On the commutative equivalence of context-free languages

Arturo Carpi[1] and Flavio D'Alessandro[2]

[1] Dipartimento di Matematica e Informatica, Università degli Studi di Perugia,
via Vanvitelli 1, 06123 Perugia, Italy.
carpi@dmi.unipg.it
[2] Dipartimento di Matematica, Università di Roma "La Sapienza"
Piazzale Aldo Moro 2, 00185 Roma, Italy.
dalessan@mat.uniroma1.it

Abstract. The problem of the commutative equivalence of context-free and regular languages is studied. In particular 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, Uniquely decipherable code, Exponential growth

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 a permutation of 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 a 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, [3, 12]). 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*. 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 in 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

proved in [5, 19] states that every context-free language is either sparse or of exponential growth. The second, proved in [16, 20], states that the class of sparse context-free languages coincides with that of *bounded languages*. We recall that a language L is termed *bounded* if there exist k non-empty 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 where remarkable theorems characterize the structure of these languages [7, 8, 14–18, 20, 21]. In [9–11] it has been given the solution (in the affirmative) of the CE Problem for sparse languages: *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 ([14, 15]) 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 implies two consequences. The first is that the answer to the CE Problem is not affirmative in general. Indeed, the generating series of a commutatively regular language must be rational, while, on the other hand, there exist context-free languages whose generating series are algebraic not rational and, even, transcendental, as proven by Flajolet in [13]. Dyck and semi-Dyck languages are the first example of such languages. 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 $\alpha_1, \alpha_2, \alpha_3 \in V^*$, V being the set of non-terminals of G. In the opposite case, G is non-expansive.

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.

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 8): the language generated by every unambiguous and non-expansive grammar G is commutatively regular, provided that there exist a code $\mathcal{W}$ of words and a bijection $f\colon P \to \mathcal{W}$, between the set of productions P

of G and $\mathcal{W}$ such that, for every production $p \in P$, the word obtained deleting all non-terminals in the right side of p 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 6).

Two ingredients play a role in the proof of Theorem 8: codes and a special structuring of the derivations of the grammar. Such structuring is based upon two objects: *minimal non-terminals* and *leftmost minimal derivations*. A non-terminal is called minimal if it is minimal with respect to the quasi-order $\leq$ defined on the set of non-terminals as follows: if X, Y are non-terminals, one has $X \leq Y$ if there is a derivation $X \Rightarrow^* \alpha_1 Y \alpha_2$. A derivation of G is called leftmost minimal if, at each step of the derivation, the production is applied to the leftmost occurrence of a minimal non-terminal of the sentential form. One of the key-feature of such structuring is that, in every leftmost minimal derivation of a non-expansive grammar, the number of occurrences of non-terminals in any of its sentential forms is bounded by an integer depending only on the grammar. This fact together with the use of codes allows us to develop a technique to deal with the problem.

The use of these objects 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 generalized automaton whose behaviour coincides with the characteristic series of the language generated by the grammar. Thus the CE Problem for unambiguous non-expansive grammars is reduced to the more 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 [6] is relevant in our study since, in the unambiguous case, the derivation process of words in such a grammar, is algebraically similar to that of words in a monoid generated by a code. 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 9 and Corollary 10). 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, the set of words uv where $X \to uXv$ is a production of G, has measure not larger than 1, with respect to every Bernoulli distribution (Proposition 11). The second result states, up to a

technical restriction, the very same characterization of codes in term of positive Bernoulli distributions (Proposition 12 and Corollary 13). 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.

2 Preliminaries

We now recall some useful terms and basic properties (see [3, 15]).

