ON n-PATH-CONTROLLED GRAMMARS

Jiří Koutný
Doctoral Degree Programme (2), FIT BUT
E-mail: ikoutny@fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

ABSTRACT

This paper discusses context-free grammars with some root-to-leaf paths in derivation trees restricted by control languages. It demonstrates that if these control languages are linear, then there are several families of generated languages depending on the common part of all restricted paths. The paper deals with the investigation of several properties of these families.

1 INTRODUCTION

Indisputably, the investigation of context-free grammars with restricted derivation trees represents an important trend in today's formal language theory (see [1], [3], [4], [5], [6], [8]). In essence, these grammas generate their languages just like ordinary context-free grammars do; in addition, however, their derivation trees have to satisfy some simple prescribed conditions. The present paper continues with the investigation of these grammars.

The authors in [5] have studied a new type of restriction in derivation: a derivation tree in a context-free grammar is accepted only if it contains a path described by a string generated by another context-free grammar. They take two context-free grammars, G and G', where G' generates a language over the total alphabet of G. A string w generated by G is accepted only if there is a derivation tree t of w with respect to G such that there exists a path in t which is marked by a string from $L(G')$. Based on this restriction, they have introduced *path-controlled grammars*, *PC* grammars for short, and they have found many properties of this formal model. As they have noticed in the final remarks, however, still many modifications remains unsolved.

The regular paths in a derivation tree of context-free grammars do not increase the generative power (see [4] and see [5], Prop. 2). In this paper, therefore, a control language is supposed to be linear (see page 597 in [7]). Here, we deal with a generalization of path-controlled grammars where the string w generated by G is accepted only if there is a derivation tree t of w with respect to G such that there exists $n \geq 0$ paths in t that are marked by the strings from linear language $L(G')$. Based on path-controlled grammars introduced in [5] and pumping lemma for linear languages (see page 120 in [5]), we establish several types of *n-path-controlled* grammars. We show that if the restricted paths satisfy the certain property, then there exist an infinite hierarchy of n-path-controlled grammars.

In conclusion, we formulate some open problems and suggest some new trends in the investigation of n-path-controlled grammars.

2 DEFINITIONS

This paper assumes that the reader is familiar with the theory of formal languages (see [7]) and the theory of regulated rewriting (see [2]). In this section, we introduce some terminology used in the sequel.

For an alphabet V, V^* denotes the letter monoid (generated by V under the operation of concatenation), ε is the unit of V^*, $V^+ = V^* - \{\varepsilon\}$. For a word $x \in V^*$, $|x|$ denotes its length.

A context-free grammar is a quadruple $G = (V, T, P, S)$, where, as usual, V is a (finite) alphabet, $T \subseteq V$ is a terminal alphabet, P is a finite set of production rules of the form $A \to x$, where $A \in V - T$, $x \in V^*$, and $S \in V - T$ is the starting symbol. Let $N = V - T$ denote a set of nonterminals. In the standard manner, we introduce the relations $\Rightarrow_G$, $\Rightarrow_G^i$, $\Rightarrow_G^+$, $\Rightarrow_G^*$ (see [7]).

Let ${}_G\triangle(x)$ denote a set of the derivation trees with frontier x with respect to the grammar G. Let $t \in {}_G\triangle(x)$ be a derivation tree. Let $path(s)$ denote the word obtained by concatenating all symbols of the path s (in order from the top, i.e. from the root of t to a leaf of t).

Borrowing the notation from [5], we generalize path-controlled grammar and introduce a n-path-controlled grammar, ${}_nPC$ grammar for short. An ${}_nPC$-grammar is a pair (G, G'), where $G = (V, T, P, S)$ is a context-free grammar and $G' = (V', V, P', S')$ is a linear grammar (see page 597 in [7]). The language generated by (G, G') is $L(G, G') = \{w \in L(G) |$ there is a set C of n different paths in $t \in {}_G\triangle(w)$ such that for all $p \in C$ it holds $path(p) \in L(G')$ and all $p \in C$ are divided in the common node of $t\}$.

