Size Complexity in Context-Free Grammar Forms

SEYMOUR GINSBURG AND NANCY LYNCH

University of Southern California, Los Angeles, California

ABSTRACT. Grammar forms are compared for their efficiency in representing languages, as measured by the sizes (i e total number of symbols, number of variable occurrences, number of productions, and number of distinct variables) of interpretation grammars For every regular set, right- and left-linear forms are essentially equal in efficiency Any form for the regular sets provides, at most, polynomial improvement over right-linear form Moreover, any polynomial improvement is attained by some such form, at least on certain languages Greater improvement for some languages is possible using forms expressing larger classes of languages than the regular sets However, there are some languages for which no improvement over right-linear form is possible

While a similar set of results holds for forms expressing exactly the linear languages, only linear improvement can occur for forms expressing all the context-free languages.

KEY WORDS AND PHRASES complexity, grammar forms, size of grammars

CR CATEGORIES 5 2

1. *Introduction*

In [1], the concept of "grammar form" was introduced to model the situation where all grammars structurally close to a given master grammar are of interest Among questions naturally formulated in this framework are many about the complexity or efficiency of grammars. For example, is there a "type of grammar" which improves the efficiency of right-linear form for defining the regular sets, and if so, by how much? Grammar forms provide a reasonable and tractable way of considering the totality of allowable expressions, thereby permitting the above question to be answered with both upper and lower bounds.

The general problem of concern to us is the following: Which grammar forms are more efficient than others for defining families of languages, and how much gain in efficiency is possible? In [2], this question is answered for efficiency measured in terms of derivation complexity. The purpose of this paper is to consider the question when "size of grammar" is the complexity measure.

There are five sections in addition to the present introductory one. Section 2 contains basic notions about context-free grammar forms, as well as definitions of four measures of grammar size (each similar to one in [3]) considered throughout the paper. The results obtained usually apply to all four measures.

Section 3 deals with forms defining exactly the regular sets. Using a "reversal" construction, it is first shown that for every regular set right- and left-linear forms are of

This work was presented at the Seventh Annual SIGACT Symposium, Albuquerque, New Mexico, May 1975.

This work was supported in part by a Guggenheim Fellowship and by the National Science Foundation under Grants 42306 and DCR 92373

Authors' addresses S Ginsburg, Computer Science Department, University of Southern California, Salvatori Computer Science Center, Los Angeles, CA 90007, N. Lynch, Departments of Mathematics and Computer Science, University of Southern California, Salvatori Computer Science Center, Los Angeles, CA 90007

equal efficiency. Next, an upper bound is given on the amount of improvement possible over right- or left-linear form. A key point in the argument is the simulation of variables embedding themselves on the right by variables embedding themselves on the left; a construction similar to the reversal mentioned above is used. Finally, it is proved that every polynomial improvement over right-linear form is actually attainable by some form defining the regular sets.

Section 4 considers grammar forms whose defining power is greater than the regular sets. For such forms, it is possible to get greater improvement than that obtained in Section 3. However, there are some regular sets for which right-linear form is optimal, that is, these sets cannot be defined more efficiently by any other form, regardless of the expressive power.

Section 5 sketches results, similar to those in Sections 3 and 4, for grammar forms defining the linear languages. In addition, it is noted that for forms defining all the context-free languages, the variation possible is much less (in fact, linear).

Section 6 discusses some open questions.

2. *Preliminaries*

We first recall some elementary notions about context-free grammar forms. Then we present four types of "sizes" with which we shall be concerned

Definition. A *(context-free) grammar form* is a 6-tuple $F = (V,\Sigma,\mathcal{V},\mathcal{S},\mathcal{P},\sigma)$, where

(i) V is an infinite set of abstract symbols,

(ii) Σ is an infinite subset of V such that $V - \Sigma$ is infinite, and

(iii) $G_F = (\mathcal{V},\mathcal{S},\mathcal{P},\sigma)$, called the *form grammar*, is a context-free grammar[1] such that $\mathcal{S} \subseteq \Sigma$ and $(\mathcal{V} - \mathcal{S}) \subseteq (V - \Sigma)$.

The reader is referred to [1] for motivation and further details about grammar forms.

Throughout, V and Σ are assumed to be fixed infinite sets satisfying conditions (i) and (ii) above. All context-free grammar forms are with respect to this V and Σ. Also, the adjective "context-free" is usually omitted from the phrase "context-free grammar form."

For our purposes, we shall henceforth assume each context-free grammar has at least one production

The purpose of a grammar form is to specify a family of grammars, each "structurally close" to the form grammar. This is done via the notion of:

Definition. An *interpretation* of a grammar form $F = (V,\Sigma,\mathcal{V},\mathcal{S},\mathcal{P},\sigma)$ is a 5-tuple $I = (\mu,V_I,\Sigma_I,P_I,S_I)$, where

(i) μ is a substitution on $\mathcal{V}^*$ such that $\mu(\mathrm{a})$ is a finite subset of Σ^* for each a in $\mathcal{S}$, $\mu(\xi)$ is a finite subset of $V - \Sigma$ for each ξ in $\mathcal{V} - \mathcal{S}$, and $\mu(\xi) \cap \mu(\eta) = \emptyset$ for each ξ and η, $\xi \neq \eta$, in $\mathcal{V} - \mathcal{S}$.

(ii) P_I is a subset of $\mu(\mathcal{P}) = \cup_{\pi \text{in} \mathcal{P}}\, \mu(\pi)$, where $\mu(\alpha \to \beta) = \{u \to v \mid u \text{ in } \mu(\alpha), v \text{ in } \mu(\beta)\}$,

(iii) S_I is in $\mu(\sigma)$, and

(iv) Σ_I (V_I) contains the set of all symbols in $\Sigma(V)$ which occur in P_I (together with S_I).

$G_I = (V_I,\Sigma_I,P_I,S_I)$ is called the *grammar* of I.

An interpretation is usually exhibited by indicating S_I, P_I, and (implicity or explicity) μ. The sets V_I and Σ_I are ordinarily not stated explicity.

A grammar form determines a family of grammars and a family of languages as follows:

Definition. For each grammar form F, $\mathcal{G}(F) = \{G_I \mid I \text{ an interpretation of } F\}$ is called the *family of grammars of* F and $\mathcal{L}(F) = \{L(G_I) \mid G_I \text{ in } \mathcal{G}(F)\}$ the *grammatical family of* F.

[1] We assume the reader is familiar with context-free grammars Here $\mathcal{V}$ is the total alphabet, $\mathcal{S}$ is the terminal alphabet, $\mathcal{P}$ is the set of productions, and σ is the start variable The empty work ϵ is allowed as the right-hand side of a production

As mentioned in the Introduction, we are concerned with certain "size" measures of various grammar forms. Four specific such forms, to be considered in the remaining sections, are the following:

Definition. The grammar form $(V,\Sigma,\{\sigma,a\},\{a\},\{\sigma \rightarrow a\sigma,\ \sigma \rightarrow a\},\sigma)$ is called *right-linear* form. The grammar form $(V,\Sigma,\{\sigma,a\},\{a\},\{\sigma \rightarrow \sigma a,\ \sigma \rightarrow a\},\sigma)$ is called *left-linear* form. The grammar form $(V,\Sigma,\{\sigma,a\},\{a\},\{\sigma \rightarrow a\sigma a,\ \sigma \rightarrow a\},\sigma)$ is called *standard linear* form The grammar form $(V,\Sigma,\{\sigma,a\},\{a\},\{\sigma \rightarrow \sigma\sigma,\ \sigma \rightarrow a\},\sigma)$ is called *Chomsky binary* form.

Note that the grammars of the interpretations of each of the above forms are well-known types of context-free grammars. Thus each grammar of an interpretation of left-linear form is a left-linear grammar (and conversely), each grammar of an interpretation of standard linear form is a linear context-free grammar (and conversely), etc.

The size measures of concern to us are now given. Each has already been considered in the literature with respect to context-free grammars [3].

Notation. For each context-free grammar G, let

(a) $S(G)$ be the total number of occurrences of variables and terminals[2] on both sides of all productions in G,

(b) $V(G)$ be the total number of occurrences of variables on both sides of all productions in G,

(c) $P(G)$ be the number of productions in G,

(d) $N(G)$ be the number of variables in G.

Clearly, $N(G) \leqq P(G) \leqq V(G) \leqq S(G)$ for each reduced context-free grammar[3] G.

3. *Forms Defining the Regular Sets*

In this section we first establish that right- and left-linear forms are of approximately equal efficiency, as measured by each of our four criteria. Then we prove that any grammar form defining the regular sets gives at most polynomial improvement over right-linear form. Finally, by exhibiting a sequence of "worst possible" languages, we show that every polynomial improvement may be realized by some form defining the regular sets.

PROPOSITION 3.1. *For each right-linear (left-linear) grammar G, there exists an equivalent*[4] *left-linear (right-linear) grammar G' such that*

$$S(G') \leqq S(G) + P(G) + 1, \quad V(G') \leqq V(G) + P(G) + 1,$$
$$P(G') = P(G) + 1, \quad \textit{and} \quad N(G') = N(G) + 1.$$

PROOF. We give the argument for the case where $G = (V_1,\Sigma_1,P_1,S_1)$ is a right-linear grammar Essentially we simulate each left-to-right derivation in G by a right-to-left derivation in G'. Specifically, let S be a new symbol not in V_1, and let P' consist of the following:

(a) For each production in P_1 of the type $A \rightarrow wB$, with B in $V_1 - \Sigma_1$ and w in $\Sigma_1{}^*$, let $B \rightarrow Aw$ be in P'.

(b) For each production in P_1 of the type $A \rightarrow w$, w in Σ_1^*, let $S \rightarrow Aw$ be in P'.

(c) $S_1 \rightarrow \epsilon$ is in P'.