Words and languages. Let A be a finite non-empty alphabet and A^* be the free monoid generated by 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. More generally, for every subset B of A, we set $|w|_B = \sum_{b\in B} |w|_b$ and $\mathrm{alph}_B(w) = \{b \in B \mid |w|_b > 0\}$. If $A = \{a_1, \ldots, a_t\}$ is an ordered alphabet of t letters, and if $w \in A^*$ is an arbitrary word, then the *Parikh vector* of w is the tuple $\psi(w)$ of $\mathbb{N}^t$ defined as $\psi(w) = (|w|_{a_1}, \ldots, |w|_{a_t})$. The function $\psi\colon A^* \to \mathbb{N}^t$, mapping w into the Parikh vector of w, is an epi-morphism of the free monoid A^* onto the free commutative additive monoid $\mathbb{N}^t$, called the *Parikh morphism (over A)*. One can introduce in A^* the equivalence relation $\sim$, called *commutative equivalence*, defined as follows: for all $u, v \in A^*$ $u \sim v$ if $\psi(u) = \psi(v)$. 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 $f : L \to L'$ such that, for every $u \in L$, $u \sim f(u)$. A set $\mathcal{X}$ over the alphabet A is said to be a *prefix set* if $\mathcal{X}A^+ \cap \mathcal{X} = \emptyset$, that is, if, for every $u, v \in \mathcal{X}$, u is not a proper prefix of v. A set $\mathcal{X}$ of words over an alphabet A is said to be *commutatively prefix* if there exists a prefix set $\mathcal{X}'$ such that $\mathcal{X}$ is commutively equivalent to $\mathcal{X}'$. A subset $\mathcal{X}$ of A^+ is a *code (over A)* if every word of $\mathcal{X}^+$ has a unique factorization as a product of words of $\mathcal{X}$.

Let B and A be alphabets with $B \subseteq A$. The *projection of A^* onto B^** is the epi-morphism $\widehat{\pi}_B : A^* \to B^*$ generated by the function $\pi_B : A \to B \cup \{\varepsilon\}$ 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.

Formal series and generalized automata. 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 usually, the semiring of non-singular series in non-commutative and commutative variables will be denoted, respectively, by $\mathbb{N}\langle\langle A\rangle\rangle$ and $\mathbb{N}[[A]]$. The coefficient of a monomial w in the series s is denoted by (s, w). With any language L on an alphabet A, we associate the *characteristic series* 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* in commutative variables of L 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 whose Parikh vector is $(n_1, \ldots, n_t)$, $n_1, \ldots, n_t \in \mathbb{N}$. Let M be a monoid. A *generalized automaton* A over M is given by a finite digraph (Q, E) whose arrows are labelled by elements of M together with two subsets I and F of Q. The elements of Q, I and F are called respectively, *states*, *initial states*, and *final states* of the automaton and the elements of E are called *transitions*. Any path in the graph (Q, E) starting in an initial state and ending in a terminal state is said to be *successful*. The *label* of such a path is the product (computed in M) of the labels of its arrows. The *behaviour* of the automaton is the formal sum $\sum_{m \in M} k_m m$, where, for all $m \in M$, k_m is the number (possibly infinite) of successful paths of A with label m. If M is the free monoid (resp., the free commutative monoid) generated by an alphabet T, then the behaviour of A is a formal power series in non-commutative (resp., commutative) variables with coefficients in $\widehat{\mathbb{N}}$ and variables in T. If, moreover, in the automaton A there is no cycle with label ε, then the behaviour of A is a non-singular series.

As is well known, a formal power series (in non-commutative or commutative variables) is the behaviour of a generalized automaton if and only if it is rational, that is, it belongs to the minimal subsemiring of $\widehat{\mathbb{N}}\langle\langle T \rangle\rangle$ (resp., $\widehat{\mathbb{N}}[[T]]$) containing all monomials and closed for the $*$-operation.

