Some Results on Small Context-free Grammars Generating Primitive Words [1]

Pál DÖMÖSI
Institute of Mathematics and Informatics, L. Kossuth University
Debrecen, Egyetem tér 1, H-4032, Hungary
e-mail: domosi@math.klte.hu

and

Dirk HAUSCHILDT
Fachbereich Informatik, Universität Hamburg
Vogt-Kölln-Strasse 30, D-22527 Hamburg
e-mail: hauschil@informatik.uni-hamburg.de

and

Géza HORVÁTH
Institute of Mathematics and Informatics, L. Kossuth University
Debrecen, Egyetem tér 1, H-4032, Hungary
e-mail: gezian@tigris.klte.hu

and

Manfred KUDLEK
Fachbereich Informatik, Universität Hamburg
Vogt-Kölln-Strasse 30, D-22527 Hamburg
e-mail: kudlek@informatik.uni-hamburg.de

Abstract

In this report we characterize all context-free grammars with not more than three nonterminals generating only primitive words.

[1]This work was supported by the German-Hungarian Research Project 102 of OMFB (Hungarian Research Foundation) and Project X 233.6 of Forschungszentrum Karlsruhe

1 Introduction

A number of recent papers investigated the language of all primitive words over an alphabet with several letters, concerning its relation to the Chomsky-hierarchy (see [1]-[7]). In [1] the authors conjectured that this language is not context-free. This conjecture is still open. To help the research on this problem, in this report we consider certain 'small' and 'maximal' context-free grammars in Chomsky normal form generating only primitive words. These grammars are small with respect to nonterminals and maximal with respect to productions. Since a necessary condition for the generated language to contain only primitive words (over terminal symbols) is that all sentential forms are also primitive words (over nonterminals) it suffices to consider only the sentential form languages.

It was the hope to deduce from the structure of such grammars some insight for a proof of the conjecture that the entire set of primitive words is not context-free by showing that there are always missing certain primitive words in the language generated by the grammar.

Another conjecture was that any such grammar (and also all non-maximal ones) generate only regular sets of primitive words. This conjecture does not hold since we show that there exists a maximal grammar with 3 nonterminals generating a nonregular infinite set of primitive words.

In this paper we characterize all context-free grammars with not more than three nonterminals generating only primitive words. It turned out that all such grammars generate infinite sets of primitive words.

2 Preliminaries

A *word* is a finite sequence of elements of some finite nonempty set Σ. We call the set Σ an *alphabet,* the elements of Σ *letters.* The set of all words over Σ is denoted by Σ^*. We put $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$, where λ denotes the *empty word* having no letters. The *length* of a word w, in symbols $|w|$, means the number of letters in w when each letter is counted as many times as it occurs. By definition, $|\lambda| = 0$. If u and v are words over an alphabet Σ, then their *catenation* uv is also a word over Σ. Especially, for any word uvw, we say that v is a *subword* of uvw.

A *language* over Σ is a set $L \subseteq \Sigma^*$. We extend the concept of catenation for the class of languages as usual. Therefore, if L_1 and L_2 are languages, then their product is $L_1L_2 = \{p_1p_2 \mid p_1 \in L_1, p_2 \in L_2\}$. Let p be a word. We put $p^0 = \lambda$ and $p^n = p^{n-1}p$ $(n > 0)$. Thus p^k $(k \geq 0)$ is the *k-th power* of p. If there is no danger of confusion, then sometimes we identify p with the singleton set $\{p\}$. Thus we will write p^* and p^+ instead of $\{p\}^*$ and $\{p\}^+$, respectively. A nonempty word is said to be *primitive* if it is not a proper power ($k > 1$) of another word. A word is *nonprimitive* if it is not primitive. Let Q_Σ denote the set of primitive words over Σ.

An *(unrestricted generative, or simply, unrestricted) grammar* is an ordered quadruple $G = (N, \Sigma, S, P)$ where N and Σ are disjoint alphabets, $S \in \Sigma$, and P is a finite set of ordered pairs (U, V) such that V is a word over the alphabet $N \cup \Sigma$ and U is a word over $N \cup \Sigma$ containing at least one letter of N. The elements of N are called *variables* or *nonterminals*, and those of Σ *terminals*. $N \cup \Sigma$ is the *total alphabet* and S is called the *start symbol*. Elements (U, V) of P are called *productions* and are written $U \to V$. If $U \to V \in P$ implies $U \in N$ then G is called *context-free*. Especially, G is a context-free grammar given in *Chomsky normal form* if all productions are of the form $A \to BC$ or $A \to a$, where A, B, C are variables and a is a terminal.

