The Version of Record is available online at https://doi.org/10.1016/j.tcs.2016.09.008

Unavoidable sets and regularity of languages generated by (1,3)-circular splicing systems ☆

Clelia De Felice, Rosalba Zizza, Rocco Zaccagnino
Università degli Studi di Salerno

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.

Keywords: Regular languages, circular splicing systems, unavoidable sets

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 [4, 12].

The *circular splicing operation* is a language-theoretic word operation introduced by Head in [17] which models a DNA recombination process on

☆Partially supported by the *FARB* Project *"Codici e sistemi splicing nella teoria dei linguaggi formali"* (University of Salerno, 2014), the *FARB* Project *"Aspetti matematici e applicativi nella teoria dei codici e linguaggi formali"* (University of Salerno, 2015), and the *MIUR* PRIN 2010-2011 grant *H41J12000190001* *"Automata and Formal Languages: Mathematical and Applicative Aspects"*.

two circular DNA molecules by means of a pair of restriction enzymes. For instance, circular splicing models the integration of a plasmid into the DNA of a host bacterium (see [14] for an overview on circular DNA in Nature).

Obviously a string of circular DNA can be represented by a circular word, i.e., by an equivalence class with respect to the conjugacy relation $\sim$, defined by $xy \sim yx$, for $x, y \in A^*$ [22]. The set of all strings equivalent to a given word w is the full linearization of the circular word ${}^{\sim}w$. A circular language C is a set of circular words. It is regular (resp. context-free, context-sensitive) if so is its full linearization $\mathrm{Lin}(C)$, i.e., the union of the full linearizations of its elements.

We deal with one of the several existing variants of the circular splicing operation, named here Păun definition. Correspondingly, 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 r, represented as quadruples of words $r = u_1\#u_2\$u_3\#u_4$ [18]. 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. The main results on the computational power of such systems will be discussed later in detail. They have been obtained in [2], first for a new variant of circular splicing, introduced in the same paper and named *flat splicing*, then easily extended to the classical model.

In this paper, we focus on $(1, 3)$-CSSH systems. *Păun circular semi-simple splicing systems* (or *CSSH systems*), previously considered in [8, 9, 27], are such that both u_1u_2, u_3u_4 have length one for any rule $u_1\#u_2\$u_3\#u_4$. A $(1, 3)$-CSSH system is a CSSH system such that $u_2 = u_4 = 1$. Therefore R is a symmetric binary relation on A. The following problem is still unsolved, even for $(1, 3)$-CSSH systems.

Problem 1.1 *Given a circular splicing system $S = (A, I, R)$, where I, R are finite sets, can we decide whether the corresponding generated language is regular?*

The above question has been positively answered for unary languages [6, 7], for *(monotone) complete systems* [4], and for *marked systems* [11]. A $(1, 3)$-CSSH system $S = (A, I, R)$ is complete if $R = A \times A$ whereas S is marked if $I = A$.

Regular languages play a central role in Formal Language theory and admit several characterizations based on different concepts. In particular, regular languages can be characterized as the *upward closed sets* of monotone well quasi-orders on a finitely generated free monoid [13]. There exist different characterizations of the notion of a *well quasi-order (wqo)*. Following one of them, 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 $x_i \leq x_j$.

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^* [10, 19, 22, 23]. 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. This theorem has been subsequently extended in [13]. 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 $L_Y = \{w \in A^* \mid 1 \leq_Y w\}$. Roughly L_Y is the smallest set of words containing Y and invariant under the *iterated insertion* operation, defined in [15].

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 [4] by the above mentioned result in [13].

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*. 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 $L_{Y,R}$, with $Y = \mathrm{Lin}(I)$ closed under conjugation. Therefore, the search of a characterization of regularity of languages generated by $(1,3)$-CSSH system is actually the search of a characterization of regularity of $L_{Y,R}$, hence a generalization of the above mentioned result in [13]. 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 [11]. 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 briefly sketch the content of this paper. In Section 4, 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 [3]. Then in Section 6, we mimic another construction given in [13] to alternatively define languages generated by $(1,3)$-CSSH systems. The latter construction is recursive and obtained by means of a new operation introduced in Section 5. In the same Section 5, we also define special marked systems associated with languages generated by the intermediate steps of this construction. We introduce our notions of R-unavoidability and strong R-unavoidability in Section 7. We prove our main result in Section 8. Finally, in Section 9 we discuss future perspectives that follow on from the above results.

2. Basics

2.1. Words and circular words

We suppose the reader familiar with classical notions in formal languages [20, 22]. 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 $w \in A^*$, $|w|$ 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. A language is *regular* if it is recognized by a finite automaton. 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 [1]. Moreover, for any language X, we set $\mathrm{alph}(X) = \cup_{w \in X} \mathrm{alph}(w)$. A X-factorization of w of length n is any n-tuple $(x_1, \ldots, x_n)$ of elements of X such that $w = x_1 \cdots x_n$. Finally, for a word u in A^*, we set $u^{-1}X = \{v \in A^* \mid uv \in X\}$. If X is regular then so is $u^{-1}X$.

For a given word $w \in A^*$, a circular word $^\sim w$ is the equivalence class of w with respect to the *conjugacy* relation $\sim$ defined by $xy \sim yx$, for $x, y \in A^*$ [22]. The notations $|^\sim w|$ and $\mathrm{alph}(^\sim w)$ will be defined as $|w|$ and $\mathrm{alph}(w)$ for any representative w of $^\sim w$.

Let $^\sim A^*$ denote the set of all circular words over A, i.e., the quotient of A^* with respect to $\sim$. Given $L \subseteq A^*$, $^\sim L = \{^\sim w \mid w \in L\}$ is the *circularization* of L, i.e., the set of all circular words corresponding to elements of L. A subset C of $^\sim A^*$ is named a *circular language* and every subset L of A^* such that $^\sim L = C$ is called a *linearization* of C. In particular, a linearization of a circular word $^\sim w$ is a linearization of $\{^\sim w\}$, whereas the *full linearization* $\mathrm{Lin}(C)$ of C is the set of all the words in A^* corresponding to the elements of C, i.e., $\mathrm{Lin}(C) = \{w' \in A^* \mid \exists\ ^\sim w \in C :\ w' \sim w\}$.

Given a family of languages FA in the Chomsky hierarchy, $FA^\sim$ is the set of all those circular languages C which have some linearization in FA. In particular, $Reg^\sim$ is the class of circular languages having a regular linearization, i.e., $Reg^\sim = \{C \subseteq\ ^\sim A^* \mid \exists L \in Reg :\ ^\sim L = C\}$. If $C \in Reg^\sim$ then C is a *regular circular language*. Analogously, we can define context-free and context-sensitive circular languages. The *rotational closure* of language X, written $\mathrm{RC}(X) = \{yx \mid x, y \in A^* \text{ and } xy \in X\}$, is the set of all words in the conjugacy classes of the elements in X. It is known that the class of regular (resp. context-free, context-sensitive) languages is closed under rotational closure [20, 21, 26]. Consequently, a circular language C is regular (resp. context-free, context-sensitive) if and only if its full linearization $\mathrm{Lin}(C)$ is regular (resp. context-free, context-sensitive).

2.2. Circular and flat splicing

A *Păun circular splicing system* is a triple $S = (A, I, R)$, where A is a finite alphabet, I is the *initial* circular language, with $I \subseteq\ ^\sim A^*$, $I \neq \emptyset$, and R is the set of *rules*, with $R \subseteq A^*\#A^*\$A^*\#A^*$ and $\#, \$ \notin A$. Given a rule, $r = u_1\#u_2\$u_3\#u_4$ and circular words $^\sim w'$, $^\sim w''$, if there are linearizations w' of $^\sim w'$, w'' of $^\sim w''$ and words h, k, such that $w' = u_2hu_1$, $w'' = u_4ku_3$, then the result of the splicing operation applied to $^\sim w'$ and $^\sim w''$ by r is the circular word $^\sim w$ such that $w = u_2hu_1u_4ku_3$. Therefore, we set $(^\sim w', ^\sim w'')\vdash_r {}^\sim w$ and we say that $^\sim w$ is generated (or spliced) starting with $^\sim w'$, $^\sim w''$ and by using a rule r. The splicing operation is extended to circular languages in order to obtain the definition of circular splicing languages. Given a Păun circular splicing system S and a circular language $C \subseteq\ ^\sim A^*$, we set $\sigma'(C) = \{w \in\ ^\sim A^* \mid \exists w', w'' \in C, \exists r \in R :\ (w', w'')\vdash_r w\}$. We also define $\sigma^0(C) = C$,

$\sigma^{i+1}(C) = \sigma^i(C) \cup \sigma'(\sigma^i(C))$, $i \geq 0$, $\sigma^*(C) = \bigcup_{i \geq 0} \sigma^i(C)$. Then, $L(S) = \sigma^*(I)$ is the circular language *generated* by S. A circular language C is *Păun generated* (or C is a *circular splicing language*) if a Păun circular splicing system S exists such that $C = L(S)$.

In this paper R will always be a finite set. Moreover we focus on finite circular splicing systems. A circular splicing system is *finite* (resp. *regular, context-free, context-sensitive*) if its initial set is finite (resp. regular, context-free, context-sensitive). We suppose that I does not contain the empty word (adding the empty word to I will only add the empty word to $L(S)$ [2, 11]). Furthermore, as observed in [4], in order to find a characterization of the circular splicing languages, there is no loss of generality in assuming that the set R of the rules is *symmetric* (i.e., for each $u_1\#u_2\$u_3\#u_4 \in R$, we have $u_3\#u_4\$u_1\#u_2 \in R$). Thus, in what follows, we assume that R is symmetric. However, for simplicity, in the examples of Păun systems, only one of either $u_1\#u_2\$u_3\#u_4$ or $u_3\#u_4\$u_1\#u_2$ will be reported in the set of rules. It is known that the corresponding class of generated circular languages is not comparable with the class of regular circular languages [6, 24, 27] and it is contained in the class of context-sensitive circular languages [2].

In [2], the authors also proved that the splicing language is context-free if it is generated by an *alphabetic context-free splicing system* (i.e., a context-free splicing system such that in any rule $u_1\#u_2\$u_3\#u_4$, the words u_j are letters or the empty word). All the above mentioned results from [2] have been obtained first for a new variant of circular splicing, introduced in the same paper and named flat splicing, then easily extended to the classical model. This new variant allow us to separate operations on formal languages and grammars from the operation of circular closure (circularization).

A *flat splicing system* is a triplet $\mathcal{S} = (A, I, R)$, where A is an alphabet, I is a set of words over A, called the *initial set*, and R is a finite set of *splicing rules*, which are quadruplets $\langle \alpha | \gamma - \delta | \beta \rangle$ of words over A. The words α, β, γ and δ are called the *handles* of the rule.

