Further remarks on fully initial grammars

ALEXANDRU MATEESCU and GHEORGHE PĂUN

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

We investigate those languages generated by (context-free) grammars in which all nonterminals are regarded as axioms (problem raised by S. Horváth, at a formal language workshop, in Budapest, 1987). Among the considered topics, we can list: motivations, necessary conditions, right/left — regular/linear variants (generative capacity and closure properties), and other questions.

1. Motivations

In a usual context-free grammar (in general, in a Chomsky grammar), a nonterminal symbol is distinguished and taken as axiom (all derivations have to start from this nonterminal). This is motivated by mathematical reasons, as well as by the "classical" applications of Chomsky grammars, namely in modelling the syntax of natural or programming languages. However, there are many circumstances where this restriction is not important. This was the reason for which S. Horváth proposed to consider grammars in which *a certain amount* of nonterminals are allowed to be axioms. In [3], [9], grammars in which *all* nonterminals are axioms are considered (they are called *fully initial*).

Besides the naturalness of this idea, many further reasons can be invoked for dealing with several-axiom grammars. Here are some of them. (1) For instance, in W-grammars (two-level grammars) [11], the meta-level is a context-free grammar for which no axiom is distinguished. (2) In pure grammars [7], one considers finite sets of axioms. (3) According to the well-known Ginsburg—Rice—Schutzenberger theorem, each context-free language is a component of the minimal solution of a system of equations on a free monoid [4]; the study of equation systems does not involve special variables ("start" variables). (4) Moreover, in [6] systems of equations in which the iteration process starts from an arbitrary n-tuple of finite sets (not from an n-tuple of empty sets, as usual) are considered; in this way a characterization of EOL languages is obtained. (5) The ADJ group [1] associates a many-sorted initial algebra with a context-free grammar so that the language generated by this grammar is the homomorphic image of a certain carrier of the initial algebra. The construction of this many-sorted initial algebra does not depend on the start symbol of the corresponding context-free grammar. (6) Generalizing the definition of hipernotions in W-grammars, in [2] H-systems are introduced and investigated; in them the start symbol is replaced by an arbitrary (not necessarily finite) language; the language generated by an H-system is then defined by using homomorphisms, not production rules.

As one can see, there are enough reasons for further investigation of grammars in which more than one (or all) nonterminals are axioms. Moreover, as an *a posteriori* reason, the problems raised and the results obtained about these grammars prove that the subject is worth considering, leading to interesting new insights about Chomsky grammars.

2. Definitions and notations

For a vocabulary V, we denote by V^* the free monoid generated by V under the operation of concatenation, and λ is the null element. The length of a string $x \in V^*$ is denoted by $|x|$. Inclusion and strict inclusion are denoted by $\subseteq$ and $\subset$, respectively.

A Chomsky grammar is a quadruple $G=(V_N, V_T, S, P)$; V_N is the nonterminal vocabulary, V_T is the terminal one, $S \in V_N$ is the axiom and P is the production set. The usual language generated by G is defined by

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

The fully initial language generated by G is

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

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

When dealing with the fully initial language only, we shall write a grammar in the form $G=(V_N, V_T, P)$, thus omitting the useless axiom.

Usually, a language is said to be of type 3 if it can be generated by a right-linear or a left-linear grammar, in the classical case. (Right-linear and left-linear grammars have the same generative power.) For fully initial grammars this is not true, therefore we shall distinguish several classes of "type-3" grammars.

A grammar $G=(V_N, V_T, P)$ 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 $\mathscr{FL}_{\text{rlin}}$, $\mathscr{FL}_{\text{llin}}$ the corresponding families of fully initial languages. Moreover, we distinguish between grammars with rules of the form $A \to xB$ with an arbitrary string $x \in V_T^*$ as above and grammars in which x must be a terminal. A grammar $G=(V_N, V_T, P)$ 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 $\mathscr{FL}_{\text{rreg}}$, $\mathscr{FL}_{\text{lreg}}$.

The above family $\mathscr{FL}_3$ is, in fact, $\mathscr{FL}_{\text{rlin}} \cup \mathscr{FL}_{\text{llin}}$. We shall also denote this family by $\mathscr{FL}_{\text{lin}}^{\cup}$ and we shall consider the following families too:

$$\mathscr{FL}_{\text{lin}}^{\cap} = \mathscr{FL}_{\text{rlin}} \cap \mathscr{FL}_{\text{llin}},$$

$$\mathscr{FL}_{\text{reg}}^{\cup} = \mathscr{FL}_{\text{rreg}} \cup \mathscr{FL}_{\text{lreg}},$$

$$\mathscr{FL}_{\text{reg}}^{\cap} = \mathscr{FL}_{\text{rreg}} \cap \mathscr{FL}_{\text{lreg}}.$$

As in many cases, we shall consider two languages identical if they differ by at most the empty string λ.

The sets of prefixes, suffixes and subwords of a given string x are denoted by $\text{Init}(x)$, $\text{Fin}(x)$, $\text{Sub}(x)$, respectively, and these notations will be extended in the natural way to languages. When considering only proper prefixes, suffixes and subwords, we shall write $\text{Initp}(x)$, $\text{Finp}(x)$ and $\text{Subp}(x)$, respectively.

For further details in formal language theory, the reader is referred to [10].

3. Necessary conditions for the context-free case

We shall consider here some necessary conditions for a language to be in $\mathscr{FL}_2$; some of these conditions will be also particularized to $\mathscr{FL}_3$ or to subfamilies of $\mathscr{FL}_3$.

Lemma 1. For each language $L \in \mathscr{FL}_2$ there is a λ-free grammar $G=(V_N, V_T, P)$ such that P does not contain chain rules (rules of the form $A \to B$, $A, B \in V_N$) and $L = L_{\text{in}}(G)$.

Proof. The same as for usual context-free languages.

Lemma 2. For each language $L \in \mathscr{FL}_2$, $L \subseteq V^*$, 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) $|vwx| \leqq q$, $|vx| > 0$,

(ii) for all $k \geqq 0$, $uv^k wx^k y \in L$ and $v^k wx^k \in L$.

Proof. The same as for usual context-free languages, with the following two remarks:

— we start from a reduced grammar, G, in the sense of Lemma 1 (see Lemma 3.1.1 in [4]), not from a Chomsky normal form grammar (as in Theorem 6.4 in [10]);

— given a derivation tree T, all subtrees having the roots in the nonterminals of T correspond to substrings of the string associated to T and which belong to the fully initial language generated by the grammar; therefore, when we have a derivation $S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} uvwxy$, then both $uv^k wx^k y$ and $v^k wx^k$ belong to $L_{\text{in}}(G)$.

Corollary 1. If $L \in \mathscr{FL}_2$, then there is a constant p such that for all $z \in L$, $|z| > p$, we have $\text{Subp}(z) \cap L \neq \emptyset$.

