INFORMATION AND CONTROL 37, 115–133 (1978)

Algebraic Properties of Operator Precedence Languages*

STEFANO CRESPI-REGHIZZI† AND DINO MANDRIOLI

Istituto di Elettrotecnica ed Elettronica, Politecnico di Milano, 20133 Milano, Piazza Leonardo da Vinci, 32, Italy

AND

DAVID F. MARTIN

Computer Science Department, University of California, Los Angeles, Los Angeles, California 90024

This paper presents new results on the algebraic ordering properties of operator precedence grammars and languages. This work was motivated by, and applied to, the mechanical acquisition or inference of operator precedence grammars. A new normal form of operator precedence grammars called *homogeneous* is defined. An algorithm is given to construct a grammar, called *max-grammar*, generating the largest language which is compatible with a given precedence matrix. Then the class of *free* grammars is introduced as a special subclass of operator precedence grammars. It is shown that operator precedence languages corresponding to a given precedence matrix form a Boolean algebra.

1. INTRODUCTION

In this paper we present a study of the algebraic properties of operator precedence grammars which share a common set of precedence relations. Operator precedence grammars (Floyd, 1963) have been extensively used to define the syntax of programming languages.

Motivation for this study arose from an attempt to improve the process of designing the syntax of new programming languages, especially when a grammar must be constructed to generate exactly those sentences (programs) intended by the language designer. To construct such a grammar by hand is a lengthy procedure, and a method for automating the construction has been proposed by Crespi-Reghizzi (1970, 1973) and subsequently applied by Fu and Booth (1975) to pattern recognition problems. The proposed method is an example of a grammatical inference algorithm, as defined in Gold (1967), Feldman (1972), and surveyed in Fu and Booth (1975).

* Work supported by CNR.

† Present address: Istituto Scienze Informazione, Pisa, Italy.

0019-9958/78/0372-0115$02.00/0

The paper is organized as follows. Notation and preliminary statements are introduced in Sect. 2. In Sect. 3 we establish a new normal form for operator precedence grammars, and we define the central notions of free grammars and maxgrammars. In Sect. 4 we discuss the lattice properties of related families of grammars, languages and precedence matrices. In Sect. 5 we prove the closure of some operator precedence languages under boolean operations. Section 6 presents the conclusion and some suggestions for extensions. App. 1 contains a complete example and App. 2 presents some special properties of free operator precedence languages.

2. Notation and Preliminary Statements

Let Z be a countable set. $\mathscr{P}(Z)$ will denote the power set of Z, $|Z|$ its cardinality, and $\varnothing$ the empty set. Posets and lattices will be denoted, respectively, as pairs (Z, R) and 4-tuples (Z, R, J, M), where R is the ordering relation, J and M are, respectively, the join and meet operations. In accordance with Hopcroft (1969) a context-free grammar (CFG) will be denoted by a 4-tuple $G = (V_N, V_T, P, S)$. In this paper, without loss of generality, V_T and S will be the same for all CFG's considered, unless otherwise stated.

The symbols $\underset{G}{\Rightarrow}$, $\overset{*}{\underset{G}{\Rightarrow}}$, $\overset{+}{\underset{G}{\Rightarrow}}$, and $\overset{h}{\underset{G}{\Rightarrow}}$ will represent G-derivations of lengths 1, $\geqslant 1$, $\geqslant 0$, and h respectively.

The symbol G will be omitted when obvious.

Unless otherwise stated, lower case latin letters at the end of the alphabet will denote strings on the terminal alphabet; lower (resp. upper) case latin letters at the beginning of the alphabet will denote elements of V_T (resp. V_N); and lower case greek letters at the beginning of the alphabet will denote strings in V^*, where $V = V_T \cup V_N$.

The empty string will be written λ.

A CFG is an *operator grammar* (OG)[1] if no right part of any production has adjacent nonterminals.

All grammars considered in the sequel (unless otherwise staded) are OG's which in addition are *reduced* and λ-free (Hopcroft, 1969).

A CFG is *invertible* if no two productions have identical right parts.

For a (not necessarily reduced) CFG G its *left terminal set* $\mathscr{L}_G(A)$ and *right terminal set* $\mathscr{R}_G(A)$ are defined by

$$\mathscr{L}_G(A) = \{a \mid A \overset{*}{\underset{G}{\Rightarrow}} Ba\alpha,\ B \in V_N \cup \{\lambda\}\},$$

$$\mathscr{R}_G(A) = \{a \mid A \overset{*}{\underset{G}{\Rightarrow}} \alpha aB,\ B \in V_N \cup \{\lambda\}\}.$$

[1] Greibach (1965) showed that every grammar is equivalent to an OG, and Gray and Harrison (1972) showed that the relation between the two grammars is a complete cover.

These definitions are extended to $\mathscr{P}(V_N)$ and to $\overline{V}^* = V^* - V^*V_N{}^2V^*$ in the following way. Let $\mathscr{C}$ stand for one of $\mathscr{L}$ or $\mathscr{R}$ (the same in each definition). Then for $Z \subseteq V_N$, $\mathscr{C}_G(Z) = \bigcup_{A \in Z} \mathscr{C}_G(A)$; for $\alpha \in V$, if $\alpha = \lambda$ then $\mathscr{C}_G = \varnothing$, else $\mathscr{C}_G(\alpha) = \mathscr{C}_{G'}(D)$, where G' is the same as G except for the addition of the production $D \to \alpha$, $D \notin V_N$. For each $a, b \in V_T$, the *operator precedence relations* $\doteq$; $<.$, $.>$ are defined via:

(1) $a \doteq b$ iff $\exists A \to \alpha a B b \beta \in P$, $B \in V_N \cup \{\lambda\}$,

(2) $a <. b$ iff $\exists A \to \alpha a D \beta \in P$, $b \in \mathscr{L}_G(D)$,

(3) $a .> b$ iff $\exists A \to \alpha D b \beta \in P$, $a \in \mathscr{R}_G(D)$.

For a given G, its *operator precedence matrix* $M = \mathrm{OPM}(G)$ is the $| V_T | \times | V_T |$ matrix which to each ordered pair (a, b) associates the set M_{ab} of operator precedence relations holding between a and b. If $M_{ab} = \phi$, we write $M_{ab} =$ "blank." For M_1, M_2 OPM's we define

$$M_1 \subseteq M_2 \text{ iff } M_{1,ab} \subseteq M_{2,ab} \ \forall a, b;$$
$$M = M_1 \cup M_2 \text{ iff } M_{ab} = M_{1,ab} \cup M_{2,ab} \ \forall a, b;$$
$$M = M_1 \cap M_2 \text{ iff } M_{ab} = M_{1,ab} \cap M_{2,ab} \ \forall a, b.$$

A grammar G is an *operator precedence grammar* (OPG) if OPM (G) is a *conflict-free* matrix, i.e., iff, $\forall a, b$, $| \mathrm{OPM}(G)_{ab} | \leqslant 1$. The language generated by an OPG is called an *operator precedence* language (OPL).

The *parenthesis grammar* $\tilde{G}$ associated with a grammar

$$G = (V_N, V_T, P, S) \text{ is}$$
$$\tilde{G} = (V_N, \tilde{V}_T, \tilde{P}, S), \text{ where } \tilde{V}_T = V_T \cup \{[,]\},$$
$$\tilde{P} = \{A \to [\alpha] \mid A \to \alpha \in P \wedge \alpha \notin V_N\} \cup \{A \to B \in P\},$$

where $[,] \notin V$.

The strings in $L(\tilde{G})$ display via their brackets the syntactic structure of corresponding strings in $L(G)$. The *renaming rules* $(A \to B)$ of P are not bracketed because they do not contribute to the structure of derivation trees in an essential way.

Two grammars G_1 and G_2 are *weakly* (resp. *structurally*) *equivalent* iff

$$L(G_1) = L(G_2) \qquad (\text{resp. } L(\tilde{G}_1) = L(\tilde{G}_2)).$$

An OPG $G = (V_N, V_T, P, S)$ is in *Fischer normal form* (FNF) (Fischer 1969) iff

(1) G is invertible;

(2) S does not occur in the right part of any production of P;

(3) P contains no renaming rules, except those with left part S (if any).

