JOURNAL OF COMPUTER AND SYSTEM SCIENCES 5, 353–364 (1971)

A Characterization of Context-free Languages

JOZEF GRUSKA*

Department of Computer Information and Control Sciences, University of Minnesota, Minneaopolis, Minnesota 55455

Received March 30, 1970

1. INTRODUCTION AND SUMMARY

The developments in the theory of languages and automata in recent years have been towards defining and studying new grammars, automata and languages, abstract families of automata, languages, translations, and so on. In spite of this fact, the central position of regular sets and context-free languages (CFL's) in the theory of languages remains, and so one can still expect their further study to be useful and maybe stimulating for the general development of language theory.

One of the important advantages in dealing with regular languages is the fact that each such language can be denoted explicitly by a regular expression, in contrast to the implicit representation of languages by grammars. A similar characterization of CFL's would perhaps be useful also. However, one can hardly expect to get a characterization of CFL's so elegant and so simple as the one we have developed for regular languages.

In Section 2 an attempt is made to give a regular-like characterization of CFL's, using union, product, and a new operation-symbol iteration. In so doing, we are compelled to use some "auxiliary" symbols. Their role and number are discussed in Section 3. The application of the results of Section 2 to semilinear languages and their subclasses is presented in Section 4.

2. REGULAR-LIKE CHARACTERIZATION OF CFL'S

In this section a regular-like characterization of CFL's is established.

We begin by introducing some basic concepts and notation of language theory.

If L_1 and L_2 are languages, then the product of L_1 and L_2, in short $L_1 \cdot L_2$ or L_1L_2, is the language $L_1 \cdot L_2 = \{z; z = xy; x \in L_1, y \in L_2\}$. For a language L and integer

* Present Address: Mathematical Institute of Slovak Academy of Sciences, Štefanikova 41, Bratislava, Czechoslovakia.

571/5/4-2

$i \geqslant 0$, L^i is defined as follows: $L^0 = \{\epsilon\}$, where ϵ is the empty word, $L^{i+1} = L^i \cdot L$ for $i \geqslant 0$. Finally, let

$$L^* = \bigcup_{i=0}^{\infty} L^i, \qquad L^+ = \bigcup_{i=1}^{\infty} L^i.$$

A context-free grammar G is a quadruple $G = \langle \Sigma_N , \Sigma_T , P, \sigma \rangle$, where Σ_N is the alphabet of nonterminal symbols, shortly variables; Σ_T is the alphabet of terminal symbols, $\Sigma_N \cap \Sigma_T = \theta$, $\sigma \in \Sigma_N$ is the initial symbol, and P is a finite set of productions of the form $A \to w$, where $A \in \Sigma_N$, $w \in (\Sigma_T \cup \Sigma_N)^*$. For two words x and y, we write $x \Rightarrow y$ if and only if, $x = uAv, y = uwv$ for some u, v in $(\Sigma_T \cup \Sigma_N)^*$ and $A \to w$ in P. Let $\overset{*}{\Rightarrow}$ be the transitive and reflexive closure of $\Rightarrow$. The set $L(G) = \{x; \sigma \overset{*}{\Rightarrow} x \in \Sigma_T{}^*\}$ is called the language generated by G. A language L is said to be context-free if $L = L(G)$ for a context-free grammar G.

The crucial operation in our regular-like characterization of context-free languages is the operation of "symbol iteration" which is similar to a special case of nested iterated substitution of Král and Greibach [5]. To define this operation, we proceed as follows:

DEFINITION 2.1. Let σ be a symbol and L, L_1 be languages. The *σ-substitution of L_1 into L*, denoted by $L \overset{\sigma}{\uparrow} L_1$, is defined by

$$L \overset{\sigma}{\uparrow} L_1 = \{z; z = x_1 y_1 x_2 \cdots x_k y_k x_{k+1} ; x_1 \sigma x_2 \sigma \cdots \sigma x_{k+1} \in L,$$

$$\sigma \text{ does not occur in the word } x_1 x_2 , \ldots, x_{k+1} \text{ and } y_i \in L_1 \text{ for } 1 \leqslant i \leqslant k\}.$$

COROLLARY 2.2. *For each symbol σ, the operation $\overset{\sigma}{\uparrow}$ is associative.*

DEFINITION 2.3. Let σ be a symbol and L be a language. The *σ-iteration* of L, denoted by L^σ is defined by

$$L^\sigma = \{z; z \in L \cup L \overset{\sigma}{\uparrow} L \cup L \overset{\sigma}{\uparrow} L \overset{\sigma}{\uparrow} L \cup \ldots, z \text{ has no occurence of } \sigma\}$$

Remarks. If L is a language over an alphabet Σ_L and σ a symbol, then L^σ is the language generated by the grammar $G = \langle \{\sigma\}, \Sigma_L^{-\{\sigma\}}, P, \sigma \rangle$ where $P = \{(\sigma \to x); x \in L\}$ is, in general, an infinite set of productions. If L is a finite language, then L^σ is a context-free language.

