On the Commutative Equivalence of Algebraic Formal Series and Languages

Arturo Carpi
Dipartimento di Matematica e Informatica,
Università degli Studi di Perugia,
via Vanvitelli 1, 06123 Perugia, Italy.
e-mail: carpi@dmi.unipg.it

Flavio D'Alessandro
Dipartimento di Matematica,
Università di Roma "La Sapienza"
Piazzale Aldo Moro 2, 00185 Roma, Italy
e-mail: dalessan@mat.uniroma1.it
and
Department of Mathematics,
Boğaziçi University
34342 Bebek, Istanbul, Turkey

Abstract

The problem of the commutative equivalence of context-free and regular languages is studied. Conditions ensuring that a context-free language of exponential growth is commutatively equivalent with a regular language are investigated.

Keywords: Commutative equivalence, Context-free language, Unique factorization code, Exponential growth

Mathematics Subject Classification 2010: 68Q45, 68Q70, 94A45

1 Introduction

In this paper, we study the commutative equivalence of context-free and regular languages. Two words are said to be *commutatively equivalent* if one is obtained from the other by rearranging the letters of the word. Two languages L_1 and L_2 are said to be *commutatively equivalent* if there exists a bijection $f\colon L_1 \to L_2$ such that every word $u \in L_1$ is commutatively equivalent to $f(u)$. This notion plays an important role in the study of several problems of Theoretical Computer Science such as, for instance, in the Theory of Codes, where it is involved in the

celebrated Schützenberger conjecture about the commutative equivalence of a maximal finite code with a prefix one (see, e.g, [5, 25, 41] and also [15] for some related problems). The question of our interest can be formulated as follows:

Commutative Equivalence Problem *Given a context-free language L_1, does there exist a regular language L_2 which is commutatively equivalent to L_1?*

In the sequel, for short, we refer to it as *CE Problem*. A language which is commutatively equivalent to a regular one will be called *commutatively regular*.

It is worth noticing that commutatively equivalent languages share the same alphabet and their generating series are equal. In particular, every commutatively regular language must be *counting regular*, that is, its generating series is rational. This remark leads us to recall that a conceptually related study was conducted by Béal and Perrin in [3], where it is proved that a formal series $s = \sum_{n\geq 0} s_n x^n$ is the generating series of a regular language over a k-letter alphabet if and only if the series s and its complementary $t = \sum_{n\geq 0}(k^n - s_n)\, x^n$ are both $\mathbb{N}$-rational.

Recently, Ibarra, McQuillan, and Ravikumar investigated this topic in view of the more general notion of *strongly counting regularity* [35]. Such notion amounts to ask, for an arbitrary language L, that the language $L \cap L_1$ is counting-regular, for every regular language L_1. A thorough analysis of the structure and of the decidability issues of this family of languages have been obtained in [35, 36].

For our discussion, the following notions are useful. Given a language L, the *growth function* g_L returns, for any non-negative integer n, the number of the words of L whose length is less than or equal to n. A language L is called *sparse* if its growth function is polynomially upper bounded. A language L is said to be *of exponential growth* if there exists a real number $k > 1$ such that $g_L(n) > k^n$ for all sufficiently large n.

Two results are relevant in this context. The first provides a "gap property" [8, 38]: every context-free language is either sparse or of exponential growth. The second states that the class of sparse context-free languages coincides with that of bounded context-free languages [34, 40]. We recall that a language L is termed *bounded* if there exist k words $u_1, \ldots, u_k$ such that $L \subseteq u_1^* \cdots u_k^*$.

Bounded context-free languages play a meaningful role in Computer Science and in Mathematics and have been widely investigated in the past so that their structure has been characterized by several theorems [6, 14, 17, 18, 22, 24, 27, 28, 32–37, 40, 42]. A characterization of regular bounded sets, based upon a combinatorial property of the factors of the words of the language, has been obtained by Restivo [42] and, subsequently, extended to context-free languages by Boasson and Restivo [6]. A survey on the relationships between bounded languages and semigroups has been given by de Luca and Varricchio in [24].

In [19–21] the solution (in the affirmative) of the CE Problem for sparse languages is given: *Every bounded context-free language L_1 is commutatively equivalent to a regular language L_2. Moreover the language L_2 can be effectively constructed starting from an effective presentation of L_1.* It is also shown that

the CE Problem can be solved in the affermative for the wider class of bounded semi-linear languages.

In view of the latter theorem and of the results mentioned above, the CE Problem remains open for the class of context-free languages of exponential growth. It should be pointed out that the techniques forged to solve the CE Problem in the bounded case cannot be used in the exponential one. This is due to the fact that such techniques are based upon the faithful representation of bounded context-free languages by means of semi-linear sets of vectors (over $\mathbb{N}$), a result due to Ginsburg and Spanier [27, 28] that does not hold in the general case.

A remark is relevant in this context: given a commutatively regular language L, its characteristic series in commutative variables – that is, the formal series $\underline{\underline{L}}$ such that the coefficient of every word w is the number of words of L commutatively equivalent to w – is rational.

This fact has two consequences. The first is that the answer to the CE Problem is not in the affirmative in general. Indeed, the generating series of a commutatively regular language L is always rational while there exist context-free languages whose generating series are algebraic but not rational.

The most natural case corresponds to the family of Dyck languages which are even deterministic and of exponential growth. Taking into account that by the well-known representation theorem by Chomsky and Schützenberger [16] every context-free language is the image of a Dyck language by a rational transduction, these languages constitute a very general example.

Note also that, as a related result relevant in this setting, another well-known theorem by Chomsky and Schützenberger [16] states that the generating series of an unambiguous context-free language is algebraic.

It is worth noting that Flajolet even provided examples of linear inherently ambiguous context-free languages with a transcendental generating series, such as the language $L = \{a^n b v_1 a^n v_2 \mid n \geq 1,\ v_1, v_2 \in A^*\}$ over the alphabet $A = \{a, b\}$ [26, Theorem 3].

In view of the previous results, solving the CE Problem seems to be highly non trivial.

The second consequence is that the study of the CE Problem can be reduced to the family of languages whose characteristic series are rational. In this context, the class of *non-expansive* grammars seems to play a relevant role. A context-free grammar G is said to be *expansive* if one has $X \Rightarrow^* \alpha_1 X \alpha_2 X \alpha_3$ for some non-terminal X and suitable words $\alpha_1, \alpha_2, \alpha_3$. In the opposite case, G is non-expansive. Non-expansive grammars have been widely studied (see [4] and reference therein). The class of languages generated by these grammars coincides with that of context-free languages of finite index.

A remarkable result by Baron and Kuich [2] provides a characterization of non-expansive grammars. In particular, an unambiguous grammar is non-expansive if and only if all non-terminals generate languages whose characteristic series are rational.

It is worth noticing that in the quoted paper it has been conjectured that all unambiguous context-free languages whose characteristic series in commutative

variable is rational are generated by non-expansive grammars. If this conjecture had been true, then all commutatively regular unambiguous context-free languages would have been generated by non-expansive grammars. However, we disprove the conjecture by exhibiting a deterministic context-free language which is commutatively regular but cannot be generated by a non-expansive grammar (see Example 1).

On the other side, it is natural to ask whether all languages generated by non-expansive grammars or, at least, by unambiguous non-expansive grammars are commutatively regular.

In the first part of this paper, we investigate the CE Problem for languages generated by non-expansive grammars. The first result we prove, can be formulated as follows (Theorem 2): the language generated by an unambiguous and non-expansive grammar G is commutatively regular if for every non-terminal X, there exists a bijection $f\colon P_X \to \mathcal{W}_X$ between the set P_X of the productions of X in G and a prefix code $\mathcal{W}_X$ such that for any production $p\colon X \to \alpha$, the word obtained deleting all non-terminals in α is commutatively equivalent to $f(p)$.

This condition is verified, in particular, if the number of terminals occurring in the right side of each production is sufficiently large (with respect to the number of productions) and they are not all equal to the same letter (see Theorem 3).

One of the key ingredients of our technique is the notion of *commutative equivalence* of grammars. Two context free grammars G and G' are commutatively equivalent if there exists a bijection $f\colon P \to P'$ between the sets P and P' of productions of G and G', respectively, such that, for every production $p \in P$, the right side components of p and of $f(p)$ are commutatively equivalent, and the left sides are equal. One can show that the characteristic series in commutative variables of two commutatively equivalent (cycle-free) grammars are equal. As a straightforward consequence of this fact, one has that commutatively equivalent unambiguous cycle-free grammars generate commutatively equivalent languages. Therefore, in order to show that a context-free language is commutatively regular, it is enough to find an unambiguous grammar generating the language which is commutatively equivalent to an unambiguous right linear one. These facts together with the use of codes allow us to develop a technique to deal with the problem.

This method also allows to get an alternative proof of the 'if' part of the theorem of Baron and Kuich. In our opinion, this proof could be of interest in itself since it furnishes a method for the construction, starting from a non-expansive grammar, of a right linear grammar with the same characteristic series in commutative variables. Thus the CE Problem for unambiguous non-expansive grammars is related to the general problem of finding a regular language with a prescribed characteristic series in commutative variables.

In the second part of the paper, we investigate the CE Problem with respect to the first non-trivial family of non-expansive grammars: the *minimal linear grammars*. A linear grammar is called minimal if it has only one non-terminal symbol. This notion, first introduced in [16] (see also [30]), is relevant in our

study since, in the unambiguous case, the derivation process of words in such a grammar, is algebraically close to the process of message encoding by variable length codes [12].

We first prove that the language generated by an unambiguous minimal linear grammar G is commutatively regular if the language of words generated by G in k steps, for some given $k \geq 1$, is a commutatively prefix set (Theorem 4 and Corollary 3). This result shows a connection between the CE Problem for unambiguous minimal linear grammars and the study of conditions that guarantee for a finite set of words to be commutatively equivalent to a code.