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

Now we introduce the concept of *leftmost minimal derivation*. Let $G = \langle V, T, P, S \rangle$ 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 usually, 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. We say that a non-terminal X occurring in a sentential form α is *minimal* if there does not exist a non-terminal $Y \in \mathrm{alph}_N(\alpha)$ such that $Y < X$. Analogously to leftmost or rightmost derivations, we may consider derivations in G where the replaced non-terminal is the leftmost minimal non-terminal occurring in a sentential form. More formally, let $\alpha, \beta \in V^*$ be such that $\alpha \Rightarrow \beta$. Then there are $\alpha_1, \alpha_2, \gamma \in V^*$, $A \in N$ such that

$$\alpha = \alpha_1 A \alpha_2, \quad \beta = \alpha_1 \gamma \alpha_2, \quad A \to \gamma \text{ in } P. \tag{1}$$

If A is a minimal non-terminal of α and all non-terminals occurring in α_1 are not minimal non-terminal of α, then we say that $\alpha \Rightarrow \beta$ is a *leftmost minimal generation*. In such a case, we write $\alpha \underset{M,p}{\Longrightarrow} \beta$, where p is the production $A \to \gamma$. Sometimes, for simplicity, the subscript p will be omitted. The reflexive and transitive closure of the relation $\underset{M}{\Longrightarrow}$ will be denoted by $\overset{*}{\underset{M}{\Longrightarrow}}$. If $\alpha \in V^*$, then any sequence

$$S = \beta_0 \underset{M,p_1}{\Longrightarrow} \beta_1 \underset{M,p_2}{\Longrightarrow} \cdots \beta_{n-1} \underset{M,p_n}{\Longrightarrow} \alpha \tag{2}$$

with $\beta_1, \ldots, \beta_{n-1} \in V^*$, $p_1, \ldots, p_n \in P$ will be called a *leftmost minimal derivation* of α. In the sequel, if no ambiguity arises, the generation (2) will be simply denoted $S \Rightarrow_r \alpha$, where $r = p_1 \cdots p_n$.

As a straightforward adaptation of a classical result, one can prove that there exists a one-to-one correspondence between leftmost minimal derivations of a word $w \in L(G)$ and parse trees of such a word.

3 Non-expansive grammars and rational series

In the sequel, we consider a context-free grammar $G = \langle V, T, P, S\rangle$. 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 investigated structures. In this context, as a related result, it's worth to mention a remarkable contribution by Beal and Perrin that provides a characterization of the series that are generating series of regular languages on alphabets of prescribed size [1]. A result of Baron and Kuich [2] provides the characterization of context-free grammars whose characteristic series are 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 said non-expansive.

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 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 leftmost derivations of words 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 of the language generated by G. We say that a context-free grammar G is *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. The following theorem characterizes non-expansive grammars.

Theorem 1. [2] *A cycle-free reduced context-free grammar is non-expansive if and only if for all non-terminal 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.*

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 the behaviour of a (generalized) $\mathbb{N}$-automaton A on the free commutative monoid. Thus, in order to investigate the commutative regularity of the language L, one is reduced to ask whether the behaviour of A is the characteristic series in commutative variables of a regular language. In other terms, one is reduced to search for an unambiguous automaton (on the monoid T^*) whose behaviour, projected into the semiring of series in commutative variables, is equal to that of A.

Now we will show how to explicitly construct a generalized automaton whose behaviour is the characteristic series of a given non-expansive grammar. Such a construction will furnish also an alternative proof of the 'if' part of Theorem 1.

Let G be a grammar. Recall that $\leq$ and $\Rightarrow_M^*$ are, respectively, the quasi-order on N and the leftmost minimal derivation relation of G introduced in Section 2.

The following lemma shows that in a leftmost minimal derivation of a non-expansive grammar the number of non-terminals occurring in the sentential forms is limited.

Lemma 3. *Let $G = \langle V, T, P, S\rangle$ be a non-expansive context-free grammar. There exists an integer $k_G > 0$ such that if $S \Rightarrow_M^* \beta$, $\beta \in V^*$, then $|\pi_N(\beta)| \leq k_G$.*

Let G be a grammar. It is useful to identify leftmost minimal generations of G with the sequences of productions used in such a derivation. More precisely, let P^* be the set of the finite sequences of productions. We denote by $D_M(G)$ the set of the sequences $p_1 p_2 \cdots p_n$ with $p_i \in P$, $1 \leq i \leq n$, $n \geq 1$ such that there exists a leftmost minimal derivation (2) with $\alpha \in T^*$. Thus, $D_M(G)$ is a language on the alphabet P which 'represents' the leftmost minimal derivations of the grammar G. Now, we assume that G is non-expansive and construct an automaton A accepting $D_M(G)$.

