Grammars with Restricted Derivation Trees

Jiří Koutný[*]

Department of Information Systems
Faculty of Information Technology
Brno Universtity of Technology
Božetěchova 1/2, 616 66 Brno, Czech Republic
ikoutny@fit.vutbr.cz

Abstract

This extended abstract of the doctoral thesis studies theoretical properties of grammars with restricted derivation trees. After presenting the state of the art concerning this investigation area, the research is focused on the three main kinds of the restrictions placed upon the derivation trees. First, it introduces completely new investigation area represented by cut-based restriction and examines the power of the grammars restricted in this way. Second, it investigates several new properties of path-based restriction placed upon the derivation trees. Specifically, it studies the impact of *erasing productions* on the power of grammars with restricted path and introduces two corresponding normal forms. Then, it describes a new relation between grammars with restricted path and some pseudoknots. Next, it presents a counterargument to the power of grammars with controlled path that has been considered as well-known so far. Finally, it introduces a generalization of path-based restriction to not just one but several paths. The model generalized in this way is studied, namely its pumping, closure, and parsing properties.

Categories and Subject Descriptors

F.4.3 [**Mathematical Logic and Formal Languages**]: Formal Languages—*Classes defined by grammars or automata, Operations on languages*

Keywords

tree controlled grammars, level controlled grammars, path controlled grammars, paths controlled grammars, cut controlled grammars, ordered cut controlled grammars, regulated rewriting, restricted derivation trees

---

1. Introduction

The *formal language theory* is an inherent part of the *theoretical computer science* particularly concerned with the study of the formal models. The formal models are mathematical objects used to describe the formal languages. The fundamental models include grammars and automata. The former are used to generate words and the latter accept them.

Grammars are the kind of the *rewriting models* that start from a specified symbol (i.e., start symbol). Then, the symbol is modified according to the given set of rewriting productions. Each production is composed of two components—the left-hand side and the right-hand side of a production. The application of a production on a word is done by rewriting a symbol equivalent to the left-hand side of a production by its right-hand side in the word. This process is known as a *derivation step*. During the computation of a derivation step, just one symbol is rewritten in the word. Given a start symbol of a grammar, a derivation step is computed repeatedly by applying the productions from the given set. Once the resulting word is composed of the symbols that cannot be rewritten anymore, the process of applying derivation steps ends and the resulting word belongs to the language of the grammar.

Essentially, the grammars are composed of a finitely many symbols that are rewritten by finitely many production in finitely many derivation steps. In this way, the grammars represent a finite description of even infinite languages. By the notion of infinite languages are meant those languages that contains infinitely many words. Since the most of the languages commonly used in practice are infinite, the grammars represent a powerful tool how to deal with them. In the formal language theory, there exists a huge variety of grammars which essentially differ in two domains. Specifically, in the complexity of the productions and in the way how to select appropriate production to be applied in a derivation step.

Generally, the complexity of rewriting productions can be seen from two angles—theoretical and practical. Theoretical viewpoint: As little as possible restrictions placed on the form of the rewriting productions in a rewriting model is desirable. More specifically, the more complex rewriting productions are, the larger class of languages may the model generate. In other words, to generate complex languages, complex productions are needed. By the notion of a form of a production, namely the number of the symbols on its left-hand side is meant. Practical viewpoint: Grammars are theoretical models that are implemented

in many practical applications. From the perspective of cost-effective implementation, the simple rewriting productions are desirable. As simple-enough productions for effective implementation, those of the form with just one symbol on their left-hand side are considered. Such kind of productions are referred to as *context-free productions*, since they can be applied without any consideration of a context of currently rewritten symbol.

Non-regulated rewriting models like grammars and automata belong to the well-known core of the formal language theory and they are frequently used in practice. Indeed, automata including its variants underlie lexical analysers (see [3] and [30]), context-free grammars represent the basis of both top-down as well as bottom-up parsers (see [3] and [4]), etc. However, the power of the models with simple productions is indeed smaller than required for usage in many practical applications. On the other hand, the models that use only simple productions are usually easier to implement. As a background, it is desirable to extend context-free grammars as in many applications there are some natural phenomena which cannot be captured by context-free rewriting. More precisely, the motivation is based on the observation that many of the languages commonly used in practice, including programming and natural languages, are not context-free (see [15], [16], [34], [35], and [39]). Consequently, that means such languages cannot be generated by a grammar with only context-free productions. For these reasons, the idea whether or not it would be possible to use grammars with only context-free productions and increase the corresponding power in some other way—without changing the form of rewriting productions.

Basically, this can be achieved by two fundamental approaches—using a kind of a regulation of rewriting or using more than one grammar with context-free productions in a model: Using a kind of a regulation of rewriting. By the notion of a *regulation*, the way how to select appropriate production to be applied is meant. Indeed, a situation is common in which, given a word, it is possible to apply several productions. Informally, the essential idea is represented by the observation that a regulation mechanism somehow prescribes the order of productions the grammar must follow. Therefore, many different kinds of such a regulation have been introduced in order to ensure selecting appropriate production. All of the resulting models based on a kind of regulation are collectively referred to as *regulated rewriting models*. Using more than one grammar with context-free productions in a model. Roughly speaking, the main underlying idea is based on the observation that from the cooperation of several simple models, we can obtain more power than from each of them if they work separately. These systems were also thoroughly studied and the corresponding investigation area is referred to as the *theory of grammar systems*. However, we will deal with them only rarely in this work.

Informally, the goal of this work is to introduce a model that generates more than context-free languages and is usable in practice. From the theoretical viewpoint it means, the model should be able to generate namely the programming languages and the languages used in linguistics (e.g., multiple repetition, cross-dependencies, and copy-language). From the practical viewpoint, it should be possible to develop sophisticated parsing methods working in a polynomial time for the model.

One way to extend the power of context-free grammars is to consider *context-sensitive grammars* where the productions are more complex. Indeed, context-sensitive grammars contain the productions with even more than a single symbol on the left-hand side. However, despite their great power, generating complex languages by context-sensitive grammars actually leads to several fundamental problems making their practical usage problematic (see [6], [8], [29], [36], [44], and [45]). Specifically, for context-sensitive grammars, many problems are undecidable, it is difficult to describe the derivation by a graph structure, etc.

