C. Câmpeanu, G. Pighizzini (eds.): Proceedings of the 10th workshop
Descriptional Complexity of Formal Systems. Charlottetown PE, Canada, July 16–18, 2008.
(Pages 145–156)

On Two Hierarchies of Subregularly Tree Controlled Languages

Jürgen Dassow Bianca Truthe

Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik
PSF 4120, D-39016 Magdeburg, Germany
dassow@iws.cs.uni-magdeburg.de truthe@iws.cs.uni-magdeburg.de

Abstract. Tree controlled grammars are context-free grammars where the associated language only contains those terminal words which have a derivation where the word of any level of the corresponding derivation tree belongs to a given regular language.

In this paper, we consider first the case that we take only such regular languages as the control set which can be represented by finite unions of monoids. We show that the corresponding hierarchy of tree controlled languages collapses already at the level 2. Moreover, we present a characterization of both levels by well-known language families generated by extended Lindenmayer systems.

Furthermore, we give some comments on the hierarchy of tree controlled languages if one restricts the number of states allowed in the accepting automaton of the regular control language.

Keywords: Tree controlled grammars, generative capacity, hierarchies

1 Introduction

It is a well-known fact that the most investigated classes of formal languages, the regular and context-free languages, are not able to cover all phenomena which are known from natural languages, programming languages etc. Thus, there have been introduced many grammars with a context-free core and some mechanism which controls the sequences of rules in a derivation or the applicability of a rule etc. (see [2] and [6]). One such control mechanism was introduced by Culik II and Maurer in [1] where the structure of the derivation trees is restricted by the requirement that all words belonging to a level of the derivation tree have to be in a given regular language. Păun proved that the generative power of these grammars, called tree controlled grammars, coincides with that of context-sensitive grammars (if erasing rules are forbidden) or arbitrary phrase structure grammars (if erasing rules are allowed). Therefore most of the classical decision problems are undecidable or NP-hard. But if one restricts the underlying context-free grammars to be unambiguous, then the membership problem can be solved in quadratic time and a lot of important non-context-free languages can be generated. Thus it is a natural question to consider restricted versions of tree controlled grammars. In [3] the generative power has been studied in those cases that one restricts the control language to special subclasses of the family of regular languages,

e. g. to monoids, nilpotent, combinational, definite, non-counting, regular suffix-closed and regular commutative languages.

In this paper, we consider first the case that we take only such regular languages as the control set which can be obtained from singletons, consisting of the empty word or a letter of the underlying alphabet, and the empty set by the use of union and Kleene closure. It follows easily that such languages can be represented by finite unions of monoids. Therefore we have an obvious hierarchy with respect to the number of unions. We first show that the corresponding hierarchy of tree controlled languages collapses already at the level 2. Moreover, we present a characterization of both levels by well-known language families generated by extended Lindenmayer systems.

Furthermore, we give some comments on the hierarchy of tree controlled languages if one restricts the number of states allowed in the accepting automaton of the regular control language.

2 Definitions

Throughout the paper, we assume that the reader is familiar with the basic concepts of formal language theory; for details we refer to [6], [5], and [2].

With any derivation in a context-free grammar G, we associate a derivation tree. With any derivation tree t of height k and any number $0 \leq j \leq k$, we associate the words of level j and the sentential form of level j which are given by all nodes of depth j read from left to right and all nodes of depth j and all leaves of depth less than j read from left to right, respectively.

Obviously, if w and v are sentential forms of two successive levels, then $w \Longrightarrow^* v$ holds and this derivation is obtained by a parallel replacement of all nonterminals occurring in w.

A *tree controlled grammar* is a quintuple $G = (N, T, P, S, R)$ where

- (N, T, P, S) is a context-free grammar with a set N of nonterminals, a set T of terminals, a set P of context-free non-erasing rules, and an axiom S,
- R is a regular set over $(N \cup T)^*$.

The language $L(G)$ generated by a tree controlled grammar $G = (N, T, P, S, R)$ consists of all words $z \in T^*$ such that there is a derivation tree t where z is the word obtained by reading the leaves from left to right and the words of all levels of t – besides the last one – belong to R.

Example 1. *We now consider the tree controlled grammar*

$$G = (\{S, A, B, C\}, \{a, b\}, P, S, R)$$

with $P = \{S \to AB, A \to aAb, B \to Ba, A \to ab, B \to a, A \to aCb, C \to Cb, C \to b\}$ *and* $R = (\{a, b, S, C\}^*\{A, B\}\{a, b, S, A, C\}^*\{B\}\{a, b, S, C\}^*)^* \cup \{a, b, S, C\}^*$. *Due to the given productions and the control set, the words of a level of a derivation tree can only be from the set* $\{S, AB, aAbBa, aba, aCba, Cb, b\}$. *Therefore any derivation has* *the form* $S \Longrightarrow AB \Longrightarrow aAbBa \Longrightarrow \cdots \Longrightarrow a^{n-1}Ab^{n-1}Ba^{n-1} \Longrightarrow a^nb^na^n$ *or*

$$\begin{aligned} S &\Longrightarrow AB \Longrightarrow aAbBa \Longrightarrow \cdots \Longrightarrow a^{n-1}Ab^{n-1}Ba^{n-1} \Longrightarrow a^nCb^na^n \\ &\Longrightarrow a^nCb^{n+1}a^n \Longrightarrow \cdots \Longrightarrow a^nCb^{n+m-1}a^n \Longrightarrow a^nb^{n+m}a^n \end{aligned}$$

with $n \geq 1$ *and* $m \geq 1$*. Thus,* $L(G) = \{a^nb^{n+m}a^n \mid n \geq 1,\ m \geq 0\}$*.*