A word W over $N \cup \Sigma$ *derives directly* a word W', in symbols, $W \xrightarrow{1} W'$, if and only if there are words W_1, U, W_2, V such that $W = W_1UW_2, W' = W_1VW_2$ and $U \to V$ belongs to P. W *derives* W', or in symbols, $W \xrightarrow{*} W'$ if and only if there is a finite sequence of words $W_0, \ldots, W_k (k \geq 0)$ over $N \cup \Sigma$ with $W_0 = W, W_k = W'$ and $W_i \xrightarrow{1} W_{i+1}$ for $0 \leq i \leq k-1$. Thus for every $W \in (N \cup \Sigma)^*$ we have $W \xrightarrow{*} W$.

The set $S(G) = \{W \mid W \in (N \cup \Sigma)^*, S \xrightarrow{*} W\}$ is called the set of *sentential forms* of G. The *language* $L(G)$ *generated by* G is defined by $L(G) = S(G) \cap \Sigma^*$. $L \subseteq \Sigma^*$ is a *context-free language* if we have $L = L(G)$ for some context-free grammar G.

The grammar $G_1 = (N_1, \Sigma_1, S_1, P_1)$ is *letter-isomorphic* to another grammar $G_2 = (N_2, \Sigma_2, S_2, P_2)$ if there exists a bijective mapping $\varphi : N_1 \cup \Sigma_1 \to N_2 \cup \Sigma_2$ such that $\varphi(S_1) = S_2, \{\varphi(A) \mid A \in N_1\} = N_2$, $\{\varphi(a) \mid a \in \Sigma_1\} = \Sigma_2$, moreover, $\{\varphi(x_1) \ldots \varphi(x_s) \to \varphi(y_1) \ldots \varphi(y_t) \mid x_1 \ldots x_s \to y_1 \ldots y_t \in P_1\} = P_2$. In this report we will not distinguish the letter-isomorphic grammars. Throughout this report by a *grammar* $G = (N, \Sigma, S, P\}$ we mean a (λ-free) context-free grammar given in Chomsky normal form.

For any terminal symbol x we consider the set $N(x) = \{X \in N \mid X \to x \in P\}$. We say that $x \in \Sigma$ is *similar* to $y \in \Sigma$ with respect to $M \subseteq N$ with $M \neq \emptyset$ if $M \subseteq N(x) \cap N(y)$. (Then we also say, in short, that x is similar to y.)

A grammar G is *reduced* if it has the following properties :

(i) For any pair x, y of terminal symbols, $N(x) = N(y)$ implies $x = y$.

(ii) For any $x \in N \cup \Sigma$, there exists a pair $W_1, W_2 \in (N \cup \Sigma)^*$ such that the word $W_1xW_2 \in S(G)$.

For $X \in N$ let $\Sigma(X) = \{x \in \Sigma \mid X \to x \in P\}$ where also $\Sigma(X) = \emptyset$ is possible.

We shall restrict our investigations to reduced grammars.

Now we define the *skeleton* of $G = (N, \Sigma, S, P)$ as $G_0 = (N, S, P_0)$ with productions $P_0 = \{A \to BC \in P \mid A, B, C \in N\}$. The set $S(G_0) = \{W \in N^+ \mid S \xrightarrow{*} W\}$ is called the *(sentential form) language* generated by the skeleton G_0. We also say that a skeleton G_0 is *maximal (with respect to the primitive words)* if $S(G_0)$ contains only primitive words. Moreover, for any $X, Y, Z \in N, X \to YZ \notin P_0$ we obtain a nonprimitive word in $S(G_0')$ with $G_0' = (N, S, P_0')$ and $P_0' = P_0 \cup \{X \to YZ\}$.

Note that $L(G) \subseteq Q_\Sigma \Rightarrow S(G_0) \subseteq Q_N$.