The following statements are either well-known properties or straightforward consequences thereof.

Statement 2.1. (Fischer, 1969). For each OPG G a structurally equivalent OPG G' in FNF can be effectively constructed. ▮

Statement 2.2 (Mc Naughton, 1967). If an OPG G is in FNF, then $A, B \neq S \wedge A \overset{*}{\underset{G}{\Rightarrow}} x \wedge B \overset{*}{\underset{G}{\Rightarrow}} x$ implies $A = B$. ▮

Statement 2.3 (Floyd, 1963). For any pair of OG's $G_1, G_2, L(\tilde{G}_1) \subseteq L(\tilde{G}_2)$ implies $\mathrm{OPM}(G_1) \subseteq \mathrm{OPM}(G_2)$. ▮

Statement 2.4 (Floyd, 1963). For any pair of OPG's G_1, G_2, $\mathrm{OPM}(G_1) = \mathrm{OPM}(G_2)$ implies that $\forall x \in L(G_1) \cap L(G_2)\ \exists! y \in L(\tilde{G}_1)$, and $\exists! z \in L(\tilde{G}_2)$ such that $y = z$ and $h(y) = h(z) = x$, where h: $\tilde{V}_T{}^* \to V_T{}^*$ is the brackets erasing homomorphism. ▮

3. Basic Definitions and Theorems

We give some new definitions and theorems which are useful for proving the Boolean closure properties of OPG's and OPL's. First we introduce a new normal form for OPG's.

Definition 3.1. An OPG G is *homogeneous* iff it is in FNF and, for any $A \to \alpha \in P$, with $A \neq S$, $\mathscr{L}_G(\alpha) = \mathscr{L}_G(A)$, $\mathscr{R}_G(\alpha) = \mathscr{R}_G(A)$. ▮

Theorem 3.2. (*Homogenous Normal Form*). *For any* OPG G, *with* $\mathrm{OPM}(G) = M$, *a structurally equivalent homogenous* OPG H *can be effectively constructed.* ▮

Proof. The following algorithm derives $H = (V_N', V_T, P', S)$ from a FNF $G = (V_N, V_T, P, S)$. The nonterminals of H, other than S, are objects of the form $\langle s, A, t\rangle \in \mathscr{P}(V_T) \times V_N \times \mathscr{P}(V_T)$. Next we construct a sequence of grammars

$$G_k = (N_k, V_T, P_k, S), \qquad k = 1, \ldots, r,$$

leading to the definition of H.

Step 1. $\forall A \to x \in P$, $A \neq S$, $x = aw = yb$, include $\langle\{a\}, A, \{b\}\rangle$ in N_1 and include $\langle\{a\}, A, \{b\}\rangle \to x$ in P_1. For any $S \to x \in P$, include $S \to x$ in P_1.

Step k. $k > 1$. Define $Q = P - \{A \to x \in P\}$ and the finite substitution f_k on $V - \{S\}$:

$$f_k(A) = \{\langle s, A, t\rangle \in N_{k-1}\},$$
$$f_k(a) = \{a\}.$$

Extend f_k to $(V - \{S\})^*$ in the natural way. Then $\forall A \to \alpha \in Q$, $A \neq S$ and $\forall \bar{\alpha} \in f_k(\alpha)$, include $\langle \mathscr{L}_{G_{k-1}}(\bar{\alpha}), A, \mathscr{R}_{G_{k-1}}(\bar{\alpha})\rangle$ in N_k, and include $\langle \mathscr{L}_{G_{k-1}}(\bar{\alpha}), A, \mathscr{R}_{G_{k-1}}(\bar{\alpha})\rangle \to \bar{\alpha}$ in P_k. In addition, $\forall S \to \alpha \in Q$ and $\forall \bar{\alpha} \in f_k(\alpha)$, include $S \to \bar{\alpha}$ in P_k. Stop at the first step, say the r-th, for which $N_r = N_{r-1}$. Then $H = G_r$. ∎

The above procedure terminates since $\forall k$, $|N_k| \leqslant |V_n| \cdot |\mathscr{P}(V_T)|$[2]. The following lemma helps to show that H has the desired properties.

LEMMA 3.3. $\forall A \in V_N - \{S\}$, $h \geqslant 1$,

$$A \overset{h}{\underset{G}{\Rightarrow}} x \qquad \textit{iff} \quad \exists s, t \in \mathscr{P}(V_T)$$

such that $\langle s, A, t\rangle \overset{h}{\underset{H}{\Rightarrow}} x$. ∎

Proof. (by induction on the length h of a derivation).

Basis: for $h = 1$, immediate from Step 1 of the construction.

Induction step: assume the lemma holds for any derivation of length $\leqslant h$. Let $A \overset{h+1}{\underset{G}{\Rightarrow}} x$ via $A \overset{1}{\underset{G}{\Rightarrow}} [\alpha] = [x_0 B_1 \cdots B_m x_m]$, $m \geqslant 0$, $B_i \in V_N - \{S\}$,[2] and $[\alpha] \overset{h}{\underset{G}{\Rightarrow}} x$, via $B_i \overset{h_i}{\underset{G}{\Rightarrow}} y_i$ such that $x_0 y_1 \cdots y_m x_m = x$, $h_i \leqslant h$.

Then by hypothesis $\forall_i \exists (s_i, t_i)$ such that $\langle s_i, B_i, t_i\rangle \overset{h_i}{\underset{G}{\Rightarrow}} y_i$. Now by construction $\exists k$ such that $\forall i \langle s_i, B_i, t_i\rangle \in f_{k+1}(B_i)$, since each $\langle s_i, B_i, t_i\rangle$ has been defined in N_k.

Consequently $\bar{\alpha} = x_0\langle s_1, B_1, t_1\rangle \cdots \langle s_m, B_m, t_m\rangle x_m$ is in $f_{k+1}(\alpha)$ and $\langle \mathscr{L}_{G_k}(\bar{\alpha}), A, \mathscr{R}_{G_k}(\bar{\alpha})\rangle \to \bar{\alpha}$ is in $P_{k+1} \subseteq P'$ and therefore $\langle \mathscr{L}_{G_k}(\bar{\alpha}), A, \mathscr{R}_{G_k}(\bar{\alpha})\rangle \overset{h+1}{\underset{H}{\Rightarrow}} x$.

Conversely, if $\langle s, A, t\rangle \overset{h+1}{\underset{H}{\Rightarrow}} x$ via $\langle s, A, t\rangle \overset{1}{\underset{H}{\Rightarrow}} [\bar{\alpha}] = [x_0\langle s_1, B_1, t_1\rangle \cdots \langle s_m, B_m, t_m\rangle x_m]$ and $\langle s_i, B_i, t_i\rangle \overset{h_i}{\underset{H}{\Rightarrow}} y_i$, then since $h_i \leqslant h$, $B_i \overset{h_i}{\underset{G}{\Rightarrow}} y_i$ and $A \to \alpha = x_0 B_1 \cdots B_m x_m$ is in P. Thus $A \overset{h+1}{\underset{G}{\Rightarrow}} x$. ∎

In order to complete the proof of the theorem, we must now prove the following points.

(a) H is an OG (obvious because G is an OG);

(b) H is reduced. The fact that $\forall \langle s, A, t\rangle \in V_N' - \{S\}$, $\langle s, A, t\rangle \overset{*}{\underset{H}{\Rightarrow}} x$ for some x, can be proved by induction on the sets N_k: in fact this is trivially true for $k = 1$, and the k-th step of the construction preserves this property.

On the other hand, it is easy to verify that if $S \overset{*}{\underset{G}{\Rightarrow}} uAv$, then

$$S \overset{*}{\underset{H}{\Rightarrow}} u\langle s, A, t\rangle v, \forall s, t$$

[2] This is certainly the case since G is in FNF.

such that $\langle s, A, t\rangle \in f_r(A)$. (Notice, however, that in general G_k, $k < r - 1$, is not reduced).

