

# Theory of Computer Science

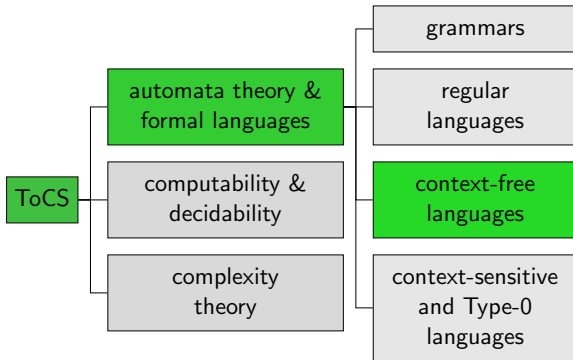
## B8. Context-free Languages: $\epsilon$ -Rules & Chomsky Normal Form

Gabriele Röger

University of Basel

March 26, 2025

# Content of the Course



# Context-free Grammars and $\varepsilon$ -Rules

# Repetition: Context-free Grammars

## Definition (Context-free Grammar)

A **context-free grammar** is a 4-tuple  $\langle V, \Sigma, P, S \rangle$  with

- 1  $V$  finite set of variables,
- 2  $\Sigma$  finite alphabet of terminal symbols (with  $V \cap \Sigma = \emptyset$ ),
- 3  $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \epsilon \rangle\}$  finite set of rules,
- 4 If  $S \rightarrow \epsilon \in P$ , then all other rules in  $V \times ((V \setminus \{S\}) \cup \Sigma)^+$ .
- 5  $S \in V$  start variable.

## Context-free Grammars: Exercise

We have used the pumping lemma for regular languages to show that  $L = \{a^n b^n \mid n \in \mathbb{N}_0\}$  is not regular.

Show that it is context-free by specifying a suitable grammar  $G$  with  $\mathcal{L}(G) = L$ .



# Repetition: Context-free Grammars

## Definition (Context-free Grammar)

A **context-free grammar** is a 4-tuple  $\langle V, \Sigma, P, S \rangle$  with

- 1  $V$  finite set of variables,
- 2  $\Sigma$  finite alphabet of terminal symbols (with  $V \cap \Sigma = \emptyset$ ),
- 3  $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \epsilon \rangle\}$  finite set of rules,
- 4 If  $S \rightarrow \epsilon \in P$ , then all other rules in  $V \times ((V \setminus \{S\}) \cup \Sigma)^+$ .
- 5  $S \in V$  start variable.

## Repetition: Context-free Grammars

### Definition (Context-free Grammar)

A **context-free grammar** is a 4-tuple  $\langle V, \Sigma, P, S \rangle$  with

- 1  $V$  finite set of variables,
- 2  $\Sigma$  finite alphabet of terminal symbols (with  $V \cap \Sigma = \emptyset$ ),
- 3  $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \epsilon \rangle\}$  finite set of rules,
- 4 If  $S \rightarrow \epsilon \in P$ , then all other rules in  $V \times ((V \setminus \{S\}) \cup \Sigma)^+$ .
- 5  $S \in V$  start variable.

Rule  $X \rightarrow \epsilon$  is only allowed if  $X = S$   
and  $S$  never occurs on a right-hand side.

## Repetition: Context-free Grammars

### Definition (Context-free Grammar)

A **context-free grammar** is a 4-tuple  $\langle V, \Sigma, P, S \rangle$  with

- 1  $V$  finite set of variables,
- 2  $\Sigma$  finite alphabet of terminal symbols (with  $V \cap \Sigma = \emptyset$ ),
- 3  $P \subseteq (V \times (V \cup \Sigma)^+) \cup \{\langle S, \epsilon \rangle\}$  finite set of rules,
- 4 If  $S \rightarrow \epsilon \in P$ , then all other rules in  $V \times ((V \setminus \{S\}) \cup \Sigma)^+$ .
- 5  $S \in V$  start variable.

Rule  $X \rightarrow \epsilon$  is only allowed if  $X = S$   
and  $S$  never occurs on a right-hand side.

With regular grammars, this restriction could be lifted.  
How about context-free grammars?



## Reminder: Start Variable in Right-Hand Side of Rules

For every type-0 language  $L$  there is a grammar where the start variable does not occur on the right-hand side of any rule.

### Theorem

*For every grammar  $G = \langle V, \Sigma, P, S \rangle$  there is a grammar  $G' = \langle V', \Sigma, P', S \rangle$  with rules  $P' \subseteq (V' \cup \Sigma)^+ \times (V' \setminus \{S\} \cup \Sigma)^*$  such that  $\mathcal{L}(G) = \mathcal{L}(G')$ .*

In the proof we constructed a suitable grammar, where the rules in  $P'$  were not fundamentally different from the rules in  $P$ :

- for rules from  $V \times (V \cup \Sigma)^+$ , we only introduced additional rules from  $V' \times (V' \cup \Sigma)^+$ , and
  - for rules from  $V \times \epsilon$ , we only introduced rules from  $V' \times \epsilon$ ,
- where  $V' = V \cup \{S'\}$  for some new variable  $S' \notin V$ .