Let $r = \langle \alpha | \gamma - \delta | \beta \rangle$ (or $\alpha\#\beta\$\delta\#\gamma$) be a splicing rule. Given two words $u = x\alpha \cdot \beta y$ and $v = \gamma z \delta$, applying r to the pair (u, v) yields the word $w = x\alpha \cdot \gamma z \delta \cdot \beta y$ (The dots are used only to mark the places of cutting and pasting, they are not parts of the words.) We denote this operation by $u, v \vdash_r w$. Note that the first word (here u) is always the one in which the second word (here v) is inserted. The *language generated* by the flat splicing system $\mathcal{S} = (A, I, R)$, written $L(\mathcal{S})$, is the smallest language containing I and closed by R.

A rule $r = \langle \alpha | \gamma - \delta | \beta \rangle$ is *alphabetic* if its four handles α, β, γ and δ are letters or the empty word. A flat splicing system is alphabetic if all its rules are alphabetic. In [2], the authors introduced a suitable "normal form" for a flat splicing system, named flat heterogeneous splicing system. They proved that, for any alphabetic circular splicing system S we may always find a flat heterogeneous splicing system $\mathcal{S}$ such that $\mathrm{Lin}(S) = L(\mathcal{S})$. Finally, they stated the following result.

Theorem 2.1 *[2] The language generated by a flat or circular alphabetic context-free splicing system is context-free.*

In this framework, the following still open questions may be asked.

Problem 2.2 *Given a splicing system, can we decide whether the corresponding generated language is regular?*

Problem 2.3 *Given a regular circular language, can we decide whether it is a splicing language?*

Problem 2.4 *Can we characterize the structure of the regular circular languages which are splicing languages?*

Problem 2.3 has been solved for alphabetic splicing systems in [2], along with a similar question for general systems. Moreover, the above problems have been solved for unary languages [6, 7]. In this paper, we tackle Problem 2.2 for a special class of alphabetic splicing systems, namely $(1,3)$-CSSH systems.

Definition 2.5 *[8, 9, 27]. A circular splicing system $S = (A, I, R)$ is a Păun circular semi-simple splicing system (or CSSH system) if S is finite and, for any rule $u_1 \# u_2 \$ u_3 \# u_4$ in R, we have $|u_1 u_2| = |u_3 u_4| = 1$. A $(1,3)$-CSSH system is a CSSH system such that $u_2 = u_4 = 1$.*

Let $S = (A, I, R)$ be a $(1,3)$-CSSH system. By Theorem 2.1, $L(S)$ is a context-free language. From now on, we will adopt the simpler notation (a_i, a_j) for the rule $a_i \# 1 \$ a_j \# 1$. Moreover, we suppose that $\mathrm{alph}(R) = \{a_i \mid (a_i, a_j) \in R\} \subseteq \mathrm{alph}(I) = A$ and $\mathrm{alph}(w) \cap \mathrm{alph}(R) \neq \emptyset$, for any $w \in I$. Indeed, omitting rules or circular words in I which do not intervene in the application of the splicing operation, will not change the language generated by a CSSH system, beyond the finite set of words removed from I. This result was incorrectly stated for Păun circular splicing systems in [11] but it is not difficult to see that it holds for CSSH systems.

Definition 2.6 *[4, 11] A* $(1,3)$*-CSSH system* $S = (A, I, R)$ *is* complete *if* $R = A \times A$*. A* $(1,3)$*-CSSH system* $S = (A, I, R)$ *is* marked *if* $I = A$*.*

Problems 2.2–2.4 have been solved for marked systems in [11]. A characterization of languages generated by marked systems will be recalled in Section 6.4. In Section 1 we mentioned the characterization of regular circular languages generated by complete systems. More generally, in [4] it has been proved that unavoidability of $\mathrm{Lin}(I)$ characterizes *monotone complete* systems $S = (A, I, R)$ generating regular circular languages, thus answering to Problem 2.2 (a monotone complete system is a CSSH system such that for two fixed integers i, j, with $1 \leq i < j \leq 4$, one has $u_i = u_j = 1$ in any rule $u_1\#u_2\$u_3\#u_4$).

3. Outline of the results

We briefly sketch the content of this paper. In [3], the connection between alphabetic circular and flat splicing systems, stated in [2] and mentioned in Section 2, has been simplified for complete systems. The full linearizations of the corresponding splicing languages have also been characterized through the insertion operation given in [15]. For languages $Z, Y \subseteq A^*$, the result of the *insertion* operation applied to Z, Y is the language $Z \leftarrow Y = \{z_1 y z_2 \mid z_1 z_2 \in Z \text{ and } y \in Y\}$. The result of the *iterated insertion* operation applied to Y, is the language $Y^{\leftarrow_*} = \cup_{i \geq 0} Y^{\leftarrow_i}$, where we inductively define $Y^{\leftarrow_0} = \{1\}$, $Y^{\leftarrow_{i+1}} = Y^{\leftarrow_i} \leftarrow Y$, for $i \geq 0$. As stated in [15], $Y^{\leftarrow_*} = L_Y = \{w \in A^* \mid 1 \leq_Y w\}$, where the quasi-order $\leq_Y$ is the reflexive and transitive closure of the relation $\{(uv, uyv) \mid y \in Y, u, v \in A^*\}$.

In Section 4, we investigate further in this direction. We introduce a generalization of the above operations, named *R-insertion* and *iterated R-insertion.* On the other hand, we know that, given a $(1,3)$-CSSH system S, there is a flat splicing system $\mathcal{S}$ such that $\mathrm{Lin}(S) = L(\mathcal{S})$ [3]. Then we show that $\mathrm{Lin}(S)$ may be alternatively defined through the iterated R-insertion.

This result allow us to work on the full linearization of the circular splicing language instead on circular languages, thus to simplify many proofs. In particular, it is of great help for stating another characterization of $\mathrm{Lin}(S)$, through a construction given in Section 6, which in turn is needed for the proof of our main result.

Regarding this second construction, we recall that a characterization of the quasi-orders $\leq_Y$ which are wqo has been given in [13]. Their proof uses a recursive construction of sets I_n, obtained starting with a finite set X and by using the star $*$ and the concatenation $\cdot$ operations on languages. In Section 6, we obtain another equivalent definition of $\mathrm{Lin}(S)$ through an extension of the above operations $*$, $\cdot$ and of sets I_n.

In our context, these operations may be extended in several ways. Our extension $*_R$ of the $*$ operation is obviously based on an extension $\cdot_R$ of the concatenation operation. Both extensions will be defined in Section 5. Loosely speaking the $\cdot_R$ operation is a concatenation between words allowed by the rules R. Moreover it allows us a "proper" insertion of an element of X between two elements of X in special factorizations defined in Section 5. The main result concerning the second construction in our paper, is that we may obtain X_R^* as the image by a substitution of a language generated by a marked system, and this substitution is regular if so is X (Sections 5.1, 5.2).

As said in Section 1, in [13] the authors proved that $\leq_Y$ is a wqo if and only if Y is *unavoidable* in A^*. The latter condition also characterizes the regularity of $L_Y = \{w \in A^* \mid 1 \leq_Y w\}$. We recall that Y is unavoidable in A^* if there exists $k_0 \in N$ such that any w in A^*, with $|w| > k_0$, has a factor in Y. This notion appeared in a paper by Schützenberger [25], then explicitly introduced in [13] and considered also by other authors. There are algorithms to check that a given finite set Y is unavoidable (see Chapter 1 in [23]). In Section 7, we extend this notion by the concepts of R-unavoidable and strong R-unavoidable sets. We also prove relations between these two notions.

In Section 8, we prove our main result. In details, let $S = (A, I, R)$ be a $(1, 3)$-CSSH system. We prove that the strong R-unavoidability of $\mathrm{Lin}(I)$ in $\mathrm{Lin}(L(S))$ and a condition on the set R of rules guarantee the regularity of $\mathrm{Lin}(L(S))$ (Section 8). The mentioned condition on R is the same condition that characterizes regularity of languages generated by marked systems.

There are several issues that follows from the results stated in this paper, they will be discussed in Section 9.

4. $(1, 3)$-CSSH systems, flat systems and the iterated R-insertion operation

We give below the notion of the R-insertion operation.

Definition 4.1 *Given $Z \subseteq A^*$, $Z' \subseteq A^+$ and a symmetric relation R over A, the result $Z \leftarrow_R Z'$ of the R-*insertion *operation applied to Z, Z', is the following language*

$$Z \leftarrow_R Z' = \begin{cases} Z' & \textit{if } Z = \{1\}, \\ \{z_1 z z_2 \mid z_1 z_2 \in Z, z_2 z_1 \in A^* a, z \in Z' \cap A^* b, (a,b) \in R\} & \textit{otherwise.} \end{cases}$$

Definition 4.2 *Given $Y \subseteq A^*$ and a symmetric relation R over A, the result $Y^{\leftarrow_{*,R}}$ of the* iterated R-insertion *operation applied to Y, is the language $Y^{\leftarrow_{*,R}} = \cup_{i \geq 0} Y^{\leftarrow_{i,R}}$, where*

$$\begin{aligned} Y^{\leftarrow_{0,R}} &= \{1\} \\ Y^{\leftarrow_{1,R}} &= Y \\ Y^{\leftarrow_{i+1,R}} &= \bigcup_{0<j\leq i} (Y^{\leftarrow_{i,R}} \leftarrow_R Y^{\leftarrow_{j,R}}) \cup (Y^{\leftarrow_{j,R}} \leftarrow_R Y^{\leftarrow_{i,R}}), \qquad \text{for} \quad i \geq 1. \end{aligned}$$

Since $Y^{\leftarrow_{*,R}} = (Y \setminus \{1\})^{\leftarrow_{*,R}}$, in what follows, we assume $Y \subseteq A^+$. Moreover, we also set $Y^{\leftarrow_{+,R}} = Y^{\leftarrow_{*,R}} \setminus \{1\}$.

Lemma 4.3 *Let Y be a finite set and let R be a symmetric relation over A. If $w_1 w_2, w \in Y^{\leftarrow_{*,R}}$, with $w_2 w_1 \in A^* a$, $w \in A^* b$, $(a,b) \in R$, then $w_1 w w_2 \in Y^{\leftarrow_{*,R}}$.*

Proof :
Let $w_1 w_2, w, (a,b)$ be as in the statement. Thus, $w_1 w_2, w \in Y^{\leftarrow_{+,R}}$. By Definition 4.2, there are $i, j > 0$ such that $w_1 w_2 \in Y^{\leftarrow_{i,R}}$, $w \in Y^{\leftarrow_{j,R}}$. Set $t = \max\{i,j\}$. Hence, again by Definition 4.2, $w_1 w w_2 \in Y^{\leftarrow_{t+1,R}} \subseteq Y^{\leftarrow_{*,R}}$.

