We could consider rightmost derivations instead of leftmost derivations. Theorem 3.1 also holds for rightmost derivations.

Thus, for each word $w \in L(G, A)$, there are as many leftmost derivations as there are rightmost derivations.

A context-free grammar $G = (N, T, S, P)$ is called proper, if it satisfies the following conditions:

(1) $\forall A \in N : L(G, A) = \{ w \in T^* \mid A \to_P^* w \} \neq \emptyset$, und

(2) $\forall A \in N \, \exists u, v \in T^* : S \to_P^* uAv$.

Thus, in a proper context-free grammar, there are no useless nonterminals.

Lemma 3.3

It is decidable whether a given context-free grammar $G = (N, T, S, P)$ is proper.

Proof of Lemma 3.3.

Inductively, we first determine the set of nonterminals $V_{\text{term}} = \{ A \in N \mid L(G, A) \neq \emptyset \}$:

$$\begin{aligned} V_1 &:= \{ A \in N \mid \exists x \in T^* : (A \to x) \in P \}; \\ V_{i+1} &:= V_i \cup \{ A \in N \mid \exists \alpha \in (V_i \cup T)^* : (A \to \alpha) \in P \} \ (i \geq 1). \end{aligned}$$

Then $V_{\text{term}} = \bigcup_{i \geq 1} V_i = V_{|N|}$.

Next, we inductively determine the set of reachable nonterminals $V_{\text{reach}} = \{ A \in N \mid \exists \alpha, \beta \in (V_{\text{term}} \cup T)^* : S \to_P^* \alpha A \beta \}$:

$$\begin{aligned} U_1 &:= \{ A \in V_{\text{term}} \mid \exists \alpha, \beta \in (V_{\text{term}} \cup T)^* : (S \to \alpha A \beta) \in P \}; \\ U_{i+1} &:= U_i \cup \{ A \in V_{\text{term}} \mid \exists B \in U_i \, \exists \alpha, \beta \in (V_{\text{term}} \cup T)^* : \\ &\qquad (B \to \alpha A \beta) \in P \} \ (i \geq 1). \end{aligned}$$

Then $V_{\text{reach}} = \bigcup_{i \geq 1} U_i = U_{|N|}$.

Now G is proper iff $V_{\text{reach}} = V_{\text{term}} = N$. □

Lemma 3.4

From any given context-free grammar $G = (N, T, S, P)$, one can construct a proper context-free grammar $G' = (N', T, S, P')$ such that $L(G') = L(G)$.

Proof.

Just take $N' = V_{\text{reach}}$ and

$$P' = \{ (A \to \alpha) \in P \mid A \in V_{\text{reach}} \text{ and } \alpha \in (V_{\text{reach}} \cup T)^* \}.$$

Then G' is a proper context-free grammar and $L(G') = L(G)$. □

A production $(\ell \to r) \in P$ is called terminal if $r \in T^*$.
It is called an ε-production if $r = \varepsilon$.

If $\varepsilon \in L(G)$, then G must contain at least one ε-production.
In fact, it can be required that $(S \to \varepsilon)$ is the only ε-production in G.

Lemma 3.5

From a given context-free grammar $G = (N, T, S, P)$, one can construct an equivalent context-free grammar $G' = (N', T, S', P')$ such that S' does not occur on the righthand side of any production in P', and G' contains no ε-production with the only possible exception of $(S' \to \varepsilon)$. In fact, $(S' \to \varepsilon) \in P'$ iff $\varepsilon \in L(G) = L(G')$.

Proof.

First we consider the case that $\varepsilon \notin L(G)$.

Task: Elimination of all ε-productions.

(1.) Determine $V_1 := \{ A \in N \mid A \to_G^* \varepsilon \}$:

$$V_1^{(1)} := \{ A \in N \mid (A \to \varepsilon) \in P \};$$
$$V_1^{(i+1)} := V_1^{(i)} \cup \{ A \in N \mid \exists r \in V_1^{(i)*} : (A \to r) \in P \}.$$

Then $V_1 = \bigcup_{i \geq 1} V_1^{(i)} = V_1^{(|N|)}$.

