Theoretical Computer Science 562 (2015) 304–319

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the commutative equivalence of bounded context-free and regular languages: The code case

Flavio D'Alessandro [a,b,*], Benedetto Intrigila [c]

[a] Dipartimento di Matematica, Università di Roma "La Sapienza", Piazzale Aldo Moro 2, 00185 Roma, Italy
[b] Department of Mathematics, Boğaziçi University, 34342 Bebek, Istanbul, Turkey
[c] Dipartimento di Ingegneria dell'Impresa, Università di Roma "Tor Vergata", via del Politecnico, 1, 00133 Roma, Italy

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form 23 September 2014
Accepted 1 October 2014
Available online 8 October 2014
Communicated by G. Ausiello

Keywords:
Bounded context-free language
Semilinear set
Commutative equivalence

ABSTRACT

This is the first paper of a group of three where we prove the following result. Let A be an alphabet of t letters and let $\psi : A^* \longrightarrow \mathbb{N}^t$ be the corresponding Parikh morphism. Given two languages $L_1, L_2 \subseteq A^*$, we say that L_1 is commutatively equivalent to L_2 if there exists a bijection $f : L_1 \longrightarrow L_2$ from L_1 onto L_2 such that, for every $u \in L_1$, $\psi(u) = \psi(f(u))$. Then every bounded context-free language is commutatively equivalent to a regular language.

1. Introduction

In this paper, we study an algebraic and combinatorial problem concerning bounded context-free languages. A strictly related notion is that of *sparse language*: a language L is termed *sparse* if its counting function, that is, the function f_L that maps every integer $n \geq 0$ into the number $f_L(n)$ of words of L of length n, is polynomially upper bounded. Sparse languages play a meaningful role both in Computer Science and in Mathematics and have been widely investigated in the past. The interest in this class of languages is due to the fact that, in the context-free case, it coincides with the one of *bounded languages*. A language L is termed *bounded* if there exist k words $u_1, \ldots, u_k$ such that $L \subseteq u_1^* \cdots u_k^*$ ([3,5,6,14,16–20,22,23,25]; see also [11] for an excellent survey on the relationships between bounded languages and monoids of polynomial growth).

The starting point of this investigation is the following result proved in [5]: for every sparse context-free language L_1, there exists a regular language L_2 such that $f_{L_1} = f_{L_2}$.

Therefore, the counting function of a sparse context-free language is always rational. This result is interesting since, as it is well known [13], the counting function of a context-free language may be transcendental.

A conceptual limit of the above-mentioned construction is that the regular language L_2 is on a different alphabet from the one of L_1. Therefore it is natural to ask whether this limitation can be removed. It is worth noticing that, from a general point of view, imposing restrictions on the alphabets makes some classical constructions more difficult but obviously more informative. In this context, as a related result, we can mention a remarkable contribution by Beal and Perrin where the problem of the length equivalence of regular languages on alphabets of prescribed size is considered [1].

Let us now describe the problem we are interested in. For this purpose, some preliminary notions are needed. Let $A = \{a_1, \ldots, a_t\}$ be an alphabet of t letters and let $\psi : A^* \longrightarrow \mathbb{N}^t$ be the corresponding Parikh morphism. Given two languages

* Corresponding author.

http://dx.doi.org/10.1016/j.tcs.2014.10.005
0304-3975/

L_1 and L_2 over the alphabet A, we say that L_1 is *commutatively equivalent to* L_2 if there exists a bijection $f : L_1 \longrightarrow L_2$ from L_1 onto L_2 such that, for every $u \in L_1$, $\psi(u) = \psi(f(u))$. This notion is important in the theory of variable-length codes since it is involved in the celebrated Schützenberger conjecture about the commutative equivalence of a maximal finite code with a prefix one (see e.g. [2]).

Now the general problem above can be formulated as follows: *given a bounded context-free language L_1, does it exist a regular language L_2 which is commutatively equivalent to L_1?*
In the sequel, for short, we refer to it as *CE (Commutative Equivalence) Problem*.

This is the first paper of a group of three (see also [9,10]) where we prove the following statement that solves in the affirmative the CE Problem.

Theorem 1. *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.*

Theorem 1 with a sketch of the proof was announced in [7]. Actually we prove that the CE problem can be solved in the affirmative for the wider class of bounded semi-linear languages. Moreover the use of such languages turns out to be of crucial importance in the solution of the problem as it makes possible to handle the languages through suitable sets of vectors of integers, namely semi-linear sets of the free commutative monoid $\mathbb{N}^k$.