■

The following result generalizes a result proved in [3]. Recall that in a circular splicing system $S = (A, I, R)$, the set R is supposed to be symmetric.

Theorem 4.4 *For any circular language L over A the following conditions are equivalent:*

(1) There exists a $(1,3)$-CSSH system $S = (A, I, R)$ such that $L = L(S)$.

(2) There exists a flat splicing system $\mathcal{S} = (A, Y, R')$ such that $L(\mathcal{S}) = \mathrm{Lin}(L)$, where $Y \subseteq A^+$ is a finite language closed under the conjugacy relation, $R' = \{\langle a|1{-}b|1\rangle \mid (a,b) \in R\}$, and R is a symmetric relation over A.

(3) *There exists a finite language $Y \subseteq A^+$ such that Y is closed under the conjugacy relation and a symmetric relation R on A such that* $\mathrm{Lin}(L) = Y^{\leftarrow_{+,R}}$.

Theorem 4.4 is a direct consequence of the following two results. The first of them, Proposition 4.5, has been proved in [3]. The second one generalizes a result proved in the same paper.

Proposition 4.5 *Let $S = (A, I, R)$ be a $(1,3)$-CSSH system. Then the flat splicing system $\mathcal{S} = (A, Y, R')$, where $Y = \mathrm{Lin}(I)$, $R' = \{\langle a|1{-}b|1\rangle \mid (a, b) \in R\}$, is such that $L(\mathcal{S}) = \mathrm{Lin}(L(S))$. Conversely, let $\mathcal{S} = (A, Y, R')$ be a flat splicing system, where $Y \subseteq A^+$ is a finite language closed under the conjugacy relation, $R' = \{\langle a|1{-}b|1\rangle \mid (a, b) \in R\}$ and R is a symmetric relation on A. Let $I = {}^{\sim}Y$ be the circularization of Y. Then $L(\mathcal{S}) = \mathrm{Lin}(L(S))$, where $S = (A, I, R)$ is a $(1,3)$-CSSH system.*

Proposition 4.6 *Let $Y \subseteq A^+$ and let R be a symmetric relation on A. Then $Y^{\leftarrow_{+,R}} = L(\mathcal{S})$, where $\mathcal{S} = (A, Y, R')$ is a flat splicing system and $R' = \{\langle a|1{-}b|1\rangle \mid (a, b) \in R\}$.*

Proof :
We prove that $L = L(\mathcal{S}) \subseteq Y^{\leftarrow_{+,R}}$. Of course $L \subseteq A^+$. The proof is by induction on the minimal number of steps used for generating $w \in L$. If the number of steps is null, we have $w \in Y \subseteq Y^{\leftarrow_{+,R}}$.

Suppose now that for any word $w \in L$ generated in at most k steps, we have $w \in Y^{\leftarrow_{+,R}}$. Let w be a word generated in at least $k + 1$ steps. By the definition of the flat splicing operation, there are two words u and v, generated in at most k steps, a rule $\langle a|1{-}b|1\rangle \in R'$ and words x, y, z such that $u = xaz$, $v = yb$, $w = xaybz$. Thus, $(a, b) \in R$. Moreover, by induction, u and v are in $Y^{\leftarrow_{+,R}}$, hence w is also in $Y^{\leftarrow_{+,R}}$, by Lemma 4.3.

Conversely, we prove that $Y^{\leftarrow_{i,R}} \subseteq L(\mathcal{S})$, by induction on i, $i \geq 1$. Clearly $Y \subseteq L(\mathcal{S})$. Let w be a word in $Y^{\leftarrow_{i+1,R}}$, $i \geq 1$. By definition there are $z_1z_2 \in Y^{\leftarrow_{j,R}}$, $w' \in Y^{\leftarrow_{k,R}}$, with $0 < j, k \leq i$, $w'' \in A^*$, and $(a, b) \in R$, such that $z_2z_1 \in A^*a$, $w' = w''b$, and $w = z_1w'z_2$. By the induction hypothesis, the nonempty words z_1z_2, w' are in $L(\mathcal{S})$. If $z_1 \neq 1$, set $z_1 = z_1'a$. Thus the word $w = z_1'aw''bz_2$ is in $L(\mathcal{S})$, by using the rule $\langle a|1{-}b|1\rangle \in R'$. If $z_1 = 1$, then $z_2 \in A^*a$ and $(b, a) \in R$. Set $z_2 = z_2'a$. Thus the word $w = w''bz_2'a$ is in $L(\mathcal{S})$, by using the rule $\langle b|1{-}a|1\rangle \in R'$. ∎

5. The $\cdot_R$ and $*_R$ operations

In this section we define two operations on languages, the $\cdot_R$ and $*_R$ operations. We begin with an informal description of them.

Let X be a language. The language X_R^* is defined below as the union of the languages $X^{i,R}$. In turn, $X^{i,R}$ coincides with X^i for $i \in \{0,1\}$. For $i > 1$, X^i is the set of the concatenations of *all* the X-factorizations of length i, whereas $X^{i,R}$ will be the set of the concatenations of *some* of the X-factorizations of length i, called *valid X-factorizations* (or valid factorizations, when the context does not make it ambiguous) of $X^{i,R}$.

A valid factorization of $X^{2,R}$ is a pair (x,y), where $x \in A^*a \cap X$, $y \in A^*b \cap X$, and $(a,b) \in R$. Then the product xy is a member of $X^{2,R}$. The set of the valid factorizations of $X^{3,R}$ is the set of the tuples obtained by inserting in any position of any valid factorization of $X^{2,R}$ either an element of X or a sequence of two elements of a valid factorization of $X^{2,R}$ (or vice versa), provided that the insertion is "allowed" by R. Then, $X^{3,R}$ is the set of words which are products of elements in a valid factorization of $X^{3,R}$. In general, the set of the valid factorizations of $X^{i+1,R}$ is the set of the tuples obtained by inserting in any position of any valid factorization of $X^{k,R}$ a sequence of elements of a valid factorization of $X^{j,R}$, with $0 \le k,j \le i$, provided that the insertion is "allowed" by R. In other words, we get the valid factorizations of $X^{i+1,R}$ by inserting a valid factorization inside another valid factorization, both of them previously obtained, and provided that the insertion is "allowed" by R. Then again, $X^{i+1,R}$ is the set of words which are products of elements in a valid factorization of $X^{i+1,R}$.

The set of the valid factorizations of $X^{n,R}$ will be denoted by $\mathcal{VF}(X^{n,R})$. If $(x_1,\ldots,x_n) \in \mathcal{VF}(X^{n,R})$, then we say that $(x_1,\ldots,x_n)$ is a valid factorization of $w = x_1 \cdots x_n$. For any i, $0 \le i \le n$, the pair (x,y), where $x = x_1 \cdots x_i$, $y = x_{i+1} \cdots x_n$ is a *valid pair* for $X^{n,R}$. The set of the valid pairs of the elements in $X^{n,R}$ is denoted by $\mathcal{VP}(X^{n,R})$.

The following example should clarify these notions and their relations with the splicing operation.

Example 5.1 Let $X = \{a, ab, aba, ba, aab, baa\}$ and $R = \{(a,b)\}$. Let $S = (A, {}^\sim X, R)$. We have $(aba)(ab) \in X^2 \cap \mathrm{Lin}(L(S))$. Moreover, (aba, ab) is in $\mathcal{VF}(X^{2,R})$ and $abaab$ is a member of $X^{2,R}$. Then, since $(a,b) \in R$, we may insert ab between (aba) and (ab) and we get $abaabab \in X^{3,R}$, and $(aba, ab, ab) \in \mathcal{VF}(X^{3,R})$. We also have $w' = (aba)(ab)(a) \in X^{3,R}$ and

$(aba, ab, a) \in \mathcal{VF}(X^{3,R})$. We cannot obtain $w = (a)(baa)(ab)(ba)$ from w' even if w' factorizes also as $(a)(baa)(ba)$ since (a, baa, ba) is a X-factorization of $w' \in X^{3,R}$ but $(a, baa, ba) \notin \mathcal{VF}(X^{3,R})$. However, $w \in X^{4,R}$ since $(a)(ab) \in X^{2,R}$, hence $(aba)(a)(ab) \in X^{3,R}$ and finally $w = (aba)(a)(ab)(ba) \in X^{4,R}$. Observe that we also have $w = (aba)(a)(ab)(ba) \in X^{3,R}$, since $(aba)(ba) \in \mathcal{VF}(X^{2,R})$, $(a)(ab) \in \mathcal{VF}(X^{2,R})$ and so $w = (aba)(\mathbf{a})(\mathbf{ab})(ba) \in \mathcal{VF}(X^{3,R})$, since $aba \in A^*a, aab \in A^*b$ and $(a, b) \in R$.

Definition 5.2 *Let R be a symmetric relation over A and let $X \subseteq A^+$ be a set. We set $X^{*_R} = \bigcup_{i \geq 0} X^{i,R}$, $X^{+_R} = \bigcup_{i>0} X^{i,R}$, and $\mathcal{VF}(X^{*_R}) = \bigcup_{i \geq 0} \mathcal{VF}(X^{i,R})$, where:*

$$\begin{aligned}
\mathcal{VF}(X^{0,R}) &= \{1\}, \quad X^{0,R} = \{1\}; \\
\mathcal{VF}(X^{1,R}) &= \{(x) | x \in X\}, \quad X^{1,R} = X;
\end{aligned}$$

and, for $i > 1$,

$$\begin{aligned}
\mathcal{VF}(X^{i+1,R}) &= \{(x_1, \ldots, x_j, x'_1, \ldots, x'_t, x_{j+1}, \ldots, x_k) \mid (x_1, \ldots, x_j, x_{j+1}, \ldots x_k) \in \mathcal{VF}(X^{k',R}), \\
&\quad 1 \leq j \leq k, (x'_1, \ldots, x'_t) \in \mathcal{VF}(X^{t',R}), 0 \leq k', t' \leq i, \\
&\quad x_{j+1} \cdots x_k x_1 \cdots x_j \in A^*a, x'_1 \cdots x'_t \in A^*b, (a, b) \in R\}
\end{aligned}$$

$$\begin{aligned}
X^{i+1,R} &= \{x_1 \cdots x_k \mid (x_1, \ldots, x_k) \in \mathcal{VF}(X^{i+1,R}), k \geq 1\} = \\
&= \bigcup_{0<j\leq i} (X^{i,R} \cdot_R X^{j,R}) \cup (X^{j,R} \cdot_R X^{i,R})
\end{aligned}$$

Thus, $X^{1,R} = X$, $X^{2,R} = \{xy \mid x \in X \cap A^*a, y \in X \cap A^*b, (a, b) \in R\}$. The language $X^{3,R}$ is the set of the words $x_1x_2x_3$ where x_1x_2 (resp. x_2x_3, x_1x_3) is in $X^{2,R}$, (x_1, x_2) (resp. (x_2, x_3), (x_1, x_3)) is valid, and x_3 (resp. x_1, x_2) is in $X^{1,R} \cup X^{2,R}$ and may be inserted thanks to a rule in R.