In view of this problem, it becomes natural to study the property of unambiguity of these grammars. By using the notion of Bernoulli distribution, we prove two results for an unambiguous minimal linear grammar which are analogous to fundamental properties of codes. The first is a "Kraft-McMillan like" inequality: in an arbitrary unambiguous minimal linear grammar, for every Bernoulli distribution μ, one has $\sum \mu(uv) \leq 1$, where the sum is extended to all productions $X \to uXv$ of G (Proposition 6). The second result states, up to a technical restriction, the very same characterization of codes in term of positive Bernoulli distributions (Proposition 7 and Corollary 4). We finally refine our results for minimal linear grammars on a binary alphabet of terminal symbols, showing a relation with the Schützenberger conjecture of codes mentioned above.

In conclusion, we point out that the general question whether all languages generated by non-expansive grammars are commutatively regular remains open, and we hope that the techniques developed here will help to find an answer.

The paper is organized as follows. In Section 2 preliminaries on context-free languages and formal power series are presented. In Section 3, we introduce and discuss the notion of commutative equivalence for grammars. Section 4 is devoted to the study of the characteristic series in commutative variables of languages generated by non-expansive grammars and the theorem by Baron and Kuich. Section 5 is devoted to the proof of our statements on the commutative equivalence of languages generated by non-expansive grammars and regular languages (Theorem 2 and Theorem 3). In Section 6, we investigate the CE Problem for the class of linear minimal grammars. Section 7 contains some concluding remarks and open problems.

Some results of this paper have been presented at DLT 2018 [10].

2 Preliminaries

We now recall some useful terms and basic properties concerning formal languages, context-free grammars, and formal power series [4, 5, 9, 27, 31, 44].

2.1 Words and languages

Let A be a finite non-empty alphabet and A^* be the free monoid generated by A. The identity of A^* is called the *empty word* and is denoted by ϵ. The set $A^* \setminus \{\epsilon\}$ is denoted by A^+. The *length* of a word $w \in A^*$ is the integer $|w|$

inductively defined by $|\epsilon| = 0$, $|wa| = |w| + 1$, $w \in A^*$, $a \in A$. If $n \in \mathbb{N}$, then $A^{\leq n}$ denotes the set of all the words of A^* of length not larger than n. For every $a \in A$, $|w|_a$ denotes the number of occurrences of the letter a in w.

One can introduce in A^* the equivalence relation $\sim$, called *commutative equivalence*, defined as follows: for all $u, v \in A^*$, one has $u \sim v$ if $|u|_a = |v|_a$ for every letter $a \in A$. Thus, one has $u \sim v$ if the word v is obtained rearranging the letters of u in a different order. Two languages L and L' are said to be *commutatively equivalent*, and one writes $L \sim L'$, if there exists a bijection $f\colon L \to L'$ such that, for every $u \in L,\ u \sim f(u)$.

A set X over the alphabet A is said to be a *prefix set* if $XA^+ \cap X = \emptyset$, that is, if, for every $u, v \in L$, u is not a proper prefix of v. A set X of words over an alphabet A is said to be *commutatively prefix* if there exists a prefix set X' such that X is commutively equivalent to X'.

A subset X of A^+ is a *code (over A)* if every word of X^+ has a unique factorization as a product of words of X.

Let B and A be alphabets with $B \subseteq A$. The *projection of A^* onto B^** is the morphism $\widehat{\pi}_B\colon A^* \to B^*$ generated by the function $\pi_B\colon A \to B \cup \{\varepsilon\}$ such that, for every $a \in A,\ \pi_B(a) = a$, if $a \in B$, and $\pi_B(a) = \varepsilon$, otherwise. In the sequel, the morphism $\widehat{\pi}_B$ will be simply denoted π_B.

Let A and B be two alphabets and let $\mathcal{P}(B^*)$ denote the power set of B^*. A map $\phi\colon A^* \to \mathcal{P}(B^*)$ is a *substitution* if for any $u, v \in A^*$ one has $\phi(uv) = \phi(u)\phi(v)$. The substitution ϕ is *regular* if for all $a \in A$, $\phi(a)$ is a regular language. As is well-known (see, e.g., [44]), if $L \subseteq A^*$ is a regular language and ϕ is a regular substitution, then $\phi(L) = \bigcup_{w \in L} \phi(w)$ is a regular language.

2.2 Formal series

We assume the reader to be familiar with the theory of formal power series. Just in order to facilitate the lecture of the paper, we recall here some notions. A comprehensive presentation of the subject can be found, for instance, in [45].

Let A be an alphabet and $\widehat{\mathbb{N}}$ be the semiring $\widehat{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$. The semiring of formal power series in non-commutative and commutative variables with coefficients in $\widehat{\mathbb{N}}$ and variables in A will be denoted, respectively, by $\widehat{\mathbb{N}}\langle\langle A \rangle\rangle$ and $\widehat{\mathbb{N}}[[A]]$. A formal power series with coefficients in $\widehat{\mathbb{N}}$ is said to be *unambiguous* (resp., *non-singular*) if all its coefficients belong to the set $\{0, 1\}$ (resp., to $\mathbb{N}$). As usual, the sub-semirings of non-singular series in non-commutative and commutative variables will be denoted, respectively, by $\mathbb{N}\langle\langle A \rangle\rangle$ and $\mathbb{N}[[A]]$ and the sub-semirings of non-singular polynomials by $\mathbb{N}\langle A \rangle$ and $\mathbb{N}[A]$. The coefficient of a monomial w in the series s is denoted by (s, w).

Let A and B be two alphabets. Any non-erasing morphism $\phi\colon A^* \to B^*$ can be extended to a semiring morphism of $\widehat{\mathbb{N}}\langle\langle A \rangle\rangle$ into $\widehat{\mathbb{N}}\langle\langle B \rangle\rangle$ by

$$\phi\Big(\sum_{w \in A^*} k_w w\Big) = \sum_{w \in A^*} k_w \phi(w),$$

$k_w \in \mathbb{N},\ w \in A^*$. In a similar way, the natural projection $c_A\colon A^* \to A^*/\sim$

is extended to a morphism of $\widehat{\mathbb{N}}\langle\langle A\rangle\rangle$ onto $\widehat{\mathbb{N}}[[A]]$. Moreover, there exists a semiring morphism $\widetilde{\phi}\colon \widehat{\mathbb{N}}[[A]] \to \widehat{\mathbb{N}}[[B]]$ such that the following diagram, where c_A and c_B denote the natural projections, commutes:

$$\begin{array}{ccc} \widehat{\mathbb{N}}\langle\langle A\rangle\rangle & \xrightarrow{\phi} & \widehat{\mathbb{N}}\langle\langle B\rangle\rangle \\ {\scriptstyle c_A}\downarrow & & \downarrow{\scriptstyle c_B} \\ \widehat{\mathbb{N}}[[A]] & \xrightarrow[\widetilde{\phi}]{} & \widehat{\mathbb{N}}[[B]] \end{array}$$

For any element s of $\widehat{\mathbb{N}}\langle\langle A\rangle\rangle$ or of $\widehat{\mathbb{N}}[[A]]$, we denote by s^* the formal power series whose coefficients, for any monomial w, are given by $(s, w) = \sum_{n=0}^{\infty}(s^n, w)$. A formal power series is *rational* if it belongs to the minimal subsemiring of $\widehat{\mathbb{N}}\langle\langle A\rangle\rangle$ (resp., $\widehat{\mathbb{N}}[[A]]$) containing all monomials and closed for the $*$-operation.

With any language L on an alphabet A, we associate the *characteristic series of* L in non-commutative variables $\underline{L} = \sum_{w\in L} w$. The natural projection of $\underline{L}$ in the commutative semiring $\widehat{\mathbb{N}}[[A]]$ will be called the *characteristic series of* L in commutative variables and will be denoted by $\underline{\underline{L}}$. Thus, for any monomial $a_1^{n_1}\cdots a_t^{n_t}$, $(\underline{\underline{L}}, a_1^{n_1}\cdots a_t^{n_t})$ gives the number of the words of L which are commutatively equivalent to $a_1^{n_1}\cdots a_t^{n_t}$.

2.3 Context-free grammars

Let $G = (V, T, P, S)$ be a context-free grammar, where V denotes the vocabulary, T denotes the set of terminals, $N = V \setminus T$ denotes the set of non-terminals, P denotes the set of productions, and $S \in V$ denotes the axiom. For every $\alpha, \beta \in V^*$, we write $\alpha \Rightarrow_G \beta$ if α *directly derives* β in G. As usual, the transitive (resp., transitive and reflexive) closure of the relation $\Rightarrow_G$ will be denoted by $\Rightarrow_G^+$ (resp., $\Rightarrow_G^*$). If no ambiguity arises $\Rightarrow_G$ (resp., $\Rightarrow_G^+$, $\Rightarrow_G^*$) is simply denoted $\Rightarrow$ (resp., $\Rightarrow^+$, $\Rightarrow^*$).

Let $\delta = p_1 p_2 \cdots p_k$ be a finite sequence of productions of G and

$$\alpha_0 \Rightarrow \alpha_1 \Rightarrow \cdots \Rightarrow \alpha_k$$

be a derivation where any α_i is obtained from α_{i-1} replacing an occurrence of the left side of p_i by the corresponding right side, $1 \le i \le k$. In such a case we write $\alpha_0 \Rightarrow_\delta \alpha_k$. Moreover, the integer k is said to be the *length* of the derivation. For any non-terminal X and any $\alpha \in V^*$, we write $X \Rightarrow^k \alpha$ if there exists a derivation $X \Rightarrow_\delta \alpha$ of length k.

We denote by $L(G)$ the language $\{u \in T^* \mid S \Rightarrow_G^* u\}$ of all the words of T^* generated by G. A grammar G is said to be *unambiguous* if every $u \in L(G)$ is generated by exactly one leftmost derivation; otherwise G is said to be *ambiguous*.