Clearly, each two paths of a derivation tree contain at least one common node. Thus, for a ${}_nPC$ grammar (G, G'), there is some $m_C \in \mathbb{N}$ that denotes a number of common nodes for all $p \in C$. Hence, for each two $p_1, p_2 \in C$ it holds that $path(p_1) = rDs_1$, $path(p_2) = rDs_2$, where $r \in N^*$, $D \in N$, $s_1, s_2 \in N^*T$ and $|rD| = m_C$.

Consider the pumping lemma for linear languages (see page 120 in [5]). Hence, each $path(p)$, where $p \in C$, such that $|path(p)| \geq k$, for some $k \geq 0$, can be written in the form $path(p) = uvwxy$ and for each $m \geq 0$ it holds that $uv^m wx^m y \in L(G')$.

We distinguish five types of ${}_nPC$ grammars depending on the value of m_C in relation to pumping lemma for $L(G')$: The types are ${}_n^I PC$ if C satisfies $0 \leq m_C \leq |u|$, ${}_n^{II} PC$ if C satisfies $|u| < m_C \leq |uv|$, ${}_n^{III} PC$ if C satisfies $|uv| < m_C \leq |uvw|$, ${}_n^{IV} PC$ if C satisfies $|uvw| < m_C \leq |uvwx|$, and ${}_n^{V} PC$ if C satisfies $|uvwx| < m_C \leq |uvwxy|$, where $uvwxy$ is the shortest path from C.

We denote the family of the languages generated by LIN, CF, PC, ${}_nPC$, ${}_n^I PC$, ${}_n^{II} PC$, ${}_n^{III} PC$, ${}_n^{IV} PC$, ${}_n^{V} PC$ grammars by **LIN**, **CF**, **PC**, ***n*-PC**, **I-*n*-PC**, **II-*n*-PC**, **III-*n*-PC**, **IV-*n*-PC**, **V-*n*-PC**, respectivelly.

3 RESULTS

Theorem 1. **PC = 1-PC = I-1-PC = II-1-PC = III-1-PC = IV-1-PC = V-1-PC.**

Proof. The equality clearly follows from the definitions of PC, ${}_nPC$, ${}_n^i PC$, for $i = I, II, III, IV, V$, grammars. □

Theorem 2. If $L \in$ **III-n-PC**, for $n = card(C) \geq 0$, then there are $p, q \in \mathbb{N}$ such that each $z \in L$ with $|z| > p$ can be written in the form $z = u_1 v_1 u_2 v_2 \ldots u_{2n+2} v_{2n+2} u_{2n+3}$, such that $0 < |v_1 v_2 \ldots v_{2n+2}| \leq q$ and $u_1 v_1^i u_2 v_2^i \ldots u_{2n+2} v_{2n+2}^i u_{2n+3} \in L$ for all $i \geq 1$.

Proof. Let (G,G') be a ${}^{III}_{n}PC$-grammar, where $G=(V,T,P,S)$ and $G'=(V',V,P',S')$. We deal directly only with grammar G—in the case of G', we deal only with the language $L(G')$.

Consider $t \in_{(G,G')} \triangle(z)$. For each $path(s) = SA_1 \dots A_k a$ of t, where $s \in C$, $S, A_1, \dots, A_k \in N$, $a \in T$, consider the rules $A_i \to x_i A_{i+1} y_i$ used when passing from A_i to A_{i+1} on this path and the rule $A_k \to x_k a y_k$ used in the last step of the derivation in G corresponding to the path s.

Consider that any $x_i y_i$, $i=1,\dots,k$, contains a nonterminal B that do not belong on any path $s \in C$. Clearly, there is substring z' of z derived from B. Since G is context-free, it follows that if $|z'| \geq k_1$, for some $k_1 \geq 0$, then there are two substrings z'_1, z'_2 of z' that can be pumped and by pumping lemma for context-free languages, z'_1, z'_2 are bounded in length.

