

## CHAPTER IV

### OTHER TOPICS

Our main concern in this book has been the study of the context-free case of form theory, namely context-free grammar, EOL and ETOL forms. However the non-context-free case of form theory has also been investigated. Therefore in this chapter we aim to give the reader a flavor of the results to be found in this area. In Section IV.1 phrase structure grammar forms are discussed and in Section IV.2 context dependent L forms. In Section IV.3 matrix forms are introduced and finally in Section IV.4 some results on controlled grammar forms are surveyed.

Where appropriate open problems are mentioned as well as some of the main results. However only sketch proofs are given, since our aim is to be discursive rather than formal.

#### IV.1 Phrase Structure Grammar Forms

Since the notion of a form can be applied to any class of re-writing systems, one obvious class to investigate is that of (phrase structure) grammar forms. A phrase structure grammar is a quadruple  $G = (V, \Sigma, P, S)$ , where  $P \subseteq (V - \Sigma)^+ \times V^*$  and  $P$  is finite, together with a modified version of the sequential rewrite relation,  $\Rightarrow$ . The notions of g- and s-interpretations clearly carry over to this situation without any difficulty since the left hand side of a production consists solely of nonterminals. One relatively longstanding problem has been:

(\*) Exhibit a (phrase structure) grammar form  $G$  such that

$$\mathcal{L}(CF) \not\subseteq \mathcal{L}_g(G, \Rightarrow) \not\subseteq \mathcal{L}(RE).$$

As we shall now demonstrate there are seemingly "simple" grammar forms which give rise to  $\mathcal{L}(RE)$  under g-interpretation. Furthermore we give a proof sketch of the assertion that given a two-symbol grammar form  $G$  for which  $\mathcal{L}_g(G, \Rightarrow) \not\subseteq \mathcal{L}(CF)$  then  $\mathcal{L}_g(G, \Rightarrow) = \mathcal{L}(RE)$ . This leads us to conjecture that this result is true in general, namely there is no grammar form satisfying (\*).

For our first example, consider  $G_1$  defined by the productions:

$$S \rightarrow ABSc; S \rightarrow ABc; BA \rightarrow AB; A \rightarrow a; B \rightarrow b.$$

Now  $L(G_1, \Rightarrow) \cap a^+b^+c^+ = \{a^n b^n c^n : n \geq 1\}$ , which is not context-free hence  $L(G_1, \Rightarrow)$  is not context-free. Furthermore  $G_1$  is a permuting grammar, that is, the only non-context-free production  $BA \rightarrow AB$  is a permuting production. It is well known that permuting grammars only generate languages whose Parikh set is semi-linear, hence one might suspect that  $\mathcal{L}_g(G_1, \Rightarrow) \not\subseteq \mathcal{L}(RE)$ .

However it can be proved that  $\mathcal{L}_g(G_1, \Rightarrow) = \mathcal{L}(RE)$ . Essentially we obtain erasing productions as interpretations of  $A \rightarrow a$  and  $B \rightarrow b$  which together with interpretations of  $BA \rightarrow AB$  yields enough context-sensitivity to give the result.

Consider the following interpretation  $G'_1$  and  $G_1$  which gives rise to a Parikh set which is not semi-linear:

$$S_0 \rightarrow A_\lambda B' S_1; S_1 \rightarrow A_\lambda B_0 S_1; S_1 \rightarrow A_1 B_\lambda S_2;$$

$$S_2 \rightarrow A_1 B_\lambda S_2; S_2 \rightarrow A_1 B_\lambda;$$

$$B_1 A_1 \rightarrow A_1 B_0; B_0 A_1 \rightarrow A_2 B_1; B_0 A_2 \rightarrow A_2 B_0;$$

$$B_1 A_2 \rightarrow A_2 B_1; B' A_2 \rightarrow A_0 B'; B' A_1 \rightarrow A_0 B_\lambda;$$

$$A_\lambda \rightarrow \lambda; B_\lambda \rightarrow \lambda; A_0 \rightarrow a; B_0 \rightarrow b$$

Now using the first three productions we obtain

$$S_0 \Rightarrow A_\lambda B' S_1 \Rightarrow^n A_\lambda B' (A_\lambda B_0)^n S_1 \Rightarrow A_\lambda B' (A_\lambda B_0)^n A_1 B_\lambda S_2$$

If we now erase the appearances of  $A_\lambda$  and  $B_\lambda$  we obtain

$$B' B_0^n A_1 S_2$$

and if we now replace  $S_2$  by  $(A_1 B_\lambda)^m$  using the fourth and fifth productions and again erasing  $B_\lambda$  we obtain:

$$B' B_0^n A_1^{m+1}$$

Now using the "permuting" productions an  $A_1$  can move left through  $B' B_0^n$  changing it to  $B_1' B_0^{n-1} B_1$  and emitting  $A_0$  at the left, that is

$$B' B_0^n A_1^{m+1} \Rightarrow^* A_0 B' B_0^{n-1} B_1 A_1^m.$$

Again an  $A_1$  can move left emitting an  $A_0$  giving  $A_0 A_0 B' B_0^{n-2} B_1 B_0 A_1^{m-1}$ , and so on. Note that the  $n$  positions given by the subscripted  $B$ 's define a binary number, initially 0. When an  $A_1$  passes through from right-to-left a 1 bit addition is simulated. Hence when  $2^n A_1$ 's have passed through the  $B'$  is erased giving

$$a^{2^n} b^n$$

when the  $A_0$  and  $B_0$  symbols are replaced. Clearly  $L(G_1, \Rightarrow)$  gives rise to a non-semi-linear Parikh set.

As a second example the use of erasing is more carefully controlled than in  $G_1$ . Thus we have  $G_2$  defined by:

$$\begin{aligned} S &\rightarrow DS_1; \\ S_1 &\rightarrow ABS_1c; S_1 \rightarrow ABDc; \\ BA &\rightarrow AB; \\ DA &\rightarrow aD; Db \rightarrow bD_b \\ D_b B &\rightarrow bD_b; D_b D \rightarrow \lambda \end{aligned}$$

In this grammar  $D$  forces the permuting production  $BA \rightarrow AB$  to be applied "completely" before any erasing takes place. This follows by observing that  $D$  becomes  $D_b$  on meeting the first  $B$  and  $D_b$  is blocked by any appearance of  $A$  to its right. This also holds true for interpretations of  $G_2$  as well. However even in this case  $\mathcal{L}_g(G_2, \Rightarrow) = \mathcal{L}(RE)$ . We leave to the interested reader the detailed proofs of these equalities for  $G_1$  and  $G_2$ .

Let us turn to the consideration of two-symbol forms. Note that under  $g$ -interpretation all terminals in a grammar form  $G = (V, \Sigma, P, S)$  can be identified with  $a$  without any loss of generality. Moreover we may assume that no production in  $G$  contains two consecutive  $a$ 's and also we may replace  $G$  by its full interpretation  $G'$  under the sub-

stitution  $\mu$  defined by  $\mu(a) = \{a, \lambda\}$  and  $\mu(A) = A$  for all  $A$  in  $V - \{a\}$ . Since  $G' \triangleleft_g G$  and  $G \triangleleft_g G'$  we have  $\mathcal{L}_g(G', \Rightarrow) = \mathcal{L}_g(G, \Rightarrow)$ . When  $G = G'$  and  $G$  also fulfills the above conditions we say  $G$  is terminally-closed. Letting  $G = (\{S, a\}, \{a\}, P, S)$  be terminally-closed it follows that  $S \Rightarrow^* a^i$ ,  $i \geq 0$  implies  $S \Rightarrow^* \lambda$ . If  $G$  does not contain a production  $S \rightarrow S^i$  for  $i > 1$  then  $\mathcal{L}_g(G, \Rightarrow) \subseteq \mathcal{L}(CF)$ , since productions with at least two  $S$  symbols on the left hand side can never be applied in a sentential derivation. Similarly if there is no production  $S^k \rightarrow S^\ell$  in  $G$  for  $k > 1$  then again  $\mathcal{L}_g(G, \Rightarrow) \subseteq \mathcal{L}(CF)$  since no non-context-free production is in  $G$ .

Since we can always transform a two-symbol grammar form into a  $g$ -form equivalent terminally-closed grammar form we only consider such forms in the following.

We now sketch the proof that for a two-symbol grammar form  $G$  either  $\mathcal{L}_g(G, \Rightarrow) \subseteq \mathcal{L}(CF)$  or  $\mathcal{L}_g(G, \Rightarrow) = \mathcal{L}(RE)$ . Assume there is a production  $S \rightarrow S^i$ , for  $i > 1$  in  $G$  and a production  $S^k \rightarrow S^\ell$  in  $G$  with  $k > 1$ , otherwise  $\mathcal{L}_g(G, \Rightarrow)$  is sub-context-free. Now there must be a production containing an  $a$  otherwise  $L(G, \Rightarrow)$  equals  $\emptyset$  and  $\mathcal{L}_g(G, \Rightarrow)$  is sub-context-free. Let this production be  $S^m \rightarrow S^r a S^t$ , for some  $m \geq 1$ ,  $r$  and  $t \geq 0$ . Assuming  $G$  is nonempty we must have a derivation

$$S \Rightarrow^* \lambda$$

in which case we may assume  $S \rightarrow \lambda$  is in  $G$ . Since  $S \rightarrow S^i$  is in  $G$  for some  $i > 1$ , then we have a derivation:

$$S \Rightarrow S^i \Rightarrow S^{i-1} \Rightarrow \dots \Rightarrow SS$$

in  $G$  and hence we may assume  $S \rightarrow SS$  is in  $G$ . Again since  $S^k \rightarrow S^\ell$  is in  $G$  for  $k > 1$  we have

$$S^k \Rightarrow S^\ell \Rightarrow \dots \Rightarrow \lambda$$

in  $G$  and hence we may also assume  $S^k \rightarrow \lambda$  is in  $G$ . Finally since  $S^m \rightarrow S^r a S^t$  is in  $G$  we also have the derivation

$$S \Rightarrow SS \Rightarrow \dots \Rightarrow S^m \Rightarrow S^r a S^t \Rightarrow \dots \Rightarrow a$$

in  $G$  and therefore we may again assume  $S \rightarrow a$  is in  $G$ .

It only remains to demonstrate that for  $F_k \triangleleft_g G$ , where  $F_k$  is defined by the productions:

$$S \rightarrow SS; S \rightarrow a; S^k \rightarrow \lambda$$

we have  $\mathcal{L}_g(F_k, \Rightarrow) = \mathcal{L}(RE)$  to give the result.

It is well known that  $F_2$  has this property, therefore we reduce the case  $k > 2$  by simulation to the case  $k=2$ . For each  $F_2' \triangleleft_g F_2$  the productions of the type  $AB \rightarrow \lambda$  are simulated by the derivation

$$A \Rightarrow A_1 C_1 \Rightarrow A_1 A_2 C_2 \Rightarrow \dots \Rightarrow A_1 A_2 \dots A_{k-2} C_{k-2} B \Rightarrow \lambda$$

in an interpretation  $F_k'$  of  $F_k$ , where all other productions are carried over unchanged.

As pointed out above this result gives strong evidence for the validity of the general conjecture, that under g-interpretations no g-grammatical family lies strictly between  $\mathcal{L}(CF)$  and  $\mathcal{L}(RE)$ .

However under s-interpretations as one might expect such families do exist.

Consider  $G_3 = (V_3, \{a\}, P_3, S_3)$  and  $G_4 = (\{S_4, a\}, \{a\}, P_4, S_4)$  where  $L(G_3, \Rightarrow) = \{a^{2^n} : n \geq 1\}$ ,  $P_4 = \{S_4 \rightarrow S_4^2; S_4 \rightarrow a\}$  and  $S_4$  is not in  $V_3$ . Note that  $\mathcal{L}_s(G_4, \Rightarrow) = \mathcal{L}(CF)$ . Construct  $G = (V_3 \cup \{S, S_4\}, \{a\}, P_3 \cup P_4 \cup \{S \rightarrow aS_3; S \rightarrow S_4\}, S)$  where  $S$  is a new nonterminal symbol. Now evidently  $\mathcal{L}(CF) \subseteq \mathcal{L}_s(G, \Rightarrow)$  and since  $\{a\}L(G_3, \Rightarrow)$  is in  $\mathcal{L}_s(G, \Rightarrow)$  we have proper containment. Finally note that  $L(G_3, \Rightarrow)$  is not in  $\mathcal{L}_s(G, \Rightarrow)$ , since taking interpretations of  $S \rightarrow aS_3$  only leads to odd length words and interpretations of  $S \rightarrow S_4$  yield only context-free languages.

This difference between g- and s-interpretations is highlighted by the results we will now discuss. First we will show that there are grammar forms which under s-interpretation generate languages which are either finite and hence context-free or infinite and non-context free. This kind of gap theorem does not appear to hold under g-interpretation. For the particular case of context-free grammar forms, a form can be exhibited which generates either finite languages or infinite languages which are non-regular under s-interpretations, whereas under g-interpretations no such gap exists.

Define  $G_5$  by the following productions:

- $S_5 \rightarrow FZ; Z \rightarrow MZ; Z \rightarrow MCCC D;$
- $MC \rightarrow CAM; MA \rightarrow AM; MD \rightarrow D;$
- $FC \rightarrow cF; FA \rightarrow aF; FD \rightarrow \lambda;$

then a typical derivation in  $G_5$  has the appearance of

$$S_5 \Rightarrow FZ \Rightarrow^* FM^{n-1}Z \Rightarrow FM^n C C C D \Rightarrow^* FM^{n-1} C A C A C A D \\ \Rightarrow^* FM^{n-1} C A^i C A^i C A^i D \Rightarrow^* F C A^n C A^n C A^n D \Rightarrow^* c a^n c a^n c a^n$$

Consider a  $G_5^i \triangleleft_s G_5(\mu)$  with  $L(G_5^i, \Rightarrow)$  infinite and a word  $x$  in  $L(G_5^i, \Rightarrow)$ . A derivation for  $x$  can be written as:

$$S_5^i \Rightarrow F_1 Z_1 \Rightarrow^* F_1 M_1 \dots M_{n-1} Z_2 \Rightarrow F_1 M_1 \dots M_n C_1 C_2 C_3 D_1 \Rightarrow^* \\ F_1 C_4 A_1 \dots A_n C_5 A_{n+1} \dots A_{2n} C_6 A_{2n+1} \dots A_{3n} D_2 \Rightarrow^* \\ c_1 a_1 \dots a_n c_n a_{n+1} \dots a_{2n} c_3 a_{2n+1} \dots a_{3n} = x,$$

where  $S_5^i$  is in  $\mu(S_5)$ ,  $F_i$  is in  $\mu(F)$ ,  $M_i$  is in  $\mu(M)$  and similiary for the other symbols.

Now assume  $L(G_5^i, \Rightarrow) \subseteq \mathcal{L}(CF)$ , then  $L = \mu^{-1}(L(G_5^i, \Rightarrow)) \subseteq \mathcal{L}(CF)$ . However  $L$  is an infinite subset of  $\{c a^i c a^i c a^i : i \geq 1\}$  and any such subset is

known to be non-context free. Hence  $G_5$  possesses the "gap" property we mentioned above.