If σ is not in Σ_L , then $L^* = (L\sigma \cup \{\epsilon\})^\sigma$ and $L^+ = (L\sigma \cup L)^\sigma$.

The following lemma summarizes some relations between the operations of symbol substitution and symbol iteration.

LEMMA 2.4. *Let L, L_1, L_2 be languages over the alphabets $\Sigma_L, \Sigma_{L_1}, \Sigma_{L_2}$, respectively, and σ, ρ be symbols. Then*

(i) $(L^\sigma)^\rho = (L^\rho)^\sigma$;

(ii) if $\sigma \notin \Sigma_L$, then

$$(L_1 \overset{\sigma}{\uparrow} L_2) \overset{\rho}{\uparrow} L = (L_1 \overset{\rho}{\uparrow} L) \overset{\sigma}{\uparrow} (L_2 \overset{\rho}{\uparrow} L);$$

(iii) if $\sigma \notin \Sigma_L$, then

$$L \overset{\rho}{\uparrow} (L_1 \overset{\sigma}{\uparrow} L_2) = (L \overset{\rho}{\uparrow} L_1) \overset{\sigma}{\uparrow} L_2;$$

(iv) if $\sigma \notin \Sigma_L$, then

$$(L_1 \overset{\rho}{\uparrow} L)^\sigma = (L_1{}^\sigma) \overset{\rho}{\uparrow} L.$$

Proof. Clearly, in order to prove (i) it is sufficient to show that $(L^\sigma)^\rho \subseteq (L^\rho)^\sigma$. If $z \in (L^\sigma)^\rho$, then there exists a sequence of words $z_1, z_2, \ldots, z_r$ such that $z_1 \in L^\sigma$, $z_k = z$, and for $i = 1, 2, \ldots, r-1$, $z_i = u_i \rho v_i$, $z_{i+1} = u_i w_{i+1} v_i$ with $w_{i+1} \in L^\sigma$. If we set $w_1 = z_1$, then for $i = 1, 2, \ldots, r$, $w_i \in L^\sigma$ and, therefore, for each such i, there exists a sequence

$$z_1^i, z_2^i, \ldots, z_{k_i}^i,$$

such that $z_1{}^i \in L$, $z_{k_i}^i = w_i$ and for $j = 1, 2, \ldots, k_i - 1$, z_{j+1}^i is obtained from $z_j{}^i$ by replacing a σ by a word from L. Now, let us form the sequence

$$(*) \qquad z_1{}^1, \ldots, z_{k_1}^1 = z_1, u_1 z_1{}^2 v_1, \ldots, u_1 z_{k_2}^2 v_1 = z_2, u_2 z_1{}^3 v_2, \ldots, u_{r-1} z_{k_r}^r v_{r-1} = z$$

which is actually "a derivation of z from $z_1{}^1$ in context-free grammar G with two non-terminals σ and ρ and the rules $\sigma \to \alpha$, $\rho \to \alpha$ for every $\alpha \in L$. Derivation $(*)$ has the property that the symbol ρ is replaced by a word from L if and only if there are no σ to be replaced. Obviously, there is a derivation tree corresponding to $(*)$ and to this derivation tree corresponds another derivation of z from $z_1{}^1$ in G with the property that σ is replaced if and only if there are no ρ to be replaced. But that means that $z \in (L^\rho)^\sigma$. Hence $(L^\sigma)^\rho \subseteq (L^\rho)^\sigma$ completing the proof of (i). (ii) to (iv) are now obvious.

DEFINITION 2.5. Let Σ be a (possibly infinite) alphabet. Denote by $\mathscr{E}_\Sigma$ the smallest class of languages satisfying the following conditions:

(1) $\{\epsilon\} \in \mathscr{E}_\Sigma$,

(2) if $a \in \Sigma$, then $\{a\} \in \mathscr{E}_\Sigma$,

(3) if L_1, L_2 are in $\mathscr{E}_\Sigma$ and $\sigma \in \Sigma$, then $L_1 \cup L_2$, $L_1 \cdot L_2$ and $L_1{}^\sigma$ are in $\mathscr{E}_\Sigma$.

Remark. We will say that L is a language over an infinite alphabet Σ if $L \subset \Sigma_L{}^*$ for some finite subset $\Sigma_L \subset \Sigma$.

LEMMA 2.6. *If Σ is an infinite alphabet, $\sigma \in \Sigma$ and L_1, L_2 are languages in $\mathscr{E}_\Sigma$, then $L_1 \overset{\sigma}{\uparrow} L_2$ is in $\mathscr{E}_\Sigma$.*

Proof. Let L_2 in $\mathscr{E}_\Sigma$ be fixed. If $a \in \Sigma$, then $\{a\} \overset{\sigma}{\uparrow} L_2$ is either $\{a\}$ or L_2 so is in $\mathscr{E}_\Sigma$. The collection Ξ of languages L such that $L \overset{\sigma}{\uparrow} L_2$ is in $\mathscr{E}_\Sigma$ is closed under union, product, and σ_0-iterations $\sigma_0 \in \Sigma$. Since Ξ contains $\{\epsilon\}$ and $\{a\}$ for every $a \in \Sigma$, we see that Ξ contains $\mathscr{E}_\Sigma$. Thus, if L_1 is in $\mathscr{E}_\Sigma$, then $L_1 \overset{\sigma}{\uparrow} L_2$ is again in $\mathscr{E}_\Sigma$. Since L_2 is arbitrary, this completes the proof.