Given a set $\mathcal{R}$ of regular languages, we denote by $\mathcal{TC}(\mathcal{R})$ the set of all languages generated by tree controlled grammars $G = (N, T, P, S, R)$ with $R \in \mathcal{R}$.

Lemma 2. [3] *If* $X \subseteq Y$ *holds for two sets* X *and* Y *of regular languages, then also the inclusion* $\mathcal{TC}(X) \subseteq \mathcal{TC}(Y)$ *holds.* □

An *extended tabled interactionless L system* (ET0L system for short) is a quadruple $G = (V, T, \mathcal{P}, \omega)$ where V is an alphabet, T is a subset of V, $\omega \in V^*$ and, $\mathcal{P} = \{P_1, P_2, \ldots, P_r\}$ for some $r \geq 1$ where, for $1 \leq i \leq r$, P_i is a finite subset of $V \times V^*$ such that, for any $a \in V$, there is at least one element (a, v) in P_i. The elements P_i, $1 \leq i \leq r$, are called tables.

As usual, we shall write $a \to v$ instead of (a, v). A word $x \in V^+$ directly derives a word $y \in V^*$ (written as $x \Longrightarrow y$), if

- $x = x_1x_2 \ldots x_n$ for some $n \geq 1$, $x_i \in V$, $1 \leq i \leq n$,
- $y = y_1y_2 \ldots y_n$ and
- there is a natural number j, $1 \leq j \leq r$ such that $x_i \to y_i \in P_j$ for $1 \leq i \leq n$.

The language $L(G)$ generated by an ET0L system G is defined as

$$L(G) = \{z \mid z \in T^*,\ \omega \Longrightarrow^* z\},$$

where $\Longrightarrow^*$ is the reflexive and transitive closure of $\Longrightarrow$.

By *ET0L* and *ET0L*$_r$ we denote the families of all languages generated by ET0L systems and ET0L systems with at most r tables, respectively. An ET0L system with only one table is also called an E0L system; we write *E0L* for the class *ET0L*$_1$.

We recall the following theorem on the hierarchy with respect to the number of tables; a proof of it can be found in [5].

Theorem 3. *For any ET0L system* G*, there is an ET0L system* G' *such that* G' *has at most two tables and* $L(G') = L(G)$*, i. e., ET0L*$_r$ *= ET0L*$_2$ *for any* $r \geq 2$*.* □

3 A Collapsing Hierarchy

Let X be an infinite set. We consider only languages $L \subset (X')^*$ where X' is a finite subset of X. Let us consider regular sets which are obtained by application of union and Kleene closure from the basic sets $\{x\}$ with $x \in X$, $\{\lambda\}$ and $\emptyset$. By $(A^*)^* = A^*$ and $(A^* \cup B^*)^* = (A \cup B)^*$, it follows easily by induction on the number of applied operations that any infinite such restricted regular set is of the form $A_1^* \cup A_2^* \cup \cdots \cup A_n^*$ for some $n \geq 1$ and some (finite) alphabets $A_i \subset X$, $1 \leq i \leq n$, i. e., it is a union of n monoids for some $n \geq 1$.

For any natural number $n \geq 1$, let MON_n be the set of all languages that can be represented in the form $A_1^* \cup A_2^* \cup \cdots \cup A_k^*$ with $1 \leq k \leq n$ where all A_i $(1 \leq i \leq k)$ are alphabets. Obviously, $MON_1 \subset MON_2 \subset \cdots \subset MON_j \subset \cdots$. By Lemma 2 and [3], we obtain the following results.

Proposition 4. $\mathcal{TC}(MON_1) \subseteq \mathcal{TC}(MON_2) \subseteq \cdots \subseteq \mathcal{TC}(MON_j) \subseteq \cdots$. □

Proposition 5. $\mathcal{TC}(MON_1) = \mathit{E0L}$. □

We now show that every language in $\mathcal{TC}(MON_k)$ $(k \geq 1)$ can be generated by an ET0L system with $k+1$ tables. In [1], it was shown that every tree controlled grammar $G = (N, T, P, S, R)$ with $R = N_1^* \cup N_2^* \cup \cdots \cup N_k^*$ where $N = N_1 \cup N_2 \cup \cdots \cup N_k$ and $N_i \cap N_j = \emptyset$ for $i \neq j$ generates an ET0L language. Our result is in that sense sharper that we allow an arbitrary control set $R \in MON_k$.

