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

Language representations starting from fully initial languages

Gh. Paun S. Vicolov*

Abstract

It is proved that each regular/linear/context-free language is the image of a fully initial regular/linear/context-free language by an inverse homomorphism, as well as the intersection of two regular/linear/context-free fully initial languages, respectively. The converse of the latter assertion is not true for linear and for context-free languages.

1 Fully initial languages

For a context-free grammar $G = (V_N, V_T, S, P)$, one usually define the generated language as

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

S. Horváth proposed to consider also the fully initial language generated by G, that is

$$L_{in}(G) = \{x \in V_T^* | A \overset{*}{\Longrightarrow} x, A \in V_N\}$$

(We denoted V^* the free monoid generated by V under the operation of concatenation; the null element of V^* is denoted by λ and $|x|$ denotes the length of $x \in V^*$. For $U \subseteq V$ and $x \in V^*$ we denote by $|x|_U$ the length of the string obtained by erasing from x all symbols not in U.)

We denote by REG, LIN, CF the families of regular, linear, context-free languages, respectively, and by FIREG, FILIN, FICF the families of fully initial languages generated by right-linear, linear and context-free grammars, respectively.

The fully initial languages were investigated in a series of papers [1], [3], [4], [6], [8]. In [3] it is proved that FICF is not closed under concatenation, intersection by regular sets and inverse homomorphisms; in fact, the proofs in [3] are true also for the family FILIN. The same nonclosure results hold also for the family FIREG (see [4]). On the other hand, FIX $\subset X$, strict inclusion, for each $X \in$ {REG, LIN, CF}, [3], [4].

The above quoted results naturally raise the question of representing languages in a family X, X as above, starting from languages in FIX and using suitable operations. One such representation (characterization, in fact) has been done in [6], where it is proved

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

Theorem 1 *A language $L \subseteq V^*$ is in $X, X \in$ {REG, LIN, CF}, if and only if there is $L' \in FIX$ such that*

$$L = h_c(L' \cap \{c\}V^*)$$

where c is a new symbol and $h_c : (V \cup \{c\})^* \longrightarrow V^*$ is the homomorphism defined by $h_c(a) = a, a \in V, h_c(c) = \lambda$.

Two open problems are then raised in [6]:

(1) Can the homomorphism in above theorem be removed, that is suffices an intersection for obtaining a representation/characterization of X starting from languages in FIX?

(2) What about inverse homomorphism characterizations?

We affirmativelly solve here both these problems: there are such representations (sometimes characterizations).

In what follows, two languages will be considered eqnal if they differ only by the null string λ.

2 Characterization and representation results

Theorem 2 $X = \{h^{-1}(L) | L \in \mathrm{FIX}\ h$ a homomorphism$\}$, $X \in$ {REG, LIN, CF}.

Proof. Each family X as above is closed under inverse homomorphisms [7], hence the inclusion $\supseteq$ is true.

Conversely, let $L \subseteq V^*$, be a context-free language. Denote by $d_a(L)$ the left derivative of L with respect to $a \in V$, that is

$$d_a(L) = \{x \in V^* | ax \in L\}.$$

We have

$$L = \bigcup_{a \in V} \{a\} d_a(L).$$

Each $d_a(L), a \in V$, is a context-free language; let $G_a = (V_{N,a}, V, S_a, P_a)$ be a λ-free grammar for $d_a(L)$. We construct the grammar

$$G = (V_N, V \cup \{c\}, S, P)$$

where c, S are new symbols,

$$V_N = \bigcup_{a \in V} V_{N,a} \cup \{S\}$$

and P contains the following rules:

(1) $S \longrightarrow ac$, if $a \in L, a \in V$,

(2) $S \longrightarrow aS_a, a \in V$,

(3) $A \longrightarrow x'$, for $A \longrightarrow x \in \bigcup_{a \in V} P_a$, and x' is obtained by replacing each terminal b in $x, b \in V$, by cb (the nonterminals in x remain unchanged),

(4) $A \longrightarrow x'$, for $A \longrightarrow x \in \bigcup_{a \in V} P_a$, and x' is obtained by replacing each terminal b in $x, b \in V$, by cb, excepting one occurrence of some $b \in V$ which is replaced by cbc (the nonterminals remain unchanged).