If $L(G)$ is infinite, the string $path(s) \in L(G')$ is arbitrarily long. Thus, if $path(s) = uvxyz$ with $|uvxyz| \geq k_2$, for some $k_2 \geq 0$, then $uvxyz$ satisfies $uv^i xy^i z \in L(G')$, for $i \geq 1$. Hence, the derivations starting from the symbols of v and y can be repeated in G. Since (G,G') is ${}^{III}_{n}PC$ grammar, it follows that the derivations starting from the symbols of v in G are common for all $s \in C$ and the derivations starting from the symbols of y in G are unique for each $s \in C$.

Consider the derivations starting from v in G. This leads to pumping of two substrings v_1, v_{2n+2} of z—one in the left-hand side, one in the right-hand side of common part of all $s \in C$.

Consider the derivations starting from y in G. This leads to pumping of two substrings of z—one in the left-hand side, one in the right-hand side of each $s \in C$. For each $s_{i+1} \in C$, denote this two substrings v_{2i+2}, v_{2i+3}, $i=0,1,\dots,n-1$. Since (G,G') is ${}^{III}_{n}PC$ grammar, we obtain $2n$ pumped substrings of z.

By pumping lemma for context-free languages, the substrings $v_1, v_2, \dots, v_{2n+2}$ are bounded in length. So, the total length of the $2n+2$ pumped substrings of z is bounded by a constant q. □

Corollary 3. III-n-PC cannot count to $2n+3$, but can count to $2n+2$, e.g. $L=\{a^i b^i c^i d^i e^i f^i g^i \mid i \geq 1\} \notin$ **III-2-PC**, but $L \in$ **III-3-PC**.

Corollary 4. There is an infinite hierarchy of $\bigcup_{i=0}^{n}$ **III-i-PC** languages; i.e. $\bigcup_{i=0}^{n}$ **III-i-PC** $\subset$ $\bigcup_{i=0}^{n+1}$ **III-i-PC**, for $n \geq 0$, is proper.

Corollary 5. III-n-PC is not closed under concatenation, e.g. $L=\{a^i a^i a^i a^i a^i a^i \mid i \geq 1\} \in$ **III-2-PC**, but $LL \notin$ **III-2-PC**.

Example 1. Let us have ${}^{III}_{n}PC$ grammar (G,G'), $n \geq 0$, where

$$
\begin{aligned}
&G_1 = (\{S\} \cup \{A_i, B_i \mid i=1,\dots,n\} \cup \{a_i \mid i=1,\dots,2n+2\}, \{a_i \mid i=1,\dots,2n+2\}, P, S),\\
&P = \{S \to a_1 S a_{2n+2}, S \to a_1 A_1 \dots A_n a_{2n+n}\} \cup\\
&\quad \{A_{i+1} \to a_{2i+2} A_{i+1} a_{2i+3}, A_{i+1} \to B_{i+1}, B_{i+1} \to a_{2i+2} a_{2i+3} \mid i=0,\dots,n-1\},\\
&L(G') = \bigcup_{i=1}^{n} \{S^k A_i^k B_i a_{2i} \mid k \geq 1\} \text{—clearly, } L(G') \in \textbf{LIN}.
\end{aligned}
$$

Consider a derivation in (G,G'):

$$
\begin{aligned}
&S \Rightarrow^k a_1^k S a_{2n+2}^k\\
&\quad \Rightarrow a_1^k a_1 A_1 \dots A_n a_{2n+2} a_{2n+2}^k\\
&\quad \Rightarrow^{n \times k} a^{k+1} a_2^k B_1 a_3^k \dots a_{2n}^k B_n a_{2n+1}^k a_{2n+2}^{k+1}\\
&\quad \Rightarrow^n a^{k+1} a_2^{k+1} a_3^{k+1} \dots a_{2n}^{k+1} a_{2n+1}^{k+1} a_{2n+2}^{k+1}
\end{aligned}
$$

Clearly, n different paths are described by $L(G')$. This way, by ${}^{III}_{n}PC$ grammar (G,G'), $n \geq 0$, we can generate the language $L(G,G') = \{a_1^k, \ldots, a_{2n+2}^k | \ k \geq 1\}$.