We now turn our attention to the second result which also displays the difference between g- and s-interpretations. Namely we exhibit a non-context-free grammar form whose language family is  $\mathcal{L}(EOL)$  under a restricted type of s-interpretation.

Assume in the following that a non-context-free grammar is less restrictive than above, namely for  $G = (V, \Sigma, P, S)$  we have  $P \subseteq V^*(V - \Sigma)V^* \times V^*$ , that is terminals may appear on the left hand side, but only in the presence of at least one nonterminal.

We first introduce the notion of smoothness of a grammar. Let  $G = (V, \Sigma, P, S)$  be a grammar and  $m > 0$  an integer, then we say  $G$  is  $m$ -smooth if all productions in  $P$  have left hand sides of length at most  $m$  and for all words  $\alpha$  over  $V$  containing one nonterminal and whose length is at most  $m$  there is at least one production  $\alpha \rightarrow \beta$  in  $P$  for some  $\beta$ .

This condition has something of the flavor of the completeness condition for L grammars. When specifying such an  $m$ -smooth grammar we assume all unspecified left hand sides  $\alpha$  have the identity production  $\alpha \rightarrow \alpha$ . Context-free grammars are essentially always 1-smooth, just add  $A \rightarrow A$  for every nonterminal. We now restrict interpretations in the following way:

If  $G$  is an  $m$ -smooth grammar form then  $G' \overset{\Delta}{\underset{SS}{S}} G$   
 iff  $G'$  is  $m$ -smooth and  $G' \overset{\Delta}{\underset{S}{S}} G$ , where " $\overset{\Delta}{\underset{SS}{S}}$ "  
 denotes "smooth s-interpretation".

To see that  $ss$ -interpretation is different from s-interpretation for non-context-free grammar forms, consider the following example grammar:

F:  $S \rightarrow AA; AA \rightarrow a; A \rightarrow a$

It is clear that  $\mathcal{L}_S(F, \Rightarrow) \subsetneq \mathcal{L}(FIN)$  and each language in  $\mathcal{L}_S(F, \Rightarrow)$  consists of words of length one and two. Hence  $\{a\}$  and  $\{aa\}$  are example languages in  $\mathcal{L}_S(F, \Rightarrow)$ .

If we assume  $F$  is 2-smooth then by our convention  $SS \rightarrow SS; SA \rightarrow SA; AS \rightarrow AS; Sa \rightarrow Sa; aS \rightarrow aS; Aa \rightarrow Aa; aA \rightarrow aA$ ; are in  $F$ . Clearly these productions neither affect the language of  $F$  nor the language family of  $F$ . Now consider  $\mathcal{L}_{SS}(F, \Rightarrow)$ . Each  $F' \overset{\Delta}{\underset{SS}{S}} F(\mu)$  must be 2-smooth hence apart from productions of the type:

$S_1 \rightarrow A_1 A_2$

it must contain productions

$A_1 \rightarrow \alpha_1; A_2 \rightarrow \alpha_2; A_1 A_2 \rightarrow \alpha_3$

for some  $\alpha_1, \alpha_2$  and  $\alpha_3$ .

This means that  $\alpha_1$  is a terminal symbol since  $A_1 \rightarrow \alpha_1$  is in  $\mu(A \rightarrow a)$ ,  $\alpha_2$  is a terminal symbol and finally since  $A_1 A_2 \rightarrow \alpha_3$  is in  $\mu(AA \rightarrow a)$ ,  $\alpha_3$  must also be a terminal symbol. Therefore  $L(F', \Rightarrow)$  contains a word of length one and a word of length two. This is true for all  $F' \stackrel{\Delta}{\underset{SS}{\triangleleft}} F$  hence  $\mathcal{L}_{SS}(F, \Rightarrow) \neq \mathcal{L}_S(F, \Rightarrow)$  and in fact  $\mathcal{L}_{SS}(F, \Rightarrow) \not\subseteq \mathcal{L}_S(F, \Rightarrow)$ .

However if we add:

$$S \rightarrow S; AA \rightarrow AA; A \rightarrow A$$

to  $F$  giving  $G$ , then in this case we do have equality, namely

$$\mathcal{L}_{SS}(G, \Rightarrow) = \mathcal{L}_S(G, \Rightarrow).$$

We next consider uniform interpretations of grammar forms, before discussing our main result.

Let  $G' = (V', \Sigma', P', S')$  and  $G = (V, \Sigma, P, S)$  be two grammar forms such that  $G' \stackrel{\Delta}{\underset{S}{\triangleleft}} G(\mu)$  for some  $\mu$ . We say  $G'$  is a uniform interpretation of  $G$  modulo  $\mu$  if  $P' \subseteq \mu_u(P)$  where  $\mu_u(P) \subseteq \mu(P)$  and  $\alpha' \rightarrow \beta'$  is in  $\mu_u(P)$  iff  $\alpha' = X_1 \dots X_m$ ,  $\beta' = Y_1 \dots Y_n$  and for all  $X_i, Y_j$ ,  $1 \leq i \leq m$ ,  $1 \leq j \leq n$ ,  $\mu^{-1}(X_i) = \mu^{-1}(Y_j)$  implies  $X_i = Y_j$ . We denote this by  $G' \stackrel{\Delta}{\underset{u}{\triangleleft}} G(\mu)$ .

To gain some insight into uniform interpretation consider the grammar form  $G_6$  defined by the productions:

$$S \rightarrow aS; S \rightarrow \lambda.$$

We claim that each non-empty language  $L$  in  $\mathcal{L}_u(G_6, \Rightarrow)$  can be expressed as  $L = \Sigma^*$ , for some  $\Sigma$ .

Consider an arbitrary  $u$ -interpretation  $G'_6 \stackrel{\Delta}{\underset{u}{\triangleleft}} G_6(\mu)$ , where  $G'_6 = (V, \Delta, P, S')$ . Then each production for  $S'$  is one of the types

$$S' \rightarrow a_i S''; S' \rightarrow \lambda.$$

But under  $u$ -interpretation  $S' \rightarrow a_i S''$  is in  $\mu_u(S \rightarrow aS)$  only if  $S' = S''$ . Hence  $G'_6$  only has "useful" productions

$$S' \rightarrow a_i S' \text{ and } S' \rightarrow \lambda.$$

If either  $S' \rightarrow \lambda$  is absent or  $S' \rightarrow a_i S'$  is not in  $P$  for any  $a_i$  then  $L(G'_6, \Rightarrow) = \emptyset$ . This leaves the case that  $S' \rightarrow \lambda$  is present and  $S' \rightarrow a_i S'$  is present for some  $a_i$  in  $\Delta$ . Let  $\Sigma \subseteq \Delta$  be those  $a_i$ 's for which there is a production  $S' \rightarrow a_i S'$ . Clearly  $L(G'_6, \Rightarrow) = \Sigma^*$  as claimed.

Let  $G$  be the grammar form specified by the productions:

- (a)  $S \rightarrow ALZR; AL \rightarrow AAL; AL \rightarrow A;$
- (b)  $AZ \rightarrow A\bar{Z}; A\bar{Z} \rightarrow ZA; A\bar{Z} \rightarrow ZZA; A\bar{Z} \rightarrow aA;$
- (c)  $AR \rightarrow R; AR \rightarrow \lambda.$

We claim that  $\mathcal{L}_u(G, \Rightarrow) = \mathcal{L}(EOL)$ .

It is easy to see that  $\mathcal{L}(EOL) \subseteq \mathcal{L}_u(G, \Rightarrow)$  since the  $A$  can be looked upon as an "activation messenger", which sweeps across a word consisting of interpretations of  $Z$  symbols enforcing rewriting of

each symbol and hence simulating a parallel rewriting step. Clearly we also have  $\mathcal{L}(EOL) \subseteq \mathcal{L}_S(G, \Rightarrow)$ ; a weaker result.

Let us look at this in slightly more detail. Consider an arbitrary EOL grammar  $F = (V, \Sigma, P, Z)$  which is binary, propagating and synchronized. Then we can construct a  $G' \triangleleft_u G$  as follows:

Let  $G' = (V', \Sigma, P', S)$  where  $V' = \{S, A, L, R\} \cup V \cup \{\bar{B} : B \text{ in } V - \Sigma\}$  such that these are three disjoint sets of symbols and  $P'$  contains the productions:

- (i)  $S \rightarrow ALZR$ ;  $AL \rightarrow AAL$ ;  $AL \rightarrow A$ ;  $AR \rightarrow R$ ;  $AR \rightarrow \lambda$
- (ii)  $AB \rightarrow A\bar{B}$ ; for all  $B$  in  $V - \Sigma$ ,
- (iii)  $A\bar{B} \rightarrow CA$ ; if  $B \rightarrow C$  is in  $P$ ,
- (iv)  $A\bar{B} \rightarrow CDA$ ; if  $B \rightarrow CD$  is in  $P$ ,
- (v)  $A\bar{B} \rightarrow aA$ ; if  $B \rightarrow a$  is in  $P$ .

A derivation in  $G'$  proceeds as follows:

$$S \Rightarrow^+ A^{n-1}LZR \Rightarrow A^nZR \Rightarrow A^{n-1}\alpha AR \Rightarrow \dots$$

Clearly  $A$  cannot begin to sweep from left-to-right until the production  $AL \rightarrow L$  is applied to remove  $L$ . Note that  $A$ 's cannot overtake each other and  $A$  cannot be destroyed during a sweep over the symbols from  $V - \Sigma$ , therefore each  $A$  must complete its sweep for a terminal word to be derived. Also an  $A$  symbol cannot sweep over a terminal symbol, hence only the last  $A$  must produce terminals if a terminal word is to be derived. Again the production  $AR \rightarrow R$  must be used for all but the last  $A$  at which time  $AR \rightarrow \lambda$  must be used. Thus in  $G'$ ,  $A^nZR \Rightarrow^* x$  in  $\Sigma^*$  iff there is a derivation  $Z \Rightarrow^* x$  in  $F$  which has exactly  $n$  steps. Hence  $L(G', \Rightarrow) = L(F, \Rightarrow)$  as desired and  $G'$  is clearly a  $u$ -interpretation of  $G$ .

Before turning to the reverse inclusion observe that a production  $B \rightarrow \alpha$  in  $F$  is simulated in two steps

$$AB \Rightarrow A\bar{B} \Rightarrow \alpha A$$

in  $G'$ . The reason for this is simply that this allows non-uniform interpretations of  $B \rightarrow \alpha$  to take place in  $G'$ , or alternatively non-uniform interpretations of the productions  $Z \rightarrow Z$ ;  $Z \rightarrow ZZ$ ;  $Z \rightarrow a$  in  $G$ . Without this "stretching" we would not obtain  $\mathcal{L}(EOL)$ .

We now claim that  $\mathcal{L}_u(G, \Rightarrow) \subseteq \mathcal{L}(EOL)$ . Consider an arbitrary interpretation  $G' \triangleleft_u G$ . We argue that  $L(G', \Rightarrow)$  is an EOL language by way of a number of observations. First consider interpretations of  $A, L, Z$  and  $R$ . Clearly there must be a finite number of such interpretations and hence a finite number of productions of the type  $S \rightarrow ALZR$ . Since  $\mathcal{L}(EOL)$  is closed under union it suffices to consider that  $G'$  has one such production  $S \rightarrow ALZR$ , say. Since we are dealing with uniform interpretations we must have:



(i) at least one of  $AL \rightarrow AAL$  and  $AL \rightarrow A$   
and

(ii) at least one of  $AR \rightarrow R$  and  $AR \rightarrow \lambda$ .

Note that if either  $AL \rightarrow A$  or  $AR \rightarrow \lambda$  are absent then  $L(G', \Rightarrow) = \emptyset$  since either generation cannot begin or termination cannot occur. On the other hand if either  $AL \rightarrow A$  is the only production for  $AL$  or  $AR$  is the only production for  $AR$ , then only one sweep over the  $Z$ -symbols can occur and hence  $L(G', \Rightarrow)$  is finite.

Thus we now assume all of  $AL \rightarrow AAL$ ,  $AL \rightarrow A$ ,  $AR \rightarrow R$  and  $AR \rightarrow \lambda$  are present.

In this case we are able to construct an EOL grammar  $F = (V, \Sigma, P, Z)$  from  $G' = (V', \Sigma, P', S)$  such that  $L(G', \Rightarrow) = L(F, \Rightarrow)$ . Let  $V = (V' - \{S, A, L, R\}) \cup \{N\}$  and  $P$  contain the productions:

(i)  $N \rightarrow N$ ;  $a \rightarrow N$ , for all  $a$  in  $\Sigma$ ,

(ii)  $B \rightarrow \sigma$ , if  $AB \rightarrow A\bar{B}$  and  $A\bar{B} \rightarrow \alpha A$  are in  $P'$ .

It should be clear that

$Z \Rightarrow^n x$  in  $F$  with  $x$  in  $\Sigma^*$ .

iff  $S \Rightarrow^+ A^n ZR \Rightarrow^+ x$  in  $G'$ .

Hence the claim is established.

Note that we have really demonstrated more than this, namely, that  $G$  is synchro-EOL-vomplete. In other words for every synchro-EOL form  $F$  there is a  $G' \triangleleft_u G$  such that  $\mathcal{L}_S(F, \Rightarrow) = \mathcal{L}_u(G', \Rightarrow)$ . We use the first construction given above which provides a  $G' \triangleleft_u G$  with  $L(G', \Rightarrow) = L(F, \Rightarrow)$  and then observe that the second construction applied to  $G'' \triangleleft_u G'$  clearly gives an  $F' \triangleleft_S F$ . The only tricky point is that we appear to need closure under union, which does not, of course, hold in general. However we can modify the second construction since union-closure is not really needed, but rather a much weaker closure result.

By suitably modifying  $G$  we can obtain every EOL grammatical family. Let  $\bar{G}$  contain the productions:

$S \rightarrow ALZR$ ;  $AL \rightarrow AAL$ ;  $AL \rightarrow A$

$AZ \rightarrow A\bar{Z}$ ;  $A\bar{Z} \rightarrow ZA$ ;  $A\bar{Z} \rightarrow ZZA$ ;  $A\bar{Z} \rightarrow aA$ ;  $A\bar{Z} \rightarrow A$

$Aa \rightarrow A\bar{a}$ ;  $A\bar{a} \rightarrow ZA$ .

Let  $F$  be a binary EOL form then a  $\bar{G}' \triangleleft_u \bar{G}$  can be constructed such that  $L(\bar{G}', \Rightarrow) = L(F, \Rightarrow)$  and moreover  $\mathcal{L}_u(\bar{G}', \Rightarrow) = \mathcal{L}_S(F, \Rightarrow)$ . We leave this to the reader. Notice that we have included all binary productions in  $\bar{G}$  to be able to obtain this result (by the results of Chapter III every EOL form  $\bar{F}$  has a form equivalent form  $F$  which is in binary normal form).