(2.) Remove all ε-productions.

Proof of Lemma 3.5 (cont.)

(3.) $\forall B \to xAy$ such that $A \in V_1$ and $xy \neq \varepsilon$:
Introduce the new production $B \to xy$.

Then $G' = (N, T, S, P')$ does not contain any ε-productions and $L(G') = L(G)$.

If $\varepsilon \in L(G)$, then first apply the construction above, which yields a context-free grammar $G' = (N, T, S, P')$ for $L(G) \setminus \{\varepsilon\}$ without ε-productions.
Then introduce a new nonterminal S' and the productions $(S' \to \varepsilon)$ and $(S' \to S)$, where S' is taken as the new start symbol.
The resulting grammar $G'' = (N, T, S', P'')$ generates $L(G)$, $(S' \to \varepsilon)$ is the only ε-production in P'', and the start symbol S' does not occur on the righthand side of any production. □

Let $G = (N, T, S, P)$ be a context-free grammar.
A production $(A \to B) \in P$, where $A, B \in N$, is called a chain rule.

Example:

A grammar for arithmetic expressions with brackets:
$G = (N, T, E, P)$, where $N = \{E, T, F\}$, $T = \{a, +, -, *, /, (,)\}$, and P contains the following productions:

$E \to T \mid E + T \mid E - T$,
$T \to F \mid T * F \mid T/F$,
$F \to a \mid (E)$.

This grammar contains the chain rules $(E \to T)$ and $(T \to F)$.

Lemma 3.6

From a given context-free grammar $G = (N, T, S, P)$, one can construct an equivalent context-free grammar G' that does not contain any chain rules.

Proof of Lemma 3.6.

By Lemma 3.5, G does not contain any ε-productions.

We define an equiv. relation $\sim$ on N: $A \sim B$ iff $A \to_P^* B$ and $B \to_P^* A$.

Let $[A] = \{ B \in N \mid A \sim B \}$ denote the equivalence class of A.

For each $A \in N$, choose a unique representative $A_0 \in [A]$, replace all occurrences of all nonterminals $B \in [A]$ within P by A_0, and delete all productions of the form $(A_0 \to A_0)$.

Let $V = \{A_1, A_2, \ldots, A_n\}$ be the remaining nonterminals. W.l.o.g. we can assume for each remaining chain rule $(A_i \to A_j)$ that $i < j$.

For $k = n-1, n-2, \ldots, 2, 1$:
If $(A_k \to A_{k'}) \in P$ (where $k' > k$) and if $A_{k'} \to x_1 \mid x_2 \mid \ldots \mid x_m$ are all $A_{k'}$-productions, then replace $(A_k \to A_{k'})$ by $A_k \to x_1 \mid x_2 \mid \ldots \mid x_m$.
(By the I.H. $|x_i| \geq 2$ or $x_i \in T$, $i = 1, 2, \ldots, m$).

The resulting grammar is equivalent to G and it does not contain any chain-rules. □

Example (cont.):

$E \to T \mid E + T \mid E - T$
$T \to F \mid T * F \mid T/F$
$F \to a \mid (E)$

Remove the chain rules $(T \to F)$ and $(E \to T)$:

There are no equivalent nonterminals.

We order the set N through $E < T < F$:

$(T \to F)$ is replaced by: $T \to a \mid (E) \mid T * F \mid T/F$,

$(E \to T)$ is replaced by: $E \to a \mid (E) \mid T * F \mid T/F \mid E + T \mid E - T$.

Thus, the new grammar has 12 productions, while the given one had only 8.

Definition 3.7

Let T be a (terminal) alphabet, and let $\overline{T} := \{ \bar{a} \mid a \in T \}$ be a set of 'copies' of T such that $T \cap \overline{T} = \emptyset$.

The Dyck language D_T^* on T is generated by the grammar $G = (\{S, A\}, T \cup \overline{T}, S, P)$, where P contains the following productions:

$$P = \{ S \to AS, S \to \varepsilon, A \to aS\bar{a} \mid a \in T \}.$$

The words from the language $D_T := L(G, A)$ are called Dyck primes.

