Acta Cybernetica, Tom. 10, Fasc. 1-2, Szeged, 1991

A note on fully initial grammars

S. Vicolov*

We (negatively) solve two conjectures of Mateescu and Paun [3], then we give characterizations in terms of syntactic semigroup of some families of regular fully initial languages.

1 Definitions and notations

For a vocabulary V, we denote by $V^*(V^+)$ the free monoid (semigroup) generated by V under the operation of concatenation; λ is the null element $(V^+ = V^* - \{\lambda\})$. The strings of V^* are called words. The length of a word $x \in V^*$ is denoted by $|x|$.

If we consider a Chomsky grammar $G = (V_N, V_T, S, P)$, then the usual language generated by G is defined by

$$L(G) = \{x \in V_T^* | S \overset{*}{\Longrightarrow} x\}.$$

The fully initial language generated by G is

$$L_{in}(G) = \{x \in V_T^* | A \overset{*}{\Longrightarrow} x \text{ for some } A \in V_N\}.$$

The study of fully initial languages was proposed by S. Horvath and has been done in a series of papers [1], [2], [3], [4].

Clearly, $L(G) \subseteq L_{in}(G)$. The family of fully initial languages generated by grammars of type $i, i = 0, 1, 2, 3$ is denoted by $\mathcal{FL}_i$.

Usually, the right-linear and the left-linear grammars generate the same family of languages. For fully initial grammars this is not true, therefore we shall distinguish several classes of "type-3" grammars.

A grammar $G = (V_N, V_T, S, P)$ is called right-linear (left-linear) if $P \subseteq V_N \times (V_T^* \cup V_T^* V_N)(P \subseteq V_N \times (V_T^* \cup V_N V_T^*))$. We denote by $\mathcal{FL}_{rlin}, \mathcal{FL}_{llin}$ the corresponding families of fully initial languages. A grammar $G = (V_N, V_T, S, P)$ is called right-regular (left-regular) if $P \subseteq V_N \times (V_T \cup V_T V_N)(P \subseteq V_N \times (V_T \cup V_N V_T))$. The corresponding families of fully initial languages are denoted by $\mathcal{FL}_{rreg}, \mathcal{FL}_{lreg}$. $\mathcal{FL}_3$ is, in fact, $\mathcal{FL}_{rlin} \cup \mathcal{FL}_{llin}$. Following [3] we shall consider the next families, too:

$$\mathcal{FL}_{reg}^{\cap} = \mathcal{FL}_{rreg} \cap \mathcal{FL}_{lreg}$$

$$\mathcal{FL}_{reg}^{\cup} = \mathcal{FL}_{rreg} \cup \mathcal{FL}_{lreg}.$$

The sets of prefixes, suffixes and subwords of a given word x are denoted by Init(x) , Fin(x), Sub(x), respectively, and these notations will be extended in the

*University of Bucharest, Faculty of Mathematics, Str. Academiei 14,70109 Bucuresti ROMANIA

natural way to languages. When considering only proper prefixes, suffixes and subwords, we shall write Initp(x), Finp(x) and Subp(x), respectively.

Let L be a language of V^+. The congruence $\sim_L$ defined over V^+ by: $u \sim_L v$ if and only if, for every $x, y \in V^*, xuy \in L \Leftrightarrow xvy \in L$, is called the syntactic congruence of L. The syntactic semigroup of L is the quotient semigroup $A^+/\sim_L$.

For further details in syntactic semigroup theory, the reader is referred to [5].

2 Necessary conditions for the context- free case

We shall reproduce here the necessary conditions for a language to be in $\mathcal{FL}_2$, which were considered in [3]. Finally we shall prove that two of the conjectures formulated there are not true.

Lemma 1 *For each language $L \in \mathcal{FL}_2, L \subseteq V^*$, there are two positive integers p, q such that each $z \in L, |z| > p$, can be written as $z = uvwxy, u, v, w, x, y \in V^*$, so that*

(i) $|vwx| \leq q, \quad |vx| > 0,$

(ii) for all $k \geq 0, uv^kwx^ky \in L$ *and* $v^kwx^k \in L$.