Observe that the CE Problem can be seen as a kind of counting problem in the class of bounded context-free languages. Despite the fact that such class has been widely investigated in the past, CE Problem appears to require some new techniques. Indeed, bounded context-free languages can be inherently ambiguous; in addition, if $u_1^* \cdots u_k^*$, $u_i \in A^+$, is the set that contains the bounded context-free language, then, in general, $u_1^* \cdots u_k^*$ is ambiguous as product of languages of A^*. Such ambiguities, which are of different nature, interfere making the study of the problem a non-trivial task.

Before explaining the main contribution of this paper, we would like to give a broader picture about the relationships between CE Problem and some classical theorems on bounded context-free languages. The first result that deserves to be mentioned is a well-known theorem by Parikh [24]. For this purpose, let us first introduce a notion. Given two languages L_1 and L_2 over the alphabet A, we say that L_1 is *Parikh equivalent to* L_2 if $\psi(L_1) = \psi(L_2)$. The theorem by Parikh states that, given a context-free language L_1, its image $\psi(L_1)$ under the Parikh map is a semi-linear set of $\mathbb{N}^t$. As a straightforward consequence of Parikh theorem, one has that there exists a regular language L_2 which is *Parikh equivalent to* L_1.

It is worth noticing that the property of Parikh equivalence by no means implies the property of commutative equivalence. Indeed, let $A = \{a, b\}$ and let $L_1 = (ab)^* \cup (ba)^*$ and $L_2 = (ab)^*$. One has that $\psi(L_1) = \psi(L_2) = (1, 1)^{\oplus}$ (the symbol $\oplus$ denotes the Kleene star operation in the monoid $\mathbb{N}^2$) so that L_1 is Parikh equivalent to L_2. On the other hand, one immediately checks that L_1 cannot be commutatively equivalent to L_2.

Another theorem that is central in this context has been proved by Ginsburg and Spanier [15]. For this purpose, let us first introduce a notion. Let $L \subseteq u_1^* \cdots u_k^*$ be a bounded language where, for every $i = 1, \ldots, k$, u_i is a word over the alphabet A. Let $\varphi : \mathbb{N}^k \longrightarrow u_1^* \cdots u_k^*$ be the map defined as: for every tuple $(\ell_1, \ldots, \ell_k) \in \mathbb{N}^k$,

$$\varphi(\ell_1, \ldots, \ell_k) = u_1^{\ell_1} \cdots u_k^{\ell_k}.$$

The map φ is called the *Ginsburg map*. Ginsburg and Spanier proved that L is context-free if and only if $\varphi^{-1}(L)$ is a finite union of linear sets, each having a stratified set of periods. Roughly speaking, a stratified set of periods corresponds to a system of well-formed parentheses. However, Ginsburg and Spanier theorem is of no help to study counting problems and, in particular, our problem, because of the ambiguity of the representation of such languages. Indeed, let $A = \{a, b, c\}$ be a three letter-alphabet and let the language $L = \{a^i b^j c^k \mid i, j, k \in \mathbb{N},\ i = j \text{ or } j = k\}$ [4]. Since L is inherently ambiguous, by [14] Theorem 6.2.1, L cannot be represented unambiguously as a finite disjoint union of a stratified set of periods. In this context, another important recent result that gives a characterization of bounded context-free languages in terms of finite unions of *Dyck loops* has been proven in [20]. However, neither this latter result can be used to deal with our problem because of the ambiguity of the representation of such languages as a finite union of Dyck loops.

The proof of Theorem 1 will be presented in its complete generality in [10]. In particular, we will prove Theorem 1, by using a refined version of the techniques of this paper together with another result proved in [9].

The goal of this paper is to describe some key elements of our technique. We will do this by proving Theorem 1 under the following assumption.

Let $L = \varphi(B)$ be a language described, via the Ginsburg map, by a semi-simple set B:

$$B = \bigcup_{i=0}^{n} B_i, \quad n \geq 1,$$

where B_0 is a finite set of vectors and, for every $i = 1, \ldots, n$, B_i is a simple set:

$$B_i = \mathbf{b}_0^{(i)} + \{\mathbf{b}_1^{(i)}, \ldots, \mathbf{b}_{k_i}^{(i)}\}^{\oplus},$$

where $k_i > 0$ is the dimension of B_i and the vectors $\mathbf{b}_0^{(i)}, \mathbf{b}_1^{(i)}, \ldots, \mathbf{b}_{k_i}^{(i)}$, form the unambiguous representation of B_i.