Theorem 6. *For all $k \geq 1$, the inclusion $\mathcal{TC}(MON_k) \subseteq \mathit{ET0L}_{k+1}$ holds.*

Proof: Let L be a language generated by a tree controlled grammar

$$G_{\mathrm{t}} = (N, T, P, S, R)$$

where R is the union of at most k monoids: $R = R_1^* \cup R_2^* \cup \cdots \cup R_k^*$.

If $S \notin R$ then $L = \emptyset$ and we take the ET0L system $G = (\{S\}, \emptyset, \{h_1\}, S)$ with $h_1(S) = S$. Since $L(G) = \emptyset$, we have $L = L(G)$. Let us now consider $S \in R$.

We construct an ET0L system $G = (V \cup \{F\}, T, \{h_1, h_2, \ldots, h_{k+1}\}, S)$ as follows:

$$\begin{aligned}
V &= N \cup T, \\
h_i'(A) &= \{ w \mid A \to w \in P \text{ and } w \in R_i^* \} \text{ for } i = 1, 2, \ldots, k \text{ and } A \in N, \\
h_{k+1}'(A) &= \{ w \mid A \to w \in P \text{ and } w \in T^* \} \text{ for } A \in N, \\
h_i(A) &= \begin{cases} h_i'(A) & \text{if } h_i'(A) \neq \emptyset, \\ \{F\} & \text{otherwise,} \end{cases} \quad \text{for } i = 1, 2, \ldots, k+1 \text{ and } A \in N, \\
h_i(a) &= \{a\} \text{ for } i = 1, 2, \ldots, k+1 \text{ and } a \in T \cup \{F\}.
\end{aligned}$$

The trap symbol F is introduced to meet the definition that $h_i(A) \neq \emptyset$ for $i = 1, 2, \ldots, k+1$ and $A \in N$. The sentential forms containing F do not contribute to the language $L(G)$. By induction on the derivation length, it can be proved that G generates the given language L. Hence, $\mathcal{TC}(MON_k) \subseteq \mathit{ET0L}_{k+1}$. □

According to Theorem 3, we even have the next result.

Corollary 7. *For all $k \geq 1$, the inclusion $\mathcal{TC}(MON_k) \subseteq \mathit{ET0L}_2$ holds.* □

We now show that the inversion holds for $k \geq 2$.

Theorem 8. *Every ET0L language can be generated by a tree controlled grammar with a control set composed of two monoids: $\mathit{ET0L} \subseteq \mathcal{TC}(MON_2)$.*

Proof: Let L be an ET0L language. Then, by Theorem 3, there is an ET0L system $G = (V, T, \{ h_1, h_2 \}, \omega)$ with two tables that generates the language L.

With every symbol $x \in V$, we associate two new symbols x_1 and x_2. We set

$$V_1 = \{ x_1 \mid x \in V \} \text{ and } V_2 = \{ x_2 \mid x \in V \}.$$

The corresponding isomorphisms are denoted by η_1 and η_2, respectively:

$$\eta_1 : V^* \to V_1^* \text{ with } \eta_1(x) = x_1 \quad \text{and} \quad \eta_2 : V^* \to V_2^* \text{ with } \eta_2(x) = x_2.$$

Additionally, let $S \notin V \cup V_1 \cup V_2$ be a new symbol.

We now construct a tree controlled grammar $G_{\mathrm{t}} = (N, T, P, S, R)$ as follows:

$$\begin{aligned} N &= \{ S \} \cup V_1 \cup V_2, \\ P &= \{ S \to \eta_j(\omega) \mid j \in \{ 1, 2 \} \} \\ &\qquad \cup \{ \eta_i(x) \to \eta_j(w) \mid i, j \in \{ 1, 2 \}, x \in V \text{ and } w \in h_j(x) \} \\ &\qquad \cup \{ \eta_i(x) \to x \mid i \in \{ 1, 2 \} \text{ and } x \in T \}, \\ R &= (V_1 \cup \{ S \})^* \cup V_2^*. \end{aligned}$$

We prove that the tree controlled grammar G_{t} generates the given ET0L language L.

(I) $L \subseteq L(G_{\mathrm{t}})$.

We show for every sentential form w of G by induction on the derivation length that $\eta_1(w)$ and $\eta_2(w)$ are words of a level of a derivation tree of G_{t}.

For the axiom ω of G, the words $\eta_1(\omega)$ and $\eta_2(\omega)$ are obtained by G_{t} using the rule $S \to \eta_1(\omega)$ or $S \to \eta_2(\omega)$. The words S, $\eta_1(\omega)$ and $\eta_2(\omega)$ belong to R, so $\eta_1(\omega)$ and $\eta_2(\omega)$ appear as words of a level of a derivation tree of G_{t}.

Let w be a sentential form of G. By induction hypothesis, $\eta_1(w)$ and $\eta_2(w)$ occur as a level in a derivation tree of G_{t}. Let $u \in h_i(w)$ be a derivative of w by a table h_i ($i \in \{ 1, 2 \}$). Due to the construction of the rule set P, we can derive $\eta_i(u)$ in G_{t}. Since $\eta_i(u) \in V_i^*$, the word $\eta_i(u)$ belongs to R and it is a level of a derivation tree of G_{t}. If $w \in T^*$, then we can derive w itself from $\eta_1(w)$ or $\eta_2(w)$ in G_{t} by applying the terminating rules and obtain $w \in L(G_{\mathrm{t}})$, which proves $L \subseteq L(G_{\mathrm{t}})$.