Definition 1 *For a given language $L \subseteq V^*$, let*

$$\mathrm{Min}(L) = \{z \in L | \mathrm{Subp}(z) \cap L = \emptyset\}$$

and define

$$R_1(L) = \mathrm{Min}(L)$$

$$R_i(L) = R_{i-1}(L) \cup \mathrm{Min}(L - R_{i-1}(L)), i \geq 2.$$

We say that L has property R if and only if all the sets $R_i(L), i \geq 1$, are finite.

Lemma 2 *If $L \in \mathcal{FL}_2$, then L has property R.*

In [3] it is also proved that none of these conditions is sufficient for a language to be in $\mathcal{FL}_2$, and one formulates the following conjectures:

(1) If L is a context-free language which fulfils the condition in Lemma 1, then $L \in \mathcal{FL}_2$.

(2) For arbitrary languages, the condition in Lemma 1 is stronger than property R.

Proposition 1 *Conjecture (2) is not true.*

Proof. Consider the languages

$$L_1 = \{cd^nae^{k_1}b \ldots e^{k_n}b | n \geq 0, k_1, \ldots k_n \geq 0\},$$

$$L = L_1 \cup \{e^nb | n \geq 0\} \cup \{d^nab^n | n \geq 0\}.$$

We shall prove that L fulfils the condition in Lemma 1. Let us take $p = 2$ and $q = 3$. For $z = e^nb$ or $z = d^nab^n$ we clearly have all conditions in lemma fulfilled. If $z = cd^nae^{k_1}b \ldots e^{k_n}b$, then $|z| > p$ implies $n \geq 1$. There are two cases.

1. For all $i, 1 \leq i \leq n, k_i = 0$. Therefore $z = cd^n ab^n$. We take $u = cd^{n-1}, v = d, w = a, x = b, y = b^{n-1}$. It follows that $z = uvwxy, |vx| > 0, |vwx| \leq q, uv^k wx^k y = cd^{n-1}d^k ab^k b^{n-1} \in L$ and $v^k wx^k = d^k ab^k \in L$ for every $k \geq 0$.

2. There is an $i, 1 \leq i \leq n$, such that $k_i \geq 1$. We consider $u = cd^n ae^{k_1}b \ldots e^{k_{i-1}}be^{k_i-1}, v = e, w = b, x = \lambda, y = e^{k_{i+1}}b \ldots e^{k_n}b$. Then $z = uvwxy, |vx| > 0, |vwx| \leq q, uv^k wx^k y = cd^n ae^{k_1}b \ldots e^{k_{i-1}}be^{k_i-1}e^k be^{k_{i+1}}b \ldots e^{k_n}b \in L$ and $v^k wx^k = e^k b \in L$ for all $k \geq 0$.

On the other hand, L does not observe property R. Indeed, it is clear that $R_1(L) = \{a, b\}$ and $R_2(L) = \{a, b, ca, eb, dab\}$. $\mathrm{Min}(L - R_2(L)) \supseteq \{cd^n a(eb)^n | n \geq 1\}$ since, for all $n \geq 1, z = cd^n a(eb)^n$ implies $z \in L - R_2(L)$, $\mathrm{Subp}(z) \cap L_1 = \emptyset$ and $\mathrm{Subp}(z) \cap (L - L_1) = \{a, b, eb\} \subseteq R_2(L)$. It follows that $R_3(L)$ is an infinite set.

In conclusion, L fulfils the condition in Lemma 1 without observing property R.

Proposition 2 *Conjecture (1) is not true.*

Proof. We shall consider the same language L as in the above proof. Let $G = (V_N, V_T, S, P)$, where $V_N = \{A, B, C, S\}, V_T = \{a, b, c, d, e\}$ and $P = \{S \longrightarrow cA, A \longrightarrow dAB, B \longrightarrow eB, A \longrightarrow a, B \longrightarrow b, S \longrightarrow B, S \longrightarrow C, C \longrightarrow dCb, C \longrightarrow a\}$. It is easy to see that $L = L(G)$. Consequently, L is a context - free language which fulfils the condition in Lemma 1. L has not property R, therefore, according to Lemma 2, $L \notin \mathcal{FL}_2$. In conclusion, the proposition is proved.

Remark 1 *Note that* $L_{in}(G) = L \cup \{d^n ae^{k_1}b \ldots e^{k_n}b | n \geq 0, k_i \geq 0, 1 \leq i \leq n\}$.