(c) H is structurally equivalent to G.

In fact if $S \overset{*}{\underset{G}{\Rightarrow}} x$, then either

(i) $S \overset{1}{\underset{G}{\Rightarrow}} B \overset{*}{\underset{G}{\Rightarrow}} x$ or

(ii) $S \overset{1}{\underset{G}{\Rightarrow}} [\alpha] \overset{*}{\underset{G}{\Rightarrow}} x$.

By Lemma 3.3, since G is in FNF, (i) implies that for some s, t, $S \overset{1}{\underset{H}{\Rightarrow}} \langle s, B, t\rangle$ and $\langle s, B, t\rangle \overset{*}{\underset{H}{\Rightarrow}} x$, whereas (ii) implies that $\alpha = x_0 B_1 \cdots B_m x_m$, $B_i \in V_N - \{S\}$ and $B_i \overset{*}{\underset{G}{\Rightarrow}} y_i$ such that $[x_0 y_1 \cdots y_m x_m] = x$.

Thus there exist s_i, t_i such that $\langle s_i, B_i, t_i\rangle \overset{*}{\underset{H}{\Rightarrow}} y_i$ and

$$S \to \bar{\alpha} = x_0\langle s_1, B_1, t_1\rangle \cdots \langle s_m, B_m, t_m\rangle x_m \text{ is in } P'.$$

Hence (c) is proved.

(d) H is an OPG with $\mathrm{OPM}(H) = M = \mathrm{OPM}(G)$ (by Statement 2.3);

(e) H is in FNF because G is in FNF and the construction preserves both the invertibility and the nonappearance of S in the right part of any production. Finally

(f) H is homogeneous. We show that $\forall\langle s, A, t\rangle \to \alpha$,

$$\mathscr{L}_H(\langle s, A, t\rangle) = s \text{ (the proof that}$$
$$\mathscr{R}_H(\langle s, A, t\rangle) = t \text{ is omitted).}$$

Clearly by the definition of left terminal set and by construction of H, we have $s \subseteq \mathscr{L}_H(\langle s, A, t\rangle)$. On the other hand, $a \in \mathscr{L}_H(\langle s, A, t\rangle)$ implies the existence of the derivation

$$\langle s, A, t\rangle \overset{1}{\underset{H}{\Rightarrow}} \langle s_1, A_1, t_1\rangle \beta_1 \overset{1}{\underset{H}{\Rightarrow}} \cdots \overset{1}{\underset{H}{\Rightarrow}} \langle s_n, A_n, t_n\rangle \beta_n$$

$$\overset{1}{\underset{H}{\Rightarrow}} Ea\gamma_n\beta_n,$$

where $E \in V_n \cup \{\lambda\}$ and $\langle s_n, A_n, t_n\rangle \to Ea\gamma_n \in P_k$ for some $k \leqslant r$. Hence $a \in s_n$, and similarly $a \in s_{n-1}, \ldots, a \in s_1$, $a \in s$. Thus $\mathscr{L}_H(\langle s, A, t\rangle) \subseteq s$. It follows immediately that $\mathscr{L}_H(\alpha) = s$. ∎

We now define some new concepts which are basic to the rest of the paper.

Definition 3.4. The class C_M of OPG's over a conflict-free precedence matrix M is defined as

$$C_M = \{G \mid \mathrm{OPM}(G) = M' \subseteq M\}. \text{ ∎}$$

DEFINITION 3.5. The class $C_{M,q}$ of OPG's with right bound $q \geqslant 1$ over a conflict-free precedence matrix M is defined as

$$C_{M,q} = \{G \mid G \in C_M \wedge (\forall A \to \alpha \in P, |f(\alpha)| \leqslant q,$$

where f: $(V_N \cup V_T)^* \to V_T$ is the homomorphism which erases nonterminals)}. ∎

Definition 3.5 excludes grammars containing productions of unbounded length, a quite reasonable assumption for practical purposes.

We mention at this time that the results reported in the sequel were originally (Crespi-Reghizzi, 1970) developed under the different and stronger hypothesis that M be $\doteq$ —acyclic, i.e., the transitive closure of $\doteq$ irreflexive.

DEFINITION 3.6. An OPG G is *free* if it is homogeneous and, $\forall A, B \neq S$, $\mathscr{L}_G(A) = \mathscr{L}_G(B) \wedge \mathscr{R}_G(A) = \mathscr{R}_G(B)$ implies $A = B$. ∎

Next we present a central result.

THEOREM 3.7. *For any conflict-free* OPM M *on* V_T, *and* $q \geqslant 1$, *a free* OPG $G_{M,q}$ *can be effectively constructed such that*

(a) $\mathrm{OMP}(G_{M,q}) = M$;

(b) $\forall G \in C_{M,q}, L(G) \subseteq L(G_{M,q})$. ∎

$G_{M,q}$ is called the *maxgrammar* and $L_{M,q} = L(G_{M,q})$ the *maxlanguage* associated with M and q.

Proof. We begin by constructing $G_{M,q} = \{\hat{V}_N, V_T, \hat{P}, S\}$.

The elements of $\hat{V}_N$ other than S will be objects of the form $\langle s, t\rangle \in (\mathscr{P}(V_T))^2$.

We now construct a sequence of grammars

$$G_k = (N_k, V_T, P_k, S), \qquad k = 1,\ldots, r,$$

leading to the definition of $G_{M,q}$.

Step 1. $\forall x = a_1 \cdots a_m \in (V_T)^m$, $1 \leqslant m \leqslant q$, such that, for $1 \leqslant i \leqslant m-1$, $a_i \doteq a_{i+1}$ in M, include $\langle\{a_1\}, \{a_m\}\rangle$ in N_1 and include $\langle\{a_1\}, \{a_m\}\rangle \to x$ in P_1.

Step k. $k > 1$. $\forall \alpha = D_0 a_1 \cdots D_{m-1} a_m D_m$, $1 \leqslant m \leqslant q$, such that

(i) if $m > 1$, $a_i \doteq a_{i+1}$ in M, $1 \leqslant i \leqslant m-1$;

(ii) D_0 is λ or an element $\langle s, t\rangle \in N_{k-1}$ such that, $\forall b \in t$, $b .> a_1$ in M;

(iii) D_m is λ or an element $\langle s, t\rangle \in N_{k-1}$ such that, $\forall b \in s$, $a_m <. b$ in M;

(iv) if $m > 1$, D_i, for $1 \leqslant i \leqslant m-1$, is λ or an element $\langle s, t\rangle \in N_{k-1}$ such that, $\forall b \in s$, $a_i <. b$ in M, $\forall c \in t$, $c .> a_{i+1}$ in M;

include $\langle \mathscr{L}_{G_{k-1}}(\alpha), \mathscr{R}_{G_{k-1}}(\alpha)\rangle$ in N_k and include

$$\langle \mathscr{L}_{G_{k-1}}(\alpha), \mathscr{R}_{G_{k-1}}(\alpha)\rangle \to \alpha \qquad \text{in} \quad P_k .$$

Stop at the first step, say the r-th, such that $N_r = N_{r-1}$. ∎

This procedure always terminates since $\forall k$, $|N_k| \leqslant |\mathscr{P}(V_T)|^2$.
Then define $\hat{V}_N = N_r \cup \{S\}$,

$$\hat{P} = P_r \cup \{S \to \langle s, t\rangle \mid \langle s, t\rangle \in N_r\}.$$

Consider now $G = (V_N, V_T, P, S) \in C_{M,q}$ with $\mathrm{OPM}(G) = M' \subseteq M$: by Theorem 3.2 we can assume that G is in HNF and that the elements of $V_N - \{S\}$ are of the form $\langle s, A, t\rangle$ with $\mathscr{L}_G(\langle s, A, t\rangle) = s$, $\mathscr{R}_G(\langle s, A, t\rangle) = t$. We first prove a preliminary lemma.

