X

Tenth Lecture 5 NOVEMBER 2024

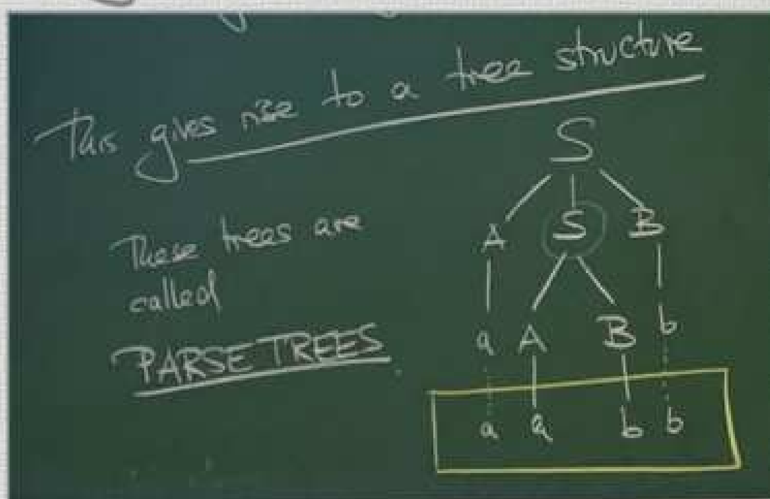
AUTOMATA & FORMAL LANGUAGES

RECAP

Chapter 3
CONTEXT-FREE LANGUAGES

Context-free rules $A \longrightarrow \alpha$
can be applied independently from context

$\rightsquigarrow$ tree structure



PARSE TREES

$\mathbb{T} = (T, \ell)$

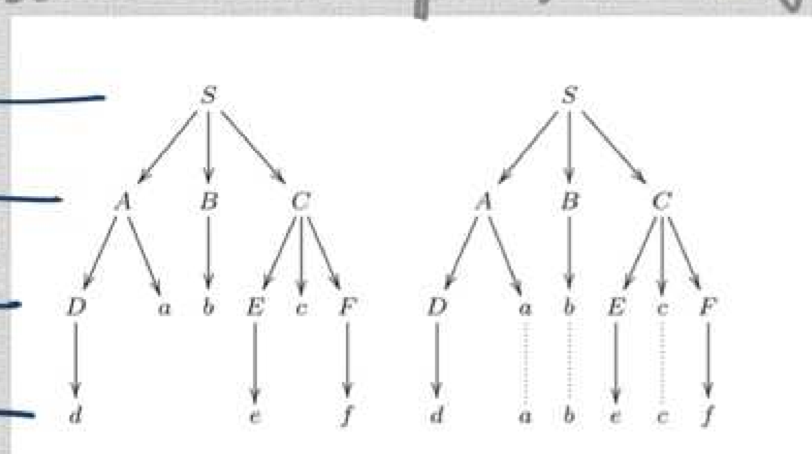
Leaves marked with letters
Non-terminal nodes marked with variables

G-parse tree if all of the branchings correspond to production rules in G.

Read off the produced word $\sigma_{\mathbb{T}}$ left-to-right.

Level 0
1
2
3

Tree of height 3.



RECAP

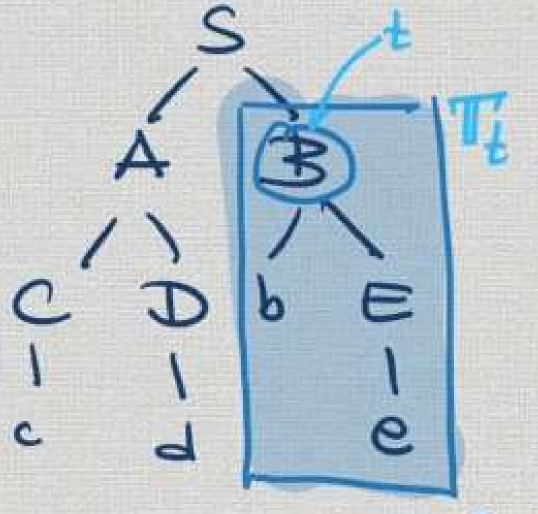
PROPOSITION 3.2

Theorem If G is context-free then TFAE

(i) $w \in \mathcal{L}(G)$

(ii) there is a G-parse tree $\mathbb{T}$ starting from S s.t. $\sigma_{\mathbb{T}} = w$.

