Return words in tree sets

Valérie Berthé* Clelia De Felice† Francesco Dolce‡ Julien Leroy(*)§
Dominique Perrin‡ Christophe Reutenauer¶ Giuseppina Rindone‡

April 2014

Abstract

Given a set F of words, one associates to each word w in F an undirected graph, called its extension graph, and which describes the possible extensions of w in F on the left and on the right. We investigate the family of sets of words defined by the property of the extension graph of each word in the set to be acyclic or connected or a tree. We prove that in a uniformly recurrent tree set, the sets of first return words are bases of the free group on the alphabet. We also study S-adic representations of such sets.

1 Introduction

This paper studies properties of classes of sets which occur as the set of factors of infinite words of linear factor complexity. It is a mix of a series of papers [5], [6], [7], [8], [9] devoted to this subject initiated in [3]. These classes of sets, called acyclic, connected or tree sets, are defined by a limitation to the possible two-sided extensions of a word of the set. We will see that Sturmian sets are tree sets. Moreover, the sets obtained by coding a regular interval exchange set are also tree sets (see [6]). Any word w in a tree set is neutral in the sense that the number of pairs (a, b) of letters such that $awb \in F$ is equal to the number of letters a such that $aw \in F$ plus the number of letters b such that $wb \in F$ minus 1. We call such a set a neutral set.

We study sets of first return words in a tree set F. For this, we use Rauzy graphs, which are restrictions of a de Bruijn graph to the set of vertices formed by the words of given length in a set F. We first show that if F is a recurrent connected set, the group described by any Rauzy graph of F containing the alphabet A, with respect to some vertex is the free group on A (Theorem 9). Next, we prove that in a uniformly recurrent connected set containing A, the set of first return words to any word in F generates the free group on A (Theorem 11). Next, we prove that if F is a uniformly recurrent tree set containing A, the set of first return words to any word of F is a basis of the free group on A (Corollary 13). The proof uses the fact that in a uniformly recurrent neutral set F containing the alphabet A, the number of first return words to any word of F is equal to $\mathrm{Card}(A)$, a result obtained in [1].

We also show that the class of uniformly recurrent tree sets is closed under decoding by return words. This means that if F is a uniformly recurrent tree set and f a coding morphism for the set of return words to a word $w \in F$, then $f^{-1}(F)$ is a uniformly recurrent tree set. This result allows us to build S-adic representations of uniformly recurrent tree sets where S is the

*CNRS, Université Paris 7
†Università degli Studi di Salerno
‡Université Paris Est, LIGM
§Université du Luxembourg
¶Université du Québec à Montréal

set of elementary automorphisms of the free group generated by the alphabet. In the case of a ternary alphabet, we get an S-adic characterization of uniformly recurrent tree sets. This characterization can be expressed by using a Büchi automaton.

2 Preliminaries

2.1 Uniformly recurrent sets and factor complexity

Let A be a finite nonempty alphabet. All words considered below, unless stated explicitly, are supposed to be on the alphabet A. We denote by 1 or by ε the empty word.

Let F be a set of words on the alphabet A. For $w \in F$, we denote

$$\begin{aligned} L(w) &= \{a \in A \mid aw \in F\} & \ell(w) &= \mathrm{Card}(L(w)) \\ R(w) &= \{a \in A \mid wa \in F\} & r(w) &= \mathrm{Card}(R(w)) \\ E(w) &= \{(a,b) \in A \times A \mid awb \in F\} & e(w) &= \mathrm{Card}(E(w)) \end{aligned}$$

A set of words F is *factorial* if $\mathrm{Fact}(F) \subset F$. It is *biessential* if it is factorial and if for all $w \in F$, one has $r(w) \geq 1$ and $\ell(w) \geq 1$. It is *recurrent* if it is biessential and for all $u, v \in F$, there is w such that $uwv \in F$. It is *uniformly recurrent sets* if it is biessential and for any $u \in F$, there is an integer $n \geq 1$ such that u is a factor of every word of F of length n.