Now we can state and prove the main result of this section, namely that each CFL can be represented by the operations $\cup$, $\cdot$ and symbol iteration.

THEOREM 2.7. *If Σ is an infinite alphabet, then L is a (nonempty)* CFL *over Σ if and only if $L \in \mathscr{E}_\Sigma$.*

Proof. In view of the Definition 2.5, in order to show that each language in $\mathscr{E}_\Sigma$ is context-free, it is sufficient to show that if L is a context-free language and σ is a symbol then L^σ is also a CFL. Let L be given, and let G be a CFG for L such that σ is not a variable of G. L^σ is now exactly the language generated by the grammar G' which has σ as the initial symbol and the productions of G' are those of G and the production $\sigma \to \sigma_0$, where σ_0 is the initial symbol of G. Thus L^σ is a CFL.

To complete the proof it remains to show that $\mathscr{E}_\Sigma$ contains all CFL's over Σ.

Let L be a nonempty CFL over an alphabet $\Sigma_L \subset \Sigma$. Since Σ is an infinite alphabet, there exists a CFG $G = \langle \Sigma_N, \Sigma_L, P, \sigma \rangle$ generating L and such that $\Sigma_N \subset \Sigma$. Let $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_n$ be an enumeration of variables of G with $\sigma_i \neq \sigma_j$ for $i \neq j$. We can assume without loss of generality that the sets $\theta_i = \{x; \sigma_i \overset{*}{\Rightarrow} x \in \Sigma_L{}^*\}$, $1 \leqslant i \leqslant n$, are nonempty.

Let us now define, for $1 \leqslant i \leqslant n$, the sets

$$V_i = \{\sigma_j ; i \leqslant j \leqslant n\},$$
$$P_i = \{(\sigma' \to \alpha); (\sigma' \to \alpha) \in P, \sigma' \in V_i\},$$

and the grammars

$$G_i = \langle V_i, \Sigma_n \cup (\Sigma_N - V_i), P_i, \sigma_i \rangle.$$

Moreover, define, for $1 \leqslant i \leqslant k \leqslant n$, the languages $\Phi_{i,k}$, by double downward induction from n as follows:

$$\Phi_{i,n} = \{\alpha; (\sigma_i \to \alpha) \in P\} \qquad \text{for} \quad 1 \leqslant i \leqslant n$$

and, for $i < k \leqslant n$,

$$(*) \qquad \Phi_{i,k-1} = \Phi_{i,k} \overset{\sigma_k}{\uparrow} \Phi_{k,k}^{\sigma_k}.$$

By Lemma 2.6, the sets $\Phi_{i,k}$ are in $\mathscr{E}_\Sigma$. Using (*), Definitions 2.1 and 2.3 and downward induction for k from n, we get, for $1 \leqslant i \leqslant k \leqslant n$,

$$\Phi_{i,k} = \{x; \text{there exists an } \alpha \in \Phi_{i,n} \text{ such that } \sigma_i \to \alpha \overset{*}{\Rightarrow} x \text{ in } G, x \text{ does not contain symbols from } V_{k+1}, \text{ and there exists a derivation of } x \text{ from } \alpha \text{ in } G \text{ which uses only the productions from } P_{k+1}\},^1$$

and

$$\Phi_{i,i}^{\sigma_i} = \{x; \sigma_i \overset{*}{\Rightarrow} x, x \text{ does not contain symbols from } V_i, \text{ and there exists a derivation of } x \text{ from } \sigma_i \text{ which uses only the productions from } P_i\}.$$

Therefore $\Phi_{i,i}^{\sigma_i} \subseteq L(G_i)$ for $1 \leqslant i \leqslant n$. To complete the proof of the theorem, it is sufficient to show that $L(G_i) \subseteq \Phi_{i,i}^{\sigma_i}$ also for $1 \leqslant i \leqslant n$. To prove that, we will use a downward induction for i from n.

It is clearly true for $i = n$. Assume now that $L(G_j) \subseteq \Phi_{j,j}^{\sigma_j}$ (and, therefore, $L(G_j) = \Phi_{j,j}^{\sigma_j}$) for $j = i+1, i+2, \ldots, n$. To show that then also $L(G_i) \subseteq \Phi_{i,i}^{\sigma_i}$, we proceed as follows.

