# Formal Languages & Compiler Design
# Homework 1

Stefan Stefanache

October 26, 2022

**1.** Given the grammar

$$G = \big(\{S, H\}, \{b, c, d, e\}, \{S \to b^2Se \mid H, H \to cHd^2 \mid cd\}, S\big) \tag{1}$$

find the language it generates.

*Proof.* Let us define the language

$$L = \{b^{2n}c^{m+1}d^{2m+1}e^n \mid n, m \in \mathbb{N}\} = \{L_{n,m} \mid n, m \in \mathbb{N}\} \tag{2}$$

and label $G$'s productions as follows:

$$S \to b^2Se \tag{3}$$
$$S \to H \tag{4}$$
$$H \to cHd^2 \tag{5}$$
$$H \to cd \tag{6}$$

Our goal is proving that $L = L(G)$ using the double inclusion technique.

**Step 1:** Show that $L \subseteq L(G)$:

Using induction, one can easily show that for all $n \in \mathbb{N}$,

$$H \underset{(5)}{\overset{n}{\Longrightarrow}} c^nHd^{2n} \underset{(6)}{\overset{1}{\Longrightarrow}} c^{n+1}d^{2n+1} \tag{7}$$
$$S \underset{(3)}{\overset{n}{\Longrightarrow}} b^{2n}Se^n \underset{(4)}{\overset{1}{\Longrightarrow}} b^{2n}He^n \tag{8}$$

Therefore, we have that

$$S \underset{(8)}{\overset{*}{\Longrightarrow}} b^{2k}He^k \underset{(7)}{\overset{*}{\Longrightarrow}} b^{2k}c^{p+1}d^{2p+1}e^k, \quad \forall p, k \in \mathbb{N}$$

As a result,

$$S \overset{*}{\Longrightarrow} w, \quad \forall w \in L$$

so

$$L \subseteq L(G) \tag{9}$$

**Step 2:** Show that $L \supseteq L(G)$:

All elements in $L(G)$ are words/sequences derived from $S$. Therefore, we have to show that all elements that can be derived from $S$ and don't contain any nonterminal symbols are contained by $L$.

We start by studying how the elements derived from $H$ look like. After applying the (5) rule for $m \in \mathbb{N}$ times, the result is of the form $c^m H d^{2m}$. To get rid of $H$, we apply (6) one time to get a word/sequence of the form $c^{m+1}d^{2m+1}$. This is the only kind of word/sequence that can be obtained starting from $H$.

Now, we continue by looking at the possible elements derived from $S$. Analogously, we apply rule (3) for $n \in \mathbb{N}$ times to get to the form $b^{2n}Se^n$. At this point, our only move is to transform the $S$ into $H$ using (4). Finally, we transform $H$ like previously discussed, to obtain a final word/sequence of the form $b^{2n}c^{m+1}d^{2m+1}e^n \in L$, with $n, m \in \mathbb{N}$.

Since all possible words/sequences derived from $S$ can be found in $L$, we have that

$$L \supseteq L(G) \tag{10}$$

Finally, from (9) and (10), we have that $L = L(G)$, so the language generated by the grammar $G$ is given by (2). □

**2.** Find grammars that generate the following languages:

**A.** $L_1 = \{x^n y^n \mid n \in \mathbb{N}\}$, with proof

**B.** $L_2 = \{a^n b^{2n} \mid n \in \mathbb{N}\}$, with proof

**C.** $L_3 = \{a^n b^m \mid n, m \in \mathbb{N}^*\}$, with proof using regular grammar

**D.** $L_4 = \{x^{2n} \mid n \in \mathbb{N}\}$, $L_4' = \{x^{2n} \mid n \in \mathbb{N}^*\}$, with proof using regular grammar

**E.** $\mathbb{N}$

**F.** All arithmetic expressions containing $a$ as operand, $+$, $*$ as operators and ().

*Proof.* For all proofs, we'll be using a similar method as in the previous exercise: find a grammar $G_1$ and then prove that $L_1 = L(G_1)$ using the double inclusion technique.

**A.** Let us define the grammar

$$G_1 = \big(\{A\}, \{x, y\}, \{A \to xAy \mid \epsilon\}, A\big) \tag{11}$$

and label its rules as

$$A \to xAy \tag{12}$$
$$A \to \epsilon \tag{13}$$

**Step 1.** Show that $L_1 \subseteq L(G_1)$:
Using induction, one can easily show that $A \underset{(12)}{\overset{n}{\Longrightarrow}} x^n A y^n$.