A word w is called *right-special* if $r(w) \geq 2$. It is called *left-special* if $\ell(w) \geq 2$. It is called *bispecial* if it is both right and left-special. For a word $w \in F$, let $m(w) = e(w) - \ell(w) - r(w) + 1$. We say that w is *strong* if $m(w) > 0$, *weak* if $m(w) < 0$ and *neutral* if $m(w) = 0$. A word w is called *ordinary* if $E(w) \subset a \times A \cup A \times b$ for some $(a,b) \in E(w)$ (see [10], Chapter 4). Any ordinary word is neutral.

A factorial set F is said to be *neutral* (resp. *weak*, resp. *strong*) if any word of F is neutral (resp. neutral of weak, resp. neutral or strong). The sequence $(p_n)_{n \geq 0}$ with $p_n = \mathrm{Card}(F \cap A^n)$ is called the *factor complexity* of F. Set $k = \mathrm{Card}(F \cap A) - 1$.

Proposition 1 *The factor complexity of a strong (resp. weak, resp. neutral) set F is at least (resp. at most, resp. exactly) equal to $kn + 1$.*

An infinite word is *episturmian* if the set of its factors is closed under reversal and contains for each n at most one word of length n which is right-special (see [3] for more references). It is a *strict episturmian* word if it has exactly one right-special word of each length and moreover each right-special factor u is such that $r(u) = \mathrm{Card}(A)$.

A *Sturmian set* is a set of words which is the set of factors of a strict episturmian word. Any Sturmian set is uniformly recurrent (see [3]).

Example 2 *Let $A = \{a, b\}$. The* Fibonacci morphism *is the morphism $f : A^* \to A^*$ defined by $f(a) = ab$ and $f(b) = a$. The* Fibonacci word *$x = abaababaabaababaababa \cdots$ is the fixpoint $x = f^\omega(a)$ of the Fibonacci morphism. It is a Sturmian word (see [19]). The set $F(x)$ of factors of x is the* Fibonacci set.

Example 3 *Let $A = \{a, b, c\}$. The Tribonacci word $x = abacabaabacababacabaabacaba \cdots$ is the fixpoint $x = f^\omega(a)$ of the morphism $f : A^* \to A^*$ defined by $f(a) = ab$, $f(b) = ac$, $f(c) = a$. It is a strict episturmian word (see [16]). The set $F(x)$ of factors of x is the* Tribonacci set.

2.2 Automata and free groups

We denote by A° the free group on the alphabet A. It is the set of all words on the alphabet $A \cup A^{-1}$ which are *reduced*, in the sense that they do not have any factor aa^{-1} or $a^{-1}a$ for $a \in A$. We extend the bijection $a \mapsto a^{-1}$ to an involution on $A \cup A^{-1}$ by defining $(a^{-1})^{-1} = a$.

For any word w on $A \cup A^{-1}$ there is a unique reduced word $\rho(w)$ equivalent to w modulo the relations $aa^{-1} \equiv a^{-1}a \equiv 1$ for $a \in A$. The product of two elements $u, v \in A^\circ$ is the reduced word w equivalent to uv, namely $\rho(uv)$. If $w = a_1 \cdots a_n$ with $a_i \in A \cup A^{-1}$ is a reduced word, its inverse is the reduced word denoted w^{-1} and defined by $w^{-1} = a_n^{-1} \cdots a_1^{-1}$.

We denote $\mathcal{A} = (Q, i, T)$ a deterministic automaton with a set Q of states, $i \in Q$ as initial state and $T \subset Q$ as set of terminal states. For $p \in Q$ and $w \in A^*$, we denote $p \cdot w = q$ if there is a path labeled w from p to the state q and $p \cdot w = \emptyset$ otherwise. The automaton is *finite* when Q is finite. The set *recognized* by the automaton is the set of words $w \in A^*$ such that $i \cdot w \in T$.