One of the many others approaches extending the power of context-free grammars is represented by *matrix grammars* introduced by Abraham in [1]. The fundamental underlying principle in a derivation step in matrix grammars is that not just one but a fixed number of context-free productions are required to be applied in a given order. This provides synchronization among different parts of a generated word and many non-context-free languages can be generated in this way (see [32], [38], and [46]).

There are lots of other well-known approaches for extending the power of context-free grammars which preserve the context-free nature of productions. Specifically, *Random Context Grammars* (see [43]), *Programmed Grammars* (see [37]), *Ordered Grammars* (see [14]), *Indian and Russian Parallel Grammars* (see [25]), *Indexed Grammar* (see [2]), and many others. However, these approaches do not represent the main topic of this work although some connections can probably be found.

2.   Derivation Tree Restricted Models

One of the power-extending approaches is represented by the restrictions placed upon the derivation trees. Given a grammar, by the notion of a *derivation tree*, a graph structure depicting the application of productions on the start symbol up to the resulting word is meant. Indisputably, the investigation of context-free grammars with restricted derivation trees represents an important trend in today's formal language theory (see [7], [9], [11], [13], [24], [17], [19], [20], [22], [26], [27], [28], [31], [33], [41], and [42]). In essence, these grammars generate their languages just as ordinary context-free grammars do but their derivation trees must satisfy some simple prescribed conditions.

The following two sections gives an informal overview of the results related to the investigation of derivation-tree-restricted grammars. Through those two sections, it is assumed that the reader is familiar with the graph theory (see [5]) and the theory of formal languages (see [29]), including the theory of regulated rewriting (see [10]).

2.1   Level Based Restriction

The idea of restrictions placed upon the derivation trees of context-free grammars is introduced by Culik and Maurer in [9] and the resulting grammars restricted in this way are referred to as *tree controlled grammars*. In essence, the notion of a tree controlled grammar is defined as follows: take a context-free grammar, G, and a regular language, R. A word, w, generated by G belongs to the language defined by G and R if there is a derivation tree, t, for w in G such that all levels of t (except the last one) are described by R. Given a tree controlled grammar, (G, R), G and R are referred to as controlled grammar and control language, respectively. Culik and Maurer investigate several

basic properties of tree controlled grammars—namely, the membership problem and the power.

Based on the original definition of a tree controlled grammar, Păun studies the modifications where many well-known types of both controlled grammars and control languages are considered in [33]. More precisely, Păun studies controlling the levels of the derivation trees of a regular grammar by several types of a control language, controlling the levels of the derivation trees of a context-free grammar without erasing productions by several types of a control language, and controlling the levels of the derivation trees of a context-free grammar by a finite language.

It is well-known that tree controlled grammars with a context-free grammar controlled by a regular language characterize the class of recursively enumerable languages. Thus, the question arises whether or not it is possible to achieve the same power as tree controlled grammars have when the levels of the derivation trees are restricted by a subregular control language. This problem is studied by Dassow and Truthe in [13], where many types of subregular languages are considered. Dassow and Truthe study primarily controlling the levels of the derivation trees of a context-free grammar by two types of a language such that one is a subset of the other and controlling the levels of the derivation trees of a context-free grammar by many different types of subregular languages. The same authors, Dassow and Truthe, also study hierarchies of subregularly tree controlled languages in [11] and [12]. They present controlling the levels of the derivation trees of a context-free grammar by the union of monoids, by regular languages with restricted descriptional complexity, and by the language accepted by a deterministic finite automaton with at most given number of states.

Stiebe in [40] states that there is a tree controlled grammar for every linearly bounded queue automaton. Then, Stiebe proves that controlling the levels of the derivation trees of a context-free grammar by the language accepted by a minimal finite automaton with at most five states characterize the class of context-sensitive languages. If, additionally, erasing productions in a controlled grammar are allowed, controlling the levels of the derivation trees of a context-free grammar by the language accepted by a minimal finite automaton with at most five states characterizes the class of recursively enumerable languages.

Turaev, Dassow, and Selamat in [41] examine tree controlled grammars with bounded nonterminal complexity and demonstrate that seven nonterminals in a tree controlled grammar are enough to generate any recursively enumerable language. Then, they establish that three nonterminals in a tree controlled grammar are enough to generate any regular language and any regular simple matrix language can be generated by a tree controlled grammar with three nonterminals. Finally, they demonstrate that three nonterminals in a tree controlled grammar are enough to generate any linear language. The same authors in [42] state several further nonterminal-complexity-related properties of tree controlled grammars.

2.2 Path Based Restriction
As an attempt to increase the power of context-free grammars without changing the basic formalism and loosing some basic properties of context-free languages (decidability, efficient parsing, etc.), Marcus, Martín-Vide, Mi-

trana, and Păun in [26] study a new type of a restriction in a derivation: a derivation tree in a context-free grammar is accepted if it contains a path described by a control language. More precisely, they consider two context-free grammars, G and G'. A word, w, generated by G belongs to the language defined by G and G' if there is a derivation tree, t, for w in G such that there exists a path of t described by the language of G'. Based on this restriction, they introduce a *path controlled grammar* and study several properties of this model. Specifically, they study controlling a path of the derivation trees of several types of grammars by a regular language and controlling a path of the derivation trees of a regular grammar by a linear or context-free language. Then, they establish two kinds of pumping properties depending on the type of a controlled grammar, some consequences to the closure properties of path controlled grammars, and a basic parsing property for path controlled grammars. They also investigate the power of path controlled grammars. However, there exists a serious problem with the correctness of the proof they present.

As a continuation of the investigation of path-based restrictions, Martín-Vide and Mitrana study parsing properties of path controlled grammars, closure properties of path controlled grammars, and several decision problems for path controlled grammars in [27] and [28].

2.3 Goals of the Thesis
As it clearly follows from the previous sections, level-based restriction is well-studied and the most of the important questions have been answered. On the other hand, in the case of path-based restriction many basic properties including the generative power have not been successfully investigated yet. Moreover, several other restrictions placed upon the derivation trees have not yet been introduced at all. Indeed, the restrictions placed upon the cuts of the derivation tree as well as upon several paths of the derivation trees represent completely new investigation areas. Thus, the goals of the doctoral thesis consist in three investigation areas.