With any non-terminal X of the grammar G one can associate the series $\underline{G}_X$ of $\widehat{\mathbb{N}}\langle\langle T\rangle\rangle$, whose coefficients $\left(\underline{G}_X, w\right)$ count the number of leftmost derivations $X \Rightarrow^* w$. The natural projection of $\underline{G}_X$ in the commutative semiring $\widehat{\mathbb{N}}[[T]]$

will be denoted by $\underline{\underline{G}}_X$. Thus, the coefficients $\left(\underline{\underline{G}}_X, w\right)$ count the number of leftmost derivations of words which are commutatively equivalent to w. The series $\underline{\underline{G}} = \underline{\underline{G}}_S$ is called the *characteristic series* (in commutative variables) of the grammar G. In particular, if G is unambiguous, then $\underline{\underline{G}}$ is the characteristic series (in commutative variables) of the language generated by G.

A context-free grammar G is said to be *cycle-free* if there is no non-terminal X such that $X \Rightarrow^+ X$. This condition ensures that any word $w \in L(G)$ has finitely many leftmost derivations in G, so that the series $\underline{G}$ and $\underline{\underline{G}}$ are non-singular.

A context-free grammar is said to be *reduced* if, for any non-terminal X, there are words $u_1, u_2, u_3 \in T^*$ such that $S \Rightarrow^* u_1 X u_2 \Rightarrow^* u_3$.

3 Commutatively Equivalent Grammars

We say that two productions of context-free grammars are *commutatively equivalent* if their left sides are equal and their right sides are commutatively equivalent over the alphabet V. Two context free grammars $G = (V, N, P, S)$ and $G' = (V, N, P', S)$ are *commutatively equivalent* if there exists a bijection $f\colon P \to P'$ such that every production $p \in P$ is commutatively equivalent to $f(p)$.

We will establish the following

Proposition 1 *The characteristic series in commutative variables of two commutatively equivalent cycle-free grammars are equal.*

As a straightforward consequence of Proposition 1, one has the following

Corollary 1 *Commutatively equivalent unambiguous (cycle-free) grammars generate commutatively equivalent languages.*

Remark 1 In the previous corollary the hypothesis that the grammars are cycle-free may be eliminated. Indeed, if G and G' are two commutatively equivalent unambiguous grammars, then the corresponding reduced grammars are commutatively equivalent and unambiguous, too. Since unambiguous reduced grammars are necessarily cycle-free, by the previous corollary we conclude that the generated languages are commutatively equivalent.

We mention that also Proposition 1 remains true without the hypothesis that the considered grammars are cycle-free. However, the proof of the general case would require some very complex combinatorial argument and is omitted, as it would be outside the scope of the paper.

In order to give a simple proof of Proposition 1, we recall some basic properties of algebraic systems of equations. For a comprehensive presentation of the subject the reader is referred to [45].

Let T be an alphabet. An *algebraic system of equations* on $\mathbb{N}\langle T\rangle$ in the unknowns $X_1, X_2, \ldots, X_k$ is a set of equations

$$X_i = \alpha_i, \quad 1 \leq i \leq k, \tag{1}$$

with $\alpha_i \in \mathbb{N}\langle T \cup \{X_1, X_2, \ldots, X_k\}\rangle$, $1 \leq i \leq k$.

The *approximating solutions* of the algebraic system (1) are the k-tuples $(X_1^{(n)}, X_2^{(n)}, \ldots, X_k^{(n)})$ of elements of $\mathbb{N}\langle T\rangle$ defined as follows:

$$X_1^{(0)} = X_2^{(0)} = \cdots = X_k^{(0)} = 0,$$

and for $n > 0$, $X_i^{(n)}$ is the polynomial obtained from α_i, replacing all occurrences of the unknowns X_j by $X_j^{(n-1)}$, $1 \leq i, j \leq k$. A k-tuple $(s_1, s_2, \ldots, s_k)$ of elements of $\mathbb{N}\langle\langle T\rangle\rangle$ is the *proper solution* of the algebraic system (1) if for all $w \in T^*$ there exists an integer n_w such that

$$(s_i, w) = (X_i^{(n)}, w), \quad 1 \leq i \leq k,\ n \geq n_w.$$

As is well-known, the proper solution of an algebraic system, if existing, is a solution of the system, that is replacing in α_i all occurrences of the unknowns $X_1, \ldots, X_k$ by $s_1, \ldots, s_k$, respectively, one obtains the series s_i, $1 \leq i \leq k$.

Algebraic systems of equations on $\mathbb{N}[T]$ and their approximated and proper solutions are defined similarly. Let c denote the natural projection

$$c\colon \mathbb{N}\langle\langle T \cup \{X_1, X_2, \ldots, X_k\}\rangle\rangle \to \mathbb{N}[[T \cup \{X_1, X_2, \ldots, X_k\}]].$$

By induction on n, one easily verifies that if $(X_1^{(n)}, \ldots, X_k^{(n)})$ is n-th approximating solution of the algebraic system (1), then $(c(X_1^{(n)}), \ldots, c(X_k^{(n)}))$ is the n-th approximating solution of the algebraic system

$$X_i = c(\alpha_i), \quad 1 \leq i \leq k, \tag{2}$$

on $\mathbb{N}[T]$. Consequently, if $(s_1, \ldots, s_k)$ is a proper solution of the algebraic system (1), then $(c(s_1), \ldots, c(s_k))$ is a proper solution of the algebraic system (2). In the sequel, we shall refer to the system (2) as the *commutative variant* of the algebraic system (1).

Let $G = (V, T, P, X_1)$ be a context-free grammar with non-terminal symbols $X_1, X_2, \ldots, X_k$ and productions

$$X_i \to \alpha_{ij}, \quad 1 \leq i \leq k,\ 1 \leq j \leq m_i.$$

With the grammar G, we associate the algebraic system

$$X_i = \sum_{j=1}^{m_i} \alpha_{ij}, \quad 1 \leq i \leq k,$$

in the unknowns $X_1, X_2, \ldots, X_k$.

As proved in [39], if the grammar G is cycle-free, then the algebraic system has a proper solution. Such a solution is given by the tuple $(\underline{G}_{X_1}, \underline{G}_{X_2}, \ldots, \underline{G}_{X_k})$. Consequently, the commutative variant of the algebraic system has the proper solution $(\underline{\underline{G}}_{X_1}, \underline{\underline{G}}_{X_2}, \ldots, \underline{\underline{G}}_{X_k})$.

Now, the proof of Proposition 1 is straightforward: the commutative variants of the algebraic systems associated with two commutatively equivalent cycle-free grammars G and G' are equal, so that they have the same proper solution. Thus, for all variables X_i, $\underline{\underline{G}}_{X_i} = \underline{\underline{G'}}_{X_i}$. In particular, $\underline{\underline{G}} = \underline{\underline{G'}}$.

4 Non-expansive Grammars and Rational Series

In the sequel, we consider a context-free grammar $G = (V, T, P, S)$. As already observed, the characteristic series (in commutative variables) and the generating series of a commutatively regular language must be rational. Rational series are well-known and their structure has been thoroughly investigated. A result of Baron and Kuich [2] provides the characterization of the context-free grammars G such that, for every non-terminal X, the series $\underline{\underline{G}}_X$ is rational. This characterization is based upon the notion of *non-expansive grammar*. A grammar G is said to be *expansive* if there is a non-terminal X such that $X \Rightarrow^* \alpha_1 X \alpha_2 X \alpha_3$ for some $\alpha_1, \alpha_2, \alpha_3 \in V^*$. In the opposite case, G is non-expansive.

Theorem 1 [2] *A cycle-free reduced context-free grammar is non-expansive if and only if for all non-terminals X, the series $\underline{\underline{G}}_X$ is rational.*

The following is a straightforward consequence of the theorem above.

Corollary 2 *The characteristic series in commutative variables of the language generated by an unambiguous non-expansive grammar is rational.*

Remark 2 Theorem 1 does not imply that an unambiguous context-free language whose characteristic series in commutative variables is rational, is generated by a non-expansive grammar. In fact, in [2] the following two conjectures were formulated:

1. a (reduced) context-free grammar G such that $\underline{\underline{G}}$ is rational is non-expansive;
2. an unambiguous context-free grammar G is non-expansive if and only if $\underline{\underline{L(G)}}$ is rational.

However, the following example shows that both conjectures are false.

Example 1 Let D_1 be the Dyck language on the alphabet $A = \{a, \overline{a}\}$ and $\overline{D}_1 = A^* \setminus D_1$. Then

$$L = bD_1 \cup \overline{D}_1 b$$

is a deterministic context-free language. It is commutatively equivalent to the rational language A^*b and therefore its characteristic series in commutative variables is rational.

Let us verify that a context-free grammar G generating L is necessarily expansive.

We can assume without loss of generality that the grammar G has no ϵ-rules and no unit rules. Indeed, as is well-known, one can effectively construct a context-free grammar $G_1 = (V, T, P_1, S)$ with no ϵ-rule and no unit rule such that $L(G_1) = L(G)$ and, moreover, for any derivation $X \Rightarrow^*_{G_1} \alpha$, one has $X \Rightarrow^*_G \alpha$. Hence, if G_1 is expansive, then G is expansive, too.

Thus, let us assume that G has no ϵ-rule and no unit rule. Let us verify that if $S \to \alpha$ is a production of G such that $\alpha \Rightarrow^* vb$ for some $v \in \overline{D}_1$, then such a

production cannot occur in a derivation of any word bu, with $u \in D_1$. Indeed, if it was not the case, one would have

$$S \stackrel{*}{\Rightarrow} \beta_1\alpha\beta_2 \stackrel{*}{\Rightarrow} bu,$$

for some $\beta_1, \beta_2 \in V^*$. Since G has no unit rule, one has $\alpha = X\alpha_1$, with $X \in V$ and $\alpha_1 \in V^+$. Taking into account that $X\alpha_1 \Rightarrow^* vb$, $\beta_1 X\alpha_1\beta_2 \Rightarrow^* bu$ and G has no ϵ-rule, one derives

$$v = v_1v_2, \quad X \stackrel{*}{\Rightarrow} v_1, \quad \alpha_1 \stackrel{*}{\Rightarrow} v_2b,$$

$$u = u_1u_2u_3, \quad \beta_1X \stackrel{*}{\Rightarrow} bu_1, \quad \alpha_1 \stackrel{*}{\Rightarrow} u_2, \quad \beta_2 \stackrel{*}{\Rightarrow} u_3,$$

with $u_1, u_2, u_3, v_1, v_2 \in A^*$. Consequently,

$$S \Rightarrow X\alpha_1 \stackrel{*}{\Rightarrow} v_1u_2 \in A^*.$$

This yields a contradiction, since the letter b does not occur in the word v_1u_2 and, therefore, $v_1u_2 \notin L$.