The opposite implication $S(G_0) \subseteq Q_N \Rightarrow L(G) \subseteq Q_\Sigma$ holds if $\Sigma(X) \cap \Sigma(Y) = \emptyset$ for all $X, Y \in N$ with $X \neq Y$.

To show this consider a binary derivation tree for $w \in L(G)$. Cutting off all leaves $x \in \Sigma^+$, generated by some productions $X \to x$, yields a binary derivation tree for $W \in N^+$ with $W \xrightarrow{*} w$. The condition $\Sigma(X) \cap \Sigma(Y) = \emptyset$ implies that for any $x \in \Sigma$ there exists a unique $X \in N$ with $x \in \Sigma(X)$. Define the letter-to-letter homomorphism $c : \Sigma \to N$ by $c(x) = X$ if $x \in \Sigma(X)$. From $w = u^k$ for $k > 1$ follows that $W = c(w) = (c(u))^k$, a contradiction.

Note that for any context-free grammar G there exists an equivalent context-free grammar G' in Chomsky normal form with this property. To show this assume G to be in Chomsky normal form, add new nonterminals $\overline{\Sigma}$, replace each production $X \to YZ$ by the productions $\{X \to YZ, X \to \bar{y}Z, X \to Y\bar{z}, X \to \bar{y}\bar{z}\}$ with $y \in \Sigma(Y), z \in \Sigma(Z)$, and productions $X \to x$ by $\bar{x} \to x$. Thus, in $G' : \Sigma(X) = \emptyset$ for $X \in N$, and $\Sigma(\bar{x}) = \{x\}$ for $x \in \Sigma$.

3 Maximal Skeletons with 1 and 2 Nonterminals

Maximal Skeleton with 1 Nonterminal

If $|N| = 1$ then the only maximal skeleton is $G_0 = (N, S, \emptyset)$, and the only reduced grammar is $G = (\{S\}, \{s\}, \{S \to s\}, S)$.

□

Maximal Skeleton with 2 Nonterminals

If $|N| = 2$ then the only maximal skeleton is G_0 with
$P_0 = \{S \to SX, S \to XS, X \to XX\}$
$S(G_0) = \{X\}^* \cdot \{S\} \cdot \{X\}^*$ (only productions $\{S \to SX, S \to XS\}$ are necessary) and $S(G_0) \subset Q_N$ is obvious.
The reduced grammars have the form
$G = (\{S, X\}, \{s, x\}, P_1 \cup P_2 \cup \{S \to s, X \to x\}, S)$, where
$P_1 \subseteq \{S \to SX, S \to XS\}, P_2 \subseteq \{X \to XX\}, P_1 \neq \emptyset$.

□

For any fixed cardinality of nonterminals we may characterize all reduced grammars by using the characterization of maximal skeletons. (If $|N| > 2$ then we have to take into consideration the similarity possibilities of terminals as well.)

4 Maximal Skeletons with 3 Nonterminals

Using an appropriate computer program (written by Géza Horváth [4]), checked and improved by Dirk Hauschildt, we found 11 different maximal skeleton candidates, up to symmetries, with 3 nonterminals (S, X, Y).

These symmetries are σ defined by $\sigma(X) = Y, \sigma(Y) = X$, and π defined by $\pi(A \to BC) = A \to CB$, with the properties $\sigma^2 = \pi^2 = 1, \pi\sigma = \sigma\pi$.

The computer program in question checked that none of these 11 skeletons generates a nonprimitive word W of nonterminals with length $|W| \leq 12$ (in the improved version it turned out that $|W| \leq 10$ suffices).

It was run in some dialogue way in several steps, using an input list of skeletons generating some nonprimitive word such that any enlarged skeleton (some productions added) could be disregarded. Another list contained only such skeletons generating no nonprimitive word with $|W| \leq 10$ such that any skeleton with a subset of productions could be disregarded. Finally we got a list of 11 candidates for maximal skeletons (with respect to primitive words, and up to symmetries). The program is given in the appendix.

In more details, given a $n \in I\!N$ the program computes a set $\mathcal{N}$ of 'minimal' skeletons generating some nonprimitive words of length $\leq n$, and a set $\mathcal{P}$ of 'maximal' skeletons not generating such words.