Then $G' = (V_1 \cup \{S\},\Sigma_1,P',S)$ satisfies the conclusions of the proposition. (Intuitively, G' simulates the action of G "in reverse," i.e. $S_1 \underset{G}{\Rightarrow} w_0A_0 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} w_0 \cdots w_rA_r \underset{G}{\Rightarrow} w_0 \cdots w_{r+1}$, each w_i in Σ_1^*, if and only if $S \underset{G'}{\Rightarrow} A_rw_{r+1} \underset{G'}{\Rightarrow} \cdots \underset{G'}{\Rightarrow} A_0w_1 \cdots w_{r+1} \underset{G'}{\Rightarrow} S_1w_0 \cdots w_{r+1} \underset{G'}{\Rightarrow} w_0 \cdots w_{r+1}$.

We now consider arbitrary forms defining exactly the regular sets. Our interest is in determining if any are substantially more efficient than right- (or left-) linear form for the

[2] Thus, no occurrence of the symbol ϵ is counted in determining $S(G)$

[3] Remember that all context-free grammars here are assumed to have at least one production

[4] Two context-free grammars G_1 and G_2 are said to be *equivalent* if $L(G_1) = L(G_2)$.

representation of particular regular sets. Several general questions arise: (1) How large a gain in efficiency can be achieved? (2) Is there a single most efficient form for the entire family of regular sets? (3) Are there pairs of forms, each more efficient than the other, for different languages? Questions (1) and (2) are answered in this section, while (3) remains open.

We begin by showing that, for three of the four measures under consideration, every form for the regular sets has a polynomial that bounds its improvement over right- or left-linear form. To do this, we need two lemmas, each transforming arbitrary forms defining the regular sets into a normal form.

LEMMA 3.2. *For each grammar form F, there exists an equivalent*[5] *form F' and positive integers c, n with the following properties*:

(1) *F' is completely reduced*[6] *and sequential*,[7] *and*

(2) *for each G in $\mathcal{G}(F)$, there exists an equivalent G' in $\mathcal{G}(F')$ such that $M(G') \leqq c[M(G)]^n$ if M is in $\{S,V,P\}$, and $N(G') \leqq N(G)$.*

PROOF The existence of an F' satisfying condition (1) is guaranteed by[8] Theorem 3.1 of [1]. To verify that F' also satisfies condition (2), we follow the constructions leading to the proof of Theorem 3.1 of [1], showing that the growth in size of interpretation grammars is bounded at each step.

Given F, we obtain, in the obvious way, an equivalent reduced grammar form F_a by Lemma 3.1 of [1]. For each G in $\mathcal{G}(F)$, there is some equivalent G' in $\mathcal{G}(F_a)$ such that $S(G') \leqq S(G)$, $V(G') \leqq V(G)$, $P(G') \leqq P(G)$, and $N(G') \leqq N(G)$.

By the proof of Lemma 3.2 of [1], we obtain a reduced, noncyclic grammar form F_b equivalent to F_a. This procedure involves at most $N(G_{F_a})$ repetitions of a construction eliminating a single maximal cycle set of F_a Let F_i denote the form resulting from F_a after i maximal cycle sets are eliminated. Then for each G in $\mathcal{G}(F_i)$ we obtain, in the natural way, an equivalent G' in $\mathcal{G}(F_{i+1})$. Since each production of G' has as its left (right) side the left (right) side of some production of G, it is straightforward to see that $S(G') \leqq [S(G)]^2$, $V(G') \leqq [V(G)]^2$, $P(G') \leqq [P(G)]^2$, and $N(G') \leqq N(G)$. Thus to each G in $\mathcal{G}(F_a)$ there corresponds an equivalent G' in $\mathcal{G}(F_b)$, with $S(G') \leqq [S(G)]^{2^{N(G_{F_a})}}$, $V(G') \leqq [V(G)]^{2^{N(G_{F_a})}}$, $P(G') \leqq [P(G)]^{2^{N(G_{F_a})}}$, and $N(G') \leqq N(G)$.

The construction of Lemma 3.3 of [1] is now applied to F_b to obtain an equivalent, reduced form F_c with no productions of the type $\xi \rightarrow \eta$, ξ and η variables. For each G in $\mathcal{G}(F_b)$, the natural equivalent G' in $\mathcal{G}(F_c)$ has $S(G') \leqq [S(G)]^2$, $V(G') \leqq [V(G)]^2$, $P(G') \leqq [P(G)]^2$, and $N(G') \leqq N(G)$.

Next, the construction of Lemma 3.4 of [1] is applied to F_c to get an equivalent, completely reduced grammar form F_d. This procedure involves at most $N(G_{F_c})$ repetitions of a construction which eliminates a single non-partially-self-embedding variable of F_c, followed by the addition of several productions to insure that each nonstart variable partially embeds itself in one step. The addition of productions involves no increase in size of interpretation grammars. Let $\bar{F}_i$ be the form resulting from F_c after i non-partially-self-embedding variables are eliminated For each i, suppose $v(i)$ is the variable eliminated in going from $\bar{F}_i$ to $\bar{F}_{i+1}$ Let $k(i)$ be the maximum number of times $v(i)$ occurs on the right of any single production of $\bar{F}_i$ Then for each G in $\mathcal{G}(\bar{F}_i)$, the natural equivalent G' in $\mathcal{G}(\bar{F}_{i+1})$ has $S(G') \leqq (k(i) + 1)[S(G)]^{k(i)+2}$, $V(G') \leqq (k(i) + 1)[V(G)]^{k(i)+2}$, $P(G') \leqq$

[5] Two forms F and F' are called *equivalent* if $\mathcal{L}(F) = \mathcal{L}(F')$

[6] A context-free grammar $G = (V_1,\Sigma_1,P,\sigma)$ is said to be *completely reduced* if (i) G is reduced, (ii) there are no variables α and β in $V_1 - \Sigma_1$ such that $\alpha \rightarrow \beta$ is in P, and (iii) for each variable α in $V_1 - (\Sigma_1 \cup \{\sigma\})$ there exist x and y in Σ_1^*, $xy \neq \epsilon$, such that $\alpha \rightarrow x\alpha y$ is in P A grammar form is said to be *completely reduced* if its form grammar is

[7] A context-free grammar $G = (V_1,\Sigma_1,P,S)$ is said to be *sequential* if the variables in $V_1 - \Sigma_1$ can be ordered ξ_1, , ξ_k, with $\xi_1 = S$, in such a way that if $\xi_i \rightarrow x\xi_j y$ is a production in P then $j \geqq i$ A grammar form is said to be *sequential* if its form grammar is

[8] Theorem 3.1 of [1] is the following result Each grammar form has an equivalent, completely reduced, sequential grammar form

$[P(G)]^{k(i)+2}$, and $N(G') \leqq N(G)$. (This is because each production of G' has as its left side the left side of some production of G, and as its right side the right side of some production of G, with at most $k(i)$ positions replaced by right sides of productions of G.) Let $k = \max\{k(i) \mid i\}$. Then for each G in $\mathscr{G}(F_c)$, an equivalent G' in $\mathscr{G}(F_d)$ can be found so that:

$$S(G') \leqq (k+1)^{(k+2)^{N(G_{F_c})}} [S(G)]^{(k+2)^{N(G_{F_c})}},$$

$$V(G') \leqq (k+1)^{(k+2)^{N(G_{F_c})}} [V(G)]^{(k+2)^{N(G_{F_c})}},$$

$$P(G') \leqq [P(G)]^{(k+2)^{N(G_{F_c})}}, \quad \text{and} \quad N(G') \leqq N(G).$$

These bounds are obtained by replacing each $k(i)$ by k in the bounds for $\bar{F}_{i+1}$ and using direct substitution. The bounds for $P(G')$ and $N(G')$ are obtained in a straightforward manner, while those for $S(G')$ and $V(G')$ require bounding a geometric series in the exponent of $k + 1$.

Finally, Theorem 3.1 of [1] is applied to F_d to obtain an equivalent, completely reduced, sequential form F'. Since $\mathscr{G}(F_d) \subseteq \mathscr{G}(F')$, there is no increase in grammar size at this step.

Using the above sequence of construction and bounds, the lemma follows.

LEMMA 3.3. *For each grammar form F defining the regular sets, there exists an equivalent form F' and positive integers c, n with the following properties*:

(1) *each production of F' is one of the types $\alpha \rightarrow \beta\gamma$, $\alpha \rightarrow w\beta$, $\alpha \rightarrow \beta w$, or $\alpha \rightarrow w$, where α, β, γ are variables and w is a terminal word*;

(2) *F' is sequential, reduced, and for every variable α of F', $\alpha \underset{G_{F'}}{\overset{*}{\Rightarrow}} w$ for some nonempty terminal word w; and*

(3) *for each G in $\mathscr{G}(F)$, there exists an equivalent G' in $\mathscr{G}(F')$ such that $M(G') \leqq c[M(G)]^n$ for all M in $\{S,V,P\}$.*

PROOF. Let $F'' = (N,\Sigma,\mathscr{V}_2,\mathscr{S}_2,\mathscr{P}_2,\sigma_2)$ be the grammar form given by the conclusion of Lemma 3.2 Informally, F' is obtained from F'' by simulating single productions with "long" right-hand sides by sequences of productions with "short" right-hand sides Formally, let $\mathscr{P}_3$ consist of the following productions:

(a) Each production in F'' which is one of the four permitted types is in $\mathscr{P}_3$.

(b) Each production in F'' not one of the four permitted types is of the type $\alpha \rightarrow x_1 \cdots x_m$, $m \geqq 3$, where each x_i is either a variable or a nonempty terminal word, and no two consecutive x_i are terminal words.