Let G_2 be the grammar obtained by G by deleting all productions $S \to \alpha$ such that $\alpha \Rightarrow^* vb$ for some $v \in \overline{D}_1$. By the result established above, one easily verifies that $L(G_2) = bD_1$. Now, let G_3 be the grammar obtained by G_2 by deleting all occurrences of b in the right sides of the productions. Clearly, one has $L(G_3) = D_1$.

As is well-known [43], the language D_1 cannot be generated by a non-expansive grammar. Thus, in G_3 there is a nonterminal X and a derivation $X \Rightarrow^* \alpha_1X\alpha_2X\alpha_3$, with $\alpha_1, \alpha_2, \alpha_3 \in V^*$. By the construction of G_3, one derives that there is a derivation $X \Rightarrow^* \beta_1X\beta_2X\beta_3$, with $\beta_1, \beta_2, \beta_3 \in V^*$ in G_2 and, consequently, in G. Hence, also G is expansive.

By the way, this example is derived from a very similar one in [2]. It is also worth noticing that the class of languages generated by non-expansive grammars coincides with that of context-free languages of finite index (see, e.g., [4, sec. VII.5]).

Clearly, the characteristic series in commutative variables of a commutatively regular language L is rational. For this reason, in view of Theorem 1, the study of the CE Problem for languages generated by non-expansive grammars is of particular interest.

Indeed, if a language L is generated by an unambiguous non-expansive grammar, then its characteristic series $\underline{\underline{L}}$ is a rational series and therefore it is a component of the proper solution of a proper linear system in commutative variables $\mathcal{S}$ (see, e.g., [45]). Thus, in order to investigate the commutative regularity of the language L, one is reduced to ask whether the components of the proper solution of $\mathcal{S}$ are the characteristic series in commutative variables of regular languages. In other terms, one is reduced to search for a proper linear system in non-commutative variables whose commutative version is equal to $\mathcal{S}$ or to a system equivalent to it and with a proper solution whose components are unambiguous series. This problem will be studied in the following section.

5 The CE Problem for Non-expansive Grammars

Let $G = (V, T, P, S)$ be a context-free grammar. One may consider the relation $\leq$ on the set N of non-terminal symbols of G defined as follows: for any $X, Y \in N$, one has $X \leq Y$ if there is a derivation $X \Rightarrow^* \alpha_1 Y \alpha_2$ in G with $\alpha_1, \alpha_2 \in V^*$. As one easily verifies, the relation $\leq$ is a quasi-order on N. As usual, if X, Y are non-terminals such that $X \leq Y$ and $Y \leq X$, then we shall write $X \equiv Y$, while if one has $X \leq Y$ but $Y \leq X$ does not hold true, then we shall write $X < Y$. The relations $<$ and $\equiv$ are respectively a partial order and an equivalence on the set N of non-terminals. With any non-terminal $X \in N$, we associate the sets

$$N_{\equiv X} = \{Y \in N \mid Y \equiv X\} \quad \text{and} \quad N_{<X} = \{Y \in N \mid Y < X\}$$

As proved in [9], a grammar G is non-expansive if and only if all its productions have the form

$$X \to uYv \quad \text{or} \quad X \to u,$$

with $X \in N$, $Y \in N_{\equiv X}$, $u, v \in (T \cup N_{<X})^*$. Grammars satisfying the condition above were called *superlinear* in [9]. We say that a context-free grammar G is *right superlinear* if all its productions have the form

$$X \to uY \quad \text{or} \quad X \to u,$$

with $X \in N$, $Y \in N_{\equiv X}$, $u \in (T \cup N_{<X})^*$. From the result of [9] quoted above one derives

Proposition 2 *Any right superlinear grammar is non-expansive.*

Now we shall prove that right superlinear grammars generate regular languages.

Proposition 3 *Let $G = (V, T, P, S)$ be a right superlinear grammar. Then $L(G)$ is a regular language.*

Proof We proceed by induction on $\mathrm{Card}(V)$. Clearly, if $\mathrm{Card}(V) = 1$ and, more generally, if $N_{<S} = \emptyset$, then G is a right linear grammar and the statement is trivially verified. Thus, we assume $N_{<S} \neq \emptyset$.

Let $G' = (V, T', P', S)$ be the grammar with

$$T' = T \cup N_{<S}, \quad P' = \{X \to \alpha \in P \mid X \in N_{\equiv S}\}.$$

and $\phi\colon T'^* \to T^*$ be the substitution defined by

$$\phi(\alpha) = \{w \in T^* \mid \alpha \stackrel{*}{\underset{G}{\Rightarrow}} w\},$$

for all $\alpha \in T'^*$. Since G is right superlinear, the grammar G' is right linear and therefore it generates a regular language $R = L(G') \subseteq T'^*$. Let us verify that

$$L(G) = \phi(R).$$

Indeed, let $w \in L(G)$, $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \Rightarrow \alpha_n = w$ be a rightmost derivation of w in G and i be the maximal index such that $S \Rightarrow^* \alpha_i$ in G'. Since G' is right linear, either $\alpha_i \in R$ or α_i has the form $\alpha_i = uX$, with $u \in T'^*$ and $X \in N_{\equiv S}$. However, in the latter case, α_{i+1} would be a direct consequence of α_i in the grammar G', contradicting the maximality of i. We conclude that $\alpha_i \in R$ and $w \in \phi(R)$. This proves the inclusion $L(G) \subseteq \phi(R)$.

Conversely, if $w \in \phi(R)$, then one has $w \in \phi(\alpha)$ for some $\alpha \in R$. Thus, $S \Rightarrow^* \alpha$ in G' and $\alpha \Rightarrow^* w$ in G. Taking into account that all productions of G' are also productions of G, one derives $S \Rightarrow^* w$ in G, so that $w \in L(G)$. This proves the inclusion $\phi(R) \subseteq L(G)$.

Now we show that ϕ is a regular substitution. Indeed, one easily verifies that for any $X \in N_{<S}$, $\phi(X)$ is the language generated by the grammar $G_X = (T', T, P'', X)$ with

$$P'' = \{Y \to \alpha \in P \mid Y \in N_{<S}\}.$$

Since such a grammar is right superlinear and $\mathrm{Card}(T') < \mathrm{Card}(V)$, we may assume, by the induction hypothesis, that $\phi(X)$ is a regular language. Moreover, for any $a \in T$, $\phi(a) = \{a\}$ is regular, as well. Thus, ϕ is a regular substitution.

In conclusion, we have proved that $L(G) = \phi(R)$, where R is a regular language and ϕ is a regular substitution. Hence $L(G)$ is regular. □

The following proposition is a consequence of Corollary 1 and Proposition 3.

Proposition 4 *An unambiguous grammar which is commutatively equivalent to a right superlinear unambiguous grammar generates a commutatively regular language.*

Now, we will show some applications of the previous proposition. Let $\mathcal{X} = (x_1, \cdots, x_m)$ be a list of words (non necessarily distinct). We will say that $\mathcal{X}$ is *commutatively prefix* if there exist pairwise distinct words $y_1, \cdots, y_m$ such that $x_i \sim y_i$, $1 \leq i \leq m$ and $\{y_1, \ldots, y_m\}$ is a prefix set.

Let G be a context-free grammar. For every non-terminal X, let $\alpha_1, \ldots, \alpha_m$ be the list of the right sides of the productions of G with X on the left side. We may consider the sequence of the projections of these words on the terminal alphabet

$$\mathcal{T}_X = \big(\pi_T(\alpha_1), \ldots, \pi_T(\alpha_m)\big).$$

Notice that this list may contain repetions.

Theorem 2 *Let $G = (V, T, P, S)$ be a non-expansive unambiguous grammar. If for every non-terminal X, the list $\mathcal{T}_X$ is commutatively prefix, then $L(G)$ is commutatively regular.*

Proof In view of Proposition 4, it is sufficient to verify that G is commutatively equivalent to a right superlinear unambiguous grammar.

With no loss of generality we may assume that G has the non-terminal symbols $X_1, X_2, \ldots, X_k$, where $S = X_1$, and the productions

$$X_i \to \alpha_{ij}, \quad 1 \leq i \leq k,\ 1 \leq j \leq m_i.$$

By hypothesis, for every $i = 1, \ldots, k$, there are pairwise distinct words $y_{i1}, \ldots, y_{im_i}$ such that $\pi_T(\alpha_{ij}) \sim y_{ij}$, $1 \leq j \leq m_i$ and $\mathcal{Y}_i = \{y_{i1}, \ldots, y_{im_i}\}$ is a prefix set.

Let $G' = (V, T, P', S)$ be the grammar with the following productions:

$$\begin{array}{ll} X_i \to y_{ij}\pi_{N\setminus\{Y\}}(\alpha_{ij})Y & \text{if } \alpha_{ij} \text{ contains a non-terminal } Y \equiv X_i, \\ X_i \to y_{ij}\pi_N(\alpha_{ij}) & \text{otherwise,} \end{array}$$

$1 \leq i \leq k$, $1 \leq j \leq m_i$. One easily verifies that G and G' are commutatively equivalent. Let us verify that G' is unambiguous.

