Neutral and tree sets of arbitrary characteristic

Francesco Dolce and Dominique Perrin

Université Paris-Est, LIGM

Abstract

We study classes of minimal sets defined by restrictions on the possible extensions of the words. These sets generalize the previously studied classes of neutral and tree sets by relaxing the condition imposed on the empty word and measured by an integer called the characteristic of the set. We present several enumeration results holding in these sets of words. These formulae concern return words and bifix codes. They generalize formulae previously known for Sturmian sets or more generally for tree sets. We also give two geometric examples of this class of sets, namely the natural coding of some interval exchange transformations and the natural coding of some linear involutions.

Keywords: Neutral Sets, Tree Sets, Bifix Codes, Interval Exchanges, Linear Involutions

1. Introduction

Sets of words of linear complexity play an important role in combinatorics on words and symbolic dynamics. This family of sets includes set of factors of Sturmian and Arnoux-Rauzy words, interval exchange sets and primitive morphic sets, that is, sets of factors of fixed points of primitive morphisms.

We study here several families of sets of words of linear complexity defined by properties of a graph $E(x)$, called the extension graph of x. This graph expresses the possible extensions of x on both sides by a letter of the alphabet A.

A set S is neutral if the Euler characteristic of the graph of any nonempty word is equal to 1. A special family of neutral sets is given by tree sets, sets such that $E(x)$ is a tree for every nonempty word and acyclic for every word. The Euler characteristic of the graph $E(\varepsilon)$ is called the characteristic of S and is denoted by $\chi(S)$. These sets were first considered in [1] and in [4].

The motivation for studying neutral and tree sets is the following: First, the family of uniformly recurrent tree sets appears as the natural closure of two known families, namely the Sturmian sets and the interval exchange sets. Next, the family of neutral sets is a naturally defined generalization of tree sets for which a number of properties true for tree sets still hold.

The factor complexity of a neutral set S on k letters is for $n \neq 1$

$$p_n = n(k - \chi(S)) + \chi(S). \tag{1}$$

We prove here several results concerning neutral sets. The first one (Theorem 4.9) is a formula giving the cardinality of a finite S-maximal bifix code X of S-degree n in a recurrent neutral set S on k letters as

$$\mathrm{Card}(X) = n(k - \chi(S)) + \chi(S). \tag{2}$$

The remarkable feature is that, for fixed S, the cardinality of X depends only on its S-degree. In the particular case where X is the set of all words of S of length n, we recover Equation (1). Formula (2) generalizes the formula proved in [2] for Sturmian sets and in [6] for neutral sets of characteristic 1.

The second one concerns return words. The set of right return words to a word x in a factorial set S, denoted by $\mathcal{R}_S(x)$, is an important notion. It is the set of words u such that xu is in S and ends with x for the first time. In several families of sets of linear complexity, the set of return words to x is known to be of fixed cardinality independent of x. This was proved for Sturmian words in [14], for interval exchange sets in [17] (see also [11]) and for neutral sets of characteristic zero in [1].

We first prove here (Theorem 5.2) that the set $\mathcal{CR}_S(X)$ of complete return words to a bifix code X (satisfying additional hypotheses) in a recurrent neutral set S on k letters satisfies $\mathrm{Card}(\mathcal{CR}_S(X)) = \mathrm{Card}(X) + k - \chi(S)$ and that this quantity is an upper bound for $\mathrm{Card}(\mathcal{CR}_S(X))$ for every neutral set. The remarkable feature here is that, for fixed S, the cardinality of $\mathcal{CR}_S(X)$ depends only on $\mathrm{Card}(X)$. When X is reduced to one element x, we have $\mathcal{CR}_S(x) = x\mathcal{R}_S(x)$ and we recover the result of [1]. When $X = S \cap A^n$, then $\mathcal{CR}_S(X) = S \cap A^{n+1}$. This implies $p_{n+1} = p_n + k - \chi(S)$ and also gives Equation (1) by induction on n. The proofs of these formulæ use a probability distribution naturally defined on a neutral set.

As a corollary of Theorem 5.2 we prove that in neutral sets the notions of recurrence and uniformly recurrence coincide (Corollary 5.3).

A third result concerns the decoding of a neutral set by a bifix code. We prove that the decoding of any recurrent neutral set S by an S-maximal bifix code is a neutral set. We find an analogous result also for tree sets. This property is proved for uniformly recurrent tree sets in [7].

We study in more detail the decoding of a neutral set S by special bifix codes called modular codes. These bifix codes have S-degree equal to the characteristic of S. We prove that the decoding of a recurrent tree set by the modular code is union of c recurrent tree sets of characteristic 1 (Theorem 7.3). The result is proved for uniformly recurrent tree sets of characteristic 2 in [8].

We finally prove two results which allows one to obtain a large family of neutral sets (actually tree sets) of geometric origin, namely using interval exchange transformations or linear involutions. More precisely, we prove that the natural coding of an interval exchange transformation without connections of length ≥ 1 is a tree set and that the natural coding of a linear involution without connections is a tree set of characteristic 2. This extends a result in [5] concerning

interval exchange without connections as well as a result of [9] concerning linear involutions without connection.

Acknowledgement.. This paper is an extended version of a conference paper [13]. This work was supported by grants from Région Île-de-France and ANR project Eqinocs ANR-13-BS02-004.

2. Extension graphs

Let A be a finite alphabet. We denote by A^* the set of all words on A. We denote by ε the empty word. A *factor* of a word x is a word v such that $x = uvw$. If both u and w are nonempty, we say that x is an *internal factor*. A set of words on the alphabet A is said to be *factorial* if it contains the factors of its elements as well as the alphabet A.

Let S be a factorial set on the alphabet A. For $w \in S$, we define

$$\begin{aligned} L_S(w) &= \{a \in A \mid aw \in S\}, \\ R_S(w) &= \{a \in A \mid wa \in S\}, \\ E_S(w) &= \{(a,b) \in A \times A \mid awb \in S\} \end{aligned}$$

and furthermore

$$\ell_S(w) = \mathrm{Card}(L_S(w)), \quad r_S(w) = \mathrm{Card}(R_S(w)), \quad e_S(w) = \mathrm{Card}(E_S(w)).$$

We omit the subscript S when it is clear from the context. A word w is *right-extendable* if $r(w) > 0$, *left-extendable* if $\ell(w) > 0$ and *biextendable* if $e(w) > 0$. A factorial set S is called *right-extendable* (resp. *left-extendable*, resp. *biextendable*) if every word in S is right-extendable (resp. left-extendable, resp. biextendable).

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 left-special and right-special. For $w \in S$, we define

$$m_S(w) = e_S(w) - \ell_S(w) - r_S(w) + 1.$$

A word w is called *neutral* if $m_S(w) = 0$. We say that a set S is *neutral* if it is factorial and every nonempty word $w \in S$ is neutral. The *characteristic* of S is the integer $\chi(S) = 1 - m_S(\varepsilon)$.

Thus, a neutral set of characteristic 1 is such that all words (including the empty word) are neutral. This is what is called a neutral set in [4].

Example 2.1 Let $A = \{a, b\}$ and let φ be the morphism from A^* to itself defined by $\varphi : a \mapsto ab, b \mapsto a$. Let S be the set of factors of the fixed point $x = \varphi^\omega(a)$. The set S, called the *Fibonacci set* is a neutral set of characteristic 1. Indeed one can prove that every word, including the empty word, is neutral.

The following example of a neutral set of characteristic larger than 1 is from [4].

Example 2.2 Let $A = \{a, b, c, d\}$ and let σ be the morphism from A^* into itself defined by $\sigma : a \mapsto ab,\ b \mapsto cda,\ c \mapsto cd,\ d \mapsto abc$. Let S be the set of factors of the infinite word $x = \sigma^\omega(a)$. One has $S \cap A^2 = \{ab, ac, bc, ca, cd, da\}$ and thus $m(\varepsilon) = -1$. It is shown in [4] that every nonempty word is neutral. Thus S is neutral of characteristic 2.

A set of words $S \neq \{\varepsilon\}$ is *recurrent* if it is factorial and for any $u, w \in S$, there is a $v \in S$ such that $uvw \in S$. An infinite factorial set is said to be *uniformly recurrent* if for any word $u \in S$ there is an integer $n \geq 1$ such that u is a factor of any word of S of length n. A uniformly recurrent set is recurrent.

The *factor complexity* of a factorial set S of words on an alphabet A is the sequence $p_n = \mathrm{Card}(S \cap A^n)$. Let $s_n = p_{n+1} - p_n$ and $b_n = s_{n+1} - s_n$ be respectively the first and second order differences sequences of the sequence p_n.

The following result is [12, Proposition 3.5] (see also [10, Theorem 4.5.4]).

Proposition 2.3 *Let S be a factorial set on the alphabet A. One has $b_n = \sum_{w \in S \cap A^n} m(w)$ and $s_n = \sum_{w \in S \cap A^n} (r(w) - 1)$ for all $n \geq 0$.*

One deduces easily from Proposition 2.3 the following result which shows that a neutral set has linear complexity.

Proposition 2.4 *The factor complexity of a neutral set on k letters is given by $p_0 = 1$ and $p_n = n(k - \chi(S)) + \chi(S)$ for every $n \geq 1$.*

Proof. Since S contains the empty word and the alphabet, we have $p_0 = 1$ and $p_1 = k$. Thus $s_0 = k - 1$

By Proposition 2.3 one has $b_0 = m(\varepsilon) = 1 - \chi(S)$ and $b_n = 0$ for every $n > 0$. Thus $s_n = k - \chi(S)$ for every $n > 0$.