Proof. Take p as in Lemma 2 and, for $z \in L$, $|z| > p$, write $z = uvwxy$ with the above properties. As $v^k wx^k \in L$ for $k \geqq 0$, when $k = 0$, we obtain $w \in L \cap \text{Sub}(z)$. Moreover, $|vx| > 0$, hence we have, in fact, $w \in L \cap \text{Subp}(z)$.

Corollary 2. If $L \in \mathcal{FL}_2$ is an infinite language, then also $L \cap \mathrm{Subp}(L)$ is infinite.

Proof. Let p, q be the constants of Lemma 2 and take $z \in L$, $|z| > p$, $z = uvwxy$. Each string $v^k w x^k$, $k \geq 0$, is in L. Clearly, $v^k w x^k \in \mathrm{Subp}(L)$ and $v^k w x^k \neq v^{k+1} w x^{k+1}$, $k \geq 0$ (we have $|vx| > 0$), therefore $L \cap \mathrm{Subp}(L)$ contains the infinite set $\{v^k w x^k | k \geq 0\}$.

Lemma 3. The conditions (properties) in the above two corollaries are independent from one another.

Proof. We consider the languages

$$L_1 = \{a\} \cup \{ab^n a | n \geq 1\}$$

and

$$L_2 = \{ba^n b, cba^n bc | n \geq 1\}.$$

The first language fulfils the condition in Corollary 1 (take $p=1$; $\mathrm{Subp}(ab^n b) \cap \cap L_1 = \{a\}$ for all $n \geq 1$), but not that in Corollary 2 ($\mathrm{Subp}(L_1) \cap L_1 = \{a\}$). The second language fulfils the condition in Corollary 2 ($L_2 \cap \mathrm{Subp}(L_2) = \{ba^n b | n \geq 1\}$), but not that in Corollary 1 (the strings $ba^n b$, irrespective of their length, have no proper subwords in L_2).

This lemma shows that none of the conditions in Corollaries 1 and 2 is sufficient for a language to be in $\mathcal{FL}_2$; even they together are insufficient for that, as it follows from the next result.

Lemma 4. The condition in Lemma 2 is strictly stronger than the conditions in Corollaries 1 and 2 together.

Proof. We consider the language

$$L = \{b\} \cup \{ba^n b, cba^n bc | n \geq 1\}.$$

It is easy to see that both conditions in Corollaries 1 and 2 are fulfilled (similarly to languages L_1, L_2 in the above proof), but that in Lemma 2 is not. Indeed, let p and q be two positive integers and take $z = ba^n b$, $|z| > p$ (there are arbitrarily long such strings in L). We must have $z = uvwxy$ such that $v^k w x^k \in L$, $k \geq 0$, $|vx| > 0$. It follows that $vx \in \{a^n | n \geq 1\}$, hence $v^k w x^k$ is of the form $a\alpha$ or of the form αa, $\alpha \in \{a, b\}^*$. Such strings cannot be in L, a contradiction.

Lemma 5. The condition in Lemma 2 is not sufficient for a language to be in $\mathcal{FL}_2$.

Proof. Let us consider the language

$$L = \{a^n | n \geq 0\} \cup \{b^n | n \geq 0\} \cup \{a^n b^{2^m} | n, m \geq 1\}.$$

The language L is not context-free; as $\mathcal{FL}_2 \subset \mathcal{L}_2$ [3], it follows that $L \notin \mathcal{FL}_2$. However, this language fulfils the condition in Lemma 2. Take, for instance, $p=1$, $q=1$. For $z=a^n$ or $z=b^n$, we clearly have all conditions in lemma fulfilled. If $z=a^n b^{2^m}$ we take $u=\lambda$, $v=a$, $w=\lambda$, $x=\lambda$, $y=a^{n-1}b^{2^m}$. Obviously, $z=uvwxy$, $|vwx| \leq q=1$, $|vx|>0$, $uv^k w x^k y = a^k a^{n-1} b^{2^m} \in L$ for all $k \geq 0$ (for $k=0$, $n=1$ we can obtain $uv^k w x^k y = b^{2^m}$, which is in L too), and $v^k w x^k = a^k \in L$ for all $k \geq 0$.

Conjecture 1. If L is a context-free language which fulfils the condition in Lemma 2, then $L \in \mathcal{F}\mathcal{L}_2$.

We consider now a necessary condition of different type, similar to the one used in the theory of Marcus contextual languages [8].

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

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

and define

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

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

We say that L has *property* R iff all the sets $R_i(L)$, $i \geqq 1$, are finite.

Lemma 6. If $L \in \mathcal{F}\mathcal{L}_2$, then L has property R.

Proof. Let $L \in \mathcal{F}\mathcal{L}_2$, $L \subseteq V^*$, be a language and take a grammar $G=(V_N, V_T, P)$ such that $L_{\mathrm{in}}(G)=L$ and G does not contain λ-rules and chain rules (Lemma 1). For a string $x \in L$, let $T(x, G)$ be the set of all derivation trees describing derivations of x in G starting from a nonterminal in V_N (which is the root of a tree). Denote by hei(T) the *height* of a given tree $T \in T(x, G)$, i.e. the maximum of lengths of paths linking the root of T to its leafs (symbols in x). For a given string x we define

$$\mathrm{hei}_G(x) = \max\{\mathrm{hei}(T) | T \in T(x, G)\}.$$

Then we have

$$R_i(L) \subseteq \{x \in L | \mathrm{hei}_G(x) \leqq i\}, \quad i \geqq 1.$$

Indeed, let $x \in \mathrm{Min}(L)$ be a string and take a derivation D: $A \overset{*}{\Rightarrow} x$ in G corresponding to a tree T. If $\mathrm{hei}(T) \geqq 2$, then the derivation D is of the form D: $A \Rightarrow \alpha_1\alpha_2 \dots \alpha_k \overset{*}{\Rightarrow} \beta_1\beta_2 \dots \beta_k = x$, $\alpha_i \in V_N \cup V_T$, $\alpha_i \overset{*}{\Rightarrow} \beta_i$, $1 \leqq i \leqq k$, $k \geqq 2$, and for some i, $1 \leqq i \leqq 2$, $\alpha_i \in V_N$. This implies $\beta_i \in L \cap \mathrm{Subp}(z)$, hence $z \notin \mathrm{Min}(L)$, a contradiction. In conclusion, $\mathrm{hei}(T)=1$, $\mathrm{hei}_G(x)=1$, and the inclusion $R_i(L) \subseteq \{x \in L | \mathrm{hei}_G(x) \leqq i\}$ holds for $i=1$.