To speed up running time the program can also read in (in advance) known elements from $\mathcal{N}$ and $\mathcal{P}$, such that only skeletons not included by an element of $\mathcal{P}$ and not including an element of $\mathcal{N}$ have to be checked.

The program was run first for $n = 6$ yielding sets $\mathcal{N}$ and $\mathcal{P}$. Some elements of $\mathcal{P}$ were checked by hand for generating only primitive words or generating some nonprimitive word of length $n > 6$. In the second case such an element was removed from $\mathcal{P}$ changing the set $\mathcal{P}$. In the next run with $n = 8$ some new elements were added to $\mathcal{P}$ (being included in some of the shifted ones).

By repeating this procedure for n=9 and n=10 we finally got 12 candidates for maximal skeletons, one of which was not reduced.

There exist no more skeletons with the property from above. In this section we prove that each of them is a maximal skeleton indeed. (We note that apart from the last case the first version of these proofs have been published by Géza Horváth [4]).

Consider $N = \{S, X, Y\}$ with the start symbol S, and denote by $Q = Q_N$ the set of all primitive words over N. We distinguish the following 11 cases.

Case 1.

$P_0 = \{S \to XY, S \to SX, S \to XS, S \to YX, X \to XX,$
$\quad Y \to SX, Y \to XS, Y \to XY, Y \to YX\}.$

$S(G_0) = (X^* \cdot \{S, Y\} \cdot X^*) \setminus \{Y\} \subset Q.$

This is shown in the following way :

Let $L = X^*(\{S\} \cup \{ Y\})X^* \setminus \{Y\} \subset Q$. Induction on $W \in L$, namely $S \in L$, and any application of a production from P_0 on some $W \in L$ yielding again some $W' \in L$, implies $S(G_0) \subseteq L$.

On the other hand, any $W \in L$ can be derived from S. $S, XY, YX \in L$ is obvious. $X^mSX^n \in L$ by $S \xrightarrow{m} X^mS \xrightarrow{n} X^mSX^n$ with productions $\{S \to XS, S \to XS\}$, and $X^mYX^n \in L$ ($m > 0$) by $S \xrightarrow{m-1} X^{m-1}S \xrightarrow{n} X^{m-1}SX^n \xrightarrow{1} X^mYX^n$. This implies $L \subseteq S(G_0)$.

Note that only productions $\{S \to XS, S \to SX, S \to XY, S \to YX\}$ have to be applied.

$S(G_0) \subseteq Q$ is obvious since any $W \in S(G_0)$ contains either only 1 S or 1 Y.

$SY \notin S(G_0)$ implies $S(G_0) \subset Q$.

$\square$

Case 2.

$P_0 = \{S \to XY, S \to SX, S \to XS, S \to YX, X \to XX, Y \to YY\}$.

$S(G_0) = (X^* \cdot (\{S\} \cup Y^+) \cdot X^*) \setminus Y^+ \subset Q$.

The proof is similar to case 1, for $S(G_0) \subseteq L$ showing by induction that any application of a production yields again an element from L, and for $L \subseteq S(G_0)$ with the difference that also $Y \to YY$ is applied.

Here, only productions $\{S \to XS, S \to SX, S \to XY, S \to YX, Y \to YY\}$ have to be applied.

$S(G_0) \subset Q$ follows from the fact that each word $W \in S(G_0)$ has the form $W = X^m SXn$ or $W = X^m Y^k X^n$, and from $SY \notin S(G_0)$.

$\square$

Case 3.

$P_0 = \{S \to XY, S \to SY, S \to XS, X \to XX, Y \to YY\}$.

$S(G_0) = (X^* \cdot (\{S\} \cup X^+) \cdot Y^*) \setminus X^+ \subset Q$.

Again, the proof is similar to case 1. To show $L \subseteq S(G_0)$, apply the derivations $S \overset{n}{\to} SY^n \overset{m}{\to} Y^m SX^n$ and $S \overset{m-1}{\to} X^{m-1}S \overset{n}{\to} X^{m-1}SY^n \overset{1}{\to} X^m Y^n$ ($m > 0$) implying $SY^n, X^m SY^n, X^m Y^n \in L$.

