THE LEFTMOST DERIVATION OF TYPE TWO IN MATRIX GRAMMARS

Ing. Oto ŠKRKAL, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: skrkal@fit.vutbr.cz

Supervised by: Dr. Alexander Meduna

ABSTRACT

This paper discusses the descriptional complexity of matrix grammars using leftmost derivation of type two with respect to the number of nonterminals and matrices with two or more productions. It proves that these matrix grammars need only nine nonterminals and six matrices of length two or more to generate recursively enumerable languages.

1 INTRODUCTION

Formal language theory has recently investigated the economical transformations of several formal models. The aim of these transformations is to reduce the number of nonterminals or number of productions in grammars without any decrease of their generative power.

Matrix grammars are one of the models that use regulated rewriting (see [1]). In this model, the productions are arranged into sequences which have to be used as a whole during one derivation step. It has been proved that the leftmost derivation of type two increases the generative power of matrix grammars with ε-productions to the family of recursively enumerable languages (see [1]). This paper proves that these grammars need no more than nine nonterminals and six matrices having more than one production to generate recursively enumerable languages.

2 DEFINITIONS

This paper assumes the reader to be familiar with the formal language theory (see [2]).

Let V be an alphabet. V^* is the free monoid generated by V under the operation of concatenation, ε is the unit of V^*. Set $V^+ = V^* \setminus \{\varepsilon\}$. An ε-production has its right-hand side equal to ε. A context production is a matrix of length two or more. (A matrix of length one is sometimes referred to as a production.) We set $\mathrm{alph}(w) = \{a \in V : w =$

xay, where $x, y \in V^*\}$ for all $w \in V^*$. For $w \in V^*$, $|w|$ denotes the length of w. For all $m, n \in Z$, $\min(m, n)$ denotes a minimum of m and n.

A context-free grammar $G = (N, T, P, S)$ is in *Greibach normal form* if every production in P has the form of $A \to a\alpha$, where $A \in N$, $a \in T$ and $\alpha \in N^*$. In case that $\varepsilon \in L(G)$, $S \to \varepsilon \in P$ and S is on the right-hand side of no production in P.

A *matrix grammar* is a quadruple, $G = (N, T, M, S)$, where N and T are terminal and nonterminal alphabet such that $N \cap T = \oslash$. $S \in N$ is an axiom, and M is a finite set of matrices. Every matrix is a sequence of the form $m : (p_1, p_2, \ldots, p_n)$, $n \geq 1$, with $p_i = A_i \to b_i$, $A_i \in N$, $b_i \in V^*$ for all $1 \leq i \leq n$, where $V = N \cup T$. Matrices can be labeled by some label m. Let $x, y \in V^*$, $(p_1, p_2, \ldots, p_n) \in M$ for some $n \geq 1$, $p_i = A_i \to b_i$ for all $i \in \langle 1, n \rangle$. If there are strings $x_0, x_1, \ldots, x_n \in V^*$ such that $x_0 = x$, $x_n = y$, $x_{i-1} = x'_{i-1} A_i x''_{i-1}$, and $x_i = x'_{i-1} b_i x''_{i-1}$ for all $i \in \langle 1, n \rangle$ and some $x'_{i-1}, x''_{i-1} \in V^*$, then x *directly derives* y in G according to $(p_1, p_2, \ldots, p_n)$, symbolically written as $x \Rightarrow_G y\ [(p_1, p_2, \ldots, p_n)]$, or $x \Rightarrow y$ for short. As usual, we extend $\Rightarrow_G$ to $\Rightarrow^i_G$ (where $i \geq 0$), $\Rightarrow^+_G$, and $\Rightarrow^*_G$. The language generated by G is $L(G) = \{x : x \in T^*, S \Rightarrow^* x\}$.