Indeed, suppose that G' is ambiguous. Then there are $\alpha, \beta_1, \beta_2 \in V^*$, $w \in T^*$ and leftmost derivations in G'

$$\alpha \Rightarrow \beta_1 \overset{*}{\Rightarrow} w, \quad \alpha \Rightarrow \beta_2 \overset{*}{\Rightarrow} w, \quad \beta_1 \neq \beta_2.$$

Taking into account the form of the productions of G', one easily obtains

$$\alpha = uX_i\alpha', \quad \beta_1 = uy_{ij_1}\gamma_1\alpha', \quad \beta_2 = uy_{ij_2}\gamma_2\alpha', \quad w = uy_{ij_1}w_1 = uy_{ij_2}w_2,$$

with $u, w_1, w_2 \in T^*$, $\gamma_1, \gamma_2, \alpha' \in V^*$, $1 \leq i \leq k$, $1 \leq j_1, j_2 \leq m_i$ and $j_1 \neq j_2$. This implies $y_{ij_1}w_1 = y_{ij_2}w_2$ which is impossible as $\mathcal{Y}_i$ is a prefix set.

We conclude that G' is an unambiguous right superlinear grammar and, consequently, $L(G)$ is commutatively regular. □

In order to prove our next theorem, we need the following

Lemma 1 *Let $\mathcal{M} = (v_1, \ldots, v_m)$ be a list of words of T^+ such that:*

1. *for $i = 1, \ldots, m$, $|v_i| \geq m$;*
2. *for every $a \in T$, there exists at most one word $v_i \in a^+$.*

Then $\mathcal{M}$ is commutatively prefix.

Proof We proceed by induction on m.

If $m = 1$, the statement is trivially true. Thus, we assume $m \geq 2$. With no loss of generality, we suppose that $|v_m| = \max_{1 \leq i \leq m} |v_i|$. By the inductive hypothesis, $(v_1, \ldots, v_{m-1})$ is commutatively prefix, so that there exists a prefix code $\mathcal{Y} = \{y_1, \ldots, y_{m-1}\}$ such that $y_i \sim v_i$, $i = 1, \ldots, m-1$. To prove the statement, it is sufficient to find a word y_m such that $y_m \sim v_m$ and no word of $\mathcal{Y}$ is a prefix of y_m.

Suppose that $v_m = a^n$ for some $a \in T$, $n \geq m$. By Condition (2), no other word of $\mathcal{M}$, and, consequently, no word of $\mathcal{Y}$ is a power of a. Thus, it is sufficient to take $y_m = v_m$.

Now, let us consider the case that v_m is not the power of a single letter. Then, we can find a factor u of v_m of length m containing at least two distinct letters. The number of the words which are commutatively equivalent to u is not smaller than m. Thus, among these words, at least one is different from all the prefixes of length m of the words of $\mathcal{Y}$. Let v be such a word. Since u is a factor of v_n and $u \sim v$, one has $v_m \sim us \sim vs$ for some $s \in T^*$ and no word of $\mathcal{Y}$ can be a prefix of vs, since otherwise, v would be a prefix of such a word. Thus, our goal is attained taking $y_m = vs$. □

As a straightforward consequence of the last two results, we get

Theorem 3 *Let G be a non-expansive unambiguous grammar. For every non-terminal symbol X, let m_X denote the number of the productions in P with X on the left side. Suppose that the following conditions are verified:*

1. *For all production $X \to \alpha$, one has $|\pi_T(\alpha)| \geq m_X$.*
2. *For all $X \in N$ and $a \in T$, there exists at most one production $X \to \alpha$ such that $\pi_T(\alpha) \in a^*$.*

Then $L(G)$ is commutatively regular.

Indeed, Lemma 1 ensures that if Conditions 1 and 2 are satisfied, then the hypotheses of Theorem 2 are verified.

The following example shows that in some cases, by conveniently manipulating the productions of an unambiguous grammar, one can reduce himself to a grammar generating the same language which satisfy the hypotheses of Theorems 2 or 3 .

Example 2 Let $G = (V, T, P, S)$ be the grammar associated with the algebraic system

$$\begin{aligned} S &= XY \\ X &= aXb + bXa + aa \\ Y &= bbYa + bba \end{aligned}$$

One easily verifies that G is unambiguous and non-expansive. Here, we cannot apply Theorem 2, because the list $\mathcal{T}_S$ is given by (ϵ, ϵ) which is not commutatively prefix. Replacing in the first equation of the system all occurrences of X and Y by the expression in the right hand side of the second and third equation, respectively, we obtain the equivalent system

$$\begin{aligned} S &= aXbbbYa + aXbbba + bXabbYa + bXabba + aabbYa + aabba \\ X &= aXb + bXa + aa \\ Y &= bbYa + bba \end{aligned}$$

In the corresponding grammar $G' = (V, T, P', S)$, the list $\mathcal{T}_S$ is given by ($abbba$, $abbba, babba, babba, aabba, aabba$), whose elements are commutatively equivalent, respectively, to the words

$$a^2b^3, abab^2, ab^2ab, ab^3a, a^3b^2, a^2b^2a,$$

which are distinct elements of a prefix set. Thus, the list $\mathcal{T}_S$ is commutatively prefix. Also the lists $\mathcal{T}_X = (ab, ba, aa)$ and $\mathcal{T}_Y = (bba, bba)$ are commutatively prefix. By Theorem 2, we conclude that the language generated by G and G' is commutatively prefix.

Theorem 2 may also be used to verify that the characteristic series in commutative variables of non-expansive cycle-free grammars are rational [2].

Proposition 5 *Let $G = (V, T, P, S)$ be a non-expansive cycle-free grammar. Then the series $\underline{\underline{G}}$ is rational.*

PROOF Let $P = \{p_1, \ldots, p_n\}$. The *annotated parenthesized grammar* associated with G is the grammar $G' = (V, T', P', S)$ obtained by replacing any production $p_i \colon X \to \alpha$ by $p'_i \colon X \to ({}_i\alpha)_i$, $1 \leq i \leq n$, where $({}_1,)_1, \ldots, ({}_n,)_n$ are new terminal symbols. As is well-known, the grammar G' is unambiguous and there is a 1-1 correspondence between the words of $L(G')$ and the leftmost derivations of G, where any word $w \in L(G')$ corresponds to a leftmost derivation of $\pi_T(w)$. It follows that $\underline{G} = \pi_T(\underline{G'})$ and, consequently,

$$\underline{\underline{G}} = \widetilde{\pi}_T(\underline{\underline{G'}}).$$

where $\widetilde{\pi}_T \colon \widehat{\mathbb{N}}[[T']] \to \widehat{\mathbb{N}}[[T]]$ is obtained from π_T as explained in Section 2.2.

Taking into account that any ${}_i$ occurs in the right side of a single production of G', one easily verifies that G' satisfies the hypotheses of Theorem 2. Thus, $\underline{\underline{G'}}$ and, consequently, $\underline{\underline{G}}$ are rational series. This completes the proof. □

6 Minimal Linear Grammars

Minimal linear grammars, first introduced by Chomsky and Schützenberger in [16], provide the first non-trivial example of grammars for which the CE Problem can be investigated. A *minimal linear grammar* is a linear grammar with only one non-terminal symbol X. Thus, the productions of a minimal linear grammar can be written as

$$X \to u_i X v_i, \ 1 \leq i \leq m, \qquad X \to w_j, \ 1 \leq j \leq n, \tag{3}$$

with $u_i, v_i, w_j \in T^*$, $1 \leq i \leq m$, $1 \leq j \leq n$. The productions $X \to w_j$ will be called *terminal*.

The derivation process of words in an unambiguous minimal linear grammar is algebraically close to the process of message encoding by variable length codes [12]. Indeed, let G be a minimal grammar with a unique terminal production p_T and $f \colon A \to P'$ be a bijection of an alphabet A onto the set $P' = P \setminus \{p_T\}$ of non-terminal productions of G. Such a map naturally defines a correspondence between A^* and $L(G)$, associating any word $w \in A^*$ to the word generated in G by the sequence of the corresponding productions, followed by p_T.

More formally, one can define the map $c_f \colon A^* \to L(G)$ as follows: if $w = a_1 a_2 \cdots a_n$, $a_i \in A$, $1 \leq i \leq n$, then $c_f(w)$ is the word v such that

$$X \underset{\delta}{\Rightarrow} v, \quad \text{where } \delta = f(a_1) f(a_2) \cdots f(a_n) p_T.$$

The map c_f is a bijection if and only if the grammar G is unambiguous. In such a case, 'decoding', that is the effective computation of the inverse mapping c_f^{-1}, is obtained by parsing the coded sentences.

The object of this section is to investigate the connections between minimal linear grammars and codes with respect to the CE Problem. Clearly, minimal linear grammars are non-expansive so that Theorem 2 and Theorem 3 apply to them. However, by exploiting the connection between such grammars and codes, new conditions ensuring that they generate a commutatively regular language can be set up.

Theorem 4 *Let G be an unambiguous minimal linear grammar with the productions (3). If the list of words $(u_iv_i)_{i=1,\dots,m}$ is commutatively prefix, then $L(G)$ is commutatively regular.*

PROOF By hypothesis, there exists a prefix set $\mathcal{Y} = \{y_1, \dots, y_m\}$ such that $y_i \sim u_iv_i$, $1 \leq i \leq m$. We shall prove that there exist words $z_1, \dots, z_n$ such that $|z_j| = |w_j|$, $1 \leq j \leq n$ and $L(G)$ is commutatively equivalent to the regular set $\mathcal{Y}^*\{z_1, \dots, z_n\}$.

The proof is by induction on n.

First suppose $n = 1$. In this case, in view of the unambiguity of G, any word of $L(G)$ can be uniquely written as

$$u_{i_1}u_{i_2}\cdots u_{i_h}w_1v_{i_h}\cdots v_{i_2}v_{i_1}$$

with $h \geq 0$, $i_1, \dots, i_h \in \{1, \dots, m\}$ and, conversely, any word of this form belongs to $L(G)$. Similarly, taking into account that $\mathcal{Y}$ is a code, any word of $\mathcal{Y}^*w_1$ can be uniquely written as

$$y_{i_1}y_{i_2}\cdots y_{i_h}w_1$$

with $h \geq 0$, $i_1, \dots, i_h \in \{1, \dots, m\}$ and, conversely, any word of this form belongs to $\mathcal{Y}^*w_1$. Since

$$u_{i_1}u_{i_2}\cdots u_{i_h}w_1v_{i_h}\cdots v_{i_2}v_{i_1} \sim y_{i_1}y_{i_2}\cdots y_{i_h}w_1,$$

we conclude that $L(G)$ and $\mathcal{Y}^*w_1$ are commutatively equivalent. Thus, the statement is verified taking $z_1 = w_1$.

