On the structure of context-free languages

Flavio D'Alessandro[1,2]

Dipartimento di Matematica "G. Castelnuovo"
La Sapienza Università di Roma
00185 Roma, Italy,

Benedetto Intrigila [3]

Dipartimento di Ingegneria dell'Impresa
Università di Roma "Tor Vergata"
00133 Roma, Italy

Abstract

We discuss the following result. 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 $u \in L_1$, $f(u)$ is obtained from u by a permutation of the letters of u. Then every bounded context-free language is commutatively equivalent to a regular language.

Keywords: Bounded context-free language; Semilinear set; Commutative equivalence

1 Main contribution

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, that is, the function that maps every word u into the vector $(|u|_{a_1}, \ldots, |u|_{a_t})$, where, for every $i = 1, \ldots, t$, $|u|_{a_i}$ is the number of occurrences of the symbol a_i in u.

Given two languages 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 $u \in L_1$, $\psi(u) = \psi(f(u))$. A language L is termed *bounded* if there exist non-empty words $u_1, \ldots, u_k$, with $k \geq 1$, such that $L \subseteq u_1^* \cdots u_k^*$. In the group of three papers [9,10,11], the proof of the following theorem is provided.

[1] This work was partially supported by MIUR project PRIN 2010/2011 "Automi e Linguaggi Formali: Aspetti Matematici e Applicativi".

[2] Email: dalessan@mat.uniroma1.it

[3] Email: intrigil@mat.uniroma2.it.

This paper is electronically published in
Electronic Notes in Theoretical Computer Science
URL: www.elsevier.nl/locate/entcs

Theorem 1.1 *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.*

Actually we prove our result for the broader class of bounded semi-linear languages whose definition is postponed in Section 3 (see Definition 3.1).

Theorem 1.2 *Every bounded semi-linear 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 1.1 was announced in [7] with a sketch of the proof.

2 Connection with context-free languages

Theorem 1.1 naturally fits in the theory of bounded context-free languages developed by Ginsburg and Spanier. 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 ([3,5,6,15,16,17,18,19,20,23]; an excellent survey on the relations between bounded languages and monoids of polynomial growth can be found in [12]). In this framework, it is worth noticing the following immediate consequence of Theorem 1.1: for every sparse context-free language L_1, there exists a regular language L_2, over the same alphabet of L_1, 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 [14], the counting function of a context-free language may be transcendental. An immediate consequence of the latter is that Theorem 1.1 does not hold for an arbitrary context-free language. Indeed, if such a language were commutatively equivalent to a regular one, then its counting function would be rational. On the other hand, there exist context-free languages whose generating functions are algebraic not rational, as for instance in the case of the Dyck languages, and, even, transcendental, as said before. In this context, as a related result, we 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]. We would like to give a broader picture about the relations between our contribution and some classical theorems on bounded context-free languages. The first result that deserves to be mentioned is a well-known theorem by Parikh [21]. 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 straighforward 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 non-empty 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 [16] 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 [19]. 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. In addition to the inherent ambiguity of context-free languages, one should observe that the set $u_1^* \cdots u_k^*$ is, in general, ambiguous as a product of languages of A^*. Such two ambiguities, which are of different nature, interfere making a non trivial task the construction of a regular language that provides the solution of the problem.

Theorem 1.2 is so surprising that it could look implausible. We have therefore decided to provide a proof of the statement, as robust as possible, writing it in all the details. The proof is long and, in order to make its reading easier, it is splitted in three papers [9,10,11]. The first one deals a special but meaningful issue of the problem: it has been written to help the reader to capture some key elements of our technique. The second and the third papers describe the solution in the full generality. In particular, the second paper shows a result on rational sets of vectors of $\mathbb{N}^k$ which is in our opinion interesting in its own right.

3 An overview of the solution

We now give a short description of some key aspects of the proof of Theorem 1.2. Let us first recall the definition of bounded semi-linear language.

Definition 3.1 A bounded language $L \subseteq u_1^* \cdots u_k^*$ is said to be *semi-linear* if there exists a set B of $\mathbb{N}^k$ such that $\varphi(B) = L$ and $B = \bigcup_{i=0}^{n} B_i, n \geq 1$, where B_0 is a finite set of vectors and, for every $i = 1, \ldots, n$, B_i is a set of dimension $k_i > 0$:

$$B_i = \mathbf{b}_0^{(i)} + \{\mathbf{b}_1^{(i)}, \ldots, \mathbf{b}_{k_i}^{(i)}\}^{\oplus}, \tag{1}$$

where $\mathbf{b}_0^{(i)}, \mathbf{b}_1^{(i)}, \ldots, \mathbf{b}_{k_i}^{(i)}$, are vectors of $\mathbb{N}^k$.

By a well-known theorem of Eilenberg and Schützenberger [13], we can assume that, in every set B_i of (1), the vectors $\mathbf{b}_1^{(i)}, \ldots, \mathbf{b}_{k_i}^{(i)}$, are linearly independent in $\mathbb{Q}^k$ and, moreover, that the sets B_i's are pairwise disjoint. The sets B_i's satisfying these last two properties are called *simple*. By a result of Honkala [17], we can always construct a semi-linear set B of $\mathbb{N}^k$ such that $\varphi(B) = L$ and φ is injective on B.