The conclusion immediately follows by induction on n. ∎

Let S be a biextendable set of words. For $w \in S$, we consider the set $E(w)$ as an undirected graph on the set of vertices which is the disjoint union of $L(w)$ and $R(w)$ with edges the pairs $(a, b) \in E(w)$. This graph is called the *extension graph* of w. We sometimes denote by $1 \otimes L(w)$ and $R(w) \otimes 1$ the copies of $L(w)$ and $R(w)$ used to define the set of vertices of $E(w)$. We note that since $E(w)$ has $\ell(w) + r(w)$ vertices and $e(w)$ edges, the number $1 - m_S(w)$ is the Euler characteristic of the graph $E(w)$ [1].

A biextendable set S is called a *tree set* of characteristic c if for any nonempty $w \in S$, the graph $E(w)$ is a tree and if $E(\varepsilon)$ is a union of c trees (the definition of tree set in [4] corresponds to a tree set of characteristic 1). Note that a tree set of characteristic c is a neutral set of characteristic c.

Example 2.5 Let S be the neutral set of Example 2.2. The graph $E(\varepsilon)$ is represented in Figure 1. It is acyclic with two connected components. It is

[1] We consider here graphs as 1-dimensional compexies and thus they have no faces.

shown in [4] that the extension graph of any nonempty word is a tree. Thus S is a tree set of characteristic 2.

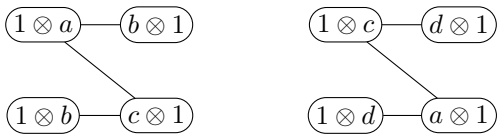

Figure 1: The two trees forming the graph $E(\varepsilon)$. Vertices correspond to letters, while edges correspond to words of length 2 in S.

Let S be a factorial set. For $x \in S$, we define

$$\rho_S(x) = e_S(x) - \ell_S(x), \quad \lambda_S(x) = e_S(x) - r_S(x).$$

Thus, when x is neutral, $\rho_S(x) = r_S(x) - 1$ and $\lambda_S(x) = \ell_S(x) - 1$. The following result shows that in a biextendable neutral set, ρ_S is a left probability distribution on S (and λ_S is a right probability), except for the value on ε which is $\rho(\varepsilon) = e(\varepsilon) - \ell(\varepsilon) = m(\varepsilon) + r(\varepsilon) - 1 = \mathrm{Card}(A) - \chi(S)$ and can be different from 1 (see [2] for the definition of a right or left probability distribution). We omit the subscript S when it is clear from the context.

Proposition 2.6 *Let S be a biextendable neutral set. Then for any $x \in S$, one has $\lambda_S(x), \rho_S(x) \geq 0$ and*

$$\sum_{a \in L(x)} \rho_S(ax) = \rho_S(x), \qquad \sum_{a \in R(x)} \lambda_S(xa) = \lambda_S(x).$$

Proof. Since S is biextendable, we have $\ell(x), r(x) \leq e(x)$. Thus $\lambda(x), \rho(x) \geq 0$. Next, $\sum_{a \in L(x)} \rho(ax) = \sum_{a \in L(x)} (r(ax) - 1) = e(x) - \ell(x) = \rho(x)$. The proof for λ is symmetric. ∎

If in a neutral set S we have $\rho(\varepsilon) = 0$, then $\rho(x) = 0$ for all $x \in S$. Otherwise, $\rho'(x) = \rho(x)/\rho(\varepsilon)$ is a left probability distribution. A symmetric result holds for λ.

3. Multiplying maps

We now introduce a construction which allows one to build tree sets of characteristic m starting from a tree set of characteristic 1.

A *transducer* is a labeled graph with vertices in a set Q and edges labeled in $\Sigma \times A$. The set Q is called the set of states, the set Σ is called the *input alphabet* and A is called the *output alphabet*. The graph obtained by erasing the output letters is called the *input automaton* (with an unspecified initial state). Similarly, the *output automaton* is obtained by erasing the input letters.

Let $\mathcal{A}$ be a transducer with set of states $Q = \{0, 1, \ldots, m-1\}$ on the input alphabet Σ and the output alphabet A. We assume that

1. the input automaton is a group automaton, that is, every letter of Σ acts on Q as a permutation,
2. the output labels of the edges are all distinct.

We define m maps $\delta_k : \Sigma^* \to A^*$ corresponding to the initial state k, for $k = 0, 1, \ldots, m-1$. Let $\delta_k(u) = v$ if the path starting at state k with input label u has output v. An m-tuple $\delta = (\delta_0, \delta_1, \ldots, \delta_{m-1})$ is called a *m-multiplying map* and the transducer $\mathcal{A}$ a *m-multiplying transducer*. The *image* of a set of words T on the alphabet Σ by the m-multiplying map δ is the set $\delta(T) = \delta_0(T) \cup \delta_1(T) \cup \cdots \cup \delta_{m-1}(T)$.

The following is a generalization of [8, Proposition 4.3].

Theorem 3.1 *For any tree set T of characteristic c on the alphabet Σ and any m-multiplying map δ, the image of T by δ is a tree set of characteristic mc.*

Proof. Set $S = \delta(T) = \delta_0(T) \cup \delta_1(T) \cup \cdots \cup \delta_{m-1}(T)$. The set S is clearly biextendable since T is biextendable by definition.

Let us consider a nonempty word $x = \delta_i(y)$, with $0 \leq i \leq m-1$. The graph $E_S(x)$ is isomorphic to the graph $E_T(y)$. Indeed, let j be the end of the path with origin i and input label y in the m-multiplying transducer. For $a_i, b_j \in A$, one has $a_i x b_j \in S$ if and only if $ayb \in T$ where a (resp. b) is the input label of the edge with output label a_i (resp. b_j) ending in i (resp. with origin j). Thus, $E_S(x)$ is a tree for any nonempty word $x \in S$.

Finally, the graph $E_S(\varepsilon)$ is, up to orientation, the union of m graphs, all of them isomorphic to $E_T(\varepsilon)$. Indeed, consider the map π from $S \cap A^2$ onto $\{0, 1, \cdots, m-1\}$ which assigns to $ab \in S \cap A^2$ the state i which is the end of the edge of $\mathcal{A}$ with output label a (and the origin of the edge with output label b). Set $S_i = \pi^{-1}(i)$. We have a partition $S \cap A^2 = S_0 \cup S_1 \cup \cdots \cup S_{m-1}$ such that each graph having S_i as set of edges is isomorphic to $E_T(\varepsilon)$. Since $E_T(\varepsilon)$ is a forest of c trees, the graph $E_S(\varepsilon)$ is a forest of mc trees.

■

Example 3.2 Let $B = \{a, b\}$ and let T be the Fibonacci set (see Example 2.1). Let δ be the doubling map given by the transducer of Figure 2.

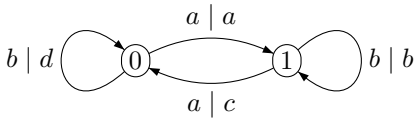

Figure 2: A doubling automaton.

The graph $E_S(\varepsilon)$ is represented in Figure 1.

4. Bifix codes

A prefix code is a set of nonempty words which does not contain any proper prefix of its elements. A suffix code is defined symmetrically. A *bifix code* is

a set which is both a prefix code and a suffix code (see [3] for a more detailed introduction). Let S be a recurrent set. A prefix (resp. bifix) code $X \subset S$ is S-maximal if it is not properly contained in a prefix (resp. bifix) code $Y \subset S$. Since S is recurrent, a finite S-maximal bifix code is also an S-maximal prefix code (see [2, Theorem 4.2.2]). For example, for any $n \geq 1$, the set $X = S \cap A^n$ is an S-maximal bifix code.

Given a set X, we define $\rho(X) = \sum_{x \in X} \rho(x)$. We prove the following result. It accounts for the fact that, in a Sturmian set S, any finite S-maximal suffix code contains exactly one right-special word [2, Proposition 5.1.5].

Proposition 4.1 *Let S be a neutral set and let X be a suffix code. Then $\rho(X) \leq \mathrm{Card}(A) - \chi(S)$ with equality if X is finite and S-maximal.*

Proof. If $\rho(\varepsilon) = 0$, then $\chi(S) = \mathrm{Card}(A)$ and thus the formula holds. Otherwise, ρ' is a left probability distribution (as seen at the end of Section 2), and the formula holds by a well-known property of suffix codes (see [2, Proposition 3.3.4]). ∎

Example 4.2 Let S be the neutral set of characteristic 2 of Example 2.2. The set $X = \{a, ac, b, bc, d\}$ is an S-maximal suffix code (its reversal is the $\tilde{S}$-maximal prefix code $\tilde{X} = \{a, b, ca, cb, d\}$). The values of ρ on X are represented in Figure 3 on the left. One has $\rho(X) = \rho(a) + \rho(bc) = 2$, in agreement with Proposition 4.1.