By Theorem 2.3 of[9] [1] (since F'' is reduced), only x_1 or x_m, but not both, can be α. If either $x_1 = \alpha$ or neither x_1 nor x_m is α, let $\beta_1, \cdots, \beta_{m-2}$ be new variables and let $\alpha \rightarrow x_1\beta_1$, $\beta_1 \rightarrow x_2\beta_2, \cdots, \beta_{m-3} \rightarrow x_{m-2}\beta_{m-2}$, and $\beta_{m-2} \rightarrow x_{m-1}x_m$ be in $\mathscr{P}_3$. If $x_m = \alpha$, let β_m, $\beta_{m-1}, \cdots, \beta_3$ be new variables and $\alpha \rightarrow \beta_m x_m$, $\beta_m \rightarrow \beta_{m-1}x_{m-1}, \cdots, \beta_4 \rightarrow \beta_3 x_3$, $\beta_3 \rightarrow x_1x_2$ be in $\mathscr{P}_3$.

Let $\mathscr{V}_3$ consist of the variables occurring in the productions of $\mathscr{P}_3$ and $F' = (V,\Sigma,\mathscr{V}_3,\mathscr{S}_2,\mathscr{P}_3,\sigma_2)$. Clearly, F' is equivalent to F, each production in F' is one of the four types in (1), and F' is sequential Although F' has no productions of the type $\xi \rightarrow \eta$, ξ and η variables, F' may not be completely reduced. However, each variable of F' derives a nonempty terminal word. (This is because σ_2 derives a nonempty terminal word and for each variable γ in F'', $\gamma \rightarrow x\gamma y$ for some terminal word $xy \neq \epsilon$ since F'' is completely reduced.) Thus (2) is satisfied. Let k be the largest value of m for which $\alpha \rightarrow x_1 \ldots x_m$ is a production of F'', where α is a variable, each x_i is either a variable or a nonempty terminal word, and no two consecutive x_i are terminal words. Then for each G in $\mathscr{G}(F'')$, the natural equivalent G' in $\mathscr{G}(F')$ has $S(G') \leqq 3[S(G)]$, $V(G') \leqq$

[9] Theorem 2 3 of [1] is the following. Let $F = (V,\Sigma,\mathscr{V},\mathscr{S},\mathscr{P},\sigma)$ be a reduced grammar form Then $\mathscr{L}(F)$ is the family of regular sets if and only if $L(G_F)$ is infinite and F contains no variable ξ such that $\xi \overset{*}{\Rightarrow} u\xi v$ for some words u, v in $\mathscr{S}^+$

$2k[V(G)]$, and $P(G') \leqq 2k[P(G)]$. These bounds, combined with the bounds for F'' arising from Lemma 3.2, yield (3), thereby completing the proof.

In order to show that each form for the regular sets can be simulated by right-linear form with at most polynomial loss of efficiency (for three of the four measures under consideration), it therefore suffices to restrict our attention to Lemma 3.3 forms For each of the Lemma 3.3 type forms, each variable may embed itself on the right or on the left, but not both The next lemma shows how to convert such a form into an equivalent one in which variables may only embed themselves on the right. The technique is similar to that used in Proposition 3.1 to convert left-linear to right-linear form. First though, we define an infinite sequence $\{F_n\}$ of forms for the regular sets, of successively greater "sequential depth," in which every variable embeds itself only on the right

Definition 3.4. For each $n \geqq 1$, let $F_n = (V,\Sigma,\{\alpha_0,\cdots,\alpha_{n-1},a\},\{a\},\mathcal{P}_n,\alpha_0)$, where $\mathcal{P}_n = \{\alpha_i \to \alpha_j\alpha_k \mid 0 \leqq i \leqq k \leqq n-1,\ i < j \leqq n-1\} \cup \{\alpha_i \to a\alpha_j \mid 0 \leqq i \leqq j \leqq n-1\} \cup \{\alpha_i \to \alpha_j a \mid 0 \leqq i < j \leqq n-1\} \cup \{\alpha_i \to a \mid 0 \leqq i \leqq n-1\}$.

Note that F_1 is a right-linear form..

We now simulate an arbitrary Lemma 3 3 form with a member of the sequence $\{F_n\}$.

LEMMA 3 5. *Let F be a grammar form for the regular sets, satisfying the following conditions*: (i) *Each production of F is one of the types $\alpha \to \beta\gamma$, $\alpha \to w\beta$, $\alpha \to \beta w$, or $\alpha \to w$, where α, β, γ are variables and w is a terminal word*; *and* (ii) *F is sequential, reduced, and for every variable α of F, $\alpha \overset{*}{\underset{G_F}{\Rightarrow}} w$ for some nonempty terminal word w. Let n be the number of variables in G_F. Then there exists a positive integer c with the following property: For every G in $\mathcal{G}(F)$ there exists some equivalent G' in $\mathcal{G}(F_n)$ such that $M(G') \leqq c[M(G)]^2$ for M in $\{S, P, V\}$ and $N(G') \leqq c[N(G)]^3$.*

PROOF Since F is sequential and generates only regular sets, the variables of the grammar G may be partitioned into levels so that all variables on a given level are either left recursive or right recursive. By means of a construction similar to that used in Proposition 3.1, we may transform each left recursive level into a right recursive level. More formally, we may assume without loss of generality that the variables of $F = (V,\Sigma,\mathcal{V},\mathcal{S},\mathcal{P},\sigma)$ are $\sigma = \alpha_0, \alpha_1, \cdots, \alpha_{n-1}$. Consider any interpretation (μ, G) of F, with $G = (V_1,\Sigma_1,P_1,S)$. There is no loss of generality in assuming that G is reduced and $V_1 - \Sigma_1 = \bigcup_{i=0}^{n-1} \mu(\alpha_i)$. Because of Theorem 2.3 of [1] and the fact that F is reduced, each production of G is one of the following types:

(1) $A \to BC$, where A and B are in $\mu(\alpha_i)$ and C is in $\mu(\alpha_j)$, for some i, j, $0 \leqq i < j \leqq n - 1$;

(2) $A \to BC$, where A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, and C is in $\mu(\alpha_k)$, for some i, j, k, $0 \leqq i < j \leqq n-1$ and $0 \leqq i \leqq k \leqq n-1$,

(3) $A \to wB$, where A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, and w is in Σ^*, for some i, j, $0 \leqq i \leqq j \leqq n-1$;

(4) $A \to Bw$, where A, B are in $\mu(\alpha_i)$ and w is in Σ^*, for some i, $0 \leqq i \leqq n-1$;

(5) $A \to Bw$, where A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, and w is in Σ^*, for some i, j, $0 \leqq i < j \leqq n-1$; and

(6) $A \to w$, where A is in $\mu(\alpha_i)$ and w is in Σ^*, for some i, $0 \leqq i \leqq n-1$.

We now define an interpretation (μ', G') of F_n Let $G' = (V_2,\Sigma_1,P_2,S)$ where V_2 consists of S and all the variables in P_2 The set P_2 consists of all productions in P_1 of types (2), (3), (5), and (6), as well as the following productions. (Productions (a)-(g) below simulate, in reverse, as in Proposition 3.1, the effect of taking types (1) and (4), in combination)

(a) $A \to w[BA]$ if A, B are in $\mu(\alpha_i)$ for some i, $0 \leqq i \leqq n-1$, w is in Σ^* and $B \to w$ is in P_1;

(b) $A \to C[BCA]_1$ and $[BCA]_1 \to w[BA]$ if A, B are in $\mu(\alpha_i)$ and C is in $\mu(\alpha_j)$ for some i, j, $0 \leqq i < j \leqq n$, w is in Σ^*, and $B \to Cw$ is in P_1;

(c) $A \to w[BCA]_2$ and $[BCA]_2 \to C[BA]$ if A, B are in $\mu(\alpha_i)$ and C is in $\mu(\alpha_j)$ for some i, j, $0 \leqq i < j \leqq n-1$, w is in Σ^*, and $B \to wC$ is in P_1;

(d) $A \rightarrow C[BCA]_3$ and $[BCA]_3 \rightarrow D[BA]$ if A, B are in $\mu(\alpha_i)$, C is in $\mu(\alpha_j)$, and D is in $\mu(\alpha_k)$ for some i,j,k, $0 \leqq i < k \leqq n - 1$, and $0 \leqq i < k \leqq n - 1$, and $B \rightarrow CD$ is in P_1;

(e) $[CA] \rightarrow D[BA]$ if A, B, C are in $\mu(\alpha_i)$ and D is in $\mu(\alpha_j)$, for some i,j, $0 \leqq i < j \leqq n - 1$, and $B \rightarrow CD$ is in P_1;

(f) $[CA] \rightarrow w[BA]$ if A, B, C are in $\mu(\alpha_i)$ for some i, $0 \leqq i \leqq n - 1$, w is in Σ^*, and $B \rightarrow Cw$ is in P_1; and

(g) $[AA] \rightarrow \epsilon$ for all variables A in V_1.

Let $\mu'(a)$ contain ϵ and every terminal word occurring in at least one production in P_2. If A is in $V_2 \cap \mu(\alpha_i)$, let A be in $\mu'(\alpha_i)$. If A is in $\mu(\alpha_i)$, then let $[BCA]_1$, $[BCA]_2$, $[BCA]_3$, and $[BA]$ be in $\mu(\alpha_i)$ for all variables $[BCA]_1$, $[BCA]_2$, $[BCA]_3$, and $[BA]$. Clearly G' is in $\mathscr{G}(F_n)$. It is straightforward to verify that the size of G' satisfies the conclusions of the lemma. It remains to show that $L(G') = L(G)$.

Intuitively, the new variables in G' play the following role. The purpose of the paired variables is to simulate from left to right a derivation which in G proceeds from right to left. Specifically, if a variable $[BA]$ is generated (from S) in a G'-derivation, then a word x, consisting entirely of terminals and of variables in V_1 which correspond to (sequentially) higher variables in F than A does, has already been generated immediately to the left of $[BA]$. Furthermore, $B \overset{*}{\underset{G}{\Rightarrow}} x$. Part of a G-derivation, from A, is being simulated, and one is waiting to see if A G-generates B (along with possibly other symbols) In particular, if a variable $[AA]$ is generated, then a word x, G-derivable from A, has already been generated, and the production $[AA] \rightarrow \epsilon$ is used. The purpose of the triple variables is to ensure that no more than two symbols occur on the right of any production in G'. If a variable $[BCA]_1$ is generated during a G'-derivation, then C has already been deposited immediately to the left of $[BCA]_1$, and some w in Σ^* such that $B \rightarrow Cw$ is in P_1 will next be deposited, along with $[BA]$. Similar remarks may be made for $[BCA]_2$ and $[BCA]_3$.