In Example 5.6, we will show that, for our aims, we cannot take a simpler definition where all the X-factorizations are taken into account. Moreover observe that, differently from $\leftarrow_R$, the operator $\cdot_R$ cannot insert words inside an element of X.

5.1. A marked system associated with X^{*_R}

Let X be a set of nonempty words, let $A = \mathrm{alph}(X)$, and let R be a symmetric relation. In this section we define a marked system $S_{X,R}$ associated with X^{*_R}. Of course, we assume $R \subseteq A \times A$. Indeed, any $(a,b) \in R \setminus (A \times A)$ does not apply in the construction of X^{*_R}. In other words $X^{*_R} = X^{*_{R_1}}$, where $R_1 = R \cap (A \times A)$. We also assume $X_a = X \cap A^* a \neq \emptyset$, for any $a \in A = \mathrm{alph}(X)$. Indeed, in our results X will be a set containing $Y = \mathrm{Lin}(I)$, where $S = (A, I, R)$ is a $(1,3)$-CSSH system and we know that we may assume $\emptyset \neq Y_a = Y \cap A^* a$. Thus X_a is also nonempty.

Definition 5.3 *Let A be an alphabet and let R be a symmetric relation on A. Let $X \subseteq A^+$ be a language such that $X_a = X \cap A^* a \neq \emptyset$, for any $a \in A$. Let B be any alphabet such that* $\mathrm{Card}(A) = \mathrm{Card}(B)$ *and let β be any bijection from A onto B. We say that the* marked system $S_{X,R} = (B, R')$, *where $R' = \{(\beta(a), \beta(b)) \mid (a,b) \in R\}$, and the* substitution $\phi : B^* \to \mathfrak{P}(A^*)$, *defined by $\phi(\beta(a)) = X_a$ are* associated *with (X, R).*

In order to simplify notations, from now on we set $\beta(a) = a'$, for any $a \in A$.

Example 5.4 Let $X = \{a, ab, aba, ba, aab, baa\}$ and $R = \{(a,b)\}$ as Example 5.1. Thus $X_a = \{a, aba, ba, baa\}$ and $X_b = \{ab, aab\}$. The marked system $S_{X,R} = (B, R')$, where $B = \{a', b'\}$ and $R' = \{(a', b')\}$, and the substitution ϕ, defined by $\phi(a') = X_a$, $\phi(b') = X_b$, are associated with (X, R).

Proposition 5.5 *For any language X of nonempty words and for any symmetric relation R over A, we have $\phi(\mathrm{Lin}(L(S_{X,R}))) = X^{+_R}$.*

Proof :
First we prove that $\phi(\mathrm{Lin}(L(S_{X,R}))) \subseteq X^{+_R}$. Let $z \in \mathrm{Lin}(L(S_{X,R}))$. Thus, by Theorem 4.4, $z \in B^{\leftarrow_{+,R'}}$. Hence there exists k', $k' > 0$, such that $z \in B^{\leftarrow_{k',R'}}$. Looking at the definition of the iterated R-insertion in this special case, we may set $z = a'_1 \cdots a'_k$, where $a'_1, \ldots, a'_k \in B$. Any w in $\phi(z) = \phi(a'_1) \cdots \phi(a'_k)$ has the form $w = w_1 \cdots w_k$, where $w_r \in \phi(a'_r)$, $1 \leq r \leq k$.

Then we prove, by induction on k', that for any $w_1, \ldots, w_k$ such that $w_r \in \phi(a'_r)$, $1 \leq r \leq k$, the k-tuple $(w_1, \ldots, w_k) \in \mathcal{VF}(X^{k',R})$ and $w = w_1 \cdots w_k \in X^{k',R}$. Consequently, $\phi(z) \subseteq X^{+_R}$.

Let $k' = 1$, i.e., $z = a' \in B$. Thus, $\phi(z) = \phi(a') = X_a \subseteq X^{1,R}$ and $(w) \in \mathcal{VF}(X^{1,R})$, for any w in $\phi(a')$.

Assume the statement for any j, with $1 \leq j \leq k'$ and let us prove it for $k'+1$. Looking again at the definition of the iterated R-insertion in this special case, if $z \in B^{\leftarrow_{k'+1,R'}}$, then there exists $j \leq k'$ such that either

$z \in B^{\leftarrow_{k',R'}} \leftarrow_{R'} B^{\leftarrow_{j,R'}}$ or $z \in B^{\leftarrow_{j,R'}} \leftarrow_{R'} B^{\leftarrow_{k',R'}}$. Suppose that the first case holds (the argument is the same in the other case). Thus, there are $a'_1, \ldots, a'_t, a'_{\ell_1}, \ldots, a'_{\ell_s} \in B$, $z' = a'_1 \cdots a'_t \in B^{\leftarrow_{k',R'}}$, $z'' = a'_{\ell_1} \cdots a'_{\ell_s} \in B^{\leftarrow_{j,R'}}$, such that $z = a'_1 \cdots a'_h a'_{\ell_1} \cdots a'_{\ell_s} a'_{h+1} \cdots a'_t$ with $1 \leq h \leq t$, $(a'_h, a'_{\ell_s}) \in R'$, and where it is understood that for $h = t$ the word on the right of z'' is empty. (Note that the case $z = z''z'$, $(a'_t, a'_{\ell_s}) \in R'$, has not been considered since in this case $z \in B^{\leftarrow_{j,R'}} \leftarrow_{R'} B^{\leftarrow_{k',R'}}$.)

By induction hypothesis, for any $w_1, \ldots, w_t$ such that $w_r \in \phi(a'_r)$, $1 \leq r \leq t$, we have $(w_1, \ldots, w_t) \in \mathcal{VF}(X^{k',R})$ and $w_1 \cdots w_t \in X^{k',R}$. Similarly, for any $w'_g \in \phi(a'_{\ell_g})$, $1 \leq g \leq s$, we have $(w'_1, \ldots, w'_s) \in \mathcal{VF}(X^{j,R})$ and $w'_1 \cdots w'_s \in X^{j,R}$. Moreover, $(a_h, a_{\ell_s}) \in R$, by Definition 5.3. By Definitions 5.2, 5.3, it is easy to conclude that $(w_1, \ldots, w_h, w'_{\ell_1}, \ldots, w'_{\ell_s}, w_{h+1}, \ldots, w_t) \in \mathcal{VF}(X^{k'+1,R})$ and, as a consequence, $w = w_1 \cdots w_h w'_{\ell_1} \cdots w'_{\ell_s} w_{h+1} \cdots w_t \in X^{k'+1,R}$.

Conversely, we prove that $X^{+_R} \subseteq \phi(\mathrm{Lin}(L(S_{X,R})))$. Let $w \in X^{+_R}$. Therefore there exists k', $k' > 0$, such that $w \in X^{k',R}$, i.e., $(w_1, \ldots, w_k) \in \mathcal{VF}(X^{k',R})$ such that $w = w_1 \cdots w_k$.

We prove, by induction on k, that there are $a'_1, \ldots, a'_k \in B$ such that $z = a'_1 \cdots a'_k \in B^{\leftarrow_{k',R'}}$ and $w_r \in \phi(a'_r)$, $1 \leq r \leq k$. Hence, $w \in \phi(z)$ and, by Theorem 4.4, $w \in \phi(\mathrm{Lin}(L(S_{X,R})))$.

Let $i = 1$, i.e., $w \in X = X^{1,R}$. Thus, there exists $a \in A$ such that $w \in X_a$. Hence $w \in \phi(z)$, where $z = a' \in B = B^{\leftarrow_{1,R'}}$.

Assume the statement for any j, with $1 \leq j \leq k'$ and let us prove it for $k' + 1$. Now $w \in X^{k'+1,R} = \cup_{0<j\leq k'} (X^{k',R} \cdot_R X^{j,R}) \cup (X^{j,R} \cdot_R X^{k',R})$.

Notice that, since the elements of X are supposed to be nonempty words, the same holds for the elements in a k-tuple in $\mathcal{VF}(X^{k',R})$. Thus, by Definition 5.2, there exist $(w_1, \ldots, w_k) \in \mathcal{VF}(X^{k',R})$, $(w'_1, \ldots, w'_t) \in \mathcal{VF}(X^{j,R})$, $(a, b) \in R$, with $w'_t \in X \cap A^*b$ such that $w = w_1 \cdots w_h w'_1 \cdots w'_t w_{h+1} \cdots w_k$, $1 \leq h \leq k$, $w_h \in A^*a$, and where it is understood that for $h = k$ the word on the right of w'_t is empty. (We do not consider the case where $w = w'_1 \cdots w'_h w_1 \cdots w_k w'_{h+1} \cdots w'_t$, with $1 \leq h \leq t$, and $w'_h \in A^*a, w_k \in A^*b$ since the argument below remains the same.)

By induction hypothesis there are $a'_1, \ldots, a'_k, a'_{\ell_1}, \ldots, a'_{\ell_t} \in B$ such that $a'_1 \cdots a'_k \in B^{\leftarrow_{k',R'}}$ and $w_r \in \phi(a'_r)$, $1 \leq r \leq k$, $a'_{\ell_1} \cdots a'_{\ell_t} \in B^{\leftarrow_{j,R'}}$ and $w'_g \in \phi(a'_{\ell_g})$, $1 \leq g \leq t$. Moreover, by Definition 5.3, $(a', b') = (a'_h, a'_{\ell_t}) \in R'$. Finally, $z = a'_1 \cdots a'_h a'_{\ell_1} \cdots a'_{\ell_t} a'_{h+1} a'_t \in B^{\leftarrow_{k'+1,R'}}$ and the proof is ended. ∎

The following example shows that Proposition 5.5 is no more true if we choose a simpler definition of X^{*_R}.

Example 5.6 Consider again $X = \{a, ab, aba, ba, aab, baa\}$ and $R = \{(a, b)\}$, as in Example 5.1. The associated marked system $S_{X,R} = (B, R')$ and substitution ϕ were given in Example 5.4 and repeated here for convenience. Hence, $B = \{a', b'\}$, $R' = \{(a', b')\}$, and ϕ is defined by $\phi(a') = X_a = \{a, aba, ba, baa\}$, $\phi(b') = X_b = \{ab, aab\}$. In Example 5.1, we observed that we cannot obtain $w = abaaabba$ from $w' = (aba)(ab)(a) \in X^{3,R}$ by inserting ab into $w' = (a)(baa)(ba)$ before the second b. If we considered a simpler definition of X^{*_R} that allows us to do that, the language $X^{4,R}$ would not match $B^{\leftarrow_{4,R'}}$ since $a'a'a' \notin B^{\leftarrow_{4,R'}}$. Notice that $w' \in \phi(a'a'a') \cap \phi(a'b'a')$, where $a'b'a' \in B^{\leftarrow_{3,R'}}$.