If $\mathbb{T}$ parse tree, t is non-terminal node, then $\mathbb{T}_t$ is the subtree starting at t:

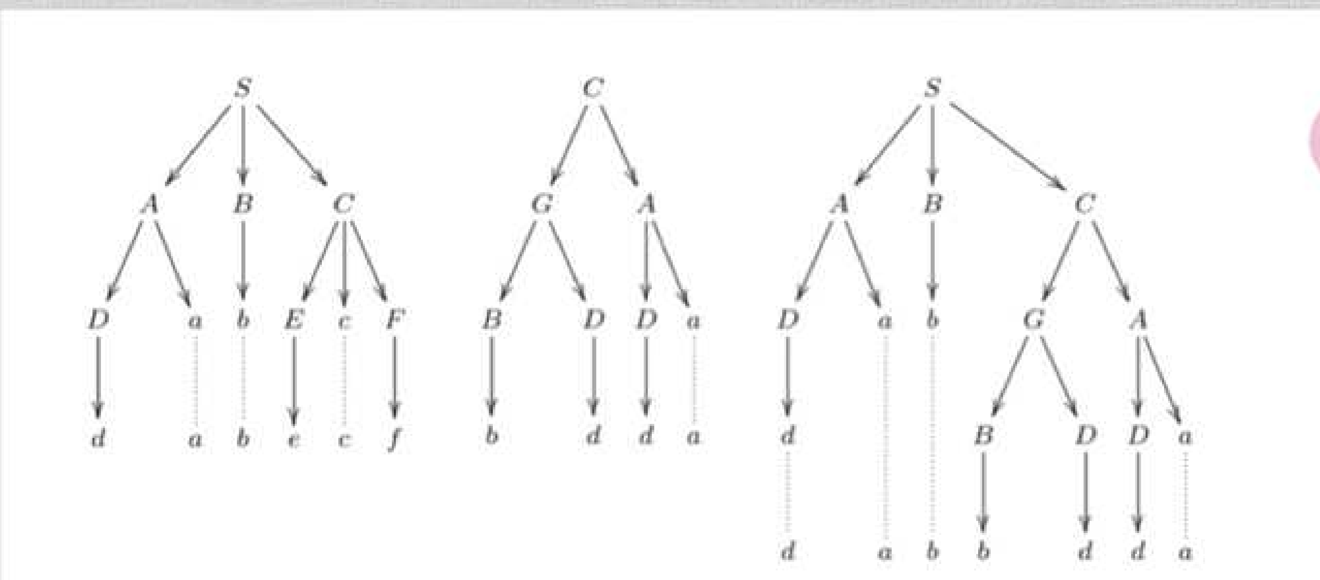


If $w = \sigma_{\mathbb{T}}$, then

$$w = v\sigma_{\mathbb{T}_t}v'.$$

Note that if $\ell(t) = B$, then $\mathbb{T}_t$ is a parse tree starting from B.

Grafting



If $\mathbb{T}$ G-parse tree, t node $\ell(t) = A$
and $\mathbb{T}'$ G-parse tree starting from A,
I can cut out $\mathbb{T}_t$ from $\mathbb{T}$ and
graft $\mathbb{T}'$ into the gap.

$$\sigma_{\mathbb{T}} = v\sigma_{\mathbb{T}_t} v'$$

$\mathbb{T}^* := \text{graft}(\mathbb{T}, t, \mathbb{T}')$;
this is a G-parse tree.

[Context-freeness guarantees this.]

$$\boxed{\sigma_{\mathbb{T}^*} = v\sigma_{\mathbb{T}'} v'.}$$

TODAY:
Chomsky's
Theorem

Chomsky in 2017

§3.2 Chomsky normal form

<u>Motivation</u> If G is regular and $w \in \mathcal{L}(G)$, every derivation of w has length $|w|$.

G is said to be in <u>Chomsky normal form</u> (CNF) if all rules are either

$A \longrightarrow a$ <u>terminal rules</u>

or $A \longrightarrow BC$ $A, B, C \in V$ <u>binary rules</u>

Every CNF grammar is context free.