We now show that $L(G) \subseteq L(G')$. Consider a G-derivation δ of some word in $L(G)$. The only productions of δ not in P_2 are either of the type $A \rightarrow BC$, where A, B are in $\mu(\alpha_i)$ and C is in $\mu(\alpha_j)$, $0 \leqq i < j \leqq n - 1$, or of the type $A \rightarrow Bw$, where A, B are in $\mu(\alpha_i)$, $0 \leqq i \leqq n - 1$, and w is in Σ^+. (A production in P_1 of the type $A \rightarrow B$, where A and B are in $\mu(\alpha_i)$, $0 \leqq i \leqq n - 1$, is in P_2 since it is of type 3.) Note that for each such production, $\alpha_i \overset{*}{\underset{G_F}{\Rightarrow}} \alpha_i z$ for some nonempty word z in $\mathscr{S}^+$, i.e. α_i is "partially self-embedding on the left." We shall see that the effect of such productions can be simulated by using productions (a)–(g) in combination. Suppose there are no such productions in δ. Then there is nothing to prove and we are done. Suppose there are such productions. Consider the first such production, say $A \rightarrow BC$ or $A \rightarrow Bw$, with A, B, C, and w as above. Rearrange δ after using this occurrence so that B is the next variable which is expanded. (Clearly, this is possible.) Let δ' be the new G-derivation. Since G is reduced, $L(G)$ is regular, B is in $\mu(\alpha_i)$, and α_i is partially self-embedding on the left in F, it follows that the only possible productions with B on the left are either $B \rightarrow DE$, $B \rightarrow Dw$, or $B \rightarrow x$, where D is in $\mu(\alpha_i)$, E is in $\mu(\alpha_j)$, $i < j \leqq n - 1$, w is in Σ^*, and x is a word of terminals and of variables corresponding to variables α_r, $r > i$. If the expansion of B is one of the first two types, then without loss of generality we may assume that D is the next variable expanded, etc. Thus we may assume (with a possible change in notation) that the expansion of A in δ' involves

$$A = A_0 \underset{G}{\Rightarrow} A_1x_1 \underset{G}{\Rightarrow} A_2x_2x_1 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} A_kx_k\cdots x_1 \underset{G}{\Rightarrow} x_{k+1}\cdots x_1, \qquad (*)$$

where A_0, A_1, $\cdots$, A_k are in $\mu(\alpha_i)$; each x_j, $1 \leqq j \leqq k$, is either a terminal word or a variable corresponding to some α_r, $r > i$; and x_{k+1} is a word of terminals and of variables corresponding to α_r, $r > 1$. Also, x_{k+1} is of a form consistent with the allowable types of productions in F. To prove that $L(G) \subseteq L(G')$, it obviously suffices (by induction) to show that $A \overset{*}{\underset{G'}{\Rightarrow}} x_{k+1} \cdots x_1$.

To see this, note that $A \overset{*}{\underset{G'}{\Rightarrow}} x_{k+1}[A_kA]$ is implemented by using one or two productions of P_2, of types (a)-(d). For each j, $1 \leqq j \leqq k$, $[A_jA] \overset{*}{\underset{G'}{\Rightarrow}} x_j\,[A_{j-1}A]$ is implemented by either a type (e) or type (f) production. Also, $[A_0A] = [AA]$, so that $[A_0A] \overset{*}{\underset{G'}{\Rightarrow}} \epsilon$ by a type (g) production. Combining, $A \overset{*}{\underset{G'}{\Rightarrow}} x_{k+1} \cdots x_1[A_0A] \overset{*}{\underset{G'}{\Rightarrow}} x_{k+1} \cdots x_1$.

We next show that $L(G') \subseteq L(G)$. Consider a G'-derivation δ of an arbitrary word in $L(G')$. The only productions occurring in δ not in P_1 are those of types (a)-(g). We shall see that the effect of such productions can be simulated in G Suppose there are no such productions in δ. Then there is nothing to prove and we are done. Suppose there are some such productions. Consider the first production of types (a)-(d) occurring in δ, with A, B in $\mu(\alpha_i)$. Without loss of generality, this production, or this production in combination with the next production applied in δ, may be assumed to cause the depositing of a word of terminals and of variables corresponding to variables α_r, $r > i$, say $A \overset{*}{\underset{G'}{\Rightarrow}} z[BA]$. Also, $[BA]$ may be assumed to be the next variable expanded (using a type (e) or (f) production), and it may likewise be assumed that the paired variables are expanded immediately until an application of $[AA] \to \epsilon$ occurs. Thus the expansion of the variable A in δ (with a possible change in notation) involves

$$A = A_0 \overset{*}{\underset{G'}{\Rightarrow}} x_{k+1}[A_kA] \underset{G'}{\Rightarrow} x_{k+1}x_k[A_{k-1}A] \underset{G'}{\Rightarrow} \cdots \underset{G'}{\Rightarrow} x_{k+1} \cdots x_1[A_0A] \underset{G'}{\Rightarrow} x_{k+1} \cdots x_1, \qquad (**)$$

where A, A_1, $\cdots$, A_k are in $\mu(\alpha_i)$, each x_j, $1 \leqq j \leqq k$, is either a terminal word or a variable corresponding to some α_r, $r > i$, and x_{k+1} is a word of terminals and of variables corresponding to variables α_r, $r > i$. Also x_{k+1} must be of a type obtainable from (a), (b), (c), or (d). It suffices (by induction) to verify that $A \overset{*}{\underset{G}{\Rightarrow}} x_{k+1} \cdots x_1$.

To see this, note from the definition of (e) and (f) that G must contain the productions $A_{j-1} \to A_jx_j$, $1 \leqq j \leqq k$. Also, from the definition of x_{k+1}, $A_k \to x_{k+1}$ is in P_1. Thus

$$A \underset{G}{\Rightarrow} A_1x_1 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} A_kx_k \cdots x_1 \underset{G}{\Rightarrow} x_{k+1} \cdots x_1,$$

i.e. $A \overset{*}{\underset{G}{\Rightarrow}} x_{k+1} \cdots x_1$. This completes the proof.

The final lemma needed to show that every grammar form for the regular sets is simulatable by right-linear form with at most polynomial increase in size (for three of the size measures) states that each form F_n has the desired simulation property.

Lemma 3.6. *For each positive integer n, there exists a positive integer c with the following property: For each G in $\mathcal{G}(F_n)$, there exists an equivalent G' in $\mathcal{G}(F_1)$ such that $M(G') \leqq c[M(G)]^{2n}$ for each M in $\{S,V,P,N\}$.*

Proof. Let $G = (V_1,\Sigma_1,P,S)$ be in $\mathcal{G}(F_n)$. We now define $G' = (V_2,\Sigma_1,P_2,[S])$ in $\mathcal{G}(F_1)$ in such a way that G' simulates leftmost derivations of G. Let

$$V_2 - \Sigma_1 = \{[X] \mid X \text{ in } \bigcup_{i=0}^{n} ((V_1 - \Sigma_1) \cup ((V_1 - \Sigma_1) \times (V_1 - \Sigma_1)))^i\},$$

i.e. the variables of G' are to be all words of length less than or equal to n in which every position is occupied by either a variable of G or by an ordered pair of variables of G. Let P_2 consist of the following productions (where A, B, C are in $V_1 - \Sigma_1$, w is in Σ_1^*, and D_1, $\cdots$, D_{n-1} are in $(V_1 - \Sigma_1) \cup ((V_1 - \Sigma_1) \times (V_1 - \Sigma_1))$):

(a) $[AD_1\cdots D_k] \to [BCD_1\cdots D_k]$ if $A \to BC$ is in P_1 and $0 \leqq k \leqq n-2$;

(b) $[AD_1\cdots D_k] \to w[BD_1\cdots D_k]$ if $A \to wB$ is in P_1 and $0 \leqq k \leqq n-1$;

(c) $[AD_1\cdots D_k] \to [B\langle A,B\rangle D_1\cdots D_k]$ and $[\langle A,B\rangle D_1\cdots D_k] \to w[D_1\cdots D_k]$ if $A \to Bw$ is in P_1 and $0 \leqq k \leqq n-2$;

(d) $[AD_1\cdots D_k] \to w[D_1\cdots D_k]$ if $A \to w$ is in P_1 and $0 \leqq k \leqq n-1$; and

(e) $[\epsilon] \to \epsilon$.

It is easy to see that G' is in $\mathscr{G}(F_1)$, $L(G') \subseteq L(G)$, and the size increase is as stated. To complete the proof it thus suffices to show that $L(G) \subseteq L(G')$. Therefore let δ be a leftmost derivation of a word x in $L(G)$, and let δ' be the natural simulation of δ using productions of the type (a)-(e), where k is allowed to be as large as necessary. It then suffices to show that:

(*) No word in δ' has brackets containing more than n symbols.