In a similar manner we can obtain  $\mathcal{L}(\text{ETOL})$ . We leave the reader to demonstrate that  $H$  defined by the productions:

$S \rightarrow \text{LZR}; L \rightarrow \text{AL}; \text{AL} \rightarrow \text{A};$   
 $\text{AZ} \rightarrow \text{AZ}\bar{Z}; \text{AZ}\bar{Z} \rightarrow \text{ZA}; \text{AZ}\bar{Z} \rightarrow \text{ZZA}; \text{AZ}\bar{Z} \rightarrow \text{aA};$   
 $\text{AR} \rightarrow \text{R}; \text{AR} \rightarrow \lambda$

has  $\mathcal{L}_u(H, \Rightarrow) = \mathcal{L}(\text{ETOL})$ .

Combining smoothness and uniform-interpretations we obtain su-interpretations denoted by  $\hat{\mathcal{L}}_{su}$ . We can exhibit a grammar form  $G$  such that  $\hat{\mathcal{L}}_{su}(G, \Rightarrow) = \mathcal{L}(\text{MAT})$  the family of context-free matrix languages or such that  $\hat{\mathcal{L}}_{su}(G, \Rightarrow) = \mathcal{L}(\text{SCAT})$ , the family of context-free scattered-context languages. However in both cases  $\hat{\mathcal{L}}_{su}(G, \Rightarrow) \subsetneq \mathcal{L}_u(G, \Rightarrow)$ . For example, consider  $G$  defined by:

- (a)  $S \rightarrow \text{AS}; \text{AS} \rightarrow \text{AZ};$
- (b)  $\text{AZ} \rightarrow \text{ZA}; \text{Aa} \rightarrow \text{aA}; \text{A} \rightarrow \hat{\text{A}};$
- (c)  $\hat{\text{A}}\text{Z} \rightarrow \hat{\text{A}}\bar{\text{Z}}; \hat{\text{A}}\bar{\text{Z}} \rightarrow \text{Z}\bar{\text{A}}; \hat{\text{A}}\bar{\text{Z}} \rightarrow \text{ZZ}\bar{\text{A}}; \hat{\text{A}}\bar{\text{Z}} \rightarrow \text{a}\bar{\text{A}};$
- (d)  $\bar{\text{A}}\text{Z} \rightarrow \text{Z}\bar{\text{A}}; \bar{\text{A}}\text{a} \rightarrow \text{a}\bar{\text{A}}; \bar{\text{A}} \rightarrow \hat{\bar{\text{A}}};$
- (e)  $\hat{\bar{\text{A}}}\bar{\text{Z}} \rightarrow \hat{\bar{\text{A}}}\bar{\text{Z}}; \hat{\bar{\text{A}}}\bar{\text{Z}} \rightarrow \text{Z}; \hat{\bar{\text{A}}}\bar{\text{Z}} \rightarrow \text{ZZ}; \hat{\bar{\text{A}}}\bar{\text{Z}} \rightarrow \text{a}.$

We claim that  $\hat{\mathcal{L}}_{su}(G, \Rightarrow) = \mathcal{L}(\text{SCAT})$ . Essentially the  $A$  is an activation messenger once more. A scattered-context grammar whose matrices have size two and which are in binary normal form can be simulated by having an  $A_m$  for each matrix  $m$ . As  $A_m$  sweeps from left-to-right it can change into  $\hat{A}_m$  which enables the first production in  $m$  to be applied. At the same time  $\hat{A}_m$  becomes  $\bar{A}_m$  which sweeps across until it changes into  $\hat{\bar{A}}_m$  at which time the second production in  $m$  is applied and  $\hat{A}_m$  disappears.

We leave to the reader the detailed proof that  $\hat{\mathcal{L}}_{su}(G, \Rightarrow) = \mathcal{L}(\text{SCAT})$ , however we observe that smoothness is essential. For example, in a  $u$ -interpretation we could omit  $\bar{\text{A}}\text{Z} \rightarrow \text{Z}\bar{\text{A}}$  and  $\bar{\text{A}}\text{a} \rightarrow \text{a}\bar{\text{A}}$  in which case the two productions of an  $m$  would have to be applied to adjacent symbols. But this means that context-sensitive productions can be simulated, hence  $\mathcal{L}_u(G, \Rightarrow) = \mathcal{L}(\text{CS})$ .

## IV.2 EIL Forms

Just as Section 1 considers context-dependent grammar forms, this current section is concerned with context-dependent L forms, better known as EIL Forms. We will present results in three areas, namely reduction results, generative capacity and undecidability. Since the EOL-like reduction results go through for EIL forms we concentrate our attention on context-reduction results. In fact it is demonstrated that EIL forms behave in a similar manner to DIL grammars in this respect; an unexpected occurrence. Unlike EOL forms, which do not generate exactly  $\mathcal{L}(CF)$  we show that there is an EIL form  $F$  which does exactly this, that is  $F$  is context-free-complete. Finally questions of completeness, goodness and form equivalence are shown to be undecidable, again in contrast to what is known for EOL forms.

Given integers  $m, n \geq 0$  we say that  $G = (V, \Sigma, P, S)$  is an  $E(m, n)L$  grammar if  $P \subseteq V^{\leq m} \times V \times V^{\leq n} \times V^*$  is finite and the modified version of  $\Rightarrow$  which is defined below is given. Moreover for all  $\alpha, \beta$  in  $V^*$  and  $X$  in  $V$  such that  $0 \leq |\alpha| \leq m, 0 \leq |\beta| \leq n$  there is a production

$$(\alpha, X, \beta, \gamma) \text{ in } P$$

for some  $\gamma$  in  $V^*$ . We usually write this as:

$$(*) (\alpha, X, \beta) \rightarrow \gamma.$$

Intuitively  $(*)$  can be understood as:  $X$  is replaced by  $\gamma$  if  $\alpha X \beta$  appears in the word currently being rewritten. If  $|\alpha| = m$  and  $|\beta| = n$  this is exactly what happens. However if  $|\alpha| = m_1 < m$  say, then the occurrence of  $X$  we replace must be the  $(m_1 + 1)$ st in the word under consideration. Thus in an  $E(2, 2)L$  grammar, the production

$$(B, A, BB) \rightarrow AA$$

is applicable to

BABB

but not to

BBABB.

Thus short contexts ( $< m, < n$ ) are only used at the beginning and end of a word.

We write  $\alpha \Rightarrow \beta$  according to  $G$  if  $\alpha = X_1 \dots X_k, k \geq 1,$

$\beta = \beta_1 \dots \beta_k$  and for all  $i, 1 \leq i \leq k,$

$$(X_{i-m} \dots X_{i-1}, X_i, X_{i+1} \dots X_{i+n}) \rightarrow \beta_i$$

is in  $P$ . We assume  $X_j = \lambda$  for all  $j \leq 0$  or  $j \geq k + 1$ . As usual we obtain  $\Rightarrow^i, \Rightarrow^+$  and  $\Rightarrow^*$ . Hence  $L(G, \Rightarrow) = \{x: x \text{ is in } \Sigma^* \text{ and } S \Rightarrow^* x\}$ . We say  $G$  is an EIL grammar if there are  $m \geq 0$  and  $n \geq 0$  such that  $G$  is an  $E(m, n)L$  grammar. Clearly the  $E(0, 0)L$  grammars are exactly the EOL grammars.

Given two  $E(m, n)L$  grammars (or forms)  $G'$  and  $G$ , where

$G' = (V', \Sigma', P', S')$  and  $G = (V, \Sigma, P, S)$  then we say that  $G'$  is an interpretation of  $G$  modulo  $\mu$ , where  $\mu$  is a dfl-substitution on  $V^*$  if the usual conditions hold. We only need to consider what is meant by  $\mu((\alpha, X, \beta) \rightarrow \gamma)$ , for some production

$$(\alpha, X, \beta) \rightarrow \gamma$$

in  $P$ . Simply stated  $(\alpha', X', \beta') \rightarrow \gamma'$  is in  $\mu((\alpha, X, \beta) \rightarrow \gamma)$  if  $\alpha'$  is in  $\mu(\alpha)$ ,  $X'$  is in  $\mu(X)$ ,  $\beta'$  is in  $\mu(\beta)$  and  $\gamma'$  is in  $\mu(\gamma)$ .

As before we write  $G' \xrightarrow[S]{\Delta} G(\mu)$  or simply  $G' \xrightarrow[S]{\Delta} G$ . In the following we always have the convention that if no production is listed for a letter  $X$  in a particular context  $(\alpha, \beta)$  this means that the production is assumed to be blocking, that is

$$(\alpha, X, \beta) \rightarrow N$$

where  $N$  is a special symbol having the production

$$(\gamma, N, \delta) \rightarrow N$$

for all contexts  $(\gamma, \delta)$ .

Note that  $F_1$  defined by the productions:

$(\alpha, S) \rightarrow S$ ;  $(\alpha, S) \rightarrow SS$ ;  $(\alpha, S) \rightarrow \lambda$ ;  $(\alpha, S) \rightarrow a$ ; where  $\alpha = S$  or  $\lambda$ , is an  $E(1,0)L$  form with  $\mathcal{L}_S(F_1, \Rightarrow) = \mathcal{L}(RE)$ . Again by convention since  $n = 0$  we have written  $(\alpha, X, \lambda)$  as  $(\alpha, X)$ . That  $\mathcal{L}_S(F_1, \Rightarrow)$  generates all recursively enumerable languages follows by two observations. First every  $L$  in  $\mathcal{L}(RE)$  can be generated by an  $E(1,0)L$  grammar and second every  $E(1,0)L$  grammar can be converted into an equivalent synchronized  $E(1,0)L$  grammar whose right hand sides are of type:  $B, BC, \lambda$  or  $a$ . Clearly such a grammar is an interpretation of  $F_1$ .

If we omit the erasing productions from  $F_1$  giving  $F_2$ , then  $\mathcal{L}_S(F_2, \Rightarrow) = \mathcal{L}(CS)$  by similar arguments.

We now turn to our first reduction result, namely, for each  $E(m,n)L$  form  $G$  with  $m \geq 2$  (respectively  $n \geq 2$ ) a form equivalent  $E(m-1,n)L$  form  $H$  (respectively  $E(m,n-1)L$  form  $H$ ) can be constructed.

We only consider the case of the reduction of left context, since the reduction of right context follows in a similar manner. The basic idea is to stretch each derivation in  $G$  into two steps such that in the first step each two adjacent symbols are represented as one compressed symbol. This implies that  $m-1$  compressed symbols now contain information regarding  $m$  of the original symbols as desired.

More precisely let the alphabet of  $H$  consist of that of  $G$  together with all additional symbols  $[X, Y]$  where  $X$  and  $Y$  are symbols of  $G$  or  $\lambda$ . The productions of  $H$  are defined as follows:

Each symbol  $Y$  of  $G$  with left neighbour  $X$  is rewritten as  $[X, Y]$  independently of the right neighbours, that is

$$(X_1 \dots X_{m-1} X, Y, Z_1 \dots Z_n) \rightarrow [X, Y]$$

is in  $H$  for all  $X_1, \dots, X_{m-1}, Z_1, \dots, Z_n$ .

Secondly if

$$(X_1 \dots X_m, X, Z_1 \dots Z_n) \rightarrow \alpha$$

is a production of  $G$ , then

$$([X_1, X_2] \dots [X_{m-1}, X_m], [X_m, X], [X, Z_1] \dots [Z_{n-1}, Z_n]) \rightarrow \alpha$$

is a production of  $H$ .

Clearly  $\mathcal{L}_S(G, \Rightarrow) \subseteq \mathcal{L}_S(H, \Rightarrow)$ . That the reverse inclusion holds follows by observing that length two derivations in any  $H' \xrightarrow{S} H$  can be used to construct the productions of an equivalent  $G' \xrightarrow{S} G$ .

Applying the result iteratively we can obtain a form equivalent  $E(1,1)L$  form if  $m \geq 1$  and  $n \geq 1$ , a form equivalent  $E(1,0)L$  form if  $m \geq 1$  and  $n = 0$  and a form equivalent  $E(0,1)L$  form if  $m = 0$  and  $n \geq 1$ .

However in the case than an  $E(1,1)L$  form is synchronized, that is the rewriting of a terminal in any context and the rewriting of a nonterminal in a context which contains a terminal both produce a blocking symbol, the reduction can be carried one step further.

In other words for each synchronized  $E(m,n)L$  form  $G$  with  $m \geq 1$  and  $n \geq 1$  a form equivalent synchronized  $E(1,0)L$  form (or  $E(0,1)L$  form)  $H$  can be constructed. For example to construct an  $E(0,1)L$  form  $H$  which is form equivalent to  $G$  proceed as follows. The alphabet of  $H$  consists of that of  $G$  together with nonterminals  $\$, \bar{\$}$  (left boundary markers),  $\bar{a}$  for each terminal  $a$  in  $G$  and  $[X, Y]$  for all  $X$  a nonterminal of  $G$  and  $Y$  a nonterminal of  $G$  except for  $X$  equal to  $\$$  and  $\bar{\$}$  if  $Y$  equals  $\lambda$ . We also assume  $H$  has a new sentence symbol yielding only  $\$S$  where  $S$  is the sentence symbol of  $G$ .

Now consider a derivation step

$$A_1 \dots A_k \Rightarrow \alpha_1 \dots \alpha_k$$

in  $G$  where  $k \geq 1$ . First assume  $\alpha_1 \dots \alpha_k$  is non-empty and all non-terminal. Then we simulate this derivation in 2 steps

$$\$A_1 \dots A_k \Rightarrow [\$A_1][A_1, A_2] \dots [A_k, \lambda] \Rightarrow \$\alpha_1 \dots \alpha_k$$

in  $H$ . Thus assuming  $G$  is an  $E(1,1)L$  form, then the production

$$(\lambda, A_1, A_2) \rightarrow \alpha_1$$

in  $G$  becomes

$$([\$, A_1], [A_1, A_2]) \rightarrow \$\alpha_1$$

in  $H$ . To accomplish the first step in the simulating derivation we accumulate right neighbours as in the previous construction, that is

$$(A, B) \rightarrow [A, B]$$

and

$$(A, \lambda) \rightarrow [A, \lambda].$$

We also have

$$([A, \lambda], \lambda) \rightarrow \lambda$$

in H for all A.

Secondly assume  $\alpha_1 \dots \alpha_k$  is non-empty and terminal (because of synchronization these are the only cases we need consider) then

$$A_1 \dots A_k \Rightarrow \alpha_1 \dots \alpha_k$$

in G is simulated in 3 steps.

$$\$A_1 \dots A_k \Rightarrow [\$A_1][A_1, A_2] \dots [A_k, \lambda] \Rightarrow \overline{\$}\alpha_1 \dots \bar{\alpha}_k \Rightarrow \alpha_1 \dots \alpha_k$$

in H. This delay of one step is to prevent the possibility that

$([\$A_1], [A_1, A_2]) \rightarrow \lambda$  is in H, which could then be applied during the

nonterminal rewriting stage causing a loss of left context. We leave

to the reader the details of the construction and the verification of

the form equivalence of G and H. As before  $\mathcal{L}_s(G, \Rightarrow) \subseteq \mathcal{L}_s(H, \Rightarrow)$  is

clear, while the following hint should be used in the proof of the

reverse inclusion. Let  $H' \triangleleft_s H$  and  $H'$  contain a production  $(A, B) \rightarrow \alpha$

where A is an interpretation of  $[A_1, A_2]$  and B of  $[A_2, A_3]$ . Then in

constructing a  $G' \triangleleft_s G$  choose  $[A, B]$  to be an interpretation of  $A_2$ ,

in which case  $(A, B) \rightarrow \alpha$  in  $H'$  becomes

$$([D, A], [A, B], [B, E]) \rightarrow \alpha$$

in  $G'$ .