Let $L' = L_{in}(G)$ and consider the homomorphism $h : V^* \longrightarrow (V \cup \{c\})^*$ defined by $h(a) = ac, a \in V$. Clearly, $Im(h) = (V\{c\})^*$ and

$$L_{in}(G) = L(G) \cup \bigcup_{A \in V_N - \{S\}} L_A(G)$$

with

$$L_A(G = \{x \in V^* | A \overset{*}{\Longrightarrow} x \; in \; G\}$$

As each $x \in L_A(G), A \neq S$, is of the form $x = cy, y \in (V \cup \{c\})^*$, it follows that $Im(h) \cap L_A(G) = \emptyset$, hence $h^{-1}(L') = h^{-1}(L(G))$. On the other hand, we have

$$L(G) = L_1 \cup L_2 \cup L_3 \cup L_4$$

where

$$L_1 = L(G) \cap \{x \in (V \cup \{c\})^* | |x|_c = |x|_V - 1\}$$

(the strings in L_1 are produced by using rules of the form (2) and (3), without using rules (1) and (4))

$$L_2 = L(G) \cap \{x \in (V \cup \{c\})^* | |x|_c > |x|_V\}$$

(the strings in L_2 are obtained by using rules of forms (2), (3) and (4), namely at least two times rules of type (4))

$$L_3 = L(G) \cap \{x \in (V \cup \{c\})^* | |x|_c = |x|_V, x = ya, a \in V, y \in (V \cup \{c\})^*\}$$

(the strings in L_3 are produced by using rules of types (2), (3) and (4), exactly one time a rule of type (4), but with *cac* not introduced on the rightmost position of the string).

$$L_4 = L(G) \cap (V\{c\})^*$$

(the strings in L_4 are produced by using rules of type (1), or of types (2), (3), (4), exactly one time a rule of type (4), with *cac* introduced on the rightmost position of the string)

Clearly, $Im(h) \cap (L_1 \cup L_2 \cup L_3) = \emptyset$, hence $h^{-1}(L(G)) = h^{-1}(L_4)$.

Moreover, $h(L) = L_4$ (from each derivation in $G_a, a \in V$, we can obtain a derivation in G and conversely, and $h(L) \subseteq Im(h) = (V\{c\})^*$), and h is an injective homomorphism, hence $h^{-1}(L_4) = L$, that is $L = h^{-1}(L_4) = h^{-1}(L(G)) = h^{-1}(L') = h^{-1}(L_{in}(G))$.

As one can see, if L is regular, then G is right-linear, and if L is linear, then G is linear too, which completes the proof.

Consider now the intersection. For LIN and CF we cannot obtain characterizations: consider the linear grammars

$$\begin{aligned}
G_1 &= (\{S, A\}, \{a, b, c\}, S, \{S \longrightarrow Sc, S \longrightarrow A, A \longrightarrow aAb, \\
&\quad A \longrightarrow ab\}), \\
G_2 &= (\{S, A\}, \{a, b, c\}, S, \{S \longrightarrow aS, S \longrightarrow A, A \longrightarrow bAc, \\
&\quad A \longrightarrow bc\}).
\end{aligned}$$

We have

$$\begin{aligned} L_{in}(G_1) &= \{a^n b^n c^m \mid n \geq 1, m \geq 0\} \\ L_{in}(G_2) &= \{a^n b^m c^m \mid n \geq 0, m \geq 1\} \end{aligned}$$

hence

$$L_{in}(G_1) \cap L_{in}(G_2) = \{a^n b^n c^n \mid n \geq 1\}$$

a language which is not context-free.

However, we can obtain representations of languages in X, $X \in$ {REG, LIN, CF}, as intersections of languages in FIX; as REG is closed under intersection, for this family we have in fact a characterization.

Theorem 3 *For each $L \in X$, $X \in$ {REG, LIN}, there are $L_1, L_2 \in FIX$, such that $L = L_1 \cap L_2$.*

Proof. We consider here only the linear case; the regular case is a particular one.