For $w \in (\Sigma_n \cup \Sigma_L)^* - \Sigma_L^*$ let $\pi(w) = \max\{j; \sigma_j \text{ occurs in } w\}$. First, we will show that

$$\text{If } \sigma_i \to \alpha \text{ is in } P \text{ and } x \text{ can be generated from } \alpha \text{ in } G_{i+1}, \text{ then } x \in \Phi_{i,i}. \tag{1}$$

Indeed, let x have the above property. Then there exists a derivation τ of x from σ_i;

$$\tau : \sigma_i = w_1, w_2, \ldots, w_l = x$$

such that $\pi(w_j) > i$ for $1 < j \leqslant l$ and in each derivation step a symbol σ_j with the maximal j is replaced. To simplify the ensuing discussion, let $\pi(j)$ to denote $\pi(w_j)$ for $1 \leqslant j \leqslant l$. Let now $k_1, k_2, \ldots, k_r$ be a sequence of integers defined as follows: $k_1 = 2$. Assume that $k_1, k_2, \ldots, k_j$ have already been defined. If $k_j = l$, then $j = r$. If $k_j < l$ and $\pi(\nu) \geqslant \pi(k_j)$ for $k_j < \nu < l$, then $k_{j+1} = l$; otherwise, $k_{j+1} = \min\{\nu; \nu > k_j, \pi(\nu) < \pi(k_j)\}$. Clearly, w_{k_2} is obtained from $w_{k_1} = w_2$ by replacing all symbols $\sigma_{\pi(2)}$ by some words from $L(G_{\pi(2)}) = \Phi_{\pi(2),\pi(2)}^{\pi(2)}$. Thus $w_{k_2} \in \Phi_{i,\pi(2)-1}$. In general, if $2 < j < l$ and $w_{k_j} \in \Phi_{i,\pi(k_{j-1})-1}$, then also $w_{k_j} \in \Phi_{i,\pi(k_j)}$ and $w_{k_{j+1}}$ is obtained from w_{k_j} by replacing all symbols $\sigma_{\pi(k_j)}$ by some words from $L(G_{\pi(k_j)})$ and, therefore, $w_{k_{j+1}} \in \Phi_{i,\pi(k_j)-1}$. Thus $x \in \Phi_{i,i}$ and (1) is proven.

Let now $x \in L(G_i)$ and again let

$$\sigma_i = w_1, w_2, \ldots, w_l = x$$

be a derivation of x such that in each derivation step a symbol σ_j with the maximal

[1] Assume for the sake of simplicity that $V_{n+1} = \theta = P_{n+1}$.

index is replaced. Let $1 = k_1 < k_2 < k_3 < \cdots < k_r = l$ be the sequence of all integers j such that $\pi(j) = i$ and $j < l$. It means by (1), that $w_{k_2} \in \Phi_{i,i}$. Moreover, for $j > 2$, w_{k_j} is obtained from $w_{k_{j-1}}$ by replacing the symbols σ_i by some words which are, by (1), in $\Phi_{i,i}$. Since x has no occurence of σ_i, x must be in $\Phi_{i,i}^{\sigma_i}$. Thus $L(G_i) \subseteq \Phi_{i,i}^{\sigma_i}$ and it completes the proof of the Theorem.

Remark. Regular languages are often defined as languages represented by regular expressions. In view of the last theorem, we can define also CFL's over an infinite alphabet Σ as languages represented by context-free expressions (CFE's) over Σ which are defined as follows (here we assume that for each $\sigma \in \Sigma$, $\boldsymbol{\sigma}$ is a special symbol not in Σ):

(1) if $a \in \Sigma \cup \{\epsilon\}$, then a is a CFE

(2) if E_1, E_2 are CFE's and $\sigma \in \Sigma$, then $(E_1 \cup E_2)$, $(E_1 \cdot E_2)$, $(E_1\boldsymbol{\sigma})$ are CFE's.

Here we assume that if a CFE E represents a language L, then $(E\boldsymbol{\sigma})$ represents L^{σ}.

COROLLARY 2.8. *If Σ is a finite alphabet, then L is a* CFL *over Σ if and only if $L \in \mathscr{E}_{\Sigma'}$, for some finite alphabet $\Sigma' \supseteq \Sigma$.*

Proof. Straightforward from Theorem 2.7.

EXAMPLE 1. If $L = \{a^i b^i, i \geqslant 1\}$, then $L = (a\sigma b \cup ab)^{\sigma}$.

EXAMPLE 2. Let L be the language generated by the grammar with the initial symbol σ_1, and the productions

$$\sigma_1 \to a\sigma_1 b, \qquad \sigma_1 \to ba^2\sigma_2 ba,$$
$$\sigma_2 \to a^2\sigma_2 b, \qquad \sigma_2 \to ba^3\sigma_3 ba,$$
$$\sigma_3 \to a^3\sigma_3 b, \qquad \sigma_3 \to ba\sigma_1 ba, \qquad \sigma_3 \to b.$$

Then $L = (ba^2(ba^3(a^3\sigma b \cup ba\rho ba \cup b)^{\sigma} ba \cup a^2\sigma b)^{\sigma} ba \cup a\rho b)^{\rho}$.

