminal node t s.t. height of t is $\geq m+1$.

Find s on the path from root to t s.t. the height of Π_s is exactly $m+1$.

The sequence leading from s to t has $m+2$ nodes, and the first $m+1$ are labelled with variables and only t has $\ell(t) \in \Sigma$.

By PHP, there are t_0, t_1; $t_0 \neq t_1$ s.t. $\ell(t_0) = \ell(t_1)$.



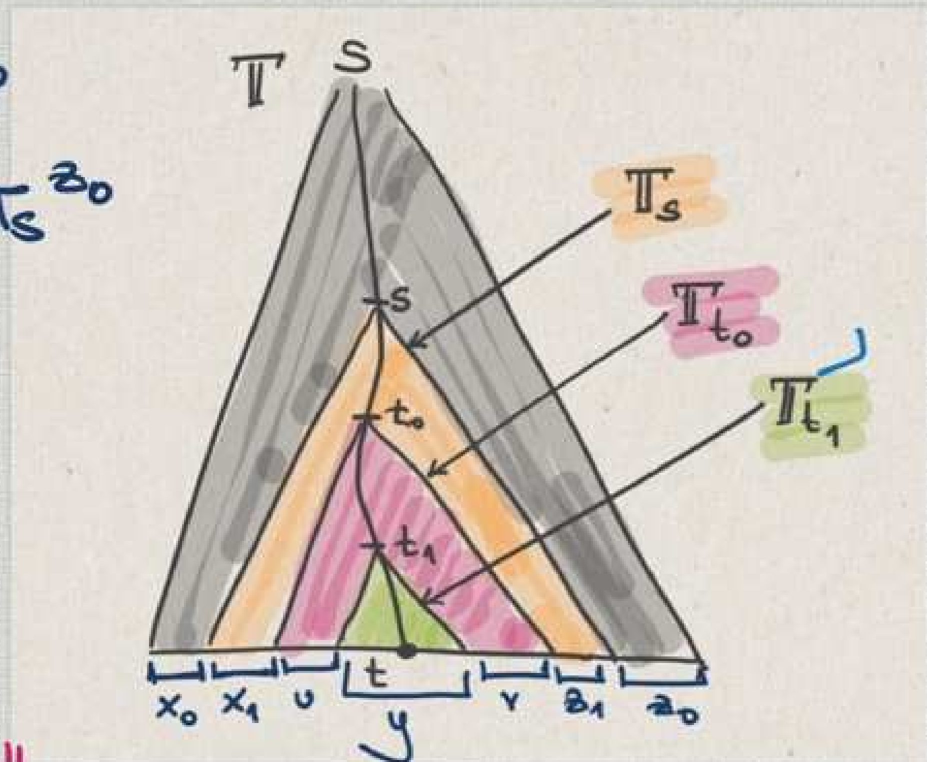
$$w = x_0 x_1 u y v z_1 z_0$$

$$\sigma_{\mathbb{T}} = w = x_0 \, \sigma_{\mathbb{T}_s} \, z_0$$

$$\sigma_{\mathbb{T}_s} = x_1 \, \sigma_{\mathbb{T}_{t_0}} \, z_1$$

$$\sigma_{\mathbb{T}_{t_0}} = u \, \sigma_{\mathbb{T}_{t_1}} \, v$$

$$\sigma_{\mathbb{T}_{t_1}} = y$$



Homework: Check that all seven of them can be empty.

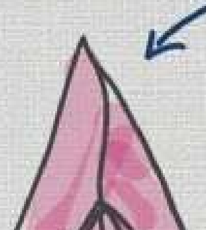
Since $t_0 \neq t_1$, we know that $|uv| > 0$.

By choice of s, $|x_1 u y v z_1| \leq n \implies |uyv| \leq n$.

$x := x_0 x_1 \qquad z := z_1 z_0$.

$\mathbb{T}_{t_0}$ $\quad$ $\mathbb{T}_{t_1}$

Parse trees starting with the same variable.