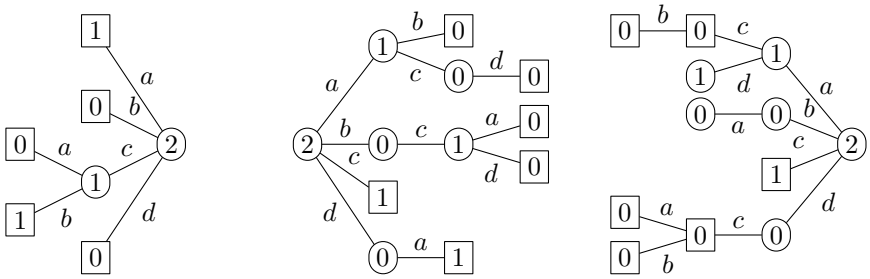

Figure 3: An S-maximal suffix code (left) and an S-maximal bifix code represented as a prefix code (center) and as a suffix code (right).

Let X be a bifix code. Let Q be the set of words without any suffix in X and let P be the set of words without any prefix in X. A *parse* of a word w with respect to a bifix code X is a triple $(q, x, p) \in Q \times X^* \times P$ such that $w = qxp$. We denote by $d_X(w)$ the number of parses of a word w with respect to X. The S-degree of X, denoted by $d_X(S)$ is the maximal number of parses with respect to X of a word of S. For example, the set $X = S \cap A^n$ has S-degree n.

Example 4.3 Let S be the neutral set of characteristic 2 of Example 2.2. The set $X = \{ab, acd, bca, bcd, c, da\}$ is an S-maximal bifix code of S-degree 2 (see Figure 3 on the center and the right).

Let S be a recurrent set and let X be a finite bifix code. By [2, Theorem 4.2.8], X is S-maximal if and only if its S-degree is finite. Moreover, in this case, a word $w \in S$ is such that $d_X(w) < d_X(S)$ if and only if it is an internal factor of a word of X.

The following is [2, Theorem 4.3.7].

Theorem 4.4 *Let S be a recurrent set and let X be a finite S-maximal bifix code of S-degree n. The set of nonempty proper prefixes of X is a disjoint union of $n-1$ S-maximal suffix codes.*

Example 4.5 Let S and X be as in Example 4.3. The set of nonempty proper prefixes of X is the S-maximal suffix code represented on the left of Figure 3.

We now recall the definitions of kernel and derived code. Let S be a recurrent set.

The *kernel* of a finite S-maximal bifix code X is the set of words of X which are an internal factor of X.

Example 4.6 Let S and X be as in Example 4.3. The kernel of X is $K = \{c\}$. Indeed c is the only word that is both an element of X and an internal factor of elements of X (namely, acd, bca and bcd).

Let now X be a finite S-maximal bifix code of S-degree $d \geq 2$. Let P' be the set of proper prefixes of X which are internal factors of X. Then P' is a non empty prefix-closed set and thus it is the set of proper prefixes of a unique prefix code X' called the *derived code* of X. It is actually an S-maximal bifix code of S-degree $d-1$ (see [2], Theorem 4.3.1 and Proposition 4.3.5). Note that the set P of proper prefixes of X is the disjoint union of P' and an S-maximal suffix code Y. Thus, consistently with Theorem 4.4, $P' \setminus \{\varepsilon\}$ is a disjoint union of $d-2$ S-maximal suffix codes and $P \setminus \{\varepsilon\}$ a disjoint union of $d-1$ S-maximal suffix codes.

Example 4.7 Let S and X be as in Example 4.3. The derived code of X is $X' = A$. We have $P = \{\varepsilon, a, b, d, ac, bc, da\}$ and $P' = \{\varepsilon\}$. Thus $P = P' \cup Y$ where S-maximal suffix code represented in Figure 3 on the left.

The following statement is closely related with a similar statement concerning the average length of a bifix code, but which requires an invariant probability distribution (see [2, Corollary 4.3.8]).

Proposition 4.8 *Let S be a recurrent neutral set and let X be a finite S-maximal bifix code of S-degree n. The set P of proper prefixes of X satisfies $\rho_S(P) = n(\mathrm{Card}(A) - \chi(S))$.*

Proof. By Theorem 4.4, we have $P \setminus \{\varepsilon\} = \cup_{i=1}^{n-1} Y_i$, where the Y_i are S-maximal suffix codes. By Proposition 4.1, we have $\rho(Y_i) = \mathrm{Card}(A) - \chi(S)$ and thus $\rho(P) = \rho(\varepsilon) + (n-1)(\mathrm{Card}(A) - \chi(S)) = n(\mathrm{Card}(A) - \chi(S))$. ∎

The following theorem is a generalization of [6, Theorem 3.6] where it is proved for a neutral set of characteristic 1. We consider a recurrent set S, and we implicitly assume that all words of S are on the alphabet A.

Theorem 4.9 *Let S be a neutral recurrent set. For any finite S-maximal bifix code X of S-degree d, one has*

$$\mathrm{Card}(X) = d(\mathrm{Card}(A) - \chi(S)) + \chi(S). \tag{3}$$

Note that we recover, as a particular case of Theorem 4.9 applied to the set X of words of length n in S, the fact that for a set S satisfying the hypotheses of the theorem, the factor complexity is $p_0 = 1$ and $p_n = n(\mathrm{Card}(A) - \chi(S)) + \chi(S)$.

Proof of Theorem 4.9. Since X is a finite S-maximal bifix code, it is an S-maximal prefix code (see Section 4). By a well-known property of trees, this implies that $\mathrm{Card}(X) = 1 + \sum_{p\in P}(r(p)-1)$ where P is the set of proper prefixes of X. Since $\rho(p) = r(p) - 1$ for p non empty and $\rho(\varepsilon) = m(\varepsilon) + r(\varepsilon) - 1$, we have

$$\begin{aligned}\mathrm{Card}(X) &= 1 + \sum_{p\in P}(r(p)-1) = 1 + \sum_{p\in P}\rho(p) - m(\varepsilon)\\ &= \rho(P) + \chi(S) = d(\mathrm{Card}(A) - \chi(S)) + \chi(S)\end{aligned}$$

since $\rho(P) = d(\mathrm{Card}(A) - \chi(S))$ by Proposition 4.8. ∎

Example 4.10 Let S be the neutral set of Example 2.2 and let X be the S-maximal bifix code of Example 4.3. We have $\mathrm{Card}(X) = 2(4-2)+2 = 6$ according to Theorem 4.9.

The following statement is a converse of Theorem 4.9.

Theorem 4.11 *Let S be a uniformly recurrent set containing the alphabet A. If every finite S-maximal bifix code of S-degree d has $d(Card(A)-c)+c$ elements, then S is neutral of characteristic c.*

To prove Theorem 4.11, we use the following result, which can be proved in the same way as Theorem 3.12 in [6], using internal transformations.

Proposition 4.12 *Let S be a uniformly recurrent set containing the alphabet A and let $d_0 \geq 2$. If all finite S-maximal bifix codes of S-degree $d \geq d_0$ have the same cardinality, then any word of length greater than or equal to $d_0 - 1$ is neutral.*

Proof of Theorem 4.11. We first apply the statement to the S-maximal bifix code $X = S \cap A^2$ which has S-degree 2. Since $\mathrm{Card}(X) = 2(\mathrm{Card}(A) - c) + c = 2\,\mathrm{Card}(A) - c$, we conclude that $m_S(\varepsilon) = 1 - c$. On the other hand, applying Proposition 4.12 with $d_0 = 2$, we conclude that every nonempty word is neutral. Thus S is neutral of characteristic c. ∎

We also note that Theorem 4.9 can be formulated in an equivalent way using the notion of derived code of a maximal bifix code.

Theorem 4.13 *Let S be a recurrent neutral set, let X be a finite S-maximal bifix code of S-degree $d \geq 2$ and let X' be the derived code of X. One has*

$$\mathrm{Card}(X) = \mathrm{Card}(X') + \mathrm{Card}(A) - \chi(S). \tag{4}$$

Indeed, since X' has degree $d_S(X) - 1$, by Theorem 4.9, we have

$$\mathrm{Card}(X) - \mathrm{Card}(X') = \mathrm{Card}(A) - \chi(S).$$

Conversely, we may prove Theorem 4.9 by induction on n, assuming Theorem 4.13. Equation (3) holds for $n = 1$ since in this case $X = A$. Next, assume that it holds for $d - 1$. Then, by Equation (4), we have

$$\begin{aligned} \mathrm{Card}(X) &= \mathrm{Card}(X') + \mathrm{Card}(A) - \chi(S) \\ &= (d-1)(\mathrm{Card}(A) - \chi(S)) + \chi(S) + \mathrm{Card}(A) - \chi(S) \\ &= d(\mathrm{Card}(A) - \chi(S)) + \chi(S). \end{aligned}$$

Example 4.14 Let S be the neutral set of Example 2.2 and let X be the S-maximal bifix code of Example 4.3. We have $X' = A$ and accordingly $\mathrm{Card}(X) = \mathrm{Card}(A) + \mathrm{Card}(A) - 2 = 6$.

5. Return words

Let S be a factorial set of words. For a set $X \subset S$ of nonempty words, a *complete return word* to X is a word of S which has a proper prefix in X, a proper suffix in X and no internal factor in X. We denote by $\mathcal{CR}_S(X)$ the set of complete return words to X. The set $\mathcal{CR}_S(X)$ is a bifix code. If S is uniformly recurrent, $\mathcal{CR}_S(X)$ is finite for any finite set X. For $x \in S$, we denote by $\mathcal{CR}_S(x)$ instead of $\mathcal{CR}_S(\{x\})$.