Let $G = (V_N, V_T, S, P)$ be a linear grammar. Without loss of generality we may assume that each rule in P is of the next forms: $A \longrightarrow aB, A \longrightarrow Ba, A \longrightarrow a$ (for, each rule $A \longrightarrow a_1 \ldots a_n B b_m \ldots b_1$ can be replaced by $A \longrightarrow a_1 A_1, A_1 \longrightarrow a_2 A_2, \ldots, A_{n-1} \longrightarrow a_n C, C \longrightarrow C_1 b_1, C_1 \longrightarrow C_2 b_2, \ldots, C_{m-1} \longrightarrow B b_m$, etc.).

Consider the new symbols S_1, S_2 and construct the grammars

$$G_i = (V_N \cup \{S_i\}, V_T, S_i, P_i), i = 1, 2,$$

with

$$\begin{aligned} P_1 = \; & \{S_1 \longrightarrow x \mid x \in L(G), |x| \leq 1\} \\ & \cup \{S_1 \longrightarrow xAy \mid S \overset{*}{\Longrightarrow} xAy \text{ in } G, |xy| \leq 2\} \\ & \cup \{A \longrightarrow xBy \mid A \overset{*}{\Longrightarrow} xBy \text{ in } G, |xy| = 2\} \\ & \cup \{A \longrightarrow a \mid A \longrightarrow a \in P, a \in V_T\}, \\ P_2 = \; & \{S_1 \longrightarrow x \mid x \in L(G), |x| \leq 2\} \\ & \cup \{S_1 \longrightarrow xAy \mid S \overset{*}{\Longrightarrow} xAy \text{ in } G, |xy| \leq 2\} \\ & \cup \{A \longrightarrow xBy \mid A \overset{*}{\Longrightarrow} xBy \text{ in } G, |xy| = 2\} \\ & \cup \{A \longrightarrow ab \mid A \overset{*}{\Longrightarrow} ab \text{ in } G, a, b \in V_T\}. \end{aligned}$$

Clearly,

$$\begin{aligned} L(G_1) &= L(G_2) = L \\ L_{in}(G_1) &= L(G_1) \cup \bigcup_{A \neq S_1} L_A(G_1) \\ L_{in}(G_2) &= L(G_2) \cup \bigcup_{A \neq S_2} L_A(G_2) \end{aligned}$$

and

$$\begin{aligned} L_A(G_1) &\subseteq \{x \in V_T^* \mid |x| = 2k+1, k \geq 0\},\ A \neq S_1, \\ L_A(G_2) &\subseteq \{x \in V_T^* \mid |x| = 2k, k \geq 1\},\ A \neq S_2. \end{aligned}$$

Therefore,

$$L_{in}(G_1) \cap L_{in}(G_2) = L(G_1) \cap L(G_2) = L.$$

A similar representation theorem can be obtained also for the context-free case.

Theorem 4 *For each $L \in$ CF, there are $L_1, L_2 \in$ FICF, such that $L = L_1 \cap L_2$.*

Proof. Let $L \subseteq V^*$ be a context-free language and consider

$$\text{even}(L) = L \cap \{ab \mid a,\ b \in V\}^*,$$

$$\text{odd}(L) = L \cap \{ab \mid a,\ b \in V\}^* V.$$

Clearly, $L =$ even$(L) \cup$ odd(L) and even(L), odd(L) are context-free languages (CF is closed under intersection by regular sets).