3. AUXILIARY SYMBOLS

For a finite alphabet Σ and an integer k, let $\mathscr{E}_{\Sigma,k}$ be the class of all CFL's over Σ which are in $\mathscr{E}_{\Sigma'}$, where $\Sigma' = \Sigma \cup \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ with $\sigma_1, \sigma_2, \ldots, \sigma_k$ being distinct symbols not in Σ. (In other words, in $\mathscr{E}_{\Sigma,k}$, k stands for the number of "auxiliary" symbols which are not in Σ but which may be used in regular-like constructions of the languages in $\mathscr{E}_{\Sigma,k}$.)

A natural question arises. Given a finite alphabet Σ, is there a k such that $\mathscr{E}_{\Sigma,k}$

contains all CFL's over Σ? We strongly suspect that if Σ has at least two symbols, then the answer is negative, but we have no proof. If Σ has only one symbol then all CFL's over Σ are regular and they are in $\mathscr{E}_{\Sigma,1}$.

Following the proof of Theorem 2.7, it would seem that the auxiliary symbols used in regular-like constructions of CFL's play the same role as variables in context-free grammars. It is known [7] that, for any integer n and any alphabet Σ with at least two symbols, there is a linear CFL L_n which cannot be generated by a CFG with less than n variables. For instance, if $\Sigma = \{a, b\}$, then the language generated by the grammar with the initial symbol σ and the productions

$$\sigma_i \to a^i\sigma_i b, \qquad \sigma_i \to ba^{i+1}\sigma_{i+1}ba, \qquad 1 \leqslant i \leqslant n-1,$$
$$\sigma_n \to a^n\sigma_n b, \qquad \sigma_n \to ba\sigma_1 ba,\ \sigma_n \to b^2a^2$$

has such property. However, using the same techniques as in the example 2, Section 1, we can show that this language is in $\mathscr{E}_{\Sigma,2}$. We can even prove the theorem.

THEOREM 3.1. *For any finite alphabet Σ, the class $\mathscr{E}_{\Sigma,2}$ contains all linear languages over Σ.*

The theorem follows instantly from the following lemma which is easier to prove.

LEMMA 3.2. *Let $G = \langle \Sigma_N, \Sigma, P, \sigma\rangle$ be a linear context-free grammar and A, B be two distinct symbols not in $\Sigma_N \cup \Sigma$. Let $G' = \langle \Sigma_N, \Sigma \cup \{A\}, P \cup P', \sigma\rangle$ be a new grammar formed from G by adding a finite set P' of productions of the form $X \to uAv$ where $X \in \Sigma_N$ and $uv \in \Sigma^*$. Then $L(G) \in \mathscr{E}_{\Sigma\cup\{A,B\}}$, $L(G') \in \mathscr{E}_{\Sigma\cup\{A,B\}}$, and both $L(G)$ and $L(G')$ may be obtained from languages $\{a\}$; $a \in \Sigma \cup \{A, B\} \cup \{\epsilon\}$ by a finite number of operations $\cup$, $\cdot$, A-iteration and B-iteration.*

Proof. We will use induction on the number of variables of G. The lemma clearly holds if G has only one variable. Assume now that the lemma holds for all linear grammars with at most $k-1$ variables and let $G = \langle \Sigma_N, \Sigma, P, \sigma\rangle$ be a linear grammar with k variables $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_k$. Define grammars G_i, $2 \leqslant i \leqslant n$, by $G_i = \langle \Sigma_N - \{\sigma\}, \Sigma \cup \{\sigma\}, \bar{P}, \sigma_i\rangle$, $\bar{P} = P - \{(\sigma \to \alpha);\ (\sigma \to \alpha) \in P\}$. Each G_i, $2 \leqslant i \leqslant n$ has $k-1$ variables and, therefore, by the second part of the induction hypothesis, $L(G_i) \in \mathscr{E}_{\Sigma\cup\{\sigma,B\}}$ and $L(G_i)$ can be obtained from languages $\{a\}$; $a \in \Sigma \cup \{\sigma, B\} \cup \{\epsilon\}$ by a finite number of operations $\cup$, $\cdot$, σ-iteration, and B-iteration.

Let now $\sigma \to \alpha_j$, $1 \leqslant j \leqslant r$, be an enumeration of all productions of G with σ on the left side. Define the languages L_j, $1 \leqslant j \leqslant r$, as follows: If $\alpha_j \in (\Sigma \cup \{\sigma\})^*$, then $L_j = \{\alpha_j\}$. If $\alpha_j = u\alpha_l v$ for some $l > 1$, then $L_j = uL(G_l)v$. Clearly, $L(G) = (\bigcup_{j=1}^{r} L_j)^\sigma$ and, therefore, $L(G) \in \mathscr{E}_{\Sigma\cup\{\sigma,B\}}$. Clearly, σ can be replaced by A and it completes the induction step for G.