LEMMA 3.8. *$\forall \langle s, A, t\rangle \in V_N - \{S\}$, if $\langle s, A, t\rangle \overset{*}{\underset{G}{\Rightarrow}} x$, then*

$$\langle s, t\rangle \overset{*}{\underset{G_{M,q}}{\Rightarrow}} x. \quad ∎$$

Proof (by induction on the length h of a derivation).

Basis: let $\langle s, A, t\rangle \overset{1}{\underset{G}{\Rightarrow}} x = [a_1 \cdots a_m]$, $q \geqslant m \geqslant 1$. Clearly $s = \{a_1\}$ and $t = \{a_m\}$. Then the relations $a_i \doteq a_{i+1}$, $i = 1,..., m-1$ (of course, if $m > 1$) are in M', hence in M. Therefore $\langle\{a_1\}, \{a_m\}\rangle \to a_1 \cdots a_m$ is in $G_{M,q}$ and $\langle s, t\rangle \overset{1}{\underset{G_{M,q}}{\Rightarrow}} x$.

Induction step: assume that if $\langle s, A, t\rangle \overset{k}{\underset{G}{\Rightarrow}} x$ and $k \leqslant h$, then $\langle s, t\rangle \overset{k}{\underset{G_{M,q}}{\Rightarrow}} x$, and consider a derivation $\langle s, A, t\rangle \overset{1}{\underset{G}{\Rightarrow}} [\alpha]$, $\alpha = D_0 a_1 \cdots D_{m-1} a_m D_m$, $D_i \in \{\lambda\} \cup (V_N - \{S\})$, $q \geqslant m \geqslant 1$, and a derivation $[\alpha] \overset{h}{\underset{G}{\Rightarrow}} x$ via $D_i \overset{h_i}{\underset{G}{\Rightarrow}} y_i$, $i = 0,..., m$. Thus, $\forall i$ such that $D_i = \langle s_i, B_i, t_i\rangle \in V_N - \{S\}$, $\langle s_i, t_i\rangle \overset{h_i}{\underset{G_{M,q}}{\Rightarrow}} y_i$, since $h_i < h$.

Now the relations $a_i \doteq a_{i+1}$, $i = 1,..., m-1$ (of course, if $m > 1$) are in M' hence in M.

Furthermore, $\forall i$, $1 \leqslant i \leqslant m$, such that $D_i = \langle s_i, B_i, t_i\rangle$, $\forall b \in s_i$, $a_i <\!. b$ is in M', and $\forall c \in t_{i-1}$, $c .\!> a_i$ is in M'. Thus, since $\langle s_i, t_i\rangle$ is in N_k for some k, the rule $\langle \mathscr{L}_{G_k}(\bar{\alpha}), \mathscr{R}_{G_k}(\bar{\alpha})\rangle \to \bar{\alpha}$, $\bar{\alpha} = E_0 a_1 \cdots E_{m-1} a_m E_m$ with $E_i = \lambda$ if $D_i = \lambda$, else $E_i = \langle s_i, t_i\rangle$ if $D_i = \langle s_i, B_i, t_i\rangle$, is in $P_{k+1} \subseteq \hat{P}$.

But $\mathscr{L}_{G_k}(\bar{\alpha}) = \mathscr{L}_G(\alpha) = s$ and $\mathscr{R}_{G_k}(\bar{\alpha}) = \mathscr{R}_G(\alpha) = t$ since G is homogenous.

Therefore, $\langle s, t\rangle \overset{1}{\underset{G_{M,q}}{\Rightarrow}} [\alpha] \overset{h}{\underset{G_{M,q}}{\Rightarrow}} x$. ∎

In order to complete the proof of Theorem 3.7 it is sufficient to show that:

(a) $\forall \mathrm{OPG}$ $G \in C_{M,q}$, $L(G) \subseteq L_{M,q}$ (which is certainly true if $L(\tilde{G}) \subseteq L(\tilde{G}_{M,q})$).

Consider the two forms that a derivation in $\tilde{G}$ can take:

(i) $S \overset{1}{\underset{\tilde{G}}{\Rightarrow}} \langle s, A, t\rangle \overset{*}{\underset{\tilde{G}}{\Rightarrow}} x$,

(ii) $S \overset{1}{\underset{\tilde{G}}{\Rightarrow}} [\alpha] \overset{*}{\underset{\tilde{G}}{\Rightarrow}} x$, $\alpha = D_0 a_1 \cdots D_{m-1} a_m D_m$

$$q \geqslant m \geqslant 1,\ D_i \in \{\lambda\} \cup V_N, \qquad i = 0,\ldots, m.$$

Case (i) is trivial: $S \to \langle s, t\rangle \in \hat{P}$ and by Lemma 3.8 $\langle s, t\rangle \overset{*}{\underset{\tilde{G}_{M,q}}{\Rightarrow}} x$.

In case (ii) a proof similar to the one of Lemma 3.8 would show that

$$\langle \mathscr{L}_{G_{M,q}}(\bar{\alpha}), \mathscr{R}_{G_{M,q}}(\bar{\alpha})\rangle \to \bar{\alpha} \in \hat{P},$$

where $\bar{\alpha}$ is the same as in the induction step of Lemma 3.8. Furthermore, $S \to \langle \mathscr{L}_{G_{M,q}}(\bar{\alpha}), \mathscr{R}_{G_{M,q}}(\bar{\alpha})\rangle \in \hat{P}$ and point (a) is proved.

(b) $G_{M,q}$ is reduced (easily verified by induction on N_k);

(c) $\forall \langle s, t\rangle \to \alpha \in \hat{P}$, $\mathscr{L}_{G_{M,q}}(\langle s, t\rangle) = s = \mathscr{L}_{G_{M,q}}(\alpha)$ and $\mathscr{R}_{G_{M,q}}(\langle s, t\rangle) = t = \mathscr{R}_{G_{M,q}}(\alpha)$.

The proof, very similar to the one of Theorem 3.2 (point (f)), is omitted. Thus $G_{M,q}$ is free.

(d) $M \supseteq \mathrm{OPM}(G_{M,q})$.

Any relation $a \doteq b$ in $\mathrm{OPM}(G_{M,q})$ is implied by a production $B \to \alpha a D b \beta \in \hat{P}$, $\alpha, \beta \in \hat{V}^*$, $D \in \{\lambda\} \cup \hat{V}_N$ which is in $\hat{P}$ only if $a \doteq b$ is in M. In addition the string $\alpha a\langle s, t\rangle \beta$ (respectively, $\alpha\langle s, t\rangle a\beta$), $\alpha, \beta \in \hat{V}^*$, is the right part of some production of $G_{M,q}$ only if, $\forall b \in s$ (respectively, $\forall b \in t$) $a <. b$ (respectively, $b .> a$) is in M. Thus $M \supseteq \mathrm{OPM}(G_{M,q})$.

In addition M can be viewed as the union of some matrices with exactly one nonblank entry.[3] Let M' one such matrix. Then, by Statement 2.3, we have $M' \subseteq \mathrm{OPM}(G_{M,q})$ and

(e) $M \subseteq \mathrm{OPM}(G_{M,q})$,

and therefore $M = \mathrm{OPM}(G_{M,q})$. Hence $G_{M,q}$ is an OPG with $\mathrm{OPM}(G_{M,q}) = M$, and it is free by virtue of (c). The theorem is now completely proved. ∎

4. The Lattice of Free Grammars

In this section some important ordering properties of free grammars are demonstrated.

[3] It is immediate to construct an OPG G such that $\mathrm{OPM}(G)$ is such a matrix.

Let M be a (conflict-free) OPM on alphabet V_T. Consider the family of free grammars

$$\mathscr{F}(M, q) = \{F = (V_N{}^F, V_T, P^F, S) \mid F \text{ is free},$$
$$(V_N{}^F - \{S\}) \subseteq \mathscr{P}(V_T)^2\} \cap C_{M,q}.$$

We shall say that nonterminals of F other than S have *standard names*, i.e., they are objects of the form $\langle s, t\rangle$, $s, t \in \mathscr{P}(V_T)$. The requirement that nonterminal names be standard prevents the inclusion in $\mathscr{F}(M, q)$ of any grammar which is isomorphic to F. Let $\mathscr{L}(M, q) = \{L(F) \mid F \in \mathscr{F}(M, q)\}$.