- First, to introduce new investigation area represented by cut-based restrictions and establish the generative power of the model restricted in this way.

- Second, to establish several new results in the investigation of one-path-restricted grammars introduced in [26].

- Third, to generalize one-path-restricted model into several paths and investigate several its properties.

3. New Results in Restrictions Placed upon Derivation Trees
First we reformulate the fundamental definitions so that all derivation-tree-based restrictions can be studied using the same notation. Then, we present new results in the derivation-tree based restrictions investigation area.

Since restriction placed upon a level, a path, and a cut is, in essence, a restriction placed upon a derivation tree, we use a slightly modified but equivalent formulation of the definitions stated in [26], [27], and [28]. Consequently, aforementioned modifications allow us to investigate all derivation-tree-based restrictions using the same terminology—i.e., restriction on the levels (see [9], [13], [33],

[41], and [42]), the paths (see [7], [17], [18], [19], [20], [?], [22], [23], [24], [26], [27], and [28]), or the cuts (see [24]). More precisely, all restrictions placed upon the derivation trees are covered by the general notion of *tree controlled grammar* that generates its language under several kinds of the restrictions.

DEFINITION 3.1. *For a sequence, s, of the nodes of a derivation tree, the* word obtained by concatenating all symbols of s is denoted as word(s).

DEFINITION 3.2. *The class of regular, linear, ε-free context-free, context-free, propagating scattered context grammars, matrix grammars, context-sensitive, and unrestricted grammars is denoted as* **REG**, **LIN**, $\mathbf{CF}_\varepsilon$, **CF**, **PSC**, **MAT**, **CS**, **RE**, *respectively.*

DEFINITION 3.3. *The set of all regular, linear, ε-free context-free, context-free, context-sensitive, and unrestricted grammars is denoted as* $\mathbb{G}_{\mathbf{REG}}$, $\mathbb{G}_{\mathbf{LIN}}$, $\mathbb{G}_{\mathbf{CF}_\varepsilon}$, $\mathbb{G}_{\mathbf{CF}}$, $\mathbb{G}_{\mathbf{CS}}$, *and* $\mathbb{G}_{\mathbf{RE}}$, *respectively.*

Note that hereafter the notion *tree controlled grammar* is used in different meaning than in the previous section, see the following definitions of a *tree controlled grammar* and the definitions of the languages as well as the classes that tree controlled grammars generate under various kinds of restrictions that are introduced in the following three sections.

DEFINITION 3.4. *A* tree controlled *grammar is a pair,* (G, R), *where* $G = (V, T, P, S)$ *is a controlled grammar, and* R *is a control language over* V.

DEFINITION 3.5. *Let* (G, R) *be a tree controlled grammar where* $G = (V, T, P, S)$, *then* $_{(G,R)}\triangle(x)$, $x \in V^*$, *denotes the set of the derivation trees with frontier x in G.*

In the research presented through this section, we do not directly deal with level-based restriction placed upon the derivation trees. However, for the sake of completeness, note the following definitions related to level-based restriction placed upon the derivation trees.

DEFINITION 3.6. *Let* (G, R) *be a tree controlled grammar. The* language that (G, R) generates under the levels control by R is denoted by $_{levels}L(G, R)$ and defined by the following equivalence:

For all $x \in T^*$, $x \in {}_{levels}L(G, R)$ if and only if there is a derivation tree, $t \in {}_G\triangle(x)$, such that for all levels, s, of t (except the last one), word$(s) \in R$.

DEFINITION 3.7. *For some language classes* **X** *and* **Y**, *the class of* tree controlled languages under the levels control *is defined as*
levels-TC(X, Y) $= \{_{levels}L(G, R) : (G, R)$ *is a tree controlled grammar in which* $G \in \mathbb{G}_{\mathbf{X}}$ *and* $R \in \mathbf{Y}\}$.

Next, we summarize the most interesting results and point out some important open questions. Based on the *State of the Art* in the area of restrictions placed upon the derivation trees summarized above, this work deals in principle with three kinds of derivation-tree based restrictions, cut-based, path-based, and several-path-based.

3.1 Cut Based Restriction

In this section, we introduce new derivation-tree-based restrictions of tree-controlled grammars which are based on the restriction placed upon the cuts. Then, we introduce several properties of grammars with restriction placed upon the cuts. Specifically, we investigate the power.

3.1.1 Definitions

DEFINITION 3.8. *Let* (G, R) *be a tree controlled grammar. The* language that (G, R) generates under the cuts control by R is denoted by $_{cut}L(G, R)$ and defined by the following equivalence:

For all $x \in T^*$, $x \in {}_{cut}L(G, R)$ if and only if there is a derivation tree, $t \in {}_G\triangle(x)$, and a set, $_xM$, of its cuts such that

1. for each $c \in {}_xM$, word$(c) \in R$, and

2. $_xM$ covers the whole t.

DEFINITION 3.9. *The class of* tree controlled languages under the cuts control *is defined as*
cut-TC(CF, REG) $= \{_{cut}L(G, R) : (G, R)$ *is a tree controlled grammar in which G is a context-free grammar and* $R \in \mathbf{REG}\}$ *and the* class of tree controlled languages with ε-free controlled grammar under cuts control *is defined as*
cut-TC$_\varepsilon$(CF$_\varepsilon$, REG) $= \{_{cut}L(G, R) : (G, R)$ *is a tree controlled grammar in which G is an ε-free context-free grammar and* $R \in \mathbf{REG}\}$.

DEFINITION 3.10. *Let* $\preceq$ *be a binary relation on a sequence,* $_xM$, *of the cuts such that for each two cuts,* $c_1, c_2 \in {}_xM$, $c_1 \preceq c_2$ *if and only if for each node, n_2, of c_2 either there is a node, n_1, of c_1 such that n_2 is a direct descendent of n_1, or $n_1 = n_2$. In other words, $n_1 \neq n_2$ implies n_2 is a direct descendent of n_1.*

DEFINITION 3.11. *Let* (G, R) *be a tree controlled grammar. The* language that (G, R) generates under the ordered-cuts control by R is denoted by $_{ord\text{-}cut}L(G, R)$ and defined by the following equivalence:

For all $x \in T^*$, $x \in {}_{ord\text{-}cut}L(G, R)$ if and only if there is a derivation tree, $t \in {}_G\triangle(x)$, and a sequence, $c_{1_x}, c_{2_x}, \ldots,$
c_{n_x}, of the cuts of t, for some $n_x \geq 1$, such that

1. for all $i = 1_x, 2_x, \ldots, n_x$, word$(c_i) \in R$,

2. $\{c_{1_x}, c_{2_x}, \ldots, c_{n_x}\}$ covers the whole t, and

3. $c_{i_x} \preceq c_{(i+1)_x}$ for all $i = 1, 2, \ldots, n - 1$.