(II) $L(G_{\mathrm{t}}) \subseteq L$.

Every derivation tree of G_{t} has as root the nonterminal symbol S. We show for every word w that occurs at a level (apart from the first and last ones) of a derivation tree of G_{t} by induction on the derivation length that there are a number $i \in \{ 1, 2 \}$ and a word w' such that $w = \eta_i(w')$ and w' is a sentential form of G.

Let w be a word of the second level of a derivation tree. Then w is a derivative of S in G_{t}. Since there are only two rules for S, there is a natural number $i \in \{ 1, 2 \}$ such that $w = \eta_i(\omega)$. Furthermore, ω is a sentential form (the axiom) of G.

Let w be a word of a further level of a derivation tree of G_{t}. By induction hypothesis, there are a number $i \in \{ 1, 2 \}$ and a word w' such that $w = \eta_i(w')$

and w' is a sentential form of G. Let u be the next level after w. Then u is a derivation of w and $u \in R \cup T^*$. If $u \in R$, then there is a number $j \in \{1, 2\}$ such that $u \in V_j^*$, because S does not occur on the right hand side of a rule. Then, according to the rules of G_{t}, there is a word u' such that $u = \eta_j(u')$ and $u' \in h_j(w')$. Hence, u' is a sentential form of G. If $u \in T^*$, then terminating rules were applied to w. According to the rules in P, we have $u = w'$. In this case, we obtain that u is a sentential form and even a terminal word of $L(G_{\mathrm{t}})$. Thus, every word generated by G_{t} is also generated by G.

Together, we obtain $L = L(G_{\mathrm{t}})$, which completes the proof that every ET0L language is also generated by a tree controlled grammar where the control set is described by two monoids. Hence, the inclusion $\mathit{ET0L} \subseteq \mathcal{TC}(\mathit{MON}_2)$ holds. □

This theorem improves the result $\mathit{ET0L}_k \subseteq \mathcal{TC}(\mathit{MON}_{k+1})$ of the paper [1].

By Proposition 4, Proposition 5, Corollary 7 and Theorem 8, we obtain the hierarchy shown in Figure 1 where an arrow from a class X to a class Y indicates $X \subseteq Y$.

$$\begin{tikzcd}
 & & & \mathit{ET0L} \arrow[ddll, bend right] & & \\
\mathit{E0L} \arrow[urrr, bend left=10] \arrow[d, leftrightarrow] & & & & & \\
\mathcal{TC}(\mathit{MON}_1) \arrow[r] & \mathcal{TC}(\mathit{MON}_2) \arrow[r] \arrow[uurr, bend right=10] & \mathcal{TC}(\mathit{MON}_3) \arrow[r] \arrow[uur, bend right] & \cdots \arrow[r] & \mathcal{TC}(\mathit{MON}_k) \arrow[r] \arrow[uul, bend right] & \cdots
\end{tikzcd}$$

Figure 1. Hierarchy of the classes $\mathcal{TC}(\mathit{MON}_k)$ and $\mathit{ET0L}$

From Figure 1, one can immediately see that the classes $\mathcal{TC}(\mathit{MON}_k)$ and $\mathit{ET0L}$ are pairwise equivalent for $k \geq 2$.

Theorem 9. *The classes $\mathcal{TC}(\mathit{MON}_k)$ for $k \geq 2$ coincide with the class ET0L.* □

For $k \geq 1$, the inclusions and equivalences hold as shown in Figure 2. An arrow from a class X to a class Y indicates the proper inclusion $X \subset Y$.

$$\begin{tikzcd}[column sep=0pt]
\mathit{CS} & \\
\mathit{ET0L} \arrow[u] & = \mathcal{TC}(\mathit{MON}_2) = \mathcal{TC}(\mathit{MON}_3) = \cdots = \mathcal{TC}(\mathit{MON}_k) = \cdots \\
\mathit{E0L} \arrow[u] & = \mathcal{TC}(\mathit{MON}_1)
\end{tikzcd}$$

Figure 2. Characterization of the classes $\mathcal{TC}(\mathit{MON}_k)$

We have shown that the languages generated by tree controlled grammars with the union of finitely many monoids as the control set can already be generated by a tree controlled grammar with the union of two monoids as the control set. We obtained a two level hierarchy where the first level (using one monoid) is characterized by the family of E0L languages and the second level by the family of ET0L languages.