Let us assume, this relation is true for $j=1, 2, \dots, i$, $i \geqq 1$, and consider $x \in R_{i+1}(L)$. If $x \in R_i(L)$, then $\mathrm{hei}_G(x) \leqq i$ by the induction hypothesis. Assume that $x \in R_{i+1}(L) - R_i(L)$, that is $x \in \mathrm{Min}(L - R_i(L))$. In other terms, $\mathrm{Subp}(x) \cap \cap (L - R_i(L)) = \emptyset$. Suppose that $\mathrm{hei}_G(x) > i+1$, and take a derivation tree $T \in T(x, G)$ such that $\mathrm{hei}(T) > i+1$. There is a derivation D, associated with this tree, having the form D: $A \Rightarrow \alpha_1\alpha_2 \dots \alpha_k \overset{*}{\Rightarrow} \beta_1\beta_2 \dots \beta_k = x$, such that $\alpha_j \in V_N \cup V_T$, $\alpha_j \overset{*}{\Rightarrow} \beta_j$, $1 \leqq j \leqq k$ ($\alpha_j = \beta_j$ if $\alpha_j \in V_T$), $k \geqq 2$, and there is an $\alpha_j \in V_N$ for some j, $1 \leqq j \leqq k$. All strings β_j, $1 \leqq j \leqq k$, belong to $\mathrm{Subp}(x) \cap L$. As $\mathrm{Subp}(x) \cap (L - R_i(L)) = \emptyset$, we must have $\beta_j \in R_i(L)$. By the induction hypothesis we get $\mathrm{hei}_G(\beta_j) \leqq i$, $1 \leqq j \leqq k$. This implies that the tree T consists of a "root level" describing the rule $A \to \alpha_1\alpha_2 \dots \alpha_j$ and of all trees associated with subderivations $\alpha_j \overset{*}{\Rightarrow} \beta_j$, for $\alpha_j \in V_N$. In conclusion, $\mathrm{hei}(T) \leqq \leqq i+1$, a contradiction. We obtain $\mathrm{hei}_G(x) \leqq i+1$, which completes the induction argument.

The sets $\{x \in L | \mathrm{hei}_G(x) \leqq i\}$, $i \geqq 1$, are clearly finite, therefore the sets $R_i(L)$, $i \geqq 1$, are finite too, and the proof is completed.

Lemma 7. The property R implies conditions in Corollaries 1, 2, but there are languages fulfilling both these conditions without having the property R.

Proof. Consider again the language L in the proof of Lemma 4 (it satisfies the conditions in Corollaries 1 and 2). We obtain

$$R_1(L) = \{b\},$$
$$R_2(L) = \{b\} \cup \{ba^n b \mid n \geqq 1\},$$

hence $R_2(L)$ is infinite, L does not have the property R.

Define now, for a given language L,

$$p = \max\{|x| \mid x \in R_1(L)\}$$

If $z \in L$, $|z| > p$, then $z \notin R_1(L)$, hence $\mathrm{Subp}(z) \cap L \neq \emptyset$. The property R implies thus the condition in Corollary 1.

Consider an infinite language L having the property R but not having the property in Corollary 2, that is $L \cap \mathrm{Subp}(L)$ is finite, $\mathrm{card}(L \cap \mathrm{Subp}(L)) = t$. As L is infinite, but all sets $R_i(L)$, $i \geqq 1$, are finite, it follows that $R_i(L) \subset R_{i+1}(L)$, $i \geqq 1$ (if $R_j(L) = R_{j+1}(L)$, then $R_j(L) = R_{j+k}(L)$, $k \geqq 1$, hence $L \subseteq R_j(L)$, a contradiction). As $R_{i+1}(L) - R_i(L) = \mathrm{Min}(L - R_i(L)) \neq \emptyset$, it follows that $R_{i+1}(L) \cap (L \cap \cap \mathrm{Subp}(L)) \neq \emptyset$ and $R_i(L) \cap (L \cap \mathrm{Subp}(L)) \subset R_{i+1}(L) \cap (L \cap \mathrm{Subp}(L))$ for all $j \geqq 1$. This implies $\mathrm{card}(R_{t+1}(L) \cap L \cap \mathrm{Subp}(L)) \geqq t+1$, therefore $\mathrm{card}(L \cap \cap \mathrm{Subp}(L)) \geqq t+1$, a contradiction.

Lemma 8. The condition R is not sufficient for a (context-free) language to be in $\mathscr{F}\mathscr{L}_2$.

Proof. We consider the language

$$L = \{a^n \mid n \geqq 1\} \cup \{ab^n a^n \mid n \geqq 1\}.$$

This is a context-free language and we have

$$R_1(L) = \{a\},$$
$$R_i(L) = \{a^j \mid 1 \leqq j \leqq i\} \cup \{ab^j a^j \mid 1 \leqq j \leqq i-1\}, \quad i \geqq 2,$$

therefore the property R is observed.

However, this language is not in $\mathscr{F}\mathscr{L}_2$. Assume the contrary, and factorize a long enough $z = ab^n a^n$ in L into $z = uvwxy$ as in Lemma 2. Then we must have $v = b^i$, $x = a^i$, $i > 0$, which implies that all $v^k w x^k = b^{ik} w a^{ik}$, $k \geqq 0$, are in L, a contradiction to the form of strings in L.

Remark 1. The above proof shows that if Conjecture 1 were proved then, for context-free languages, the condition in Lemma 2 would be stronger than property R.

Conjecture 2. For arbitrary languages, the condition in Lemma 2 is stronger than property R.

Remark 2. If in condition (ii) of Lemma 2 we take $k \geqq 1$ instead of $k \geqq 0$ (sometimes, the pumping lemma is formulated in this weaker form; see [4], for instance), then the modified condition will be independent of condition R. The language

L in the above proof supports one of the implications; the other one can be proved using the language

$$L = \{ba^n ba^n b^m a | n,\ m \geqq 1\} \cup \{a^n ba^n | n \geqq 1\}.$$

Taking $p=1,\ q=3$ we obtain the modified property in Lemma 2, but we have

$$R_1(L) = \{aba\},$$

$$R_2(L) = \{aba,\ a^2ba^2\} \cup \{babab^m a | m \geqq 1\},$$

hence property (condition) R is not satisfied.

Lemma 2 has some particular forms for right/left linear grammars.

Lemma 9. (i) If $L \in \mathcal{FL}_{rlin}$, then there are two positive integers p, q such that, for all $z \in L$, $|z| > p$, we can write $z = uvw$, $0 < |v| \leqq q$ and $uv^i w \in L$, $v^i w \in L$, for all $i \geqq 0$.

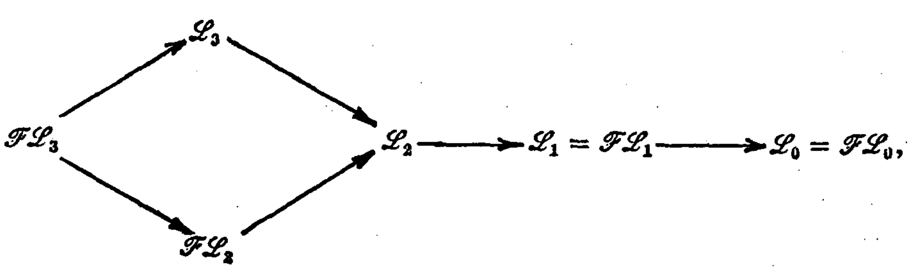
(ii) If $L \in \mathcal{FL}_{llin}$, then there are two positive integers p, q such that, for all $z \in L$, $|z| > p$, we can write $z = uvw$, $0 < |v| \leqq q$ and $uv^i w \in L$, $uv^i \in L$, for all $i \geqq 0$.