Let now $G' = \langle \Sigma_N, \Sigma \cup \{A\}, P \cup P', \sigma\rangle$ where P' is a finite set of productions of

the form $X \to uAv$; $X \in \Sigma_N$, $uv \in \Sigma$. Define, for $2 \leqslant i \leqslant k$, the grammars G_i and G_i' as follows:

$G_i = \langle \Sigma_N - \{\sigma\}, \Sigma \cup \{A\}, P_i, \sigma_i \rangle$, where P_i is obtained from $P \cup P'$ by deleting all productions which contain the symbol σ (either on the left or the right side).

$G_i' = \langle \Sigma_N - \{\sigma\}, \Sigma \cup \{\sigma\}, P_i', \sigma_i \rangle$, where P_i' is obtained from P by deleting all productions with σ on the left side.

According to the induction hypothesis, both $L(G_i)$ and $L(G_i')$ are in $\mathscr{E}_{\Sigma \cup \{\sigma, A\}}$ and can be obtained from languages $\{a\}$, $a \in \Sigma \cup \{\sigma, A\} \cup \{\varepsilon\}$ using only $\cup$, $\cdot$, σ-iteration, and A-iteration.

Now let $\sigma \to \alpha_j$, $1 \leqslant j \leqslant r$ be an enumeration of all productions of G with σ on the left side. Define the languages L_j, $1 \leqslant j \leqslant r$ as follows. If $\alpha_j \in (\Sigma \cup \{\sigma, A\})^*$, then $L_j = \{\alpha_j\}$; if $\alpha_j = u\sigma_l v$ for some $l > 1$, then $L_j = u(L(G_l) \cup L(G_l'))v$. Clearly, $L(G') = (\bigcup_{j=1}^{r} L_j)^{\sigma}$ and the lemma holds for G', too. This completes the induction step and thereby the proof of the lemma.

Remark. Theorem 3.1 also holds for semilinear languages (see Section 4) because the class of semilinear languages is the smallest class containing linear languages and closed under substitution.

Theorem 3.1 implies that from the fact that we need infinitely many variables to generate all CFL's we cannot conclude that we also need infinitely many auxiliary symbols in regular-like constructions of CFL's and, therefore, auxiliary symbols play a slightly different role.

Some auxiliary symbols are necessary because the class $\mathscr{E}_{\Sigma}$, where Σ is a finite alphabet does not contain all regular languages over Σ. However, they are in $\mathscr{E}_{\Sigma,1}$.

There are also nonlinear CFL's which cannot be generated by grammars with less than n nonterminal symbols but they are in $\mathscr{E}_{\Sigma,2}$. For example, the language generated by the grammar with the initial symbol σ_1 and the rules

$$\sigma_i \to a^i \sigma_i b, \qquad \sigma_i \to ca\sigma_1 bca^2\sigma_2 bc \cdots ca^n\sigma_n bc, \qquad \sigma_i \to c, \qquad 1 \leqslant i \leqslant n$$

has such a property.

4. Application to Semilinear Languages

In view of the results of Section 2, each CFL can be characterized in a simple form using only operations of union, product and symbol iterations. The secret here lies in the fact that symbol iterations are quite complicated and not very transparent operations. In the case of linear and semilinear languages, symbol iterations have much

simpler and transparent form and these languages can be characterized by simpler operations. The results of this section are based on the report [6] and have much in common with results of Yntema [8] and Brzozowski [2].

THEOREM 4.1. *Let Σ be an alphabet. Let $\mathbf{S} \subset 2^{\Sigma^*}$, $\mathbf{D} \subset 2^{\Sigma^* \times \Sigma^*}$ be the smallest classes such that the following conditions are satisfied*:

(1) if $a, b \in \Sigma \cup \{\epsilon\}$, then $\{a\} \in \mathbf{S}$, $\{\langle a, b\rangle\} \in \mathbf{D}$

(2) if $X, Y \in \mathbf{S}$, $P, Q \in \mathbf{D}$, then

(2a) $X \cup Y \in \mathbf{S}$, $P \cup Q \in \mathbf{D}$,

(2b) $P \otimes X = \{p_1 x p_2 ; \langle p_1, p_2\rangle \in \mathbf{D}, x \in X\} \in \mathbf{S}$,

(2c) $P \odot Q = \{\langle p_1 g_1, g_2 p_2\rangle ; \langle p_1, p_2\rangle \in P; \langle g_1, g_2\rangle \in Q\} \in \mathbf{D}$,

(2d) $P^+ = P \cup P \odot P \cup P \odot P \odot P \cup \cdots \in \mathbf{D}$.

Then a language L over Σ is linear if and only if $L \in \mathbf{S}$.

Proof. To prove the "if", it is sufficient to show that for every $L \in \mathbf{S}$, $P \in \mathbf{D}$ and σ, a symbol not in Σ, L and $P \otimes \{\sigma\}$ are linear languages. It is certainly true if $L = \{a\}$, and $P = \{\langle a, b\rangle\}$, where a, b are in $\Sigma \cup \{\epsilon\}$. By (2a) through (2d), if it is true for X, Y in $\mathbf{S}$ and P, Q in $\mathbf{D}$, then it is also true for $X \cup Y$, $P \cup Q$, $P \otimes X$, $P \odot Q$, P^+ and, therefore, for all $X \in \mathbf{S}$, $P \in \mathbf{D}$.