5.2. Regularity and non-regularity of X^{*_R}

In this section we consider conditions under which the language X^{*_R} is regular. We use the same notations as in the previous section.

Let R' be a symmetric relation over B, represented by an undirected graph $G' = (B, R')$, where B is the vertex set and R' is the edge set. In an undirected graph, *self-loops* - edges from a vertex to itself - are forbidden but here we do not make this assumption. As in [3, 5], G' will be referred to as the graph *associated* with the marked system $S = (B, R')$. A path in a graph is *simple* if all vertices in the path are distinct. A graph G is *simple* if there are no self-loops in G and the simple graph *underlying* G is the graph obtained by dropping the self-loops in G.

We state below a characterization of the marked systems S generating regular circular languages by means of a property of the graph G' associated with S. This characterization was proved in [11], then reviewed in a graph theoretical setting in [5]. The involved property of the graph G' is given by means of the well known graph $P_4 = (V, E)$, where $V = \{a_1, a_2, a_3, a_4\}$, and $E = \{(a_1, a_2), (a_2, a_3), (a_3, a_4)\}$. We also recall that a *P_4-free* graph G is a graph such that every connected subgraph of the simple graph underlying G, which is induced by a set of four vertices of G, is not P_4.

Theorem 5.7 *Let $S = (B, R')$ be a marked system, let G' be the graph associated with S. The following conditions are equivalent:*

(1) $L(S)$ is a regular circular language.

(2) The simple graph underlying graph G' is P_4-free.

Two graphs which contain the same number of graph vertices connected in the same way are said to be isomorphic. Formally, two graphs G and H with graph vertices $V_n = \{1, 2, ..., n\}$ are said to be isomorphic if there is a permutation p of V_n such that $\{u, v\}$ is in the set of graph edges $E(G)$ if and only if $\{p(u), p(v)\}$ is in the set of graph edges $E(H)$. The following result is a direct consequence of the definitions.

Proposition 5.8 *Let $X \subseteq A^+$ be a set of nonempty words and let R be a symmetric relation over A. Let $S_{X,R} = (B, R')$ be the marked system associated with (X, R), where $R' = \{(a', b') \mid (a, b) \in R\}$, let G' be the graph associated with $S_{X,R}$. The graphs G' and $G = (A, R)$ are isomorphic. In particular, G' is P_4-free if and only if G is P_4-free.*

The following is a direct consequence of the above results.

Corollary 5.9 *Let $X \subseteq A^+$ be a regular language of nonempty words such that $X_a = X \cap A^*a \neq \emptyset$ for any $a \in A$. Let R be a symmetric relation over A. If $G = (A, R)$ is P_4-free, then X^{*_R} is regular.*

PROOF :
Let X, R be as in the statement. Consider the marked system $S_{X,R} = (B, R')$ and the substitution ϕ associated with (X, R) (Definition 5.3). If $G = (A, R)$ is P_4-free, then $L(S_{X,R})$ is a regular language, by Theorem 5.7 and Proposition 5.8. Since the class of regular languages is closed under intersection, ϕ is a regular substitution. Finally, by Proposition 5.5, we have $\phi(\mathrm{Lin}(L(S_{X,R}))) = X^{+_R}$. Since regular languages are closed under regular substitution, the languages X^{+_R} and X^{*_R} are both regular. ∎

6. Another construction of splicing languages

In this section, R will be a symmetric relation over A. Moreover, we assume that $Y \subseteq A^+$ is a finite set closed under the conjugacy relation, and such that $Y \cap A^*a \neq \emptyset$, for any $a \in A$. Aimed to provide an alternative construction of splicing languages, we first give some definitions.

Definition 6.1 *Let $A = \{a_1, \ldots, a_n\}$ and $\bar{A} = \{b_1, \ldots, b_n\}$ be two disjoint alphabets such that* $\mathrm{Card}(A) = \mathrm{Card}(\bar{A})$. *We consider the morphism $\phi_C : A^* \to \bar{A}^*$ defined by $\phi_C(a_i) = b_i$, for any i, $1 \leq i \leq n$. We set $\bar{R} = R \cup \{(\phi_C(a), c), (c, \phi_C(a)) \mid (a, c) \in R\}$.*

Definition 6.2 *We set*

$$\mathrm{Base}(I_0) = Y, \quad I_0 = Y^{*_R}$$

and, for $n \geq 0$, for each $a_i \in A$,

$$\begin{aligned}
\bar{I}_n(a_i) &= \phi_C^{-1}(a_i)((\{\phi_C(a_i)\} \cup I_n)^{*_{\bar{R}}} \ \cap \ (\phi_C(a_i)A^* \setminus A^+\phi_C(a_i)A^*)) \\
\mathrm{Base}(I_{n+1}) &= \{a_1 w_1 \cdots a_k w_k \mid a_1 \cdots a_k \in Y, \ w_i \in \bar{I}_n(a_i) \cup \{1\}\ , 1 \leq i \leq k\} \\
I_{n+1} &= (\mathrm{Base}(I_{n+1,Y}))^{*_R}
\end{aligned}$$

In the previous definition, it is clear that $a_j \in A$, for $1 \leq j \leq k$, and the operation between the letters a_i and the words w_i is the usual concatenation of words. Therefore, any element in I_n may be written as $w = w_1 \cdots w_n$, where $w_j \in \mathrm{Base}(I_n)$ and $(w_1, \ldots, w_n) \in \mathcal{VF}(I_n)$. Next, when the context does not make it ambiguous, we write I_n instead of $I_{n,Y}$.

In the following subsections we describe the properties of I_n sets. Briefly, they form a non decreasing sequence of sets, with respect to the order of set inclusion (Proposition 6.4). Their union is the splicing language (Theorem 6.9), hence they are closed with respect to the R-insertion operation (Proposition 6.7). Finally, under an appropriate hypothesis, if I_n is regular, then so is I_{n+1} (Proposition 6.13).

6.1. Inclusion

The following result is a direct consequence of Definition 5.2.

Proposition 6.3 *Let $X_1, X_2 \subseteq A^*$. If $X_1 \subseteq X_2$, then $\mathcal{VF}(X_1^{*_R}) \subseteq \mathcal{VF}(X_2^{*_R})$ and, consequently, $X_1^{*_R} \subseteq X_2^{*_R}$.*

Proposition 6.4 *For any $n \geq 0$,*

(i) $\mathrm{Base}(I_n) \subseteq \mathrm{Base}(I_{n+1})$, *and consequently* $I_n \subseteq I_{n+1}$,

(ii) $\bar{I}_n(a) \subseteq \bar{I}_{n+1}(a)$ *for each* $a \in A$.

Proof :
By Proposition 6.3, if $\mathrm{Base}(I_n) \subseteq \mathrm{Base}(I_{n+1})$, then $I_n \subseteq I_{n+1}$. We prove (i) and (ii) together, by mutual induction on n.

(**Basis**) Let $n = 0$.

(i) Let $w = a_1 \cdots a_k \in \mathrm{Base}(I_0) = Y$, where $a_i \in A$, $1 \leq i \leq k$. Thus, by definition, $w = a_1 \cdot 1 \cdots a_k \cdot 1 \in \mathrm{Base}(I_1)$. Therefore (i) holds for $n = 0$.

(ii) Suppose that $w \in \bar{I}_0(a)$ with $a \in A$. By Definition 6.2, we have

$$w \in \phi_C^{-1}(a) \left((\{\phi_C(a)\} \cup I_0)^{*_{\bar{R}}} \;\; \cap \;\; \left(\phi_C(a)A^* \setminus A^+\phi_C(a)A^*\right)\right).$$

By the above argument, $I_0 \subseteq I_1$. Thus it is easy to see that $\bar{I}_0(a) \subseteq \bar{I}_1(a)$, once again by Proposition 6.3.

(**Induction**) Now we assume that (i) and (ii) are true for $n \geq 0$ and we prove them for $n + 1$.

(i) Assume $w \in \mathrm{Base}(I_n)$. Then, by definition, $w = a_1 w_1 a_2 w_2 \cdots a_k w_k$, with $a_1 \cdots a_k \in Y$, and $w_i \in \bar{I}_{n-1}(a_i) \cup \{1\}$ for each i, $1 \leq i \leq k$. By inductive hypothesis of *(ii)*, if $w_i \neq 1$, then $w_i \in \bar{I}_n(a_i)$. Hence, $w = a_1 w_1 a_2 w_2 \cdots a_k w_k \in \mathrm{Base}(I_{n+1})$.

(ii) Let $w \in \bar{I}_n(a)$ with $a \in A$. By Definition 6.2,

$$w \in \phi_C^{-1}(a) \left((\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}} \;\; \cap \;\; \left(\phi_C(a)A^* \setminus A^+\phi_C(a)A^*\right)\right).$$

By the above argument, $I_n \subseteq I_{n+1}$. Thus it is easy to see that $\bar{I}_n(a) \subseteq \bar{I}_{n+1}(a)$, once again by Proposition 6.3.

■

6.2. Insertion

Proposition 6.5 and Lemma 6.6 are needed for stating Proposition 6.7.

Proposition 6.5 *Let $a \in A$, $w \in A^*$. The word w is in $\bar{I}_n(a)$ if and only if there is $(\phi_C(a), x_1, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}})$ such that $w = x_1 \cdots x_m$ and $x_j \in I_n$, $1 \leq j \leq m$. In this case, $(a, x_1, \ldots, x_m) \in \mathcal{VF}((a \cup I_n)^{*_R})$.*

PROOF :
We preliminary observe that $I_n \subseteq A^*$ for any n. Therefore, for every x in $(\{\phi_C(a)\} \cup I_n)$, either $x = \phi_C(a)$ or $\phi_C(a) \notin \mathrm{alph}(w)$. Assume $w \in \bar{I}_n(a)$. By Definition 6.2, there is $(x_0, x_1, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}})$ such that $x_0 x_1 \cdots x_n \in (\phi_C(a)A^* \setminus A^+\phi_C(a_i)A^*)$ and $w = \phi_C^{-1}(a)\{x_0 x_1 \cdots x_n\}$. By our preliminary observation, $x_0 = \phi_C(a)$, $w = x_1 \cdots x_m$, and $x_j \in I_n$, $1 \leq j \leq m$.