Proof. Obvious particularizations of the proof of Lemma 2 to right/left linear grammars.

4. Fully initial languages in the Chomsky hierarchy

As we have mentioned, in [3] it is proved that $\mathcal{FL}_2 \subset \mathcal{L}_2$. A more precise (and more general) result is true, namely we have.

Theorem 1. The following diagram holds:



where $\rightarrow$ indicates a strict inclusion; the families $\mathcal{L}_3$, $\mathcal{FL}_2$ are incomparable.

Proof. As $\{ba^n b | n \geqq 1\}$ is not in $\mathcal{FL}_2$ (it fulfils no necessary condition in the previous section), it follows that $\mathcal{L}_3 - \mathcal{FL}_2 \neq \emptyset$, hence also $\mathcal{L}_2 - \mathcal{FL}_2 \neq \emptyset$, $\mathcal{L}_3 - \mathcal{FL}_3 \neq \emptyset$. On the other hand, $\{a^n b^n | n \geqq 1\}$ is in $\mathcal{FL}_2 - \mathcal{L}_3$, hence $\mathcal{FL}_2 - \mathcal{FL}_3 \neq \emptyset$ and $\mathcal{L}_3, \mathcal{FL}_2$ are incomparable.

Consider now a grammar G of arbitrary type, $G = (V_N, V_T, P)$ and construct the grammar $G' = (V_N \cup \{S'\}, V_T, S', P \cup \{S' \rightarrow A | A \in V_N\})$. Clearly, G' is of the same type as G and $L(G') = L_{in}(G)$, hence $\mathcal{FL}_i \subseteq \mathcal{L}_i$, $i = 0, 1, 2, 3$.

In order to complete the proof, we have to prove that $\mathscr{L}_i \subseteq \mathscr{FL}_i$, $i=0, 1$. Take a language $L \in \mathscr{L}_i$, $L \subseteq V^*$. We can write

$$L = \bigcup_{a \in V} \{a\}\partial_a L \cup \{x \in L \mid |x| \leqq 2\}$$

($\partial_a L$ is the left derivative of L with respect to a). As $\mathscr{L}_i$, $i=0, 1$, are closed under left derivative, $\partial_a L \in \mathscr{L}_i$. Let $G_a=(V_{N,a}, V, S_a, P_a)$ be a type-i grammar for $\partial_a L$. Assume the $V_{N,a}$ are pairwise disjoint and define $G=(V_N, V, S, P)$ with

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

$$P = \{S \to x | x \in L,\ |x| \leqq 2\} \cup \{S \to X_a S_a | a \in V\} \cup$$

$$\cup \{\alpha u' \to \alpha v' | \alpha \in V_N,\ u \to v \in P_a,\ a \in V\} \cup$$

$$\cup \{X_a b' \to ab | a, b \in V\} \cup \{ab' \to ab | a, b \in V\}$$

where u' is the string obtained from u by replacing each $a \in V$ by $a' \in V_N$. It is easy to see that no derivation $A \overset{*}{\Rightarrow} w$, $A \in V_N$, is possible in G unless $A=S$, therefore $L(G)=L_{\text{in}}(G)$. Moreover, $L(G)=L$. In conclusion, $L \in \mathscr{FL}_i$, $i=0, 1$, and the proof is ended.

This theorem shows that families $\mathscr{FL}_0$ and $\mathscr{FL}_1$ request no further investigations.

5. Type-3 fully initial languages

First, let us consider some characterizations and representations of languages in $\mathscr{FL}_{\text{rreg}}$, $\mathscr{FL}_{\text{lreg}}$, $\mathscr{FL}_{\text{rlin}}$, $\mathscr{FL}_{\text{llin}}$.

Lemma 10. (i) $L \in \mathscr{FL}_{\text{rreg}}$ if and only if $L \in \mathscr{L}_3$ and $L=\text{Fin}(L)$. (ii) $L \in \mathscr{FL}_{\text{lreg}}$ if and only if $L \in \mathscr{L}_3$ and $L=\text{Init}(L)$. (iii) $L \in \mathscr{FL}_{\text{reg}}^{\cap}$ if and only if $L \in \mathscr{L}_3$ and $L=\text{Sub}(L)$.

Proof. (i) Let $L \in \mathscr{FL}_{\text{rreg}}$ be a language such that $L=L_{\text{in}}(G)$, $G=(V_N, V_T, P)$. Clearly, $L \in \mathscr{L}_3$ and $L \subseteq \text{Fin}(L)$. Take a string $w \in \text{Fin}(L)$. There is a $u \in V_T^*$ such that $uw \in L$. Therefore, there is a derivation $A \overset{*}{\Rightarrow} uw$ in G. As G is a right-regular grammar, there is a $B \in V_N$ such that $A \overset{*}{\Rightarrow} uB \overset{*}{\Rightarrow} uw$, which implies $w \in L_{\text{in}}(G)=L$. In conclusion, $w \in L$, $\text{Fin}(L) \subseteq L$.

Conversely, let $L \in \mathscr{L}_3$, $L=\text{Fin}(L)$, and consider a reduced right-regular grammar G, $G=(V_N, V_T, S, P)$, without useless nonterminals, $L=L(G)$, $P \subseteq V_N \times \times (V_T \cup V_T V_N)$. Clearly, $L(G) \subseteq L_{\text{in}}(G)$. Take a string $w \in L_{\text{in}}(G)$. There is a derivation $A \overset{*}{\Rightarrow} w$ in G, $A \in V_N$. As G is reduced, there is a derivation $S \overset{*}{\Rightarrow} uA$, $u \in V_T^*$, therefore $S \overset{*}{\Rightarrow} uA \overset{*}{\Rightarrow} uw$ is possible in G. This implies $w \in \text{Fin}(L(G))=L$, that is $w \in L$, hence $L_{\text{in}}(G) \subseteq L(G)$. In conclusion, $L_{\text{in}}(G)=L$, $L \in \mathscr{FL}_{\text{rreg}}$ and (i) is proved.

(ii) Analogously.

(iii) Follows from the definition of $\mathscr{FL}_{\text{reg}}^{\cap}$, the above parts (i) and (ii) and the relations $\text{Sub}(L)=\text{Fin}(\text{Init}(L))=\text{Init}(\text{Fin}(L))=\text{Init}(\text{Sub}(L))=\text{Fin}(\text{Sub}(L))$.

Denote by $\text{Mi}(w)$ the mirror image of a string w and extend this operation to languages.

Lemma 11. (i) $L\in\mathcal{F}\mathcal{L}_{\rm rreg}$ if and only if $\mathrm{Mi}(L)\in\mathcal{F}\mathcal{L}_{\rm lreg}$. (ii) $L\in\mathcal{F}\mathcal{L}_{\rm rlin}$ if and only if $\mathrm{Mi}(L)\in\mathcal{F}\mathcal{L}_{\rm llin}$.