It can be shown that when an  $E(m, n)L$  form with  $m, n \geq 1$  is not synchronized the context cannot be reduced any further than  $(1, 1)$ ,

hence the above results are the best possible. For example consider

the  $E(1, 1)L$  form  $F_3$ :

$$(\lambda, S, \lambda) \rightarrow ab; (\lambda, S, \lambda) \rightarrow aba;$$

$$(\lambda, a, b) \rightarrow \lambda; (b, a, \lambda) \rightarrow \lambda;$$

$$(a, b, \lambda) \rightarrow ababa; (a, b, a) \rightarrow abab.$$

$F_3$  has no form equivalent  $E(1, 0)L$  or  $E(0, 1)L$  form, indeed it has no

form equivalent  $E(m, 0)L$  or  $E(0, n)L$  form for any  $m, n \geq 0$ . This can be

proved by noting that the only possible derivations in  $F_3$  are:

$$S \Rightarrow ab \Rightarrow ababa \Rightarrow \text{blocking}$$

and

$$S \Rightarrow aba \Rightarrow abab \Rightarrow \text{blocking}.$$

Thus for every  $F'_3 \triangleleft_s F_3$ ,  $LS(F'_3, \Rightarrow) \subseteq \{2, 3, 4, 5\}$  and moreover if 2 is in

$LS(F'_3, \Rightarrow)$  then 5 is in  $LS(F'_3, \Rightarrow)$  and if 3 is in  $LS(F'_3, \Rightarrow)$  4 is in

$LS(F'_3, \Rightarrow)$ . Now it can be shown that any form equivalent  $E(m, 0)L$  form

H must contradict one of these restrictions and similarly for a form

equivalent  $E(0, n)L$  form.

We now turn our attention to the generation of  $\mathcal{L}(CF)$  by EIL

forms. While it is not known whether there is an EIL form F with

$\mathcal{L}_s(F, \Rightarrow) = \mathcal{L}(CF)$ , we can, under a natural modification of the notion

of interpretation, obtain  $\mathcal{L}(CF)$ . We conjecture that there is no EIL

form  $F$  with  $\mathcal{L}_S(F, \Rightarrow) = \mathcal{L}(CF)$ , however this is almost certainly a difficult conjecture to resolve.

We approach the result in two steps. First we show that the "restricted EOL languages" are exactly the context-free languages and second we introduce "constant-context" interpretations. Together these two results enable us to demonstrate the claimed result.

Considering the first step let us define restricted EOL languages. Consider the EOL form  $F$  defined by the productions

$$S \rightarrow S; S \rightarrow a; S \rightarrow aS; S \rightarrow aSS; a \rightarrow a;$$

and consider an arbitrary interpretation  $F' = (V', \Sigma', P', S')$  of  $F$ . We say that productions in  $P'$  which are interpretations of  $S \rightarrow a$ ,  $S \rightarrow aS$  or  $S \rightarrow aSS$  are active productions and productions which are interpretations of  $S \rightarrow S$  or  $a \rightarrow a$  are cycling productions. We now define a modification of  $\Rightarrow$  for such interpretations. Let  $\alpha$  be in  $\Sigma'^*(V' - \Sigma)^+$  and  $\beta$  be in  $V'^*$ . We say

$$\alpha \stackrel{\text{L}}{\Rightarrow} \beta$$

iff  $\alpha = a_1 \dots a_j A_{i+1} \dots A_m$ ,  $\beta = b_1 \dots b_j \alpha_{i+1} B_{i+2} \dots B_m$ ,  $a_j \rightarrow b_j$  is in  $P'$ ,  $1 \leq j \leq i$ ,  $A_j \rightarrow B_j$  is in  $P'$ ,  $i+2 \leq j \leq m$  and  $A_{i+1} \rightarrow \alpha_{i+1}$  is in  $P'$ . In other words each symbol is rewritten by cycling productions except for the leftmost nonterminal, which is rewritten by an active production. Note that each nonterminal which can appear in the leftmost spot must have an active production otherwise such an appearance cannot be rewritten.  $\stackrel{\text{L}}{\Rightarrow}$  can be extended to  $\stackrel{\text{L}}{\Rightarrow}^*$  and  $\stackrel{\text{L}}{\Rightarrow}^+$  in the usual way. The restricted EOL language of  $F'$ , denoted  $L(F', \stackrel{\text{L}}{\Rightarrow})$ , is defined by:

$$L(F', \stackrel{\text{L}}{\Rightarrow}) = \{x: x \text{ is in } \Sigma'^* \text{ and } S' \stackrel{\text{L}}{\Rightarrow}^+ x\}.$$

Thus we may speak of  $\mathcal{L}_S(F, \stackrel{\text{L}}{\Rightarrow})$  for the particular  $F$  defined above, in which case we claim that  $\mathcal{L}(CF) = \mathcal{L}_S(F, \stackrel{\text{L}}{\Rightarrow})$ .

That  $\mathcal{L}(CF) \subseteq \mathcal{L}_S(F, \stackrel{\text{L}}{\Rightarrow})$  follows by observing that each context-free language can be generated by a grammar  $G$  in Greibach two-standard form, that is its productions are interpretations of the active productions in  $F$ . Now add  $A \rightarrow A$  to  $G$  for all nonterminals and  $a \rightarrow a$  for all terminals giving  $\bar{G}$ , then  $\bar{G} \stackrel{\text{L}}{\Leftarrow} F$  and  $L(\bar{G}, \stackrel{\text{L}}{\Rightarrow}) = L(G, \Rightarrow)$  since  $G$  simulates a leftmost derivation in  $G$  and vice versa.

We will sketch the proof that  $\mathcal{L}_S(F, \stackrel{\text{L}}{\Rightarrow}) \subseteq \mathcal{L}(CF)$ , that is for all  $F' \stackrel{\text{L}}{\Leftarrow} F$   $L(F', \stackrel{\text{L}}{\Rightarrow})$  is context-free. We can assume that the only cycling productions for terminals are  $a \rightarrow a$  for each  $a$  in  $\Sigma'$ . If this is not the case an  $F'' \stackrel{\text{L}}{\Leftarrow} F$  and a gsm  $M$  can be constructed such that  $L(F', \stackrel{\text{L}}{\Rightarrow}) = M(L(F'', \stackrel{\text{L}}{\Rightarrow}))$  using the technique of Theorem III.3.28 with  $F''$  having  $a \rightarrow a$  as the only terminal production for each  $a$  in  $\Sigma'$ .

Consider the nonterminals of  $F'$ ,  $S' = A_1, \dots, A_n$  say. Let  $T_i(A)$  denote the set of nonterminals reachable in  $i$  steps from  $A$  using only cycling productions and  $T_i = (T_i(A_1), \dots, T_i(A_n))$ . Then the sequence

$$T_1, T_2, \dots$$

is clearly ultimately periodic, that is there is a  $t \geq 1$  and a  $p \geq 1$  such that for all  $i \geq t$

$$T_{i+p} = T_i.$$

To give the reader the flavor of the proof consider the case  $t = 1$ , that is the sequence is periodic. We construct a context-free grammar  $G = (V, \Sigma', P, S)$  where  $V = \Sigma' \cup \{S\} \cup \{A_i^j : 1 \leq i \leq n, 1 \leq j \leq p\}$  and  $P$  contains the following productions:

- (i)  $S \rightarrow A_1^j$  for each  $j, 1 \leq j \leq p$ ,
- (ii)  $B^j \rightarrow aC^kE^\ell$  for each  $j, 1 \leq j \leq p$  if  $B \rightarrow aCD$  is in  $P'$ ,  $E$  is in  $T_k(D)$  and  $j \equiv k + \ell + 1 \pmod{p}$ , where  $1 \leq k, \ell \leq p$ ,
- (iii)  $B^j \rightarrow aC^k$  for each  $j, 1 \leq j \leq p$ , if  $B \rightarrow aC$  is in  $P'$  where  $1 \leq k \leq p$  and  $k \equiv j-1 \pmod{p}$ ,
- (iv)  $B^1 \rightarrow a$  is in  $P$  if  $B \rightarrow a$  is in  $P'$ .

The superscript of a nonterminal indicates the number of active derivation steps modulo  $p$  before a complete terminal word is derived from it. Because of the periodicity of the  $T_i$ -sequence it suffices to keep track of the length of the cycling sequence modulo  $p$ . In type (ii) productions the leftmost nonterminal when the  $C^k$  symbol has produced a terminal word is a member of  $T_k(D)$  as constructed.

It can be shown that  $L(G, \Rightarrow) = L(F', \stackrel{L}{\Rightarrow})$  by an inductive proof on the length of terminal words. Essentially in the inductive step two cases occur corresponding to an initial choice of either a type (ii) or (iii) production.

It remains to extend the proof to the case that  $t > 1$ . We once more leave the reader to extend the proof technique.

This leaves the task of defining "constant-context" interpretations or cc-interpretations, before tackling the main result.

Let  $F = (V, \Sigma, P, S)$  and  $F' = (V', \Sigma', P', S')$  be two  $E(m, n)L$  forms such that  $F' \triangleleft_S F(\mu)$  for some  $\mu$ . We say  $F'$  is a cc-interpretation of  $F$  modulo  $\mu$  if additionally  $P'$  satisfies the condition:

Whenever  $(\alpha', X', \beta') \rightarrow \gamma'$  is in  $P'$ ,  $\alpha''$  is in  $\mu(\alpha)$ ,  $\beta''$  is in  $\mu(\beta)$  then

$$(\alpha'', X', \beta'') \rightarrow \gamma' \text{ is in } P'.$$

We write  $F' \triangleleft_{cc} F(\mu)$ . In other words we take full interpretations of the context of each production taken into  $P'$ . This prohibits making further contextual restrictions under cc-interpretation. In particular



if a production  $X \rightarrow \alpha$  is applicable in all contexts in the original form  $F$  then it is still applicable in all contexts in any interpretation  $F' \stackrel{\Delta}{\underset{CC}{\Leftarrow}} F$ . In the usual way we define  $\mathcal{L}_{CC}(F, \Rightarrow)$ .

We now claim that the E(1,0)L form  $F_4$  defined by the productions:

- (1)  $(\alpha, S) \rightarrow a$ ;  $(\alpha, S) \rightarrow aS$ ;  $(\alpha, S) \rightarrow aSS$ ;  
 (2)  $(\alpha, a) \rightarrow a$ ;  $(S, S) \rightarrow S$ ;

where  $\alpha$  ranges over  $\lambda$  and  $a$ , satisfies

$$\mathcal{L}(CF) = \mathcal{L}_{CC}(F_4, \Rightarrow).$$

Note that productions of type (1) correspond to active productions and those of type (2) to cycling productions. Clearly  $\mathcal{L}(CF) \subseteq \mathcal{L}_{CC}(F_4, \Rightarrow)$ . That the reverse inclusion also holds follows from two observations. First every  $F_4 \stackrel{\Delta}{\underset{CC}{\Leftarrow}} F_4$  does not introduce more contextual restrictions. This means that the cycling a's and S's are rewritten in EOL-like cycling mode, that is context is irrelevant. Second in such an  $F_4$  there must be type (1) productions for every nonterminal in every terminal or  $\lambda$  context. Thus the leftmost nonterminal is always rewritten using active productions, which are again independent of the nonterminal's left context. Therefore  $\mathcal{L}_{CC}(F_4, \Rightarrow) \subseteq \mathcal{L}(CF)$  and equality is proved.

To close this section we briefly consider some decidability issues. These are all based on the following two reachability results. Given an EIL form  $F = (V, \Sigma, P, S)$  we say  $X$  in  $V$  is reachable if there is a derivation  $S \Rightarrow^* \alpha X \beta$  for some  $\alpha, \beta$  in  $V^*$ . First, given an  $F' = (V', \Sigma', P', S') \stackrel{\Delta}{\underset{S}{\Leftarrow}} F(\mu)$  we have:

if  $X$  is reachable in  $F'$  then  $\mu^{-1}(X)$  is reachable in  $F$ . Second, given  $F$  and an  $X$  in  $V$  it is undecidable whether or not  $X$  is reachable in  $F$ .

As we would expect almost everything turns out to be undecidable for EIL forms. We will only consider E(1,1)L forms in the sequel for simplicity rather than for results. The various undecidability results can all be obtained by reducing the reachability problem to the problem under consideration. Consider the following examples of how this may be carried out.

We first demonstrate the undecidability of completeness of an E(1,1)L form. Let  $F = (V, \Sigma, P, S)$  be an arbitrary E(1,1)L form and  $A$  an arbitrary letter in its alphabet.

We construct an E(1,1)L form  $G = (U, \Delta, Q, S)$  from  $F$ . Let all symbols in  $V$  be nonterminals in  $G$ ,  $Q$  contain all productions in  $P$ , the productions

$$(A, B, C) \rightarrow \lambda; (C, B, A) \rightarrow \lambda; (\lambda, A, \lambda) \rightarrow Z$$

for all  $B, C$  in  $V$  and the productions

$(B,C,D) \rightarrow C$

for all  $B,C,D$  in  $V$ . Furthermore  $Q$  also contains the productions of some complete  $E(1,1)L$  form  $H$  over terminal alphabet  $\Delta$  and nonterminal alphabet  $U - (V \cup \Delta)$ , where  $Z$  is its sentence symbol.

Clearly  $G$  is complete iff  $A$  is reachable in  $F$ , since if  $A$  is reachable in  $F$  then there is a derivation  $S \Rightarrow^* A \Rightarrow Z$  in  $G$ .

Immediately by replacing  $H$  by a good  $E(1,1)L$  form we may observe that goodness is undecidable. Letting  $H$  be some arbitrary  $E(1,1)L$  form we have  $\mathcal{L}_S(H, \Rightarrow) = \mathcal{L}_S(G, \Rightarrow)$  iff  $A$  is reachable in  $F$ . Hence form equivalence is undecidable.

Since the technique is also appropriate for cc-interpretations replacing  $H$  by  $F_4$  shows that it is undecidable whether  $\mathcal{L}_{cc}(G, \Rightarrow) = \mathcal{L}(CF)$  or not.

Many other undecidability results can be demonstrated by this reduction technique. We leave these to the reader.

### IV.3 Context-free Matrix Forms

Not surprisingly not only has the notion of interpretation been applied to the two most important rewriting mechanisms, namely ETOL and context-free grammars and their non-context-free variants, but also it has been applied to other rewriting systems in the literature. We will discuss two of these generalizations. In the next section we will examine controlled grammar forms, while in the present section we examine matrix forms.