Example 5.1 Let $n \geq 1$ and let $X = S \cap A^n$. Then $\mathcal{CR}_S(X) = S \cap A^{n+1}$.

Theorem 5.2 *Let S be a neutral set. For any finite nonempty bifix code $X \subset S$ with empty kernel, we have*

$$\mathrm{Card}(\mathcal{CR}_S(X)) \leq \mathrm{Card}(X) + \mathrm{Card}(A) - \chi(S) \tag{5}$$

with equality if S is recurrent.

Proof. Let P be the set of proper prefixes of $\mathcal{CR}_S(X)$. For $q \in P$, we define $\alpha(q) = \mathrm{Card}\{a \in A \mid qa \in P \cup \mathcal{CR}_S(X)\} - 1$. For $P' \subset P$, we set $\alpha(P') = \sum_{p \in P'} \alpha(p)$.

Since $\mathcal{CR}_S(X)$ is a finite prefix code, we have, by a well-known property of trees, $\mathrm{Card}(\mathcal{CR}_S(X)) \leq 1 + \alpha(P)$ with equality if $\mathcal{CR}_S(X)$ is nonempty (that is, if S is recurrent).

Let P' be the set of words in P which are proper prefixes of X and let $Y = P \setminus P'$. Since P' is the set of proper prefixes of X, we have $\alpha(P') = \mathrm{Card}(X) - 1$.

Since $P \cup \mathcal{CR}_S(X) \subset S$, one has $\alpha(q) \leq \rho_S(q)$ for any $q \in P$. Moreover, if S is recurrent, and since X has empty kernel, any word of S with a prefix in X is comparable for the prefix order with a word of $\mathcal{CR}_S(X)$. This implies that for any $q \in Y$ and any $b \in R_S(q)$, one has $qb \in P \cup \mathcal{CR}_S(X)$. Consequently, we have $\alpha(q) = \rho_S(q)$ for any $q \in Y$. Thus we have shown that

$$\mathrm{Card}(\mathcal{CR}_S(X)) \leq 1 + \alpha(P') + \rho(Y) \leq \mathrm{Card}(X) + \rho(Y)$$

with equality if S is recurrent. Let us show that Y is a suffix code which is S-maximal if S is recurrent. This will imply our conclusion by Proposition 4.1. Suppose that $q, uq \in Y$ with u nonempty. Since q is in Y, it has a proper prefix in X. But this implies that uq has an internal factor in X, a contradiction. Thus Y is a suffix code. Assume next that S is recurrent. Consider $w \in S$. Then, for any $x \in X$, there is some $u \in S$ such that $xuw \in S$. Let y be the shortest suffix of xuw which has a proper prefix in X. Then $y \in Y$. This shows that Y is an S-maximal suffix code. ∎

Since a recurrent set S is uniformly recurrent if and only if the set of return words is finite (see, for example, [7, Proposition 4.2]), we have the following consequence of Theorem 5.2.

Corollary 5.3 *A recurrent neutral set is uniformly recurrent.*

Proof. By Theorem 5.2, the set $\mathcal{CR}_S(x)$ is finite for any $x \in X$. ∎

Let S be a factorial set. A *right return word* to x in S is a word w such that xw is a word of S which ends with x and has no internal factor equal to x (thus xw is a complete first return word to x). We denote by $\mathcal{R}_S(x)$ the set of right return words to x in S. Since $\mathcal{CR}_S(x) = x\mathcal{R}_S(x)$, the sets $\mathcal{CR}_S(x)$ and $\mathcal{R}_S(x)$ have the same number of elements. Thus we have the following consequence of Theorem 5.2.

Corollary 5.4 *Let S be a uniformly recurrent neutral set. For any $x \in S$, the set $\mathcal{R}_S(x)$ has* $\mathrm{Card}(A) - \chi(S) + 1$ *elements.*

Example 5.5 Consider again the neutral set S of Example 2.2. We have $\mathcal{R}_S(a) = \{bca, bcda, cad\}$.

The following statement, which holds under fairly general hypotheses, shows an interesting connection between complete return words to a bifix code and the derived code (see Section 4). It explains the similarity between Formulae (4) and (5) (with equality).

Proposition 5.6 *Let S be a recurrent set. Let X be a finite S maximal bifix code, let X' be the derived code of X and let K, K' be the kernels of X and X' respectively. Then*

$$\mathcal{CR}_S(X' \setminus K) = X \setminus K. \tag{6}$$

Proof. Let us first show the inclusion from right to left. Let $x \in X \setminus K$. Then x has a proper prefix in $X' \setminus K$, namely the shortest prefix of x which is not an internal factor of X (see [2, Lemma 4.3.3]). Similarly, x has a proper suffix which is in $X' \setminus K$. Moreover x cannot have an internal factor in $X' \setminus K$. Indeed, by definition of X', the words in $X' \setminus K$ are not internal factors of X. This shows that $x \in \mathcal{CR}_S(X' \setminus K)$.

Conversely, consider $x \in \mathcal{CR}_S(X' \setminus K)$. Let P be the set of proper prefixes of X. Let y (resp. z) be the proper prefix (resp. suffix) of x which is in $X' \setminus K$. Since x' is in X', it is in P. We cannot have $x \in P$ since otherwise z would be in K. Thus x has a prefix yu in X. By the first part of the proof, yu has a suffix in $\mathcal{CR}_S(X' \setminus K$, and thus x has an internal factor in $X' \setminus K$, a contradiction unless $x = yu$. Thus $x \in X$. ∎

If S is assumed to be uniformly recurrent and neutral, Formulæ (4) and (5) (with equality) show that both sides of Equation (6) have the same cardinality. Thus the inclusion implies the equality.

Example 5.7 Let S and X be as in Example 4.3. We have $K = \{c\}$ and

$$X \setminus K = \{ab, acd, bca, bcd, da\} = \mathcal{CR}_S(\{a, b, d\})$$

in agreement with Proposition 5.6.

6. Bifix decoding

In this section we show some closure properties for the families of neutral and tree sets.

Let S be a factorial set and let X be a finite S-maximal bifix code. A *coding morphism* for X is a morphism $f : B^* \to A^*$ which maps bijectively an alphabet B onto X. The set $f^{-1}(S)$ is called a *maximal bifix decoding* of S.

Theorem 6.1 *Any maximal bifix decoding of a recurrent neutral set is a neutral set with the same characteristic.*

Let S be a factorial set. For two sets of words X, Y and a word $w \in S$, we set $L_S^X(w) = \{x \in X \mid xw \in S\}$, $R_S^Y(w) = \{y \in Y \mid wy \in S\}$, $E_S^{X,Y}(w) = \{(x, y) \in X \times Y \mid xwy \in S\}$, and furthermore

$$e_S^{X,Y}(w) = \mathrm{Card}(E_S^{X,Y}(w)),\ \ell_S^X(w) = \mathrm{Card}(L_S^X(w)),\ r_S^Y(w) = \mathrm{Card}(R_S^Y(w)).$$

Finally, for a word w, we define $m_S^{X,Y}(w) = e_S^{X,Y}(w) - \ell_S^X(w) - r_S^Y(w) + 1$. Note that $E_S^{A,A}(w) = E_S(w)$, $m_S^{A,A}(w) = m_S(w)$, and so on.

Proposition 6.2 *Let S be a neutral set, let X be a finite S-maximal suffix code and let Y be a finite S-maximal prefix code. Then $m_S^{X,Y}(w) = m_S(w)$ for every $w \in S$.*

Proof. We use an induction on the sum of the lengths of the words in X and in Y.

If X, Y contain only words of length 1, since X (resp. Y) is an S-maximal suffix (resp. prefix) code, we have $X = Y = A$ and there is nothing to prove.

Assume next that one of them, say Y, contains words of length at least 2. Let p be a nonempty proper prefix of Y of maximal length. Set $Y' = (Y \setminus pA) \cup p$. If $wp \notin S$, then $m^{X,Y}(w) = m^{X,Y'}(w)$ and the conclusion follows by induction hypothesis. Thus we may assume that $wp \in S$. Then

$$m^{X,Y}(w) - m^{X,Y'}(w) = e^{X,A}(wp) - \ell^X(wp) - r^A(wp) + 1 = m^{X,A}(wp).$$

By induction hypothesis, we have $m^{X,Y'}(w) = m(w)$ and $m^{X,A}(wp) = 0$, whence the conclusion. ∎

Proof of Theorem 6.1. Let S be a recurrent neutral set and let $f : B^* \to A^*$ be a coding morphism for a finite S-maximal bifix code X. Set $U = f^{-1}(S)$. Let $v \in U \setminus \{\varepsilon\}$ and let $w = f(v)$. Then $m_U(v) = m_S^{X,X}(w)$. Since S is recurrent, X is an S-maximal suffix code and prefix code. Thus, by Proposition 6.2, $m_U(v) = m_S(w)$, which implies our conclusion. ∎

The following example shows that the maximal decoding of a uniformly recurrent neutral set need not be recurrent.

Example 6.3 Let S be the set of factors of the infinite word $(ab)^\omega$. The set $X = \{ab, ba\}$ is a bifix code of S-degree 2. Let $f : u \mapsto ab$, $v \mapsto ba$. The set $f^{-1}(S)$ is the set of factors of $u^\omega \cup v^\omega$ and it is not recurrent.