The proof of Theorem 1.2 is essentially based upon two main tools:

- **The first tool** has been conceived to prove the theorem under the assumption that, for every $i = 1, \ldots, n$ and for every $j = 1, \ldots, k_i$, the word $\varphi(\mathbf{b}_j^{(i)})$, that represents via φ the vector $\mathbf{b}_j^{(i)}$ of (1), contains at least two distinct letters.

  From an intuitive point of view, the idea underlying this tool is the following. First, by using a technique of geometrical nature (inspired to our work [8]), we provide a new decomposition of B into simple sets, every one of each fulfills the following property: the words that represent, *via* the Ginsburg map φ, the generators of such simple sets are sufficiently long.

  Afterwards, by using techniques of Combinatorics on words [2], we codify the latter words with words of a unique factorization code. This allows us to cope with the ambiguity (as a product of languages) of the set $u_1^* \cdots u_k^*$ and to construct the regular language which is commutatively equivalent to the context-free one.
- **The second tool** provides the solution of Theorem 1.2 in the opposite case, that is, under the assumption that there exists a letter $a \in A$ such that, for every $i = 1, \ldots, n$, all the words $\varphi(\mathbf{b}_1^{(i)}), \ldots, \varphi(\mathbf{b}_{k_i}^{(i)})$ are powers of a.

  We treat such last case by reducing the study of commutative equivalence for languages to that of the *commutative equivalence for semi-linear sets of vectors*. More precisely, given two subsets S_1, S_2 of $\mathbb{N}^k$, we say that S_1 is commutatively equivalent to S_2 if there exists a bijection $f : S_1 \longrightarrow S_2$ from S_1 onto S_2 such that, for every $\mathbf{v} \in S_1$, $|\mathbf{v}| = |f(\mathbf{v})|$, where $|\mathbf{v}|$ denotes the sum of the components of $\mathbf{v}$. In [10] we prove that every semi-linear set of $\mathbb{N}^k$ is commutatively equivalent to a subset which is recognizable in $\mathbb{N}^k$ in the classical sense of Elgot and Mezei.
- By eventually combining the two tools, we then provide the proof of Theorem 1.2, by treating all the other intermediate cases, that is, the cases where the words, representing, *via* the Ginsburg map φ, the generators of the simple sets of (1) either contain more than two letters or are powers of a given letter.

4 Open problems

Finally, we mention an open problem pointed out in [22]. Also in the theory of formal series there is a notion of commutative equivalence (see [2], Ch. 14). Given a $\mathbb{N}$-series $\sigma \in \mathbb{N}\langle\langle A \rangle\rangle$ over an alphabet A of non-commutative variables, the *commutative image* of σ is the $\mathbb{N}$-series $\psi(\sigma) \in \mathbb{N}[[A]]$ over the commutative variables $\psi(A)$ defined as: for every $u \in \psi(A^*)$, $(\psi(\sigma), u) = \sum_{\psi(w)=u} (\sigma, w)$. Given two series $\sigma_1, \sigma_2 \in \mathbb{N}\langle\langle A \rangle\rangle$, we say that σ_1 is *commutatively equivalent to* σ_2 if they have the same commutative image. In this context, as a possible extension of Theorem 1.1, one can ask whether every $\mathbb{N}$-algebraic series with bounded support is commutatively equivalent to a rational one. Theorem 1.1 is a first step of the study of this problem.

References

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

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

[3] L. Boasson, A. Restivo, Une Caractérisation des Langages Algébriques Bornés, ITA **11**, 203–205 (1977).

[4] 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).

[5] F. D'Alessandro, B. Intrigila, S. Varricchio, On the structure of the counting function of context-free languages, Theoret. Comput. Sci. **356**, 104–117 (2006).

[6] 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).

[7] F. D'Alessandro, B. Intrigila, The commutative equivalence of bounded context-free and regular languages, in International Conference on words and formal languages, Words 2011, Electronic Proceedings in Theoretical Computer Science, p. 1-21, doi: 10.4204/EPTCS.63 (2011).

[8] F. D'Alessandro, B. Intrigila, S. Varricchio, Quasi-polynomials, Semi-linear set, and Linear Diophantine equations, 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, S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages,* Springer–Verlag, Berlin, (1999).

[13] S. Eilenberg, M. -P. Schützenberger, Rational sets in commutative monoids, J. of Algebra **13**, 173–191 (1969).

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

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

[16] S. Ginsburg, *The mathematical theory of context-free languages*, Mc Graw- Hill, New York, (1966).

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

[18] O. Ibarra, B. Ravikumar, On sparseness, ambiguity and other decision problems for acceptors and transducers, LNCS, vol. 210, pp. 171–179, Springer-Verlag, Berlin, (1986).

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

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

[21] R. J. Parikh, On context-free languages, J. ACM **13**, 570–581 (1966).

[22] Perrin D., *private communication,* (2014).

[23] A. Restivo, A characterization of bounded regular sets, LNCS, vol. 33, pp. 239–244, Springer-Verlag, Berlin, (1975).