Examples:

$ab\bar{a}\bar{b}\bar{a}, a\bar{a}, ba\bar{a}\bar{b} \in D_T$ and $ab\bar{b}\bar{a}ba\bar{a}\bar{b}, \varepsilon \in D_T^* \setminus D_T$.

As $T = \{a, b\}$ contains just two letters, D_T and D_T^* are denoted as D_2 and D_2^*.

Corollary 3.8

REG $\subsetneq$ CFL.

Proof.

As observed before, REG $\subseteq$ CFL.

On the other hand,
$D_T^* \cap \{ a^n \bar{a}^m \mid n, m \geq 1 \} = \{ a^n \bar{a}^n \mid n \geq 1 \}$ is not regular.

The class REG is closed under intersection (Theorem 2.8).

If $D_T^* \in$ REG, then also $\{ a^n \bar{a}^n \mid n \geq 1 \} \in$ REG, a contradiction.

Thus, $D_T^* \in$ CFL $\setminus$ REG. $\square$

3.2. Normal Forms of Context-Free Grammars

A context-free grammar $G = (N, T, S, P)$ is in weak Chomsky Normal Form, if $r \in N^* \cup T \cup \{\varepsilon\}$ for each production $(\ell \to r) \in P$.

The grammar G is in Chomsky Normal Form (CNF), if $r \in N^2 \cup T \cup \{\varepsilon\}$ for each production $(\ell \to r) \in P$).

Theorem 3.9 (Chomsky 1959)

Given a context-free grammar G, one can effectively construct an equivalent context-free grammar G' that is in Chomsky Normal Form. In addition, it can be ensured that the start symbol S' of G' does not occur on the righthand side of any production and that G' does not contain any ε-production apart from possibly $(S' \to \varepsilon)$.

Proof of Theorem 3.9.

From the previous subsection we recall that we can construct a context-free grammar $G_1 = (N_1, T, S_1, P_1)$ from G such that $L(G_1) = L(G)$ and G_1 contains no chain rules and satisfies the condition on ε-productions.

Let $N_1' := \{ A_a \mid a \in T \}$ be a new alphabet of nonterminals.
We take $G_2 := (N_2, T, S_2, P_2)$, where

$$N_2 := N_1 \cup N_1', \ S_2 := S_1, \text{ and } P_2 := P_{2,1} \cup \{ A_a \to a \mid a \in T \}.$$

Here $P_{2,1}$ is obtained from P_1 by replacing in each righthand side r, where $|r| > 1$, each occurrence of a terminal symbol $a \in T$ by $A_a \in N_1'$.

Then P_2 contains three types of productions:

Proof of Theorem 3.9 (cont.).

(1.) possibly the ε-production $(S_2 \to \varepsilon)$,

(2.) the terminal productions of the form $(A \to b) \in P_1$, and the new terminal productions $(A_a \to a)$ $(a \in T)$,

(3.) nonterminal productions of the form $(A \to r)$, where $r \in N^*$ and $|r| \geq 2$.

Thus, G_2 is in weak Chomsky Normal Form.

If $(A \to B_1 B_2 \cdots B_m)$ is a nonterminal production s.t. $m > 2$, then we introduce new nonterminals $A^{(1)}, A^{(2)}, \ldots, A^{(m-2)}$, and we replace this production by the following ones:

$$(A \to B_1 A^{(1)}), (A^{(1)} \to B_2 A^{(2)}), \ldots, (A^{(m-2)} \to B_{m-1} B_m).$$

We obtain a grammar G' in CNF that is equivalent to G_2 and therewith to G. In addition, G' satisfies the condition on ε-productions. $\square$

Example:

Let $G = (\{S, A, B\}, \{a, b\}, S, P)$, where P is the following system:

$$P := \{S \to bA \mid aB, A \to bAA \mid aS \mid a, B \to aBB \mid bS \mid b\}.$$

Then G is a proper context-free grammar.