We prove (*) by showing inductively that

(**) In each word of δ', the number of symbols in the brackets is at most n, and for each i, $1 \leqq i \leqq n$, the symbol in position i from the right, if it is a single variable, is in $\mu(\alpha_j)$ for some $j \geqq i-1$

Now (**) is certainly true for the first word, $[S]$, of δ'. Assume it is true for some word in δ' and that $[E_k\cdots E_1]$ is the variable being expanded by the next production. If $k=0$, the induction clearly follows Suppose $1 \leqq k \leqq n$ and for all i, $1 \leqq i \leqq k$, E_i is in $\mu(\alpha_j)$ for some $j \geqq i-1$ if E_i is a single variable If E_k is a single variable and $k \neq n$, then the next production in δ' is one of the types:

$[E_k\cdots E_1] \to [BCE_{k-1}\cdots E_1]$, where $E_k \to BC$ is in P_1, or
$[E_k\cdots E_1] \to w[BE_{k-1}\cdots E_1]$, where $E_k \to wB$ is in P_1 for some w in Σ_1^*, or
$[E_k\cdots E_1] \to [B\langle E_k,B\rangle E_{k-1}\cdots E_1]$, where $E_k \to Bw$ is in P_1 for some w in Σ_1^*, or
$[E_k\cdots E_1] \to w[E_{k-1}\cdots E_1]$, where $E_k \to w$ is in P_1 and w is in Σ_1^*

In all cases the induction follows. If E_k is a single variable and $k=n$, then since E_n is in $\mu(\alpha_{n-1})$ the next production in δ' can only be one of the types

$[E_n\cdots E_1] \to w[BE_{n-1}\cdots E_1]$, where $E_n \to wB$ is in P_1 and w is in Σ_1^*, or
$[E_n\cdots E_1] \to w[E_{n-1}\cdots E_1]$, where $E_n \to w$ is in P_1 and w is in Σ_1^*.

In either case, the induction follows. If E_k is a double variable $\langle A,B\rangle$, then the next production is of the type

$$[\langle A,B\rangle E_{k-1}\cdots E_1] \to w[E_{k-1}\cdots E_1] \text{ for some } w \text{ in } \Sigma_1^*,$$

and again the induction follows. Thus the induction holds in all cases. Therefore $L(G) \subseteq L(G')$ and the lemma is proved.

Combining Lemmas 3.3, 3.5, and 3 6, we have:

THEOREM 3.7 *For each grammar form F defining the regular sets, there exist positive integers c and n with the following property: For every G in $\mathscr{G}(F)$, there exists an equivalent G' in right-linear form such that $M(G') \leqq c[M(G)]^n$ for each M in $\{S,V,P\}$.*

Because of Proposition 3.1, Theorem 3.7 is also true for left-linear form.

Theorem 3 7 is not stated for $N(G)$ since the proof of Lemma 3.3 does not hold in this case However, Theorem 3.7 is still true for $N(G)$. Roughly, we apply Lemma 3 2 to an arbitrary form for the regular sets, obtaining an equivalent form in which $N(G)$ does not increase. We then use a reversal construction similar to the one in the proof of Lemma 3.5, to insure that all embedding takes place on the right Finally, we simulate the resulting form by right-linear form using a construction similar to that used in the proof of Lemma 3.6. Here, longer strings of variables may be needed than in the simpler "binary" case, and the possible cases to consider are notationally more complicated, but the ideas are essentially the same.

We now turn to the second major result of the section, namely, that any polynomial improvement may actually be obtained by some form defining the regular sets, at least on an infinite set of languages We begin by introducing a special family of languages and establish two lemmas about it

Definition 3.8. For all positive integers n and k, let $L_{n,k} = 0^*(10^*)^{k^n}$.

Thus $L_{n,k}$ is the set of all words in $\{0, 1\}^*$ which have exactly k^n occurrences of 1.

The first lemma states that each $L_{n,k}$ is definable by a grammar in F_n of size at most linear in k.

LEMMA 3.9. *For all positive integers n and k, there is a grammar G in* $\mathscr{G}(F_n)$ *such that* $L(G) = L_{n,k}$ *and* $M(G) \leqq (4n + 6)k$ *for each M in* $\{S,V,P,N\}$.

PROOF. Let $V_1 = \{0, 1\} \cup \{A_{i,j} \mid 0 \leqq i \leqq n - 1,\ 0 \leqq j \leqq k\}$ and $G = (V_1,\{0, 1\},P,A_{0,0})$, where

$$\begin{aligned} P = \{A_{i,j} \to A_{i+1,0}A_{i,j+1} \mid 0 \leqq i \leqq n - 2, 0 \leqq j \leqq k - 1\} \cup \{A_{i,k} \to \epsilon \mid 0 \leqq i \leqq n - 1\} \\ \cup \{A_{n-1,j} \to 0A_{n-1,j} \mid 0 \leqq j \leqq k\} \cup \{A_{n-1,j} \to 1A_{n-1,j+1} \mid 0 \leqq j \leqq k - 1\}. \end{aligned}$$

Clearly, G is in $\mathscr{G}(F_n)$, with each $A_{i,j}$ corresponding to α_i. It is readily seen that every $A_{i,0}$ generates $L_{n-i,k}$. (The second index of $A_{i,j}$ is used to count up to k.) Thus $L(G) = L_{n,k}$ Finally, it is a straightforward matter to verify the size bounds.

The second lemma asserts that each grammar, in right-linear form, which defines $L_{n,k}$ is of size at least k^n.

LEMMA 3 10. *For each positive integer s and each grammar G in right-linear form, with* $L(G) = 0^*(10^*)^s$, $M(G) \geqq s + 1$ *for each M in* $\{N,P,V,S\}$.

PROOF. It suffices to show that $N(G) \geqq s + 1$. The argument is of a standard type and consists of showing that if $N(G) < s + 1$ then an incorrect word is generated. Consider a word $x = 0^m(10^m)^s$, where m is some integer larger than the maximum number of terminal symbols in each production of P_1. Let

$$\delta\colon\quad S \Rightarrow w_1A_1 \Rightarrow w_1w_2A_2 \Rightarrow \cdots \Rightarrow w_1 \cdots w_lA_l \Rightarrow w_1 \cdots w_{l+1} = x$$

be a G-derivation of x, where each A_i, $1 \leqq i \leqq l$, is a variable, and each w_i, $1 \leqq i \leqq l + 1$, is in $\{0, 1\}^*$. From the choice of m, exactly s distinct words $w_{i_1}, \cdots, w_{i_s}$ contain a single occurrence of the symbol 1, and w_{l+1} is not one of them To complete the proof, it suffices to show that $A_{i_1-1}, \cdots, A_{i_s-1}, A_l$ are all different variables. Suppose $A_{i_t-1} = A_{i_u-1}$ for some t and u, $t < u$. Then

$$\begin{aligned} S \Rightarrow w_1A_1 \Rightarrow \cdots \Rightarrow w_1 \cdots w_{i_u-1}\, A_{i_u-1} = w_1 \cdots w_{i_u-1}A_{i_t-1} \\ \overset{*}{\Rightarrow} w_1 \cdots w_{i_u-1}\, w_{i_t} \cdots w_lw_{l+1} \end{aligned}$$

is a derivation of a word x' in $L(G)$. But x' has at least $s + 1$ occurrences of 1, contradicting the nature of the words in $L(G)$. A similar argument shows A_l to be distinct from the remaining variables. Thus $A_{i_1-1}, \cdots, A_{i_s-1}, A_l$ are all different.

From (i) Lemmas 3.9 and 3 10, (ii) the fact that for each context-free grammar $G_1 = (V_1,\Sigma_1,P_1,S_1)$ there is an equivalent reduced grammar $G_2 = (V_2,\Sigma_1,P_2,S_1)$, with $V_2 \subseteq V_1$ and $P_2 \subseteq P_1$, and (iii) the fact that $S(G)$ is the largest, and $N(G)$ the smallest, of the four measures for each reduced context-free grammar G; we get

THEOREM 3.11. *For each positive integer n, there exists a grammar form F for the regular sets and a positive integer c with the following property: For every integer* $k \geqq 1$, *there is a grammar G in* $\mathscr{G}(F)$ *such that for each M in* $\{S,V,P,N\}$, (1) $M(G) \leqq ck$, *and* (2) $M(G') \geqq k^n$ *for every equivalent grammar G' in right-linear form.*

From Theorem 3.11, Theorem 3.7, and the comment following Theorem 3.7, we see that every form for the regular sets may be similarly improved by any polynomial. In other words:

COROLLARY 3.12. *For every form F' defining the regular sets and every positive integer n, there exists a form F defining the regular sets and a positive integer c with the following property: For every integer* $k \geqq 1$, *there is a grammar G in* $\mathscr{G}(F)$ *such that for* *each M in $\{S,V,P,N\}$, (1) $M(G) \leqq ck$ and (2) $M(G') \geqq k^n$ for every equivalent grammar G' in $\mathscr{G}(F')$.*

The effect of Corollary 3.12 is that there is no "best" form for the regular sets.

4. *Forms Defining More than the Regular Sets*

In Section 3 we discussed the improvement possible using forms which define exactly the regular sets. In this section we examine the effect of allowing forms which define more than the regular sets. Specifically, we establish two results. The first asserts that by permitting such forms, we can achieve more than polynomial improvement on an infinite family of regular sets. The second says that even permitting such forms, there is an infinite family of regular sets for which no improvement over right-linear form is possible.

For the first result, we have:

THEOREM 4.1. *For each recursive function f and for arbitrarily large positive integers k, there is a grammar G in Chomsky binary form, defining a regular set, such that for each M in $\{S,V,P,N\}$, (1) $M(G) \leqq k$, and (2) $M(G') \geqq f(k)$ for each equivalent grammar G' in right-linear form.*

The proof is an easy corollary of [4, Prop. 7], the bounded simulation of right-linear grammars by one-way finite state acceptors, and the bounded simulation of any context-free grammar by one in Chomsky binary form.