4 A Hierarchy with Respect to the Size of Automata

Let R be a regular language. Its descriptional complexity $c(R)$ is defined as the number of states of a minimal finite deterministic automaton that accepts R (which is unique up to isomorphism of finite deterministic automata). For any $n \geq 1$, by REG_n we denote the family of regular languages R such that $c(R) \leq n$. It is known that

$$REG_1 \subset REG_2 \subset REG_3 \subset \cdots \subset REG_n \subset \cdots \ .$$

Obviously, by Lemma 2,

Lemma 10. $\mathcal{TC}(REG_1) \subseteq \mathcal{TC}(REG_2) \subseteq \mathcal{TC}(REG_3) \subseteq \cdots \subseteq \mathcal{TC}(REG_n) \subseteq \cdots$. □

If an automaton with input alphabet X has exactly one state z, then the accepted set is the empty set (if the set of accepting states is empty) or X^* (if the set of accepting states is $\{z\}$). Since a tree controlled grammar $G = (N, T, P, S, \emptyset)$ generates the empty set, and the empty set is also generated by any tree controlled grammar $G' = (N, T, P, S, V^*)$ with $V \cap (N \cup T) = \emptyset$, we get the following statement.

Theorem 11. $\mathcal{TC}(REG_1) = \mathcal{TC}(MON_1) = E0L$. □

We consider the language

$$L(G) = \{a^n b^{n+m} a^n \mid n \geq 1,\ m \geq 0\}$$

from Example 1. By [5], Corollary 4.7, $L(G) \notin E0L$. Hence, by Theorem 11, we obtain $L(G) \notin \mathcal{TC}(REG_1)$.

Moreover, it is easy to see from Example 1 that $L(G)$ is in $\mathcal{TC}(REG_2)$ since the language R from Example 1 is accepted by the deterministic finite automaton

$$\mathcal{A} = (\{z_0, z_1\}, \{a, b, S, A, B, C\}, z_0, \delta, \{z_0\})$$

where the transition function δ corresponds to the graph given in Figure 3 [1].

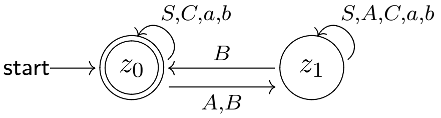

Figure 3. Transition graph of $\mathcal{A}$

Therefore the following result holds.

Theorem 12. $\mathcal{TC}(REG_1) \subset \mathcal{TC}(REG_2)$. □

Let L be a language and $V = alph(L)$ be the minimal alphabet of L. We say that L is *combinational* if and only if it can be represented in the form $L = V^*A$ for some subset $A \subseteq V$. By $COMB$, we denote the family of all combinational languages.

With respect to the hierarchy obtained in [3], we immediately get the following result.

[1] In all figures displaying transition graphs of automata in this paper, the word start points to the start state and all accepting states are marked by a surrounding double circle.

Theorem 13. *$\mathcal{TC}(COMB) \subseteq \mathcal{TC}(REG_2)$.*

Proof: Every combinational language L can be represented as V^*A with $V = alph(L)$ and $A \subseteq V$. Such a language can be accepted by a deterministic finite automaton with two states and the transition function shown in Figure 4.

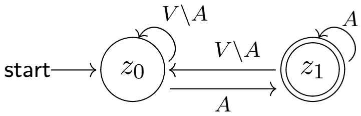

Figure 4. Automaton for accepting the language V^*A

Hence, $COMB \subseteq REG_2$. By Lemma 2, we obtain $\mathcal{TC}(COMB) \subseteq \mathcal{TC}(REG_2)$. □

Theorem 14. *$ET0L \subseteq \mathcal{TC}(REG_4)$.*

Proof: By Theorem 9, for any ET0L language L, there is a tree controlled grammar $G = (N, T, P, S, A_1^* \cup A_2^*)$ where $A_1 \subseteq N \cup T$ and $A_2 \subseteq N \cup T$ such that $L(G) = L$. The finite automaton

$$\mathcal{A}' = (\{z_0, z_1, z_2, z_3\}, N \cup T, z_0, \delta', \{z_0, z_1, z_2\})$$

with the transition function δ' defined according to the transition graph shown in Figure 5 accepts the language $A_1^* \cup A_2^*$.

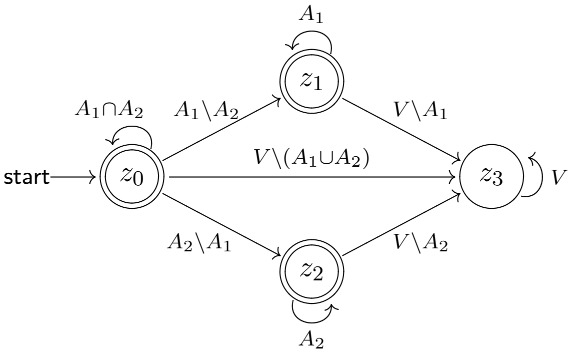

Figure 5. Transition graph of $\mathcal{A}'$ (with $V = N \cup T$)

Therefore $L = L(G) \in \mathcal{TC}(REG_4)$. □

We now show that this inclusion is strict. We give a language that is not generated by an ET0L system but by a tree controlled grammar with a control set which is accepted by a deterministic finite automaton with four states.