Here, only productions $\{S \to XS, S \to SY, S \to XY\}$ have to be used.

$S(G_0) \subseteq Q$ is obvious since all $W \in S(G_0)$ have the forms $W = X^m SY^n$ or $W = X^m Y^n$, and since $SX \notin S(G_0)$, also $S(G_0) \subset Q$.

$\square$

Case 4.

$P_0 = \{S \to XS, S \to YS, S \to SX, S \to SY, X \to XX, X \to XY,$
$X \to YX, X \to YY, Y \to XX, Y \to XY, Y \to YX, Y \to YY\}$.

$S(G_0) = \{X, Y\}^* \cdot \{S\} \cdot \{X, Y\}^* \subset Q$.

Similar to case 1 again. To show $L \subseteq S(G_0)$, any USV with $U, V \in \{X, Y\}^*$ is derived by using only productions $\{S \to XS, S \to YS, S \to SX, S \to SY\}$.

Since any $W \in S(G_0$ contains exactly 1 S, and $XY \notin S(G_0)$, follows that $S(G_0) \subset Q$.

$\square$

Case 5.

$P_0 = \{S \to XS, X \to SY, X \to XX, X \to YY\}$.

$S(G_0) = \{X, SY, YY\}^* \cdot \{S\}$.

Again similar to case 1 ($SY \overset{1}{\to} X \cdot SY$ and at the end $S \overset{1}{\to} X \cdot S$). $L \subseteq S(G_0)$ follows from the derivations $S \overset{1}{\to} XS, S \overset{1}{\to} XS \overset{1}{\to} XXS, S \overset{1}{\to} XS \overset{1}{\to} SYS$. Here, only productions $\{S \to XS, X \to XX, X \to SY\}$ are used.

$W = U^k \in S(G_0)$ with $k > 1$ implies $U = YU'S$, a contradiction, since U must end in S but the next U start with Y. Therefore, since also $SX \notin S(G_0)$, follows that $S(G_0) \subset Q$.

$\square$

Case 6.

$P_0 = \{S \to XS, X \to SY, X \to XX, X \to XY, Y \to YX, Y \to YY\}$

$S(G_0) = \{X, SY\} \cdot \{X, Y, SY\}^* \cdot \{S\} \cup \{S\} \subset Q$.

Also similar to case 1 ($SY \xrightarrow{1} X \cdot SY$). $L \subseteq S(G_0)$ follows from the derivations $S \xrightarrow{1} XS, X \xrightarrow{1} XX, X \xrightarrow{1} XY$, and $X \xrightarrow{1} SY \xrightarrow{1} XSY$.

Only productions $\{S \to XS, X \to XX, X \to XY, X \to SY\}$ have to be applied.

Again, $W = U^k \in S(G_0)$ with $k > 1$ implies $U = YU'S$, a contradiction. Since also $SX \notin S(G_0)$ follows that $S(G_0) \subset Q$.

□

Case 7.

$P_0 = \{S \to XY, S \to SY, Y \to XS, Y \to YY\}$.

$S(G_0) = \{S, X\} \cdot \{Y, XS, XX\}^* \cdot \{Y, XS\} \cup \{S\} \subset Q$.

(a) $S(G_0) \subseteq L$ is shown similar to case 1.

(b) $L \subseteq S(G_0)$ is a consequence from the following derivations.

$S \xrightarrow{1} SY, S \xrightarrow{1} XY, S \xrightarrow{n} SY^n, Y \xrightarrow{1} XS \xrightarrow{1} XSY, Y \xrightarrow{1} XS \xrightarrow{1} XXY, Y \xrightarrow{1} YY$ where all productions are used.

(c) To show $S(G_0) \subset Q$ 4 possibilities have to be considered. For this assume $W \in S(G_0) \setminus Q$, i.e. $W = U^k$ with $k > 1$.

(ca) $W = SVS$. Then $U = SU'XS$, a contradiction since SS is not a subword of any $W \in S(G_0)$.

(cb) $W = SVY$. Then $U = SU'Y$. But the number $n_i(X)$ of X in all blocks starting with S, except the last one, is $n_i(X) = 2m + 1$, whereas in the last one it is $n_j(X) = 2n$. A contradiction.