To show the "only if", it is sufficient to repeat the second part of the proof of Theorem 2.7 and to notice that if G is a linear grammar, then all sets $\phi_{i,k}$ have the form $\phi_{i,k} = \bigcup_{j=1}^{k} P_j^{(i,k)} \otimes \{\sigma_j\} \cup L_{i,k}$, where

$$P_j^{(i,k)} \in \mathbf{D}, L_{i,k} \in \mathbf{S} \qquad \text{and} \qquad \phi_{i,i}^{\sigma_1} = \bigcup_{j=1}^{i-1} P_j^{(i)} \otimes \{\sigma_j\} \cup L_i,$$

where again $P_j^{(i)} \in \mathbf{D}$ and $L_i \in \mathbf{S}$.

Thus $L(G) = \phi_{1,1}^{\sigma_1} \in \mathbf{S}$.

A similar characterization can be obtained for semilinear languages.

DEFINITION 4.2. A grammar $G = \langle \Sigma_N, \Sigma, P, \sigma\rangle$ is said to be semilinear if for every $A \in \Sigma_N$ and $x \in (\Sigma \cup \Sigma_N)^*$, $A \overset{*}{\Rightarrow} x$ implies that there is at most one occurence of A in x. A language L is said to be semilinear if L is generated by a semilinear grammar.

There is a great variety of terms used for semilinear languages and grammars: standard matching choice sets [8], nonexpansive grammars, derivation-bounded languages [4], superlinear grammars and languages [2]. The motivation for introducing one more term is Lemma 4.4 which is also useful to get further results.

DEFINITION 4.3. Let $G = \langle \Sigma_N , \Sigma_T , P, \sigma \rangle$ be a CFG. Let $\equiv$ be a relation on P which is defined as follows: $(A \to \alpha) \equiv (B \to \beta)$ if and only if either $A = B$ or $A \overset{*}{\Rightarrow} u_1 B v_1$ and $B \overset{*}{\Rightarrow} u_2 A v_2$ for some u_1 , u_2 , v_1 , v_2 in $(\Sigma_N \cup \Sigma_T)^*$. $\equiv$ is the equivalence relation on P. The equivalence classes on P with regard to $\equiv$ are called grammatical levels of G.

Remark. A grammatical level G_0 of a CFG G is a set of context-free productions of G. We can think of G_0 as being a CFG without an initial symbol, the variables of which are symbols on the left sides of productions in G_0 and all other symbols occurring in productions of G_0 are terminal regarding G_0. Thus a non terminal symbol of G may be a terminal symbol of G_0.

If we consider only reduced CFG's, i.e., the grammars in which each nonterminal symbol is reachable from the initial symbol and each nonterminal symbol generates a nonempty set of terminal words, then using definitions 4.3 and 4.2 we can easily prove the following lemma.

LEMMA 4.4. *A grammar G is semilinear if and only if every grammatical level G_0 of G is linear (with respect to nonterminal symbols of G_0).*

Proof. If G is not semilinear, then there exists a nonterminal A in G such that $A \overset{*}{\Rightarrow} uAwAv$ for some words, u, w, v and, therefore, the grammatical level of G containing the rules with A on left side is not linear. On the other hand, if a grammatical level of G is not linear, then G contains a rule $A \to uBwCv$, where B and C are nonterminals such that the rules with B and C on left side are in the same grammatical level as $A \to uBwCv$. Hence, there exist words u_1 , v_1 , u_2 , v_2 such that $B \overset{*}{\Rightarrow} u_1 A v_1$, $C \overset{*}{\Rightarrow} u_2 A v_2$, and, therefore, G is not semilinear.

From the foregoing lemma follows also the well-known fact that the class of semilinear languages is closed under substitution. An analog of Theorem 4.1 for semilinear languages has the form:

THEOREM 4.5. *Let Σ be an alphabet. Define $\mathbf{S} \subset 2^{\Sigma^*}$, $\mathbf{D} \subset 2^{\Sigma^* \times \Sigma^*}$ to be the smallest classes for which conditions* (1) *to* (2d) *of Theorem* 4.1 *and the condition*

$$\langle X, Y \rangle = \{\langle x, y \rangle ; x \in X, y \in Y\} \in \mathbf{D} \tag{2e}$$

are satisfied. Then L is a semilinear language over Σ if and only if $L \in \mathbf{S}$.

Proof. Notice, that from (2e) it follows that if $X, Y \in \mathbf{S}$, then so is XY. The proof of the "if" goes through silimarly as for Theorem 4.1. Now let L_0 be a linear language over an alphabet $\Sigma' \supset \Sigma$, and let L be a language which is obtained from L_0 by applying a substitution ϕ where $\phi(a) = \{a\}$ for $a \in \Sigma$ and $\phi(a) \in \mathbf{S}$ for $a \notin \Sigma$. It follows from Theorem 4.1 and from (2e) that L is in $\mathbf{S}$. Thus $\mathbf{S}$ contains all languages over Σ which arise from linear languages by substitution and therefore $\mathbf{S}$ contains all semilinear languages over Σ.