All automata considered in this paper are deterministic and we simply call them 'automata' to mean 'deterministic automata'. The automaton $\mathcal{A}$ is *trim* if for any $q \in Q$, there is a path from i to q and a path from q to some $t \in T$. An automaton is called *simple* if it is trim and if it has a unique terminal state which coincides with the initial state. The set recognized by a simple automaton is a right unitary submonoid of A^*.

Let $\mathcal{A} = (Q, i, T)$ be an automaton. A *generalized path* is a sequence $(p_0, a_1, p_1, a_2, \ldots, p_{n-1}, a_n, p_n)$ with $a_i \in A \cup A^{-1}$ and $p_i \in Q$, such that for $1 \leq i \leq n$, one has $p_{i-1} \cdot a_i = p_i$ if $a_i \in A$ and $p_i \cdot a_i^{-1} = p_{i-1}$ if $a_i \in A^{-1}$. The *label* of the generalized path is the reduced word equivalent to $a_1 a_2 \cdots a_n$. It is an element of the free group A°. The set *described* by the automaton is the set of labels of generalized paths from i to a state in T. Since a path is a particular case of a generalized path, the set recognized by an automaton $\mathcal{A}$ is a subset of the set described by $\mathcal{A}$. The set described by a simple automaton is a subgroup of A°. It is called the *subgroup described* by $\mathcal{A}$.

2.3 Return words

Let F be a set of words. For $w \in F$, let $\Gamma_F(w) = \{x \in F \mid wx \in F \cap A^+ w\}$ be the set of *right return words* Let $\mathcal{R}_F(w) = \Gamma_F(w) \setminus \Gamma_F(w)A^+$ be the set of *first right return words.*

The following result has been proved for neutral sets in [1].

Theorem 4 *Let F be a uniformly recurrent set containing the alphabet A. If F is strong (resp. weak, resp. neutral), then for every $w \in F$, the set $\mathcal{R}_F(w)$ has at least (resp. at most, resp. exactly)* $\mathrm{Card}(A)$ *elements.*

3 Acyclic, connected and tree sets

Let F be a set of words. For a word $w \in F$, we consider an undirected graph $G(w)$ called its *extension graph* in F and defined as follows. The set of vertices is the disjoint union of $L(w)$ and $R(w)$ and its edges are the pairs $(a, b) \in E(w)$.

Example 5 *Let F be the Tribonacci set (see Example 3). The graphs $G(\varepsilon)$ and $G(ab)$ are represented in Figure 1.*

We say that F is an *acyclic* (resp. a connected, resp. a tree) set if it is biessential and if for every word $w \in F$, the graph $G(w)$ is acyclic (resp. connected, resp. a tree). Obviously, a tree set is acyclic and connected.

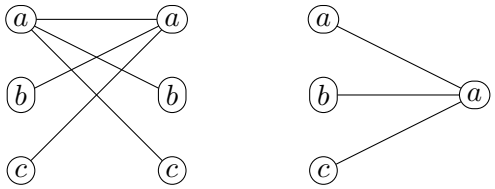

Figure 1: The extension graphs $G(\varepsilon)$ and $G(ab)$ in the Tribonacci set.

Note that a biessential set F is acyclic (resp. connected) if and only if the graph $G(w)$ is acyclic (resp. connected) for every bispecial word w. Indeed, if w is not bispecial, then $G(w) \subset a \times A$ or $G(w) \subset A \times a$, thus it is always acyclic and connected.

If the extension graph $G(w)$ of w is acyclic, then $m(w) \leq 0$. Thus w is weak or neutral. More precisely, one has in this case, $m(w) = -c + 1$ where c is the number of connected components of the graph $G(w)$.

Similarly, if $G(w)$ is connected, then w is strong or neutral. Thus, if F is an acyclic (resp. a connected, resp. a tree) set, then F is a weak (resp. strong, resp. neutral) set.

Example 6 *A Sturmian set F is a tree set. Indeed, any word $w \in F$ is ordinary, which implies that $G(w)$ is a tree.*

Since a tree set is neutral, we deduce from Proposition 1 the following statement, where $k = \mathrm{Card}(F \cap A) - 1$.