Now suppose $n \geq 2$. With no loss of generality, we may assume that $|w_n| = \max_{1\leq j\leq n}|w_j|$. Let G' be the grammar obtained by G deleting the production $X \to w_n$. By the induction hypothesis, we assume that there are $z_1, \dots, z_{n-1}$ such that $|z_i| = |w_i|$, $1 \leq i \leq n-1$ and $\mathcal{Y}^*\{z_1, \dots, z_{n-1}\}$ is commutatively equivalent to $L(G')$. Since $w_n \notin L(G')$, necessarily there exists a word

$$z_n \notin \mathcal{Y}^*\{z_1, \dots, z_{n-1}\} \tag{4}$$

such that $z_n \sim w_n$. The language $L(G) \setminus L(G')$ is generated by the grammar obtained by G deleting the productions $X \to w_i$, $i = 1, \dots, n-1$. Thus, by an argument similar to that used in the case $n = 1$, one can prove that $L(G)\setminus L(G')$ is commutatively equivalent to $\mathcal{Y}^*z_n$. Let us verify that

$$\mathcal{Y}^*\{z_1, \dots, z_{n-1}\} \cap \mathcal{Y}^*z_n = \emptyset.$$

Indeed, if it was not the case, one would have $yz_n = y'z_j$ for some $y, y' \in \mathcal{Y}^*$ $1 \leq j \leq n-1$. By our assumption on the maximality of $|w_n|$, one has $|z_n| \geq |z_j|$ so that $y' = yx$, $z_n = xz_j$, for a suitable word x. Taking into account that $\mathcal{Y}$ is a prefix code, one derives $x \in \mathcal{Y}^*$ and therefore $z_n \in \mathcal{Y}^* z_j$, contradicting (4).

In conclusion, $\mathcal{Y}^*\{z_1, \ldots, z_{n-1}\}$ and $\mathcal{Y}^* z_n$ are disjoint sets commutatively equivalent to $L(G')$ and $L(G) \setminus L(G')$, respectively. Thus, $\mathcal{Y}^*\{z_1, \ldots, z_n\}$ is commutatively equivalent to $L(G)$. This concludes the proof. □

For any $k \in \mathbb{N}$, denote by L_k the set of words $\{w \in T^* \mid X \Rightarrow^{k+1} w\}$. If the production $X \to \epsilon$ is present in G, the previous theorem takes a simpler form.

Corollary 3 *Let G be an unambiguous minimal linear grammar. Assume that $X \to \epsilon$ is a production of G and, for some $k \in \mathbb{N}$, L_k is commutatively prefix. Then $L(G)$ is commutatively regular.*

Proof Let $G' = (V, T, P', X)$ be the grammar with the following productions:

$$\begin{array}{ll} X \to uXv & \text{if } X \Rightarrow^k uXv \text{ in } G, \\ X \to w & \text{if there exists } \ell \leq k \text{ such that } X \Rightarrow^\ell w \text{ in } G,\ w \in T^*. \end{array}$$

One easily checks that G' is unambiguous and $L(G') = L(G)$. Indeed, as any production of G' corresponds to a derivation of G, one has $L(G') \subseteq L(G)$ and, moreover, no word of $L(G')$ can have two distinct derivations in G' since, otherwise, it would have two derivations in G. Conversely, any derivation of a word $w \in L(G)$ in G can be decomposed as

$$X \overset{k}{\Rightarrow} \alpha_1 \overset{k}{\Rightarrow} \alpha_2 \overset{k}{\Rightarrow} \cdots \overset{k}{\Rightarrow} \alpha_q \overset{r}{\Rightarrow} w$$

with $q \geq 0$ and $0 < r \leq k$. Taking into account the linearity of G, one obtains that in G' there is the derivation $X \Rightarrow \alpha_1 \Rightarrow \cdots \Rightarrow \alpha_q \Rightarrow w$. This proves the inclusion $L(G) \subseteq L(G')$.

Moreover,

$$\{uv \mid X \to uXv \in P'\} \subseteq L_k$$

and for any word $w \in L_k$ there is at most one production $X \to uXv$ of G' such that $w = uv$. Since L_k is commutatively prefix, this ensures that G' satisfies the hypotheses of Theorem 4. The conclusion follows. □

Example 3 Let G be the linear minimal grammar whose set of productions is $P = \{X \to abXb,\ X \to abXaa,\ X \to baXb,\ X \to \varepsilon\}$. One easily checks that G is unambiguous. Moreover, $L_1 = \{abb, abaa, bab\}$ is a prefix set and, therefore, $L(G)$ is commutatively equivalent to L_1^*.

A natural problem arising from the previous results is to figure out which minimal linear grammars satisfy the hypotheses of Theorem 4 and Corollary 3. In view of this problem, an essential element of the study of the CE Problem is the property of unambiguity of the grammar. We thus investigate conditions that force these grammars to satisfy that property. These conditions mimic for minimal linear grammars well-known properties of codes.

6.1 Measure of a minimal linear grammar

Let T be a finite alphabet and $\mathbb{R}_+$ be the set of non-negative real numbers. A *Bernoulli distribution* μ on T is any map

$$\mu\colon T \to \mathbb{R}_+,$$

such that $\sum_{a\in T} \mu(a) = 1$. A Bernoulli distribution is *positive* if, for all $a \in T$, $\mu(a) > 0$. Any Bernoulli distribution μ over T is extended to a unique morphism (still denoted μ) of T^* into the multiplicative monoid $\mathbb{R}_+$. One then extends μ to the family of subsets of T^* by setting, for every $X \subseteq T^*$, $\mu(X) = \sum_{x\in X} \mu(x)$. The following holds.

Proposition 6 *Let G be an unambiguous minimal linear grammar with the productions (3). One has $\sum_{i=1}^m \mu(u_iv_i) \leq 1$ for all positive Bernoulli distributions μ on T.*

Proof Set $p = \sum_{i=1}^m \mu(u_iv_i)$. Let $k \geq 1$. As one easily verifies, a word w belongs to L_k if and only if it can be factorized

$$w = u_isv_i \quad \text{with } s \in L_{k-1},\ 1 \leq i \leq m.$$

Moreover, such a factorization is unique by the unambiguity of G. One derives

$$\mu(L_k) = \sum_{i=1}^m \sum_{s\in L_{k-1}} \mu(u_isv_i) = \sum_{i=1}^m \mu(u_iv_i) \sum_{s\in L_{k-1}} \mu(s) = p\mu(L_{k-1}).$$

From the equation above, one obtains

$$\mu(L_k) = p^kq, \quad k \geq 0, \tag{5}$$

where $q = \mu(L_0) > 0$.

Now, let ℓ be the maximal length of the right sides of the productions of G. One easily verifies that, for all $k \geq 0$, the maximal length of the words of L_k is smaller than or equal to $(k+1)\ell$. Consequently,

$$\mu(L_k) \leq \sum_{i=0}^{(k+1)\ell} \mu(T^i) = (k+1)\ell + 1, \quad k \geq 0. \tag{6}$$

From Equations (5) and (6) one obtains $p^kq \leq (k+1)\ell + 1$ for all $k \geq 0$. This necessarily implies $p \leq 1$. □

Remark 3 Taking into account that the sum $\sum_{i=1}^m \mu(u_iv_i)$ is equal to the value of the polynomial $\sum_{i=1}^m \underline{\underline{u_iv_i}}$ when each letter $a \in T$ is replaced by its probability $\mu(a)$, the statement of the proposition above may be extended, by continuity, to all (not necessarily positive) Bernoulli distributions.

Now we give a characterization of unambiguous minimal linear grammars.

Proposition 7 *Let G be a minimal linear grammar and μ be a positive Bernoulli distribution on T. Then G is unambiguous if and only if the following two conditions are satisfied:*

1. *no word of $L(G)$ has two distinct derivations of length 2.*
2. *for all $k \geq 1$, one has*

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) = \sum_{i=0}^{k} \left(\frac{\mu(L_1)}{\mu(L_0)}\right)^i \mu(L_0). \tag{7}$$

PROOF For any $i \geq 0$, let Δ_i be the set of all derivations of G of length $i+1$ and for any $\delta \in \Delta_i$, denote by w_δ the word such that $S \Rightarrow_\delta^* w_\delta$. Since $L_i = \{w_\delta \mid \delta \in \Delta_i\}$, for all $k \geq 0$ one has

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) \leq \sum_{i=0}^{k} \sum_{\delta \in \Delta_i} \mu(w_\delta), \tag{8}$$

where the $=$ sign holds if and only if the words w_δ occurring in the equation are pairwise distinct.

Let us evaluate $\sum_{\delta \in \Delta_i} \mu(w_\delta)$. Note that $w_\delta = u_1 u_2 \cdots u_i w v_i \cdots v_2 v_1$, where δ is the sequence of the productions $X \to u_1 X v_1$, $X \to u_2 X v_2$, $\ldots$, $X \to u_i X v_i$, $X \to w$. Since $\mu(w_\delta) = \mu(u_1 v_1)\mu(u_2 v_2) \cdots \mu(u_i v_i)\mu(w)$, one easily derives

$$\sum_{\delta \in \Delta_i} \mu(w_\delta) = \left(\sum_{X \to uXv \text{ in } P} \mu(uv)\right)^i \sum_{X \to w \text{ in } P} \mu(w). \tag{9}$$

As one easily verifies, the last factor in the right hand side of the equation above is equal to $\mu(L_0)$.