The first step yields the grammar $G_2 = (N_2, \{a, b\}, S, P_2)$, where $N_2 = \{S, A, B, A_a, A_b\}$ and

$$\begin{aligned} P_2 \quad = \quad & \{ S \to A_bA \mid A_aB, A \to A_bAA \mid A_aS \mid a, B \to A_aBB \mid A_bS \mid b, \\ & A_a \to a, A_b \to b \}, \end{aligned}$$

which is in weak CNF.

Example (cont.):

From G_2 we obtain the grammar $G' = (N', \{a, b\}, S, P')$, where $N' = \{S, A, B, A_a, A_b, C_1, C_2\}$ and

$$\begin{aligned} P' \;=\; \{ & S \to A_b A \mid A_a B,\; A \to A_b C_1 \mid A_a S \mid a,\; C_1 \to AA, \\ & B \to A_a C_2 \mid A_b S \mid b,\; C_2 \to BB, \\ & A_a \to a,\; A_b \to b\}. \end{aligned}$$

This grammar is in CNF, and it is equivalent to G. □

A context-free grammar $G = (N, T, S, P)$ is in

Greibach Normal Form,

if $r \in T \cdot N^*$ holds for each production $(\ell \to r) \in P$.

Theorem 3.10 (Greibach 1965)

Given a context-free grammar G such that $\varepsilon \notin L(G)$, one can effectively construct an equivalent context-free grammar G' that is in Greibach Normal Form.

Lemma 3.11

Let $G = (N, T, S, P)$ be a context-free grammar, let $(A \to \alpha_1 B \alpha_2) \in P$, and let $(B \to \beta_1), (B \to \beta_2), \ldots, (B \to \beta_r) \in P$ be the set of all B-produktions in G. Further, let $G_1 = (N, T, S, P_1)$ be the grammar that is obtained G by replacing the production $(A \to \alpha_1 B \alpha_2)$ by the productions

$$(A \to \alpha_1 \beta_1 \alpha_2), (A \to \alpha_1 \beta_2 \alpha_2), \ldots, (A \to \alpha_1 \beta_r \alpha_2).$$

Then $L(G_1) = L(G)$.

Lemma 3.12

Let $G = (N, T, S, P)$ be a context-free grammar and let

$$(A \to A\alpha_1), (A \to A\alpha_2), \ldots, (A \to A\alpha_r) \in P$$

be the set of A-produktions the righthand side of which has the prefix A. Let $(A \to \beta_1), (A \to \beta_2), \ldots, (A \to \beta_s) \in P$ be the other A-productions of G. The grammar $G_1 = (N \cup \{B\}, T, S, P_1)$ is obtained from G by introducing the new nonterminal B and by replacing the A-productions of G by the following productions:

$$\begin{array}{llllll} (A \to \beta_1), & \ldots, & (A \to \beta_s), & (A \to \beta_1 B), & \ldots, & (A \to \beta_s B), \\ (B \to \alpha_1), & \ldots, & (B \to \alpha_r), & (B \to \alpha_1 B), & \ldots, & (B \to \alpha_r B). \end{array}$$

Then $L(G_1) = L(G)$.

Proof of Theorem 3.10.

Let G be in CNF with $N = \{A_1, A_2, \ldots, A_m\}$.

(1.) Modify the productions such that $(A_i \to A_j\alpha) \in P$ implies $i < j$:

FOR $i := 1$ TO m DO
  FOR $j := 1$ TO $i - 1$ DO
    FOR ALL $(A_i \to A_j\alpha) \in P$ DO
      Let $A_j \to \beta_1 \mid \beta_2 \mid \ldots \mid \beta_n$ be all A_j-productions.
      Add the productions $A_i \to \beta_1\alpha \mid \beta_2\alpha \mid \ldots \mid \beta_n\alpha$
      and delete $(A_i \to A_j\alpha)$
    END;
  END;

(2.) Remove left recursive productions using Lemma 3.12:

  IF there is a production of the form $(A_i \to A_i\alpha)$ THEN
    use Lemma 3.12 with the new nonterminal B_i
  END
END.