Remark 2 *The negative answer of these two conjectures raises another problem: a context-free language which satisfies simultaneously the condition in Lemma 1 and the condition R, is in* $\mathcal{FL}_2$*?*

Proposition 3 *The condition R and the condition in Lemma 1 fulfilled in the same time, are not sufficient for a context-free language to be in* $\mathcal{FL}_2$.

Proof. Consider the language

$$L_2 = \{cd^n ae^{k_1}b \ldots e^{k_n}b | n \geq 0, k_1, \ldots k_n \geq 0\} \cup \{d^n ab^n | n \geq 0\} \cup \{e, b\}^+.$$

Note that $L_2 = L \cup \{e, b\}^+$, where L is the language used in the above proofs. L and $\{e, b\}^+$ are context-free languages. Consequently, L_2 is a context-free language, too. We have pointed out in the proof of Proposition 1 that L satisfies the condition in Lemma 1; it is easy to see that $\{e, b\}^+$ also satisfies this condition. In conclusion, L_2 fulfils the condition in Lemma 1.

L_2 observes property R. Indeed, $R_1(L_2) = \{a, e, b\}$ and $R_i(L_2) = \{cd^n ae^{k_1}b \ldots e^{k_n}b | 0 \leq n \leq i - 2, 0 \leq n + k_1 + \ldots + k_n \leq i - 1\} \cup \{d^n ab^n | 0 \leq n \leq i - 1\} \cup \{u \in \{e, b\}^+, |u| \leq i\}, i \geq 2$.

The last equality can be obtained by induction. We denote by A_i the right term of the equality. It is clear that $R_2(L_2) = A_2$. Suppose that $R_j(L_2) = A_j$, for an arbitrary $j \geq 2$. We must show that $R_{j+1}(L_2) = A_{j+1}$. According to definition and to the above supposition we have $R_{j+1}(L_2) = R_j(L_2) \cup \mathrm{Min}(L_2 - R_j(L_2)) = A_j \cup \mathrm{Min}(L_2 - A_j)$. Also using the inclusions $A_{j+1} \subseteq L_2$ and $R_{j+1}(L_2) \subseteq L_2$, we conclude that it is sufficient to prove that $z \in A_{j+1}$ iff $z \in A_j \cup \mathrm{Min}(L_2 - A_j)$, for all $z \in L_2$. There are three cases.

(1) $z = cd^n ae^{k_1}b \ldots e^{k_n}b$. $z \in A_{j+1}$ if $n \leq j-1$ and $n + k_1 + \ldots + k_n \leq j$. Obviously, $\mathrm{Subp}(z) \cap L_2 = \mathrm{Sub}(e^{k_1}b \ldots e^{k_n}b) \cup \{d^t ab^t | 1 \leq n, k_1 + \ldots + k_t = 0\}$.

Suppose that $z \in A_{j+1}$. We obtain $\mathrm{Subp}(z) \cap L_2 \subseteq \{u \in \{e,b\}^+ || u| \leq j\} \cup \{d^t ab^t | t \leq j-1\} \subseteq A_j$. It follows that $z \in A_j \cup \mathrm{Min}(L_2 - A_j)$.

Conversely, suppose that $z \in A_j \cup \mathrm{Min}(L_2 - A_j)$. If $z \in A_j$, then $z \in A_{j+1}$. If $z \in \mathrm{Min}(L_2 - A_j)$, we obtain $\mathrm{Subp}(z) \cap L_2 \subseteq A_j$. This implies $\mathrm{Sub}(e^{k_1}b \ldots e^{k_n}b) \subseteq A_j$. Hence $n + k_1 + \ldots + k_n \leq j$ and $n \leq j$. If $n = j$, we have $k_1 + \ldots + k_n = 0$ and $d^j ab^j \in (\mathrm{Subp}(z) \cap L_2) - A_j$, which is a contradiction. Consequently, $n \leq j-1$ and $n + k_1 + \ldots + k_n \leq j$.

Thus we proved that, in this case, $z \in A_{j+1}$ iff $z \in R_{j+1}(L_2)$.

(2) $z = d^n ab^n$. $z \in A_{j+1}$ iff $n \leq j$. $n \leq j$ iff $\mathrm{Subp}(z) \cap L_2 = \{d^k ab^k | k \leq j-1\} (\subseteq A_j)$ iff $z \in A_j \cup \mathrm{Min}(L_2 - A_j)$.