Suppose that G is unambiguous. Then Condition 1 is trivially verified. Consequently, one has $\mu(L_1) = \sum_{\delta \in \Delta_1} \mu(w_\delta)$. Thus, (9) for $i = 1$ becomes

$$\mu(L_1) = \left(\sum_{X \to uXv \text{ in } P} \mu(uv)\right) \mu(L_0).$$

Solving with respect to the factor in parentheses and replacing in (9) one obtains

$$\sum_{\delta \in \Delta_i} \mu(w_\delta) = \left(\frac{\mu(L_1)}{\mu(L_0)}\right)^i \mu(L_0). \tag{10}$$

Taking into account that in this case (8) holds with the equal sign, one obtains (7). This proves that if G is unambiguous, Condition 2 is satisfied.

Conversely, suppose that G is ambiguous but Condition 1 is satisfied. Then one obtains again (10). In this case there is k for which (8) holds with the $<$ sign. Thus, one obtains

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) < \sum_{i=0}^{k} \left(\frac{\mu(L_1)}{\mu(L_0)}\right)^i \mu(L_0),$$

so that Condition 2 is not satisfied. This proves that if G is ambiguous, at least one of Conditions 1 and 2 is not satisfied. □

In the case where the only terminal production is $X \to \epsilon$, one has $\mu(L_0) = 1$ so that we obtain the following.

Corollary 4 *Let G be a minimal linear grammar such that the only terminal production is $X \to \epsilon$, and μ be a positive Bernoulli distribution on T. Then G is unambiguous if and only if the following two conditions are satisfied:*

1. *no word of $L(G)$ has two distinct derivations of length 2.*
2. *for all $k \geq 1$, one has*

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) = \sum_{i=0}^{k} (\mu(L_1))^i . \tag{11}$$

We notice that, by Proposition 6, $\mu(L_1) \leq 1$. If $\mu(L_1) < 1$, then Eq. (11) can be written as:

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) = \frac{1 - \mu(L_1)^{k+1}}{1 - \mu(L_1)}.$$

This equation holds true also in the case $\mu(L_1) = 1$, assuming that, by continuity, $(1 - x^{k+1})/(1 - x) = k - 1$, for $x = 1$.

6.2 Languages with k b's

In the case of a binary alphabet, the previous propositions can be refined. For this purpose, let us prove a statement which is a natural extension of [5, Example 6.3]. We recall that, given a language L, the *growth function* g_L of L returns, for any non-negative integer n, the number of the words of L whose length is at most n.

Proposition 8 *Let $k > 0$ be an integer. A subset L of $(a^*b)^k a^*$ is commutatively prefix if and only if its growth function g_L satisfies the inequality*

$$g_L(n) \leq \binom{n}{k}, \quad n \geq k. \tag{12}$$

Proof Let $T = \{a, b\}$. If L is a prefix subset of $(a^*b)^k a^*$, then, for any n, the words

$$xa^{n-|x|}, \quad x \in L,\ |x| \leq n$$

are pairwise distinct words of the set

$$\mathcal{Z}_n = \{w \in T^* \mid |w|_b = k,\ |w|_a = n - k\}.$$

One derives

$$g_L(n) \leq \mathrm{Card}(\mathcal{Z}_n) = \binom{n}{k}.$$

This proves that the condition is necessary.

In order to prove sufficiency, assuming that (12) is satisfied, we will construct a chain of prefix sets $\mathcal{Y}_0 \subseteq \mathcal{Y}_1 \subseteq \cdots \subseteq \mathcal{Y}_n \subseteq \cdots$ such that $\mathcal{Y}_n$ is commutatively equivalent to $L \cap T^{\leq n}$, $n \geq 0$. For $n < k$, it is sufficient to take $\mathcal{Y}_n = \emptyset$. Now, let $n \geq k$. Proceeding inductively, we assume that there is a prefix set $\mathcal{Y}_{n-1}$ commutatively equivalent to $L \cap T^{\leq n-1}$. Since $\mathrm{Card}(\mathcal{Z}_n) = \binom{n}{k}$ and in view of (12), one has

$$\mathrm{Card}(L \cap T^n) = g_L(n) - g_L(n-1) \leq \mathrm{Card}(\mathcal{Z}_n) - \mathrm{Card}(\mathcal{Y}_{n-1}).$$

Taking into account that any word of $\mathcal{Y}_{n-1}$ is a prefix of a unique word of $\mathcal{Z}_n$, one can find $\mathrm{Card}(L \cap T^n)$ distinct words of $\mathcal{Z}_n$ with no prefix in $\mathcal{Y}_{n-1}$. These words are commutatively equivalent to those of the set $L \cap T^n$. Thus, adding these words to $\mathcal{Y}_{n-1}$, one obtains a prefix set $\mathcal{Y}_n$ commutatively equivalent to $L \cap T^{\leq n}$. We conclude that $\mathcal{Y} = \bigcup_{n \geq 0} \mathcal{Y}_n$ is a prefix set commutatively equivalent to L. □

As an immediate consequence of Corollary 3 and Proposition 8, we get:

Corollary 5 *Let G be an unambiguous minimal linear grammar with the terminal production $X \to \epsilon$. Assume that, for some $h, k \in \mathbb{N}$, $L_h \subseteq (a^*b)^k a^*$. If the growth function g_{L_h} of L_h satisfies the inequality $g_{L_h}(n) \leq \binom{n}{k}$, $n \geq k$, then $L(G)$ is commutatively regular.*

We recall that a set L of words is said to be a *Bernoulli set* if, for every Bernoulli distribution μ, $\mu(L) = 1$. In [25], a remarkable result of de Luca shows that every Bernoulli set contained in $a^* \cup a^*ba^* \cup a^*ba^*ba^*$, is commutatively prefix. As a consequence of this result, Corollary 3 and Corollary 4 we get the following.

Corollary 6 *Let G be a minimal linear grammar with the sole terminal production $X \to \epsilon$. Assume that $L_1 \subset a^* \cup a^*ba^* \cup a^*ba^*ba^*$ and no word of $L(G)$ has two distinct derivations of length 2. If for every Bernoulli distribution μ, $\mu(\bigcup_{i=0}^{k} L_i) = k + 1$, then $L(G)$ is commutatively regular.*

Proof By the hypothesis on G, one has that: 1) L_1 is commutatively prefix, by de Luca's result [25]; 2) G is unambiguous by Corollary 4. Then the claim follows by applying Corollary 3. □

The interest of Corollary 6 lies on the fact that one does not need to suppose that the grammar G is unambiguous in order to prove that $L(G)$ is commutatively regular. This is emphasized by the following example.

Example 4 Let G be the linear minimal grammar with terminal alphabet $T = \{a, b\}$ and with the productions

$$X \to a^2X, \quad X \to abXa, \quad X \to abXb, \quad X \to bXa, \quad X \to bXb, \quad X \to \epsilon.$$

One has $L_1 = \{aa, ba, bb, aba, abb\} \subseteq a^* \cup a^*ba^* \cup a^*ba^*ba^*$ and no word of L_1 has two derivations of length 2. Moreover, as one easily verifies,

$$\bigcup_{i=0}^{k} L_i = \{uv \mid u \in \mathcal{Z}^{\leq k},\ v \in T^{|u|_b}\}, \quad \text{where } \mathcal{Z} = \{a^2, ab, b\}$$

so that $\mu\big(\bigcup_{i=0}^{k} L_i\big) = k+1$, for all Bernoulli distribution μ. Thus, by the above corollary, $L(G)$ is commutatively regular.

The problem whether every maximal finite code is commutatively prefix, is still open. The conjecture was originally formulated by Schützenberger at the end of 50's for the case of finite codes (see [5, 25, 41]). The conjecture in this formulation has been disproved by Shor [46]. Indeed, the set L defined as:

$$L = \{b, ba, ba^7, ba^{13}, ba^{14}, a^3b, a^3ba^2, a^3ba^4, a^3ba^6, \\ a^8b, a^8ba^2, a^8ba^4, a^8ba^6, a^{11}b, a^{11}ba, a^{11}ba^2\}$$

is a code which is not commutatively prefix. However a simple computation shows that the growth function g_{L^2} of L^2 satisfies the inequality $g_{L^2}(n) \leq \binom{n}{2}$ for all $n \geq 2$ and, therefore, L^2 is commutatively prefix. Thus one may ask whether any finite code $\mathcal{Y}$ has a power $\mathcal{Y}^n$ which is commutatively prefix. In [25], a positive answer to the latter question has been conjectured in the case of finite complete codes.

7 Concluding Remarks

In this paper, we have studied conditions ensuring that a language generated by an unambiguous non-expansive context-free grammar is commutatively regular. The choice of focusing on non-expansive grammars was motivated by the fact that the characteristic series in commutative variables of their languages are rational and this is a necessary condition for commutative regularity.

Actually, a language is commutatively regular if and only if its characteristic series in commutative variables is also the characteristic series of a regular language. Thus, one may consider the general problem of studying the characteristic series of regular languages (in commutative variables). In other terms, we would like some criterion to determine whether a rational series in commutative variables is the natural projection of an unambiguous rational series in non-commutative variables.

If s is the characteristic series in commutative variables of any language on the alphabet $A = \{a_1, \ldots, a_k\}$, then one has

$$\big(s,\, a_1^{n_1} \cdots a_k^{n_k}\big) \leq \frac{(n_1 + \cdots + n_k)!}{n_1! \cdots n_k!}\,. \tag{13}$$

for all $n_1, \ldots, n_k \in \mathbb{N}$. Indeed, the multinomial coefficient in the right hand side of (13) gives the number of the words of A^* which are commutatively equivalent

to $a_1^{n_1} \cdots a_k^{n_k}$. As far as we know, the question whether there exist rational series satisfying (13) which are not the characteristic series of a regular set is open. As already mentioned in the introduction, a related problem, namely characterizing generating series of regular languages on a fixed alphabet, was studied in [3].