An interesting corollary of Theorem 6.1 is the following.

Corollary 6.4 *Any maximal bifix decoding of a recurrent tree set is a tree set with the same characteristic.*

Proof. Let S be a recurrent tree set of characteristic c and let $f : B^* \to A^*$ be a coding morphism for a finite S-maximal bifix code X. By definition S is acyclic. By [4, Theorem 3.11], the set $U = f^{-1}(S)$ is also acyclic. From Proposition 6.2, we have that $m_U(f^{-1}(w)) = m_S^{X,X}(w) = m_S(w)$ for every $w \in S$. Thus $m_U(u) = 0$ for every nonempty word u and $m_U(\varepsilon) = \chi(S)$. By an elementary result of graph theory it follows that $E_U(u)$ is a tree for every nonempty $u \in U$ and $E_U(\varepsilon)$ is a forest of $\chi(S)$ trees. Hence U is a tree set of characteristic $\chi(U) = \chi(S)$. ∎

7. Modular codes

For some special bifix code, we can give a more precise description of the bifix decoding.

Let S be a tree set of characteristic c. Since S is biextendable, any letter $a \in A$ occurs exactly twice as a vertex of $E(\varepsilon)$, one as an element of $L(\varepsilon)$ and one as an element of $R(\varepsilon)$.

Denote by $\mathcal{T}_0, \dots \mathcal{T}_{c-1}$ the c trees such that $E(\varepsilon) = \mathcal{T}_0 \cup \cdots \cup \mathcal{T}_{c-1}$. We define the *modular weight* of a letter a as $\|a\| = j - i \pmod c$, where $\mathcal{T}_i$ is the tree containing a as a left extension and $\mathcal{T}_j$ the tree containing a as a right extension.

Given a word $w = a_0 a_1 \cdots a_m$, we define the *modular weight* of w as $\|w\| = \sum_{k=0}^{m} \|a_k\| \pmod c$.

Note that the modular weight of a word depends on the choice of the order for the trees $\mathcal{T}_i$.

The set of words having modular weight equal to zero has the form $X^* \cap S$ for some special bifix code $X \subset S$ called the *modular code*. The set X is the set of words having modular weight 0 such that all nonempty prefixes (or suffixes) have positive modular weight. It is easy to see that X is actually a S-maximal bifix code.

Another way to define the modular code is by using the *modular graph*. This graph is defined as the directed graph $\mathcal{G}$ with vertices $0, 1, \dots, c-1$ and edges all triples (i, a, j) for $0 \leq i, j \leq c-1$ and $a \in A$ such that $(1 \otimes b, a \otimes 1) \in \mathcal{T}_i$ and $(1 \otimes a, c \otimes 1) \in \mathcal{T}_j$ for some $b, c \in A$. Observe that for every letter $a \in A$ there is exactly one edge labeled a because a appears exactly one as a left (resp. right) vertex in $E(\varepsilon)$.

Note that, when S is a tree set of characteristic c obtained by a multiplying map using a transducer $\mathcal{A}$, the modular graph of S is the output automaton of $\mathcal{A}$.

Example 7.1 Let S be the tree set of characteristic 2 of Example 3.2. The modular graph of S is represented in Figure 4. It is the output automaton of the 2-multiplying transducer of Figure 2.

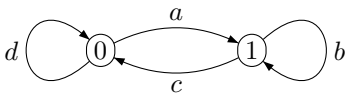

Figure 4: The modular graph.

The following is a generalization of [8, Proposition 4.5].

Proposition 7.2 *Let S be a tree set of characteristic c and let $\mathcal{G}$ be its modular graph. Let $S_{i,j}$ be the set of words in S which are the label of a path from i to j in the graph $\mathcal{G}$.*

(1) The family $(S_{i,j} \setminus \{\varepsilon\})_{0 \leq i,j \leq c-1}$ is a partition of $S \setminus \{\varepsilon\}$.
(2) For $u \in S_{i,j} \setminus \{\varepsilon\}$ and $v \in S_{k,\ell} \setminus \{\varepsilon\}$, if $uv \in S$, then $j = k$.
(3) $\|w\| = 0$ if and only if $w \in S_{k,k}$ for some $0 \leq k \leq c-1$.

Proof. We first note that for $a, b \in A$ such that $ab \in S$, there is a path in $\mathcal{G}$ labeled ab. Since $(a, b) \in E(\varepsilon)$, there is a k such that $(1 \otimes a, b \otimes 1) \in \mathcal{T}_k$. Then

we have $a \in S_{i,k}$ and $b \in S_{k,j}$ for some $0 \leq i, j \leq c-1\}$. This shows that ab is the label of a path from i to j in $\mathcal{G}$.

Let us prove by induction on the length of a nonempty word $w \in S$ that there exists a unique pair i, j such that $w \in S_{i,j}$. The property is true for a letter, by definition of the extension graph $E(\varepsilon)$ and for words of length 2 by the above argument. Let next $w = ax$ be in S with $a \in A$ and x nonempty. By induction hypothesis, there is a unique pair (k, j) such that $x \in S_{k,j}$. Let b be the first letter of x. Then the edge of $\mathcal{G}$ with label b starts in k. Since ab is the label of a path, we have $a \in S_{i,k}$ for some i and thus $ax \in S_{i,j}$. The other assertions follow easily. ∎

Note that point (3) of Proposition 7.2 says that the modular code does not depend on the choice of the order of the states in the modular graph (or of the trees $\mathcal{T}_i$ in $E(\varepsilon)$).

The following theorem improves Corollary 6.4 in the case of a bifix decoding by the modular code. The same result is proved in [8, Theorem 4.6] for uniformly recurrent tree sets of characteristic 1 (recall that by Corollary 5.3 a recurrent tree set is uniformly recurrent).

Theorem 7.3 *The decoding of a recurrent tree set S of characteristic c by the modular code is a union of c recurrent tree sets of characteristic* 1*. More precisely, if f is the coding morphism for the modular code, then $f^{-1}(S_{0,0})$, $f^{-1}(S_{1,1})$, …, $f^{-1}(S_{c-1,c-1})$ are recurrent tree sets of characteristic* 1*.*

Proof. Let us define $T_k = f^{-1}(S_{k,k})$ for every $0 \leq k \leq c-1$. Fixed a k, we show that T_k is a recurrent tree set of characteristic 1.

First, it is easy to verify that T_k is biextendable.

Next, since S is recurrent, for every $u, v \in S_{k,k} \subset S$ there exists a $w \in S$ such that $uwv \in S$. From point (2) of Proposition 7.2 follows that $w \in S_{k,k}$. Thus T_k is recurrent.

Let now X be the modular code and set $X_k = X \cap S_{k,k}$. In order to prove that T_k is a tree set it is enough to show that $E_{S_{k,k}}(w) = E_S^{X_k,X_k}(w)$ is a tree for any $w \in S_{k,k}$. Note first that $E_{S_{k,k}}(w) = E_S^{X,X}(w)$ for any $w \in S_{k,k} \setminus \{\varepsilon\}$. Indeed, for $w \in S_{k,k}$ and $x, y \in X$ such that $xwy \in S$, one has $x, y \in X_k$ and thus $xwy \in S_{k,k}$.

According to Proposition [4, Proposition 3.9], the graph $E_S^{X,X}(w)$ is a tree for any word $w \in S \setminus \{\varepsilon\}$, whence the result.

Next, let us show that the graph $E_S^{X_k,X_k}(\varepsilon)$ is also a tree. First, since a tree set is acyclic, the graph $E_S^{X,X}(\varepsilon)$ is acyclic by Proposition [4, Proposition 3.7] and so is its subgraph $E_S^{X_k,X_k}$.

Let us prove that for every $x, y \in S_{k,k}$ there is a path in $E_S^{X_k,X_k}(\varepsilon)$ from x to y.

If $x, y \in A$, then there is a path from x to y in $E(\varepsilon)$ and thus a there is a path from x to y in $E_S^{X_k,X_k}(\varepsilon)$ obtained by replacing an edge $(a, b) \in A \times A$ of

the path by an edge (z,t) in $X_S^{X_k,X_k} \times X_S^{X_k,X_k}$ such that z ends with a and t begins with b.

Otherwise, assume for example that $y = au$ with u nonempty. Set $Y = \{v \in S \mid av \in X_k\}$. Since Y is an $a^{-1}S$-maximal prefix code, by [4, Proposition 3.9], the graph $E_S^{X_k,Y}(a)$ is a tree. Since $u \in Y$, there is a path in $E_S^{X_k,Y}(a)$ from x to u. This implies that there is a path from x to y in $E_S^{X_k,X_k}(\varepsilon)$. Thus $E_S^{X_k,X_k}(\varepsilon)$ is connected. ∎

8. Interval exchanges

In this section we define interval exchange sets and we show that they are tree sets.

Let $I =]\ell, r[$ be a nonempty open interval of the real line and A a finite ordered alphabet. For two intervals Δ, Γ, we write $\Delta < \Gamma$ if $x < y$ for any $x \in \Delta$ and $y \in \Gamma$. A partition $(I_a)_{a\in A}$ of I (minus $\mathrm{Card}(A) - 1$ points) in open intervals is *ordered* if $a < b$ implies $I_a < I_b$.