DEFINITION 3.12. *The class of* tree controlled languages under the ordered cuts control *is defined as*

ord-cut-TC($\mathbf{CF}, \mathbf{REG}$) = {$_{ord\text{-}cut}L(G, R)$: (G, R) is a tree controlled grammar in which G is a context-free grammar and $R \in \mathbf{REG}$} and the class of tree controlled languages with ε-free controlled grammar under ordered cuts control is defined as
ord-cut-TC$_\varepsilon$($\mathbf{CF}_\varepsilon, \mathbf{REG}$) = {$_{ord\text{-}cut}L(G, R)$: (G, R) is a tree controlled grammar in which G is an ε-free context-free grammar and $R \in \mathbf{REG}$}.

3.1.2 Results

Concerning cut-based restriction placed upon the derivation trees, we have introduced two fundamental types of such kind of a restriction and thus, we have opened a new investigation area in derivation-tree-restricted models. Next, we have proved that both restrictions increase the power of context-free grammars so they characterize $\mathbf{RE}$:

$$\mathbf{ord\text{-}cut\text{-}TC(CF, REG)} = \mathbf{cut\text{-}TC(CF, REG)} = \mathbf{RE}.$$

An important open problem consists of the investigation of cut controlled grammars where ε-productions are forbidden. Consequently, the grammars restricted in this way should be placed into the relation with some other well-known language families, such as $\mathbf{CS}$, and the deciding the question whether or not:

$$\mathbf{ord\text{-}cut\text{-}TC}_\varepsilon(\mathbf{CF}_\varepsilon, \mathbf{REG}) = \mathbf{CS},$$

$$\mathbf{cut\text{-}TC}_\varepsilon(\mathbf{CF}_\varepsilon, \mathbf{REG}) = \mathbf{CS}.$$

Next problem is represented by the descriptional complexity of $\mathbf{ord\text{-}cut\text{-}TC(CF, REG)}$ and $\mathbf{cut\text{-}TC(CF, REG)}$. The results presented above are based on the transformation of an unrestricted grammar into Pentonnen normal form. However, using Geffert normal form, the number of nonterminals in the resulting cut controlled grammar would be reduced.

Another future research idea is represented by the controlling the cuts of the derivation trees in which several types of subregular control languages are considered. In this way, the question whether or not a kind of a subregular language is enough to increase the power of controlled grammar properly. Consequently, the relation between the power of level-based and cut-based models restricted in this way would be founded out.

3.2 Path Based Restriction

In this section, we introduce a path-based restriction on tree-controlled grammars that is equivalent to the model introduced in [26], [27], and [28]. Next, we formally define the pseudoknot structure represented as a language. We introduce new results related to the *normal forms* and the presence of erasing productions in a controlled grammar. Then, this section presents a relationship between biology and the formal language theory in the form of word representation of pseudoknots generated by path controlled grammars. Last section of this part being a counterargument against the proof of the power of path controlled grammars that has been considered as correct so far.

3.2.1 Definitions

DEFINITION 3.13. *Let (G, R) be a tree controlled grammar. The* language that (G, R) generates under the path

control by R is denoted by $_{path}L(G, R)$ and defined by the following equivalence:

For all $x \in T^*$, $x \in {}_{path}L(G, R)$ if and only if there is a derivation tree, $t \in {}_G\triangle(x)$, such that there is a path, p, of t with $\mathrm{word}(p) \in R$.

DEFINITION 3.14. *For $\mathbf{X}, \mathbf{Y} \in \{\mathbf{LIN}, \mathbf{CF}\}$, the* class of tree controlled languages under the path control *is defined as*
path-TC($\mathbf{X}, \mathbf{Y}$) = {$_{path}L(G, R)$: (G, R) is a tree controlled grammar in which $G \in \mathbb{G}_\mathbf{X}$ and $R \in \mathbf{Y}$} and the class of tree controlled languages with ε-free controlled grammar under path control is defined as
path-TC$_\varepsilon$($\mathbf{CF}_\varepsilon, \mathbf{Y}$) = {$_{path}L(G, R)$: (G, R) is a tree controlled grammar in which G is an ε-free context-free grammar and $R \in \mathbf{Y}$}.

DEFINITION 3.15. *Let (G, R) be a tree controlled grammar that generates the language under path control by R, where $G = (V, T, P, S)$. (G, R) is in 1st normal form if every production, $r : A \to x \in P$, is of the form $A \in V - T$ and $x \in T \cup (V - T) \cup (V - T)^2$.*

DEFINITION 3.16. *Let (G, R) be a tree controlled grammar that generates the language under path control by R, where $G = (V, T, P, S)$. (G, R) is in 2nd normal form if every production, $r : A \to x \in P$, is of the form $A \in V - T$ and $x \in T \cup ((V \cup \{E\}) - T)^2$ where $E \cap V = \emptyset$ and $E \to \varepsilon \in P$. The alphabet of G should now include E, with $E \notin V$.*

DEFINITION 3.17. *Let Σ be an alphabet. The following languages over Σ are pseudoknots:*

1. $\{xyx^Ry^R : x, y \in \Sigma^*\}$,
   $\{u_1xu_2yu_3x^Ru_4y^Ru_5 : x, y, u_i \in \Sigma^*, 1 \le i \le 5\}$,

2. $\{xyx^Rzz^Ry^R : x, y, z \in \Sigma^*\}$,
   $\{u_1xu_2yu_3x^Ru_4zu_5z^Ru_6y^Ru_7 : x, y, z, u_i \in \Sigma^*, 1 \le i \le 7\}$,

3. $\{xyx^Rzy^Rz^R : x, y, z \in \Sigma^*\}$,
   $\{u_1xu_2yu_3x^Ru_4zu_5y^Ru_6z^Ru_7 : x, y, z, u_i \in \Sigma^*, 1 \le i \le 7\}$,

4. $\{xyzx^Ry^Rz^R : x, y, z \in \Sigma^*\}$,
   $\{u_1xu_2yu_3zu_4x^Ru_5y^Ru_6z^Ru_7 : x, y, z, u_i \in \Sigma^*, 1 \le i \le 7\}$.