(3) $z \in \{e,b\}^+ . z \in A_{j+1}$ iff $|z| \leq j+1$ iff $\mathrm{Subp}(z) \cap L_2 \subseteq \{u \in \{e,b\}^+ | \; |u| \leq j\} (\subseteq A_j)$ iff $z \in A_j \cup \mathrm{Min}(L_2 - A_j)$.

In conclusion, L_2 is a context-free language which satisfies both the condition in Lemma 1 and the condition R.

On the other hand, $L_2 \notin \mathcal{FL}_2$. Assume the contrary and consider a type-2 grammar $G = (V_N, V_T, S, P)$ such that $L_{in}(G) = L_2$. Since $L_2 = \{cd^n ae^{k_1}b \ldots e^{k_n}b | n \geq 0, k_1, \ldots, k_n \geq 0\} \cup \{d^n ab^n | n \geq 0\} \cup \{e,b\}^+$, we conclude that, for generating the strings of the form $cd^n ae^{k_1}b \ldots e^{k_n}b$, we need derivations such as: $X \stackrel{*}{\Longrightarrow} d^j XB^j, j \geq 1, X \in V_N, B \in V_N, B \stackrel{*}{\Longrightarrow} e^k b, k \geq 1, X \stackrel{*}{\Longrightarrow} w, w \in T_T^+$. It follows that $d^j w(e^k b)^j \in L_{in}(G) - L_2$, which is a contradiction.

Thus, the proof is completed.

3 Characterizations of languages in $\mathcal{FL}_{rreg}$, $\mathcal{FL}_{lreg}$, $\mathcal{FL}^{\cap}_{reg}$

We shall consider here a characterization of these families in terms of the syntactic semigroup. For proving it we shall use the following lemma, presented in [3].

Lemma 3 *(i) $L \in \mathcal{FL}_{rreg}$ if and only if L is regular and $L = Fin(L)$.*
(ii) $L \in \mathcal{FL}_{lreg}$ if and only if L is regular and $L = Init(L)$.
(iii) $L \in \mathcal{FL}^{\cap}_{reg}$ if and only if L is regular and $L = Sub(L)$.

We also shall use two well-known results in the theory of syntactic semigroups [5]:

Lemma 4 *Let $L \subseteq V^+$. L is regular if and only if its syntactic semigroup is finite.*

Lemma 5 *Let $L \subseteq V^+$ be a language and denote by φ the canonical homomorphism φ: $V^+ \longrightarrow V^+ / \sim_L$. Then $V^+ - L = \varphi^{-1}(\varphi(V^+ - L))$.*

We shall consider below that L, $\mathrm{Fin}(L)$, $\mathrm{Init}(L)$ and $\mathrm{Sub}(L)$ do not contain the null word λ.

Proposition 4 *Let L be a language over V. Denote by S the syntactic semigroup of L, by φ the canonical homomorphism $\varphi : V^+ \longrightarrow V^+/\sim_L = S$ and $P = \varphi(L)$. Then, we have:*

(i) $L \in \mathcal{FL}_{rreg}$ if and only if S is finite and $S(S-P) \subseteq S-P$.

(ii) $L \in \mathcal{FL}_{lreg}$ if and only if S is finite and $(S-P)S \subseteq S-P$.

(iii) $L \in \mathcal{FL}^{\cap}_{reg}$ if and only if S is finite, S has a zero, 0, and $S - P = \{0\}$.