Recall that a (context-free) matrix grammar  $G$  is a quadruple  $(V, \Sigma, M, S)$  where  $V, \Sigma$  and  $S$  are the same as in a context-free grammar.  $M$  is a finite set of matrices of (context-free) productions, where each matrix  $m$  in  $M$  is a finite nonempty sequence of (context-free) productions. A matrix  $m$  is usually written as:

$$[A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n], n \geq 1.$$

By  $P(M)$  we denote the set of productions used in the matrices of  $M$ , hence we may consider a matrix to be a word over  $P(M)$ .

Depending on whether  $P(M)$  is a set of right linear, linear,  $\lambda$ -free productions, etc., we say that  $G$  is a right linear, linear,  $\lambda$ -free matrix grammar, etc.

Given a matrix grammar  $G = (V, \Sigma, M, S)$  and words  $\alpha, \beta$  in  $V^*$  we write  $\alpha \xRightarrow{m} \beta$  (or simply  $\alpha \Rightarrow \beta$ ) iff there exists an integer  $n \geq 1$ , words  $\gamma_1, \dots, \gamma_{n+1}, \delta_1, \dots, \delta_n, \eta_1, \dots, \eta_n$  and a matrix  $m = [A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n]$  in  $M$  such that

$$\alpha = \gamma_1, \beta = \gamma_{n+1}, \gamma_i = \delta_i A_i \eta_i \text{ and } \gamma_{i+1} = \delta_i \alpha_i \eta_i \text{ for every } i, 1 \leq i < n.$$

In other words  $\alpha$  derives  $\beta$  in  $G$  if there is a matrix  $m$  in  $M$  such that  $\alpha$  derives  $\beta$  in the usual context-free grammar sense using the productions of  $m$  in the order specified by  $m$ .

In the standard manner we can define the transitive and reflexive transitive closure of  $\Rightarrow$  and hence the language generated by  $G$ , denoted by  $L(G, \Rightarrow)$ , is defined as  $\{x: x \text{ is in } \Sigma^*, S \Rightarrow^* x \text{ in } G\}$ .

We can now define  $\xrightarrow{g}$  and  $\xrightarrow{s}$  for matrix grammars in a similar manner to their definition for context-free grammars. It should be clear that if  $G' \xrightarrow{s} G(\mu)$  for some  $\mu$  then each matrix  $m'$  in  $G'$  is an image under  $\mu$  of some matrix  $m$  in  $G$ , that is  $m'$  is in  $\mu(m)$ . This immediately implies that the  $i$ th production in  $m'$  is an image of the  $i$ th production in  $m$ . From this fact we may observe that a derivation in  $G'$  is always an image of a derivation in  $G$ , simply because non-terminals are not destroyed under interpretation and matrices retain their basic structure. This also applies to  $g$ -interpretation.

Clearly every context-free grammar may be viewed as a matrix grammar and hence as a matrix form by simply letting each production in the grammar be a matrix. So immediately the matrix form  $H_1$ :

$$[S \rightarrow aS], [S \rightarrow a]$$

is known to generate all regular languages under both s- and g-interpretation. However the matrix form  $H_2$ :

$$[S \rightarrow S, S \rightarrow aS, S \rightarrow S], [S \rightarrow S, S \rightarrow a]$$

also generates exactly the regular languages. This though not proved here is not too surprising since the productions  $S \rightarrow S$  appearing in the matrices can be considered to be identity productions. Thus  $\mathcal{L}_S(\text{REG}) \subseteq \mathcal{L}_S(H_2, \Rightarrow)$  is easily seen. Conversely since only one non-terminal occurs in any derivation of  $H_2$  an arbitrary production

$$[A_1 \rightarrow A_2, A_3 \rightarrow aA_4, A_5 \rightarrow A_6]$$

in an interpretation  $H_2'$  of  $H_2$  can only be applied if  $A_2 = A_3$  and  $A_4 = A_5$ , in which case it has the same effect as  $[A_1 \rightarrow aA_6]$ . We may argue similarly for an arbitrary production  $[A_1 \rightarrow A_2, A_3 \rightarrow a]$ .

Naturally the most interesting results for matrix forms are those that are not simple generalizations of results for grammar forms. It turns out that most of the reduction results for matrix forms are different. For example given a matrix form  $G = (V, \Sigma, M, S)$  a new matrix form  $F = (U, \Sigma, N, S)$  can be constructed such that the matrices in  $N$  are one of three simple kinds, namely

$$(i) [A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n] \text{ for } n \geq 1 \text{ and the } \alpha_i \text{ are in } (V - \Sigma)^+, \\ 1 \leq i \leq n,$$

$$(ii) [A \rightarrow \lambda],$$

$$(iii) [A \rightarrow a], \text{ where } a \text{ is in } \Sigma$$

and moreover  $\mathcal{L}_S(G, \Rightarrow) = \mathcal{L}_S(F, \Rightarrow)$ .

We call such a matrix form separated. In such a form the erasing of a nonterminal and the introduction of a terminal is carried out in a context-free manner. This simple result has to some extent the flavour of synchronization of EOL grammars. In a separated matrix form any sentential derivation can be partitioned into a sequence of applications of type (i) matrices followed by a sequence of applications of matrices of types (ii) and (iii).

Whether productions of type (ii) can be removed completely while preserving the language family is an open question, whose resolution would imply that the families of context-free matrix languages and  $\lambda$ -free context-free matrix languages are equal. Since the latter result is a long-standing open problem, the problem of  $\lambda$ -removal for matrix forms also remains open.

Let  $\text{maxr}(G) = \text{maxr}(P(M))$ , where  $G = (V, \Sigma, M, S)$  is a matrix grammar and  $\text{maxr}(P(M))$  is the length of the longest right hand side of a production in  $P(M)$ . It is not too surprising that for each matrix form  $G$  a form equivalent short matrix form  $F$  can be constructed, where short means  $\text{maxr}(F) \leq 2$ . However the construction is of some interest since it depends on the following simulation result.

Lemma

Let  $G = (V, \Sigma, M, S)$  be a matrix form,  $m = [A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n]$  a matrix in  $M$  and  $k$  an integer,  $1 \leq k \leq n$ . Let  $F = (V_F, \Sigma, M_F, S)$  where  $M_F = (M - \{m\}) \cup \{[A_1 \rightarrow \alpha_1, \dots, A_{k-1} \rightarrow \alpha_{k-1}, A_k \rightarrow B_1 B_2, B_1 \rightarrow \alpha_{k1}, B_2 \rightarrow \alpha_{k2}, A_{k+1} \rightarrow \alpha_{k+1}, \dots, A_n \rightarrow \alpha_n]\}$ ,  $V_F = V \cup \{B_1, B_2\}$  and  $\alpha_k = \alpha_{k1} \alpha_{k2}$  for some  $\alpha_{k1}, \alpha_{k2}$ .

Then  $\mathcal{L}_S(F, \Rightarrow) = \mathcal{L}_S(G, \Rightarrow)$ .

We leave the proof of this result to the reader. Using this simulation lemma we sketch the proof of the following theorem.

Theorem

Let  $G = (V, \Sigma, M, S)$  be a matrix form. Then there exists a form equivalent matrix form  $F$  such that  $\text{maxr}(F) \leq 2$ .

Proof Sketch: If  $\text{maxr}(G) \leq 2$  then there is nothing to prove. Therefore assume  $\text{maxr}(G) = \ell > 2$ . Let  $\#(G)$  denote the sum of the lengths of the right hand sides of the appearances of productions in  $M$ , whose right hand sides have length  $> 2$ . Consider a matrix

$m = [A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n]$  in  $M$  with a position  $k$  in  $m$  such that

$|\alpha_k| > 2$ . Replace  $m$  in  $M$  by

$[A_1 \rightarrow \alpha_1, \dots, A_{k-1} \rightarrow \alpha_{k-1}, A_k \rightarrow B_1 B_2, B_1 \rightarrow \alpha_{k1}, B_2 \rightarrow \alpha_{k2}, A_{k+1} \rightarrow \alpha_{k+1}, \dots, A_n \rightarrow \alpha_n]$

where  $B_1$  and  $B_2$  are new nonterminals,  $|\alpha_{k1}| = |\alpha_k| - 2$ ,  $|\alpha_{k2}| = 2$  and  $\alpha_k = \alpha_{k1} \alpha_{k2}$ . Then letting this new matrix form be denoted by  $F$ , by the simulation lemma  $\mathcal{L}_S(F, \Rightarrow) = \mathcal{L}_S(G, \Rightarrow)$  and  $\#(G) > \#(F)$  since  $|\alpha_{k1}| < |\alpha_k|$  and no productions with right hand sides longer than two have been introduced.

Now either  $\text{maxr}(F) \leq 2$  or the procedure is repeated. Clearly this is a finite process, hence the Theorem.  $\square$

Notice in the proof of the above theorem that while the matrix form  $G$  is indeed shortened it is shortened at the expense of increasing the size of some of the matrices, that is the number of pro-

ductions in a matrix. For matrix grammars it is well known that matrix grammars whose sizes are at most two need be considered, where for a matrix grammar of form  $G = (V, \Sigma, M, S)$ ,  $\text{size}(G)$  is the maximum size of the matrices in  $M$ . This raises the natural question, namely can a matrix form be reduced to an equivalent matrix form which is both short and whose size is at most 2.

Our approach is the following. Assume  $G = (V, \Sigma, M, S)$  is short and  $\text{size}(G) > 2$ . Let  $m$  be a matrix in  $M$  whose size is greater than two, where

$$m = [A_1 \rightarrow \alpha_1, \dots, A_n \rightarrow \alpha_n], \quad n > 2,$$

then we first replace  $m$  by the matrices

$$m_1 = [A_1 \rightarrow [m, 1], A_2 \rightarrow [m, 2]]$$

$$m_2 = [[m, 2] \rightarrow \alpha_2, A_3 \rightarrow \alpha_3, \dots, A_{n-1} \rightarrow \alpha_{n-1}, A_n \rightarrow [m, n]]$$

$$m_3 = [[m, 1] \rightarrow \alpha_1, [m, n] \rightarrow \alpha_n],$$

giving  $F$ .

If the symbols in  $\alpha_1$  are not equal to any of the  $A_i$ ,  $2 \leq i \leq n$ , this replacement is all that is necessary, since none of the  $A_i \rightarrow \alpha_i$ ,  $2 \leq i \leq n$  is applicable to any symbol in  $\alpha_1$ . Now  $m_3$  cannot be applied before  $m_2$  has been applied and similarly  $m_2$  cannot be applied until  $m_1$  has been applied. Furthermore whenever  $m$  is applicable  $m_1$ ,  $m_2$  and  $m_3$  are applicable and conversely if both  $m_1$  and  $m_2$  are applicable then  $m$  is applicable. Thus in this case it is easy to verify that  $L(F, \Rightarrow) = L(G, \Rightarrow)$  and  $\mathcal{L}_S(F, \Rightarrow) = \mathcal{L}_S(G, \Rightarrow)$ .

When the symbols  $A_2, \dots, A_n$  occur in  $\alpha_1$  we have a more complicated situation, which fortunately we are able to resolve. Consider the case that  $\alpha_1$  is in  $V - \Sigma$ , which only leaves the case  $\alpha_1$  is in  $(V - \Sigma)(V - \Sigma)$ , since  $G$  is short.

This means that when  $m$  is applied in a sentential derivation in  $G$  one of two situations occur. Given  $S \Rightarrow^* \gamma_1 \Rightarrow \gamma_2 \Rightarrow \dots \Rightarrow \gamma_{n+1}$ , where  $\gamma_i = \delta_i A_i \eta_i$  and  $\gamma_{i+1} = \delta_i \alpha_i \eta_i$ ,  $1 \leq i < n$ , we have either  $\alpha_1$  in  $\delta_1 \alpha_1 \eta_1$  is not rewritten using any  $A_i \rightarrow \alpha_i$ ,  $2 \leq i \leq n$  or it is so rewritten. The productions  $m_1$ ,  $m_2$  and  $m_3$  take care of the first situation hence we need to add further productions to  $F$  to take care of the second situation. This we do by adding the following productions:

for all  $k$  such that  $2 \leq k \leq n$  and  $A_k = \alpha_1$  add

$$[A_1 \rightarrow [\bar{m}, k]] \text{ and}$$

$$[A_2 \rightarrow \alpha_2, \dots, A_{k-1} \rightarrow \alpha_{k-1}, [\bar{m}, k] \rightarrow \alpha_k, A_{k+1} \rightarrow \alpha_{k+1}, \dots, A_n \rightarrow \alpha_n].$$

This leaves only the final case, namely  $\alpha_1 = B_1 C_1$ , where  $B_1$  and  $C_1$  are in  $V - \Sigma$ . Essentially the same technique is used except that the difficulty is compounded because of the possible situations that may occur.

These are:

- (a) neither  $B_1$  nor  $C_1$  are rewritten further
- (b)  $B_1$  is rewritten by an  $A_k$ ,  $2 \leq k \leq n$  but  $C_1$  is not
- (c)  $B_1$  is rewritten by an  $A_k$ ,  $2 \leq k \leq n$  and  $C_1$  is rewritten by an  $A_\ell$ ,  $2 \leq \ell \leq n$ ,  $k \neq \ell$ .
- (d)  $B_1$  is not rewritten but  $C_1$  is rewritten by some  $A_\ell$ ,  $2 \leq \ell \leq n$ .

Situation (a) has already been taken care of while (d) is similar to the case that  $\alpha_1$  is in  $V-\Sigma$ . Thus let us consider (b), leaving the analogous case (c) to the reader.

We add the following productions to  $F$ :

for all  $k$  such that  $2 \leq k < n$  and  $A_k = B_1$  add

$[A_1 \rightarrow [m,k,1][m,k,2]]$

$[A_2 \rightarrow \alpha_2, \dots, A_{k-1} \rightarrow \alpha_{k-1}, [m,k,1] \rightarrow \alpha_k, \dots, A_n \rightarrow [m,n]]$

$[[m,1,2] \rightarrow C_1, [m,n] \rightarrow \alpha_n]$

For the case  $A_n = B_1$  we also add

$[A_1 \rightarrow [m,n,1][m,1,2]]$

$[A_2 \rightarrow \alpha_2, \dots, A_{n-1} \rightarrow \alpha_{n-1}, [m,n,1] \rightarrow [m,n]]$

$[[m,1,2] \rightarrow C_1, [m,n] \rightarrow \alpha_n]$

To prove that  $F$  and  $G$  are  $s$ -form equivalent easily reduces to proving that  $\mathcal{L}_s(F, \Rightarrow) \subseteq \mathcal{L}_s(G, \Rightarrow)$ . For an arbitrary interpretation  $F' \triangleleft_S F$  we construct an interpretation  $G' \triangleleft_S G$  such that interpretations of all matrices in  $M - \{m\}$  are retained while interpretations of the new matrices in  $M_F - M$  are replaced by their "compactified" versions, that is we carefully "reverse" the above construction to give interpretations of  $m$ .

This demonstrates that we only need consider short matrix forms which have size at most two.