Note that presented pseudoknots form obviously non-context-free languages.

3.2.2 Results

As a continuation of the investigation of path-based restrictions introduced in [26] and studied in [27] and [28], we have considered the impact of ε-productions in path controlled grammars to the power and we have stated that ε-productions can be removed from a path controlled grammar without affecting its language:

$$\mathbf{path\text{-}TC(CF, CF)} = \mathbf{path\text{-}TC}_\varepsilon(\mathbf{CF}_\varepsilon, \mathbf{CF}).$$

We have established two Chomsky-like normal forms for path controlled grammars and we have formulated algorithms that transform a path controlled grammar into its normal form:

- Let $L \in$ **path-TC**(**CF**, **CF**). Then, there exists a tree controlled grammar, (G, R), in 1st normal form such that $L = {}_{path}L(G, R)$.

- Let $L \in$ **path-TC**(**CF**, **CF**). Then, there exists a tree controlled grammar, (G, R), in 2nd normal form such that $L = {}_{path}L(G, R)$.

A future investigation idea consists of the modifying a general parsing methods that are based on Chomsky normal form such that it will be able to parse path controlled grammars in a polynomial time.

Another practically motivated idea is represented the relation between path controlled grammars and the theory of pseudoknots. We have demonstrated several typical pseudoknots used in biology represented by the words of non-context-free languages. We have demonstrated some pseudoknots belong to **path-TC**(**LIN**, **LIN**):

$$\{xyx^R y^R : x, y \in \Sigma^*\} \in \textbf{path-TC}(\textbf{LIN}, \textbf{LIN}),$$
$$\{xyx^R z z^R y^R : x, y, z \in \Sigma^*\} \in \textbf{path-TC}(\textbf{LIN}, \textbf{LIN}),$$
$$\{xyx^R z y^R z^R : x, y, z \in \Sigma^*\} \in \textbf{path-TC}(\textbf{LIN}, \textbf{LIN}).$$

Apparently, there is a huge variety of another pseudoknot structures in biology. For example, $\{xyzx^R y^R z^R : x, y, z \in \Sigma^*\}$ and it is an open question whether or not those pseudoknots can be generated by tree controlled grammars with linear components that generate the language under path control.

The last presented result deals with a reflection on the power of path controlled grammars that has been considered as well-known (see [26]) for more than last ten years. However, we have presented an argument against the correctness of the proof given in [26] that states

$$\textbf{path-TC}(\textbf{CF}, \textbf{CF}) \subseteq \textbf{MAT}.$$

We have concluded the counterargument by stating that the aforementioned inclusion still may hold; however, it cannot be proved in the way given in [26]. More precisely, we have found the language tha can be generated by a grammar with controlled path. However, this language cannot be generated by the grammar obtained by the construction introduced in [26]. Apparently, the power of path controlled grammar still represents an open problem.

3.3   Several Paths Based Restriction

This section being a generalization of path-restricted rewriting model to a restriction placed upon not just one but several paths. Then, it presents several properties of the model restricted in this way. Specifically, the power of a kind of all-paths-restricted rewriting model, closure and pumping properties in relation to the number of controlled paths, and the approximation of the power for n-path restricted model. The last section of this part presents an application related result concerning the parsing of path restricted languages.

3.3.1   Definitions

DEFINITION 3.18. *Let (G, R) be a tree controlled grammar. The language that (G, R) generates under the all-paths control by R is denoted by ${}_{all\text{-}paths}L(G, R)$ and defined by the following equivalence:*

For all $x \in T^*$, $x \in {}_{all\text{-}paths}L(G, R)$ if and only if there is a derivation tree, $t \in {}_G\triangle(x)$, such that for all paths, s, of t, word$(s) \in R$.

DEFINITION 3.19. *The class of tree controlled languages under all paths control is defined as*
all-path-TC(**CF**, **REG**) $= \{{}_{all\text{-}paths}L(G, R) : (G, R)$ *is a tree controlled grammar in which G is a context-free grammar and $R \in$ **REG**$\}$.*

DEFINITION 3.20. *Let (G, R) be a tree controlled grammar. The language of tree controlled grammar under not common n-path control by R, $n \geq 1$, is denoted by ${}_{nc\text{-}n\text{-}path}L(G, R)$ and defined by the following equivalence:*

For all $x \in T^*$, $x \in {}_{nc\text{-}n\text{-}path}L(G, R)$ if there exists a derivation tree, $t \in {}_G\triangle(x)$, such that there is a set, Q_t, of n paths of t such that for each path, $p \in Q_t$, word$(p) \in R$.

DEFINITION 3.21. *For $\mathbf{X}, \mathbf{Y} \in \{\textbf{REG}, \textbf{LIN}, \textbf{CF}\}$, the class of tree controlled languages under not-common n-path control is defined as*
nc-n-path-TC(**X**, **Y**) $= \{{}_{nc\text{-}n\text{-}path}L(G, R) : (G, R)$ *is a tree controlled grammar with $G \in \mathbb{G}_{\mathbf{X}}$ and $R \in \mathbf{Y}\}$.*

DEFINITION 3.22. *Let p_1, p_2 be any different two paths of a derivation tree, t. Then, p_1 and p_2 contain at least one common node (the root of t, root(t)), and p_1 ends in a different leaf of t than p_2. Let $\text{cmn}(p_1, p_2)$ denote the maximal number of consecutive common nodes of p_1 and p_2.*

DEFINITION 3.23. *Let Q_t be a nonempty set of some paths of a derivation tree, t. The paths of Q_t are divided in a common node of t if and only if for some $k \geq 1$, $\text{cmn}(p_1, p_2) = k$ for every pair $(p_1, p_2) \in Q_t^2$. Let all paths of Q_t be divided in a common node of t. If $Q_t = \{p\}$, then $m_{Q_t} = |\text{word}(p)|$, otherwise $m_{Q_t} \geq 1$ denotes the maximal number of consecutive common nodes of all paths in Q_t.*

DEFINITION 3.24. *Let (G, R) be a tree controlled grammar. The language of tree controlled grammar under n-path control by R, $n \geq 1$, is denoted by ${}_{n\text{-}path}L(G, R)$ and defined by the following equivalence:*

For all $x \in T^*$, $x \in {}_{n\text{-}path}L(G, R)$ if there exists a derivation tree, $t \in {}_G\triangle(x)$, such that there is a set, Q_t, of n paths of t that are divided in a common node of t and for each path, $p \in Q_t$, word$(p) \in R$.