We first prove a useful lemma. Since the bound q can be considered fixed, in the remainder of the section, we drop q from the notation, still keeping in mind that this bound is implicitly in force.

LEMMA 4.1. *Let* $F_1, F_2 \in \mathscr{F}(M)$. *Then* $L(F_1) \subseteq L(F_2)$ *iff* $P_1 \subseteq P_2$ (*and therefore* $L(F_1) = L(F_2)$ *iff* $P_1 = P_2$). ∎

Proof. Obviously $P_1 \subseteq P_2$ implies $L(F_1) \subseteq L(F_2)$.

Conversely assume $L(F_1) \subseteq L(F_2)$. Then since $F_1, F_2 \in \mathscr{F}(M) \subset C_M$, $L(\tilde{F}_1) \subseteq L(\tilde{F}_2)$ by point (a) of the proof of Theorem 3.7. Thus it can be easily proved by induction in a manner similar to the proofs of Lemmas 3.3 and 3.8 that if $S \overset{*}{\underset{\tilde{F}_1}{\Rightarrow}} x\langle s, t\rangle z \overset{1}{\underset{\tilde{F}_1}{\Rightarrow}} x[\alpha]z \overset{*}{\underset{\tilde{F}_1}{\Rightarrow}} x[y]z$, then $\langle s, t\rangle \to \alpha$ is also in F_2 and $S \overset{*}{\underset{\tilde{F}_2}{\Rightarrow}} x\langle s, t\rangle z \overset{1}{\underset{\tilde{F}_2}{\Rightarrow}} x[\alpha]z \overset{*}{\underset{\tilde{F}_2}{\Rightarrow}} x[y]z$. ∎

DEFINITION 4.2. Let $F_1, F_2 \in \mathscr{F}(M)$. The notation $F_1 \leqslant F_2$ denotes $P^{F_1} \subseteq P^{F_2}$, and $L(F_1) \leqslant L(F_2)$ denotes $L(F_1) \subseteq L(F_2)$. ∎

Note that the systems $(\mathscr{F}(M), \leqslant)$ and $(\mathscr{L}(M), \leqslant)$ are posets.

DEFINITION 4.3. The binary operations *join* $(+)$ and *meet* $(*)$ are defined on $\mathscr{F}(M)$ by

(a) $F_1 + F_2 = (V_N^{F_1} \cup V_N^{F_2}, V_T, P^{F_1} \cup P^{F_2}, S)$,

(b) $F_1 * F_2 = R((V_N^{F_1} \cap V_N^{F_2}, V_T, P^{F_1} \cap P^{F_2}, S))$ where $R(G)$ denotes the reduced form of G. ∎

The *empty grammar*[4] $(\varnothing, V_T, \varnothing, \varnothing)$ is denoted F_0, where $L(F_0) = \varnothing$.

The reader can easily verify that $F_1 + F_2, F_1 * F_2 \in \mathscr{F}(M)$.

LEMMA 4.4. *For* $F_1, F_2 \in \mathscr{F}(M)$

(a) $L(F_1 + F_2) \supseteq L(F_1) \cup L(F_2)$

(b) $L(F_1 * F_2) = L(F_1) \cap L(F_2)$. ∎

[4] The use of the empty set $\varnothing$ to denote the axiom is not consistent with our notation, but is consistent with equivalent definitions of a grammar using a finite set $S \subseteq V_N$ of axioms. This special notation was adopted to make F_0 be reduced.

Proof. (a) immediately follows from Lemma 4.1.

In order to prove (b), notice that trivially $L(F_1 * F_2) \subseteq L(F_1) \cap L(F_2)$ and if $x \in L(F_1) \cap L(F_2)$ then the derivations $S \overset{*}{\underset{F_1}{\Rightarrow}} x$ and $S \overset{*}{\underset{F_2}{\Rightarrow}} x$ are identical.

Therefore the productions involved are in $P^{F_1} \cap P^{F_2}$, hence in the production set of $F_1 * F_2$. ∎

DEFINITION 4.5. The binary operations join $(+)$ and meet $(*)$ are extended to $\mathscr{L}(M)$ via

(a) $L(F_1) + L(F_2) = L(F_1 + F_2)$,

(b) $L(F_1) * L(F_2) = L(F_1 * F_2)$. ∎

From Lemma 4.4, $L(F_1) + L(F_2) \supseteq L(F_1) \cup L(F_2)$ and $L(F_1) * L(F_2) = L(F_1) \cap L(F_2)$.

We can now state the following central result:

THEOREM 4.6. *The systems* $(\mathscr{F}(M), \leqslant, +, *)$ *and* $(\mathscr{L}(M), \leqslant, +, *)$ *are isomorphic lattices.* ∎

Proof. First we establish that the second system is a lattice. In fact, if $L_1 = L(F_1)$ and $L_2 = L(F_2)$ are in $\mathscr{L}(M)$, $L_1 * L_2 = L(F_1) \cap L(F_2)$ is clearly the g.l.b. of L_1 and L_2. Moreover, obviously $L_1 \leqslant L_1 + L_2$, $L_2 \leqslant L_1 + L_2$. Suppose now that $\exists L' \in \mathscr{L}(M)$, $L' = L(F')$, such that $L_1 \leqslant L' \leqslant L_1 + L_2$ and $L_2 \leqslant L' \leqslant L_1 + L_2$. Then by Lemma 4.1 $F_1 \leqslant F'$, $F_2 \leqslant F'$, and since any $x \in L_1 + L_2$ is generated using productions in $P^{F_1} \cup P^{F_2}$, $x \in L'$. Hence $L' = L_1 + L_2$, i.e., $L_1 + L_2$ is the l.u.b. of L_1, L_2.

Finally the correspondence $f: \mathscr{F}(M) \to \mathscr{L}(M)$ defined by $f(F) = L(F)$ is a one-to-one mapping (because of the standard form of grammars in $\mathscr{F}(M)$) which preserves the relation $\leqslant$ (Lemma 4.1). From Definition 4.5 it follows that the two systems are isomorphic lattices.

The null elements of the two systems are, respectively, F_0 and $\varnothing$, while the universal elements are, respectively, G_M (the maxgrammar of M) and $L_M = L(G_M)$. ∎

COROLLARY 4.7. *Let* $F \in \mathscr{F}(M)$ *and define the subsets*

$$\mathscr{F}(F) = \{G \in \mathscr{F}(M) \mid G \leqslant F\},$$
$$\mathscr{L}(F) = \{L(G) \mid G \in \mathscr{F}(F)\}.$$

Then the systems $(\mathscr{L}(F), \leqslant, +, *)$, $(\mathscr{F}(F), \leqslant, +, *)$ *are isomorphic sublattices of* $\mathscr{L}(M)$ *and* $\mathscr{F}(M)$. ∎

Similar properties can also be proved for the maxgrammars. Let M be an OPM, and define the sets

$$\mathscr{M}(M) = \{M' \mid M' \subseteq M\},$$

$\mathscr{G}_{\max}(M)$[5] $= \{G_{M'} \mid G_{M'}$ is the maxgrammar with OPM $M' \subseteq M\}$,

$$\mathscr{L}_{\max}(M) = \{L(F) \mid F \in \mathscr{G}_{\max}(M)\}.$$

Notice that $\mathscr{G}_{\max}(M) \subseteq \mathscr{F}(M)$ (maxgrammars are free and have standard nonterminal names).

