RAIRO
INFORMATIQUE THÉORIQUE

GHEORGHE PĂUN

On simple matrix languages versus scattered context languages

RAIRO – Informatique théorique, tome 16, nº 3 (1982), p. 245-253.

<http://www.numdam.org/item?id=ITA_1982__16_3_245_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 16, n° 3, 1982, p. 245 à 253)

ON SIMPLE MATRIX LANGUAGES VERSUS SCATTERED CONTEXT LANGUAGES (*)

by Gheorghe PĂUN ([1])

Communicated by J. BERSTEL

Abstract. — *We prove that the family of simple matrix languages of Ibarra is strictly included in the family of scattered context languages of Greibach and Hopcroft.*

Résumé. — *Nous prouvons que la famille des langages simplement matriciels de Ibarra est strictement contenue dans la famille des « scattered context » langages de Greibach et Hopcroft.*

1. INTRODUCTION

In formal language theory, many restrictions in the derivation of context-free grammars were introduced in order to increase the generative capacity of these grammars. Despite the great attention paid to the relationships between different such restrictions, there still are many open problems in this area. The present paper aims to clarify the relation between two known and important restrictions, namely, the simple matrix grammars of Ibarra [4] and the scattered context grammars of Greibach and Hopcroft [2]. We prove that each simple matrix language is a scattered context one, but the converse is not true. The result is an expected one taking into account the large generative capacity of scattered context grammars [2].

2. DEFINITIONS

Following [4], a simple matrix grammar of order n is an $(n+3)$-tuple $G=(V_1, V_2, \ldots, V_n, \Sigma, S, M)$, where:

(a) $V_1, \ldots, V_n, \Sigma$ are mutually disjoint nonempty vocabularies, the elements

(*) Received in January 1980, revised in January 1981.

[1] Institute of Mathematics, Str. Academiei 14, Bucuresti R-70109 Romania.

of the set $V_N = \bigcup_{i=1}^{n} V_i$ being called nonterminals and the elements of Σ being called terminals;

(*b*) $S \notin V_N \cup \Sigma$ is the start symbol of the grammar;

(*c*) M is a finite set of matrices of the form $(S \to x_1 x_2 \ldots x_n)$ or of the form $(A_1 \to x_1, \ldots, A_n \to x_n)$, $A_i \in V_i$, $x_i \in (V_i \cup \Sigma)^*$, $i=1, 2, \ldots, n$, and the number of nonterminals in x_i (denoted by $N(x_i)$) is equal to the number of nonterminals in x_j for all values of i, j. (For a vocabulary V we denote by V^* the free monoid generated by V under the operation of concatenation and the null element λ.) The vector of the j-th nonterminals of each x_i as above is denoted by $poz_j(x_1, \ldots, x_n)$.

For y_i, $z_i \in (V_i \cup \Sigma)^*$, $i=1, 2, \ldots, n$, one writes $y_1 y_2 \ldots y_n \Rightarrow z_1 z_2 \ldots z_n$ iff $y_i = u_i A_i v_i$, $z_i = u_i x_i v_i$, $u_i \in \Sigma^*$, $v_i \in (V_i \cup \Sigma)^*$ for each $i=1, 2, \ldots, n$, and $(A_1 \to x_1, \ldots, A_n \to x_n) \in M$. Also, we write $S \Rightarrow x_1 x_2 \ldots x_n$ iff $(S \to x_1 x_2 \ldots x_n) \in M$. The language generated by the grammar G is $L(G) = \{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \}$ where $\overset{*}{\Rightarrow}$ is the reflexive transitive closure of $\Rightarrow$. We denote by $\mathscr{S}\mathscr{M}(n)$ the family of languages generated by simple matrix grammars of order n and by $\mathscr{S}\mathscr{M}$ the family of all simple matrix languages.

In words, a simple matrix grammar has the nonterminal vocabulary partitioned into n disjoint sets and the i-th rules in its matrices contain only nonterminals in the i-th set of this partition. Moreover, the rules are used in the leftmost manner for the sentential form subwords containing nonterminals in the corresponding set.