Proof. (i) According to Lemma 3, part (i), $L \in \mathcal{FL}_{rreg}$ if and only if L is regular and $L = \mathrm{Fin}(L)$. Since we always have $L \subseteq \mathrm{Fin}(L)$, we deduce that $L = \mathrm{Fin}(L)$ is equivalent to "for all $u, v \in V^+, uv \in L \Longrightarrow v \in L$", statement which is also equivalent to "for all $u \in V^+$ and $v \in V^+ - L, uv \in V^+ - L$", i.e. $V^+(V^+ - L) \subseteq V^+ - L$. It follows from the last inclusion that $\varphi(V^+(V^+ - L)) \subseteq \varphi(V^+ - L)$ and hence $\varphi^{-1}(\varphi(V^+(V^+ - L))) \subseteq \varphi^{-1}(\varphi(V^+ - L))$. In turn, the last inclusion implies $V^+(V^+ - L) \subseteq V^+ - L$, since $V^+(V^+ - L) \subseteq \varphi^{-1}(\varphi(V^+(V^+ - L)))$ and $\varphi^{-1}(\varphi(V^+ - L)) = V^+ - L$ (Lemma 5). Consequently, $V^+(V^+ - L) \subseteq V^+ - L$ if and only if $\varphi(V^+)\varphi(V^+ - L) \subseteq \varphi(V^+ - L)(\varphi(V^+(V^+ - L)) = \varphi(V^+)\varphi(V^+ - L)$ since φ is homomorphism of semigroups) if and only if $S(S-P) \subseteq S-P$ (use $\varphi(V^+) = S$ and $\varphi(V^+ - L) = S - P$, from Lemma 5). Thus we proved the equivalence between $L = \mathrm{Fin}(L)$ and $S(S-P) \subseteq S-P$. Using the result in Lemma 4, too, we conclude the proof.

(ii) The proof is symmetrical.

(iii) Suppose that $L \in \mathcal{FL}^{\cap}_{reg}$. According to Lemma 3, part (iii), L is regular and $\mathrm{Sub}(L) = L$. From the last equality it follows that "$u \notin L \Longrightarrow xuy \notin L$, for all $x, y \in V^*$ and $u \in V^+$" (assuming the contrary, we have $xuy \in L$, hence $u \in \mathrm{Sub}(L) = L$, which is a contradiction to $u \notin L$). Take u, v arbitrary in V^+ such that $u \notin L$. From the above statement we obtain $uv \notin L, vu \notin L$ and: "$xuy \notin L, xuvy \notin L, xvuy \notin L$, for every $x, y \in V^+$". Consequently $u \sim_L uv \sim_L vu$ and hence we have $\varphi(u) = \varphi(uv) = \varphi(vu)$, i.e. $\varphi(u) = \varphi(u)\varphi(v) = \varphi(v)\varphi(u)$. Since v is an arbitrary word of $V^+, \varphi(v)$ is an arbitrary element of $\varphi(V^+) = S$. Therefore we deduce that $\varphi(u)$ is a zero of S. A semigroup may contain only one zero. As u is arbitrary in $V^+ - L$ and $\varphi(V^+ - L) = S - P$, we conclude that $S - P$ contains only one element, which is the zero of S. Since L is regular, S is finite. Thus, one of the implications is proved.

Conversely, suppose that S is finite, S has a zero, 0, and $S - P = \{0\}$. Clearly, $(S-P)S \subseteq S-P$ and $S(S-P) \subseteq S-P$. According to the parts (i) and (ii) of this Proposition, it follows that $L \in \mathcal{FL}^{\cap}_{reg}$.

Corollary 1 *Let L be a language of V^+ whose syntactic semigroup is commutative. If $L \in \mathcal{FL}^{\cup}_{reg}$, then in fact L is in $\mathcal{FL}^{\cap}_{reg}$.*

Proof. $L \in \mathcal{FL}^{\cup}_{reg}$ implies $L \in \mathcal{FL}_{rreg}$ of $L \in \mathcal{FL}_{lreg}$. We use Proposition 4, parts (i), (ii), and we obtain $S(S-P) \subseteq S-P$ or $(S-P)S \subseteq S-P$. Since S is commutative, these inclusions hold simultaneously. Using again Proposition 4, parts (i), (ii), we conclude that $L \in \mathcal{FL}^{\cap}_{reg}$.

References

[1] T. Balanescu, M. Gheorghe, Gh. Paun, On fully initial grammars with regulated rewriting, Acta Cybernetica, 9, 2(1989), 157-165.

[2] J. Dassow, On fully initial context-free languages, Papers on Automata and Languages (Ed. I. Peák), X(1988), 3-6.

[3] A. Mateescu, Gh. Paun, Further remarks on fully initial grammars, Acta Cybernetica, 9, 2 (1989), 143-156.

[4] Gh. Paun, A note on fully initial context-free languages, Papers on Automata and Languages, X (1988), 7-11.

[5] J.E. Pin, Varieties of formal languages, North Oxford Academic Publishers, London, 1986, 5-24.

(Received July 7, 1990)