Let G_{M_1} and $G_{M_2} \in \mathscr{G}_{\max}(M)$ be the maxgrammars with OPM's M_1 and M_2, respectively, and let $L_1 = L(G_{M_1}), L_2 = L(G_{M_2})$. Define the following operations and relations:

(1a) $M_1 \oplus M_2$ denotes $M_1 \cup M_2$,

(1b) $M_1 \circledast M_2$ denotes $M_1 \cap M_2$,

(1c) $M_1 \leqslant M_2$ denotes $M_1 \subseteq M_2$,

(2a) $G_{M_1} \oplus G_{M_2}$ denotes $G_{M_1 \oplus M_2}$ (the maxgrammar with OPM $M_1 \oplus M_2$),

(2b) $G_{M_1} \circledast G_{M_2}$ denotes $R(V^{G_{M_1}} \cap V^{G_{M_2}}, V_T, P^{G_{M_1}} \cap P^{G_{M_2}}, S)$,

(2c) $G_{M_1} \leqslant G_{M_2}$ denotes $P^{G_1} \subseteq P^{G_2}$,

(3a) $L_1 \oplus L_2$ denotes $L(G_{M_1} \oplus G_{M_2})$,

(3b) $L_1 \circledast L_2$ denotes $L_1 \cap L_2$,

(3c) $L_1 \leqslant L_2$ denotes $L_1 \subseteq L_2$.

LEMMA 4.8. *$G_{M_1} \circledast G_{M_2}$ is the maxgrammar with* OPM $M_1 \circledast M_2$. ∎

Proof. Any precedence relation in $M_1 \cap M_2 = M_1 \circledast M_2$ is implied by both $P^{G_{M_1}}$ and $P^{G_{M_2}}$ and conversely. Furthermore, if $S \overset{*}{\underset{G}{\Rightarrow}} x$, for some G such that OPM$(G) \leqslant M_1 \cap M_2$, then $S \overset{*}{\underset{G_{M_1}}{\Rightarrow}} x$ and $S \overset{*}{\underset{G_{M_2}}{\Rightarrow}} x$, and the three derivations above are identical as already observed (Lemmas 4.1 and 4.4). Hence

$$S \overset{*}{\underset{G_{M_1} \circledast G_{M_2}}{\Rightarrow}} x. \quad ∎$$

Finally we have

COROLLARY 4.9. *The systems* $(\mathscr{M}(M), \leqslant, \oplus, \circledast)$, $(\mathscr{G}_{\max}(M), \leqslant, \oplus, \circledast)$, $(\mathscr{L}_{\max}(M), \leqslant, \oplus, \circledast)$ *are isomorphic lattices.* ∎

Proof. The first system is clearly the lattice of submatrices of a finite OPM M, with universal element M and null element the empty OPM.

Furthermore $M_1 \leqslant M_2$ iff $G_{M_1} \leqslant G_{M_2}$ (by construction of maxgrammars), $G_{M_1} \leqslant G_{M_2}$ iff $L_1 \leqslant L_2$ (Lemma 4.1) and the definitions of $\oplus$ and $\circledast$ preserve the natural one-to-one correspondence between the elements of $\mathscr{M}(M)$, $\mathscr{G}_{\max}(M)$ and $\mathscr{L}_{\max}(M)$. Therefore the three systems are isomorphic lattices. The null

[5] Recall that the bound q is omitted since it is constant throughout this section.

and universal elements of the two latter lattices are, respectively, $(\{S\}, V_T, \{S \to a \mid a \in V_T\}, S)$ and G_M (the maxgrammar with OPM M), V_T and $L(G_M)$. ∎

In Appendix I a complete example lattice is presented.

5. Closure Properties of OPL's

We have demonstrated that there are finitely many free grammars over a given OPM and with an upper bound q on the number of terminal instances occurring in any production. Admittedly free grammars are only a subfamily, though an interesting one (Crespi-Reghizzi, 1973), of OPG's. We shall next discuss the relationship between OPG's and free grammars and deduce interesting algebraic properties of OPL's. Each grammar will be assumed to be in homogeneous normal form with the same terminal alphabet and nonterminal names (excepted S) of the form $\langle s, A, t\rangle$, as in Theorem 3.2. We first define a many-to-one mapping W that maps homogeneous OPG's onto free grammars.[6]

Definition 5.1. Let $\mathscr{H}$ the class of homogeneous OPG's and let $H = (V_N, V_T, P, S) \in \mathscr{H}$ where $V_N \subseteq \{S\} \cup \mathscr{P}(V_T) \times Z \times \mathscr{P}(V_T)$, Z is any set of abstract symbols (denoted as upper case Latin letters), and

$$\mathscr{L}_H(\langle s, A, t\rangle) = s, \qquad \mathscr{R}_H(\langle s, A, t\rangle) = t.$$

$F = W(H)$ is defined as follows:

$F = (V_N', V_T, P', S)$, where

$$V_N' = \{\langle s, t\rangle \mid \langle s, A, t\rangle \in V_N\} \cup \{S\},$$

$$\begin{aligned} P' = \{&\langle s, t\rangle \to x_0\langle s_1, t_1\rangle \cdots \langle s_n, t_n\rangle x_n \mid n \geqslant 0, \\ &\langle s, A, t\rangle \to x_0\langle s_1, B_1, t_1\rangle \cdots \langle s_n, B_n, t_n\rangle x_n \in P\} \\ &\cup \{S \to \langle s, t\rangle \mid S \to \langle s, A, t\rangle \in P\}. \end{aligned}$$ ∎

Intuitively, $F = W(H)$ is obtained by *merging* those nonterminals of H which have identical first and third components (left terminal set, right terminal set). Notice that the inverse $W^{-1}(F)$ is an infinite set since all possible sets Z of abstract symbols must be considered.

Definition 5.2. Define, for bound q,

$$\mathscr{F}_q = \bigcup_M \mathscr{F}(M, q),$$

where the union is taken over the set of all OPM's on V_T;

$$\mathscr{F} = \bigcup_{q=1}^{\infty} \mathscr{F}_q. \quad ∎$$

The following lemmas can be readily verified by the reader.

[6] W is a special case of grammatical covering defined by Reynolds (1968).

LEMMA 5.3. *$\forall H \in \mathscr{H}$, $F = W(H)$ is a free grammar (with standard nonterminal names) such that $\mathscr{L}_F(\langle s, t\rangle) = s$, $\mathscr{R}_F(\langle s, t\rangle) = t$ and* $\mathrm{OPM}(F) = \mathrm{OPM}(H)$. ∎

The proof of the first assertion is similar to the proof of point (f) of Theorem 3.2. The equality of the OPM's can then be shown in a manner similar to that of points (d) and (e) of the proof of Theorem 3.7.

LEMMA 5.4. *$L(\tilde{H}) \subseteq L(\tilde{F})$, $H \in \mathscr{H}$, $F \in \mathscr{F}$, implies $W(H) \leqslant F$.* ∎

Proof. By induction on the length h of an $\tilde{H}$-derivation (and recalling Statement 2.2), it can be proved that if $\langle s, A, t\rangle \overset{1}{\underset{\tilde{H}}{\Rightarrow}} [x_0\langle s_1, B_1, t_1\rangle \cdots \langle s_n, B_n, t_n\rangle x_n] \overset{h}{\underset{\tilde{H}}{\Rightarrow}} x$, then, since $x \in L(\tilde{F})$, $\langle s, t\rangle \overset{1}{\underset{\tilde{F}}{\Rightarrow}} [x_0\langle s_1, t_1\rangle \cdots \langle s_n, t_n\rangle x_n] \overset{h}{\underset{\tilde{F}}{\Rightarrow}} x$. ∎

Therefore it follows that

$$W(\mathscr{H}) = \{w(H) \mid H \in \mathscr{H}\} = \mathscr{F}.$$

We can now state the following result.