Example: Let $G = (V_N, \Sigma, S, P)$ be a context-free grammar and let c be a symbol not in Σ. The language

$$D(L(G)) = \{ xcx \mid x \in L(G) \}$$

is not context-free for all G, but it can be generated by the simple matrix grammar of order 2:

$$G' = (V_1, V_2, \Sigma \cup \{c\}, S, M),$$

where

$$V_1 = \{ A' \mid A \in V_N \},\ V_2 = \{ A'' \mid A \in V_N \}$$

and $M = \{ (S \to S' c S'') \} \cup \{ (A' \to x', A'' \to x'') \mid A \to x$ is a rule in P and x', x'' are obtained by replacing all nonterminal symbols B in x by B', respectively, by $B'' \}$. Indeed, the leftmost derivation "on V_1 and on V_2" ensures the equality of the strings generated in the left and in the right hand side of the symbol c.

According to [2], a scattered context grammar is a quadruple $G=(V_N, \Sigma, S, M)$, where V_N, Σ are finite nonempty disjoint vocabularies (V_N is the nonterminal and Σ is the terminal vocabulary), $S \in V_N$ is the start symbol and M is a finite set of matrix rules of the form $(A_1 \to x_1, \ldots, A_n \to x_n)$, $n \geqq 1$, $A_i \in V_N$, $x_i \in (V_N \cup \Sigma)^*$. If $z=z_1 A_1 z_2 \ldots z_n A_n z_{n+1}$, $w=z_1 x_1 z_2 x_2 \ldots z_n x_n z_{n+1}$, $z_i \in (V_N \cup \Sigma)^*$ for each i and $(A_1 \to x_1, \ldots, A_n \to x_n) \in M$, then we write $z \Rightarrow w$. The language generated by the grammar G is $L(G)=\{x \in \Sigma^* \mid S \overset{*}{\Rightarrow} x\}$. Let $\mathscr{S}$ be the family of languages generated by λ-free scattered context grammars.

The key feature of the scattered context grammars is the matrices use manner: the rules of a matrix replace occurrences of their left-hand sides *in the order in which the rules appear* in that matrix.

Example: Let us consider a language $L \subseteqq V^*$. For $x, y \in V^*$ we define:

$$\operatorname{Shuf}(x, y)=\{x_1 y_1 x_2 y_2 \ldots x_n y_n \mid n \geqq 1,\\ x_i, y_i \in V^*, x=x_1 \ldots x_n, y=y_1 \ldots y_n\}.$$

For $L_1, L_2 \subseteqq V^*$ we put:

$$\operatorname{Shuf}(L_1, L_2)=\bigcup_{\substack{x \in L_1\\ y \in L_2}} \operatorname{Shuf}(x, y)$$

and we define:

$$\operatorname{Shuf}^*(L)=\bigcup_{n=1}^{\infty} \operatorname{Shuf}^n(L),$$

where $\operatorname{Shuf}^1(L)=\operatorname{Shuf}(L, \{\lambda\})$, $\operatorname{Shuf}^{n+1}(L)=\operatorname{Shuf}(\operatorname{Shuf}^n(L), L)$, $n \geqq 1$.

For any finite language L, the language $\operatorname{Shuf}^*(L)$ is in the family $\mathscr{S}$. Indeed, for given L, the scattered context grammar $G=(\{S\}, V, S, M)$ with $M=\{(\mathrm{S} \to \mathrm{SS})\} \cup \{(\mathrm{S} \to \mathrm{a}_1, S \to a_2, \ldots, S \to a_n) \mid n \geqq 1, a_i \in V, a_1 a_2 \ldots a_n \in L\}$, clearly generates the language $\operatorname{Shuf}^*(L)$. (For simple languages L, the language $\operatorname{Shuf}^*(L)$ is not context free. For example, $\operatorname{Shuf}^*(\{abc\}) \cap \{a^n b^m c^p \mid n, m, p \geqq 1\}=\{a^n b^n c^n \mid n \geqq 1\}$ hence $\operatorname{Shuf}^*(\{abc\})$ is not a context-free language although $\{abc\}$ is a singleton.)