## $\epsilon$ -Rules

### Theorem

*For every grammar  $G$  with rules  $P \subseteq V \times (V \cup \Sigma)^*$   
there is a context-free grammar  $G'$  with  $\mathcal{L}(G) = \mathcal{L}(G')$ .*

## $\varepsilon$ -Rules

### Theorem

For every grammar  $G$  with rules  $P \subseteq V \times (V \cup \Sigma)^*$   
there is a context-free grammar  $G'$  with  $\mathcal{L}(G) = \mathcal{L}(G')$ .

### Proof.

Let  $G = \langle V, \Sigma, P, S \rangle$  be a grammar with  $P \subseteq V \times (V \cup \Sigma)^*$ .

Let  $G' = \langle V', \Sigma, P', S \rangle$  be a grammar with  $\mathcal{L}(G) = \mathcal{L}(G')$  with  $P' \subseteq V' \times ((V' \setminus S) \cup \Sigma)^*$ .

Let  $V_\varepsilon = \{A \in V' \mid A \Rightarrow_{G'}^* \varepsilon\}$ . We can find this set  $V_\varepsilon$  by first collecting all variables  $A$  with rule  $A \rightarrow \varepsilon \in P'$  and then successively adding additional variables  $B$  if there is a rule  $B \rightarrow A_1 A_2 \dots A_k \in P'$  and the variables  $A_i$  are already in the set for all  $1 \leq i \leq k$ .

...

## $\varepsilon$ -Rules

### Theorem

For every grammar  $G$  with rules  $P \subseteq V \times (V \cup \Sigma)^*$   
there is a context-free grammar  $G'$  with  $\mathcal{L}(G) = \mathcal{L}(G')$ .

### Proof (continued).

Let  $P''$  be the rule set that is constructed from  $P'$  by

- adding rules that obviate the need for  $A \rightarrow \varepsilon$  rules:  
for every existing rule  $B \rightarrow w$  with  $B \in V'$ ,  $w \in (V' \cup \Sigma)^+$ ,  
let  $I_\varepsilon$  be the set of positions where  $w$  contains a variable  
 $A \in V_\varepsilon$ . For every non-empty set  $I' \subseteq I_\varepsilon$ , add a new rule  
 $B \rightarrow w'$ , where  $w'$  is constructed from  $w$  by removing  
the variables at all positions in  $I'$ .
- removing all rules of the form  $A \rightarrow \varepsilon$  ( $A \neq S$ ).

Then  $G'' = \langle V', \Sigma, P'', S \rangle$  is context-free and  $\mathcal{L}(G) = \mathcal{L}(G'')$ . □

# Example

Consider  $G = \langle \{X, Y, Z, S\}, \{a, b\}, R, S \rangle$  with rules:

$$S \rightarrow \epsilon \mid XY$$

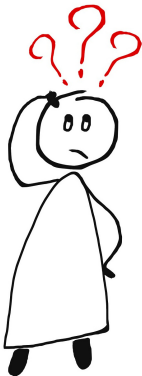
$$X \rightarrow aXYbX \mid YZ$$

$$Y \rightarrow \epsilon \mid b$$

$$Z \rightarrow \epsilon \mid a$$

↪ blackboard

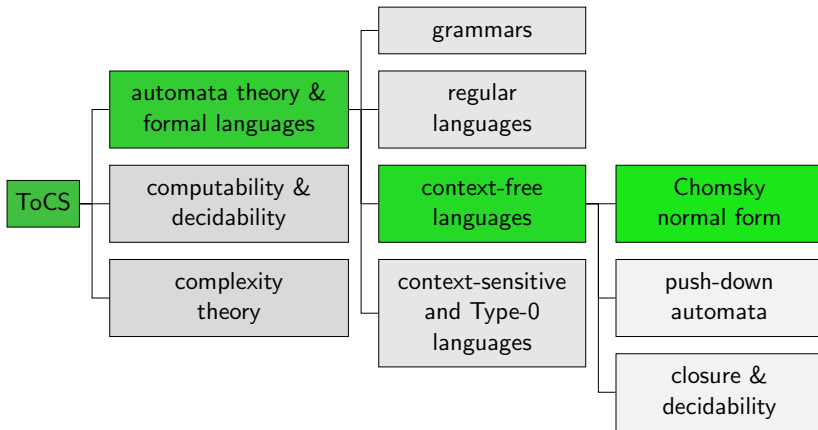
# Questions



Questions?