If condition (2e) is replaced by

$$XY \in \mathbf{S}, \tag{2e}'$$

then $L \in \mathbf{S}$ if and only if L is an ultralinear language.

Ultralinear languages [3] are languages generated by nonterminal bounded grammars [1]. A context-free grammar $G = \langle \Sigma_N , \Sigma_T , P, \sigma \rangle$ is nonterminal bounded if there is an integer k such that if $\sigma \overset{*}{\Rightarrow} x$, then x has at most k nonterminal symbols. It is easy to show that a context-free grammar G is nonterminal bounded if and only each grammatical level of G is linear; and, moreover, if $A \to xByCz$ is a rule of G with variables B, C, then no rule with B or C on the left side is in the same grammatical level of G as $A \to xByCz$. Using this fact and Theorem 4.1 we can get our result easily.

By introducing a new operation on sets of pairs of words we can get a more elegant characterization of semilinear languages.

THEOREM 4.6. *Let Σ be an alphabet. Define $\mathbf{D}_0 \subset 2^{\Sigma^* \times \Sigma^*}$ to be the smallest class such that the following conditions hold*:

(i) if $a \in \Sigma$, then $\{\langle a, \epsilon \rangle\} \in \mathbf{D}_0$, $\{\langle \epsilon, \epsilon \rangle\} \in \mathbf{D}_0$

(ii) if P, Q are in $\mathbf{D}_0$, then

(iia) $P \cup Q$, $P \odot Q$, $P^+ \in \mathbf{D}_0$

(iib) $P \circledast Q = \{\langle p_1 p_2 , g_1 g_2 \rangle ; \langle p_1 , p_2 \rangle \in P, \langle g_1 g_2 \rangle \in Q\} \in \mathbf{D}_0$.

Put $\mathbf{S}_0 = \{P \otimes \{\epsilon\}; P \in \mathbf{D}_0\}$. Then $L \in \mathbf{S}_0$ if and only if L is a semilinear language over Σ.

Proof. It is sufficient to show that $\mathbf{S} = \mathbf{S}_0$, $\mathbf{D} = \mathbf{D}_0$, where $\mathbf{S}$ and $\mathbf{D}$ are defined as in Theorem 4.5. To show that $\mathbf{S} \subseteq \mathbf{S}_0$, $\mathbf{D} \subseteq \mathbf{D}_0$ it is sufficient to show that the conditions (1) to (2e) in Theorem 4.5 hold for $\mathbf{S}_0$, $\mathbf{D}_0$. To see that it is sufficient to show that if a, b are in $\Sigma \cup \{\epsilon\}$ and X, Y in $\mathbf{S}_0$, then $\langle a, b \rangle \in \mathbf{D}_0$, $X \cup Y \in \mathbf{S}_0$, $\langle X, Y \rangle \in \mathbf{D}_0$. But $\langle a, b \rangle = \langle a, \epsilon \rangle \circledast \langle b, \epsilon \rangle \in \mathbf{D}_0$ and if X, Y are in $\mathbf{S}_0$, then $X = P \otimes \{\epsilon\}$, $Y = Q \otimes \{\epsilon\}$ with P, Q in $\mathbf{D}_0$ and, therefore, $X \cup Y = (P \cup Q) \otimes \{\epsilon\} \in \mathbf{S}_0$, $\langle X, Y \rangle = P \odot Q \in \mathbf{D}_0$. In a similar way, the relations $\mathbf{S}_0 \subseteq \mathbf{S}$, $\mathbf{D}_0 \subseteq \mathbf{D}$ can be shown. Hence the Theorem.

REFERENCES

1. R. B. BANERJI, Phrase structure languages, finite machines and channel capacity, *Information and Control* **6** (1963), 153–162.
2. Y. A. BRZOZOWSKI, "Regular-Like Expressions for Some Irregular Languages," IEEE Conference record of 1968, Ninth Annual Symposium on Switching and Automata Theory, pp. 278–280, 1968.

3. S. GINSBURG AND E. H. SPANIER, Finite-turn pushdown automata, *SIAM J. Control* **4** (1966), 429–453.
4. S. GINSBURG AND E. H. SPANIER, "Derivation-Bounded Languages," *J. Comput. System Sci.* **2** (1968), 228–250.
5. S. A. GREIBACH, "Full AFL's and Nested Iterated Substitution," IEEE Conference record of 1969 Tenth Annual Symposium on Switching and Automata Theory, pp. 222–230, 1969.
6. J. GRUSKA, "Generalization of Regular Sets," Abstracts of the Mathematics Congress, Moscow, 1966.
7. J. GRUSKA, On a Classification of context-free grammars, *Kybernetika Prague* **3** (1967), 22–29.
8. M. K. YNTEMA, Inclusion relations among families of context-free languages, *Information and Control* **10** (1967), 572–597.