3. EACH SIMPLE MATRIX LANGUAGE IS A SCATTERED CONTEXT LANGUAGE

In [4] one claims, without proof, that each simple matrix language is a matrix language. We feel that this does not hold. For instance, we believe that there are context-free languages L for which the language $D(L)$ defined as in the above first example is not a matrix one (*see* similar arguments in [6]). Moreover, in [4] it is shown that the family $\mathcal{S}\mathcal{M}$ is included in the family of deterministic context-sensitive languages. As the relation between deterministic context-sensitive languages and the scattered context languages is an open problem, the relation between the families $\mathcal{S}\mathcal{M}$ and $\mathcal{S}$ is a significant question.

The problem is answered below. We shall prove that the family $\mathcal{S}\mathcal{M}$ is strictly included in $\mathcal{S}$. A lemma proved in [5] is needed in this aim.

A simple matrix grammar is called purely leftmost iff for any matrix $(A_1 \rightarrow x_1, \ldots, A_n \rightarrow x_n)$ in M and for any $w_1, w_2, \ldots, w_n$ such that $S \overset{*}{\Rightarrow} w_1 w_2 \ldots w_n$, $w_i \in (V_i \cup \Sigma)^*$, we have either :

$$poz_1(w_1, \ldots, w_n) = (\mathrm{A}_1, \ldots, \mathrm{A}_n)$$

or :

$$poz_1(w_1, \ldots, w_n) = (B_1, \ldots, B_n),\ A_i \neq B_i \text{ for all } i = 1, 2, \ldots, n.$$

LEMMA 1 [5]: *For any simple matrix grammar G there is a purely leftmost simple matrix grammar G′ such that* $L(G) = L(G')$.

On the other hand, by extending to simple matrix grammars the arguments used for eliminating the rules of the form $A \rightarrow \lambda$ and $A \rightarrow B$ from context-free grammars [9], one can easily see that for any simple matrix grammar G there is an equivalent simple matrix grammar G' which does not contain matrices of the form $(A_1 \rightarrow \lambda, \ldots, A_n \rightarrow \lambda)$ or of the form $(A_1 \rightarrow B_1, \ldots, A_n \rightarrow B_n)$, $A_i, B_i \in V_i$. This assertion can be also obtained as a consequence of the proof of Theorem 1.1 in [4]: by Lemma 1.2 [4], for each $L \in \mathcal{S}\mathcal{M}(n)$ we can construct an n-context-free language $L' \subseteq [V^*]^n$ such that $L = \{x_1 \ldots x_n \mid (x_1, \ldots, x_n) \in L'\}$ and L' is the image of a usual context-free language L'' by a certain operator $\mathcal{T}_n$. Giving a context-free grammar G for L'' which does not contain rules of the form $A \rightarrow \lambda$, $A \rightarrow B$, by means of the construction in Lemma 1.1 [4], we can get a simple matrix grammar of order n for the language L which does not contain matrices of the form $(A_1 \rightarrow \lambda, \ldots, A_n \rightarrow \lambda)$, $(A_1 \rightarrow B_1, \ldots, A_n \rightarrow B_n)$.

We shall prove the inclusion $\mathcal{S}\mathcal{M} \subsetneqq \mathcal{S}$ in two phases: first we shall reduce the problem to the inclusion $\mathcal{S}\mathcal{M}(2) \subseteq \mathcal{S}$, then we shall simulate the leftmost derivation in a simple matrix grammar of order 2 by means of matrices in a scattered context grammar.

LEMMA 2: *Any language in* $\mathcal{S}\mathcal{M}(n)$, *n given, is the homomorphic image of the intersection of n languages in* $\mathcal{S}\mathcal{M}(2)$.

Proof: Let $G=(V_1, \ldots, V_n, \Sigma, S, M)$ be a purely leftmost simple matrix grammar of order n and let $m_1, \ldots, m_r$ be distinct labels associated to the matrices in M. Clearly, in a purely leftmost simple matrix grammar the control word describing a derivation precisely identifies that derivation. So, for each $i=1, 2, \ldots, n$ we shall construct a simple matrix grammar of order 2 which generates the "i-th component" of a string in $L(G)$ together with the control word associated to this derivation (and some other arbitrary symbols). Intersecting such languages (which are of order 2), we shall obtain the strings in $L(G)$ together with the control words of the corresponding derivations (and some auxiliary symbols). Erasing the control word (and the auxiliary symbols) by a homomorphism, we shall get the language $L(G)$. Following these ideas, let us consider the grammars:

$$G_i=(V_1^i, V_2^i, \Sigma \cup C, S, M_i); \qquad i=1, 2, \ldots, n,$$

where:

$$V_1^i=V_i \cup \{X, Y\}, \qquad V_2^i=\{Z\},$$
$$C=\{c_1, \ldots, c_n, d_1, \ldots, d_n\},$$

and M_i is constructed in the following way:

(1) If $1<i<n$, then for each matrix $m_j : (S \rightarrow w_1 w_2 \ldots w_n)$ in M we introduce in M_i the matrix:

$$(S \rightarrow X c_i w_i d_i Y m_j Z^{N(w_i)+2}).$$

(Remember that $N(w)$ is the number of nonterminal occurrences in the string w.)

If $i=1$, then we introduce the matrix:

$$(S \rightarrow c_1 w_1 d_1 Y m_j Z^{N(w_1)+1}),$$

and if $i=n$, then we introduce the matrix:

$$(S \rightarrow X c_n w_n d_n m_j Z^{N(w_n)+1}).$$

(2) The following matrices belong to M_i for $1<i<n$:

$$(X \rightarrow a X, Z \rightarrow Z), \qquad a \in \Sigma \cup \{c_1, \ldots, c_{i-1}, d_1, \ldots, d_{i-2}\},$$
$$(X \rightarrow d_{i-1}, Z \rightarrow \lambda),$$
$$(Y \rightarrow a Y, Z \rightarrow Z), \qquad a \in \Sigma \cup \{c_{i+1}, \ldots, c_n, d_{i+1}, d_{n-1}\},$$
$$(Y \rightarrow d_n, Z \rightarrow \lambda).$$

Moreover, for $i=1$ we introduce in M_i the above matrices containing the symbol Y and for $i=n$ we introduce the matrices containing the symbol X.

(3) For each matrix $m_j : (A_1 \to x_1, \ldots, A_n \to x_n)$ in M we introduce in M_i the matrix:

$$(A_i \to x_i, Z \to m_j Z^{N(x_i)}).$$

Obviously, we have:

$$L(G_i) = \{ x c_i y d_i z w \mid x \in (\Sigma \cup \{ c_1, \ldots, c_{i-1}, d_1, \ldots, d_{i-1} \})^*, \; y \in \Sigma^*,$$
$$z \in (\Sigma \cup \{ c_{i+1}, \ldots, c_n, d_{i+1}, \ldots, d_n \})^*, \; w \in \{ m_1, \ldots, m_r \}^*,$$
$$w \text{ is the control word associated to the derivation of } y \text{ according to the grammar } G_i \}.$$

As the grammar G is a purely leftmost one, the control word w precisely identifies the derivation and the string y, hence we have:

$$\bigcap_{i=1}^{n} L(G_i) = \{ c_1 y_1 d_1 c_2 y_2 d_2 \ldots c_n y_n d_n w \mid y_1 y_2 \ldots y_n \text{ can be derived from } S \text{ in the grammar } G \text{ and } w \text{ is the control word associated to such a derivation} \}.$$

Let us denote by H the above language and let h be the homomorphism which erases all symbols c_i, d_i *and* m_j. We obtain the equation:

$$L(G) = h(H)$$

and the lemma is proved.

LEMMA 3: $\mathscr{S}\mathscr{M}(2) \subseteqq \mathscr{S}$.

Proof: Let $G = (V_1, V_2, \Sigma, S, M)$ be a simple matrix grammar of order 2 without matrices containing two λ-rules. We construct the following scattered context grammar:

$$G' = (V_N, \Sigma \cup \{ c \}, S, M'),$$

where:

$$V_N = V_1 \cup V_2 \cup \{ S, c' \} \cup \{ a' \mid a \in V_1 \cup V_2 \cup \Sigma \},$$

and the set M' is obtained in the following way:

For each matrix $(S \to w) \in M$, $w \in \Sigma^*$, we introduce in M' the matrix $(S \to w)$ and for each nonterminal matrix $(S \to axby) \in M$, $a \in V_1 \cup \Sigma$, $x \in (V_1 \cup \Sigma)^*$, $b \in V_2 \cup \Sigma$, $y \in (V_2 \cup \Sigma)^*$, we introduce the matrix $(S \to a' x b' y)$.