Let $G = (N, T, M, S)$ be a matrix grammar, $V = N \cup T$, and let there be $x, y \in V^*$, and $(p_1, p_2, \ldots, p_n) \in M$ for some $n \geq 1$, with $p_i = A_i \to b_i$ for all $i \in \langle 1, n \rangle$. If there are strings $x_0, x_1, \ldots, x_n \in V^*$ such that $x_0 = x$, $x_n = y$, $x_{i-1} = x'_{i-1} A_i x''_{i-1}$, $x_i = x'_{i-1} b_i x''_{i-1}$ for all $i \in \langle 1, n \rangle$ and some $x'_{i-1}, x''_{i-1} \in V^*$, $A_i \notin \mathrm{alph}(x'_{i-1})$ for all $i \in \langle 1, n \rangle$, and there is no matrix $(q_1, q_2, \ldots, q_m) \in M$, with $q_j = B_j \to g_j$, $j \in \langle 1, m \rangle$, $m \geq 1$, applicable to x such that $B_1 \in \mathrm{alph}(x'_0)$, then the derivation $x \Rightarrow_G y\ [(p_1, p_2, \ldots, p_n)]$, is called *leftmost of type 2*.

For details about left-most derivation of type 1 and 3, see [1].

3 RESULTS

Lemma 1. *For every recursively enumerable language L over an alphabet T, there exist two context-free grammars $G_1 = (N_1, T', P_1, S_1)$ and $G_2 = (N_2, T', P_2, S_2)$ and a homomorphism $h : T' \to T^*$ such that $L = \{h(x) : x \in L(G_1) \cap L(G_2)\}$.*

Proof. See [3]. □

Theorem 1. *Every recursively enumerable language can be defined by a matrix grammar with at most nine nonterminals and six context productions and using the leftmost derivation of type 2.*

Proof. Let $G_1 = (N_1, T, P_1, S_1)$ and $G_2 = (N_2, T, P_2, S_2)$ be two context-free grammars in Greibach normal form such that $N_1 \cap N_2 = \oslash$, h a homomorphism from T to T'^*, and $L = \{h(x) : x \in L(G_1) \cap L(G_2)\}$ a recursively enumerable language defined by them.

Define arbitrary bijective projections $f : N_1 \cup N_2 \to 0^*$ and $f' : N_1 \cup N_2 \to 0'^*$ such that $|f(A)| = |f'(A)|$ for all $A \in N_1 \cup N_2$, and a bijective homomorphism $f_1 : (N_1 \cup N_2)^* \to \{0, 1\}^*$ such that $f_1(\varepsilon) = \varepsilon$, $f_1(A) = f(A)1$, $f_1(A\alpha) = f_1(A) f_1(\alpha)$ for all $A \in N_1 \cup N_2$ and $\alpha \in (N_1 \cup N_2)^*$.

Define also bijective homomorphisms $\hat{g} : T^* \to \{\hat{0}, \hat{1}\}^*$ and $\bar{g} : T^* \to \{\bar{0}, \bar{1}\}^*$ over T such that $\hat{g}(\varepsilon) = \bar{g}(\varepsilon) = \varepsilon$, $\hat{g}(a) \in \hat{0}^* \hat{1}$, $\bar{g}(a) \in \bar{0}^* \bar{1}$, $|\hat{g}(a)| = |\bar{g}(a)|$, $\hat{g}(a\alpha) = \hat{g}(a) \hat{g}(\alpha)$ and $\bar{g}(a\alpha) = \bar{g}(a) \bar{g}(\alpha)$ for all $a \in T$ and $\alpha \in T^*$.

Introduce the matrix grammar $G = (N, T', M, S)$ with the leftmost derivation of type 2 constructed in the following way. $N = \{S, 0, 1, 0', 1', \hat{0}, \hat{1}, \overline{0}, \overline{1}\}$, $N \cap (N_1 \cup N_2) = \emptyset$, and M satisfies

1. $S \to f_1(S_1) f_1(S_2) \in M$;
2. $S \to \varepsilon \in M$ if $S_1 \to \varepsilon \in P_1$ and $S_2 \to \varepsilon \in P_2$;
3. $1 \to f'(A)0'1'\hat{g}(a) f_1(\alpha) \in M$ for all $A \to a\alpha \in P_1$, where $A \in N_1$, $a \in T$, $\alpha \in N_1^*$;
4. $1 \to f'(A)0'1'\overline{g}(a)h(a) f_1(\alpha) \in M$ for all $A \to a\alpha \in P_2$, where $A \in N_2$, $a \in T$, $\alpha \in N_2^*$;
5. $(0 \to \varepsilon, 0' \to \varepsilon) \in M$, $(0' \to \varepsilon, 1' \to \varepsilon) \in M$, $0' \to 0' \in M$, $1' \to 1' \in M$;
6. $(\hat{0} \to \varepsilon, \overline{0} \to \varepsilon, \overline{1} \to \overline{1}) \in M$, $(\hat{1} \to \varepsilon, \overline{1} \to \varepsilon) \in M$, $(\hat{0} \to \hat{0}, \overline{1} \to \overline{1}) \in M$, $(\hat{0} \to \hat{0}, \overline{0} \to \overline{0}) \in M$.