Proof. (i) Take a language $L\in\mathcal{F}\mathcal{L}_{\rm rreg}$, generated by $G=(V_N, V_T, P)$ and define $G'=(V_N, V_T, \{A\to\mathrm{Mi}(x)|A\to x\in P\})$. Clearly, $L_{\rm in}(G')=\mathrm{Mi}(L(G))=\mathrm{Mi}(L)$, hence $\mathrm{Mi}(L)\in\mathcal{F}\mathcal{L}_{\rm lreg}$. The converse implication is analogous.

(ii) Similar.

Lemma 12. (i) Each language in $\mathcal{F}\mathcal{L}_{\rm rlin}$ is a homomorphic image of a language in $\mathcal{F}\mathcal{L}_{\rm rreg}$. (i) Each language in $\mathcal{F}\mathcal{L}_{\rm llin}$ is a homomorphic image of a language in $\mathcal{F}\mathcal{L}_{\rm lreg}$.

Proof. (i) Let $L\subseteq V^*$, $L\in\mathcal{F}\mathcal{F}_{\rm rlin}$, be a language generated by the grammar $G=(V_N, V, P)$. We define the grammar $G'=(V_N, V', P')$ by

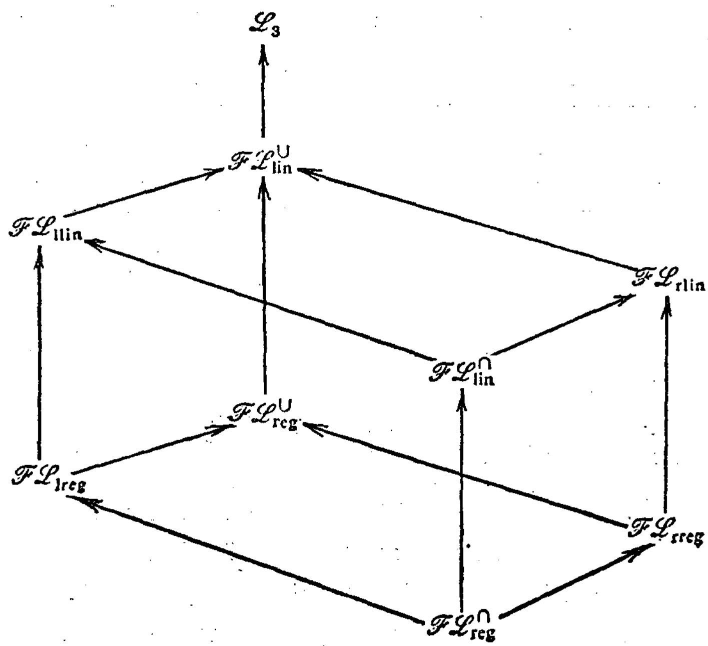
$$V' = \{[\alpha]|X\to\alpha Y \text{ or } X\to\alpha \text{ is in } P,\ \alpha\in V^*,\ X, Y\in V_N\},$$

$$P' = \{X\to[\alpha]Y|X\to\alpha Y\in P\}\cup\{X\to[\alpha]|X\to\alpha\in P\}.$$

Consider also the homomorphism $h\colon V'^*\to V^*$ defined by $h([\alpha])=\alpha$, $[\alpha]\in V'$. Clearly, G' is a right-regular grammar and $h(L_{\rm in}(G'))=L$.

(ii) Analogously.

Theorem 2. The inclusion relations between the above discussed families of type-3 fully initial languages are those in the next diagram (→ indicates a strict inclusion; the unlinked families are incomparable).



Proof. All inclusions are obvious. Moreover, we have: $ba^* \in \mathcal{FL}_{\mathrm{lreg}} - \mathcal{FL}_{\mathrm{rreg}}$, $a^*b \in \mathcal{FL}_{\mathrm{rreg}} - \mathcal{FL}_{\mathrm{lreg}}$ (use Lemma 10, parts (i), (ii)). This settles the relations on the bottom face of the "cube" in the diagram. Moreover, $c(ab)^* \in \mathcal{FL}_{\mathrm{llin}} - (\mathcal{FL}_{\mathrm{rlin}} \cup \cup \mathcal{FL}_{\mathrm{lreg}})$ and $(ab)^*c \in \mathcal{FL}_{\mathrm{rlin}} - (\mathcal{FL}_{\mathrm{llin}} \cup \mathcal{FL}_{\mathrm{rreg}})$. This settles the relations on the upper face of the "cube", as well as those indicated by the vertical edges, except $\mathcal{FL}^{\cap}_{\mathrm{reg}} \subset \mathcal{FL}^{\cap}_{\mathrm{lin}}$. This, however, follows from $(ab)^* \in \mathcal{FL}^{\cap}_{\mathrm{lin}} - \mathcal{FL}^{\cap}_{\mathrm{reg}}$ (use condition (iii) in Lemma 10). The inclusion $\mathcal{FL}^{\cup}_{\mathrm{lin}} = \mathcal{FL}_3 \subset \mathcal{L}_3$ was shown in Theorem 1.

Theorem 3. The closure properties of the above discussed families of type-3 fully initial languages are as presented in Table 1 (Y indicates a positive closure result, N points out a negative closure result).

Table 1.

	$\mathcal{FL}^{\cup}_{\mathrm{lin}}$	$\mathcal{FL}_{\mathrm{llin}}$	$\mathcal{FL}_{\mathrm{rlin}}$	$\mathcal{FL}^{\cap}_{\mathrm{lin}}$	$\mathcal{FL}^{\cup}_{\mathrm{reg}}$	$\mathcal{FL}_{\mathrm{lreg}}$	$\mathcal{FL}_{\mathrm{rreg}}$	$\mathcal{FL}^{\cap}_{\mathrm{reg}}$
Union	N	Y	Y	Y	N	Y	Y	Y
Complementation	N	N	N	N	N	N	N	N
Intersection	N	N	N	N	N	Y	Y	Y
Concatenation	N	N	N	N	N	N	N	N
Kleene closure	Y	Y	Y	Y	Y	Y	Y	Y
Homomorphism	Y	Y	Y	Y	N	N	N	N
Inverse homomorphism	N	N	N	N	Y	Y	Y	Y
Mirror image	Y	N	N	Y	Y	N	N	Y
Right quotient	N	N	Y	N	N	N	Y	N
Left quotient	N	Y	N	N	N	Y	N	N
Init, Fin, Sub	Y	Y	Y	Y	Y	Y	Y	Y
gsm mapping	N	N	N	N	N	N	N	N
Inverse gsm mapping	N	N	N	N	N	N	N	N
Intersection with regular sets	N	N	N	N	N	N	N	N