Proposition 7 *The factor complexity of a tree set is $kn + 1$.*

We now give an example of a set of complexity $2n + 1$ on an alphabet with three letters which is not neutral (hence not a tree set).

Example 8 *Let $A = \{a, b, c\}$. The* Chacon word *on three letters is the fixpoint $x = f^\omega(a)$ of the morphism f from A^* into itself defined by $f(a) = aabc$, $f(b) = bc$ and $f(c) = abc$. Thus $x = aabcaabcbcabc \cdots$. The* Chacon set *is the set F of factors of x. It is of complexity $2n + 1$ (see [15] Section 5.5.2). It contains strong, neutral and weak words. Indeed, $F \cap A^2 = \{aa, ab, bc, ca, cb\}$ and thus $m(\varepsilon) = 0$ showing that the empty word is neutral. Next $E(abc) = \{(a, a), (c, a), (a, b), (c, b)\}$ shows that $m(abc) = 1$ and thus abc is strong. Finally, $E(bca) = \{(a, a), (c, b)\}$ and thus $m(bca) = -1$ showing that bca is weak.*

4 Return words in tree sets

4.1 Stallings foldings of Rauzy graphs

Let F be a factorial set. The *Rauzy graph* of F of order $n \geq 0$ is the following labeled graph $G_n(F)$. Its vertices are the words in the set $F \cap A^n$. Its edges are the triples (x, a, y) for all $x, y \in F \cap A^n$ and $a \in A$ such that $xa \in F \cap Ay$.

When F is recurrent, all Rauzy graphs $G_n(F)$ are strongly connected. Thus, the Rauzy graph $G_n(F)$ of a recurrent set F with a distinguished vertex v can be considered as a simple automaton $\mathcal{A} = (Q, v, v)$ with set of states $Q = F \cap A^n$ (see Section 2.2).

Let G be a labeled graph on a set Q of vertices. The group described by G with respect to a vertex v is the subgroup described by the simple automaton (Q, v, v).

A *stalling folding* of an automaton $\mathcal{A} = (Q, i, T)$ consists in merging two distinct states $p, q \in Q$ and $a \in A$ such that $p \cdot a = q \cdot a$. The next result is obtained by stallings foldings of Rauzy graphs.

Theorem 9 *Let F be a recurrent connected set containing the alphabet A. The group described by a Rauzy graph of F with respect to any vertex is the free group on A.*

The following example shows that Theorem 9 is false for sets which are not connected.

Example 10 *Consider again the Chacon set (see Example 8). The Rauzy graph $G_1(F)$ corresponding to the Chacon set is represented in Figure 2. The group described by $G_1(F)$ with respect to the state a is the subgroup of A° generated by $\{a, bc\}$.*

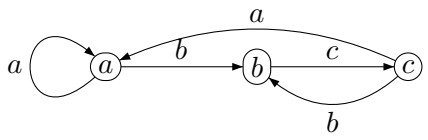

Figure 2: The graphs $G_1(F)$.

4.2 Return words and bases of free groups

We prove the following result.

Theorem 11 *Let F be a uniformly recurrent connected set containing the alphabet A. For any $w \in F$, the set $\mathcal{R}_F(w)$ generates the free group on A.*

Example 12 *Let F be the Fibonacci set. We have $\mathcal{R}_F(aa) = \{baa, babaa\}$ which generates the free group as $a = (baa)(babaa)^{-1}(baa)$ and $b = (baa)a^{-1}a^{-1}$.*

Note that Theorem 11 implies that $\mathrm{Card}(\mathcal{R}_F(w)) \geq \mathrm{Card}(A)$. This is also a consequence of Theorem 4. When F is a tree set, Theorem 4 implies that $\mathrm{Card}(\mathcal{R}_F(w)) = \mathrm{Card}(A)$. Thus we have the following corollary.