To be more concrete, consider ${}^{III}_{2}PC$ grammar (G,G'), where

$G = (\{S,X,Y,U,V,a,b,c,d,e,f\},\{a,b,c,d,e,f\},P,S)$,
$P = \{S \to aSf, S \to aXYf, X \to bXc, X \to U, U \to bc, Y \to dYe, Y \to V, V \to de\}$,
$L(G') = \{S^n X^n U b \cup S^n Y^n V d | \ n \geq 1\}$,
$L(G,G') = \{a^n b^n c^n d^n e^n f^n | \ n \geq 1\}$.

Example 2. Let $m \geq 0$ with $m \bmod 2 = 0$. Let us have ${}^{III}_{n}PC$ grammar (G,G'), $n \geq 0$, where

$G = (\{A_j, B_j, a_j | \ j = 1, \ldots, m\} \cup \{C\}, \{a_j | \ j = 1, \ldots, m\}, P, A_1)$,
$P = \{A_1 \to a_1 A_1, A_1 \to a_1 A_2, B_1 \to B_1 a_1, B_1 \to C, C \to a_1, A_m \to A_m a_m, A_m \to \{B_m\}^n\} \cup$
$\{A_i \to A_i a_i, A_i \to A_{i+1} | \ i = 2, \ldots, m-1 \text{ with } i \bmod 2 = 0\} \cup$
$\{A_i \to a_i A_i, A_i \to A_{i+1} | \ i = 3, \ldots, m-1 \text{ with } i \bmod 2 = 1\} \cup$
$\{B_i \to a_i B_i, B_i \to B_{i-1} | \ i = 2, \ldots, m \text{ with } i \bmod 2 = 0\} \cup$
$\{B_i \to B_i a_i, B_i \to B_{i-1} | \ i = 3, \ldots, m \text{ with } i \bmod 2 = 1\}$,
$L(G') = \{A_1^{k_1} A_2^{k_2} \ldots A_m^{k_m} B_m^{k_m} B_{m-1}^{k_{m-1}} \ldots B_2^{k_2} B_1^{k_1} C a_1 | \ k_i \geq 0, i = 1, \ldots, m\}$—clearly, $L(G') \in \mathbf{LIN}$.

Consider a derivation in (G,G'), for some $m \geq 0$ with $m \bmod 2 = 0$:

$$
\begin{aligned}
A_1 &\Rightarrow^{k_1} a_1^{k_1} A_1 \\
&\Rightarrow a_1^{k_1+1} A_2 \\
&\Rightarrow^{k_2} a_1^{k_1+1} A_2 a_2^{k_2} \\
&\Rightarrow a_1^{k_1+1} A_3 a_2^{k_2} \\
&\Rightarrow^{*} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} A_m a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{B_m\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{n \times k_m} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} B_m\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{n} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} B_{m-1}\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{n \times k_{m-1}} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} B_{m-1} a_{m-1}^{k_{m-1}}\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{*} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} a_{m-2}^{k_{m-2}} \ldots a_2^{k_2} B_1 a_1^{k_1} \ldots a_{m-3}^{k_{m-3}} a_{m-1}^{k_{m-1}}\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{n} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} a_{m-2}^{k_{m-2}} \ldots a_2^{k_2} C a_1^{k_1} \ldots a_{m-3}^{k_{m-3}} a_{m-1}^{k_{m-1}}\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2} \\
&\Rightarrow^{n} a_1^{k_1+1} a_3^{k_3} a_5^{k_5} \ldots a_{m-1}^{k_{m-1}} \{a_m^{k_m} a_{m-2}^{k_{m-2}} \ldots a_2^{k_2} a_1^{k+1} \ldots a_{m-3}^{k_{m-3}} a_{m-1}^{k_{m-1}}\}^n a_m^{k_m} \ldots a_6^{k_6} a_4^{k_4} a_2^{k_2}
\end{aligned}
$$