<u>Lemma 3.3</u> If G is CNF & $w \in \mathcal{L}(G)$. Then every G-der. of w has length $2|w|-1$.

<u>Pf</u>

	Length	$\#V$	$\#\Sigma$
Terminal	0	-1	$+1$
Binary	$+1$	$+1$	0

Thus we used $|w|-1$ binary rules to get $\alpha \in V^{|w|}$ and then $|w|$ terminal rules to get w. q.e.d.

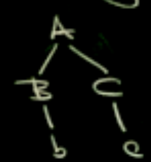
Parse trees of CNF grammars

Only two situations of branching in the tree.

binary: 

terminal 

Height 1 

Height 2 

Height 3 

We observe that if T is a G-parse tree, the number of terminal nodes is at most 2^{h-1} where h is the height of the tree. [Lemma 3.4]

Chomsky's Theorem (T 3.5)

If G is context-free, then there is G' CNF s.t. $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof idea List all possible violations of CNF and remove them w/o changing language.

Three obstacles
1. unit productions $A \longrightarrow B$ $\quad A, B \in V$
2. bad productions $A \longrightarrow \alpha$ $\quad \alpha \in V^*, |\alpha| \geq 3$
3. very bad productions $A \longrightarrow \alpha$ $\quad |\alpha| \geq 2$, α contains letters

Clearly, removing ① to ③ gives CNF, but need to be careful not to change language.

Lemma 3.6 If G is context-free, there is G' without very bad productions s.t. $\mathcal{L}(G) = \mathcal{L}(G')$.

Proof Trick from Lemma 1.24.

For each $a \in \Sigma$, add variable X_a. $\quad V' := V \cup \{X_a ; a \in \Sigma\}$

Write $X(\alpha)$ for α where all $a \in \Sigma$ are replaced by X_a

$P^* := P \cup \{A \to X(\alpha);\ A \to \alpha \in P\} \cup \{X_a \to a;\ a \in \Sigma\}$

$G^* = (\Sigma, V', S, P^*)$. As before, $\mathcal{L}(G^*) = \mathcal{L}(G)$.

$P' := P^* \setminus \{A \to \alpha;\ A \to \alpha \text{ is very bad}\}$

Since every very bad rule can be emulated by new rules, $G' = (\Sigma, V', S, P')$ produces the same words as G^* (and G)

qed

<u>Remark</u> G' may have added additional unit rules or bad rules.

<u>Def.</u> G is <u>unit closed</u> if whenever $A \to B \in P$, $B \to \alpha \in P$, then $A \to \alpha \in P$.

<u>Lemma 37</u> If G is context free, there is a unit closed G' s.t. $\mathcal{L}(G) = \mathcal{L}(G')$.

Pf. Quite obvious.

close under the closure operation.

Note that you need to do this iteratively.

$$A \to B$$
$$B \to C$$
$$C \to \alpha$$

This process will end in at most $|V||P|$ steps. (Homework!)

q.e.d.

Remark The process will only add very bad rules if G had very bad rules. So if G has no very bad rules, then so does G'.

Lemma 3.8 If G is context free and unit closed and

$$P' := P \setminus \{A \to B;\ A \to B \in P\}$$

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

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

We'll show that if $S \xrightarrow{G} w$, then $S \xrightarrow{G'} w$.

We show this by showing: if $S \xrightarrow{G} w$ uses unit productions, then there is a shorter derivation.

[The shortest one is a G'-derivation.]

$$S \xrightarrow{G} w$$

$$S \xrightarrow{G} \underbrace{\alpha A\beta \xrightarrow{G} \alpha B\beta}_{\text{uses } \boxed{A \to B}} \xrightarrow{G} \underbrace{\gamma B\delta \xrightarrow{G} \gamma\xi\delta}_{\text{uses } \boxed{B \to \xi}} \xrightarrow{G} w \qquad (*)$$

suppose no unit production

Since B is untouched here
we get $\alpha A\beta \xrightarrow{G} \gamma A\delta$

By unit closure $A \to \xi \in P$.

$$S \xrightarrow{G} \alpha A\beta \xrightarrow{G} \gamma A\delta \xrightarrow{G} \gamma\xi\delta \xrightarrow{G} w$$

This is precisely one step shorter than (*)

q.e.d.