# Chomsky Normal Form

# Content of the Course





# Chomsky Normal Form: Motivation

As in logical formulas (and other kinds of structured objects), **normal forms** for grammars are useful:

- they show which aspects are critical for defining grammars and which ones are just syntactic sugar
- they allow proofs and algorithms to be restricted to a limited set of grammars (inputs): those in normal form

Hence we now consider a **normal form** for context-free grammars.

# Chomsky Normal Form: Definition

## Definition (Chomsky Normal Form)

A context-free grammar  $G$  is in **Chomsky normal form** (**CNF**) if all rules have one of the following three forms:

- $A \rightarrow BC$  with variables  $A, B, C$ , or
- $A \rightarrow a$  with variable  $A$ , terminal symbol  $a$ , or
- $S \rightarrow \epsilon$  with start variable  $S$ .

in short:

rule set  $P \subseteq (V \times (V'V' \cup \Sigma)) \cup \{\{S, \epsilon\}\}$  with  $V' = V \setminus \{S\}$

German: Chomsky-Normalform

# Chomsky Normal Form: Theorem

## Theorem

*For every context-free grammar  $G$  there is a context-free grammar  $G'$  in Chomsky normal form with  $\mathcal{L}(G) = \mathcal{L}(G')$ .*

# Chomsky Normal Form: Theorem

## Theorem

*For every context-free grammar  $G$  there is a context-free grammar  $G'$  in Chomsky normal form with  $\mathcal{L}(G) = \mathcal{L}(G')$ .*

## Proof.

The following algorithm converts the rule set of  $G$  into CNF:

**Step 1: Eliminate rules of the form  $A \rightarrow B$**  with variables  $A, B$ .

If there are sets of variables  $\{B_1, \dots, B_k\}$  with rules

$B_1 \rightarrow B_2, B_2 \rightarrow B_3, \dots, B_{k-1} \rightarrow B_k, B_k \rightarrow B_1,$

then replace these variables by a new variable  $B$ .

Define a strict total order  $<$  on the variables such that  $A \rightarrow B \in P$  implies that  $A < B$ . Iterate from the largest to the smallest variable  $A$  and eliminate all rules of the form  $A \rightarrow B$  while adding rules  $A \rightarrow w$  for every rule  $B \rightarrow w$  with  $w \in (V \cup \Sigma)^+$ . ...

# Chomsky Normal Form: Theorem

## Theorem

*For every context-free grammar  $G$  there is a context-free grammar  $G'$  in Chomsky normal form with  $\mathcal{L}(G) = \mathcal{L}(G')$ .*

## Proof (continued).

**Step 2: Eliminate rules with terminal symbols on the right-hand side that do not have the form  $A \rightarrow a$ .**

For every terminal symbol  $a \in \Sigma$  add a new variable  $A_a$  and the rule  $A_a \rightarrow a$ .

Replace all terminal symbols in all rules that do not have the form  $A \rightarrow a$  with the corresponding newly added variables. ...

# Chomsky Normal Form: Theorem

## Theorem

For every context-free grammar  $G$  there is a context-free grammar  $G'$  in Chomsky normal form with  $\mathcal{L}(G) = \mathcal{L}(G')$ .

## Proof (continued).

**Step 3:** Eliminate rules of the form  $A \rightarrow B_1 B_2 \dots B_k$  with  $k > 2$

For every rule of the form  $A \rightarrow B_1 B_2 \dots B_k$  with  $k > 2$ , add new variables  $C_2, \dots, C_{k-1}$  and replace the rule with

$$A \rightarrow B_1 C_2$$

$$C_2 \rightarrow B_2 C_3$$

$$\vdots$$

$$C_{k-1} \rightarrow B_{k-1} B_k$$



# Example

Consider  $G = \langle \{Y, Z, S\}, \{a, b\}, R, S \rangle$  with rules:

$$S \rightarrow aZbY \mid Y \mid ab$$

$$Y \rightarrow Z \mid b$$

$$Z \rightarrow Y \mid bSa$$

→ blackboard

# Chomsky Normal Form: Length of Derivations

## Observation

Let  $G$  be a grammar in Chomsky normal form,  
and let  $w \in \mathcal{L}(G)$  be a non-empty word generated by  $G$ .  
Then all derivations of  $w$  have exactly  $2|w| - 1$  derivation steps.

Why?



# Questions



Questions?

# Summary

# Summary

- The restriction of  $\epsilon$ -occurrences in rules is not necessary to characterize the set of context-free languages.
- Every context-free language has a grammar in **Chomsky normal form**. All rules have form
  - $A \rightarrow BC$  with variables  $A, B, C$ , or
  - $A \rightarrow a$  with variable  $A$ , terminal symbol  $a$ , or
  - $S \rightarrow \epsilon$  with start variable  $S$ .