Without any loss of generality, assume that, if possible, there is allways used one of the matrices producing a sentential form different from the original one when more matrices are applicable in the next derivation step.

We prove that G generates the language $L(G) = L = \{h(x) : x \in L(G_1) \cap L(G_2)\}$.

Basic idea: Roughly speaking, G first simulates the leftmost derivation in G_1 using the matrices of (3) and (5). Then, it simulates the leftmost derivation in G_2 using the matrices of (4) and (5). During the simulation of G_2, it checks whether the string being generated by G_2 corresponds to the string generated by G_1. To do this, it uses the matrices of (6).

First, we prove the following claims.

Claim 1. *Let $x = u0^m 0'^n 1'vw$, where $u \in (\{\hat{0}, \hat{1}\} \cup T')^*$, $v \in \hat{0}^*\hat{1} \cup T'^*$, $w \in \{0,1\}^*$, $m, n \geq 0$, and let $j = \min(m, n)$. Then, $x \Rightarrow^j y$, where $y = u0^{m-j}0'^{n-j}1'vw$, and there are no $z \neq y$ such that $x \Rightarrow^j z$, and no $z' \in T'^*$ such that $x \Rightarrow^k z'$, $0 < k < j$.*

Proof. Let $x = u0^m 0'^n 1'vw$, where $u \in (\{\hat{0}, \hat{1}\} \cup T')^*$, $v \in \hat{0}^*\hat{1} \cup T'^*$, $w \in \{0,1\}^*$, $m \geq 0$ and $n \geq 0$. The claim is proven by induction on $\min(m, n)$.

Basis: Let m and n be such that $\min(m, n) = 0$. Then, $u0^m 0'^n 1'vw \Rightarrow^0 u0^m 0'^n 1'vw$.

Induction hypothesis: Suppose that the claim holds for all pairs of m and n, $m, n \geq 0$, such that $\min(m, n) \leq k$, for some $k \geq 0$.

Induction step: Let us consider m and n such that $\min(m, n) = k + 1$. Since $k + 1 \geq 1$, $m > 1$ and $n > 1$ and we can express the string $x = u00^{m-1}0'0'^{n-1}1'vw$. Due to the leftmost derivation of type 2, the only applicable matrix is $(0 \to \varepsilon, 0' \to \varepsilon)$. By its application we obtain $u0^{m-1}0'^{n-1}1'vw$ having $\min(m-1, n-1) = k$. By induction hypothesis $u0^{m-1}0'^{n-1}1'vw \Rightarrow^k u0^{m-1-k}0'^{n-1-k}1'vw$. That is, $x \Rightarrow^{k+1} u0^{m-(k+1)}0'^{n-(k+1)}1'vw$ and the claim holds. □

Claim 2. *Let $x = u0^m 0'^n 1'vw$, where $u \in (\{\hat{0}, \hat{1}\} \cup T')^*$, $v \in \hat{0}^*\hat{1} \cup T'^*$, $w \in \{0,1\}^*$, $m, n \geq 0$. If $n = m+1$, then $x \Rightarrow^{m+1} y$, $y = uvw$, and there is no $z \neq y$ such that $x \Rightarrow^{m+1} z$; if $n > m+1$, then for all $k \geq m+1$, $x \Rightarrow^k y$, $y = u0'^{n-(m+1)}vw$, and there is no $z \neq y$ such that $x \Rightarrow^k z$; if $n < m+1$, then for all $k \geq n$, $x \Rightarrow^k y$, $y = u0^{m-n}1'vw$, and there is no $z \neq y$ such that $x \Rightarrow^k z$.*

Proof. Let $x = u0^m0'^n1'vw$, where $u \in (\{\hat{0},\hat{1}\} \cup T')^*$, $v \in \hat{0}^*\hat{1} \cup T'^*$, $w \in \{0,1\}^*$, $m \geq 0$, and $n \geq 0$. We examine all relations between m and n.

