Doctopic: Algorithms, automata, complexity and games ARTICLE IN PRESS TCS:10922

Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Unavoidable sets and circular splicing languages ☆

Clelia De Felice *, Rocco Zaccagnino, Rosalba Zizza

Università degli Studi di Salerno, Italy

ARTICLE INFO

Article history:
Received 15 November 2015
Received in revised form 8 September 2016
Accepted 11 September 2016
Available online xxxx

Keywords:
Regular languages
Circular splicing systems
Unavoidable sets

ABSTRACT

Circular splicing systems are a formal model of a generative mechanism of circular words, inspired by a recombinant behaviour of circular DNA. They are defined by a finite alphabet A, an initial set I of circular words, and a set R of rules. In this paper, we focus on the still unknown relations between regular languages and circular splicing systems with a finite initial set and a finite set R of rules represented by a pair of letters ($(1,3)$-CSSH systems). When $R = A \times A$, it is known that the set of all words corresponding to the splicing language belongs to the class of pure unitary languages, introduced by Ehrenfeucht, Haussler, Rozenberg in 1983. They also provided a characterization of the regular pure unitary languages, based on the notions of unavoidable sets and well quasi-orders. We partially extend these notions and their results in the more general framework of the $(1,3)$-CSSH systems.

1. Introduction

In this paper we deal with connections between *unavoidable sets* and regularity of languages generated by *circular splicing systems*, continuing a research initiated in [3,11].

The *circular splicing operation* is a language-theoretic word operation introduced by Head in [15] which models a DNA recombination process on two circular DNA molecules by means of a pair of restriction enzymes.

A string of circular DNA can be represented by a circular word, i.e., a sequence of letters written on a circle (see Section 2 for the definition). Thus a circular word $^{\sim}w$ corresponds to the set of all cyclic permutations of w. A circular language C is a set of circular words. Its full linearization $\mathrm{Lin}(C)$ is the set of all cyclic permutations of its elements. C is regular (resp. context-free, context-sensitive) if $\mathrm{Lin}(C)$ is regular (resp. context-free, context-sensitive).

We deal with one of the several existing variants of the circular splicing operation, given in [16] and called the Păun circular splicing operation, since it is the counterpart of the Păun linear splicing operation in the circular context. Given a splicing rule r, the idea of this operation is to cut two circular words at specific sites and then to join appropriately the pieces, thus yielding a new circular word. A rule is represented as a quadruple of words $r = u_1\#u_2\$u_3\#u_4$, which specifies where to cut and how to join. A Păun circular splicing system is a triple $S = (A, I, R)$ where A is a finite alphabet, I is the *initial* circular language, and R is the set of *rules*. Both I, R will be supposed to be finite sets. The *circular language* generated by a circular splicing system S (splicing language) is the smallest language which contains I and is invariant under iterated splicing by rules in R.

☆ Partially supported by the *FARB* Project *"Codici e sistemi splicing nella teoria dei linguaggi formali"* (University of Salerno, 2014) and the *FARB* Project *"Aspetti matematici e applicativi nella teoria dei codici e linguaggi formali"* (University of Salerno, 2015).

* Corresponding author.
E-mail address: defelice@dia.unisa.it (C. De Felice).

http://dx.doi.org/10.1016/j.tcs.2016.09.008

Please cite this article in press as: C. De Felice et al., Unavoidable sets and circular splicing languages, Theoret. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.tcs.2016.09.008

The main results on the computational power of such systems will be discussed later in detail (see [16] for a more general discussion on this topic). They have been obtained in [1], first for a new variant of circular splicing, introduced in the same paper and named *flat splicing*, then extended to the classical model of Păun circular splicing systems.

We consider the still open problem of deciding whether a circular language generated by a circular splicing system is regular. This problem has been solved for some classes of splicing systems (*monotone complete systems* [3], *marked systems* [10]). Moreover, we can decide whether a splicing unary language is regular [5,6]. In this paper, we focus on this problem restricted to $(1,3)$-CSSH systems, i.e., Păun circular splicing systems such that, for each rule $u_1\#u_2\$u_3\#u_4$ in R, u_1 and u_3 are letters and u_2 and u_4 are the empty word. Therefore, R is a symmetric binary relation on A.

A famous Higman's Theorem states that the subword ordering over a finitely generated free monoid A^* is a well quasi-order (wqo) on A^* [9,20]. The subword ordering on A^* is the quasi-order where, for words u, v over A, $u \leq v$ if v can be obtained from u by inserting zero or more letters in u. A quasi-order is a wqo on a set X if, for each infinite sequence $\{x_i\}$ of elements in X, there exist $i < j$ such that $x_i \leq x_j$. Higman's Theorem has been subsequently extended in [12]. Loosely speaking, the authors considered insertions of words from a fixed finite set $Y \subseteq A^*$ instead of letters. They defined the quasi-order $\leq_Y$ as the reflexive and transitive closure of the relation $\{(uv, uyv) \mid y \in Y, u, v \in A^*\}$. They proved that $\leq_Y$ is a wqo if and only if Y is *unavoidable*, i.e., $A^* \setminus A^*YA^*$ is a finite set. This condition also characterizes regularity of the language $L_Y = \{w \in A^* \mid 1 \leq_Y w\}$, where 1 is the empty word. Roughly L_Y is the smallest set of words containing Y and invariant under the *iterated insertion* operation, defined in [13]. (The insertion operation was independently studied in another context in [18].)