Having these normal form results enables us to answer questions concerning completeness and vocomplete. For example, it should now be clear that  $F_1$  given by the following matrices:

$[S \rightarrow S, S \rightarrow S], [S \rightarrow S, S \rightarrow SS], [S \rightarrow SS, S \rightarrow S],$

$[S \rightarrow SS, S \rightarrow SS], [S \rightarrow a], [S \rightarrow \lambda], [S \rightarrow S], [S \rightarrow SS]$

is matrix-complete. This follows by observing that every short size two matrix form  $G$  whose only terminal introducing and erasing matrices are of size one is an interpretation of  $F_1$  and also that every matrix form  $H$  can be transformed into such a matrix form  $G$ , which is  $s$ -form equivalent to  $H$ . We can however obtain "smaller" matrix-complete forms by first obtaining stronger normal form results.

Let  $G = (V, \Sigma, M, S)$  be a matrix form such that  $G \triangleleft_S F_1$ , then we can construct an  $H \triangleleft_S F_2$  which is  $s$ -form equivalent to  $G$ , where  $F_2$  is defined by the productions

$[S \rightarrow S, S \rightarrow S]$

$[S \rightarrow a], [S \rightarrow \lambda], [S \rightarrow SS]$

We now construct  $H$  from  $G$  in the following manner. Let  $m$  be in  $M$  and  $m = [A_1 \rightarrow \alpha_1, A_2 \rightarrow \alpha_2]$ . If both  $\alpha_1$  and  $\alpha_2$  are single symbols then  $m$  is added to  $H$ . Otherwise if  $\alpha_2$  consists of two symbols first replace  $m$  in  $M$  by

$m_1 = [A_1 \rightarrow \alpha_1, A_2 \rightarrow [m]]$  and  $m_2 = [[m] \rightarrow \alpha_2]$ .

Thus we only need consider the remaining situation, namely  $\alpha_1$  consists of two symbols and  $\alpha_2$  of exactly one symbol.

Let  $m = [A_1 \rightarrow B_1 C_1, A_2 \rightarrow \alpha_2]$ , then we have four possibilities:

- (a)  $A_2 \neq B_1$  and  $A_2 \neq C_1$ . Add  
 $[A_1 \rightarrow [\bar{m}], A_2 \rightarrow \alpha_2]$  and  $[[\bar{m}] \rightarrow B_1 C_1]$  to  $H$
- (b)  $A_2 = B_1$  and  $A_2 \neq C_1$ . Add  
 $[A_1 \rightarrow [m,1][m,2]], [[m,1] \rightarrow A_2, A_2 \rightarrow [m,3]],$   
 $[[m,2] \rightarrow C_1, [m,3] \rightarrow \alpha_2]$
- (c)  $A_2 \neq B_1$  and  $A_2 = C_1$ . Add  
 $[A_1 \rightarrow [m,1][m,2]], [[m,2] \rightarrow A_2, A_2 \rightarrow [m,3]]$   
 $[[m,1] \rightarrow B_1, [m,3] \rightarrow \alpha_2]$
- (d)  $A_2 = B_1 = C_1$ . Add  
 $[A_1 \rightarrow [m,1][m,1]], [[m,1] \rightarrow A_2, A_2 \rightarrow [m,3]]$   
 $[[m,1] \rightarrow A_2, [m,3] \rightarrow \alpha_2]$

Now the only matrices not of the desired form are those of type  $[A \rightarrow B]$ . However we replace  $m = [A \rightarrow B]$  by  $[A \rightarrow [m], [m] \rightarrow B]$  which is then of the desired form. All other productions in  $G$  can be taken into  $H$  unchanged. Clearly  $G$  and  $H$  are  $s$ -form equivalent and moreover  $H \stackrel{\Delta}{s} F_2$  as desired. Thus  $F_2$  is matrix-complete. By similar techniques we can show that

$F_3: [S \rightarrow S, S \rightarrow SS], [S \rightarrow a], [S \rightarrow \lambda]$

and

$F_4: [S \rightarrow SS, S \rightarrow S], [S \rightarrow a], [S \rightarrow \lambda]$

are matrix-complete. For example, consider  $F_3$ . Let  $G$  be an arbitrary matrix form with  $G \stackrel{\Delta}{s} F_2$ . Construct a new matrix form  $H$   $s$ -form equivalent to  $G$  by first taking matrices of types  $[A \rightarrow a]$  and  $[A \rightarrow \lambda]$  into  $H$  unchanged. Secondly, for every matrix  $m = [A \rightarrow BC]$  in  $G$  take  $[A \rightarrow [m], [m] \rightarrow BC]$  into  $H$  and for every matrix  $m = [A \rightarrow B, C \rightarrow D]$  in  $G$  take  $[A \rightarrow B, C \rightarrow DE]$  and  $[E \rightarrow \lambda]$  into  $H$ . Clearly  $H \stackrel{\Delta}{s} F_3$  and  $L(H, \Rightarrow) = L(G, \Rightarrow)$ . Hence  $F_3$  is matrix-complete. The matrix-completeness of  $F_4$  can be shown similarly.

Notice that each  $F_i$ ,  $i = 1, 2, 3, 4$  is also matrix-vomplete. This involves showing that the transformations used not only preserve the language of the matrix form but also its language family.



The major open problem here is whether or not there exist matrix forms with only two matrices which are matrix-complete. For example are either  $F_5$  or  $F_6$  matrix-complete, where

$$F_5: [S \rightarrow S, S \rightarrow SS], [S \rightarrow a]$$

and

$$F_6: [S \rightarrow SSS, S \rightarrow a], [S \rightarrow \lambda].$$

Under g-interpretation the question is trivially answered since  $\mathcal{L}_S(G, \Rightarrow) \subseteq \mathcal{L}_G(G, \Rightarrow)$  and  $F_3 \triangleleft_g F_5$ .

We close this section on matrix forms by considering context-free-complete and -sufficient matrix forms. Since  $S \rightarrow SS, S \rightarrow a$  is a context-free-complete grammar form, we immediately have that  $[S \rightarrow SS], [S \rightarrow a]$  is a context-free-complete matrix form. This "matrixing" of context-free-complete grammar forms in fact always yield context-free-complete matrix forms, hence our main interest is in two-symbol matrix forms containing matrices of size two. Consider the following seemingly innocent example:

$$F: [S \rightarrow SS], [S \rightarrow a, S \rightarrow \lambda].$$

First observe that

$$S \Rightarrow SS \Rightarrow aS \Rightarrow a \text{ in } F$$

hence we say that  $[S \rightarrow a]$  can be generated by  $[S \rightarrow SS]$  and  $[S \rightarrow a, S \rightarrow \lambda]$ . It is not difficult to show that this implies that for  $G$  defined by:

$$[S \rightarrow SS], [S \rightarrow a]$$

we have  $\mathcal{L}_S(G, \Rightarrow) \subseteq \mathcal{L}_S(F, \Rightarrow)$ . However since  $G$  is context-free-complete we have  $\mathcal{L}(CF) \subseteq \mathcal{L}_S(F, \Rightarrow)$ . Surprisingly this containment is proper as is seen by studying the following interpretation  $F'$  of  $F$ :

$$\begin{aligned} F': [S &\rightarrow S_1 S_4], \\ [S_1 &\rightarrow S_a S_A], [S_1 \rightarrow S_a A] \\ [S_2 &\rightarrow S_b S_B], [S_2 \rightarrow S_b B] \\ [S_3 &\rightarrow S_c S_C], [S_3 \rightarrow S_c C] \\ [S_4 &\rightarrow S_2 S_3] \\ [S_a &\rightarrow a, B \rightarrow \lambda] \\ [S_b &\rightarrow b, C \rightarrow \lambda] \\ [S_c &\rightarrow c, A \rightarrow \lambda] \\ [S_A &\rightarrow S_1 A] \\ [S_B &\rightarrow S_2 B] \\ [S_C &\rightarrow S_3 C]. \end{aligned}$$

Now  $L(F', \Rightarrow) = \{a^n b^n c^n : n \geq 1\}$  hence  $\mathcal{L}(CF) \not\subseteq \mathcal{L}_S(F, \Rightarrow)$ .

In fact there is no context-free-complete  $\{S, a_1, \dots, a_m\}$ -matrix form containing a matrix of size two as we will now demonstrate.

Assume the contrary, namely that there is a matrix form  $G = (\{S, a_1, \dots, a_m\}, \{a_1, \dots, a_m\}, M, S)$ , where  $m > 0$ ,  $\mathcal{L}_S(G, \Rightarrow) = \mathcal{L}(CF)$  and  $M$  contains a matrix of size two.

First observe that  $M$  must contain at least one matrix  $m$  such that  $S \xRightarrow{m} \alpha S \beta S \gamma$  where  $\alpha \beta \gamma$  is in  $\{S, a_1, \dots, a_m\}^*$ . If this is not so sentential derivations never include more than one  $S$ , thus  $\mathcal{L}_S(G, \Rightarrow) \subseteq \mathcal{L}(LIN)$  and hence  $G$  does not generate all context-free languages. Thus we may assume that there is such a matrix  $m$  and furthermore that  $m = [S \rightarrow \beta_1 S \beta_2 S \beta_3]$ .

Second since  $\{a\}$  is a context-free language there is a derivation  $S \Rightarrow^* a_i$  in  $G$ , for some  $i$  and hence we may assume that the matrix  $[S \rightarrow a_i]$  is in  $M$ .

Third let  $r = [S \rightarrow \alpha_1, \dots, S \rightarrow \alpha_{k+2}]$  be a matrix in  $M$  with  $k \geq 0$ . Either the  $\alpha_i$  are completely terminal or they contain at least one  $S$ . Consider the following derivation sequence in  $G$ :

$S \xRightarrow{m}^{k+1} \alpha$ , where  $\alpha$  contains at least  $k+2$   $S$  symbols, and

$\alpha \Rightarrow \alpha'_1$  using  $S \rightarrow a_1$   
 $\alpha'_1 \Rightarrow \alpha'_2$  using  $S \rightarrow a_2$   
 $\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$   
 $\alpha'_k \Rightarrow \alpha'_{k+1}$  using  $S \rightarrow \alpha_{k+1}$

where  $\alpha'_{k+1}$  contains at least one  $S$ , since we have replaced at most  $k+1$   $S$  symbols with terminal words. We may assume  $M$  contains the matrix

$t = [S \rightarrow \alpha'_{k+1}, S \rightarrow \alpha_{k+2}]$

since this is "generated" by  $k+1$  applications of  $m$  followed by an application of  $r$ . Either  $\alpha_{k+2}$  is completely terminal or it contains at least one  $S$ . We will sketch the proof of the fact that when  $\alpha_{k+2}$  is terminal  $\mathcal{L}_S(G, \Rightarrow)$  contains a non-context-free language. The case that  $\alpha_{k+2}$  contains at least one  $S$  symbol follows similarly.

Recalling that  $\alpha'_{k+1}$  contains at least one  $S$ , we write  $\alpha'_{k+1}$  as  $\beta_4 S \beta_5$ . We define an interpretation of  $G$  which only refers to  $m = [S \rightarrow \beta_1 S \beta_2]$ ,  $t = [S \rightarrow \beta_4 S \beta_5, S \rightarrow \alpha_{k+2}]$  and  $[S \rightarrow a_i]$  for some  $i$ ,  $1 \leq i \leq m$ .

Let  $a$ ,  $b$  and  $c$  be new terminal symbols and consider  $G'$  defined by the following matrices:

$$\begin{aligned}
m_1 &= [S \rightarrow \beta_1 A_1 \beta_2 K \beta_3], & m_2 &= [K \rightarrow \beta_1 B_1 \beta_2 C_1 \beta_3], \\
m_3 &= [A_1 \rightarrow \beta_1 A \beta_2 E_1 \beta_3], & m_4 &= [E_1 \rightarrow \beta_1 A_1 \beta_2 A_2 \beta_3], \\
m_5 &= [E_1 \rightarrow \beta_1 E \beta_2 A_2 \beta_3], & m_6 &= [B_1 \rightarrow \beta_1 B_2 \beta_2 D_1 \beta_3], \\
m_7 &= [D_1 \rightarrow \beta_1 B_1 \beta_2 B_3 \beta_3], & m_8 &= [D_1 \rightarrow \beta_1 E \beta_2 B_3 \beta_3], \\
m_9 &= [C_1 \rightarrow \beta_1 C_2 \beta_2 C_1 \beta_3], & m_{10} &= [C_1 \rightarrow \beta_1 C_2 \beta_2 E \beta_3], \\
m_{11} &= [B_2 \rightarrow \beta_4 B \beta_5, A_2 \rightarrow \alpha_{k+2}], & m_{12} &= [C_2 \rightarrow \beta_4 C \beta_5, B_3 \rightarrow \alpha_{k+2}], \\
& [A \rightarrow a], [B \rightarrow b], [C \rightarrow c], [D \rightarrow a_i], [E \rightarrow a_i].
\end{aligned}$$

Now consider an arbitrary derivation in  $G'$ . Essentially using  $m$ , both  $A_1$  and  $K$  are introduced and then applying  $m_3$  and  $m_4$  successfully using  $m_5$  to terminate the appearance of  $A_1$ ,  $n_1$   $A$  symbols and  $n_1$   $A_2$  symbols are introduced. Now  $K$  gives rise to  $B_1$  and  $C_1$ . Successive applications of  $m_6$  and  $m_7$  using  $m_8$  to terminate the appearance of  $B_1$  produce  $n_2$   $B_2$  symbols and  $n_2$   $B_3$  symbols. Similarly  $C_1$  together with  $m_9$  terminating with  $m_{10}$  produces  $n_3$   $C_2$  symbols. Now note that  $m_{11}$  removes a  $B_2$  and an  $A_2$  symbol on each application and  $m_{12}$  removes a  $C_2$  and a  $B_3$  symbol on each application. Furthermore  $m_{11}$  introduces a  $B$  and  $m_{12}$  a  $C$  on each application. Hence the derivation can only terminate if  $n_1 = n_2 = n_3 = n$  say. And in this case

$S \Rightarrow^* u_1(u_2 a u_3)^n u_4(u_5 b u_6)^n u_7(u_8 c u_9)^n u_{10}$ , for some  $u_i$  over  $\{a_1, \dots, a_m\}$ . Now if  $L(G', \Rightarrow)$  is context-free then  $h(L(G', \Rightarrow))$  is also context-free where  $h$  is the homomorphism defined by  $h(a_i) = \lambda$ ,  $1 \leq i \leq m$  and is the identity homomorphism elsewhere. This implies that  $\{a^n b^n c^n : n \geq 1\}$  is context-free, which provides a contradiction.

When  $\alpha_{k+2}$  contains at least one  $S$  symbol we modify the above interpretation such that each  $S$  symbol in  $\alpha_{k+2}$  is renamed as  $\bar{S}$ , say, where the only production in which  $\bar{S}$  appears is  $[\bar{S} \rightarrow a_i]$ .