Proof of Theorem 3.10 (cont.).

(3.) The righthand side of each A_m-production begins with a terminal symbol. We now enforce this also for all A_i-productions, $i = m-1, m-2, \ldots, 1$:

FOR $i := m-1$ DOWNTO 1 DO
  FOR ALL $(A_i \to A_j\alpha) \in P$, $j > i$, DO
    Let $A_j \to \beta_1 \mid \beta_2 \mid \ldots \mid \beta_n$ be all A_j-productions.
    Add the productions $A_i \to \beta_1\alpha \mid \beta_2\alpha \mid \ldots \mid \beta_n\alpha$
    and delete $(A_i \to A_j\alpha)$
  END
END

(4.) Modify the B_i-productions $(B_i \to A_j\alpha)$ $(1 \leq i \leq n)$:

Let $A_j \to \beta_1 \mid \beta_2 \mid \ldots \mid \beta_n$ be all A_j-productions.
Add the productions $B_i \to \beta_1\alpha \mid \beta_2\alpha \mid \ldots \mid \beta_n\alpha$
and delete $(B_i \to A_j\alpha)$. □

Example:

Let $G = (\{A_1, A_2, A_3\}, \{a, b\}, A_1, P)$, where

$$P : (A_1 \to A_2A_3), (A_2 \to A_3A_1), (A_2 \to b), (A_3 \to A_1A_2), (A_3 \to a).$$

Step 1: The A_1- and A_2-productions are already in the correct form,

$(A_3 \to A_1A_2)$ is replaced by $(A_3 \to A_2A_3A_2)$,

$(A_3 \to A_2A_3A_2)$ is replaced by

$(A_3 \to A_3A_1A_3A_2)$ and $(A_3 \to bA_3A_2)$.

Step 2: Lemma 3.12 is applied to the A_3-productions:

$(A_3 \to bA_3A_2), (A_3 \to a), (A_3 \to bA_3A_2B_3), (A_3 \to aB_3),$

$(B_3 \to A_1A_3A_2), (B_3 \to A_1A_3A_2B_3).$

Example (cont.):

Step 3: All A_i-productions are brought into Greibach form:

$$(A_3 \to bA_3A_2), (A_3 \to a), (A_3 \to bA_3A_2B_3), (A_3 \to aB_3),$$
$$(A_2 \to bA_3A_2A_1), (A_2 \to aA_1), (A_2 \to bA_3A_2B_3A_1),$$
$$(A_2 \to aB_3A_1), (A_2 \to b),$$
$$(A_1 \to bA_3A_2A_1A_3), (A_1 \to aA_1A_3),$$
$$(A_1 \to bA_3A_2B_3A_1A_3), (A_1 \to aB_3A_1A_3), (A_1 \to bA_3).$$

Step 4: The B_3-productions are brought into Greibach form:

$$(B_3 \to bA_3A_2A_1A_3A_3A_2), (B_3 \to aA_1A_3A_3A_2),$$
$$(B_3 \to bA_3A_2B_3A_1A_3A_3A_2), (B_3 \to aB_3A_1A_3A_3A_2),$$
$$(B_3 \to bA_3A_3A_2), (B_3 \to bA_3A_2A_1A_3A_3A_2B_3),$$
$$(B_3 \to aA_1A_3A_3A_2B_3), (B_3 \to bA_3A_2B_3A_1A_3A_3A_2B_3),$$
$$(B_3 \to aB_3A_1A_3A_3A_2B_3), (B_3 \to bA_3A_3A_2B_3).$$

While G has only 5 short productions, G' has 24 long ones. □

A context-free grammar $G = (N, T, S, P)$ is in

quadratic Greibach Normal Form,

if $r \in T \cdot N^*$ and $|r| \leq 3$ for each production $(\ell \to r) \in P$, that is, r contains at most two nonterminals.

Theorem 3.13 (Rosenkrantz 1967)

Given a context-free grammar G such that $\varepsilon \notin L(G)$, one can effectively construct an equivalent context-free grammar G' that is in quadratic Greibach Normal Form.