Lemma 15. *The language $L = \{\, c^n(ab^m)^n \mid n \geq m \geq 1 \,\}$ is not generated by an ET0L system.*

Proof: Suppose, there is an ET0L system generating L. Then also the language $L' = \{\, (ab^m)^n \mid n \geq m \geq 1 \,\}$ is an ET0L language because the family of ET0L languages is closed under homomorphisms. But L' is not an ET0L language ([5]). So, neither is L. □

Theorem 16. *The language $L = \{\, c^n(ab^m)^n \mid n \geq m \geq 1 \,\}$ is generated by a tree controlled grammar where the control set is accepted by a deterministic finite automaton with four states.*

Proof: Let $\mathcal{A} = (\{\, z_0, z_1, z_2, z_3 \,\}, \{\, S, A, B, B', b, c \,\}, z_0, \delta, \{\, z_0, z_3 \,\})$ be a deterministic finite automaton where the transition function is defined according to the diagram of Figure 6. The language accepted by this automaton is denoted by $T(\mathcal{A})$.

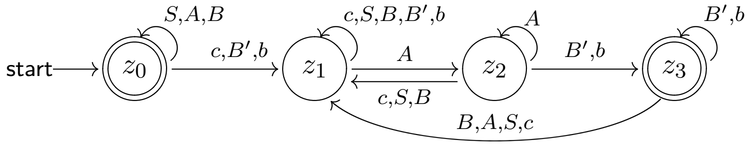

Figure 6. Transition graph of $\mathcal{A}$

Let $T = \{a, b, c\}$ and $G = (\{S, A, B, B'\}, T, P, S, T(\mathcal{A}))$ be a tree controlled grammar with the rule set

$$P = \{S \to ASB, S \to AB, A \to A, A \to c, B \to B, B \to B'b, B \to ab, B' \to B'b, B' \to ab\}.$$

We now prove $L = L(G)$.

(I) $L \subseteq L(G)$.
Let n, m be two natural numbers with $1 \leq m \leq n$. One derivation tree of the word $c^n(ab^m)^n$ is given in Figure 7.

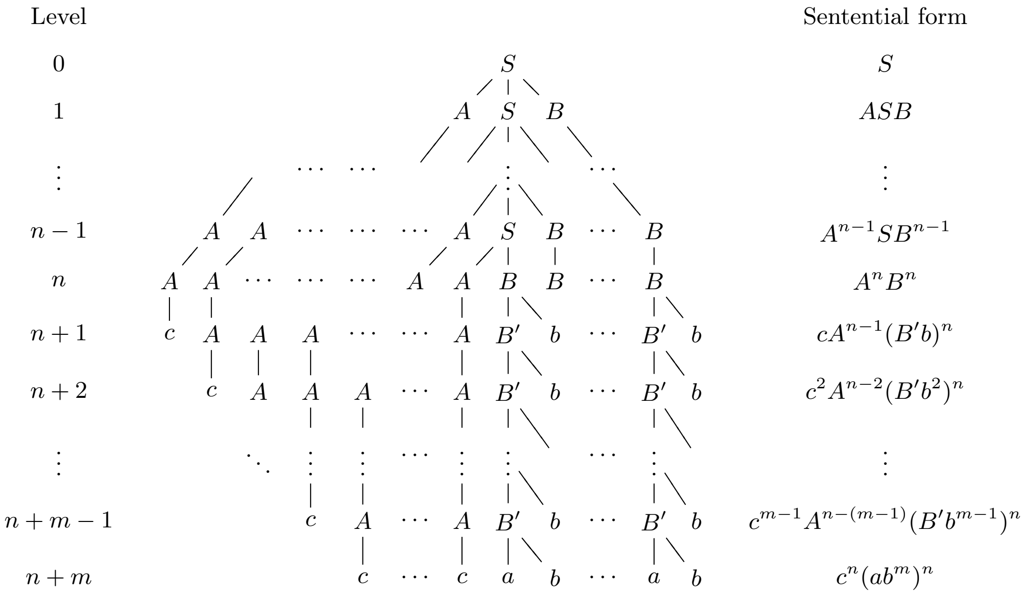

Figure 7. Derivation tree of the word $c^n(ab^m)^n$

The words of the levels 0 to n are accepted in state z_0; the words of the levels $n+1$ to $n+m-1$ are accepted in state z_3. The last level contains a terminal word.

(II) $L(G) \subseteq L$.

We show inductively which words can be derived and belong to the control set R. The start symbol S is accepted by the automaton $\mathcal{A}$. From S, the words AB and ASB can be derived (and only these two). Both words are also accepted by $\mathcal{A}$. The sentential form AB derives the words

- $cab \in T^*$, which belongs to the language L, too,
- Aab, $cB'b$, $AB'b$, cB, which are not accepted by $\mathcal{A}$, and
- $AB \in T(\mathcal{A})$.

The sentential form ASB derives the words

- $cA\sigma Bab$, $AA\sigma Bab$, $cA\sigma BB'b$, $AA\sigma BB'b$, $cA\sigma BB$ with $\sigma \in \{ S, \lambda \}$, but all these words are not accepted by $\mathcal{A}$, and
- $AA\sigma BB \in T(\mathcal{A})$ with $\sigma \in \{ S, \lambda \}$.