For the second result, we have:

THEOREM 4.2. *For each positive integer k, there is a grammar G in right-linear form such that for each M in $\{S,V,P,N\}$, (1) $M(G) \leqq 10k$, and (2) $M(G') \geqq k$ for every equivalent context-free grammar G'.*

PROOF. Let $G = (V_1,\{a_1, \cdots, a_{2k}\}, P, A_1)$, where $V_1 = \{a_1, \cdots, a_{2k}\} \cup \{A_1, \cdots, A_{2k}\}$ and $P = \{A_i \rightarrow a_iA_i, A_i \rightarrow A_{i+1} \mid 1 \leqq i \leqq 2k-1\} \cup \{A_{2k} \rightarrow a_{2k}A_{2k}, A_{2k} \rightarrow \epsilon\}$. Then G is in right-linear form, $M(G) \leqq 10k$, and

$$L(G) = a_1^*a_1^* \cdots a_{2k-1}^*a_{2k}^*.$$

Suppose that $G' = (V_2, \{a_1, \cdots, a_{2k}\}, P_2, S_2)$ is a context-free grammar equivalent to G There is no loss of generality in assuming that G' is reduced. To complete the proof it suffices to show that for each i, $1 \leqq i \leqq 2k$, there exists a variable B_i in G' such that

(*) Either $B_i \overset{*}{\underset{G'}{\Rightarrow}} u_iB_iv_ia_iw_i$ or $B_i \overset{*}{\underset{G'}{\Rightarrow}} u_ia_iv_iB_iw_i$, for some u_i, v_i, w_i in $\{a_1, \cdots, a_{2k}\}^*$.

For suppose (*) holds. Then there are no three indices i,j,l, with $i < j < l$, such that $B_i = B_j = B_l$. (For assume there are. If $B_j \overset{*}{\underset{G'}{\Rightarrow}} u_jB_jv_ja_jw_j$, then since $B_l = B_j$, a word in $L(G')$ is obtained with a_l to the left of a_j, a contradiction. If $B_j \overset{*}{\underset{G'}{\Rightarrow}} u_ja_jv_jB_jw_j$, then since $B_j = B_i$, a word in $L(G')$ is obtained with a_j to the left of a_i, a contradiction.) Hence at least half of the variables $B_1, \cdots, B_{2k}$ are different, i.e. $N(G') \geqq k$.

To see (*), we shall show that generation in G' of long words requires variables with the given property. Let r be the number of variables in G' and i a given integer, $1 \leqq i \leqq 2k$ Let m be a positive integer such that every derivation tree in G' of a_i^m has a path with at least $r+2$ nodes. Now consider a shortest G'-derivation of a_i^m, and let T be the associated derivation tree Let π be a longest path in T, and let $B_0, \cdots, B_r$ be the first $r+1$ node names on π. Thus $B_0 = S'$. Let $B_j \rightarrow x_{j+1}B_{j+1}y_{j+1}$, $0 \leqq j \leqq r-1$ and all x_{j+1} and y_{j+1} in V_2^*, be the productions in P_2 which realize the first r node names on π. Since the sequence $\{B_j\}_{0 \leqq j \leqq r}$ is of length $r+1$ and G' has only r variables, $B_s = B_t$ for some $s < t$. Then

$$x_{s+1} \cdots x_t y_t \cdots y_{s+1} \overset{*}{\underset{G'}{\Rightarrow}} z_1a_iz_2$$

for some terminal words z_1 and z_2, since T is a tree for a shortest G'-derivation of a_i^m. From this, (*) immediately follows.

5. *Forms for the Linear and the Context-Free Languages*

In this section we examine the complexity situation for forms whose expressive power is exactly the linear languages or exactly the context-free languages. The results for the linear languages parallel those for the regular sets, whereas the context-free languages only permit linear improvement rather than arbitrary polynomial improvement.

Analogous to Lemma 3.3, we have

LEMMA 5.1. *For each grammar form F defining the linear languages, there exists an equivalent form F' and positive integers c, n with the following properties:*

(1) Every production of F' is one of the types $\alpha \to \beta\gamma$, $\alpha \to w\beta$, $\alpha \to \beta w$, $\alpha \to w$, $\alpha \to \beta\alpha\gamma$, $\alpha \to w\alpha\gamma$, $\alpha \to \beta\alpha w$, or $\alpha \to w_1\alpha w_2$, where α, β, γ are variables and w, w_1, w_2, are terminal words;

(2) F' is sequential, reduced, and for every variable α of F', $\alpha \overset{*}{\underset{G_{F'}}{\Rightarrow}} w$ for some nonempty terminal word w; and

(3) for each G in $\mathscr{G}(F)$ there exists an equivalent G' in $\mathscr{G}(F')$ such that $M(G') \leqq c[M(G)]^n$ for each M in $\{S,V,P\}$.

PROOF As in the proof of Lemma 3.3, we obtain F'' from Lemma 3 2. Note that F'' satisfies conditions (2) and (3) We transform F'' into the required F' by the following procedure. Productions of F'' of the eight permitted types are put into F'. Each remaining F'' production whose left-hand variable does not occur on its right-hand side is treated exactly as in Lemma 3.3 The remaining productions of F'' are of the type $\alpha \to x_m \cdots x_1\alpha y_1 \cdots y_n$, where either m or n (or both) is at least 2, α is a variable, each x_i and y_j is either a variable different from α or a nonempty terminal word, and no two consecutive x_i or y_j are terminal words (At this point, [1, Th. 2 4] is used.[10]) If m and n are both at least two, then put into F' the set of productions

$$\{\alpha \to \beta_0\alpha\gamma_0,\ \beta_0 \to \beta_1x_1,\ \beta_1 \to \beta_2x_2,\ \cdots,\ \beta_{m-3} \to \beta_{m-2}x_{m-2},\ \beta_{m-2} \to x_mx_{m-1}\}$$
$$\cup\ \{\gamma_1 \to y_1\gamma_1,\ \gamma_1 \to y_2\gamma_2,\ \cdots,\ \gamma_{n-3} \to y_{n-2}\gamma_{n-2},\ \gamma_{n-2} \to y_{n-1}y_n\},$$

where the β_i, γ_j are new variables, chosen to be different for each new production. If either m or n is at most 1, then make the obvious modification. The rest of the proof is straightforward.

. Similar to Definition 3.4, we define an infinite sequence $\{J_n\}$ of forms for the linear languages.

Definition 5 2. For each $n \geqq 1$, let $J_n = (V,\Sigma,\mathscr{V}_n,\{a\},\mathscr{P}_n,\alpha_0)$, where

$$\mathscr{V}_n = \{a\} \cup \{\alpha_0, \cdots, \alpha_{n-1}, \beta_1, \cdots, \beta_{n-1}, \gamma_1, \cdots, \gamma_{n-1}\}$$

and

$$\begin{aligned}\mathscr{P}_n = {} & \{\alpha_i \to \beta_j\alpha_i\gamma_k,\ \alpha_i \to \beta_j\alpha_ia,\ \alpha_i \to a\alpha_i\gamma_k,\ \alpha_i \to a\alpha_ia \mid 0 \leqq i \leqq n-1,\ 1 \leqq j, k \leqq n-1\}\\ & \cup \{\alpha_i \to \beta_j\gamma_k,\ \alpha_i \to \beta_ja,\ \alpha_i \to a\gamma_k,\ \alpha_i \to a \mid 0 \leqq i \leqq n-1,\ 1 \leqq j, k \leqq n-1\}\\ & \cup \{\alpha_i \to \alpha_j\gamma_k,\ \alpha_i \to \alpha_ja,\ \alpha_i \to \beta_l\alpha_j,\ \alpha_i \to a\alpha_j \mid 0 \leqq i < j \leqq n-1,\ 1 \leqq k, l \leqq n-1\}\\ & \cup \{\beta_i \to \beta_j\beta_k \mid 1 \leqq i \leqq k \leqq n-1,\ i < j\} \cup \{\beta_i \to a\beta_k,\ \beta_i \to \beta_ja \mid 1 \leqq i \leqq k \leqq n-1,\\ & \quad 1 \leqq i < j \leqq n-1\}\\ & \cup \{\beta_i \to a,\ \gamma_i \to a \mid 1 \leqq i \leqq n-1\} \cup \{\gamma_i \to \gamma_j\gamma_k \mid 1 \leqq i \leqq j \leqq n-1,\ i < k\}\\ & \cup \{\gamma_i \to \gamma_ja,\ \gamma_i \to a\gamma_k \mid 1 \leqq i \leqq j \leqq n-1,\ 1 \leqq i \leqq k \leqq n-1\}.\end{aligned}$$

Intuitively, the α's provide n sequential levels of self-embedding variables, and the β's

[10] Theorem 2 4 of [1] is the following. Let $F = (V,\Sigma,\mathscr{V},\mathscr{S},\mathscr{P},\sigma)$ be a reduced grammar form. Call a variable *self-embedding* if there are words u, v in $\mathscr{S}^+$ such that $\xi \overset{*}{\Rightarrow} u\xi v$ Then $\mathscr{L}(F)$ is the family of linear languages if and only if (i) F has a self-embedding variable and (ii) if $\sigma \overset{*}{\Rightarrow} u_1\xi u_2\eta u_3$, with u_1, u_2, u_3 in $\mathscr{V}^*$ and ξ, η in $\mathscr{V} - \mathscr{S}$, then ξ and η are not both self-embedding variables

and γ's are strictly right recursive and strictly left recursive, respectively. In particular, J_1 is standard linear form.

Paralleling Lemma 3.5, we obtain:

LEMMA 5.3. *Let F be a grammar form for the linear languages, having the following conditions:* (i) *Each production of F is one of the types* $\alpha \to \beta\gamma$, $\alpha \to w\beta$, $\alpha \to \beta w$, $\alpha \to w$, $\alpha \to \beta\alpha\gamma$, $\alpha \to w\alpha\gamma$, $\alpha \to \beta\alpha w$, *or* $\alpha \to w_1\alpha w_2$, *where* α, β, γ *are variables and* w, w_1, w_2 *are terminal words; and* (ii) *F is sequential, reduced, and for every variable* α *of F,* $\alpha \overset{*}{\underset{G_F}{\Rightarrow}} w$ *for some nonempty terminal word w. Let n be the number of variables in* G_F.

Then there exists a positive integer c such that for every G in $\mathcal{G}(F)$, *an equivalent G' in* $\mathcal{G}(J_n)$ *can be found satisfying* $M(G') \leqq c[M(G)]^2$ *for M in* $\{S,V,P\}$, *and* $N(G') \leqq c[N(G)]^3$.