Conversely, let $(\phi_C(a), x_1, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}})$, with $x_j \in I_n$, $1 \leq j \leq m$. The word $\phi_C(a)x_1 \cdots x_m$ is clearly in $(\phi_C(a)A^* \setminus A^+\phi_C(a_i)A^*)$. Hence, by Definition 6.2, $w = x_1 \cdots x_m \in \bar{I}_n(a)$. The second part of the statement is clear. ∎

Lemma 6.6 *Let $x_1, \ldots, x_k \in X$. Let $z_h = x'_h t x''_h$, where x'_h, x''_h are nonempty words such that $x'_h x''_h = x_h$, $t \in A^*$, $1 \leq j \leq k$. If $z_h \in X$ and $(x_1, \ldots, x_k) \in \mathcal{VF}(X^{*_R})$, then $(x_1, \ldots, x_{h-1}, z_h, x_{h+1}, \ldots, x_k) \in \mathcal{VF}(X^{*_R})$.*

Proof :
We prove the statement by induction of $k \geq 0$. It is clearly true for $k = 1$. Otherwise, by Definition 5.2, there are $(y_1, \ldots, y_{k'}) \in \mathcal{VF}(X^{*_R})$, $(v_1, \ldots, v_t) \in \mathcal{VF}(X^{*_R})$ such that $(x_1, \ldots, x_k) = (y_1, \ldots, y_j, v_1, \ldots, v_t, y_{j+1}, \ldots, y_{k'})$ and $y_{j+1} \cdots y_{k'} y_1 \cdots y_j \in A^*a$, $v_1 \cdots v_t \in A^*b$, $(a, b) \in R$. Then either $x_h = y_{h'}$, with $1 \leq h' \leq k'$, or $x_h = v_{h'}$, with $1 \leq h' \leq t$. In both cases, by the induction hypothesis and since x'_h, x''_h are nonempty, the conclusion holds. ∎

Proposition 6.7 *If $uv, w \in I_n$, with $vu \in A^*a$, $w \in A^*b$ and $(a, b) \in R$, then $uwv \in I_t$, $t \geq n$. Moreover, if $|u| \leq n$ and $w \in Y$, then $uwv \in I_n$.*

Proof :
We preliminary observe that we may assume $u \neq 1$. Indeed, $w \in I_n$, for $n \geq 0$, and if $u = 1$, then $uwv = wv$ is still in I_n, by definition (and by using $(b, a) \in R$). This show also the second part of the statement for $n = 0$.

Since $uv \in I_n$ there are $y_1, \ldots, y_k \in \mathrm{Base}(I_n)$, $(y_1, \ldots, y_k) \in \mathcal{VF}(I_n)$ such that $uv = y_1 \cdots y_k$. Since $w \in I_n$ there are $z_1, \ldots, z_h \in \mathrm{Base}(I_n)$, $(z_1, \ldots, z_h) \in \mathcal{VF}(I_n)$ such that $w = z_1 \cdots z_h$. If $u = y_1 \cdots y_j$, $v = y_{j+1} \cdots y_k$, then $(y_1, \ldots, y_j, z_1, \ldots, z_h, y_{j+1}, \ldots, y_k) \in \mathcal{VF}(I_n)$ and $uwv = y_1 \cdots y_j z_1 \cdots z_h$ $y_{j+1} \cdots y_k \in I_n$.

Otherwise, $u = y_1 \cdots y'_j$, $v = y''_j \cdots y_k$, for nonempty words y'_j, y''_j such that $y_j = y'_j y''_j$. If we prove that $y'_j w y''_j \in I_m$, $m \geq n$, the first part of the statement follows by Proposition 6.4 and Lemma 6.6. As for the second part, we prove that if $|y'_j| \leq n$ and $w \in Y$, then $y'_j w y''_j \in I_n$, so, again by Lemma 6.6, $uwv \in I_n$.

We prove that $y'_j w y''_j \in I_m$, $m \geq n$, by induction on n. For our convenience, we set $y'_j = u$, $y''_j = v$. Let $n = 0$. Thus, $uv \in \mathrm{Base}(I_0) = Y$, then $uv = a_1 \cdots a_k \in Y$, where $a_j \in A$, $1 \leq j \leq k$. Moreover, $w \in I_0$,

$vu \in A^*a$, $w \in I_0 \cap A^*b$, and $(a,b) \in R$. Hence, there exists i, with $1 \leq i \leq k$ such that $u = a_1 \cdots a_i$ and $v = a_{i+1} \cdots a_k$ and $(a_i, b) \in R$. It is clear that $(a_i, z_1, \ldots, z_h) \in \mathcal{VF}((a \cup I_0)^{*_R})$, i.e., $(\phi_C(a_i), z_1, \ldots, z_h) \in \mathcal{VF}((\phi_C(a_i) \cup I_0)^{*_{\bar{R}}})$. This implies, by Definition 6.2, $w \in \bar{I}_0(a_i)$, thus $uwv = a_1 \cdots a_i w a_{i+1} \cdots a_k \in \mathrm{Base}(I_1) \subseteq I_1$. Regarding the basis for the second part of the statement, if $n = 0$, then $u = 1$ and $v \in I_0 = Y^{*_R}$. By definition, $uyv = yv \in I_0$.

Suppose that the statement is true for n', $0 \leq n' < n$, let us prove it for n. Let $uv \in \mathrm{Base}(I_n)$, $w \in I_n$. By Definition 6.2, we have $uv = a_1 w_1 a_2 w_2 \cdots a_k w_k$, with $a_1 \cdots a_k \in Y$, and, for each $w_i \neq 1$, $1 \leq i \leq k$, with $w_i \in \bar{I}_{n-1}(a_i)$. Hence for some i, $1 \leq i \leq k$, $u = a_1 \cdots a_i w_i'$ and $v = w_i'' a_{i+1} \cdots w_{k-1} a_k w_k$ where w_i', w_i'' are words such that $w_i' w_i'' = w_i$. By Proposition 6.4, we also have $w_j \in \bar{I}_n(a_j)$ for each $w_j \neq 1$, $1 \leq j \leq k$.

If $w_i' = w_i'' = 1$, then $a_i = a$, $w \in I_n \cap A^*b$ with $(a_i, b) \in R$ and so, by definition, $w \in \bar{I}_n(a_i)$. Thus, $uwv = a_1 w_1 a_2 \cdots a_i w \cdots w_{k-1} a_k w_k \in I_{n+1}$. Of course, if $w \in Y \cap A^*b$, then $w \in \bar{I}_{n-1}(a_i)$ and thus, $uwv = a_1 w_1 a_2 \cdots a_i w \cdots w_{k-1} a_k w_k \in I_n$.

Otherwise, we distinguish two cases, depending on whether $w_i' = 1$ or not. If $w_i' = 1$, then we know that $w_i = w_i'' \in \bar{I}_n(a_i)$ and $w \in I_n \cap A^*b$, with $(a,b) = (a_i, b) \in R$. Looking at Definition 6.2, we conclude that $ww_i = ww_i'' \in \bar{I}_n(a_i)$ and $uwv = a_1 w_1 a_2 \cdots a_i w w_i \cdots w_{k-1} a_k w_k \in I_{n+1}$. Moreover, if $w \in Y$, then $ww_i = ww_i'' \in \bar{I}_{n-1}(a_i)$, so $uwv = a_1 w_1 a_2 \cdots a_i w w_i \cdots w_{k-1} a_k w_k \in I_n$.

Now, assume $w_i' \neq 1$ and $w_i = x_1 \cdots x_m$ with $(\phi_C(a_i), x_1, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a_i)\} \cup I_{n-1})^{*_{\bar{R}}})$. Hence, there exists $1 \leq j \leq m$ such that $w_i' = x_1 \cdots x_{j-1} x_j'$, $w_i'' = x_j'' x_{j+1} \cdots x_m$ and $x_j = x_j' x_j''$. Let us consider the word $x_j' w x_j''$. If $x_j' = 1$, it is easy to see that $(\phi_C(a_i), x_1, \ldots, x_{j-1}, w, x_j, x_{j+1}, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a_i)\} \cup I_{n-1})^{*_{\bar{R}}})$, thus $w_i' w w_i'' \in \bar{I}_{n-1}(a_i)$. A similar argument holds if $x_j'' = 1$. In both cases, $uwv = a_1 w_1 a_2 \cdots a_i (w_i' w w_i'') \cdots w_{k-1} a_k w_k \in I_n$ and the two parts of the statement are proved.

Otherwise, $x_j' \neq 1$, $x_j' \neq 1$ and, by induction hypothesis, we have that $x_j' w x_j'' \in I_{t'}$ with $t' \geq n-1$. By Proposition 6.4 and by Lemma 6.6, we also have $(\phi_C(a_i), x_1, \ldots, x_{j-1}, x_j' w x_j'', x_{j+1}, \ldots, x_m) \in \mathcal{VF}((\{\phi_C(a_i)\} \cup I_{t'})^{*_{\bar{R}}})$, i.e., $w_i' w w_i'' \in \bar{I}_{t'}(a_i)$. Let $t = \max\{t', n\}$. Again by Proposition 6.4, $w_j \in \bar{I}_t(a_j)$ for each $w_j \neq 1$, $1 \leq j \leq k$, $j \neq i$, and $w_i' w w_i'' \in \bar{I}_t(a_i)$. Therefore $uwv \in \mathrm{Base}(I_t)$, with $t \geq n$ and the proof of the first part of the statement is ended. Regarding the second part of the statement, if $|u| \leq n$, then $|x_j'| \leq |w_i'| \leq n-1$ and, by the induction hypothesis, we have that $x_j' w x_j'' \in I_{n-1}$. Therefore, by the above argument, $w_i' w w_i'' \in \bar{I}_{n-1}(a_i)$ and $uwv =$

$a_1w_1a_2\cdots a_i(w_i'ww_i'')\cdots w_{k-1}a_kw_k \in I_n$. ■

6.3. Generation

We are now ready to prove that the collection of sets I_n is a splicing language.

Lemma 6.8 *If* $X \subseteq \mathrm{Lin}(L(S))$*, then* $X^{*_R} \subseteq \mathrm{Lin}(L(S))$*.*

PROOF :
The conclusion follows directly by using induction on i such that $w \in X^{i,R}$ and Definition 5.2. ■

Theorem 6.9 *Let* $S = (A, I, R)$ *be a* $(1,3)$*-CSSH system and let* $Y = \mathrm{Lin}(I)$*. Then* $\mathrm{Lin}(L(S)) = (\cup_{n\geq 0} I_n) \setminus \{1\}$*.*

PROOF :
First, we prove that $\mathrm{Lin}(L(S)) \subseteq (\cup_{n\geq 0} I_n) \setminus \{1\}$. Let $w \in \mathrm{Lin}(L(S))$. Clearly $w \neq 1$. By Theorem 4.4, there is i, $i > 0$, such that $w \in Y^{\leftarrow_{i,R}}$. We prove, by induction on i, that $w \in (\cup_{n\geq 0} I_n) \setminus \{1\}$. Of course, the conclusion holds for $i = 1$, i.e., if $w \in Y$. Now, assume that the statement holds for $i \geq 1$ and let us prove it for $i+1$. Let $w \in Y^{\leftarrow_{i+1,R}}$. Then, by definition, $w = w_1w'w_2$, where $w_1w_2 \in Y^{\leftarrow_{t,R}}$, $w_2w_1 \in A^*a$, $w' \in Y^{\leftarrow_{t',R}} \cap A^*b$, $(a,b) \in R$ and $t, t' \leq i$. By induction hypothesis, there are $n, n' \in \mathbb{N}$ such that $w_1w_2 \in I_n$, $w' \in I_{n'}$. Let $m = max\{n, n'\}$. By Proposition 6.4, we have that $w_1w_2, w' \in I_m$. Hence, in virtue of Proposition 6.7, we have that $w \in I_l$, $l \geq m$.