For each matrix of the form $(A_1 \rightarrow x_1, A_2 \rightarrow x_2)$, we replace x_1 by c if $x_1=\lambda$ and we replace x_2 by c if $x_2=\lambda$. Let $(A_1 \rightarrow bx, A_2 \rightarrow ey)$ be the obtained matrix (the initial matrix if $x_1 \neq \lambda$, $x_2 \neq \lambda$), $b \in V_1 \cup \Sigma \cup \{c\}$, $x \in (V_1 \cup \Sigma)^*$, $e \in V_2 \cup \Sigma \cup \{c\}$, $y \in (V_2 \cup \Sigma)^*$.

We introduce in M' the following matrices:

$$(a' \rightarrow a, A_1 \rightarrow b'x, d' \rightarrow d, A_2 \rightarrow e'y),\ a, d \in \Sigma \cup \{c\},$$
$$(A_1' \rightarrow b'x, d' \rightarrow d, A_2 \rightarrow e'y),\ d \in \Sigma \cup \{c\},$$
$$(a' \rightarrow a, A_1 \rightarrow b'x, A_2' \rightarrow e'y),\ a \in \Sigma \cup \{c\},$$
$$(A_1' \rightarrow b'x, A_2' \rightarrow e'y).$$

We consider also the terminal matrices:

$$(a' \rightarrow a, b' \rightarrow b),\ a, b \in \Sigma \cup \{c\}.$$

Clearly, the grammar G' is a λ-free one. Let h be the homomorphism which erases the letter c. We have:

$$L(G)=h(L(G')).$$

(Any sentential form according to the scattered context grammar G' contains exactly two primed symbols. The use of a matrix in M' is allowed only in the right hand side of these symbols. This restriction ensures the simulation of a leftmost application of the corresponding two-rules matrices in M. If a matrix is not used in the leftmost manner – that is, the primes do not circulate correctly – then the derivation cannot be terminated.)

As each matrix in M contains at most a λ-rule, it follows that $|h(x)| \geqq 1/2\,|x|$. (We denoted by $|z|$ the length of the string z.) Consequently, h is a linear erasing homomorphism. As the family $\mathcal{S}$ is closed under such homomorphisms [2], it follows that $h(L(G')) \in \mathcal{S}$, hence $L(G) \in \mathcal{S}$ and the lemma is proved.

THEOREM: *The family $\mathcal{S}\mathcal{M}$ is strictly included in the family $\mathcal{S}$.*

Proof: According to the above lemmas, any language $L \in \mathcal{S}\mathcal{M}(n)$ can be written as:

$$L=h\left(\bigcup_{i=1}^{n} L_i\right), \qquad L_i \in \mathcal{S} \text{ for each } i,$$

h being the homomorphism in the proof of Lemma 2. The family $\mathcal{S}$ is closed under intersection [2]. The homomorphism h is a linear erasing one. Indeed, each nonterminal matrix in a simple matrix grammar increases the length of the rewritten string at least by one symbol and there are no λ-matrices. Thus, the homomorphism h erases $2n$ symbols c_i, d_i and a control word shorter than the longest string derived between $c_i, d_i, i=1, 2, \ldots, n$. As the family $\mathcal{S}$ is closed under linear erasing homomorphisms [2], it follows that $L \in \mathcal{S}$ hence $\mathcal{S}\mathcal{M} \subseteq \mathcal{S}$.

The inclusion is, obviously, proper. For instance, the language:

$$L=\{a^n b^n c^n \mid n \geqq 1\}^*,$$

is not in $\mathcal{S}\mathcal{M}$ [4], but it is generated by the scattered context grammar:

$$G=(\{S, A, B, C\}, \{a, b, c\}, S, M),$$

with:

$$M=\{(S \to ABC), (S \to \lambda), (S \to AB), (A \to aAb, B \to cB), (A \to ab, B \to c, C \to ABC), (A \to ab, B \to c), (A \to ab, B \to c, C \to AB)\}.$$

4. FINAL REMARKS

Let $\mathcal{S}\mathcal{M}_f$ and $\mathcal{S}_f$ be the families of finite index languages in $\mathcal{S}\mathcal{M}$, respectively, in $\mathcal{S}$. (*See* the index definition in [1].) Let us observe that the above proofs of Lemmas 2 and 3 do not modify the index finiteness. Consequently, $\mathcal{S}\mathcal{M}_f \subseteqq \mathcal{S}_f$. The inclusion is proper since the language $\{a^n b^n c^n \mid n \geqq 1\}^*$ belongs to the family $\mathcal{S}_f$.