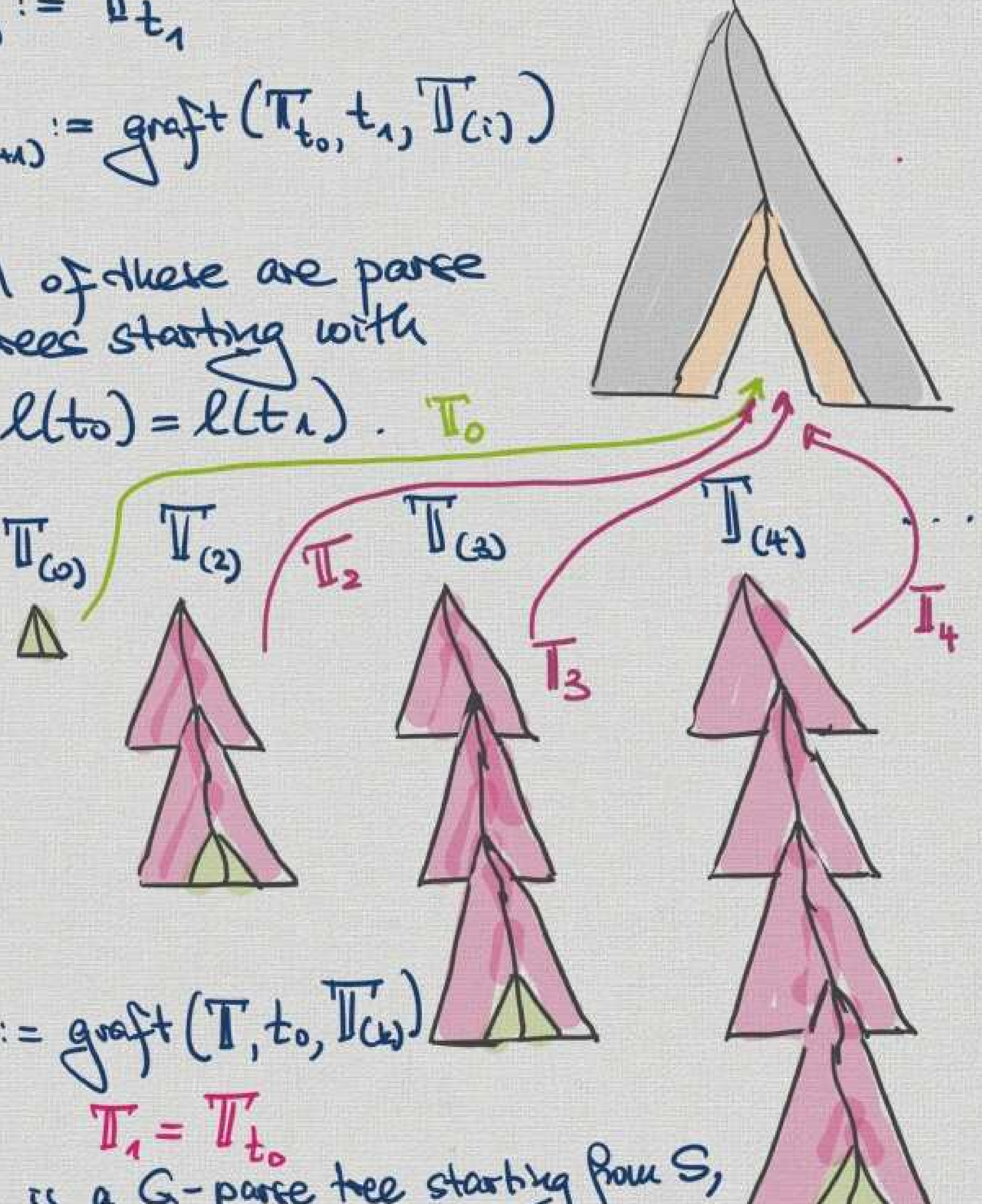
This one graft($\mathbb{T}_{t_0}, t_1, \mathbb{T}_{t_0}$) is another parse tree starting with the same symbol.

So, you can graft $\mathbb{T}_{t_0}$ into t_1

$\mathbb{T}_{(0)} := \mathbb{T}_{t_1}$

$\mathbb{T}_{(i+1)} := \text{graft}(\mathbb{T}_{t_0}, t_1, \mathbb{T}_{(i)})$

All of these are parse trees starting with $\ell(t_0) = \ell(t_1)$.



$\mathbb{T}_k := \text{graft}(\mathbb{T}, t_0, \mathbb{T}_{(k)})$

$\mathbb{T}_1 = \mathbb{T}_{t_0}$

This is a G-parse tree starting from S, so $\sigma_{\mathbb{T}_k} \in L$.

$\sigma_{\mathbb{T}_k} = x_0 x_1 u^k y v^k z_1 x_1 = x u^k y v^k z$. qed.