DEFINITION 3.25. *For $\mathbf{X}, \mathbf{Y} \in \{\textbf{REG}, \textbf{LIN}, \textbf{CF}\}$, the class of tree controlled languages under n-path control is defined as*
n-path-TC(**X**, **Y**) $= \{{}_{n\text{-}path}L(G, R) : (G, R)$ *is a tree controlled grammar with $G \in \mathbb{G}_{\mathbf{X}}$ and $R \in \mathbf{Y}\}$.*

Conventions
Hereafter, tree controlled grammars that generate the language under the n-path control are referred to as *n-path tree controlled grammars.*

Note that if we consider 0 controlled paths (i.e., $n = 0$ and consequently card$(Q_t) = 0$) in the definition of $_{n\text{-}path}L(G, R)$, then, clearly, the power of such a model equals **CF**.

DEFINITION 3.26. *Let (G, R) be a tree controlled grammar. Consider $_{n\text{-}path}L(G, R)$, for $n \geq 1$. If for each word, $z \in {}_{n\text{-}path}L(G, R)$, there exist a derivation tree, $t \in {}_G\triangle(z)$, a set of its paths, Q_t, $m_{Q_t} \geq 1$, and a partition, word$(p) = uvwxy$, for each path, $p \in Q_t$, satisfying the premise of the pumping lemma for linear languages such that it holds*

- $1 \leq m_{Q_t} \leq |u|$,
  then $_{n\text{-}path}L(G, R)$ is $_{I\text{-}n\text{-}path}L(G, R)$,

- $|u| < m_{Q_t} \leq |uv|$,
  then $_{n\text{-}path}L(G, R)$ is $_{II\text{-}n\text{-}path}L(G, R)$,

- $|uv| < m_{Q_t} \leq |uvw|$,
  then $_{n\text{-}path}L(G, R)$ is $_{III\text{-}n\text{-}path}L(G, R)$,

- $|uvw| < m_{Q_t} \leq |uvwx|$,
  then $_{n\text{-}path}L(G, R)$ is $_{IV\text{-}n\text{-}path}L(G, R)$,

- $|uvwx| < m_{Q_t} \leq |uvwxy|$,
  then $_{n\text{-}path}L(G, R)$ is $_{V\text{-}n\text{-}path}L(G, R)$.

DEFINITION 3.27. *For $i \in \{$**I, II, III, IV, V**$\}$, $n \geq 1$, the class of i-n-path tree controlled languages is defined as*
i-**n-path-TC(CF, LIN)** $= \{_{i\text{-}n\text{-}path}L(G, R) : (G, R)$ *is tree controlled grammar in which G is a context-free grammar and $R \in$ **LIN**$\}$.*

3.3.2   Results
It is well-know that path controlled grammars where the controlling grammar is regular characterize the same language class as its controlled grammar (see [26]) do. We have proved that the power of context-free grammars remains unchanged even if we restrict all paths in their derivation trees by regular languages:

$$\mathbf{CF} = \mathbf{all\text{-}path\text{-}TC(CF, REG)}.$$

Since for each context-free grammar, there is a regular language that describes all paths in all its derivation trees; and there is no regular language which increases its power when used to restrict the paths, if we consider tree controlled grammars (G, R) with $R \in$ **REG**, then, obviously, the power of such a model equals **CF** for any $n \geq 1$. Therefore, we investigate the properties of tree controlled grammar with non-regular control language. More specifically, we study tree controlled grammars that generates their languages under n-path control with linear control languages.

We have introduced a generalization of path controlled grammars so that they generate the language under the restriction placed on not just one but several paths. Consequently, we have found some subsets of n-path controlled grammars so their languages satisfy pumping premises similar to well-known premises stated by pumping lemmata for **CF**, **LIN**, and **REG**—more precisely:

- If $L \in$ **I-n-path-TC(CF, LIN)**, $n \geq 1$, then there are two constants, $k, q \geq 0$, such that each word, $z \in L$, with $|z| \geq k$ can be written as
  $$z = u_1 v_1 u_2 v_2 \ldots u_{4n} v_{4n} u_{4n+1}$$
  with $0 < |v_1 v_2 \ldots v_{4n}| \leq q$ and for all $i \geq 1$, $u_1 v_1^i u_2 v_2^i \ldots u_{4n} v_{4n}^i u_{4n+1} \in L$.

- If $L \in$ **III-n-path-TC(CF, LIN)**, $n \geq 1$, then there are two constants, $k, q \geq 0$, such that each word, $z \in L$, with $|z| \geq k$ can be written as
  $$z = u_1 v_1 u_2 v_2 \ldots u_{2n+2} v_{2n+2} u_{2n+3}$$
  with $0 < |v_1 v_2 \ldots v_{2n+2}| \leq q$ and for all $i \geq 1$, $u_1 v_1^i u_2 v_2^i \ldots u_{2n+2} v_{2n+2}^i u_{2n+3} \in L$.

- If $L \in$ **V-n-path-TC(CF, LIN)**, $n \geq 1$, then there are two constants, $k, q \geq 0$, such that each word, $z \in L$, with $|z| \geq k$ can be written as
  $$z = u_1 v_1 u_2 v_2 u_3 v_3 u_4 v_4 u_5$$
  with $0 < |v_1 v_2 v_3 v_4| \leq q$ and for all $i \geq 1$, $u_1 v_1^i u_2 v_2^i u_3 v_3^i u_4 v_4^i u_5 \in L$.

A natural question that still remains open is whether or not there are similar pumping properties also for the languages of **II-n-path-TC(CF, LIN)** and **IV-n-path-TC (CF, LIN)**.