It turns out that, when Y is closed under the conjugacy relation, the same holds for the language L_Y. Moreover the family of these languages L_Y coincides with the class of the full linearizations of the circular languages generated by complete splicing systems. Thus, regular circular languages generated by complete systems have been characterized in [3] by the above-mentioned result in [12].

In this paper, we consider a further generalization of this situation. We have a fixed finite set Y of words over a finite alphabet A and a symmetric relation $R \subseteq A \times A$. We introduce a generalization of the above operation, the *iterated R-insertion*. The idea of the R-insertion operation is that, given (a, b) in R, a word z ending with b may be inserted in a word containing a, after the letter a. We consider the language $L_{Y,R}$, defined as the smallest set of words containing Y and invariant under the iterated R-insertion operation. Of course $L_{Y,R}$ and L_Y agree when $R = A \times A$. We show that, once again, when Y is closed under the conjugacy relation the same holds for the language $L_{Y,R}$. Moreover, we prove that languages $\mathrm{Lin}(C)$, where C is generated by a $(1,3)$-CSSH system $S = (A, I, R)$, are exactly those languages $L_{Y,R}$ for which $Y = \mathrm{Lin}(I)$ is closed under conjugation. Therefore, the search for a characterization of regularity of languages generated by $(1,3)$-CSSH system is actually the search for a characterization of regularity of $L_{Y,R}$, hence a generalization of the above-mentioned result in [12]. In this paper we give partial results in this direction, described below.

Marked systems generating regular languages have been characterized by a property of the set of rules in [10]. As a main result of this paper, we prove that this property of the set of rules, along with strong R-unavoidability of the language $\mathrm{Lin}(I)$, ensures the regularity of the language generated by a $(1,3)$-CSSH system $S = (A, I, R)$. Of course, the notion of strong R-unavoidability extends the classical one. The results proved in this paper show that there are relations between wqo, unavoidability and regularity of languages generated by $(1,3)$-CSSH systems which are not thoroughly investigated.

This paper is organized as follows. Basics on words and splicing are collected in Section 2. In Section 3, we extend to the languages generated by $(1,3)$-CSSH systems the relation between insertion, circular splicing operation and flat splicing previously proved for complete systems in [2]. This result allows us to work on the full linearization of the circular splicing language instead of on the circular language, thus simplifying many proofs. In particular, it is of great help for stating results in Section 5, where we mimic another construction, given in [12], to alternatively define languages generated by $(1,3)$-CSSH systems. In turn, this alternative definition is needed for the proof of our main result. In [12] the above-mentioned construction is recursive and obtained by means of the star $*$ operation on languages. We generalize this construction by means of an extension $+_R$ of the operation $+$ (Section 4.2). We define X^{+_R} as the image by a substitution of a language generated by a special marked system, and this substitution is regular if X is regular (Section 4). We introduce our notions of R-unavoidability and strong R-unavoidability in Section 6. We prove our main result in Section 7. There are several issues that follow from the results stated in this paper, they will be discussed in Section 8.

2. Basics

2.1. Words and circular words

We suppose the reader is familiar with classical notions in formal languages [14,17,20]. We denote by A^* the free monoid over a finite alphabet A and we set $A^+ = A^* \setminus 1$, where 1 is the empty word. For a word $w \in A^*$ and a letter $a \in A$, $|w|_a$ is the number of the occurrences of a in w, $|w|$ is the length of w, and $\mathrm{alph}(w) = \{a \in A \mid |w|_a > 0\}$. A word $x \in A^*$ is a *factor* of $w \in A^*$ if there are $u_1, u_2 \in A^*$ such that $w = u_1 x u_2$. If $u_1 = 1$ then x is a *prefix* of w. If $u_1 \neq 1$ and $u_2 \neq 1$, then x is an *internal factor* of w. A language is *regular* if it is recognized by a finite automaton. Let A, B be finite alphabets. A substitution ϕ from B^* into A^* is a (monoid) morphism from B^* into the powerset $\mathfrak{P}(A^*)$ of A^*. It is called *regular* if $\phi(b)$ is a regular language for all $b \in B$. Regular languages are closed under regular substitution [14]. Moreover, for any language X, we set

Please cite this article in press as: C. De Felice et al., Unavoidable sets and circular splicing languages, Theoret. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.tcs.2016.09.008