We consider now two total orders $<_1$ and $<_2$ on A and two partitions $(I_a)_{a\in A}$ and $(J_a)_{a\in A}$ of I in open intervals ordered respectively by $<_1$ and $<_2$ and such that for every a, I_a and J_a have the same length λ_a. Let $\gamma_a = \sum_{b<_1 a} \lambda_b$ and $\delta_a = \sum_{b<_2 a} \lambda_a$.

An *interval exchange transformation* (with flips) relative to $(I_a)_{a\in A}$ and $(J_a)_{a\in A}$ is a map $T : I \to I$ such that for every $a \in A$, its restriction to I_a is either a translation or a symmetry from I_a to J_a (see for example [5] and [16] for interval exchanges with flips).

Observe that γ_a is the left boundary of I_a and that δ_a is the left boundary of J_a. If $\mathrm{Card}(A) = s$, we say that T is an s-interval exchange transformation.

Example 8.1 Let $A = \{a, b, c\}$. Consider the rotation of angle α with α irrational as a 3-transformation relative to the partition $(I_a)_{a\in A}$ of the interval $]0, 1[$, where $I_a =]0, 1-2\alpha[$, $I_b =]1-2\alpha, 1-\alpha[$ and $I_c =]1-\alpha, 1[$, while $J_c =]0, \alpha[$, $J_a =]\alpha, 1-\alpha[$ and $J_b =]1-\alpha, 1[$ (see Figure 5). Then, for each letter a, the restriction to I_a is a translation to J_a. Note that one has $a <_1 b <_1 c$ and $c <_2 a <_2 b$.

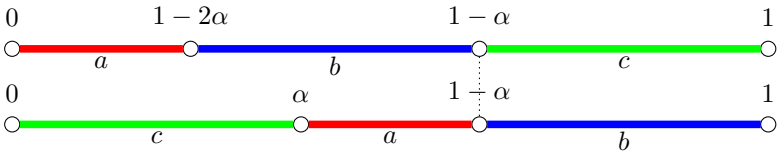

Figure 5: A 3-interval exchange transformation.

For a word $w = b_0 b_1 \cdots b_m$ let I_w be the set

$$I_w = I_{b_0} \cap T^{-1}(I_{b_1}) \cap \cdots \cap T^{-m}(I_{b_m}).$$

Set $J_w = T^{|w|}(I_w)$. We set by convention $I_\varepsilon = J_\varepsilon =]\ell, r[$. Note that each I_w is an open interval and so is each J_w (see [5]).

Let T be an interval exchange transformation on $I =]\ell, r[$. For a given $z \in I$, the *natural coding* of T relative to z is the infinite word $\Sigma_T(z) = a_0 a_1 \cdots$ on the alphabet A defined by $a_n = a$ if $T^n(z) \in I_a$. We denote by $\mathcal{L}(T)$ the set of factors of the natural codings of T. We also say that $\mathcal{L}(T)$ is the *natural coding* of T or the *interval exchange set* arising from T. Note that, for every $w \in \mathcal{L}(T)$, the interval I_w is the set of points z such that $\Sigma_T(z)$ starts with w, while the interval J_w is the set of points z such that $\Sigma_T\left(T^{-|w|}(z)\right)$ starts with w. Moreover, it is easy to prove that a word u is in $\mathcal{L}(T)$ if and only if $I_u \neq \emptyset$ (and thus if and only if $J_u \neq \emptyset$).

Example 8.2 Let T be the interval exchange transformation of Example 8.1. The first element of $\mathcal{L}(T)$ are represented in Figure 6 (right-special words are colored).

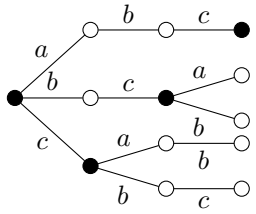

Figure 6: The words of length ≤ 3 of $\mathcal{L}(T)$.

A *connection* of an interval exchange transformation T is a triple (x, y, n) where x is a singularity of T^{-1}, y is a singularity of T, $n \geq 0$ and $T^n(x) = y$. We also say that (x, y, n) is a connection of length n ending in y. When $n = 0$, we say that $x = y$ is a connection.

Interval exchange transformations without connections, also called *regular* interval exchange transformations, are well studied (see, for example, [15] and [5]).

Example 8.3 Let T be the transformation of Example 8.1. The point γ_c is a connection of length 0. This connection is represented with a dotted line in Figure 5.

Let T be an interval exchange transformation with exactly c connections all of length 0. Denote by $\gamma_{k_0} = \ell$ and $\gamma_{k_1}, \ldots, \gamma_{k_c}$ the c connections of T. For every $0 \leq i < c$ the interval $]\gamma_{k_i}, \gamma_{k_{i+1}}[$ is called a *component* of I.

Example 8.4 Consider again the transformation T of Example 8.1. The two components of $]0, 1[$ are the two intervals $]0, 1 - \alpha[$ and $]1 - \alpha, 1[$.

In the next statement we generalize a result of [4] and show that the natural coding of an interval exchange is acyclic. This result generalizes the corresponding result for regular interval exchange (see [4]).

Theorem 8.5 *Let T be an interval exchange transformation with exactly c connections, all of length 0. Then $\mathcal{L}(T)$ is a tree set of characteristic $c+1$.*

In order to prove Theorem 8.5 we need some preliminary result.

Lemma 8.6 *Let T be an interval exchange transformation. For every nonempty word w and letter $a \in A$, one has*

(i) $a \in L(w) \iff I_w \cap J_a \neq \emptyset$,
(ii) $a \in R(w) \iff I_a \cap J_w \neq \emptyset$.

Proof. A letter a is in the set $L(w)$ if and only if $aw \in \mathcal{L}(T)$. As we have seen before, this is equivalent to $J_{aw} \neq \emptyset$. One has $J_{aw} = T(I_{aw}) = T(I_a) \cap I_w = J_a \cap I_w$, whence point (i). Point (ii) is proved symmetrically. ∎

We say that a path in a graph is *reduced* if it does not use twice consecutively the same edge.

Lemma 8.7 *Let T be an interval exchange transformation over I without connection of length ≥ 1. Let $w \in \mathcal{L}(T)$ and $a, b \in L(w)$ (resp. $a, b \in R(w)$). Then $1 \otimes a, 1 \otimes b$ (resp. $a \otimes 1, b \otimes 1$) are in the same connected component of $E(w)$ if and only if J_a, J_b (resp. I_a, I_b) are in the same component of I.*

Proof. Let $a \in L(w)$. Since the set $\mathcal{L}(T)$ is biextendable, there exists a letter c such that $(1 \otimes a, c \otimes 1) \in E(w)$. Using the same reasoning as that in Lemma 8.6, one has $J_a \cap I_{wc} \neq \emptyset$. Since $I_{wc} \subset I_w$, one has in particular $J_a \cap I_w \neq \emptyset$. This proves that J_a and I_w belong to the same component of I for every $a \in L(w)$.

Conversely, suppose that $a, b \in L(w)$ are such that J_a and J_b belong to the same component of I. We may assume that $a <_2 b$. Then, there is a reduced path $(1 \otimes a_1, b_1 \otimes 1, \ldots, b_{n-1} \otimes 1, 1 \otimes a_n)$ in $E(w)$ (see Figure 7) with $a = a_1$, $b = a_n$, $a_1 <_2 \cdots <_2 a_n$ and $wb_1 <_1 \cdots <_1 wb_{n_1}$. Indeed, by hypothesis, we have no connection of length ≥ 1. Thus, for every $1 \leq i < n$, one has $J_{a_i} \cap I_{wb_i} \neq \emptyset$ and $J_{a_{i+1}} \cap I_{wb_i} \neq \emptyset$. Therefore, a and b are in the same connected component of $E(w)$.

The symmetrical statement is proved similarly. ∎

We can now prove the main result of this section.

Proof of Theorem 8.5. Let us first prove that for any $w \in \mathcal{L}(T)$, the graph $E(w)$ is acyclic. Assume that $(1 \otimes a_1, b_1 \otimes 1, \ldots, 1 \otimes a_n, b_n \otimes 1)$ is a reduced path in $E(w)$ with $a_1, \ldots, a_n \in L(w)$ and $b_1, \ldots, b_n \in R(w)$. Suppose that $n \geq 2$ and that $a_1 <_2 a_2$. Then one has $a_1 <_2 \cdots <_2 a_n$ and $wb_1 <_1 \cdots <_1 wb_n$ (see Figure 7). Thus one cannot have an edge (a_1, b_n) in the graph $E(w)$.

Let us now prove that the extension graph of the empty word is a union of $c+1$ trees. Let $a, b \in A$. If J_a and J_b are in the same component of I, then $1 \otimes a, 1 \otimes b$ are in the same connected component of $E(\varepsilon)$ by Lemma 8.7. Thus $E(\varepsilon)$ is a union of $c+1$ trees.

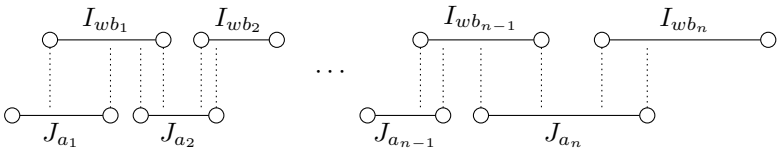

Figure 7: A path from a_1 to a_n in $E(w)$.