#### IV.4 Controlled Grammar Forms

To some extent Section 3 has prepared the way for a general investigation of controlled grammar forms. Matrix forms are a specific instance of such forms, since a matrix specifies a localized specific order of application of productions. Letting  $G = (V, \Sigma, P, S)$  be a context-free grammar form we assume in this section that each production in  $P$  is uniquely named, that is

$$p: A \rightarrow \alpha$$

for each  $A \rightarrow \alpha$  in  $P$ . Often it is convenient to consider  $A \rightarrow \alpha$  as its own name. Let

$$S = \beta_0 \Rightarrow \beta_1 \Rightarrow \dots \Rightarrow \beta_m \text{ in } \Sigma^*$$

be a derivation in  $G$ , where  $p_i: A_i \rightarrow \alpha_i$  is the production used in the step  $\beta_i \Rightarrow \beta_{i+1}$ ,  $0 \leq i < m$ . Then we say that

$$\pi = p_0 \dots p_{m-1}$$

is a control word and write  $S \Rightarrow^\pi \beta_m$ . The Szilard language of  $G$ , the set of all control words, is denoted by  $Sz(G, \Rightarrow)$  and defined by:

$$Sz(G, \Rightarrow) = \{\pi: S \Rightarrow^\pi x, \text{ for some } x \text{ in } \Sigma^*\}.$$

Similarly using leftmost derivations we obtain the left Szilard language of  $G$ ,  $Sz(G, \xrightarrow{L}) = \{\pi: S \xrightarrow{L}^\pi x \text{ for some } x \text{ in } \Sigma^*\}$ .

Alternatively we can restrict the derivations in  $G$  by providing a control set  $C \subseteq P^*$ . We say  $(G, C)$  is a controlled context-free grammar (form) if  $G = (V, \Sigma, P, S)$  is a grammar and  $C \subseteq P^*$ . The language of  $(G, C)$  denoted  $L(G, C, \Rightarrow)$  is defined by:

$$L(G, C, \Rightarrow) = \{x: x \text{ is in } \Sigma^* \text{ and } S \xrightarrow{C} x \text{ for some control word } c \text{ in } C\}.$$

As before we may also define  $L(G, C, \xrightarrow{L})$ .

Returning to grammar forms our interest in controlled grammar forms is two-fold, first given a specific grammar form what is its associated class of (left) Szilard languages and second given a grammar form and a family of control sets what properties does the corresponding family of languages have.

To make this more precise, let  $G = (V, \Sigma, P, S)$  be a grammar form.

Then

$$SZILARD(G, \Rightarrow) = \{Sz(G', \Rightarrow): G' \triangleleft_g G\}$$

and

$$SZILARD(G, \xrightarrow{L}) = \{Sz(G', \xrightarrow{L}): G' \triangleleft_g G\}.$$

Note that using  $g$ -interpretation rather than  $s$ -interpretation has no effect on SZILARD, since it is only concerned with (left) control words (or terminating sentential derivations). Moreover we may assume  $\Sigma = \emptyset$  without any loss of generality.

Similarly letting  $\mathcal{L}$  be a family of languages closed under renaming (to avoid difficulties in meaning) then

$$\text{CONTROL}(G, \mathcal{L}, \Rightarrow) = \{L(G', C, \Rightarrow) : G' \xrightarrow{g} G \text{ and } C \text{ is in } \mathcal{L}\}$$

and

$$\text{CONTROL}(G, \mathcal{L}, \xrightarrow{L}) = \{L(G', C, \xrightarrow{L}) : G' \xrightarrow{g} G \text{ and } C \text{ is in } \mathcal{L}\}.$$

In this situation we can consider  $G$  and hence  $\mathcal{L}_g(G, \Rightarrow)$  to be operating on  $\mathcal{L}$ . Thus we speak of a G-control operator. Note that g- and s-interpretation do have different aspects here. However in this case we can always assume that full interpretations are taken since for specific  $G' \xrightarrow{g} G$  and  $C$  in  $\mathcal{L}$  there is a  $G'' \xrightarrow{fg} G$  such that  $G'$  is a subgrammar of  $G''$  and  $L(G'', C, \Rightarrow) = L(G', C, \Rightarrow)$ . Hence  $\text{CONTROL}(G, \mathcal{L}, \Rightarrow) = \{L(G', C, \Rightarrow) : G' \xrightarrow{fg} G \text{ and } C \text{ is in } \mathcal{L}\}$  and similarly for  $\text{CONTROL}(G, \mathcal{L}, \xrightarrow{L})$ .

Let us consider some examples to help gain some understanding of these various families.

Let  $G_1$  be defined by:

$$p: S \rightarrow ASB; q: S \rightarrow \lambda; r: A \rightarrow \lambda; s: B \rightarrow \lambda$$

then

$Sz(G_1, \Rightarrow) = \{w : w \text{ is in } \{p, q, r, s\}^*, |w|_p = |w|_r = |w|_s, |w|_q = 1 \text{ and } q \text{ appears after all the } p\text{'s}\}$ , which is a non-context-free language since

$$Sz(G_1, \Rightarrow) \cap p^*r^*s^*q = \{p^n r^n s^n q : n \geq 0\}.$$

Therefore  $SZILARD(G_1, \Rightarrow)$  is non-context-free also. We claim that every infinite language  $L$  in  $SZILARD(G_1, \Rightarrow)$  is also non-context-free. Let

$L = Sz(G'_1, \Rightarrow)$  for some  $G'_1 \xrightarrow{g} G_1(\mu)$  and  $L$  be infinite. Then letting  $\{p_1, \dots, p_m\} = \mu(p)$ , since  $L$  is infinite there must be a derivation

$$S_1 \Rightarrow A_1 S_2 B_1 \Rightarrow A_1 A_2 S_3 B_2 B_1 \Rightarrow \dots \Rightarrow A_1 A_2 \dots A_k S_{k+1} B_k \dots B_1$$

for some  $k$  and a derivation

$$S_{k+1} \Rightarrow A'_1 S_{k+2} B'_1 \Rightarrow \dots \Rightarrow A'_1 \dots A'_r S_{k+1+r} B'_r \dots B'_1$$

for some  $r$  such that  $S_{k+1} = S_{k+r+1}$ . Otherwise  $Sz(G'_1, \Rightarrow)$  would be finite.

Now let us denote the control words of these derivations by:

$$p_{i_1} \dots p_{i_k} \text{ and } p_{j_1} \dots p_{j_r}.$$

Without loss of generality we can assume each of  $p_{i_1}, \dots, p_{i_k}, p_{j_1}, \dots, p_{j_r}$  are different, otherwise we could find shorter derivations with this property.

Finally consider  $L' = L \cap p_{i_1} \dots p_{i_k} (p_{j_1} \dots p_{j_r})^* \mu(r) \mu(s)^* \mu(q)$ . Clearly  $L'$  is infinite and non-context-free.

On the other hand  $SZILARD(G_1, \xrightarrow{L})$  is context-free since  $Sz(G_1, \xrightarrow{L})$  is always context-free. In this case

$$Sz(G, \Rightarrow) = \{(pr)^n qs^n : n \geq 0\}$$

which is linear. Once more infinite languages in  $SZILARD(G_1, \xrightarrow{L})$  can be shown to be linear and non-regular.

Little study of the SZILARD operator has been made, however this example indicates that such families are highly unstable under typical grammatical operations. For example consider  $G_2$  obtained from  $G_1$  by replacing A and B by  $\lambda$ :

$$p: S \rightarrow S; q: S \rightarrow \lambda$$

then  $Sz(G_2, \Rightarrow) = \{p^n q : n \geq 0\}$ , which is regular. Hence  $SZILARD(G_2, \Rightarrow) \subseteq \mathcal{L}(REG)$ .

Perhaps the most interesting questions here are when is  $SZILARD(F_1, \Rightarrow) = SZILARD(F_2, \Rightarrow)$  and when is  $SZILARD(F_1, \Rightarrow) \subseteq SZILARD(F_2, \Rightarrow)$ . These have recently been shown to be decidable.

We turn therefore to the study of  $CONTROL(G, \mathcal{L}, \Rightarrow)$  and  $CONTROL(G, \mathcal{L}, \overset{\perp}{\Rightarrow})$  in the following. Consider

$$G_3: S \rightarrow aS; S \rightarrow a$$

and

$$G_4: S \rightarrow Sa; S \rightarrow a$$

We have shown earlier that  $\mathcal{L}_g(G_3, \Rightarrow) = \mathcal{L}_g(G_4, \Rightarrow) = \mathcal{L}(REG)$ . However in general

$$CONTROL(G_3, \mathcal{L}, \Rightarrow) \neq CONTROL(G_4, \mathcal{L}, \Rightarrow)$$

since as will be shown, for any full semi-AFL  $\mathcal{L}$ ,  $CONTROL(G_3, \mathcal{L}, \Rightarrow) = \mathcal{L}$  and  $CONTROL(G_4, \mathcal{L}, \Rightarrow)$  is the family of reversals of members of  $\mathcal{L}$ . But not all full semi-AFLs are closed under reversal.

Consider  $G_5$  defined by:

$$S \rightarrow aS; S \rightarrow B; B \rightarrow BC; B \rightarrow a; C \rightarrow \lambda$$

then evidently  $\mathcal{L}_g(G_5, \Rightarrow) = \mathcal{L}(REG)$ . But it can be shown that  $CONTROL(G_5, \mathcal{L}(LIN), \overset{\perp}{\Rightarrow}) = \mathcal{L}(RE)$ .

Essentially the production  $B \rightarrow BC$  in  $G_5$  causes this blow-up, since it introduces an unbounded number of C's (all of which derive the empty word). In other words  $G_5$  is not left derivation bounded. Essentially whenever a grammar for  $G$  is not left derivation bounded then  $CONTROL(G, \mathcal{L}(LIN), \overset{\perp}{\Rightarrow}) = \mathcal{L}(RE)$ .

Consider the grammar  $G_6$ :

$$S \rightarrow D; D \rightarrow E; D \rightarrow DB; D \rightarrow \lambda; E \rightarrow aE; E \rightarrow \lambda; B \rightarrow \lambda$$

again  $\mathcal{L}_g(G_6, \Rightarrow) = \mathcal{L}(REG)$ . However as we shall now prove  $CONTROL(G_6, \mathcal{L}(LIN), \overset{\perp}{\Rightarrow}) = \mathcal{L}(RE)$ .

Given an arbitrary  $G'_6 \xrightarrow{g} G_6$  and an arbitrary  $C$  in  $\mathcal{L}(LIN)$  it is straightforward to show that  $L(G'_6, C, \overset{\perp}{\Rightarrow})$  is in  $\mathcal{L}(RE)$ , therefore we only need to consider the reverse inclusion.

Let  $G = (V, \Sigma, P, Z)$  be a phrase structure grammar, where  $V \cap \{S, D, E\} = \emptyset$ . We will construct a  $G'_6 \xrightarrow{g} G_6(\mu)$  and a linear language  $C$  such that  $L(G'_6, C, \overset{\perp}{\Rightarrow}) = L(G, \Rightarrow)$ , which gives the desired result.

Let  $\$1$ ,  $\$2$ ,  $\phi$ ,  $\bar{c}$  and  $\hat{c}$  be new symbols and let  $\bar{h}$  and  $\hat{h}$  be homomorphisms mapping each symbol  $X$  in  $V$  into new symbols  $\bar{X}$  and  $\hat{X}$  respectively.

Let  $G'_6 = (V', \Sigma, P', S)$  where  $V' = \bar{h}(V) \cup \Sigma \cup \{S, D, E, \bar{c}\}$  and  $P'$  is defined as follows:

$$\$1: S \rightarrow D; \phi: D \rightarrow E;$$

$$\bar{X}: D \rightarrow D\bar{X}, \text{ for all } X \text{ in } V \cup \{\phi\};$$

$$\hat{X}: \bar{X} \rightarrow \lambda, \text{ for all } X \text{ in } V \cup \{\phi\};$$

$$b: E \rightarrow bE, \text{ for all } b \text{ in } \Sigma;$$

$$\$2: E \rightarrow \lambda.$$

Let  $\mu(a) = \Sigma$ ,  $\mu(S) = \{S\}$ ,  $\mu(D) = \{D\}$ ,  $\mu(E) = \{E\}$  and  $\mu(B) = \bar{h}(V) \cup \{\bar{c}\}$ . It is clear that  $G'_6 \stackrel{g}{\sim} G_6(\mu)$ . Let  $C$  consist of all words  $\$1\alpha\$2\beta$  where

$$\alpha = \bar{h}(\alpha_{n-1})\bar{c} \dots \bar{h}(\alpha_1)\bar{c}\bar{h}(Z)\bar{c}\phi\alpha_n,$$

$$\beta = \hat{c}\hat{h}(mi(\beta_1)) \dots \hat{c}\hat{h}(mi(\beta_n)),$$

for  $n \geq 1$ ,  $\beta_i \Rightarrow \alpha_i$  in  $G$ ,  $1 \leq i \leq n$  and  $\alpha_n$  is in  $\Sigma^*$ . We claim that  $C$  is a linear language and we leave the proof of this fact to the reader.

Consider a word  $\delta$  in  $V^*$ , noting that an application of the control word  $\bar{h}(\delta) = \bar{\delta}$  to  $D$  results in  $Dmi(\bar{\delta})$ , that is

$$D \xrightarrow{L\bar{\delta}} Dmi(\bar{\delta})$$

in  $G'_6$ .

Therefore consider the effect of  $\$1\alpha\$2\beta$  on  $S$  in  $G'_6$ , where  $\alpha$  and  $\beta$  are assumed to have the form specified above.

We obtain

$$\begin{aligned} S &\xrightarrow{L\$1} D \xrightarrow{L\bar{h}(\alpha_{n-1})\bar{c}} D\bar{c}mi(\bar{h}(\alpha_{n-1})) \\ &\xrightarrow{L\bar{h}(Z)\bar{c}} D\bar{c}\bar{Z}\bar{c}mi(\bar{h}(\alpha_1)) \dots mi(\bar{h}(\alpha_{n-1})) \\ &\xrightarrow{L\phi\alpha_n\$2} \alpha_n\bar{c}\bar{Z} \dots mi(\bar{h}(\alpha_{n-1})) \end{aligned}$$

The control word  $\beta$  can be applied to the resulting intermediate word to give a terminal word only if all barred symbols are erased. This occurs only if  $\bar{h}(mi(\beta_1)) = \bar{Z}$ , since  $\phi$ ,  $\bar{c}$ ,  $\hat{c}$  are not members of  $V$ . Additionally we must have by a similar argument that

$$\bar{h}(mi(\beta_{i+1})) = mi(\bar{h}(\alpha_i)), \text{ for all } i, 1 \leq i < n$$

in other words  $\beta_1 = Z$  and  $\beta_{i+1} = \alpha_i$ ,  $1 \leq i < n$ . However this implies

$$Z = \beta_1 \Rightarrow \alpha_1 = \beta_2 \Rightarrow \alpha_2 \dots \beta_{n-1} \Rightarrow \alpha_{n-1} = \beta_n \Rightarrow \alpha_n$$

that is  $Z \Rightarrow^* \alpha_n$  in  $G$ , in other words  $\alpha_n$  is in  $L(G, \Rightarrow)$ . By a similar

argument we may also show that if  $Z \Rightarrow^* \alpha_n$  in  $G$  with  $\alpha_n$  in  $\Sigma^*$  then there are  $\alpha$  and  $\beta$  such that  $S \xRightarrow{L \alpha \$ \beta} \alpha_n$  in  $(G'_6, C)$ .