Statement 5.5. *$\mathscr{H}$ is partitioned by W into mutually disjoint classes.*[7] *For each class $W^{-1}(F)$, $F \in \mathscr{F}$, $L(H) \leqslant L(F)$, $\forall H \in W^{-1}(F)$.* ∎

Assume now, similar to Corollary 4.7, $\mathscr{F}(F) = \{G \in \mathscr{F} \mid G \leqslant F\}$, and define the sets

$$\mathscr{H}(F) = \{H \in \mathscr{H} \mid W(H) \in \mathscr{F}(F)\} \subseteq \mathscr{H}.$$[8]

Our first closure result is then:

THEOREM 5.6. *The set of languages $\{L(H) \mid H \in \mathscr{H}(F)\}$ is closed under union.* ∎

Proof. Let $H_i = (V_i, V_T, P_i, S) \in W^{-1}(F_i)$, $F_i \in \mathscr{F}(F)$, $i = 1, 2$. Assume, without loss of generality, $V_1 \cap V_2 = \{S\}$ (the middle element of the triples $\langle s, A, t\rangle$ can be changed as needed). Let $G' = (V_1 \cup V_2, V_T, P_1 \cup P_2, S)$; clearly $L(G') = L(H_1) \cup L(H_2)$. Now since $\mathrm{OPM}(H_i) = M_i \leqslant \mathrm{OPM}(F)$, $L(\tilde{H}_1) \cup L(\tilde{H}_2) \subseteq L(\tilde{F}_1) \cup L(\tilde{F}_2) \subseteq \widetilde{L(F_1 + F_2)} \subseteq L(\tilde{F})$ and $\mathrm{OPM}(G') \leqslant \mathrm{OPM}(F)$. Now construct, by applying Statement 2.1 and Theorem 3.2, a homogeneous grammar H such that $L(\tilde{H}) = L(\tilde{G}')$. Since $L(\tilde{H}) \subseteq L(\tilde{F})$, the theorem follows from Lemma 5.4. ∎

With a little additional effort the following result could be proved.

[7] If we consider the subset of $\mathscr{H}$ characterized by a fixed bound q, the number of equivalence classes is finite.

[8] For any finite number of OPG's, $G_1, \ldots, G_n$, s.t. $\bigcup_{i=1}^{n} \mathrm{OPM}(G_i)$ is conflict-free, a suitable $F \in \mathscr{F}$ can be found s.t. $L(G_i) \in \{L(H) \mid H \in \mathscr{H}(F)\}$.

COROLLARY 5.7. *The set of languages* $\{L(H) \mid H \in W^{-1}(F)\}$ *is closed under union.* ∎

The second closure property is given by

THEOREM 5.8. *The set of languages* $\{L(H) \mid H \in \mathscr{H}(F)\}$ *is closed under complementation with respect to* $L(F)$. ∎

Proof. Let $H \in \mathscr{H}(F)$, then $L(H) \subseteq L(F)$. Since F and H are unambiguous, $\bar{L} = L(F) - L(H) = h(L(\tilde{F}) - L(\tilde{H}))$, where h is the homomorphism which erases "["and"]."

It is well known (Mac Naughton, 1967) that a grammar G can be effectively constructed such that $L(\tilde{G}) = L(\tilde{F}) - L(\tilde{H})$.[9] Obviously G must be an OG and since $L(\tilde{G}) \subseteq L(\tilde{F})$, $\mathrm{OPM}(G) \subseteq \mathrm{OPM}(F)$; thus G is an OPG, from which a homogeneous OPG H' can be constructed such that $L(\tilde{H}') = L(\tilde{G})$ and therefore $L(H') = \bar{L}$. Finally by Lemma 5.3 we have $W(H') \leqslant F$. ∎

By De Morgan's law we conclude this final result.

THEOREM 5.9. *The class of languages* $\{L(H) \mid H \in \mathscr{H}(F)\}$ *form a Boolean algebra with universal element* $L(F)$ *and null element* $\varnothing$. ∎

In particular it is immediate that $L(F) = V_T^+$ if F is the maxgrammar such that $\mathrm{OPM}(F)$ has neither blanks nor $\doteq$ —cycles (see App. 2).

6. CONCLUSION

It seems possible to extend our results in two directions. First, one could consider OPM's with conflicts, and extend our study to OG's which generate precedence conflicts, by proving the existence of a maxgrammar even in this case. Lattice properties should then follow. Second, one could drop the requirement that q be finite and extend the results to grammars with an unbounded number of terminal instances in each production. (by using the weaker notion of covering (Gray and Harrison, 1972) instead of structural equivalence).

Two major conclusions relevant to the problem of inferring a grammar from a given sample of sentences stem from the previous results.

It has been proved by Gold (1967) that any class of languages containing all the finite languages and an infinite language cannot be identified in the limit by any grammatical inference procedure, unless the informant is allowed to indicate which strings (nonsentences) should not be in the language, in addition to providing examples of strings in the language. This negative result implies that not even finite-state languages can be identified in the limit under the same

[9] Although the results of McNaughton (1967) assume in the definition of a grammar that $S \subseteq V_N$ rather than $S \in V_N$ (as we do), they can still be applied to our case since we have stated in Sect. 2 that renaming rules are not bracketed.

assumptions. On the other hand, free operator precedence grammars represent a new class of languages which can be identified in the limit without requiring availability of nonsentences to the identifying procedure (Crespi-Reghizzi, 1970).

Their generative capability appears fairly adequate (Crespi-Reghizzi, 1973) for generating arithmetic-expressionlike languages, although free operator precedence languages do not include all finite-state languages (see App. 2). The fact that any operator precedence language admits a standard homogeneous form significantly reduces the numeration needed to identify a grammar which is a satisfactory generalization of the given sample of sentences. The lattice properties of homogeneous grammars allow the elimination from consideration of large sets of grammars whenever a certain grammar is found to be incompatible with a given sample of sentences or nonsentences.

APPENDIX 1. An Example

Consider the maxgrammar $F_{M,2}$ corresponding to the precedence relations $b \doteq a$ and $b < . b$. Its productions are (they are indexed for reference purposes):

1. $\langle\{a\}, \{a\}\rangle \rightarrow a$,
2. $\langle\{b\}, \{b\}\rangle \rightarrow b$,
3. $\langle\{b\}, \{a\}\rangle \rightarrow ba$,
4. $\langle\{b\}, \{b\}\rangle \rightarrow b\langle\{b\}, \{b\}\rangle$,
5. $\langle\{b\}, \{a, b\}\rangle \rightarrow b\langle\{b\}, \{a\}\rangle$,
6. $\langle\{b\}, \{a, b\}\rangle \rightarrow b\langle\{b\}, \{a, b\}\rangle$,
7. $S \rightarrow \langle\{a\}, \{a\}\rangle$,
8. $S \rightarrow \langle\{b\}, \{b\}\rangle$,
9. $S \rightarrow \langle\{b\}, \{a\}\rangle$,
10. $S \rightarrow \langle\{b\}, \{a, b\}\rangle$.

F_M[10] is homogeneous, and generates the language

$$L_{M,2} = \{a\} \cup \{b\} \cup \{b\}^+ \{b\} \cup \{ba\} \cup \{bba\} \cup \{b\}^+ \{bba\}$$
$$= \hat{L}_1 \cup \hat{L}_2 \cup \cdots \cup \hat{L}_6 .$$

We now list the grammars and languages in $\mathscr{F}(M)$ and $\mathscr{L}(M)$. The grammars are derived from F_M by eliminating one or more of its productions and reducing (where necessary) the resulting grammar. Each grammar is identified by the set

[10] We omit the subscript 2 since $F_{M,q} = F_{M,2}$, for any $q \geqslant 2$, because M is $\doteq$ —acyclic.

of indices of its productions, and each corresponding language by the set of indices of the sublanguages $\hat{L}_i$, $1 \leqslant i \leqslant 6$, of which it is a union.

$F_1 = \{2, 3, 4, 5, 6, 8, 9, 10\}$; $L_1 = \{2, 3, 4, 5, 6\}$,
$F_2 = \{1, 3, 5, 6, 7, 9, 10\}$; $L_2 = \{1, 4, 5, 6\}$,
$F_3 = \{1, 2, 4, 7, 8\}$; $L_3 = \{1, 2, 3\}$,
$F_4 = \{1, 2, 3, 5, 6, 7, 8, 9, 10\}$; $L_4 = \{1, 2, 4, 5, 6\}$,
$F_5 = \{1, 2, 3, 4, 7, 8, 9\}$; $L_5 = \{1, 2, 3, 4\}$,
$F_6 = \{1, 2, 3, 4, 5, 7, 8, 9, 10\}$; $L_6 = \{1, 2, 3, 4, 5\}$.