Finally, if $w \in \mathcal{L}(T)$ is a nonempty word and $a, b \in L(w)$, then J_a and J_b are in the same component of I, by Lemma 8.6, and thus a and b are in the same connected component of $E(w)$ by Lemma 8.7. Thus $E(w)$ is a tree. ∎

Example 8.8 Let T be the interval exchange transformation of Example 8.1. $\mathcal{L}(T)$ is a tree set of characteristic 2. In Figure 8 are represented the extension graphs of the empty word (left) and of the letters a (center) and b (right).

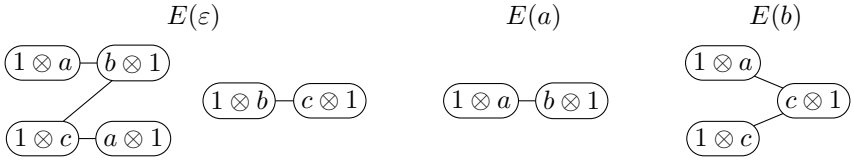

Figure 8: Some extension graphs.

9. Linear involutions

In this section, we define linear involutions, which are a generalization of interval exchange transformations (see also [9]).

We consider two copies $I \times \{0\}$ and $I \times \{1\}$ of an open interval I of the real line and set $\hat{I} = I \times \{0, 1\}$. We call the sets $I \times \{0\}$ and $I \times \{1\}$ the two *components* of $\hat{I}$. We consider each component as an open interval.

A *generalized permutation* on A of type (ℓ, m), with $\ell + m = k$, is a bijection $\pi : \{1, 2, \ldots, k\} \to A$. We represent it by a two line array

$$\pi = \begin{pmatrix} \pi(1)\ \pi(2)\ \ldots \pi(\ell) \\ \pi(\ell+1)\ \ldots \pi(\ell+m) \end{pmatrix}$$

A *length data* associated with (ℓ, m, π) is a nonnegative vector $\lambda \in \mathcal{R}_+^A = \mathcal{R}_+^k$ such that

$$\lambda_{\pi(1)} + \ldots + \lambda_{\pi(\ell)} = \lambda_{\pi(\ell+1)} + \ldots + \lambda_{\pi(2k)} \text{ and } \lambda_a = \lambda_{a^{-1}} \text{ for all } a \in A.$$

We consider a partition of $I \times \{0\}$ (minus $\ell - 1$ points) in ℓ open intervals $I_{\pi(1)}, \ldots, I_{\pi(\ell)}$ of lengths $\lambda_{\pi(1)}, \ldots, \lambda_{\pi(\ell)}$ and a partition of $I \times \{1\}$ (minus $m-1$ points) in m open intervals $I_{\pi(\ell+1)}, \ldots, I_{\pi(\ell+m)}$ of lengths $\lambda_{\pi(\ell+1)}, \ldots, \lambda_{\pi(\ell+m)}$. Let Σ be the set of $k-2$ *division points* separating the intervals I_a for $a \in A$.

The *linear involution* on I relative to these data is the map $T = \sigma_2 \circ \sigma_1$ defined on the set $\hat{I} \setminus \Sigma$, formed of $\hat{I}$ minus the $k-2$ division points, and which is the composition of two involutions defined as follows.

(i) The first involution σ_1 is defined on $\hat{I} \setminus \Sigma$. It is such that for each $a \in A \cup A^{-1}$, its restriction to I_a is either a translation or a symmetry from I_a onto $I_{a^{-1}}$.

(ii) The second involution exchanges the two components of $\hat{I}$. It is defined for $(x, \delta) \in \hat{I}$ by $\sigma_2(x, \delta) = (x, 1 - \delta)$. The image of z by σ_2 is called the *mirror image* of z.

We also say that T is a linear involution on I and relative to the alphabet A or that it is a k-linear involution to express the fact that the alphabet A has k elements.

Example 9.1 Let $A = \{a, b, c, d, a^{-1}, b^{-1}, c^{-1}, d^{-1}\}$ and

$$\pi = \begin{pmatrix} a & b & a^{-1} & c \\ c^{-1} & d^{-1} & b^{-1} & d \end{pmatrix}$$

Let T be the 8-linear involution corresponding to the length data represented in Figure 10 (we represent $I \times \{0\}$ above $I \times \{1\}$) with the assumption that the restriction of σ_1 to I_a and I_d is a symmetry while its restriction to I_b, I_c is a translation. We indicate on the figure the effect of the transformation T on a

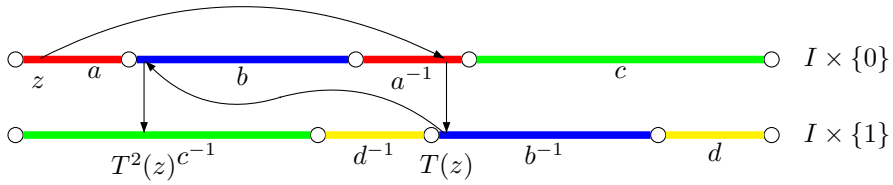

Figure 9: A linear involution.

point z located in the left part of the interval I_a. The point $\sigma_1(z)$ is located in the right part of $I_{a^{-1}}$ and the point $T(z) = \sigma_2\sigma_1(z)$ is just below on the left of $I_{b^{-1}}$. Next, the point $\sigma_1 T(z)$ is located on the left part of I_b and the point $T^2(z)$ just below.

The notion of linear involution is an extension of the notion of interval exchange transformation in the following sense. Assume that $\ell = k$ and that $A = \{\pi(1), \ldots, \pi(k)\}$.

Then, the restriction of T to $I \times \{0\}$ is an interval exchange (and so is its restriction to $I \times \{1\}$ which is the inverse of the first one). Thus, in this case, T is a pair of mutually inverse interval exchange transformations.

A linear involution T is a bijection from $\hat{I} \setminus \Sigma$ onto $\hat{I} \setminus \sigma_2(\Sigma)$. Since σ_1, σ_2 are involutions and $T = \sigma_2 \circ \sigma_1$, the inverse of T is $T^{-1} = \sigma_1 \circ \sigma_2$.

The set Σ of division points is also the set of singular points of T and their mirror images are the singular points of T^{-1}. Note that these singular points z may be 'false' singularities, in the sense that T can have a continuous extension to an open neighborhood of z.

As for interval exchanges, we define a *connection* of a linear involution T as a triple (x, y, n) such that x is a singularity of T^{-1}, y is a singularity of T, $n \geq 0$ and $T^n(x) = y$.

Example 9.2 Let us consider the linear involution T which is the same as in Example 9.1, but such that the restriction of σ_1 to I_c is a symmetry. We assume that $I =]0,1[$, that $\lambda_a = \lambda_d$. Let $x = (1-\lambda_d, 0)$ and $y = (\lambda_a, 0)$.

Then x is a singularity of T^{-1} ($\sigma_2(x)$ is the left endpoint of I_d), y is a singularity of T (it is the right endpoint of I_a) and $T(x) = y$. Thus $(x, 1, y)$ is a connection (see Figure **??**).

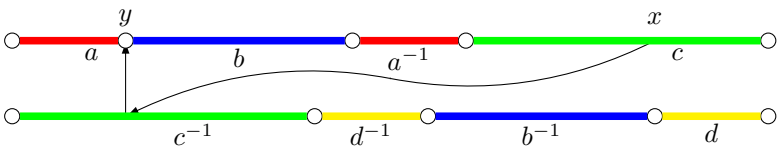

Figure 10: A linear involution.

Example 9.3 Let T be the linear involution on $I =]0,1[$ represented in Figure 11. We assume that the restriction of σ_1 to I_a is a translation whereas the restriction to I_b and I_c is a symmetry. We choose $(3-\sqrt{5})/2$ for the length of the interval I_c (or I_b). With this choice, T has no connection.

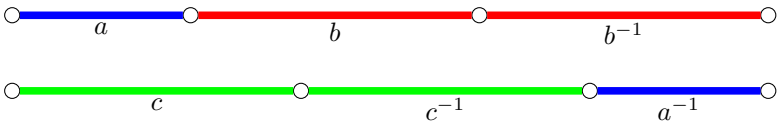

Figure 11: A linear involution on $A = \{a, b, c\}$.

Let T be a linear involution without connection. Let

$$O = \bigcup_{n \geq 0} T^{-n}(\Sigma) \quad \text{and} \quad \hat{O} = O \cup \sigma_2(O) \tag{7}$$

be respectively the negative orbit of the singular points and its closure under mirror image. Then T is a bijection from $\hat{I} \setminus \hat{O}$ onto itself. Indeed, assume that $T(z) \in \hat{O}$. If $T(z) \in O$ then $z \in O$. Next if $T(z) \in \sigma_2(O)$, then $T(z) \in \sigma_2(T^{-n}(\Sigma)) = T^n(\sigma_2(\Sigma))$ for some $n \geq 0$. We cannot have $n = 0$ since $\sigma_2(\Sigma)$ is not in the image of T. Thus $z \in T^{n-1}(\sigma_2(\Sigma)) = \sigma_2(T^{-n+1}(\Sigma)) \subset \sigma_2(O)$. Therefore in both cases $z \in \hat{O}$. The converse implication is proved in the same way.

Let T be a linear involution on I, let $\hat{I} = I \times \{0,1\}$ and let $\hat{O}$ be the set defined by Equation (7).