Proof. Union. If L_1, L_2 are in $\mathcal{FL}_{\mathrm{rreg}}$, $\mathcal{FL}_{\mathrm{lreg}}$ or $\mathcal{FL}^{\cap}_{\mathrm{reg}}$ then $L_1 \cup L_2$ belongs to the same families, as it easily follows from Lemma 10 ($L_1 \cup L_2 \in \mathcal{L}_3$ and $L_1 \cup L_2 = \mathrm{Fin}(L_1 \cup L_2)$, $L_1 \cup L_2 = \mathrm{Init}(L_1 \cup L_2)$, $L_1 \cup L_2 = \mathrm{Sub}(L_1 \cup L_2)$, respectively). $\mathcal{FL}^{\cup}_{\mathrm{reg}}$ is not closed under union, because, for instance, $L_1 = a^*b$, $L_2 = ba^*$ are in $\mathcal{FL}^{\cup}_{\mathrm{reg}}$, but $L_1 \cup L_2$ is not in $\mathcal{FL}^{\cup}_{\mathrm{lin}}$ ($L_1 \cup L_2$ is neither in $\mathcal{FL}_{\mathrm{rlin}}$ nor in $\mathcal{FL}_{\mathrm{llin}}$: use Lemma 9). The closure of $\mathcal{FL}_{\mathrm{rlin}}$, $\mathcal{FL}_{\mathrm{llin}}$, $\mathcal{FL}^{\cap}_{\mathrm{lin}}$ can be proved by direct, standard constructions.

Complementation. The language $L = a^*b^*$ is in $\mathcal{FL}^{\cap}_{\mathrm{reg}}$, but $\{a, b\}^* - L$ is not in $\mathcal{FL}^{\cup}_{\mathrm{lin}}$ (use Lemma 9).

Intersection. The closure of $\mathcal{FL}_{\mathrm{rreg}}$, $\mathcal{FL}_{\mathrm{lreg}}$, $\mathcal{FL}^{\cap}_{\mathrm{reg}}$ can again be proved using Lemma 10 ($\mathrm{Fin}(L_1 \cap L_2) \subseteq \mathrm{Fin}(L_1) \cap \mathrm{Fin}(L_2) = L_1 \cap L_2$ hence $L_1 \cap L_2 = \mathrm{Fin}(L_1 \cap L_2)$, $L_1 \cap L_2 \in \mathcal{L}_3$ etc.). For $\mathcal{FL}^{\cup}_{\mathrm{reg}}$ take $L_1 = a^*b^+$, $L_2 = a^+b^*$, both in this family; $L_1 \cap L_2 = a^+b^+$ does not belong to $\mathcal{FL}^{\cup}_{\mathrm{reg}}$. For the other families take

$$L_1 = c(aab)^*c \cup (aab)^*c \cup c(aab)^* \cup (aab)^*,$$

$$L_2 = ca(aba)^*abc \cup (aba)^*abc \cup ca(aba)^* \cup (aba)^*.$$

They belong to $\mathcal{FL}^{\cap}_{\mathrm{lin}}$, but $L_1 \cap L_2 = ca(aba)^*abc$ is not in $\mathcal{FL}^{\cup}_{\mathrm{lin}}$.

Concatenation. The languages $L_1=a^+$, $L_2=b^+$ are in $\mathcal{FL}^{\cap}_{\text{reg}}$, but $L_1L_2=a^+b^+$ is not in $\mathcal{FL}^{\cup}_{\text{lin}}$, which settles all cases.

Kleene closure. Given a right-regular or a right-linear grammar $G=(V_N, V_T, P)$, construct the grammar $G'=(V_N, V, P')$ with $P'=P\cup\{X\to\alpha Y|X\to\alpha\in P,\ \alpha\in V_T^*,\ X, Y\in V_N\}$. It is easy to see that $L_{\text{in}}(G')=L(G)^+$. The left-regular and left-linear cases can be treated similarly.

Homomorphism. A standard construction proves the positive results. For regular families take $L=a^+$ (it belongs to $\mathcal{FL}^{\cap}_{\text{reg}}$) and the homomorphism h: $a^*\to\{a, b\}^*$ defined by $h(a)=ab$. The language $h(L)=(ab)^+$ is not $\mathcal{FL}^{\cup}_{\text{reg}}$, which implies the nonclosure cases in Table 1.

Inverse homomorphism. Let h: $V^*\to V'^*$ be a homomorphism and $L\subseteq V^*$ a language in $\mathcal{FL}^{\cup}_{\text{reg}}$. According to Lemma 10, $L\in\mathcal{L}_3$ and $L=\text{Sub}(L)$. Clearly, $h^{-1}(L)\in\mathcal{L}_3$ and $h^{-1}(L)\subseteq\text{Sub}(h^{-1}(L))$. Consider now a string u in $\text{Sub}(h^{-1}(L))$. There are v, $w\in V'^*$ such that $vuw\in h^{-1}(L)$, hence $h(v)h(u)h(w)\in L$. This implies $h(u)\in\text{Sub}(L)=L$, hence $h(u)\in L$, that is $u\in h^{-1}(L)$. In conclusion, $\text{Sub}(h^{-1}(L))\subseteq h^{-1}(L)$, which shows that $\text{Sub}(h^{-1}(L)=h^{-1}(L)$, hence $h^{-1}(L)\in\mathcal{FL}^{\cap}_{\text{reg}}$ (Lemma 10, part (iii)). Similar arguments hold for $\mathcal{FL}_{\text{rreg}}$, $\mathcal{FL}_{\text{lreg}}$, $\mathcal{FL}^{\cup}_{\text{reg}}$.

Consider now the language $L=(ab)^*c\cup c(ab)^*\cup(ab)^*$. It belongs to $\mathcal{FL}^{\cap}_{\text{lin}}$, but $h^{-1}(L)=ab^*c$, for h defined by $h(a)=a$, $h(b)=ba$, $h(c)=bc$; this language is not in $\mathcal{FL}^{\cup}_{\text{lin}}$, which implies nonclosure under inverse homomorphism for $\mathcal{FL}_{\text{rlin}}$, $\mathcal{FL}_{\text{llin}}$, $\mathcal{FL}^{\cup}_{\text{lin}}$, $\mathcal{FL}^{\cap}_{\text{lin}}$.

Mirror image. The closure cases follow from Lemma 11, the nonclosure ones are settled by examples of the form: $a^+b\in\mathcal{FL}_{\text{rreg}}$, $\text{Mi}(a^+b)=ba^+\notin\mathcal{FL}_{\text{llin}}$.

Right quotient. We have $L=\{abc, ab, bc, a, b, c\}\in\mathcal{FL}^{\cap}_{\text{reg}}$, but $L/\{c\}=\{ab, b\}\notin\mathcal{FL}_{\text{lreg}}$, hence these families are not closed under right quotient. Similarly, $L=\{abc, ab, a\}\in\mathcal{FL}^{\cup}_{\text{reg}}$, but $L/\{c\}=\{ab\}\notin\mathcal{FL}^{\cup}_{\text{reg}}$. Similar languages can be constructed for $\mathcal{FL}^{\cap}_{\text{lin}}$, $\mathcal{FL}_{\text{llin}}$, $\mathcal{FL}^{\cup}_{\text{lin}}$ (take $L=a^+bc\cup a^+b\cup bc\cup a^+\cup\{b, c\}\in\mathcal{FL}^{\cap}_{\text{lin}}$, respectively, $L=ba^+bc\cup ba^+\in\mathcal{FL}^{\cup}_{\text{lin}}$).