1. Let $n = m+1$. By Claim 1, $u0^m0'^{m+1}1'vw \Rightarrow^m u0'1'vw$. Due to the leftmost derivation of type 2 the only applicable matrices are $(0' \to \varepsilon, 1' \to \varepsilon)$ and $0' \to 0'$. As agreed above, we allways use a matrix producing sentential form different from the current one, whenever it is possible. Thus, $u0'1'vw \Rightarrow uvw\ [(0' \to \varepsilon, 1' \to \varepsilon)]$; that is, $x \Rightarrow^{m+1} uvw$.

2. Let $n > m+1$. By Claim 1, $u0^m0'^n1'vw \Rightarrow^m u0'^{n-m}1'vw$. Since $n > m+1$ implies $n-m \geq 2$, we can express $u0'^{n-m}1'vw = u0'0'0'^{n-m-2}1'vw$. The only applicable matrices are $(0' \to \varepsilon, 1' \to \varepsilon)$ and $0' \to 0'$. Therefore, $u0'0'0'^{n-m-2}1'vw \Rightarrow u0'0'^{n-m-2}vw\ [(0' \to \varepsilon, 1' \to \varepsilon)]$. The only applicable matrix is $0' \to 0'$. Since this matrix lefts the sentential form unchanged, $u0'0'^{n-m-2}vw \Rightarrow^* u0'0'^{n-m-2}vw$. Thus, $x \Rightarrow^k u0'^{n-m-1}vw$ for all $k \geq m+1$.

3. Let $n < m+1$. By Claim 1, $u0^m0'^n1'vw \Rightarrow^n u0^{m-n}1'vw$. The only applicable matrix is $1' \to 1'$. This matrix lefts the sentential form unchanged, again. Therefore, $u0^{m-n}1'vw \Rightarrow^* u0^{m-n}1'vw$, and $x \Rightarrow^k u0^{m-n}1'vw$ for all $k \geq n$. □

Claim 3. *Let $x = \hat{0}^m\hat{1}u\bar{0}^n\bar{1}v$, where $u \in \{\hat{0},\hat{1}\}^*T'^*\{0,0',1'\}^*$, $v \in T'^*\{0,1\}^*$, $m \geq 0$ and $n \geq 0$, and let $j = \min(m,n)$. Then, $x \Rightarrow^j y$, where $y = \hat{0}^{m-j}\hat{1}u\bar{0}^{n-j}\bar{1}v$, and there are no $z \neq y$ such that $x \Rightarrow^j z$, and no $w \in T'^*$ such that $x \Rightarrow^k w$, $0 < k < j$.*

Proof. By analogy with the proof of the Claim 1, this claim can be proven by induction on $\min(m,n)$. □

Claim 4. *Let $x = \hat{0}^m\hat{1}u\bar{0}^n\bar{1}v$, where $u \in \{\hat{0},\hat{1}\}^*T'^*\{0,0',1'\}^*$, $v \in T'^*\{0,1\}^*$, $m,n \geq 0$. If $m = n$, then $x \Rightarrow^{m+1} y$, $y = uv$, and there is no $z \neq y$ such that $x \Rightarrow^{m+1} z$; if $m > n$, then for all $k \geq n$, $x \Rightarrow^k y$, $y = \hat{0}^{m-n}\hat{1}u\bar{1}v$, and there is no $z \neq y$ such that $x \Rightarrow^k z$; if $m < n$, then $x \Rightarrow^{m+1} y$, $y = u\bar{0}^{n-m}v$, and there is no $z \neq y$ such that $x \Rightarrow^{m+1} z$; furthermore, $x \Rightarrow^k y$ for all $k \geq m+1$ if $\hat{0} \in \mathrm{alph}(u)$.*

Proof. Let $x = \hat{0}^m\hat{1}u\bar{0}^n\bar{1}v$, where $u \in \{\hat{0},\hat{1}\}^*T'^*\{0,0',1'\}^*$, $v \in T'^*\{0,1\}^*$, $m,n \geq 0$. We examine all relations between m and n.