The new sentential forms accepted by $\mathcal{A}$ are A^2SB^2 and A^2B^2. The sentential form A^iSB^i with $i \geq 2$ leads to

- the word $A^iS'B^i \in T(\mathcal{A})$ with $S' \in \{ ASB, AB \}$,
- a word $wS'v$ with $w \in \{ c, A \}^*$, $S' \in \{ ASB, AB \}$, $v \in \{ B, B', a, b \}^*$, and $\#_c(w) > 0$ or $\#_{\{a,B'\}}(v) > 0$, but then $wS'v \notin T(\mathcal{A})$.

Hence, the only sentential forms derived from A^iSB^i and accepted by $\mathcal{A}$ are $A^{i+1}SB^{i+1}$ and $A^{i+1}B^{i+1}$. The sentential form A^iB^i with $i \geq 2$ leads to

- the word $c^i(ab)^i \in T^*$, which belongs to the language L, too,
- the word $A^iB^i \in T(\mathcal{A})$,
- a word $wA(B'b)^i \in T(\mathcal{A})$ with $w \in \{ c, A \}^*$ and $\#_c(w) > 0$,
- another word $wv \notin T^*$ with $w \in \{ c, A \}^*$ and $v \in \{ B, B', a, b \}^*$, but then $wv \notin T(\mathcal{A})$.

Hence, the only new sentential form that is accepted by $\mathcal{A}$ is $wA(B'b)^i$ with $w \in \{ c, A \}^*$, $\#_c(w) > 0$, and $i \geq 2$.

We now consider a word $wA(B'b)^i$ with $w \in \{ c, A \}^*$, $\#_c(w) > 0$, and $i \geq 2$ that occurs at some level of a derivation tree. It corresponds to a sentential form $sA(B'b^k)^i$ with the following properties: $k < i$, $s \in \{ c, A \}^*$, $|sA| = i$, there are letters $x_1, x_2, \ldots, x_n$ and words $y_1, y_2, \ldots, y_{n+1}$ such that $w = x_1x_2 \ldots x_n$ and $s = y_1x_1y_2x_2y_3 \ldots y_nx_ny_{n+1}$ (w is a scattered subword of s) and the remaining subword $s - w = y_1y_2 \ldots y_{n+1}$ does not contain the letter A (for the induction base, we have $w = s$ and $k = 1$).

Such a word $wA(B'b)^i$ ($w \in \{ c, A \}^*$, $\#_c(w) > 0$, $i \geq 2$) with a corresponding sentential form $sA(B'b^k)^i$ ($k < i$) derives

- the word $c^j(ab)^i$ with $j = \#_A(w) + 1$ and the corresponding sentential form $c^i(ab^{k+1})^i$, which is a word of the language L,
- a word $w'A(B'b)^i \in T(\mathcal{A})$ with $\#_c(w') > 0$ and the corresponding sentential form is $s'A(B'b^{k+1})^i$ with $i \geq 3$ (in this case, w contains at least one c and at least one A to produce a c in w', hence $|wA| \geq 3$), $k+1 < i$, $s' \in \{ c, A \}^*$, $|s'A| = i$, w' is a scattered subword of s' and the remaining subword $s - w$ belongs to the set $\{ c \}^*$,
- another word $wv \notin T^*$ with $w \in \{c, A\}^*$ and $v \in \{B', a, b\}^*$, but then $wv \notin T(\mathcal{A})$.

Hence, we obtain again a word of the form $wA(B'b)^i$ with a corresponding sentential form $sA(B'b^k)^i$ or a terminal word that belongs to the language L.

Thus, all terminal words generated by G are also words of the language L.

Together, we obtain $L = L(G)$ which completes the proof that the tree controlled grammar $G \in \mathcal{TC}(REG_4)$ generates the non-ET0L language L. □

Together with Theorem 14, we obtain the strict inclusion.

Corollary 17. *$ET0L \subset \mathcal{TC}(REG_4)$.*

Finally, we give the following statement.

Theorem 18. *Every language that is generated by a context-sensitive grammar with exactly r non-context-free rules $p_1, p_2, \ldots, p_r$ and n_i symbols on the left hand side of the rule p_i $(1 \leq i \leq r)$ is also generated by a tree controlled grammar where the control set is accepted by a deterministic finite automaton with at most $2 + \sum_{i=1}^{r}(n_i - 1)$ states.*

Proof: Let L be a language generated by a context-sensitive grammar $G = (N, T, P, S)$ with exactly r non-context-free rules $p_1, p_2, \ldots, p_r$ and n_i symbols on the left hand side of the rule p_i $(1 \leq i \leq r)$. For each terminal symbol, we introduce a non-terminal to postpone the termination. Let $N_T = \{\, T_a \mid a \in T \,\}$. Further, let $V = N \cup T$, $W = N \cup N_T$ and $\eta : V^* \to W^*$ be the isomorphism defined by $\eta(A) = A$ and $\eta(a) = T_a$.