On the other hand,

$$L = \bigcup_{a \in V} \{a\} d_a(L)$$

and

$$\text{even}(L) = \bigcup_{a \in V} \{a\} \text{ odd}(d_a(L)),$$

$$\text{odd}(L) = \bigcup_{a \in V} \{a\} \text{ even}(d_a(L)).$$

Therefore

$$\begin{aligned} L &= \text{even}(L) \cup \bigcup_{a \in V} \{a\} \text{ even}(d_a(L)) \\ &= \text{odd}(L) \cup \bigcup_{a \in V} \{a\} \text{ odd}(d_a(L)). \end{aligned}$$

All languages even$(d_a(L))$, odd$(d_a(L))$, $a \in V$, are context-free. In view of the super-normal form theorem in [2], [5], there are the grammars

(i) $G_1 = (V_{N,1}, V_T, S_1, P_1)$

$G_{a,1} = (V_{N,a,1}, V_T, S_{a,1}, P_{a,1})$

such that $L(G_1) =$ even(L), $L(G_{a,1}) =$ even$(d_a(L)), a \in V$, and the nonterminal rules in $P_1, P_{a,1}, a \in V$, are in the (2,0,0) normal form (of type $A \longrightarrow xBC$, $x \in V_T^*$, $|x| = 2, A, B, C$ nonterminals), whereas the terminal rules $A \longrightarrow w$ have $|w|$ in the length set of the generated language, that is $|w|$ is even;

(ii) $G_2 = (V_{N,2}, V_T, S_2, P_2)$

$G_{a,2} = (V_{N,a,2}, V_T, S_{a,2}, P_{a,2})$

such that $L(G_2) =$ odd(L), $L(G_{a,2}) =$ odd$(d_a(L))$, $a \in V$, and the nonterminal rules in $P_2, P_{a,2}, a \in V$, are in the (1,0,0) normal form (of type $A \longrightarrow bBC$, $b \in$

V, A, B, C nonterminals), whereas the terminal rules $A \longrightarrow w$ have $|w|$ in the length set of the generated language, that is $|w|$ is odd.

Now, it is easy to see that $L_{in}(G_1), L_{in}(G_{a,1}),\ a \in V$, contain only strings of even lengths, whereas $L_{in}(G_2), L_{in}(G_{a,2}),\ a \in V$, contain only strings of odd lengths (induction on the number of rules used in a derivation).

Assume all vocabularies $V_{N,i}, V_{N,a,i}, i = 1, 2$, pairwise disjoint and construct the grammars

$$G'_i = (V'_{N,i}, V_T, S'_i, P'_i),\ i = 1, 2,$$

with

$$\begin{aligned} V'_{N,i} &= V_{N,i} \cup \bigcup_{a \in V} V_{N,a,i} \cup \{S'_i\}, \\ P'_i &= P_i \cup \bigcup_{a \in V} P_{a,i} \cup \{S'_i \longrightarrow S_i\} \cup \{S'_i \longrightarrow aS_{a,i} | a \in V\}. \end{aligned}$$

From the above relations we have

$$\begin{aligned} L(G'_1) &= L(G'_2) = L, \\ L_{in}(G'_i) &= L_{in}(G_i) \cup \bigcup_{a \in V} \{a\} L_{in}(G_{a,i}),\ i = 1, 2 \end{aligned}$$

and, from the construction of G'_i, we obtain

(a) if $w \in L_{in}(G'_1) - L(G'_1)$, then $|w|$ is even,

(b) if $w \in L_{in}(G'_2) - L(G'_2)$, then $|w|$ is odd.

In conclusion, $L_{in}(G'_1) \cap L_{in}(G'_2) = L(G'_1) \cap L(G'_2) = L$, and the proof is over.

References

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

[2] M. Blattner, S. Ginsburg, Canonical forms of context-free grammars and position restricted grammar forms, *Lect. Notes Computer Sci.,* 56, Springer-Verlag, 1979, 49-53.

[3] J. Dassow, On fully initial context-free languages, *Papers on Automata and Languages,* 10 (1988), 3-6.

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

[5] H. A. Maurer, A. Salomaa, D. Wood, A super normal form theorem for context-free grammars, *Report F63,* Institut für Inform., Techn. Univ. Graz, 1981.

[6] Gh. Paun, A note on fully initial context-free languages, *Papers on Automata and languages,* 10 (1988), 7-11.

[7] A. Salomaa, *Formal languages,* Academic Press, New York, London, 1973.

[8] S. Vicolov, A note on fully initial grammars, *Acta Cybernetica,* 10, 1-2 (1991), 113-118.

(Received November 13, 1990)