Consider now $L\in\mathcal{FL}_{\text{rreg}}$ and an arbitrary language L'. According to Lemma 10, we have $L=\text{Fin}(L)$. As L/L' is a regular language, we have only to prove that $\text{Fin}(L/L')=L/L'$. Let $u\in\text{Fin}(L/L')$ be an arbitrary string. There is a v such that $vu\in L/L'$, hence there is a $w\in L'$ such that $vuw\in L$. Therefore $uw\in\text{Fin}(L)=L$, that is $u\in L/L'$. In conclusion, $\text{Fin}(L/L')\subseteq L/L'$, hence $\text{Fin}(L/L')=L/L'$, and $\mathcal{FL}_{\text{rreg}}$ is closed under right quotient (with arbitrary languages).

Finally, consider a language $L\in\mathcal{FL}_{\text{rlin}}$, $L=L_{\text{in}}(G)$, $G=(V_N, V, P)$; let L' be an arbitrary language. For $X\in V_N$ set $L_X=L(G_X)$, $G_X=(V_N, V, X, P)$. We define the grammar $G'=(V_N, V, P')$ by

$$P' = \big(P-\{X\to\alpha|\alpha\in V^*,\ X\in V_N\}\big)\cup$$

$$\cup\{X\to\alpha|X\to\alpha\beta\in P,\ \text{for some } \alpha, \beta\in V^*,\ \beta\in L',\ X\in V_N\}$$

$$\cup\{X\to\alpha|X\to\alpha\beta Y\in P,\ \text{for some } \alpha, \beta\in V^*,\ X\in V_N,\ \{\beta\}L_Y\cap L'\neq\emptyset\}.$$

It is easy to see that $L_{\text{in}}(G')=L/L'$, which completes the proof.

Left quotient. Simmetrically.

Init, Fin, Sub. Let $L \in \mathcal{FL}_{rreg}$; in view of Lemma 10, we have $\mathrm{Fin}(L)=L$, $L \in \mathcal{L}_3$. Clearly, $\mathrm{Init}(L)$, $\mathrm{Fin}(L)$, $\mathrm{Sub}(L)$ are regular languages. As $L=\mathrm{Fin}(L)$, we have $\mathrm{Fin}(\mathrm{Init}(L))=\mathrm{Init}(\mathrm{Fin}(L))=\mathrm{Init}(L)$, $\mathrm{Fin}(\mathrm{Fin}(L))=\mathrm{Fin}(L)$, $\mathrm{Fin}(\mathrm{Sub}(L))= =\mathrm{Sub}(L)$. This implies that $\mathrm{Init}(L)$, $\mathrm{Fin}(L)$, $\mathrm{Sub}(L)$ are in $\mathcal{FL}_{rreg}$, too. Similarly for $\mathcal{FL}_{lreg}$, hence also $\mathcal{FL}_{reg}^{\cup}$, $\mathcal{FL}_{reg}^{\cap}$ are closed. The family $\mathcal{FL}_{rlin}$ is closed under right quotient; as $\mathrm{Init}(L)=L/V^*$, we obtain the closure under Init. Consider now $L \in \mathcal{FL}_{rlin}$, $L=L_{in}(G)$, $G=(V_N, V, P)$, and define the grammar $G'=(V_N', V, P')$ by $V_N'=V_N \cup V_N''$, $P'=P \cup P''$, where, for each production r: $X \to a_1a_2 \ldots a_nY \in P$, $a_i \in V$, $1 \leqq i \leqq n$, $Y \in V_N \cup \{\lambda\}$, we introduce in P'' all productions $[X, r, j] \to a_{j+1} \ldots a_nY$, $1 \leqq j \leqq n-1$, simultaneously introducing the new symbols $[X, r, j]$ in V_N''. Clearly, $L_{in}(G')=\mathrm{Fin}(L)$, hence $\mathcal{FL}_{rlin}$ is closed under Fin. Now the closure under Sub follows from the closure under Init.

Similar arguments show that $\mathcal{FL}_{llin}$, hence also $\mathcal{FL}_{lin}^{\cup}$ and $\mathcal{FL}_{lin}^{\cap}$ are closed under Init, Fin, Sub.

Gsm mapping. $L=a^+$ is in $\mathcal{FL}_{reg}^{\cup}$; it is easy to construct a gms g such that $g(L)=ba^+b$. This language is not in $\mathcal{FL}_2$ (Corollary 1), hence none of the above families is closed under gsm mappings.

Inverse gsm mapping. Consider the gsm $g=(\{q_0, q_1, q_2\}, \{a, b\}, \{a\}, q_0, \{q_2\}, \{q_0b \to aq_1, q_1a \to aq_1, q_1b \to aq_2\})$. We have $g^{-1}(a^+)=ba^*b \notin \mathcal{FL}_2$ (Corollary 1), hence none of the above families is closed under inverse gsm mappings.

Intersection with regular sets. As $V^* \in \mathcal{FL}_{reg}^{\cap}$, for each V, but $\mathcal{L}_3 - \mathcal{FL}_{lin}^{\cup} \neq \emptyset$, the assertion is obvious.

6. Further questions

In the proof of inclusions $\mathcal{FL}_i \subseteqq \mathcal{L}_i$, $i=0, 1, 2, 3$, in Theorem 1, starting from the grammar G, used in fully initial manner, we constructed a grammar G' such that $\mathrm{Prod}(G')=\mathrm{Prod}(G)+\mathrm{Var}(G)$. (For an arbitrary grammar $G=(V_N, V_T, S, P)$ we denote, as in [5], $\mathrm{Prod}(G)=\mathrm{card}\, P$, $\mathrm{Var}(G)=\mathrm{card}\, V_N$.) Can the difference between $\mathrm{Prod}(G')$ and $\mathrm{Prod}(G)$ be diminished? More generally, given a language $L \in \mathcal{FL}_i$, define

$$\mathrm{Prod}(L) = \inf\{\mathrm{Prod}(G) | L = L(G)\},$$

$$\mathrm{Prod}_{in}(L) = \inf\{\mathrm{Prod}(G) | L = L_{in}(G)\}.$$

What is the relation between $\mathrm{Prod}(L)$ and $\mathrm{Prod}_{in}(L)$? The construction in the proof of Theorem 1 (used also in [3]) shows that $\mathrm{Prod}(L) \leqq \mathrm{Prod}_{in}(L)+\mathrm{Var}_{in}(L)$. We shall prove that this relation cannot be essentially improved (which shows that, in some sense, the fully initial mode of generating a language is more economical than the usual mode, at least for certain languages).