(cc) $W = XVS$. Then $U = XU'XS$. Here $n_1(X) = 2m$ in the first block ending with S, but $n_i(X) = 2n + 1$ in all other such blocks. Again a contradiction.

(cd) $W = XVY$. Then $U = XU'Y$. Here $n_1(X) = n_1(S) + 2m + 1$ in the first block ending with Y, but $n_i(X) = n_i(S) + 2n$ in all other such blocks. Also a contradiction.

Therefore $S(G_0) \subset Q$ since also $SX \notin S(G_0)$.

□

Case 8.

$P_0 = \{S \to XY, S \to YX, X \to SS\}$.

From the productions follows that the numbers $(n(S), n(X), n(Y))$ of S, X, Y, starting with $(1, 0, 0)$, fulfill the identity $n(Y) = 2n(X) + n(S) - 1$. Now, if $W = U^k$ with $k > 1$, then $n(S) = kn_S, n(X) = kn_X, n(Y) = kn_Y$, a contradiction.

Thus, $S(G_0) \subset Q$, since $SX \notin S(G_0)$ again.

□

Case 9.

$P_0 = \{S \to XY, X \to XS, Y \to SY\}$.

From the productions follows for the numbers $n(X), n(Y)$ of X, Y the identity $n(X) = n(Y)$. Furthermore, any $W \in S(G_0)$ with $W \neq S$ has the form $W = XVY$.

By induction follows for any proper prefix of any $W \in S(G_0) : n(X) > n(Y)$. This is obvious for $W = XY$. Applying $X \to XS$ or $Y \to SY$ does not change the number of x or Y. The application of $S \to XY$ either increases the number of X in a prefix by 1 or increases both the numbers of X and Y by 1.

Now, if $W = U^k$ with $k > 1$, then $n(X) = kn_X, n(Y) = kn_Y$ implying $n_X = n_Y$ where n_X, n_Y are the numbers of X, Y in the prefix U. A contradiction.

Again, $SX \notin S(G_0)$. Thus $S(G_0) \subset Q$.

In this case the set $S(G_0)$ is not regular, namely
$(S(G_0) \cap \{X, Y\}^*) \cup \{\lambda\} = h(S(G_0)) = \{X\} \cdot D(X, Y) \cdot \{Y\} \cup \{\lambda\}$,
where $h : \{S, X, Y\} \to \{X, Y\}$ is defined by $h(S) = \lambda$, $h(X) = X$, $h(Y) = Y$, and $D(X, Y)$ is the Dyck language over $\{X, Y\}$.

Any $W \in D(X, Y)$ with $W \neq \lambda$ has a unique representation as

$$W = \prod_{i=1}^{k} (X \cdot U_i \cdot Y)$$

with $U_i \in D(X, Y)$.

Now $S \to XY$ and the induction assumption $XUY \in S(G_0)$ forall $U \in D(X, Y)$ with $|U| < |W|$ gives

$$S \to XY \xrightarrow{*} XS^kY \xrightarrow{*} X \cdot (\prod_{i=1}^{k} XU_iY) \cdot Y$$

implying $XWY \in S(G_0) \cap \{X, Y\}^*$ for all $W \in D(X, Y)$.
Thus $\{X\} \cdot D(X, Y) \cdot \{Y\} \subseteq S(G_0) \cap \{X, Y\}^*$.
On the other hand : $h(S(G_0)) \subseteq \{X\} \cdot D(X, Y) \cdot \{Y\} \cup \{\lambda\}$.
This follows by induction on the number of derivation steps :
$h(S) = \lambda$ and $h(XY) = XY \in D(X, Y)$. Now assume $h(W) \in D(X, Y)$.
If $W = USV \to UXYV = W'$ then $h(W') \in D(X, Y)$,
if $W = UXV \to UXSV = W'$ then $h(W') = h(W) \in D(X, Y)$,
and if $W = UYV \to USYV = W'$ then $h(W') = h(W) \in D(X, Y)$.