Thus  $\text{CONTROL}(G_6, \mathcal{L}(\text{LIN}), \vdash) = \mathcal{L}(\text{RE})$  and the claim is established.

Observe that in  $G_6$   $D$  is right expansive,  $E$  is partially self-embedding and  $E$  is reachable from  $D$ . Based on this observation it can be proved by an exhaustive case analysis with constructions similar to the one used above that:

If  $G$  is a reduced grammar which contains a right expansive symbol  $D$ , a partially self-embedding symbol  $E$  and one of the following holds:

- (i)  $E$  is reachable from  $D$ .
- (ii)  $D$  is reachable from  $E$ .
- (iii) both  $D$  and  $E$  are simultaneously reachable from the sentence symbol.

Then  $\text{CONTROL}(G, \mathcal{L}(\text{LIN}), \vdash) = \mathcal{L}(\text{RE})$ .

Note that  $G_5$  above fulfills these conditions.

Moreover based on this result it can be further shown that for any reduced grammar form  $G$  either  $\text{CONTROL}(G, \mathcal{L}(\text{LIN}), \vdash) = \mathcal{L}(\text{RE})$  or there exists a left derivation bounded grammar  $H$  such that  $\text{CONTROL}(G, \mathcal{L}, \vdash) = \text{CONTROL}(H, \mathcal{L}, \vdash)$  for every full semi-AFL  $\mathcal{L}$ . In this latter case  $\text{CONTROL}(G, \mathcal{L}(\text{LIN}), \vdash) \not\subseteq \mathcal{L}(\text{RE})$ .

Turning to some of the basic properties of the  $G$ -control operator it is straightforward to prove results analogous to those for grammar forms in Chapter II. Thus when  $G' \xrightarrow{g} G$  we have  $\mathcal{L}_g(G', \Rightarrow) \subseteq \mathcal{L}_g(G, \Rightarrow)$  and analogously we have:

$$\text{CONTROL}(G', \mathcal{L}, \Rightarrow) \subseteq \text{CONTROL}(G, \mathcal{L}, \Rightarrow),$$

and replacing  $\Rightarrow$  by  $\vdash$  gives

$$\text{CONTROL}(G', \mathcal{L}, \vdash) \subseteq \text{CONTROL}(G, \mathcal{L}, \vdash).$$

For grammar forms  $G$  and a derivation  $A \Rightarrow^+ \alpha$  in  $G$  adding the production  $A \rightarrow \alpha$  to  $G$  to give  $\bar{G}$  does not affect its generative power. This is not always valid for controlled grammar forms. However if  $\mathcal{L}$  is a family of languages closed under  $\lambda$ -free finite substitution, for example a full semi-AFL, then

$$\text{CONTROL}(G, \mathcal{L}, \Rightarrow) \subseteq \text{CONTROL}(\bar{G}, \mathcal{L}, \Rightarrow)$$

First assume  $A \Rightarrow^2 \alpha$  in  $G$  where

$$A \Rightarrow \alpha_1 B \alpha_3 \Rightarrow \alpha_1 \alpha_2 \alpha_3 = \alpha$$

in  $G$ . Clearly  $\text{CONTROL}(G, \mathcal{L}, \Rightarrow) \subseteq \text{CONTROL}(\bar{G}, \mathcal{L}, \Rightarrow)$  since  $G \xrightarrow{g} \bar{G}$ . Thus we need to demonstrate the reverse inclusion, namely given a  $\bar{G}' \xrightarrow{g} \bar{G}(\mu)$  and a  $C$  in  $\mathcal{L}$  that  $L(\bar{G}', C, \Rightarrow)$  is in  $\text{CONTROL}(G, \mathcal{L}, \Rightarrow)$ . Construct  $G' \xrightarrow{g} G$  as follows: For each  $p': A' \rightarrow \alpha'$  in  $\mu(p: A \rightarrow \alpha)$  take



$$p_1': A' \rightarrow \alpha_1'[B, p']\alpha_3' \text{ and } p_2': [B, p'] \rightarrow \alpha_2'$$

where  $\alpha_1'\alpha_2'\alpha_3' = \alpha'$  and  $[B, p']$  is a new nonterminal. Also take all other productions in  $\bar{G}'$  unchanged into  $G'$ . Now take as  $C'$  all words in  $C$  in which each appearance of a  $p'$  is replaced by  $p_1'p_2'$ . Then  $L(G', C', \Rightarrow) = L(\bar{G}', C, \Rightarrow)$  and the result follows.

In the case of  $\stackrel{L}{\Rightarrow}$  rather than  $\Rightarrow$  the above construction is complicated since while

$$A \Rightarrow \alpha_1 B \alpha_3 \Rightarrow \alpha_1 \alpha_2 \alpha_3 = \alpha$$

in  $G$  it is not necessarily true that

$$A \stackrel{L}{\Rightarrow} \alpha_1 B \alpha_3 \stackrel{L}{\Rightarrow} \alpha_1 \alpha_2 \alpha_3 = \alpha.$$

Therefore  $C'$  is formed by replacing  $p'$  everywhere by  $p_2' * p_1 p_2'^*$  and all  $p \neq p'$  by  $p_2' * p p_2'^*$ . This ensures that for each terminating derivation in  $\bar{G}'$  there is some control word in  $C'$  which will enable the "corresponding" derivation in  $G'$  to terminate. The converse also holds.

Having proved the result for length two derivations it can then be extended to arbitrary length derivations by induction. The details we leave to the interested reader.

We now restrict our discussion in two ways.

First we only consider  $\stackrel{L}{\Rightarrow}$  from hereon in and second we only consider left derivation bounded grammar forms. Under these restrictions the following result can be proved:

(\*) Let  $G$  be a left derivation bounded grammar form with a production  $A \rightarrow \alpha_1 B \alpha_2$  and  $B \rightarrow \beta_1, \dots, B \rightarrow \beta_r$  be the set of  $B$ -productions. Letting  $\bar{G}$  be the grammar form given by replacing  $A \rightarrow \alpha_1 B \alpha_2$  in  $G$  by the productions  $A \rightarrow \alpha_1 \beta_1 \alpha_2, \dots, A \rightarrow \alpha_1 \beta_r \alpha_2$  then  $\text{CONTROL}(G, \mathcal{L}, \stackrel{L}{\Rightarrow}) = \text{CONTROL}(\bar{G}, \mathcal{L}, \stackrel{L}{\Rightarrow})$  for every full semi-AFL  $\mathcal{L}$ .

This enables us to prove normal form results for controlled left derivation bounded grammar forms analogous to those for grammar forms. This follows since most of the normal form results depend on the fact that substitution holds, that is (\*) above.

For example erasing productions can be removed so that control equivalent  $\lambda$ -free grammar forms can be found. Similarly we can obtain a Chomsky normal form grammar control equivalent to a given grammar form, where by control equivalence of  $G$  and  $H$  we mean: for all full semi-AFLs  $\mathcal{L}$ ,  $\text{CONTROL}(G, \mathcal{L}, \stackrel{L}{\Rightarrow}) = \text{CONTROL}(H, \mathcal{L}, \stackrel{L}{\Rightarrow})$ .

Remarkably, when considering control families which are full semi-AFLs, we are able to restrict interpretations even further, while retaining the same control grammatical family.

We say a  $g$ -interpretation  $G' \stackrel{g}{\triangleleft} G(\mu)$  is simple if  $\mu(A) = \{A\}$  for each nonterminal  $A$  in  $G$ . Thus we may assume that  $G'$  and  $G$  only differ as far as terminal words in their productions are concerned.

We write  $G' \xrightarrow{sg} G(\mu)$  in this case.

For all full semi-AFLs  $\mathcal{L}$  and left derivation bounded grammar forms  $G$ :

$$\text{CONTROL}(G, \mathcal{L}, \underline{\mathbb{L}}) = \{L(G', C, \underline{\mathbb{L}}) : G' \xrightarrow{sg} G \text{ and } C \text{ in } \mathcal{L}\},$$

which we denote by  $s\text{-CONTROL}(G, \mathcal{L}, \underline{\mathbb{L}})$ .

Clearly  $s\text{-CONTROL}(G, \mathcal{L}, \underline{\mathbb{L}}) \subseteq \text{CONTROL}(G, \mathcal{L}, \underline{\mathbb{L}})$  for all  $\mathcal{L}$  since  $G' \xrightarrow{sg} G$  implies  $G' \xrightarrow{g} G$ . Conversely, consider  $G' \xrightarrow{fg} G(\mu)$ , by earlier remarks only full interpretations need be considered. Also let  $C$  be an arbitrary control set in  $\mathcal{L}$ . We need to show that  $L(G', C, \underline{\mathbb{L}})$  is in  $s\text{-CONTROL}(G, \mathcal{L}, \underline{\mathbb{L}})$ .

Define  $\bar{\mu}$  by  $\bar{\mu}(A) = \{A\}$  for all nonterminals  $A$  in  $G$  and  $\bar{\mu}(a) = \mu(a)$  for all terminals  $a$  in  $G$ . Let  $H$  be defined by:

$$H \xrightarrow{sg} G(\bar{\mu}).$$

We now construct an a-transducer  $M$  such that  $L(G', C, \underline{\mathbb{L}}) = L(H, M(C), \underline{\mathbb{L}})$ . Since  $G$  is a left derivation bounded grammar form there is a  $k > 0$  such that any leftmost derivation in  $G$  has at most  $k$  nonterminals at each derivation step. Let  $\pi = p_1 \dots p_n$  be in  $C$  such that

$S \xrightarrow{L\pi} x$   
 in  $G'$  where  $S$  is the sentence symbol of  $G'$  and  $x$  is a terminal word.  $M$  is defined in such a way that given  $\pi$  it produces

$$\bar{p}_1 \dots \bar{p}_n$$

where  $\bar{p}_i$  in  $H$  is defined as follows:

Letting  $p_i$  be  $A_0 \rightarrow x_0 A_1 x_1 \dots A_r x_r$ , where the  $x_j$  are terminal words and the  $A_j$  nonterminal symbols, then  $q_i = \mu^{-1}(p_i)$  equals  $B_0 \rightarrow y_0 B_1 y_1 \dots B_r y_r$  in  $G$ .

Thus we take

$$\bar{p}_i : B_0 \rightarrow x_0 B_1 \dots B_r x_r$$

in  $H$ . Clearly this transformation can be affected by a homomorphism, hence we can define  $h(p_i) = \bar{p}_i$  using the construction above. However  $L(H, h(C), \underline{\mathbb{L}})$  is not necessarily equal to  $L(G', C, \underline{\mathbb{L}})$  since derivations may be permitted by  $(H, h(C))$  which were not permitted in  $(G', C)$ . However these unacceptable derivations have the form

$$S \xrightarrow{L\bar{p}_1 \dots \bar{p}_{i-1}} u B_0 \bar{\alpha} \xrightarrow{L\bar{p}_i} u x_0 B_1 \dots B_r x_r \bar{\alpha}$$

in  $H$ , when

$$S \xrightarrow{L p_1 \dots p_{i-1}} u C_0 \alpha \text{ and } p_i : A_0 \rightarrow x_0 A_1 \dots A_1 x_r$$

are in  $G'$  where  $C_0 \neq A_0$  but  $\mu^{-1}(C_0) = \mu^{-1}(A_0) = B_0$ . Therefore using the fact that  $G'$  is left derivation bounded  $M$  keeps track of the non-terminal subsequence of the word at the current derivation step in the derivation  $\pi = p_1 \dots p_n$ . Since  $G'$  is left derivation bounded  $M$  can

keep this information and check that the  $p_i$  currently being read is applicable to the leftmost nonterminal. If it is then  $\bar{p}_i$  is output and otherwise  $M$  blocks.

In closing this section we consider the closure properties of  $\text{CONTROL}(G, \mathcal{L}, \frac{L}{\Rightarrow})$  for  $\mathcal{L}$  a full semi-AFL. It is not difficult to see that  $\text{CONTROL}(G, \mathcal{L}, \frac{L}{\Rightarrow})$  is closed under union if  $\mathcal{L}$  is closed under union and is trivially closed under homomorphism. Closure under intersection with regular sets follows if  $\mathcal{L}$  is closed under  $\lambda$ -free finite substitutions. The proof of this closure result follows the usual cross-product construction. Finally to obtain inverse homomorphism closure it suffices to consider closure under  $c$ -substitutions, that is each terminal  $a$  is replaced by  $c^*ac^*$ , where  $c$  is a new symbol.

Now if  $G$  is infinite and strongly partially self-embedding (strongly pse), that is for all nonterminals  $A$  apart from the sentence symbol  $A \Rightarrow^+ xAy$  for some terminal words  $x$  and  $y$  not both empty and  $\mathcal{L}$  is a full semi-AFL, it can be shown that  $\text{CONTROL}(G, \mathcal{L}, \frac{L}{\Rightarrow})$  is closed under  $c$ -substitution and hence is itself a full semi-AFL.

Finally, for each left derivation bounded grammar form  $G$  a left derivation bounded grammar form  $H$  can be constructed such that  $G$  and  $H$  are control equivalent and  $H$  is strongly pse. For consider  $G$  to have  $n$  nonterminals. If  $n = 1$  then  $G$  is trivially strongly pse. If  $n > 1$  consider some nonterminal  $A$  which is not the sentence symbol and not pse. We may assume  $G$  is  $\lambda$ -free and has no chain productions without any loss of generality.  $A$  derives some nonempty terminal word. In this case, whenever

$$A \xRightarrow{L^+} \alpha \xRightarrow{L^*} x$$

where  $x$  is a nonempty terminal word, then  $\alpha$  does not contain  $A$ , otherwise  $A$  would be pse. Thus wherever  $A$  appears in productions we can replace it by all right hand sides of all  $A$ -productions. The grammar obtained in this way is control equivalent to  $G$  by our earlier observations. Moreover in this new grammar  $A$  is no longer reachable from the sentence symbol, so both the  $A$  and the  $A$ -productions might just as well be omitted. This new grammar has  $n-1$  nonterminal symbols and hence we obtain the required result by induction on the number of nonterminal symbols.

Thus we have:

Let  $G$  be an infinite left derivation bounded grammar form and  $\mathcal{L}$  a full semi-AFL, then  $\text{CONTROL}(G, \mathcal{L}, \frac{L}{\Rightarrow})$  is a full semi-AFL.

Can this result be strengthened to all infinite grammar forms? It is known, see Section II.4.3 that  $\mathcal{L}_g(G, \Rightarrow)$  is a full principal semi-AFL for all infinite  $G$ . Is this true for  $\text{CONTROL}(G, \mathcal{L}, \frac{L}{\Rightarrow})$  also?

Alternatively when is it true?

One natural family to consider may be defined as  $\text{CONTROL}(G, \mathcal{L}, \overset{L}{\Rightarrow})$  where  $\mathcal{L} = \text{CONTROL}(H, \mathcal{K}, \overset{L}{\Rightarrow})$  for some  $H$  and  $\mathcal{K}$ . A special case of this general definition is given by  $\mathcal{L} = \mathcal{L}_g(H, \overset{L}{\Rightarrow})$  for some  $H$ . What can be said about such families?