Thus, n different paths are described by $L(G')$. This way, by ${}^{III}_{n}PC$ grammar (G,G'), $n \geq 0$, $m \geq 0$ with $m \bmod 2 = 0$, we can generate the language

$$L(G,G') = \{(a_1^{k_1+1} a_3^{k_3} \ldots a_{m-1}^{k_{m-1}} a_m^m a_{m-2}^{k_{m-2}} a_{m-4}^{k_{m-4}} \ldots a_2^{k_2})^{n+1} | \ k_i \geq 0, i = 1, \ldots, m\}.$$

To be more concrete, consider $m = 4$ and ${}^{III}_{3}PC$ grammar (G,G'), where

$G = (\{A,B,C,D,E,F,G,H,I,a,b,c,d\},\{a,b,c,d\},P,A)$,
$P = \{A \to aA, A \to aB, B \to Bb, B \to C, C \to cC, C \to D, D \to Dd, D \to HHH, E \to Ea,$
$E \to I, F \to bF, F \to E, G \to Gc, G \to F, H \to dH, H \to G, I \to a\}$,
$L(G') = \{A^r B^s C^t D^u H^u G^t F^s E^r I a | \ r,s,t,u \geq 0\}$,
$L(G,G') = \{a^r c^t d^u b^s a^r c^t d^u b^s a^r c^t d^u b^s a^r c^t d^u b^s | \ r > 0, s,t,u \geq 0\}$.

4 CONCLUSION

We have considered a new type of restriction in derivation: ${}_nPC$ grammars as a generalization of *PC* grammars introduced in [5]. In relation to pumping lemma for linear languages, we have demonstrated that there are several types of such ${}_nPC$ grammars. We have found several properties of such model, especially pumping property for ${}^{III}_{n}PC$ grammars and some consequences that follow this property.

It seems that the most useful is the family of **III-n-PC** languages and for this family, there are still many questions to by answered—for instance, generative power, closure properties, decidability properties, parsing properties, and descriptional complexity. For **I-n-PC** and **V-n-PC**, however, also many properties can be found. For **II-n-PC** and **IV-n-PC**, there are currently no answers, only questions. As the requirements for a set C of all controlled paths looks fairly restrictive, weaker control could be of interest, i.e. the request for C that all controlled paths have to be divided in just one node can be weakened. The formal study of such variants remains to be carried out; we hope to return to this topic in a forthcoming paper.

ACKNOWLEDGEMENT

This work was partially supported by the FRVŠ MŠMT grant FR2581/2010/G1, the BUT FIT grant FIT-10-S-2, and the research plan MSM0021630528.

REFERENCES

[1] K. Čulik and H. A. Maurer. Tree controlled grammars. *Computing*, 19:129–139, 1977.

[2] J. Dassow and Gh. Păun. *Regulated Rewriting in Formal Language Theory*. Springer, Berlin, 1989.

[3] J. Dassow and B. Truthe. Subregularly tree controlled grammars and languages. In *Automata and Fromal Languages - 12th International Conference AFL 2008, Balatonfured*, pages 158–169. Computer and Automation Research Institute of the Hungarian Academy of Sciences, 2008.

[4] J. Koutný. Regular paths in derivation trees of context-free grammars. In *Proceedings of the 15th Conference and Competition STUDENT EEICT 2009 Volume 4*, pages 410–414. Faculty of Information Technology BUT, 2009.

[5] S. Marcus, C. Martín-Vide, V. Mitrana, and Gh. Păun. A new-old class of linguistically motivated regulated grammars. In *CLIN*, pages 111–125, 2000.

[6] C. Martín-Vide and V. Mitrana. Further properties of path-controlled grammars. In *Formal Grammar / Mathematics of Language 2005*, pages 219–230. Edimburgh, 2005.

[7] A. Meduna. *Automata and Languages: Theory and Applications*. Springer Verlag, 2005.

[8] Gh. Păun. On the generative capacity of tree controlled grammars. *Computing*, 21(3):213–220, 1979.