Therefore
$\{X\} \cdot D(X, Y) \cdot \{Y\} \subseteq S(G_0) \cap \{X, Y\}^* \subseteq h(S(G_0)) \subseteq \{X\} \cdot D(X, Y) \cdot \{Y\} \cup \{\lambda\}$,
yielding
$S(G_0) \cap \{X, Y\}^* \cup \{\lambda\} = h(S(G_0)) = \{X\} \cdot D(X, Y) \cdot \{Y\} \cup \{\lambda\}$.
This implies that $S(G_0)$ is not regular.

□

Case 10.

$P_0 = \{S \to XS, S \to SX, X \to YS, X \to SY, X \to XX, Y \to XY, Y \to YX\}$.

From the productions follows for the numbers $n(S), n(Y)$ of S, Y the identity $n(Y) = n(S) - 1$. Furthermore, any $W \in S(G_0)$ contains at least 1 S. Now, if $W = U^k$ with $k > 1$, then $n(S) = kn_S, n(Y) = kn_Y$, a contradiction.

Here, $SY \notin S(G_0)$. Therefore, $S(G_0) \subset Q$.

□

Case 11.

$P_0 = \{S \to XY, X \to SY, Y \to XS\}$.

(a) Let $P_1 = \{S \to SYY, Y \to SYS\}$, $G_1 = (N, S, P_1)$, and the homomorphism h be defined by $h(X) = SY, h(S) = S, h(Y) = Y$.

(b) $S(G_1) = S(G_0) \cap \{S, Y\}^* = h(S(G_0))$.

$S(G_1) \subseteq S(P_0) \cap \{S, Y\}^*$ follows from the fact that each production of G_1 is a derivation in $G_0 : S \xrightarrow{1} XY \xrightarrow{1} SYY, Y \xrightarrow{1} XS \xrightarrow{1} SYS$.

On the other hand, consider any $W \in S(G_0) \cap \{S, Y\}^*$ and its derivation tree. Any node with X is either generated by $S \to XY$ or $Y \to XS$. Since no leaf is labelled by X all internal nodes with X have successors S, Y generated by $X \to SY$. But this can be combined into 3 successors of S or Y generated in G_1 either by $S \to SYY$ or $Y \to SYS$, respectively, yielding a ternary derivation tree for G_1. Thus, $S(G_0) \cap \{S, Y\}^* \subseteq S(G_1)$.

$h(G_0)) \subseteq S(G_0) \cap \{S, Y\}^*$ since $h(X) = SY$ has the same effect as applying the production $X \to SY$ to any X.

$S(G_1) = S(G_0) \cap \{S, Y\}^* \subseteq h(S(G_0))$ since $S(G_1) \subseteq S(G_0)$, and therefore also $S(G_1) = h(S(G_1)) \subseteq h(S(G_0))$.

From this follows that $S(G_0) \subset Q \Leftrightarrow S(G_1) \subset Q_{\{S,Y\}}$.

(c) Now, each $W \in S(G_1)$ has the form

$$W = (\prod_{i=0}^{m-1} S^{t_i} \cdot Y) \cdot S^{t_m}$$

with $m \geq 0$ and $t_i \geq 0$, where the product stands for catenation.

Furthermore, by induction on the application of productions, it follows that $n(Y) = m = 2n_Y$ and $n(S) = 2n_S + 1$ (since $S \to SYY$ increases $n(Y)$ by 2, and $Y \to SYS$ $n(S)$ by 2). To generate W the production $S \to SYY$ is applied n_Y times.

Let $t_{i,j}$ with $t_{0,0} = 1$ denote the number of S in a S-block, i giving the index of Y and j the derivation steps. In a derivation any application of $Y \to SYS$ yields $t_{i,j+1} = t_{i,j} + 1$ and $t_{i+1,j+1} = t_{i+1} + 1$ for some $i \geq 0$. Thus, 1 block of S with even index and 1 block of S with odd index is increased by 1. Any application of $S \to SYY$ yields $t_{k,j+1} = t_{k,j}$ for $k < i$, $t_{i,j+1} + t_{i+2,j+1} = t_{i,j}$, $t_{i+1,j+1} = 0$, $t_{k+2,j+1} = t_{k,j}$ for $k > i$, for some i. Hence, the property of an index i to be even or odd is not changed.