Let k_G be as in Lemma 3. The set of states of A will be $Q = N^{\leq k_G}$. The transitions will be the triples (α, p, β) such that $\alpha, \beta \in Q$, $p \in P$ and there is a leftmost minimal generation

$$\alpha \underset{M,p}{\Longrightarrow} \beta', \quad \text{with} \quad \beta' \in V^* \text{ and } \beta = \pi_N(\beta')\,.$$

The unique initial state is S and the unique terminal state is ε. The following holds

Proposition 4. *The automaton A accepts the language $D_M(G)$.*

Now, we introduce the morphism $\phi_G : P^* \to \widehat{\mathbb{N}}[[T]]$ defined as follows. If p is the production $X \to \alpha$, then $\phi_G(p) = \underline{\underline{\pi_T(\alpha)}}$. The following holds.

Lemma 5. *Let G be a context-free grammar. If one has $S \Rightarrow_r \alpha$, $r \in P^*$, $\alpha \in V^*$, then $\phi_G(r) = \underline{\underline{\pi_T(\alpha)}}$.*

If G is non-expansive, by Lemma 5 one derives that whenever (2) is verified, one has

$$\phi_G(p_1 \cdots p_n) = \underline{\underline{w}}. \tag{3}$$

and thus one has $\phi_G\big(\underline{D_M(G)}\big) = \underline{\underline{G}}$. More precisely, replacing the labels p of the transitions of the automaton A by $\phi(p)$, one obtains a generalized automaton whose behaviour is $\underline{\underline{G}}$. By the way, this gives a proof of the 'if' part of Theorem 1.

4 The CE Problem for non-expansive grammars

Tha aim of this section is to study the CE Problem for non-expansive grammars. We provide a condition that assures that the language generated by such a grammar is commutatively regular. More precisely, we will establish the following.

Theorem 6. *Let $G = \langle V, T, P, S\rangle$ be an unambiguous, non-expansive grammar. Assume that the following properties are satisfied:*

i) for every production $X \to \alpha$ of P, $|\alpha|_T \geq 2\,\mathrm{Card}(P)$;
ii) for every $a \in T$, there exists at most one production $X \to \alpha$ of P such that $\mathrm{alph}_T(\alpha) = \{a\}$.

Then $L(G)$ is commutatively regular.

In order to prove our theorem, we need the following

Lemma 7. *Let $\mathcal{M} = (v_1, \ldots, v_m)$ be a list of words of T^+ such that:*

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

Then there exists a prefix code $\mathcal{W} = \{w_1, \ldots, w_m\}$ of m distinct words such that, for every $i = 1, \ldots, m$, one has $w_i \sim v_i$.

Now we prove the following

Theorem 8. *Let G be an unambiguous non-expansive grammar. Assume that there exist a code $\mathcal{W}$ and a bijection $f\colon P \to \mathcal{W}$ such that, for every production $p = (X \to \alpha)$ one has $\psi(f(p)) = \psi(\pi_T(\alpha))$. Then $L(G)$ is commutatively regular.*

Proof. We extend f to a morphism $\widehat{f}\colon P^* \to T^*$ and set $R = \widehat{f}(D_M(G))$. By Proposition 4, $D_M(G)$ and, consequently, R are regular sets. We will show that $L(G) \sim R$. Let $g\colon L(G) \to R$ be the map defined as follows: for all $w \in L(G)$, $g(w) = \widehat{f}(r)$, where $S \Rightarrow_r w$ is the unique rightmost minimal derivation of w in G. The map g is a bijection. Indeed, $g(L(G)) = \widehat{f}(D_M(G)) = R$. Moreover, taking into account that $\mathcal{W}$ is a code, one has that $\widehat{f}$ and, consequently, g are injective. To complete the proof, it is sufficient to verify that for all $w \in L(G)$,