In [29], it has been shown that the language $L = \{a^{n_1}ba^{n_2}b \cdots a^{n_{k-1}}ba^{n_k} \mid n_1 \leq n_2 \leq \cdots \leq n_k, k \geq 1\}$, over the alphabet $A = \{a, b\}$, which is accepted by a deterministic non-erasing stack automaton, is of intermediate growth, and its generating series is transcendental. This language can be generated by a grammar of a special type called *indexed* [1] (see also [31]). Recently, in [22,23], a subclass of such grammars called *uncontrolled finite-index* has been investigated. An indexed grammar is called uncontrolled finite-index if there is an integer k such that in every successful derivation the number of occurrences of non-terminals in each sentential form is bounded by k. These grammars, that extend the context-free ones, inherit several properties of them and, in particular, the fact that the corresponding family of languages is a semi-linear full trio. It would be interesting to investigate the CE problem and the related issues for these generating systems.

Incidentally, it is maybe of interest to note that the authors investigated recently the Kleene closure of bounded semi-linear languages in connection with the CE Problem. In [11, 13] the commutative regularity of these languages has been established, in the case that the bounded semi-linear language is a code satisfying some restriction on the number of letters of its words.

In Section 6, we have considered in particular minimal linear grammars. We stressed some analogies between these grammars and variable-length codes. Thus, it seems interesting to investigate what properties of variable-length codes can be extended to minimal linear grammars. This study has been started in [12].

Here, we limit ourselves to propose the following open question: a theorem proven in [7] (see also [25]) states that, for every set L of words of A^+, any two of the following three conditions imply the remaining one: (i) L is a code; (ii) L is a complete set, that is, for every $w \in A^*$, $A^*wA^* \cap L^* \neq \emptyset$; (iii) $\mu(L) = 1$, where μ is a positive Bernoulli distribution. This result provides a remarkable relation among the properties of codicity, completeness and measure of the set. It would be interesting to get a similar characterization of the property of unambiguity for minimal linear grammars.

References

[1] A. V. Aho, Indexed grammars – an extension of context-free grammars, J. ACM, **15** (4), 647–671 (1968).

[2] G. Baron, W. Kuich, The characterization of nonexpansive grammars by rational power series, *Information and Control* **48**, 109–118 (1981).

[3] M.-P. Béal, D. Perrin, On the generating sequences of regular languages on k symbols, *J. ACM* **50**, 955–980 (2003).

[4] J. Berstel, *Transductions and Context-Free Languages*, Teubner, Stuttgart, 1979.

[5] J. Berstel, D. Perrin, Ch. Reutenauer, *Codes and Automata*, Encyclopedia of Mathematics and its Applications No. 129, Cambridge University Press, Cambridge, 2009.

[6] L. Boasson, A. Restivo, Une caractérisation des langages algébriques bornés, *RAIRO Inform. Théor. Appl.* **11**, 203–205 (1977).

[7] J. M. Boë, A. de Luca, A. Restivo, Minimal complete sets of words, *Theoret. Comput. Sci.* **12**, 325–332, (1980).

[8] M. R. Bridson, R. H. Gilman, Context-free languages of sub-exponential growth, *J. Comput. System Sci.* **64**, 308–310 (1999).

[9] J. A. Brzozowski, Regular-like expressions for some irregular languages, in *9th Annual Symposium on Switching and Automata Theory (swat 1968)*, IEEE Computer Society, 1968, pp. 278–286.

[10] A. Carpi, F. D'Alessandro, On the commutative equivalence of context-free languages, in: M. Hoshi, S. Seki (Eds), *Developments in Language Theory 2018*, Lecture Notes in Computer Science, 11088, Springer, Berlin, 2018, pp. 169–181.

[11] A. Carpi, F. D'Alessandro, On the commutative equivalence of bounded semi-linear codes, In: R. Mercas, D. Reidenbach (Eds), *Combinatorics on Words 12th International Conference WORDS 2019*, Lecture Notes in Computer Science, 11682, Springer, Berlin, 2019, pp. 119-132.

[12] A. Carpi, F. D'Alessandro, Coding by minimal linear grammars, *Theoret. Comput. Sci.* **834**, 14–25 (2020).

[13] A. Carpi, F. D'Alessandro, On bounded linear codes and the commutative equivalence, Theoret. Comput. Sci., in press, available on line, https://doi.org/10.1016/j.tcs.2020.11.031 (2020).

[14] A. Carpi, F. D'Alessandro, O. H. Ibarra, I. McQuillan, Relationships between bounded languages, counter machines, finite-index grammars, ambiguity, and commutative regularity, Theoret. Comput. Sci., in press, available on line, https://doi.org/10.1016/j.tcs.2020.10.006 (2020).

[15] L. Carter, J. Gill, Conjectures on uniquely decipherable codes, *IRE Trans. Inform. Theory*, **IT-7**, 27–38, (1961).

[16] N. Chomsky, M.-P. Schützenberger, The algebraic theory of context-free languages, in P. Braffort and D. Hirschberg (eds.), *Computer Programming and Formal Systems*, North Holland Publishing Company, Amsterdam, 1963, pp. 118–161.

[17] F. D'Alessandro, B. Intrigila, S. Varricchio, The Parikh counting functions of sparse context-free languages are quasi-polynomials, *Theoret. Comput. Sci.* **410**, 5158–5181 (2009).

[18] F. D'Alessandro, B. Intrigila, S. Varricchio, Quasi-polynomials, linear Diophantine equations and semi-linear sets, *Theoret. Comput. Sci.* **416**, 1–16 (2012).

[19] F. D'Alessandro, B. Intrigila, On the commutative equivalence of bounded context-free and regular languages: the code case, *Theoret. Comput. Sci.* **562**, 304–319 (2015).

[20] F. D'Alessandro, B. Intrigila, On the commutative equivalence of semilinear sets of $\mathbb{N}^k$, *Theoret. Comput. Sci.* **562**, 476–495 (2015).

[21] F. D'Alessandro, B. Intrigila, On the commutative equivalence of bounded context-free and regular languages: the semi-linear case, *Theoret. Comput. Sci.* **572**, 1–24 (2015).

[22] F. D'Alessandro, O. H. Ibarra, I. McQuillan, On finite-index indexed grammars and their restrictions, In: F. Drewes, B. Truthe, and C. Martin Vides (Eds), *Language and Automata Theory and Applications 2017*, Lecture Notes in Computer Science, 10168, Springer, Heidelberg, 2017, pp. 287–298.

[23] F. D'Alessandro, O. H. Ibarra, I. McQuillan, On finite-index indexed grammars and their restrictions, *Information and Computation*, in press, available on line, https://doi.org/10.1016/j.ic.2020.104613, (2020).

[24] A. de Luca, S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages*, Springer, Berlin, 1999.

[25] A. de Luca, Some combinatorial results on Bernoulli sets and codes, *Theoret. Comput. Sci.* **273**, 143–165 (2002).

[26] P. Flajolet, Analytic models and ambiguity of context-free languages, *Theoret. Comput. Sci.* **49**, 283–309 (1987).

[27] S. Ginsburg, *The Mathematical Theory of Context-Free Languages*, Mc Graw-Hill, New York, 1966.

[28] S. Ginsburg E. H. Spanier, Semigroups, Presburger formulas, and languages, *Pacific J. Math.* **16**, 285–296 (1966).

[29] R. I. Grigorchuk, A. Machí, An example of an indexed language of intermediate growth. *Theoret. Comput. Sci.* **215**, 325–327 (1999).

[30] M. Gross, Inherent ambiguity of minimal linear grammars, *Inf. and Cont.* **7**, 366–368 (1964).

[31] M. A. Harrison, *Introduction to Formal Language Theory*, Addison-Wesley Publishing Co., Reading, Mass., 1978.

[32] J. Honkala, Decision problems concerning thinness and slenderness of formal languages, *Acta Inf.* **35**, 625–636 (1998).

[33] J. Honkala, On Parikh slender context-free languages, *Theoret. Comput. Sci.* **255**, 667–677 (2001).

[34] O. H. Ibarra, B. Ravikumar, On sparseness, ambiguity and other decision problems for acceptors and transducers, In: B. Monien, G. Vidal-Naquet (Eds), *3rd Annual Symposium on Theoretical Aspects of Computer Science*, Lecture Notes in Computer Science, 210, Springer, Berlin, 1986, pp. 171–179.

[35] O. H. Ibarra, I. McQuillan, B. Ravikumar, On counting functions of languages, In: M. Hoshi, S. Seki (Eds), *Developments in Language Theory 2018*, Lecture Notes in Computer Science, 11088, Springer, Berlin, 2018, pp. 429–440.

[36] O. H. Ibarra, I. McQuillan, B. Ravikumar, On counting functions and slenderness of languages, *Theoret. Comput. Sci.* **777**, 356–378 (2019).

[37] L. Ilie, G. Rozenberg, A. Salomaa, A characterization of poly-slender context-free languages, *RAIRO Inform. Théor. Appl.* **34**, 77–86 (2000).

[38] R. Incitti, The growth function of context-free languages, *Theoret. Comput. Sci.* **255**, 601–605 (2001).

[39] W. Kuich, Cycle-free $\mathbb{N}$-algebraic systems, In: P. Deussen, (Ed), *Theoretical Computer Science, 5th GI-Conference*, Lecture Notes in Computer Science, 104, Springer, Berlin, 1981, pp. 5–12.

[40] M. Latteux, G. Thierrin, On bounded context-free languages, *Elektron. Inform. Verarb. u. Kybern.* **20**, 3–8 (1984).

[41] D. Perrin, M.-P. Schützenberger, Un problème élémentaire de la théorie de l'information, in *Théorie de l'Information*, Colloq. Internat. CNRS No. 276, 1977, pp. 249-260.

[42] A. Restivo, A characterization of bounded regular sets, In: H. Brakhage (Ed), *Automata Theory and Formal Languages*, Lecture Notes in Computer Science, 33, Springer, Berlin, 1975, pp. 239–244.

[43] A. Salomaa, On the index of a context-free grammar and language, *Information and Control* **14**, 474–477 (1969).

[44] A. Salomaa, *Formal Languages*, Academic Press, New York, 1973.

[45] A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series*, Springer, New York, Heidelberg, Berlin, 1978.

[46] P. W. Shor, A counterexample to the triangle conjecture, *J. Combin. Theory Ser. A.*, **38**, 110–112 (1983).