The proof involves a complicated construction and is given in Appendix A Roughly speaking, we augment each interpretation grammar of F to contain both a left and a right recursive equivalent of every non-self-embedding variable Since the right-hand side of each production contains at most one self-embedding variable, we may replace each variable to the left (right) of a self-embedding variable with its right (left) recursive equivalent, thereby producing a grammar in form J_m.

The next preliminary result corresponds to Lemma 3.6.

LEMMA 5.4. *For each positive integer n, there exists a positive integer c with the following property: For each G in* $\mathcal{G}(J_n)$, *there exists an equivalent G' in* $\mathcal{G}(J_1)$ *such that* $M(G') \leqq c[M(G)]^{4n}$ *for each M in* $\{S,V,P,N\}$.

The proof is similar to that of Lemma 3.6 and is given in Appendix B. The idea here is that G' simulates "outermost" derivations of G

Combining Lemmas 5.1, 5.3, and 5.4, we obtain:

THEOREM 5.5. *For each grammar form F defining the linear languages, there exist positive integers c and n with the following property: For every G in* $\mathcal{G}(F)$, *there exists an equivalent G' in standard linear form such that* $M(G') \leqq c[M(G)]^n$ *for each M in* $\{S,V,P\}$.

Remarks similar to those following Theorem 3.7 indicate how Theorem 5.5 can be extended to the measure $N(G)$.

The attainability of any polynomial improvement is seen using the family $\{L_{n,k} \mid n \geqq 1, k \geqq 1\}$ of languages defined in Section 3.

As in Lemma 3.9, so (proof omitted) we have

LEMMA 5.6. *For every positive integer n, there is a positive integer c with the following property: For each positive integer k, there is a grammar G in* $\mathcal{G}(J_{n+1})$ *such that* $L(G) = L_{n,k}$, *and* $M(G) \leqq ck$ *for each M in* $\{S,V,P,N\}$.

Parallel to Lemma 3.10, we have

LEMMA 5.7. *For each positive integer s and each grammar G in standard linear form, with* $L(G) = 0^*(10^*)^s$, $M(G) \geqq \lceil (s-1)/2 \rceil + 1$ *for*[11] *each M in* $\{S,V,P,N\}$.

PROOF. It suffices to show that $N(G) \geqq \lceil (s-1)/2 \rceil + 1$. Let $x = 0^m(10^m)^s$, where m is an integer larger than the maximum number of occurrences of terminals in each production of P_1. Let $G = (V,\Sigma_1,P_1,S)$ and

$$\delta\colon\ S = A_0 \Rightarrow w_1A_1x_1 \Rightarrow w_1w_2A_2x_2x_1 \Rightarrow \cdots$$
$$\Rightarrow w_1\cdots w_lA_lx_l\cdots x_1 \Rightarrow w_1\cdots w_lw_{l+1}x_l\cdots x_1 = x$$

be a derivation of x, with each A_i a variable and w_i, x_i in $\{0, 1\}^*$. From the definition of m, each w_t and each x_j can have at most one occurrence of the symbol 1. Let $i_1, \cdots, i_r$ be those indices t for which either w_t or x_t contains 1. Clearly, $r \geqq \lceil (s-1)/2 \rceil$. By an argument as in Lemma 3.10, $A_{i_j-1} \neq A_{i_k-1}$ for $j \neq k$, and A_l is distinct from all A_{i_j}. Hence $N(G) \geqq \lceil (s-1)/2 \rceil + 1$.

[11] $\lceil (s-1)/2 \rceil$ denotes the smallest integer equal to or greater than $(s-1)/2$

Combining the previous lemmas, we get

THEOREM 5.8 *For each positive integer n, there exists a grammar form F for the linear languages and a positive integer c with the following property: For every integer $k \geqq 1$, there is a grammar G in $\mathcal{G}(F)$ such that for each M in $\{S,V,P,N\}$,* (1) *$M(G) \leqq ck$, and* (2) *$M(G') \geqq k^n$ for every equivalent grammar G' in standard linear form.*

The analogy to Corollary 3.12 clearly holds.

Finally, we note that results of the kind obtained for the regular sets and the linear languages do not hold in general. For the case of forms defining all the context-free languages, we can show by a straightforward simulation:

THEOREM 5.9 *For each form F defining the context-free languages, there exists a positive integer c with the following property: For every G in $\mathcal{G}(F)$, there exists an equivalent G' in Chomsky binary form such that $M'(G') \leqq c[M(G)]$ for each M in $\{S,V,P\}$.*

Thus arbitrary polynomial improvement is not possible for the forms defining the context-free languages.

6. *Open Questions*

Many open questions remain, a few of which are now mentioned. Are there results similar to those in Sections 3 and 5 for grammatical families other than those of the regular sets and linear languages? (As yet, there is no "canonical form," analogous to right-linear form, for each grammatical family. Nevertheless, perhaps it can be shown that every two forms with the same expressive power can simulate each other with at most polynomial loss of efficiency.) Do there exist two forms for the regular sets, each of which is more efficient than the other for some languages? What are the tradeoffs between derivation complexity [2] and size complexity? And finally, what can be said about complexity of forms which are not context-free?

Appendix A

We now present a formal proof for Lemma 5 3.

Consider those variables α of G_F such that $\alpha \overset{*}{\underset{G_F}{\Rightarrow}} w_1\beta w_2 \overset{*}{\underset{G_F}{\Rightarrow}} w_1w_3\beta w_4w_2$ for some variable β, words w_1, w_2 in Σ^*, and words w_3, w_4 in Σ^+. Denote these variables, in sequential order, by $\alpha_0, \cdots, \alpha_m$, where $m \leqq n - 1$ Since F defines the linear languages, $m \geqq 0$. Clearly $F = (V,\Sigma,\mathcal{V},\mathcal{S},\mathcal{P},\alpha_0)$. Denote the remaining variables of G_F, in sequential order, by $\beta_1, \beta_2, \cdots, \beta_{n-1-m}$

Let (μ, G) be an arbitrary interpretation of F, with $G = (V_1,\Sigma_1,P_1,S)$. There is no loss of generality in assuming G is reduced. We shall define an interpretation (μ',G') of J_n, with $G' = (V_2,\Sigma_1,P_2,S)$ The variables of G' are: (α) each variable of G in $\mu(\alpha_i)$ for each i, (β) distinct symbols A_L and A_R for each variable A of G in $\bigcup_{i=1}^{n-1} \mu(\beta_i)$; and ($\gamma$) distinct symbols $[B_LA_L]$, $[C_LB_LA_L]_1$, $[C_LB_LA_L]_2$, $[C_LB_LA_L]_3$, $[A_RB_R]$, $[A_RB_RC_R]_1$, $[A_RB_RC_R]_2$, and $[A_RB_RC_R]_3$ for all variables A_L, B_L, C_L, A_R, B_R, C_R. The set P_2 consists of the following productions (where w,w_1,w_2 are inΣ_1^*, $0 \leqq i,i_1,i_2 \leqq n - 1$, and $1 \leqq j,j_1,j_2 \leqq n - 1$).

(a) $A \rightarrow B_LCD_R$ if A,C are in $\mu(\alpha_i)$, B is in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, and $A \rightarrow BCD$ is in P_1.

(b) $A \rightarrow B_LCw$ if A,C are in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, and $A \rightarrow BCw$ is in P_1.

(c) $A \rightarrow wCD_R$ if A,C are in $\mu(\alpha_i)$, D is in $\mu(\beta_j)$, and $A \rightarrow wCD$ is in P_1.

(d) $A \rightarrow w_1Cw_2$ if A,C are in $\mu(\alpha_i)$ and $A \rightarrow w_1Cw_2$ is in P_1.

(e) $A \rightarrow B_LC_R$ if A is in $\mu(\alpha_i)$, B is in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, and $A \rightarrow BC$ is in P_1.

(f) $A \rightarrow B_Lw$ if A is in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, and $A \rightarrow Bw$ is in P_1.

(g) $A \to wC_R$ if A is in $\mu(\alpha_i)$, C is in $\mu(\beta_j)$, and $A \to wC$ is in P_1.

(h) $A \to w$ if A is in $\mu(\alpha_i)$ and $A \to w$ is in P_1.

(i) $A \to BC_R$ if A is in $\mu(\alpha_{i_1})$, B is in $\mu(\alpha_{i_2})$, C is in $\mu(\beta_j)$, $i_1 < i_2$, and $A \to BC$ is in P_1.

(j) $A \to Bw$ if A is in $\mu(\alpha_{i_1})$, B is in $\mu(\alpha_{i_2})$, $i_1 < i_2$, and $A \to Bw$ is in P_1.

(k) $A \to B_LC$ if A is in $\mu(\alpha_{i_1})$, C is in $\mu(\alpha_{i_2})$, B is in $\mu(\beta_j)$, $i_1 < i_2$, and $A \to BC$ is in P_1.

(l) $A \to wB$ if A is in $\mu(\alpha_{i_1})$, B is in $\mu(\alpha_{i_2})$, $i_1 < i_2$, and $A \to wB$ is in P_1.

(m) $B_L \to C_LD_L$ if B is in $\mu(\beta_j)$, C is in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j \leqq j_2$, $j < j_1$, and $B \to CD$ is in P_1.

(n) $B_L \to wC_L$ if B is in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 \leqq j_2$, and $B \to wC$ is in P_1.

(o) $B_L \to C_Lw$ if B is in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $B \to Cw$ is in P_1.

(p) $B_L \to w$ and $B_R \to w$ if B is in $\mu(\beta_j)$ and $B \to w$ is in P_1.

(q) $B_R \to C_RD_R$ if B is in $\mu(\beta_j)$, C is in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j \leqq j_1$, $j < j_2$, and $B \to CD$ is in P_1

(r) $B_R \to wC_R$ if B is in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $B \to wC$ is in P_1.

(s) $B_R \to C_Rw$ if B is in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 \leqq j_2$, and $B \to Cw$ is in P_1.