Corollary 13 *Let F be a uniformly recurrent tree set containing the alphabet A. Then for any $w \in F$, the set $\mathcal{R}_F(w)$ is a basis of the free group on A.*

4.3 Tame bases

An automorphism of the free group on A is *tame* if it belongs to the submonoid generated by the permutations of A and the automorphisms $\alpha_{a,b}$, $\tilde{\alpha}_{a,b}$ defined for $a, b \in A$ with $a \neq b$ by

$$\alpha_{a,b}(c) = \begin{cases} ab & \text{if } c = a \\ c & \text{otherwise} \end{cases}, \quad \tilde{\alpha}_{a,b}(c) = \begin{cases} ba & \text{if } c = a \\ c & \text{otherwise} \end{cases}$$

The above automorphisms and the permutations of A are called the *elementary* automorphisms on A. A basis X of the free group is *tame* if there exists a tame automorphism α such that $X = \alpha(A)$.

Example 14 *The set $X = \{ba, cba, cca\}$ is a tame basis of the free group on $\{a, b, c\}$. Indeed,one has $(b, c, a) \xrightarrow{\alpha_{c,b}} (b, cb, a) \xrightarrow{\tilde{\alpha}^2_{a,c}} (b, cb, cca) \xrightarrow{\alpha_{b,a}} (ba, cba, cca)$*

Example 15 ([21]) *The set $X = \{ab, acb, acc\}$ is a basis of the free group on $\{a, b, c\}$ but it is not a tame basis.*

Theorem 16 *Any basis of the free group included in a uniformly recurrent tree set is tame.*

Corollary 17 *If F is a uniformly recurrent tree set, then for any $w \in F$, the set $\mathcal{R}_F(w)$ is a tame basis of A°.*

5 S-adic representations

5.1 Derived sets of tree sets

Let F be a uniformly recurrent tree set on A and let $w \in F$. A *coding morphism* for $\mathcal{R}_F(w)$ is a morphism $f : A^* \to \mathcal{R}_F(w)^*$ which maps A bijectively onto $\mathcal{R}_F(w)$. If f is a coding morphism for $\mathcal{R}_F(w)$, then $f^{-1}(F)$ is called a *derived set* of F(see [12]).

The following closure property of the family of uniformly recurrent tree sets. generalizes the fact that the derived word of a Sturmian word is Sturmian (see [16]).

Theorem 18 *Any derived set of a uniformly recurrent tree set is a uniformly recurrent tree set.*

5.2 S-adic representation of tree sets

Let S be a set of morphisms and $\mathbf{s} = (\sigma_n)_{n\in\mathbb{N}}$ be a sequence in $S^{\mathbb{N}}$ with $\sigma_n : A_{n+1}^* \to A_n^*$. We let $F_{\mathbf{s}}$ denote the set of words $\bigcap_{n\in\mathbb{N}} \mathrm{Fact}(\sigma_0 \cdots \sigma_n(A_{n+1}^*))$. We call a factorial set F an *S-adic set* if there exists $\mathbf{s} \in S^{\mathbb{N}}$ such that $F = F_{\mathbf{s}}$. In this case, the sequence $\mathbf{s}$ is called an *S-adic representation* of F.

A sequence of morphisms $(\sigma_n)_{n\in\mathbb{N}}$ is said to be *primitive* if for all $r \geq 0$ there exists $s > r$ such that all letters of A_r occur in all images $\sigma_r \cdots \sigma_{s-1}(a)$, $a \in A_s$.

A uniformly recurrent set F is said to be *aperiodic* if it contains at least one right special factor of each length. The next proposition is based on return words.

Proposition 19 *An aperiodic factorial set $F \subset A^*$ is uniformly recurrent if and only if it has a primitive S-adic representation for some (possibly infinite) set S of morphisms.*

Even for uniformly recurrent sets with linear factor complexity, the set of morphisms S of Proposition 19 usually is infinite as well as the sequence of alphabets $(A_n)_{n\in\mathbb{N}}$ usually is unbounded (see [13]). For tree sets F, the next theorem significantly improves the only if part of Proposition 19: For such sets, the set S can be replaced by the set $\mathcal{S}_e$ of elementary positive automorphisms. In particular, A_n is equal to A for all n.