Next we demonstrate that $(\cup_{n\geq 0} I_n) \setminus \{1\} \subseteq \mathrm{Lin}(L(S))$. We show that $I_n \setminus \{1\} \subseteq \mathrm{Lin}(L(S))$, by induction on n. Let $w \in I_n \setminus \{1\}$. By Lemma 6.8, we may assume $w \in \mathrm{Base}(I_n)$. If $n = 0$, then $w \in Y \subseteq \mathrm{Lin}(L(S))$. Otherwise, $w = a_1w_1a_2w_2\cdots a_kw_k$, with $a_1\cdots a_k \in Y$, and, $w_i \in \bar{I}_n(a_i)$, for each $w_i \neq 1$, $1 \leq i \leq k$. As a preliminary remark, notice that if $za_i \in \mathrm{Lin}(L(S))$, then za_iw_i is also in $\mathrm{Lin}(L(S))$. This claim may be easily obtained by considering that $w_i \in \bar{I}_n(a_i)$ and by looking at Definition 6.2. It immediately yields $a_2\cdots a_ka_1w_1 \in \mathrm{Lin}(L(S))$ and so, $a_1w_1a_2a_3\cdots a_{k-1}a_k \in \mathrm{Lin}(L(S))$. Moreover, if $a_1w_1\cdots a_{i-1}w_{i-1}a_ia_{i+1}\cdots a_k \in \mathrm{Lin}(L(S))$, then $a_{i+1}\cdots a_ka_1w_1\cdots a_{i-1}w_{i-1}a_i$ is also in $\mathrm{Lin}(L(S))$ and, by the above claim, $a_{i+1}\cdots a_ka_1w_1\cdots a_{i-1}w_{i-1}a_iw_i \in \mathrm{Lin}(L(S))$. Of course, this implies $a_1w_1\cdots a_{i-1}w_{i-1}a_iw_ia_{i+1}\cdots a_k \in \mathrm{Lin}(L(S))$. The above arguments demonstrate, by induction on i, that $a_1w_1\cdots a_{i-1}w_{i-1}a_iw_ia_{i+1}\cdots a_k$ is in $\mathrm{Lin}(L(S))$, for any i, $1 \leq i \leq k$. For $i = k$, this implies $w \in \mathrm{Lin}(L(S))$. ■

Proposition 6.10 will be used in Section 8.

Proposition 6.10 *For any $n \geq 1$, if $w \in I_n \setminus (I_{n-1} \cup \{1\})$, then $|w| \geq n$.*

PROOF :
The proof is by induction on n. It is clear that the conclusion holds for $n = 1$. Suppose the statement true for each m, $0 < m < n$, and let us prove it for n. Let $w \in I_n \setminus (I_{n-1} \cup \{1\})$.

By Definition 6.2, $w = w_1 \cdots w_t$, where $w_j \in \mathrm{Base}(I_n)$, $t \geq 1$, and $(w_1, \ldots, w_t) \in \mathcal{VF}(I_n)$. Furthermore, there exists i, $0 \leq i \leq t$, such that $w_i \notin \mathrm{Base}(I_{n-1}) \cup \{1\}$, otherwise $w \in I_{n-1}$. If we prove that $|w_i| \geq n$, then $|w| \geq |w_i| \geq n$, which completes the proof.

Set $w_i = z$. Then, by definition, $z = a_1 z_1 a_2 z_2 \cdots a_k z_k$, with $a_1 \cdots a_k \in Y$, and, for each $z_i \neq 1$, $1 \leq i \leq k$, with $z_i \in (\bar{I}_{n-1}(a_i) \cup \{1\})$. Moreover there exists t, $1 \leq t \leq k$, such that $z_t \neq 1$, (otherwise $z \in Y = \mathrm{Base}(I_0) \subseteq \mathrm{Base}(I_{n-1})$) and $z_t \notin \bar{I}_{n-2}(a_i)$ (otherwise $z \in \mathrm{Base}(I_{n-1})$). Thus, $z_t \in I_{n-1} \setminus (I_{n-2} \cup \{1\})$ and, by induction hypothesis, $|z_t| \geq n-1$ which yields $|z| \geq n$.
■

6.4. Regularity and non-regularity of I_n

In this section we assume that $Y \subseteq A^+$ is a finite set closed under the conjugacy relation, and such that $Y \cap A^*a \neq \emptyset$, for any $a \in A$. The following statement is a direct consequence of Corollary 5.9.

Proposition 6.11 *Let R be a symmetric relation over A. If $G = (A, R)$ is P_4-free, then $I_0 = Y^{*_R}$ is regular.*

Proposition 6.12 *Let R be a symmetric relation over A. Let $A' = A \cup \{\bar{a}\}$, where $\bar{a} \notin A$, $a \in A$, and let $\bar{R} = R \cup \{(\bar{a}, c), (c, \bar{a}) \mid (a, c) \in R\}$. If $G = (A, R)$ is P_4-free, then $G' = (A', \bar{R})$ is also P_4-free.*

PROOF :
On the contrary, assume that $G = (A, R)$ is P_4-free and $G' = (A', \bar{R})$ is not P_4-free. Therefore, there are four different letters $a_1, a_2, a_3, a_4 \in A'$ such that the simple graph underlying G', which is induced by this set of vertices of G', is P_4, i.e., $P_4 = (V, E)$, where $V = \{a_1, a_2, a_3, a_4\}$, and $E = \{(a_1, a_2), (a_2, a_3), (a_3, a_4)\}$. Of course, $\bar{a} \in V$, otherwise $G = (A, R)$ would not be P_4-free. For the same reason, a is a member of V since, otherwise, we may substitute in V the vertex $\bar{a}$ with a and the simple graph

underlying G, which is induced by this new set of vertices of G, is still P_4. If $(a, \bar{a}) \in E$, say $(a, \bar{a}) = (a_1, a_2)$, then $(a, a_3) \in \bar{R} \setminus E$, and we obtain a contradiction. The same argument applies when $(a, \bar{a}) = (a_2, a_3)$ (since $(a, a_4) \in \bar{R} \setminus E$) or when $(a, \bar{a}) = (a_3, a_4)$ (since $(\bar{a}, a_2) \in \bar{R} \setminus E$). Finally, if $(a, \bar{a}) \notin E$, there is j, $1 \leq j \leq 4$, such that $a_j \neq a$, $a_j \neq \bar{a}$, and only one between (a, a_j) and $(\bar{a}, a_j)$ is in E, which contradicts the definition of $\bar{R}$. Thus, $G' = (A', \bar{R})$ is P_4-free.

Proposition 6.13 *For any $n \geq 0$, if $G = (A, R)$ is P_4-free and I_n is regular, then I_{n+1} is regular.*

PROOF :
Let I_n be a regular set and $G = (A, R)$ a P_4-free graph. As a preliminary step, we prove that regularity of I_n implies regularity of $\bar{I}_n(a)$, for any $a \in A$. Indeed, by Definition 6.2, $\bar{I}_n(a) = \phi_C^{-1}(a)((\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}} \cap (\phi_C(a)A^* \setminus A^+\phi_C(a)A^*))$. If I_n is regular, then the same holds for $(\{\phi_C(a)\} \cup I_n)$ and, by Corollary 5.9 and Proposition 6.12, for the language $(\{\phi_C(a)\} \cup I_n)^{*_{\bar{R}}}$ too. Thus, regularity of $\bar{I}_n(a)$ follows by the known closure properties of regular sets. Next, recall that Y is a finite set, and moreover

$$\mathrm{Base}(I_{n+1}) = \bigcup_{a_1 \cdots a_k \in Y} a_1(\bar{I}_n(a_1) \cup \{1\}) \cdots a_k(\bar{I}_n(a_k) \cup \{1\})$$

Consequently, $\mathrm{Base}(I_{n+1})$ is regular too. Finally, by Corollary 5.9, $I_{n+1} = (\mathrm{Base}(I_{n+1}))^{*_R}$ is regular. ∎

The following statement is a direct consequence of Propositions 6.11, 6.13.

Proposition 6.14 *If $G = (A, R)$ is P_4-free, then I_n is regular for any $n \geq 0$.*

7. *R*-unavoidable and strong *R*-unavoidable sets

In this section we introduce our notions of R-unavoidable and strong R-unavoidable sets.

Definition 7.1 *Let A be an alphabet, let X, Y subsets of A^+ and let R be a symmetric relation over A. Y is R-*unavoidable *in X if there exists $k_0 \in N$ such that for any x in X, with $|x| > k_0$, there exists $y \in Y$ which is a R-factor of x, i.e., $x = x_1 y x_2$, $x_2 x_1 \in A^* a$, $y \in A^* b$ and $(a, b) \in R$. The smallest k_0 satisfying the above condition is called the* avoidance bound *for Y.*

Next proposition shows that R-unavoidability is a decidable property under suitable hypotheses.

Proposition 7.2 *Let A be an alphabet, let X, Y subsets of A^+ and let R be a symmetric relation over A. If Y is a regular language and X is a context-free language, then it is decidable whether Y is R-unavoidable in X.*

PROOF :
Let Y be a regular language and let X be a context-free language. For a letter b, we set $Y_b = Y \cap A^*b$. Then, Y is R-unavoidable in X if and only if $Z = X \setminus \cup_{(a,b)\in R}(A^*aY_bA^* \cup A^*Y_bA^*a)$ is a finite set. Indeed, if Z is finite, for any word x in X, longer than any word in Z, we have $x \in \cup_{(a,b)\in R}(A^*aY_bA^* \cup A^*Y_bA^*a)$, hence x has a R-factor in Y. Conversely, if Y is R-unavoidable in X and k_0 is a subword avoidance bound for Y, no word of length greater than or equal to k_0 belongs to Z. Since R is finite, the language $\cup_{(a,b)\in R}(A^*aY_bA^* \cup A^*Y_bA^*a)$ is regular. Therefore, $Z = X \setminus \cup_{(a,b)\in R}(A^*aY_bA^* \cup A^*Y_bA^*a)$ is a context-free language [16, 20]. Since there are algorithms to determine whether a context-free language is finite [20], the conclusion holds. ∎

By Theorem 2.1, if S is a $(1,3)$-CSSH system, then $\mathrm{Lin}(L(S))$ is context-free. Moreover, it is known that if X is a context-free language and Y is a regular set, then $XY^{-1} = \{w \in A^* \mid wy \in X \text{ for some } y \in Y\}$ is context-free [16]. Thus, the set of the prefixes of $\mathrm{Lin}(L(S))$ is also context-free. Hence, by Proposition 7.2, it is decidable whether $Y = \mathrm{Lin}(I)$ is R-unavoidable in $\mathrm{Lin}(L(S))$ and in the set of the prefixes of $\mathrm{Lin}(L(S))$.