1. Let $m = n$. By Claim 3, $\hat{0}^m\hat{1}u\bar{0}^m\bar{1}v \Rightarrow^m \hat{1}u\bar{1}v$. The only applicable matrix is $(\hat{1} \to \varepsilon, \bar{1} \to \varepsilon)$. Thus, $\hat{1}u\bar{1}v \Rightarrow uv\ [(\hat{1} \to \varepsilon, \bar{1} \to \varepsilon)]$, and $x \Rightarrow^{m+1} uv$.

2. Let $m > n$. By Claim 3, $\hat{0}^m\hat{1}u\bar{0}^m\bar{1}v \Rightarrow^n \hat{0}^{m-n}\hat{1}u\bar{1}v$. Because $m > n$ implies $m-n \geq 1$, the only matrix that can be applied and has to be applied is $(\hat{0} \to \hat{0}, \bar{1} \to \bar{1})$. Since this matrix lefts the sentential form unchanged, $\hat{0}^{m-n}\hat{1}u\bar{1}v \Rightarrow^* \hat{0}^{m-n}\hat{1}u\bar{1}v$. Thus, $x \Rightarrow^k \hat{0}^{m-n}\hat{1}u\bar{1}v$ for all $k \geq n$.

3. Let $m < n$. By Claim 3, $\hat{0}^m\hat{1}u\bar{0}^n\bar{1}v \Rightarrow^m \hat{1}u\bar{0}^{n-m}\bar{1}v$. The only possible derivation step is $\hat{1}u\bar{0}^{n-m}\bar{1}v \Rightarrow u\bar{0}^{n-m}v\ [(\hat{1} \to \varepsilon, \bar{1} \to \varepsilon)]$; that is, $x \Rightarrow^{m+1} u\bar{0}^{n-m}v$. If $\hat{0} \in \mathrm{alph}(u)$, then we can express $u\bar{0}^{n-m}v = y\hat{0}z\bar{0}^{n-m}v$, where $y \in \hat{1}^*$, $z \in \{\hat{0},\hat{1}\}^*T'^*\{0,0',1'\}^*$. Observe that $n-m \geq 1$, so the matrix $(\hat{0} \to \hat{0}, \bar{0} \to \bar{0})$ can be applied. Furthermore,

due to the leftmost derivation of type 2 it is also the only applicable matrix. This matrix lefts the sentential form unchanged; therefore, $y\hat{0}z\overline{0}^{n-m}v \Rightarrow^* y\hat{0}z\overline{0}^{n-m}v$, and $x \Rightarrow^k u\overline{0}^{n-m}v$ for all $k \geq m+1$. $\square$

Define bijective homomorphisms $\delta_1 : (N_1 \cup T)^* \rightarrow \{0,1,\hat{0},\hat{1}\}^*$, and $\delta_2 : (N_2 \cup T)^* \rightarrow (\{0,1\} \cup T')^*$ as $\delta_1(\varepsilon) = \varepsilon$, $\delta_1(a) = \hat{g}(a)$, $\delta_1(A) = f_1(A)$, $\delta_1(X\alpha) = \delta_1(X)\delta_1(\alpha)$, $\delta_2(\varepsilon) = \varepsilon$, $\delta_2(a) = h(a)$, $\delta_2(B) = f_1(B)$, and $\delta_2(Y\beta) = \delta_2(Y)\delta_2(\beta)$ for all $a \in T$, $A \in N_1$, $X \in N_1 \cup T$, $\alpha \in (N_1 \cup T)^*$, $B \in N_2$, $Y \in N_2 \cup T$, and $\beta \in (N_2 \cup T)^*$.

Next, we discuss the derivation in G. Assume that any derivation in G_1 or G_2 in the next paragraphs is the leftmost derivation.

First derivation step is allways $S \Rightarrow f_1(S_1)f_1(S_2)$. In case that $S \rightarrow \varepsilon \in M$, the derivation step $S \Rightarrow \varepsilon$ is also possible. We can express $f_1(S_1)f_1(S_2)$ as $\delta_1(S_1)\delta_2(S_2)$.