The non-context-free rules have the form $p_i = A_{i,1}A_{i,2} \ldots A_{i,n_i} \to B_{i,1}B_{i,2} \ldots B_{i,n_i}$ with $A_{i,k} \in V$ and $B_{i,k} \in V^+$ for $1 \leq i \leq r$, $1 \leq k \leq n_i$. Let

$$X_{i,k} = (\eta(A_{i,k}), p_i, k, \eta(B_{i,k}))$$

for $1 \leq i \leq r$, $1 \leq k \leq n_i$, $p_i = A_{i,1}A_{i,2} \ldots A_{i,n_i} \to B_{i,1}B_{i,2} \ldots B_{i,n_i}$ and

$$M = \{\, X_{i,k} \mid 1 \leq i \leq r,\ 1 \leq k \leq n_i \,\}.$$

The language L can be generated by a tree controlled grammar similar to [2, Theorem 2.3.2].

The terminal symbols are derived in the last step (because otherwise it is not to be seen whether two adjacent symbols in a word of some level are neighbours in the sentential form, if they wrongly would be regarded as adjacent, then a rule could be applied that is not applicable to the sentential form).

We construct a tree controlled grammar $G' = (N', T, P', S, R)$ as follows:

$$\begin{aligned}
N' &= W \cup M,\\
P' &= \{\, A \to \eta(w) \mid A \to w \in P \text{ for } A \in N \text{ and } w \in V^+ \,\}\\
&\quad \cup \{\, A \to X_{i,k} \mid X_{i,k} = (A, p_i, k, B) \in M \,\}\\
&\quad \cup \{\, X_{i,k} \to B \mid X_{i,k} = (A, p_i, k, B) \in M \,\}\\
&\quad \cup \{\, A \to A \mid A \in W \,\} \cup \{\, T_a \to a \mid a \in T \,\},\\
R &= W^* \cup (W^*\{\, X_{i,1}X_{i,2} \ldots X_{i,n_i} \mid 1 \leq i \leq r \,\}W^*)^*.
\end{aligned}$$

In each step, context-free rules, chain rules and rules simulating non-context-free rules can be applied simultaneously. It is easy to see that the tree controlled grammar G' generates the language L.

The control set R is set is accepted by a finite automaton with $2+\sum_{i=1}^{r}(n_i-1)$ states. The transition graph is illustrated in Figure 8.

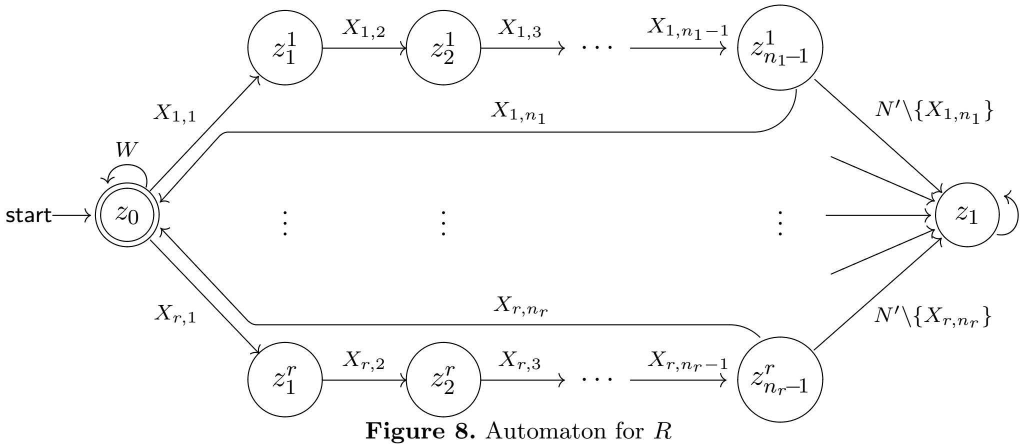

Figure 8. Automaton for R

The start state z_0 is the only accepting state, we have one 'trap' state z_1 for errors and further states z_k^i after reading a symbol $X_{i,k}$ $(1 \leq i \leq r,\ 1 \leq k \leq n_i - 1)$ – after reading X_{i,n_i} correctly, we return to z_0. □

However, the sets $\mathcal{TC}(REG_n)$, $n \geq 1$, do not form an infinite hierarchy. As shown by Stiebe, every context-sensitive language language can be generated by a tree controlled grammar whose control language is accepted by a deterministic finite automaton with at most five states ([7]).

References

[1] K. Culik II and H. Maurer, Tree controlled grammars. *Computing* **19** (1977) 129–139.

[2] J. Dassow and Gh. Păun, *Regulated Rewriting in Formal Language Theory.* EATCS Monographs on Theoretical Computer Science 18, Springer-Verlag, 1989.

[3] J. Dassow and B. Truthe, Subregularly tree controlled grammars and languages. In: E. Csuhaj-Varjú and Z. Ésik (eds.), *Automata and Formal Languages. 12th International Conference, AFL 2008, Balatonfüred, Hungary, May 27–30, 2008. Proceedings.* Computer and Automation Research Institute, Hungarian Academy of Sciences, 2008, 158–169.

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

[5] G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems.* Academic Press, 1980.

[6] G. Rozenberg and A. Salomaa (Eds.), *Handbook of Formal Languages*, Vol. I−III. Springer-Verlag, Berlin, 1997.

[7] R. Stiebe, The hierarchy $\mathcal{TC}(REG_n)$ collapses. Manuscript, 2008.