Since $x^n A y^n \underset{(13)}{\overset{1}{\Longrightarrow}} x^n y^n$, we have that

$$A \overset{*}{\Longrightarrow} x^n y^n \in L_1 \tag{14}$$

Therefore, $A \overset{*}{\Longrightarrow} w$, $\forall w \in L_1$, so

$$L_1 \subseteq L(G_1) \tag{15}$$

**Step 2.** Show that $L_1 \supseteq L(G_1)$:
We have to prove that each element derived from $A$ that doesn't contain any nonterminal symbols is contained by $L_1$. Starting from $A$, we apply rule (12) for $n \in \mathbb{N}$ times and reach a result of the form $x^n A y^n$. To get rid of the nonterminal symbol $A$, we use rule (13) to obtain the word/sequence $x^n y^n$.
Since this is the only possible way of sequencing transformations (note that this also works for $n = 0$), all elements derived from $A$ with no nonterminal terms are of the form $x^n y^n \in L_1$. Therefore,

$$L_1 \supseteq L(G_1) \tag{16}$$

Using (15) and (16), we have that $L_1 = L(G_1)$, so we've proved that the language $L_1$ is generated by the grammar given by (11).

**B.** Let us define the grammar

$$G_2 = \big(\{B\}, \{a, b\}, \{B \to aBb^2 \mid ab^2\}, B\big) \tag{17}$$

and label its rules as

$$B \to aBb^2 \tag{18}$$
$$B \to ab^2 \tag{19}$$

**Step 1.** Show that $L_2 \subseteq L(G_2)$:

Let $n \in \mathbb{N}^*$. Using induction, one can easily show that

$$B \underset{(18)}{\overset{n-1}{\Longrightarrow}} a^{n-1}Bb^{2n-2}$$

Since $a^{n-1}Bb^{2n-2} \underset{(19)}{\overset{1}{\Longrightarrow}} a^n b^{2n}$, we have that

$$B \overset{*}{\Longrightarrow} a^n b^{2n} \in L_2 \tag{20}$$

Therefore, $B \overset{*}{\Longrightarrow} w$, $\forall w \in L_2$, so

$$L_2 \subseteq L(G_2) \tag{21}$$

**Step 2.** Show that $L_2 \supseteq L(G_2)$:

We have to prove that each element derived from $B$ that doesn't contain any nonterminal symbols is contained by $L_2$. Starting from $B$, we apply rule (18) for $n-1 \in \mathbb{N}$ times and reach a result of the form $a^{n-1}Bb^{2n-2}$. To get rid of the nonterminal symbol $B$, we use rule (19) to obtain the word/sequence $a^n b^{2n}$.

Since this is the only possible way of sequencing transformations (note that this also works for $n = 0$), all elements derived from $B$ with no nonterminal terms are of the form $a^n b^{2n} \in L_2$. Therefore,

$$L_2 \supseteq L(G_2) \tag{22}$$

Using (21) and (22), we have that $L_2 = L(G_2)$, so we've proved that the language $L_2$ is generated by the grammar given by (17).

**C.** Let us define the **regular** grammar

$$G_3 = \big(\{A, B\}, \{a, b\}, \{A \to aA \mid aB, B \to bB \mid b\}, A\big) \tag{23}$$

and label its rules as

$$A \to aA \tag{24}$$
$$A \to aB \tag{25}$$
$$B \to bB \tag{26}$$
$$B \to b \tag{27}$$

**Step 1.** Show that $L_3 \subseteq L(G_3)$:

Let $n, m \in \mathbb{N}^*$. Using induction twice, one can easily show that

$$A \underset{(24)}{\overset{n-1}{\Longrightarrow}} a^{n-1}A \underset{(25)}{\overset{1}{\Longrightarrow}} a^n B \underset{(26)}{\overset{m-1}{\Longrightarrow}} a^n b^{m-1} B \underset{(27)}{\overset{1}{\Longrightarrow}} a^n b^m$$

Therefore, we have that

$$A \overset{*}{\Longrightarrow} a^n b^m \in L_3 \tag{28}$$

Therefore, $A \overset{*}{\Longrightarrow} w$, $\forall w \in L_3$, so

$$L_3 \subseteq L(G_3) \tag{29}$$

**Step 2.** Show that $L_3 \supseteq L(G_3)$:

We have to prove that each element derived from $A$ that doesn't contain any nonterminal symbols is contained by $L_3$. Starting from $A$, we apply rule (24) for $n-1 \in \mathbb{N}$ times and reach a result of the form $a^{n-1}A$. To get rid of the nonterminal symbol $A$, we use rule (19) to obtain the new form $a^n B$.

Now, we can apply rule (26) for $m-1 \in \mathbb{N}$ times to reach the form $a^n b^{m-1} B$. Getting rid of the nonterminal $B$ is done by using the rule (27) and reaching the final form $a^n b^m \in L_3$.

Since this is the only possible way of sequencing transformations, all elements derived from $A$ with no nonterminal terms are of the form $a^n b^m \in L_3$. Therefore,

$$L_3 \supseteq L(G_3) \tag{30}$$

Using (29) and (30), we have that $L_3 = L(G_3)$, so we've proved that the language $L_3$ is generated by the grammar given by (23).