Suppose that $S_1 \Rightarrow^i_{G_1} xA\alpha$, where $x \in T^*$, $A \in N_1$, $\alpha \in N_1^*$, and $S \Rightarrow^j_G \delta_1(xA\alpha)\delta_2(S_2)$ for some $i \geq 0$, $j \geq 1$. We can express $\delta_1(xA\alpha)\delta_2(S_2)$ as $\delta_1(x)f(A)1\delta_1(\alpha)\delta_2(S_2)$. The matrices of 3 and 4 are the only applicable. It follows from Claim 2 that there can be used only a matrix simulating some $A \rightarrow a\beta \in P_1$, $a \in T$, $\beta \in N_1^*$. By an application of such matrix we obtain $\delta_1(x)f(A)f'(A)0'1'\hat{g}(a)f_1(\beta)\delta_1(\alpha)\delta_2(S_2)$. By Claim 2, this derives $\delta_1(x)\hat{g}(a)f_1(\beta)\delta_1(\alpha)\delta_2(S_2)$. The newly generated string can be expressed as $\delta_1(xa\beta\alpha)\delta_2(S_2)$. The derivation $\delta_1(xA\alpha)\delta_2(S_2) \Rightarrow^+_G \delta_1(xa\beta\alpha)\delta_2(S_2)$ corresponds to $xA\alpha \Rightarrow_{G_1} xa\beta\alpha$ $[A \rightarrow a\gamma]$. At this point, we see that $S \Rightarrow^+_G \delta_1(x)\delta_2(S_2)$ for all $x \in L(G_1)$.

Express $x \in L(G_1)$ as $x = uav$, where $a \in T$, $u, v \in T^*$. Suppose that $S_2 \Rightarrow^i_{G_2} uA\alpha$, $A \in N_2$, $\beta \in N_2^*$, and $\delta_1(x)\delta_2(S_2) \Rightarrow^j_G \delta_1(av)\delta_2(uA\alpha)$ for some $i, j \geq 0$. We can express $\delta_1(av)\delta_2(uA\alpha)$ as $\hat{g}(a)\delta_1(v)\delta_2(u)f(A)1\delta_2(\alpha)$. The only applicable matrices are those of (3) and (4). It follows from claims 2 and 4 that only a matrix simulating some $A \rightarrow a\beta \in P_2$, $\beta \in N_2^*$, can be applied. Otherwise, terminal string cannot be derived. By an application of such matrix we obtain $\hat{g}(a)\delta_1(v)\delta_2(u)f(A)f'(A)0'1'\overline{g}(a)h(a)f_1(\beta)\delta_2(\alpha)$. By claims 4 and 2, this derives $\delta_1(v)\delta_2(u)h(a)f_1(\beta)\delta_2(\alpha) = \delta_1(v)\delta_2(ua\beta\alpha)$. The derivation $\delta_1(av)\delta_2(uA\alpha) \Rightarrow^+_G \delta_1(v)\delta_2(ua\beta\alpha)$ corresponds to $uA\alpha \Rightarrow_{G_2} ua\beta\alpha$ $[A \rightarrow a\beta]$. At this point, we see that $S \Rightarrow^+_G \delta_1(x)\delta_2(S_2) \Rightarrow^+_G \delta_2(x)$, where $x \in L(G_1)$, and $x \in L(G_2)$. Since $\delta_2(x) = h(x)$, $S \Rightarrow^+_G h(x)$, $x \in L(G_1) \cap L(G_2)$. Therefore, the theorem holds. $\square$

Though this paper proves that matrix grammars with no more than nine nonterminals and six context productions using the leftmost derivation of type 2 describe recursively enumerable languages, there is given no algorithm that transforms a Chomsky grammar of type 0 to a reduced matrix grammar.

REFERENCES

[1] Dassow, J., Păun, G.: *Regulated Rewriting in Formal Language Theory*, Springer-Verlags Berlin, 1989.

[2] Meduna, A.: *Automata and Languages: Theory and Applications*, Springer, London, 1999.

[3] Salomaa, A.: *Formal Languages*, Academic Press, New York, 1973.