We have also proved some closure properties, that is

- for $n \geq 1$, $i \in \{$**I, II, III, IV, V**$\}$ and $T_G, T_R \in \{$**REG, LIN, CF**$\}$,
  $$\mathbf{n\text{-}path\text{-}TC}(T_G, T_R), i\text{-}\mathbf{n\text{-}path\text{-}TC}(T_G, T_R)$$
  are closed under intersection with regular languages, union, and non-erasing homomorphism,

- for $n \geq 1$, $i \in \{$**I, III, V**$\}$, i-**n-path-TC(CF, LIN)** are not closed under concatenation, intersection, and complement.

Since n-path controlled grammars are a natural generalization of grammars with just one path controlled, we have studied several properties that are well-known for path controlled grammars in the case of controlling given n paths. Most importantly, we have tried to establish the power for n-path controlled grammar. We have found the approximation of the power that can be applied also on grammars with just one path controlled. However, this approximation does not say too much since it is well-known that **PSC** is not closed under erasing homomorphism. Thus, we have informally concluded that: *"We either have the power to check what we need but not to remove it (using **PSC**) or vice versa (using **MAT**)."* More precisely, we have stated the following:

Let $L \in$ **n-path-TC(CF, CF)**, for $n \geq 1$. Then there exists $L' \in$ **PSC** with $L = h(L')$ for some homomorphism h.

Finally, we have studied several parsing properties of path-based restriction which is indisputably one of the most important language-class-characterizing property from the practical viewpoint. Formally, we have studied a polynomial time parsing possibilities and we have stated that:

- For a tree controlled grammar, (G, R) with an unambiguous context-free grammar, G, and a linear control language, R, the membership

$$x \in {}_{nc\text{-}n\text{-}path}L(G, R),$$

  $n \geq 1$, is decidable in $O(|x|^k)$, for some $k \geq 0$.

- For a tree controlled grammar, (G, R), where G is a context-free grammar and $R \in \mathbf{LIN}$, there is a tree controlled grammar, (G', R'), such that G' does not contain unit productions and ${}_{nc\text{-}n\text{-}path}L(G, R) = {}_{nc\text{-}n\text{-}path}L(G', R')$, $n \geq 1$.

- For tree controlled grammar (G, R) where G is m-ambiguous LR grammar, $m \geq 1$, and an unambiguous language $R \in \mathbf{LIN}$, the membership $x \in {}_{nc\text{-}n\text{-}path}L(G, R)$, $n \geq 1$, is decidable in $O(|x|^k)$, for some $k \geq 0$.

The significant disadvantage concerning n-path tree controlled grammars is that the number of n paths satisfying the properties of Def. 3.24 is strictly limited by the length of the right-hand sides of the productions of underlying context-free grammar. That is, given a general context-free grammar, G, and a linear language, R, controlling the paths, the membership of a certain language might be decidable. However, given the same context-free language as $L(G)$ as a context-free grammar, H, in Chomsky normal form together with R, we might not be able to find suitable path restriction to obtain the same language. On the other hand, the derivation trees of tree controlled grammars that generates their languages under n-path control by a linear language are constructed exactly as in context-free grammars and, in addition, we have to check some of their paths. Thus, there is actually great possibility to use well-known parsing methods for context-free languages to construct the derivation trees and to check their paths. However, in this viewpoint, n-path tree controlled grammars seems to be a quite fragile formalism since it requires a context-free grammar to have a production with at least n nonterminals on the right-hand side which ensures the division of n paths in a common node. Moreover, it means that any attempt to use a parsing method that transforms a context-free grammar into Chomsky normal form will basically destroy any path restriction with $n \geq 3$. Moreover, several nice properties of context-free grammars have been lost—e.g., decomposition based on pumping lemma for linear languages is potentially ambiguous and thus, the membership problem for i-$\mathbf{n}$-$\mathbf{path}$-$\mathbf{TC}$, $i \in \{\mathbf{I}, \mathbf{II}, \mathbf{III}, \mathbf{IV}, \mathbf{V}\}$, is potentially ambiguous also.

There are still many questions to be answered, namely the power of grammars with path or paths controlled non-regularly, further closure properties, decision properties, etc. However, there are several other more general variants of path-based restriction. Indeed, a modification of the formalism such that the paths do not have to be divided in a common node of a derivation tree, or a variant where a path in a tree does not have to start in the root and end in a leaf of the tree have not been investigated yet.

References

[1] S. Abraham. Some questions of language theory. In *Proceedings of the 1965 conference on Computational linguistics*, 1965.

[2] A. V. Aho. Indexed grammars - an extension of context-free grammars. *Journal of the ACM*, 15:25, 1968.

[3] A. V. Aho, R. Sethi, and J. D. Ullman. *Compilers principles, techniques, and tools*. Addison-Wesley, 1986.

[4] A. V. Aho and J. D. Ullman. *The theory of parsing, translation, and compiling*. Prentice-Hall, Inc., 1972.

[5] A. Bondy. *Graph Theory*. Springer, 2010.

[6] R. V. Book. On the structure of context-sensitive grammars. *International Journal of Computer and Information Sciences*, 2:11, 1973.

[7] M. Čermák, J. Koutný, and A. Meduna. Parsing based on n-path tree-controlled grammars. *Theoretical and Applied Informatics*, 2011:16, 2012.

[8] A. B. Cremers. Normal forms for context-sensitive grammars. *Acta Informatica*, 3:15, 1973.

[9] K. Čulik and H. A. Maurer. Tree controlled grammars. *Computing*, 19:11, 1977.

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

[11] J. Dassow, R. Stiebe, and B. Truthe. Two collapsing hierarchies of subregularly tree controlled languages. *Theoretical Computer Science*, 410:11, 2009.

[12] J. Dassow and B. Truthe. On two hierarchies of subregularly tree controlled languages. In *10th International Workshop on Descriptional Complexity of Formal Systems, DCFS 2008, Charlottetown, Prince Edward Island, Canada, July 16-18, 2008*.

[13] J. Dassow and B. Truthe. Subregularly tree controlled grammars and languages. In *Automata and Formal Languages - 12th International Conference AFL 2008, Balatonfured*, 2008.

[14] I. Friš. Grammars with partial ordering of the rules. *Information and Control*, 12:11, 1968.