(t) $A_L \to w[B_LA_L]$ and $A_R \to [A_RB_R]w$ if A,B are in $\mu(\beta_j)$ and $B \to w$ is in P_1.

(u) $A_L \to w[C_LB_LA_L]_1$, $[C_LB_LA_L]_1 \to C_L[B_LA_L]$, $A_R \to [A_RB_RC_R]_1C_R$, and $[A_RB_RC_R]_1 \to [A_RB_R]w$ if A,B are in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $B \to wC$ is in P_1.

(v) $A_L \to C_L[C_LB_LA_L]_2$, $[C_LB_LA_L]_2 \to w[B_LA_L]$, $A_R \to [A_RB_RC_R]_2w$, and $[A_RB_RC_R]_2 \to [A_RB_R]C_R$ if A,B are in $\mu(\beta_{j_1})$, C is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $B \to Cw$ is in P_1.

(w) $A_L \to C_L[C_LB_LA_L]_3$ and $[C_LB_LA_L]_3 \to D_L[B_LA_L]$ if A,B are in $\mu(\beta_j)$, C is in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j < j_1, j < j_2$, and $B \to CD$ is in P_1.

(x) $A_R \to [A_RB_RC_R]_3C_R$ and $[A_RB_RC_R]_3 \to [A_RB_R]D_R$ if A,B are in $\mu(\beta_j)$, C is in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j < j_1$, $j < j_2$, and $B \to DC$ is in P_1.

(y) $[B_LA_L] \to w[C_LA_L]$ if A,B,C are in $\mu(\beta_j)$ and $C \to Bw$ is in P_1

(z) $[B_LA_L] \to D_L[C_LA_L]$ if A,B,C are in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $C \to BD$ is in P_1.

(a′) $[A_LA_L] \to \epsilon$ and $[A_RA_R] \to \epsilon$ for each variable A of G.

(b′) $[A_RB_R] \to [A_RC_R]w$ if A,B,C are in $\mu(\beta_i)$ and $C \to wB$ is in P_1.

(c′) $[A_RB_R] \to [A_RC_R]D_R$ if A,B,C are in $\mu(\beta_{j_1})$, D is in $\mu(\beta_{j_2})$, $j_1 < j_2$, and $C \to DB$ is in P_1.

The substitution μ' is defined as follows. Let $\mu'(a)$ contain ϵ and every terminal word occurring in at least one production in P_2. For each variable A in $\mu(\alpha_i)$, let A be in $\mu'(\alpha_i)$. For each variable A in $\mu(\beta_j)$, let A_L, $[B_LA_L]$, $[C_LB_LA_L]_1$, $[C_LB_LA_L]_2$, and $[C_LB_LA_L]_3$ be in $\mu'(\beta_j)$, and A_R, $[A_RB_R]$, $[A_RB_RC_R]_1$, $[A_RB_RC_R]_2$, and $[A_RB_RC_R]_3$ be in $\mu'(\gamma_j)$.

Clearly, the size conditions are satisfied. That $L(G') = L(G)$ follows in a similar way to Lemma 3.5 Derivations in G' proceed as in G, except that certain variables are "reversed." In particular, variables to the left of a self-embedding variable[12] embed themselves on the right only, while variables to the right of a self-embedding variable embed themselves on the left only. Since a variable A of G might occur on *both* sides of a self-embedding variable, two copies of A, A_L for the left and A_R for the right, are introduced. (A_L embeds itself only on the right and A_R only on the left.) A formal argument along the lines of that in Lemma 3.5 is left to the reader

Appendix B

Here we establish Lemma 5 4. Let (μ,G) be an interpretation of J_n, with G =

[12] A variable ξ in a grammar $G = (V_1,\Sigma_1,P_1,S)$ is called *self-embedding* if there are words u and v in Σ_1^+ such that $\xi \overset{*}{\Rightarrow} u\xi v$

(V_1,Σ_1,P_1,S). We now define an interpretation (μ',G') of J_1 in which $G' = (V_2,\Sigma_1,P_2,[S])$ simulates "outermost" derivations of G

The set $V_2 - \Sigma_1$ consists of the symbols:

(1) $[B_r \cdots B_1 A C_1 \cdots C_s]$, where $0 \leqq r,\ s \leqq n-1$, A is in $\mu(\alpha_i)$ for some i, each B_j is either a variable in $\mu(\beta_l)$ for some l or else a pair $\langle D_j,E_j\rangle$ of variables D_j in $\mu(\beta_{k_1})$ and E_j in $\mu(\beta_{k_2})$ for some k_1,k_2, and each C_j is either a variable in $\mu(\gamma_l)$ for some l or else a pair $\langle D_j',E_j'\rangle$ of variables D_j' in $\mu(\gamma_{k_1})$ and E_j' in $\mu(\gamma_{k_2})$ for some k_1,k_2.

(2) $[B_r \cdots B_1 C_1 \cdots C_s]$, where everything is as in (1) except there is no variable A.

The set P_2 consists of the following productions (where w,w_1,w_2 are words in Σ_1^*, and i,j,k,r,s are integers whose quantification will be clear in each case):

(a) $[A] \to [BCD]$ if A, C are in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, C is in $\mu(\gamma_k)$, and $A \to BCD$ is in P_1.

(b) $[A] \to [BC]w$ if A, C are in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, and $A \to BCw$ is in P_1.

(c) $[A] \to w[CD]$ if A, C are in $\mu(\alpha_i)$, D is in $\mu(\gamma_j)$, and $A \to wCD$ is in P_1.

(d) $[A] \to w_1[C]w_2$ if A, C are in $\mu(\alpha_i)$ and $A \to w_1Cw_2$ is in P_1.

(e) $[A] \to [BC]$ if A is in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, C is in $\mu(\gamma_k)$, and $A \to BC$ is in P_1.

(f) $[A] \to [B]w$ if A is in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, and $A \to Bw$ is in P_1.

(g) $[A] \to w[B]$ if A is in $\mu(\alpha_i)$, B is in $\mu(\gamma_j)$, and $A \to wB$ is in P_1.

(h) $[A] \to w$ if A is in $\mu(\alpha_i)$ and $A \to w$ is in P_1.

(i) $[A] \to [BC]$ if A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, $i < j$, C is in $\mu(\gamma_k)$, and $A \to BC$ is in P_1.

(j) $[A] \to [B]w$ if A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, $i < j$, and $A \to Bw$ is in P_1.

(k) $[A] \to [BC]$ if A is in $\mu(\alpha_i)$, B is in $\mu(\beta_j)$, C is in $\mu(\alpha_k)$, $i < k$, and $A \to BC$ is in P_1.

(l) $[A] \to w[B]$ if A is in $\mu(\alpha_i)$, B is in $\mu(\alpha_j)$, $i < j$, and $A \to wB$ is in P_1.

(m) $[B_r \cdots B_1 A C_1 \cdots C_s] \to [B_{r+2}B_{r+1}B_{r-1} \cdots B_1 A C_1 \cdots C_s]$ if B_r is in $\mu(\beta_i)$, B_{r+2} is in $\mu(\beta_j)$, $i < j$, B_{r+1} is in $\mu(\beta_k)$, $i \leqq k$, $B_r \to B_{r+2}B_{r+1}$ is in P_1, and the remaining symbols are as in (1).

(n) $[B_r \cdots B_1 C_1 \cdots C_s] \to [B_{r+2}B_{r+1}B_{r-1} \cdots B_1 C_1 \cdots C_s]$, with the quantification as in (m) and symbols as in (2).

(o) $[B_r \cdots B_1 A C_1 \cdots C_s] \to w[B_{r+1}B_{r-1} \cdots B_1 A C_1 \cdots C_s]$ if B_k is in $\mu(\beta_i)$, B_{r+1} is in $\mu(\beta_j)$, $i \leqq j$, $B_r \to wB_{r+1}$ is in P_1, and the remaining symbols as in (1).

(p) $[B_r \cdots B_1 C_1 \cdots C_s] \to w[B_{r+1}B_{r-1} \cdots B_1 C_1 \cdots C_s]$, with the quantification as in (o), symbols as in (2).

(q) $[B_r \cdots B_1 A C_1 \cdots C_s] \to w[B_{r-1} \cdots B_1 A C_1 \cdots C_s]$ if B_r is in $\mu(\beta_i)$, $B_r \to w$ is in P_1, and the remaining symbols are as in (1)

(r) $[B_r \cdots B_1 C_1 \cdots C_s] \to w[B_{r-1} \cdots B_1 C_1 \cdots C_s]$, with the quantification as in (q) and symbols as in (2).

(s) $[B_r \cdots B_1 A C_1 \cdots C_s] \to [D\langle DB_r\rangle B_{r-1} \cdots B_1 A C_1 \cdots C_s]$ and $[\langle DB_r\rangle B_{r-1} \cdots B_1 A C_1 \cdots C_s] \to w[B_{r-1} \cdots B_1 A C_1 \cdots C_s]$ if B_r is in $\mu(\beta_i)$, D is in $\mu(\beta_j)$, $i < j$, $B_r \to Dw$ is in P_1, and the remaining variables are as in (1).

(t) The same as (s), with the variable A omitted (symbols as in (2)).

(u) Symmetric versions of (m)-(t), expanding the C variables.

(v) $[\,] \to \epsilon$.

The argument that G' has the desired properties is similar to that in Lemma 3.6, and is omitted.

REFERENCES

1 Cremers, A B , and Ginsburg, S Context-free grammar forms *J Comput Syst Sci 11* (1975), 86-117

2. GINSBURG, S , AND LYNCH, N Derivation complexity in context-free grammar forms *SIAM J Comput.* (to appear).
3 GRUSKA, J. On the size of context-free grammars *Kybernetica 8* (1972), 213-218.
4 MEYER, A R., AND FISCHER, M J. Economy of description by automata, grammars, and formal systems. 12th Annual Symp. on Switching and Automata Theory, 1971, pp 188-191

RECEIVED JUNE 1975, REVISED MARCH 1976