Theorem 20 *If F is a uniformly recurrent tree set over an alphabet A, then it has a primitive $\mathcal{S}_e$-adic representation.*

5.3 The case of a ternary alphabet

Recall that a *Büchi automaton* is an automaton with a condition of acceptance adapted to infinite words. An infinite word is accepted by such an automaton if it labels an infinite path starting in an initial state and visiting infinitely often terminal states.

Using an S-adic characterization of uniformly recurrent sets of complexity $p_n = 2n+1$ obtained in [18], we obtain the following result.

Theorem 21 *There exists a Büchi automaton $\mathcal{A}$ over the alphabet $\mathcal{S}_3$ such that F is a uniformly recurrent tree set if and only if it has an $\mathcal{S}_3$-adic representation accepted by $\mathcal{A}$.*

References

[1] L. Balková, E. Pelantová, and W. Steiner. Sequences with constant number of return words. *Monatsh. Math.*, 155(3-4):251–263, 2008.

[2] L. Bartholdi and P. Silva. Rational subsets of groups. In *Handbook of Automata*. European science Foundation, 2011.

[3] J. Berstel, C. De Felice, D. Perrin, C. Reutenauer, and G. Rindone. Bifix codes and Sturmian words. *J. Algebra*, 369:146–202, 2012.

[4] J. Berstel, D. Perrin, and C. Reutenauer. *Codes and Automata*. Cambridge University Press, 2009.

[5] V. Berthé, C. De Felice, F. Dolce, J. Leroy, D. Perrin, C. Reutenauer, and G. Rindone. Acyclic, connected and tree sets. 2013.

[6] V. Berthé, C. De Felice, F. Dolce, J. Leroy, D. Perrin, C. Reutenauer, and G. Rindone. The finite index basis property. 2013.

[7] V. Berthé, C. De Felice, F. Dolce, J. Leroy, D. Perrin, C. Reutenauer, and G. Rindone. Maximal bifix decoding. 2013.

[8] V. Berthé, C. De Felice, V. Delecroix, F. Dolce, D. Perrin, C. Reutenauer, and G. Rindone. Natural coding of linear involutions. 2013.

[9] V. Berthé, C. De Felice, F. Dolce, D. Perrin, C. Reutenauer, and G. Rindone. Two-sided rauzy induction. 2013. `http://arxiv.org/abs/1305.0120`.

[10] V. Berthé and M. Rigo, editors. *Combinatorics, automata and number theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010.

[11] J. Cassaigne. Complexité et facteurs spéciaux. *Bull. Belg. Math. Soc. Simon Stevin*, 4(1):67–88, 1997. Journées Montoises (Mons, 1994).

[12] F. Durand. A characterization of substitutive sequences using return words. *Discrete Math.*, 179(1-3):89–101, 1998.

[13] F. Durand, J. Leroy, and G. Richomme. Do the properties of an S-adic representation determine factor complexity? *J. Integer Seq.*, 16(2):Article 13.2.6, 30, 2013.

[14] S. Ferenczi. Rank and symbolic complexity. *Ergodic Theory Dynam. Systems*, 16(4):663–682, 1996.

[15] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.

[16] J. Justin and L. Vuillon. Return words in Sturmian and episturmian words. *Theor. Inform. Appl.*, 34(5):343–356, 2000.

[17] I. Kapovich and A. Myasnikov. Stallings foldings and subgroups of free groups. *J. Algebra*, 248(2):608–668, 2002.

[18] J. Leroy. An S-adic characterization of minimal subshifts with first difference of complexity $1 \leq p(n+1) - p(n) \leq 2$. *DMTCS*, 16 (1),2014.

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

[20] R. C. Lyndon and P. E. Schupp. *Combinatorial Group Theory*. Classics in Mathematics. Springer-Verlag, 2001. Reprint of the 1977 edition.

[21] B. Tan, Z.-X. Wen, and Y. Zhang. The structure of invertible substitutions on a three-letter alphabet. *Adv. in Appl. Math.*, 32(4):736–753, 2004.