Indeed, consider the context-free grammar $G=(\{A_1, A_2, \ldots, A_n\}, \{a_1, a_2, \ldots \ldots, a_n, b\}, P)$ with

$$P = \{A_i \to a_iA_ia_i | 1 \leqq i \leqq n\} \cup$$

$$\cup \{A_i \to a_iA_{i+1}a_i | 1 \leqq i \leqq n-1\} \cup \{A_n \to a_nba_n\}.$$

We have

$$L_{\text{in}}(G) = \{a_i^{k_i} a_{i+1}^{k_{i+1}} \dots a_n^{k_n} b a_n^{k_n} \dots a_{i+1}^{k_{i+1}} a_i^{k_i} | 1 \leqq i \leqq n,\ k_j \geqq 1,\ 1 \leqq j \leqq n\}.$$

Consequently, $\text{Prod}_{\text{in}}(L_{\text{in}}(G)) \leqq 2n$, $\text{Var}_{\text{in}}(L_{\text{in}}(G)) \leqq n$. It is easy to see that, in fact, we have $\text{Var}_{\text{in}}(L_{\text{in}}(G)) = n$ (for each i we need a derivation $X_i \overset{*}{\Rightarrow} a_i^j X_i a_i^j$, $j \geqq 1$), hence also $\text{Prod}_{\text{in}}(L_{\text{in}}(G)) = 2n$.

Consider now a usual context-free grammar $G' = (V_N, V_T, S, P')$ such that $L(G') = L_{\text{in}}(G)$. Again, for each i, $1 \leqq i \leqq n$, we need a derivation $X_i \overset{*}{\Rightarrow} a_i^j X_i a_i^j$, $j \geqq 1$, one of the form $X_i \overset{*}{\Rightarrow} a_i^j a_{i+1}^k X_{i+1} a_{i+1}^m a_i^p$, $j, k, m, p \geqq 0$, as well as one of the form $S \overset{*}{\Rightarrow} a_i^j X_i a_i^k$, $j, k \geqq 0$. Two symbols X_i, X_j cannot be identical when $i \neq j$ (otherwise strings containing both substrings $a_i a_j$, $a_j a_i$ on the same side of b could be obtained). Moreover, the axiom S must differ from every X_i, $i \geqq 2$. In conclusion, $\text{Prod}\,(G') \geqq 3n - 1 = \text{Prod}(G) + \text{Var}(G) - 1$, therefore $\text{Prod}\,(L_{\text{in}}(G)) \geqq \text{Prod}_{\text{in}}(L_{\text{in}}(G)) + \text{Var}_{\text{in}}(L_{\text{in}}(G)) - 1$.

Consider now another question. Given a language L and a grammar G for it, $L = L(G)$, what one can say about $L_{\text{in}}(G)$? For example, taking $L = \{a^n b^n | n \geqq 1\} \cdot \{a, b\}^*$ and the grammar $G = (\{S, A, B\}, \{a, b\}, S, \{S \to AB, A \to aAb, A \to ab, B \to aB, B \to bB, B \to \lambda\})$ we obtain $L(G) = L \in \mathscr{L}_2 - \mathscr{L}_3$, $L_{\text{in}}(G) = \{a, b\}^* \in \mathscr{L}_3$.

Are there languages L for which this is not possible (no grammar G, $L = L(G)$ with $L_{\text{in}}(G)$ regular)? The answer is affirmative: take $L = \{a^n b^n | n \geqq 1\}$ and consider a context-free grammar $G = (V_N, \{a, b\}, S, P)$ such that $L = L(G)$ and G is reduced. Clearly, each recursive derivation $X \overset{*}{\Rightarrow} \alpha X \beta$, $\alpha, \beta \in \{a, b\}^*$ must have $\alpha = a^i$, $\beta = b^i$, $i \geqq 1$. For each symbol $A \in V_N$, consider the set $L_A = \{w \in \{a, b\}^* | A \overset{*}{\Rightarrow} w \text{ in } G\}$. If L_A is finite for some A, then, replacing each occurrence of A in the right-hand sides of rules in P by a string in L_A (and removing all rules $A \to \gamma$), we obtain a grammar G', $L(G) = L(G')$, $L_{\text{in}}(G) - L_{\text{in}}(G')$ is finite. The grammar G' obtained in this way be removing all $A \in V_N$ with finite L_A is linear. (If rule $X \to x_1 Y x_2 Z x_3$ is in G', then L_Y, L_Z must be infinite, hence must involve recursive derivations in the generation of their strings, hence L_X contains strings of the form $z_1 a^i b^i z_2 a^j b^j z$, i, $j \geqq 1$, a contradiction.) If $L_{\text{in}}(G)$ is regular, then $L_{\text{in}}(G')$ is regular too (it differs from $L_{\text{in}}(G)$ by a finite set). However, each derivation in G', besides its maximal recursive subderivations, contains at most card V_N further steps. These steps introduce at most $\pi = \text{card}\, V_N \cdot \max\,\{|x|\,|A \to x \in P\}$ occurrences of a and of b. In conclusion, each string in $L_{\text{in}}(G')$ is of the form $a^{n+p} b^{n+q}$, $n \geqq 1$, $p \leqq \pi$, $q \leqq \pi$. This implies $L_{\text{in}}(G') \notin \mathscr{L}_3$, a contradiction.

A further situation which can be looked for is the following. Are there languages $L \in \mathscr{L}_2 - \mathscr{L}_3$ such that each context-free grammar $G, L = L(G)$, has $L_{\text{in}}(G) \in \mathscr{L}_3$? (Such a language can be called *inherently fully initial regular*, whereas the above $L = \{a^n b^n | n \geqq 1\}$ can be called *inherently fully initial context-free*.) This last problem remains *open*.

References

[1] ADJ. — J. A. GOGUEN, J. W. THATCHER, E. G. WAGNER, J. B. WRIGHT, Initial algebra semantics and continuous algebra, *Journal of ACM*, 24, 1 (1977), 68—95.

[2] J. ALBERT, L. WEGNER, Languages with homomorphic replacements, *Lect. Notes Computer Sci.*, 85, Automata, Languages and Programming, Ed. J. W. de Bakker, J. van Leeuwen, Springer-Verlag, 1980, 19—29.

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

[4] S. GINSBURG, *The mathematical theory of context-free languages*, McGraw Hill Book Comp., New York, 1966.

[5] J. GRUSKA, Descriptional complexity of context-free languages, *Proc. Symp. Math. Found. Computer Sci.*, High Tatras, 1973.

[6] G. T. HERMAN, A biologically motivated extension of ALGOL-like languages, *Inform. Control*, 22, 5 (1973), 487—502.

[7] H. A. MAURER, A. SALOMAA, D. WOOD, Pure grammars, *Inform. Control*, 49 (1980), 47—72.

[8] GH. PĂUN, *Contextual grammars*, The Publ. House of the Romanian Acad. of Sci., Bucureşti, 1982 (in Roumanian).

[9] GH. PĂUN, A note on fully initial context-free languages, *Papers on Automata and Languages* (Ed. I. Peák), X (1988), 7—11.

[10] A. SALOMAA, *Formal languages*, Academic Press, New York, London, 1973.

[11] A. VAN WIJNGAARDEN, *Orthogonal design and description of formal languages*, Math. Centrum, Amsterdam, 1965.

(Received June 21, 1988)