Thus,

$$\sum_{i=0}^{n_Y} t_{2i} = 1 + \sum_{i=0}^{n_Y} t_{2i+1}.$$

(d) Now, if $W = U^k$ with $k > 1$ for some $U \in \{S, Y\}^+$, then

$$U = (\prod_{i=0}^{n-1} S^{t_i} \cdot Y) \cdot S^{t_m}$$

with $m = kn \equiv 0 \ (mod\ 2)$.

Since $n_W(S) = kn_U(S) \equiv 1 \ (mod\ 2)$ and $n_W(Y) = kn_U(Y) \equiv 0(mod\ 2)$ it follows that $k \equiv 1 \ (mod\ 2)$, $n_U(S) \equiv 1 \ (mod\ 2)$, and $n = n_U(Y) \equiv 0 \ (mod\ 2)$. Thus $n_U(Y) = 2n'$.

Furthermore,
$\forall\ 0 < i < n \forall\ 0 \leq j < k : t_i = t_{i+jn}$, and
$\forall\ 0 \leq j < k : t_{n+jn} = t_0 + t_m$
Since $W = U^k$ it follows that

$$\sum_{i=0}^{n_Y} t_{2i+1} = k \cdot \sum_{i=0}^{n'} t_{2i+1}$$

and

$$\sum_{i=0}^{n_Y} t_{2i} = k \cdot (\ (t_0 + t_m) + \sum_{i=0}^{n'-1} t_{2i})$$

yielding

$$\sum_{i=0}^{n_Y} t_{2i} - \sum_{i=0}^{n_Y} t_{2i+1} \equiv 0(mod\ k),$$

a contradiction to

$$\sum_{i=0}^{n_Y} t_{2i} - \sum_{i=0}^{n_Y} t_{2i+1} \equiv 1(mod\ k).$$

Again, $SX \notin S(G_0)$. Thus $S(G_1) \subset Q_{\{S,Y\}}$, and therefore $S(G_0) \subset Q$.

□

To these 11 cases can be added the nonreduced maximal skeleton consisting of the maximal skeleton with 2 nonterminals S, X enlarged by all 9 productions with Y on the left hand side.

5 Maximal Skeletons with 4 Nonterminals

Finally, the program was run for the case of 4 nonterminals, starting with n = 6 and repeating the procedure for n = 8, 9, 10, 12, 14, 15, 16, 18, 20, 21, 22, 24, 25, 26, and 27. The program produced 413 candidates for maximal skeletons which number is to big to prove for all of them to be maximal skeletons indeed.

References

1. Dömösi, P., Horváth, S., Ito, M., Formal languages and primitive words, *a.) Proc. First Conf. on Scientific. Communication, Univ. Oradea, Romania, 1991; b.) Publ. Math. (Debrecen), 42(1993), 315-321.*

2. Dömösi, P., Horváth, S., Ito, M., Kászonyi, L., Katsura, M., Some combinatorial properties of words, and the Chomsky hierarchy, *Proc. 2nd Int. Coll. Words, Languages and Combinatorics, Kyoto, Japan, 25-28 Aug., 1992, ed.: M. Ito and H. Jürgensen, World Scientific Publishers, Singapore, 105-123, 1994.*

3. Dömösi, P., Horváth, S., Ito, M., Kászonyi, L., Katsura, M., Formal languages consisting of primitive words, *Proc. Conf. FCT'93, ed.: Z. Ésik, Springer LNCS 710, 194-203, 1993.*

4. Horváth, G., A három nemterminálist tartalmazó, Chomsky-féle normál alakú, primitív szavakat generáló nyelvtanok (Determination of grammars having Chomsky normal form with three nonterminals generating primitive words), *Tudományos Diákkori pályamunka, Debrecen, 1994, p. 11.*

5. Horváth, S., Strong interchangeability, nonlinearity and related properties of primitive words, *manuscript, Budapest, August, 1994. Report FBI-HH 183/96, FB Informatik, Universität Hamburg, 1996.*

6. Ito, M., Katsura, M., Shyr, H. J., Yu, S. S., Automata accepting primitive words, *Semigroup Forum, 37 (1988), 45-52.*

7. Petersen, H., The ambiguity of primitive words, *Proc. STACS'94, Springer LNCS 775, 679-690, 1994.*