$w \sim g(w)$. Let $S \Rightarrow_r w$ be the rightmost minimal derivation of w in G. By (3), one has $\phi_G(r) = \underline{\underline{w}}$. As one easily checks, for all $p \in P$, one has $\phi_G(p) = \underline{\underline{f(p)}}$. This implies, in particular, $\phi_G(r) = \underline{\underline{\widehat{f}(r)}} = \underline{\underline{g(w)}}$. Thus, $\underline{\underline{w}} = \underline{\underline{g(w)}}$, that is, $w \sim g(w)$. We conclude that $L(G)$ is commutatively equivalent to R. □

Now we are ready to prove the main result of this section.

Proof (of Theorem 6). Let $\alpha_1, \ldots, \alpha_m$ be the list of the right sides of the productions of G, $m = \mathrm{Card}(P)$. One easily checks that the words $v_i = \pi_T(\alpha_i)$, $1 \leq i \leq m$, satisfy the hypotheses of Lemma 7. Thus, there exists a prefix code $\mathcal{W} = \{w_1, \ldots, w_m\}$ such that $w_i \sim v_i$, $1 \leq i \leq m$. Moreover, the function $f\colon P \to \mathcal{W}$ mapping any production $p_i = (X \to \alpha_i)$ into the word w_i is a bijection. The statement then follows from Theorem 8. □

5 Minimal Linear grammars

Minimal linear grammars, first introduced by Chomsky and Schützenberger in [6], 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

$$X \to u_i X v_i,\ 1 \leq i \leq m, \qquad X \to w_j,\ 1 \leq j \leq n, \tag{4}$$

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 that of words in a monoid generated by a code. The object of this section is to investigate the connections between such grammars and codes with respect to the CE Problem. Clearly, minimal linear grammars are non-expansive so that Theorems 6 and 8 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 9. *Let G be an unambiguous minimal linear grammar with the productions (4). Suppose that there exists a prefix set $\mathcal{Y} = \{y_1, \ldots, y_m\}$ such that $\psi(y_i) = \psi(u_i v_i)$, $1 \leq i \leq m$. Then there exist words $z_1, \ldots, z_n$ such that $\psi(z_i) = \psi(w_i)$, $1 \leq i \leq n$ and $L(G)$ is commutatively equivalent to the regular set $\mathcal{Y}^*\{z_1, \ldots, z_n\}$.*

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 10. *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.*

A natural problem arising from the previous results is to figure out which minimal linear grammars satisfy the hypotheses of Theorem 9 and Corollary 10. Indeed, these grammars generate commutatively regular languages. In view of the latter 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.

5.1 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 11. *Let G be an unambiguous minimal linear grammar with the productions (4). For any Bernoulli distribution μ, one has $\sum_{i=1}^{m} \mu(u_i v_i) \leq 1$.*

Now we give a characterization of unambiguous minimal linear grammars.

Proposition 12. *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*

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

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 13. *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*

$$\mu\left(\bigcup_{i=0}^{k} L_i\right) = \sum_{i=0}^{k} \left(\mu(L_1)\right)^i. \tag{5}$$

By expressing the sum in the right side of (5) in term of the rational function $(1 - x^{k+1})/(1 - x)$ one gets

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

that, for $\mu(L_1) = 1$, by continuity, will take the value $k + 1$.

Now, we present an application of the result above. Let k be a positive integer. Generalizing a result of [3, Example 6.3], one can show that a subset L of $(a^*b)^k a^*$ is commutatively prefix if and only if its growth function g_L satisfies the inequality $g_L(n) \leq \binom{n}{k}$ for all $n \geq k$. Thus, as an immediate consequence of Corollary 10, we get:

Corollary 14. *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 $g_{L_k}(n) \leq \binom{n}{k}$, $n \geq k$, then $L(G)$ is commutatively regular.*

We recall that a set L of words is said a *Bernoulli set* if, for every Bernoulli distribution μ, $\mu(L) = 1$. In [12], 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 10 and Corollary 13 we get the following.