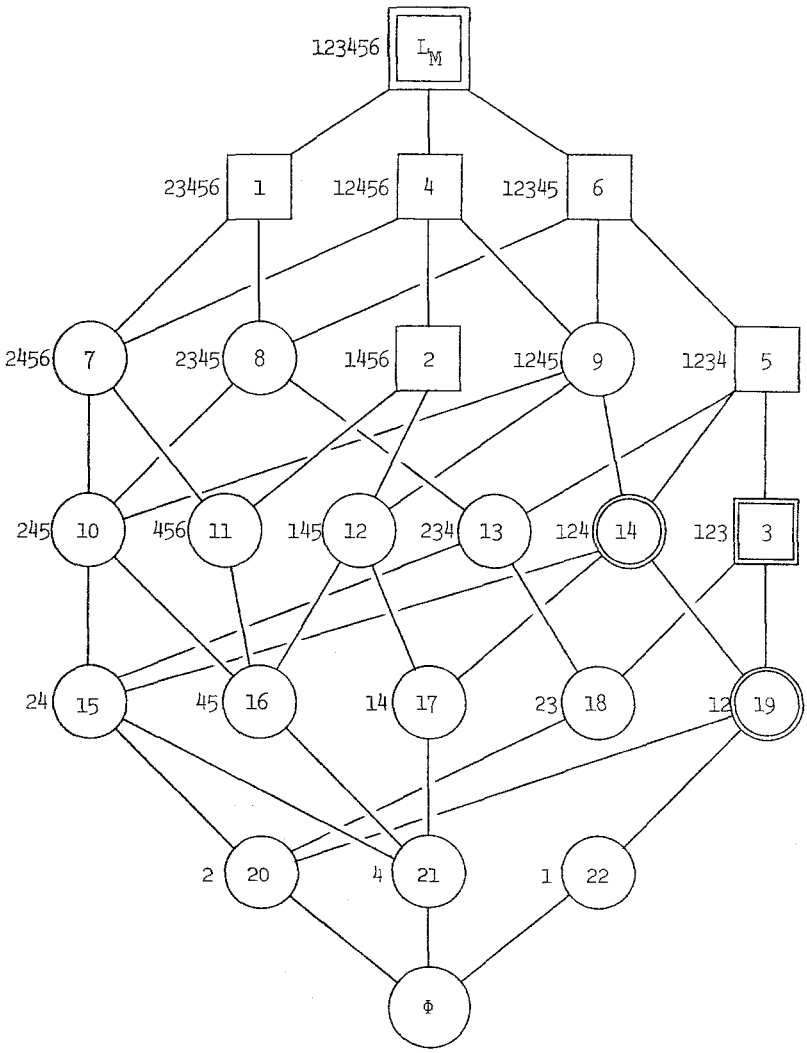

FIG. 1. The lattice of free languages and grammars corresponding to the Example.

The set $\{L_1, L_2, ..., L_6\}$ is a poset whose maximal elements are L_1, L_4, and L_6 since $L_2 \leqslant L_4$ and $L_3 \leqslant L_5 \leqslant L_6$. Applying the meet ($*$) operation to $\{L_1, L_4, L_6\}$ yields

$$F_7 = F_1 * F_4 = \{2, 3, 5, 6, 8, 9, 10\}; \qquad L_7 = \{2, 3, 5, 6\},$$
$$F_8 = F_1 * F_6 = \{2, 3, 4, 5, 8, 9, 10\}; \qquad L_8 = \{2, 3, 4, 5\},$$
$$F_9 = F_4 * F_6 = \{1, 2, 3, 5, 7, 8, 9, 10\}; \qquad L_9 = \{1, 2, 4, 5\}.$$

The set $\{L_2, L_3, L_5, L_7, L_8, L_9\}$ is a poset; the set of maximal elements is $\{L_2, L_5, L_7, L_8, L_9\}$. Again apply the meet operation to this latter set as done above and so forth, until all the grammars obtained generate the empty language. The lattice thus obtained is shown in Fig. 1; its poset is $\{\varnothing, L_1, L_2, ..., L_{22}, L_M\}$. Each node in the lattice bears the index of one of the elements in this poset, and the indices of the corresponding sublanguages $\hat{L}_i$, $1 \leqslant i \leqslant 6$, appear to the left of each node.

The grammars and corresponding languages represented by square nodes are obtained from F_M by eliminating a single production. The set of lattice elements represented by square nodes is a set of $*$-generators for the lattice, i.e., each lattice element can be derived by repeated application of $*$ to these generators.

Four maxlanguages (L_M, L_{14}, L_3, L_{19}) are contained in the lattice; they are represented by double circles and squares. Corollary 4.2 shows that these elements form a lattice. Note that the lattice of Fig. 1 is not closed under complementation with respect to L_M, since $L_M - L_4 = \hat{L}_3 = \{b\}^+\{b\}$ is not in the lattice.

APPENDIX 2. Other Properties of Free Grammars

We state without proof (proofs can be found in Crespi-Reghizzi (1970)) some other intuitive properties of maxgrammars and free grammars which, although not directly related to the rest of the paper, seem to be worth mentioning.

Statement A.1. Let G be an OPG and $L(G) = V_T^+$. Then OPM(G) is $\doteq$ —acyclic [i.e., there are no chains $a_1 \doteq a_2 \doteq \cdots \doteq a_k \doteq a_1$, $k > 1$, in OPM(G)]. ∎

Statement A.2. Let G_M be a maxgrammar with OPM M.
Then $L(G) = V_T^+$ iff M is $\doteq$ —acyclic and complete (i.e., $M_{ab} = 1, \forall a, b$). ∎

Statement A.3. No free grammar generates the regular language $L = \{a^+ - \{aa\}\}$.

Statement A.4. Languages generated by free-grammars are *noncounting* according to the definition given by Crespi-Reghizzi (1976). ∎

Received: July 30, 1973; revised; May 13, 1977

REFERENCES

CRESPI-REGHIZZI, S. (1970), The Mechanical Acquisition of Precedence Grammars (Ph.D. Dissertation), Report No. UCLA-ENG-7054, School of Engineering and Appl. Science, Univ. of California, Los Angeles, June 1970.

CRESPI-REGHIZZI, S. (1976), Non-counting Languages and Learning, *in* "Computer Oriented Learning Processes" (J. C. Simon, Ed.), Noordhoff, Leyden.

CRESPI-REGHIZZI, S., MELKANOFF, M., AND LICHTEN, L. (1973), The Use of Grammatical Inference for Designing Programming Languages, *Communications of the ACM* **16**, 83.

FELDMAN, J. (1972), Some decidability results on grammatical inference and complexity, *Inform. Contr.* **20**, 244.

FISCHER, M. (1969), Some Properties of Precedence Languages, Conf. Record of ACM Symp. on Theory of Computing, Marina Del Rey, Calif., pp. 181–190.

FLOYD, R. W. (1963), Syntactic analysis and operator precedence, *J. Assoc. Comput. Mach.* **10**, 316.

FU, K. S., AND BOOTH, T. L. (1975), Grammatical inference: Introduction and Survey—Part I, *IEEE Trans. Systems, Man and Cybernetics* **SMC-5**, 95–111.

GOLD, M. (1967), Language identification in the limit, *Inform. Contr.* **10**, 447.

GRAY, J., AND HARRISON, M. (1972), On the covering and reduction problems for context-free grammars, *J. Assoc. Comput. Mach.* **12**, 42–52.

GREIBACH, S. (1965), A new normal form for Context-free Grammars, *J. Assoc. Comput. Mach.* **12**, 42–52.

HOPCROFT, J., AND ULLMAN, J. (1969), Formal Languages and Their Relation to Automata, Addison-Wesley, Reading, Mass.

MC NAUGHTON, R. (1967), Parenthesis grammars, *J. Assoc. Comput. Mach.* **14**, 490–500.

REYNOLDS, J. C. (1968), Grammatical Covering, Tech. Memo, No. 96, Argonne National Laboratory, Applied Math. Div., March 1968.