The inclusion $\mathcal{S}\mathcal{M}_f \subset \mathcal{S}\mathcal{M}$ is a proper one. Indeed, let us consider the context-free language $L=\{c\}(D\{c\})^+$, where D is the Dyck language over the vocabulary $\{a, b\}$. As in [8] it was proved, D has an infinite index according to the context-free grammars. Let us assume that $L \in \mathcal{S}\mathcal{M}_f$ and let $G=(V_1, \ldots, V_n, \{a, b, c\}, S, M)$ be a simple matrix grammar of finite index generating the language L. Let G_i be the context-free grammars $G_i=(V_i \cup \{S\}, \{a, b, c\}, S, \{S \to w_i \mid (S \to w_1 w_2 \ldots w_n) \in M, w_i \in (V_i \cup \{a, b, c\})^*\} \cup \{A_i \to x_i \mid (A_1 \to x_1, \ldots, A_n \to x_n) \in M, A_i \in V_i, x_i \in (V_i \cup \{a, b, c\})^*\})$. Let g be a gsm which maps a string w into x providing that $w=ycxcz$, $x \in \{a, b\}^*$, $y, z \in \{a, b, c\}^*$. It is easy to see that we have $D=\bigcup_{i=1}^{n} g(L(G_i))$. As G is a finite index simple matrix grammar (we assume it to be purely leftmost), it follows that each grammar G_i has a finite index too. The family of finite index context-free languages is a full-AFL [3]. It follows D is a finite index context-free language. Contradiction with the result in [8]. Consequently, $L \in \mathscr{S}\mathscr{M}(1) - \mathscr{S}\mathscr{M}_f$.

From the above remarks it follows that the inclusion $\mathscr{S}\mathscr{M}_f \subset \mathscr{S}_f$ does not imply $\mathscr{S}\mathscr{M}(k) \subset \mathscr{S}_f$, for some $k \geqq 1$. In fact, we feel that such an inclusion does not hold.

Open problem: Is the family $\mathscr{S}\mathscr{M}$ included in the family of languages generated by λ-free context-free matrix grammars in the appearance checking mode (*See* Chap. V in [9] or the monograph [7] for the theory of matrix grammars and languages.)

Note: Useful remarks by the referee are acknowledged, allowing us to make clearer some parts of the paper.

REFERENCES

1. B. Brainerd, *An Analog of a Theorem about Context-Free Languages*, Information and Control, Vol. 11, 1968, pp. 561-568.
2. S. Greibach and J. Hopcroft, *Scattered Context Grammars*, J. of Computer and System Science, Vol. 3, 1969, pp. 232-247.
3. J. Gruska, *A Few Remarks on the Index of Context-Free Grammars and Languages*, Information and Control, Vol. 19, 1971, pp. 216-223.
4. O. Ibarra, *Simple Matrix Languages*, Information and Control, Vol. 17, 1970, pp. 259-294.
5. Gh. Păun, *On the Generative Capacity of Simple Matrix Grammars of Finité Index*, Information Processing Letters, Vol. 7, No. 2, 1978, pp. 100-102.
6. Gh. Păun, *On the Family of Finite Index Matrix Languages*, J. of Computer and System Science, Vol. 18, 1979, pp. 267-280.
7. Gh. Păun, *Matrix Grammars*, The Scientific and Enciclopaedic Publishing House, Bucharest, 1981 (in Romanian).
8. A. Salomaa, *On the Index of Context-Free Languages*, Information and Control, Vol. 14, 1969, pp. 474-477.
9. A. Salomaa, *Formal Languages*, Academic Press, New York and London, 1973.