Definition 7.3 *Let A be an alphabet, let X, Y subsets of A^+ and let R be a symmetric relation over A. The set Y is* strong R-unavoidable *in X if there exists $k_0 \in \mathbb{N}$, $k_0 > 0$, such that for any word $x \in X$ of length at least k_0, there are $y \in Y$ and $x_1, x_2 \in A^*$ such that $x_1x_2 \in X$, $x = x_1yx_2$, $|x_1| \leq k_0$, $x_2x_1 \in A^*a$, $y \in A^*b$ and $(a,b) \in R$. The smallest k_0 satisfying the above condition is called the* strong avoidance bound *for Y.*

Of course, if Y is strong R-unavoidable in X, then Y is R-unavoidable in X. We do not know whether a converse of this statement holds, by eventually adding supplementary hypotheses.

8. Sufficient conditions for regularity

Lemma 8.1 *Let $S = (A, I, R)$ be a $(1,3)$-CSSH system and let $Y = \mathrm{Lin}(I)$. If Y is strong R-unavoidable in $\mathrm{Lin}(L(S))$ with strong avoidance bound k_0,* *then* $\mathrm{Lin}(L(S)) = (\cup_{n\geq 0} I_n) \setminus \{1\} = I_{k_0} \setminus \{1\}$.

PROOF :
By Theorem 6.9, $\mathrm{Lin}(L(S)) = (\cup_{n\geq 0} I_n) \setminus \{1\}$. Hence it suffices to show that $\cup_{n\geq 0} I_n = I_{k_0}$. By contradiction, assume $\cup_{n\geq 0} I_n \setminus I_{k_0} \neq \emptyset$. Let x be a shortest word in $\cup_{n\geq 0} I_n \setminus I_{k_0}$. Hence, by Propositions 6.4, 6.10, $|x| > k_0$ and, of course, $x \notin Y$.

Since k_0 is the strong avoidance bound for Y and $x \in (\cup_{n\geq 0} I_n) \setminus \{1\} = \mathrm{Lin}(L(S))$, there are $y \in Y$ and $x_1, x_2 \in A^*$ such that $x_1x_2 \in \mathrm{Lin}(L(S))$, $x = x_1yx_2$, $|x_1| \leq k_0$, $x_2x_1 \in A^*a$, $y \in A^*b$ and $(a, b) \in R$. Since x was of minimal length, $x_1x_2 \in I_{k_0}$. If $x_1 = 1$, then $x = yx_2 \in I_{k_0}$, a contradiction with the hypothesis. Otherwise, by Proposition 6.7, the word x is again in I_{k_0}, contrary to hypothesis. ■

Theorem 8.2 *Let* $S = (A, I, R)$ *be a* $(1, 3)$*-CSSH system and let* $Y = \mathrm{Lin}(I)$*, with* $A = \mathrm{alph}(Y)$*. If* Y *is strong* R*-unavoidable in* $\mathrm{Lin}(L(S))$ *and* $G = (A, R)$ *is* P_4*-free, then* $\mathrm{Lin}(L(S))$ *is regular.*

PROOF :
Let $S = (A, I, R)$ and Y be as in the statement. By Lemma 8.1, if Y is strong R-unavoidable in $\mathrm{Lin}(L(S))$ with strong avoidance bound k_0, then $\mathrm{Lin}(L(S)) = (\cup_{n\geq 0} I_n) \setminus \{1\} = I_{k_0} \setminus \{1\}$. Hence, the conclusion follows by Proposition 6.14. ■

9. Future Perspectives

In this paper we have presented a sufficient condition for the regularity of languages generated by $(1, 3)$-CSSH systems $S = (A, I, R)$ (Theorem 8.2). There are several issues that follows from this and the other results stated in this paper. Undoubtedly, the main open question is whether we may decide the strong R-unavoidability of $\mathrm{Lin}(I)$ in $\mathrm{Lin}(L(S))$. The other main question is whether a converse of Theorem 8.2 may be stated.

Regarding that, we recall that in [12], it has been proved that if $\mathrm{Lin}(L(S))$ is regular, then $\mathrm{Lin}(I)$ is unavoidable in the set $\mathrm{Pref}(\mathrm{Lin}(L(S)))$ of the prefixes of $\mathrm{Lin}(L(S))$. In particular, for the subclass of hybrid systems, defined in the same paper by a condition on R, the regularity of the splicing language implies that $\mathrm{Lin}(I)$ is unavoidable in A^*. We do not know whether this result may be strengthened. More precisely, we do not know if, at least for hybrid systems, R-unavoidability of $\mathrm{Lin}(I)$ (in A^* or in $\mathrm{Pref}(\mathrm{Lin}(L(S)))$ or in $\mathrm{Lin}(L(S))$) is a part of a set of necessary and sufficient conditions for the regularity of the splicing language.

Another question is the connection with wqo. Indeed, as already said in [13], the authors proved that a language is regular if and only if it is upward closed with respect to a monotone wqo. A quasi-order $\leq$ on A^* is monotone if, for any words u, u', v, v', $u \leq v$, $u' \leq v'$ implies $uu' \leq vv'$. Thus the problem of finding conditions under which $\mathrm{Lin}(L(S))$ is upward closed with respect to a monotone wqo arises.

Finally, we focused on $(1,3)$-CSSH systems. The notions presented here could be extended in order to eventually obtain more general results concerning regularity of languages generated by CSSH systems.

References

[1] J. Berstel, *Transductions and Context-free Languages*, B. G. Teubner, Stuttgart, (1979).

[2] J. Berstel, L. Boasson, I. Fagnot, Splicing systems and the Chomsky hierarchy, *Theor. Comp. Science* **436** (2012) 2-22.

[3] L. Boasson, P. Bonizzoni, C. De Felice, I. Fagnot, G. Fici, R. Zaccagnino, R. Zizza, Splicing Systems from Past to Future: Old and New Challenges. In Gh. Paun, G. Rozenberg, A. Salomaa (eds.), Discr. Math. and Comp. Science, (2014) 51-76.

[4] P. Bonizzoni, C. De Felice, R. Zizza, A characterization of (regular) circular languages generated by monotone complete splicing systems, *Theor. Comp. Science* **411** (2010) 4149-4161.

[5] P. Bonizzoni, C. De Felice, G. Fici, R. Zizza, On the regularity of circular splicing languages: a survey and new developments, *Nat. Computing* **9** (2010) 397-420.

[6] P. Bonizzoni, C. De Felice, G. Mauri, R. Zizza, Circular splicing and regularity, *Theor. Inform. and Appl.* **38** (2004) 189-228.

[7] P. Bonizzoni, C. De Felice, G. Mauri, R. Zizza, On the power of circular splicing, *Discr. Appl. Math.* **150** (2005) 51-66.

[8] R. Ceterchi, C. Martín-Vide, K. G. Subramanian, On some classes of splicing languages, *in*: N. Jonoska, G. Păun, G. Rozenberg (Eds.), Aspects of Molecular Computing: Essays in Honor of the 70th Birthday of Tom Head, LNCS **2950**, 83-104 (2004).

[9] R. Ceterchi, K. G. Subramanian, Simple circular splicing systems, Rom. J. of Inf. Science and Tech. 6 (2003) 121-134.

[10] C. Choffrut, J Karhumáki, Combinatorics on Words, *in*: (G. Rozenberg, A. Salomaa, Eds.) Handbook of Formal Languages, Vol. 1, 329-438, Springer Verlag, 1996.

[11] C. De Felice, G. Fici, R. Zizza, A characterization of regular circular languages generated by marked splicing systems, *Theor. Comp. Science* **410** (2009) 4937-4960.

[12] C. De Felice, R. Zaccagnino, R. Zizza, Unavoidable Sets and Regularity of Languages Generated by (1,3)-Circular Splicing Systems, *in*: TPNC 2014, Adrian Horia Dediu,Manuel Lozano and Carlos Martín-Vide (Eds.), LNCS **8890** (2014) 169-180.

[13] A. Ehrenfeucht, D. Haussler, G. Rozenberg, On regularity of context-free languages, *Theor. Comp. Science* **27** (1983) 311-332.

[14] R. Grossi, C. S. Iliopoulos, R. Mercas, N. Pisanti, S. P. Pissis, A. Retha, and F. Vayani, Circular Sequence Comparison with q-grams, *in*: Algorithms in Bioinformatics (WABI), M. Pop, H. Touzet, Eds., LNCS **9289** (2015) 203-216.

[15] D. Haussler, Insertion languages, *Inf. Sciences* **31** (1983) 77-90.

[16] M. A. Harrison, *Introduction to Formal Language Theory*, Addison Wesley Publishing Company, 1978.

[17] T. Head, Splicing schemes and DNA, in: *Lindenmayer Systems: Impacts on Theoretical Computer Science and Developmental Biology*, Springer-Verlag, Berlin (1992) 371-383.

[18] T. Head, G. Păun, D. Pixton, Language theory and molecular genetics: generative mechanisms suggested by DNA recombination, *in*: (G. Rozenberg, A. Salomaa, Eds.) Handbook of Formal Languages, Vol. 2, 295-360, Springer Verlag, 1996.

[19] G. Higman, Ordering by divisibility in abstract algebras, *Proc. London Math. Soc.* **3** (2) (1952) 326-336.

[20] J.E. Hopcroft, R. Motwani, J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison Wesley Pearson Education, 3rd Edition, 2008.

[21] M. Kudlek, Languages of cyclic words, in: N. Jonoska, G. Păun, G. Rozenberg (Eds.), *Aspects of Molecular Computing*, Essays dedicated to Tom Head, LNCS **2950**, 278-288, Springer, 2004.

[22] M. Lothaire, Combinatorics on Words, Cambridge University Press, 1997.

[23] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, 2002.

[24] D. Pixton, Regularity of splicing languages, *Discr. Appl. Math.* **69** (1996) 101-124.

[25] M.-Paul Schützenberger, On the synchronizing properties of certain prefix codes, *Inf. and Control*, **7** (1) (1964) 23-36.

[26] M.Sipser, *Introduction to Theory of Computation*, Cengage Learning, 3rd Edition, 2013.

[27] R. Siromoney, K.G. Subramanian, A. Dare, Circular DNA and splicing systems, in: *Proc. of ICPIA*, LNCS **654**, 260-273. Springer, 1992.