**D.** Let us define the **regular** grammar

$$G_4 = \big(\{S, A, B\}, \{x\}, \{S \to xA \mid \epsilon, A \to xB \mid x, B \to xA\}, S\big) \tag{31}$$

and label its rules as

$$S \to xA \tag{32}$$
$$S \to \epsilon \tag{33}$$
$$A \to xB \tag{34}$$
$$A \to x \tag{35}$$
$$B \to xA \tag{36}$$

One trick we're using here is making sure that $A$ is always accompanied by an odd power of $x$. That way, we either use (35) to transform into something of the form $x^{2k}$, or transform it into $xB$ using (34) and cycle back with (36), getting something of the form $x^{2k+1}A$ again. By doing it this way, we make sure that only odd powers are covered.

**Step 1.** Show that $L_4 \subseteq L(G_4)$:

To enable the cycle trick discussed above, we'll prove by induction that

$$A \overset{*}{\Longrightarrow} x^{2n+1},\ \forall n \in \mathbb{N} \tag{37}$$

The base case for $n = 0$ holds since $A \underset{(32)}{\overset{1}{\Longrightarrow}} x$. Now, let's take $k \in \mathbb{N}$ and assume that

$$A \overset{*}{\Longrightarrow} x^{2k+1} \tag{38}$$

Then,

$$A \underset{(34)}{\overset{1}{\Longrightarrow}} xB \underset{(36)}{\overset{1}{\Longrightarrow}} x^2 A \underset{(38)}{\overset{*}{\Longrightarrow}} x^{2k+3}$$

This proves that $P(k) \implies P(k+1)$ is true, where

$$P(n) : A \overset{*}{\Longrightarrow} x^{2n+1}$$

Therefore, since both $P(0)$ and $P(k) \implies P(k+1)$ hold we've proved that (37) is true. Now, we have that

$$S \underset{(32)}{\overset{1}{\Longrightarrow}} xA \underset{(37)}{\overset{*}{\Longrightarrow}} x^{2n} \in L_4$$

Therefore, $S \overset{*}{\Longrightarrow} w$, $\forall w \in L_4$, so

$$L_4 \subseteq L(G_4) \tag{39}$$

**Step 2.** Show that $L_4 \supseteq L(G_4)$:

We have to prove that each element derived from $S$ that doesn't contain any nonterminal symbols is contained by $L_4$. Starting from $S$, we apply rule (32) to obtain $xA$. Now, we sequentially apply the rules (34) and (36) for $k \in \mathbb{N}$ times to obtain something of the form $x^{2k+1}A$. Now, we can apply (35) to get rid of the nonterminal symbol $A$ and get the final sequence of the form $x^{2k} \in L_4$.

We can also start from $S$ and apply (33) to obtain $\epsilon = x^0 \in L_4$.

Since these are the only possible way of sequencing transformations, all elements derived from $A$ with no nonterminal terms are of the form $a^n b^m \in L_3$. Therefore,

$$L_4 \supseteq L(G_4) \tag{40}$$

Using (39) and (40), we have that $L_4 = L(G_4)$, so we've proved that the language $L_4$ is generated by the grammar given by (31).

We do exactly the same thing for $L'_4$, but by removing the (33) rule from the production set of the grammar, since we're now dealing with only positive powers of $x$. One can define the adapted **regular** grammar

$$G'_4 = \big(\{S, A, B\}, \{x\}, \{S \to xA, A \to xB \mid x, B \to xA\}, S\big) \tag{41}$$

and follow the same steps as above to prove that $L'_4 = L(G'_4)$.

**E.** Let us define the grammar

$$G_5 = \big(\{S, A, B, C, D\}, \Sigma, P, S\big) \tag{42}$$

where the nonterminal symbols are the 10 digits,

$$\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \tag{43}$$

and the production set is given by

$$P = \{S \to 0 \mid A, A \to B \mid BC, C \to D \mid DC, D \to 0 \mid B, B \to 1 \mid 2 \mid 3 \mid \ldots \mid 9\} \tag{44}$$

This grammar generates the set of natural numbers $\mathbb{N}$ ($L(G_5) = \mathbb{N}$) and its rules can be described as follows: $B$ denotes a non-zero digit, $D$ denotes any digit, $C$ denotes a sequence of digits, and $A$ denotes a natural positive number. $S$ can be transformed to a positive number ($A$) or 0.

**F.** Let us define the grammar

$$G_6 = \big(\{S, A, b\}, \{a, +, *, (, )\}, P, S\big) \tag{45}$$

Using the production set

$$P = \{S \to S + a \mid A, A \to A * a \mid B, B \to (S) \mid a\} \tag{46}$$

we have that $L(G_6)$ covers all arithmetic expressions containing $a$ as an operand, +, * as operators and (). Compound statements can be enabled by the cyclical nature of the rules $S \to A \to B \to S \to \ldots$. The first rule enables sums, the second products, and the third compounds statements.

□