Corollary 15. *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.*

The problem whether every finite and complete 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 [3, 12]). The conjecture in this formulation has been disproved by Shor [22]. 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 [12], a positive answer to the latter question has been conjectured in the case of finite complete codes.

We close the paper with an open question. A theorem proven in [4] (see also [12]) 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 characterization of the property of codicity in term of the notion of completeness and a measure of the set. It would be interesting to get a similar characterization, of the property of unambiguity for minimal linear grammars.

References

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

2. G. Baron, W. Kuich, The Characterization of Nonexpansive Grammars by Rational Power Series, *Information and Control* **48**, 109–118 (1981).
3. J. Berstel, D. Perrin, C. Reutenauer, *Codes and Automata*, Encyclopedia of Mathematics and its Applications No. 129, Cambridge University Press, Cambridge, (2009).
4. J. M. Boë, A. de Luca, A. Restivo, Minimal complete sets of words, Theoret. Comput. Sci. **12**, 325–332, (1980).
5. M. R. Bridson, R. H. Gilman, Context-free languages of sub-exponential growth, J. Comput. System Sci. **64**, 308–310, (1999).
6. N. Chomsky, M. -P. Schützenberger, The Algebraic Theory of Context-free Languages, in P. Braffort and D. Hirschberg (eds.), "Computer Programming and Formal Systems", pp. 118–161, North Holland Publishing Company, Amsterdam, (1963).
7. F. D'Alessandro, B. Intrigila, and S. Varricchio, The Parikh counting functions of sparse context-free languages are quasi-polynomials, Theoret. Comput. Sci. **410**, 5158–5181 (2009).
8. F. D'Alessandro, B. Intrigila, and S. Varricchio, Quasi-polynomials, linear Diophantine equations and semi-linear sets, Theoret. Comput. Sci. **416**, 1–16 (2012).
9. 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).
10. F. D'Alessandro, B. Intrigila, On the commutative equivalence of semi-linear sets of $\mathbb{N}^k$, Theoret. Comput. Sci. **562**, 476–495 (2015).
11. 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).
12. A. de Luca, Some combinatorial results on Bernoulli sets and codes, Theoret. Comput. Sci. **273**, 143–165 (2002).
13. P. Flajolet, Analytic models and ambiguity of context-free languages, Theoret. Comput. Sci. **49**, 283–309 (1987).
14. S. Ginsburg E. H. Spanier, Semigroups, Presburger formulas, and languages, Pacific J. Math., **16**, 285–296 (1966).
15. S. Ginsburg, *The mathematical theory of context-free languages*, Mc Graw- Hill, New York, (1966).
16. O. H. Ibarra, B. Ravikumar, On sparseness, ambiguity and other decision problems for acceptors and transducers, LNCS, vol. 210, pp. 171–179, Springer-Verlag, Berlin, (1986).
17. O. H. Ibarra, B. Ravikumar, On bounded languages and reversal-bounded automata, Inf. Comput. **246**, 30–42 (2016).
18. L. Ilie, G. Rozenberg, A. Salomaa, A characterization of poly-slender context-free languages, RAIRO Inform. Théor. Appl. **34**, 77–86 (2000).
19. R. Incitti, The growth function of context-free languages, Theoret. Comput. Sci. **255**, 601–605 (2001).
20. M. Latteux, G. Thierrin, On bounded context-free languages, Elektron. Inform. Verarb. u. Kybern. **20**, 3–8 (1984).
21. A. Restivo, A characterization of bounded regular sets, LNCS, vol. 33, pp. 239–244, Springer-Verlag, Berlin, (1975).
22. P. W. Shor, A counterexample to the triangle conjecture, J. Combin. theory Ser. A., **38**, 110–112 (1983).