[15] J. Higginbotham. English is not a context-free language. *Linguistic Inquiry*, 15:10, 1984.

[16] A. K. Joshi. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions. Technical report, University of Pennsylvania, 1985.

[17] J. Koutný. Regular paths in derivation trees of context-free grammars. In *Proceedings of the 15th Conference STUDENT EEICT 2009 Volume 4*, 2009.

[18] J. Koutný. On n-path-controlled grammars. In *Proceedings of the 16th Conference STUDENT EEICT 2010 Volume 5*, 2010.

[19] J. Koutný. Syntax analysis of tree-controlled languages. In *Proceedings of the 17th Conference STUDENT EEICT 2011 Volume 3*, 2011.

[20] J. Koutný. On path-controlled grammars and pseudoknots. In *Proceedings of the 18th Conference STUDENT EEICT 2012 Volume 3*, 2012.

[21] J. Koutný, Z. Křivka, and A. Meduna. On grammars with controlled paths. Acta Cybernetica, submitted.

[22] J. Koutný, Z. Křivka, and A. Meduna. Pumping properties of path-restricted tree-controlled languages. In *7th Doctoral Workshop on Mathematical and Engineering Methods in Computer Science*, 2011.

[23] J. Koutný and A. Meduna. On normal forms and erasing rules in path controlled grammars. Schedae Informaticae, in press.

[24] J. Koutný and A. Meduna. Tree-controlled grammars with restrictions placed upon cuts and paths. *Kybernetika*, 48:11, 2012.

[25] M. K. Levitina. On some grammars with rules of global replacement. *Scientific-technical information (Nauchno-tehnichescaya informacii), Series 2*, page 5, 1972.

[26] S. Marcus, C. Martín-Vide, V. Mitrana, and Gh. Păun. A new-old class of linguistically motivated regulated grammars. In *Walter Daelemans, Khalil Sima'an, Jorn Veenstra, Jakub Zavrel (Eds.): Computational Linguistics in the Netherlands 2000, Selected Papers from the Eleventh CLIN Meeting, Tilburg*, 2000.

[27] C. Martín-Vide and V. Mitrana. Further properties of path-controlled grammars. In *Proceedings of FG-MoL 2005: The 10th Conference on Formal Grammar and The 9th Meeting on Mathematics of Language*, 2005.

[28] C. Martin-Vide and V. Mitrana. Decision problems on path-controlled grammars. *International Journal of Foundations of Computer Science*, 18:11, 2007.

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

[30] A. Meduna. *Elements of Compiler Design*. Taylor and Francis Informa plc, 2008.

[31] P. Navrátil. Parsing based on regulated grammars. Master's thesis, Faculty of Information Technology, Brno University of Technology, 2003.

[32] G. Păun. On the generative capacity of simple matrix grammars of finite index. *Information Processing Letters*, 7:3, 1978.

[33] Gh. Păun. On the generative capacity of tree controlled grammars. *Computing*, 21:8, 1979.

[34] G. K. Pullum. On two recent attempts to show that english is not a CFL. *Computational Linguistics*, 10:5, 1984.

[35] G. K. Pullum and G. Gazdar. Natural languages and context-free languages. *Linguistics and Philosophy*, 4:34, 1982.

[36] Jr. R. L. Cannon. An algebraic technique for context-sensitive parsing. *International Journal of Computer and Information Sciences*, 5:20, 1976.

[37] D. J. Rosenkrantz. Programmed grammars and classes of formal languages. *Journal of the ACM*, 16:25, 1969.

[38] A. Salomaa. Matrix grammars with a leftmost restriction. *Information and Control*, 20:7, 1972.

[39] S. Shieber. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:11, 1985.

[40] R. Stiebe. On the complexity of the control language in tree controlled grammars. In *Colloquium on the Occasion of the 50th Birthday of Victor Mitrana*, 2008.

[41] S. Turaev, J. Dassow, and M. H. Selamat. Language classes generated by tree controlled grammars with bounded nonterminal complexity. In *Descriptional Complexity of Formal Systems - 13th International Workshop, DCFS 2011, Gießen/Limburg, Germany, July 25-27, 2011. Proceedings*, 2011.

[42] S. Turaev, J. Dassow, and M. H. Selamat. Nonterminal complexity of tree controlled grammars. *Theoretical Computer Science*, 412:7, 2011.

[43] A. P. J. van der Walt. Random context grammars. In *Proceedings of Symposium on Formal Languages*, 1970.

[44] B. Wegbreit. A generator of context-sensitive languages. *Journal of Computer and System Sciences*, 3:6, 1969.

[45] W. A. Woods. Context-sensitive parsing. *Communications of the ACM*, 13:9, 1970.

[46] G. Zetzsche. Toward understanding the generative capacity of erasing rules in matrix grammars. *International Journal of Foundations of Computer Science*, 22:16, 2011.

Selected Papers by the Author

M. Čermák, J. Koutný, and A. Meduna. Parsing based on n-path tree-controlled grammars. *Theoretical and Applied Informatics*, 2011:16, 2012.

J. Koutný and A. Meduna. On normal forms and erasing rules in path controlled grammars. Schedae Informaticae, in press.

J. Koutný and A. Meduna. Tree-controlled grammars with restrictions placed upon cuts and paths. *Kybernetika*, 48:11, 2012.

J. Koutný, Z. Křivka, and A. Meduna. On grammars with controlled paths. Acta Cybernetica, submitted.

J. Koutný, Z. Křivka, and A. Meduna. Pumping properties of path-restricted tree-controlled languages. In *7th Doctoral Workshop on Mathematical and Engineering Methods in Computer Science*, 2011.

J. Koutný. Syntax analysis of tree-controlled languages. In *Proceedings of the 17th Conference STUDENT EEICT 2011 Volume 3*, 2011.

J. Koutný. On n-path-controlled grammars. In *Proceedings of the 16th Conference STUDENT EEICT 2010 Volume 5*, 2010.

J. Koutný. On path-controlled grammars and pseudoknots. In *Proceedings of the 18th Conference STUDENT EEICT 2012 Volume 3*, 2012.

J. Koutný. Regular paths in derivation trees of context-free grammars. In *Proceedings of the 15th Conference STUDENT EEICT 2009 Volume 4*, 2009.