Given $z \in \hat{I} \setminus \hat{O}$, the *infinite natural coding* of T relative to z is the infinite word $\Sigma_T(z) = a_0 a_1 \ldots$ on the alphabet A defined by

$$a_n = a \quad \text{if} \quad T^n(z) \in I_a.$$

We first observe that the infinite word $\Sigma_T(z)$ is reduced, that is, there is no factor of the form aa^{-1} or $a^{-1}a$ for $a \in A$. Indeed, assume that $a_n = a$ and $a_{n+1} = a^{-1}$ with $a \in A$. Set $x = T^n(z)$ and $y = T(x) = T^{n+1}(z)$. Then $x \in I_a$ and $y \in I_{a^{-1}}$. But $y = \sigma_2(u)$ with $u = \sigma_1(x)$. Since $x \in I_a$, we have $u \in I_{a^{-1}}$. This implies that $y = \sigma_2(u)$ and u belong to the same component of $\hat{I}$, a contradiction.

We denote by $L(T)$ the set of factors of the infinite natural codings of T. We say that $L(T)$ is the *natural coding* of T.

Example 9.4 Let T be the linear involution of Example 9.3. The words of length at most 3 of $S = \mathcal{L}(T)$ are represented in Figure 12.

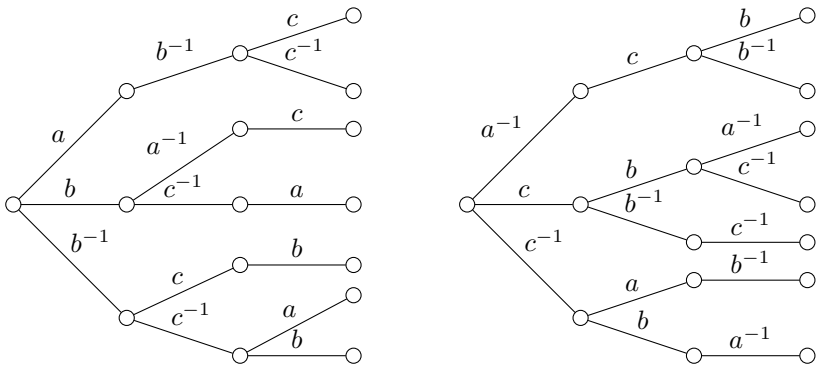

Figure 12: The words of length at most 3 of S.

The following theorem is proved in a similar way as Theorem 8.5.

Theorem 9.5 *The natural coding of a linear involution without connection is a tree set set of characteristic* 2.

In order to prove Theorem 9.5 we need some preliminary results. The following one is proved in the same way as Lemma 8.6 (see also Figure 13).

Lemma 9.6 *Let T be a linear involution. For every nonempty word w and letter $a \in A$, one has*

(i) $a \in L(w) \Leftrightarrow \sigma_2(I_{a^{-1}}) \cap I_w \neq \emptyset$,
(ii) $a \in R(w) \Leftrightarrow \sigma_2(I_a) \cap I_{w^{-1}} \neq \emptyset$.

Proof. By [9, Lemma 5.2], we have $a \in L(w)$ if and only if $I_{aw} \neq \emptyset$ which is also equivalent to $T(I_{aw}) \neq \emptyset$. As for interval exchanges, one has $T(I_{aw}) = T(I_a) \cap I_w$. Since $T = \sigma_2 \circ \sigma_1$ and since $\sigma_1(I_a) = I_{a^{-1}}$, $a \in L(w)$ if and only if $\sigma_2(I_{a^{-1}}) \cap I_w \neq \emptyset$. Next, since $\mathcal{L}(T)$ is closed under taking inverses (see [9, Proposition 5.3]), $aw \in S$ if and only if $w^{-1}a^{-1} \in S$. Thus $a \in R(w)$ if and only if $a^{-1} \in L(w^{-1})$, whence the second equivalence. ∎

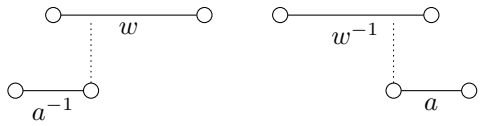

Figure 13: An illustration of $a \in L(w)$ and $a \in R(w)$.

For two subsets I, J of the real line, we write $I < J$ if $x < y$ for any $x \in I$ and $y \in J$.

Given a linear involution T on I, we introduce two orders on $\mathcal{L}(T)$ as follows. For any $u, v \in \mathcal{L}(T)$, one has

(i) $u <_R v$ if and only if $I_u < I_v$,
(ii) $u <_L v$ if and only if $I_{u^{-1}} < I_{v^{-1}}$.

The following lemma is proved in the same way as Lemma 8.7.

Lemma 9.7 *Let T be a linear involutions on I without connection. Let $w \in \mathcal{L}(T)$ and $a, a' \in L(w)$ (resp. $b, b' \in R(w)$). Then $1 \otimes a, 1 \otimes a'$ (resp. $b \otimes 1, b' \otimes 1$) are in the same connected component of $E(w)$ if and only if $I_{a^{-1}}, I_{a'^{-1}}$ (resp. $I_b, I_{b'}$) are in the same component of I.*

Proof. If $(1 \otimes a, b \otimes 1) \in E(w)$, then $\sigma_2(I_{a^{-1}}) \cap I_{wb} \neq \emptyset$. Thus $I_{a^{-1}}$ and I_{wb} belong to distinct components of $\hat{I}$. Consequently, if $a, a' \in L(w)$ (resp. $R(w)$) belong to the same connected component of $E(w)$, then $I_{a^{-1}}, I_{a'^{-1}}$ (resp. $I_{wa}, I_{wa'}$) belong to the same component of $\hat{I}$.

Conversely, let $a, a' \in L(w)$ be such that a, a' belong to the same component of $\hat{I}$. We may assume that $a <_L a'$. There is a reduced path (i.e., it does not use twice consecutively the same edge) in $E(w)$ from a to a' which is the sequence $a_1, b_1, \ldots, b_{n-1}, a_n$ with $a_1 = a$ and $a_n = a'$ with $a_1 <_L a_2 <_L \cdots <_L a_n$, $wb_1 <_R wb_2 <_R \cdots <_R wb_{n-1}$ and $\sigma_2(I_{a_i^{-1}}) \cap I_{wb_i} \neq \emptyset$, $\sigma_2(I_{a_{i+1}^{-1}}) \cap I_{wb_i} \neq \emptyset$ for $1 \leq i \leq n-1$ (see Figure 14 for an illustration).

Note that the hypothesis that T is without connection is needed since otherwise the right boundary of $\sigma_2(I_{a_i^{-1}})$ could be the left boundary of I_{wb_i}.

The assertion concerning $b, b' \in R(w)$ is a consequence of the first one since $b, b' \in R(w)$ if and only if $b^{-1}, b'^{-1} \in L(w^{-1})$ (see [9, Proposition 5.3]). ∎

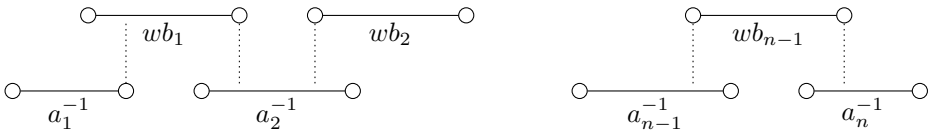

Figure 14: A path from a_1 to a_n in $E(w)$.

We can now prove the main result of this section.

Proof.[of Theorem 9.5] Let T be a linear involution on I without connection and let $S = \mathcal{L}(T)$. Let us first prove that for any $w \in \mathcal{L}(T)$, the graph $E(w)$ is acyclic. Assume that $(1 \otimes a_1, b_1 \otimes 1, \ldots, 1 \otimes a_n, b_n \otimes 1)$ is a path in $E(w)$

with $a_1, \dots, a_n \in L(w)$ and $b_1, \dots, b_n \in R(w)$. We may assume that the path is reduced, that $n \geq 2$ and also that $a_1 <_L a_2$. It follows that $a_1 <_L \dots <_L a_n$ and $wb_1 <_R \dots <_R wb_n$ (see Figure 14). Thus it is not possible to have an edge (a_1, b_n), which shows that $E(w)$ is acyclic.

Let $a, a' \in A$. If $I_{a^{-1}}$ and $I_{a'^{-1}}$ are in the same component of $\hat{I}$, then $1 \otimes a, 1 \otimes a'$ are in the same connected component of $E(\varepsilon)$. Thus $E(\varepsilon)$ is a union of two trees with $2\,\mathrm{Card}(A)$ vertices.

If $w \in S$ is nonempty and $1 \otimes a, 1 \otimes a' \in L(w)$, then $I_{a^{-1}}$ and $I_{a'^{-1}}$ are in the same component of $\hat{I}$ (by Lemma 9.6), and thus $1 \otimes a, 1 \otimes a'$ are in the same connected component of $E(w)$. Thus $E(w)$ is a tree. ∎

Example 9.8 Let T be the linear involution of Example 9.4. $\mathcal{L}(T)$ is a tree set of characteristic 2 over the alphabet $\{a, b, c, a^{-1}, b^{-1}, c^{-1}\}$. In Figure 15 are represented the extension graphs of the empty word (left) and of letters a (center) and c^{-1} (right) (where we note $\bar{a}$ instead of a^{-1}).

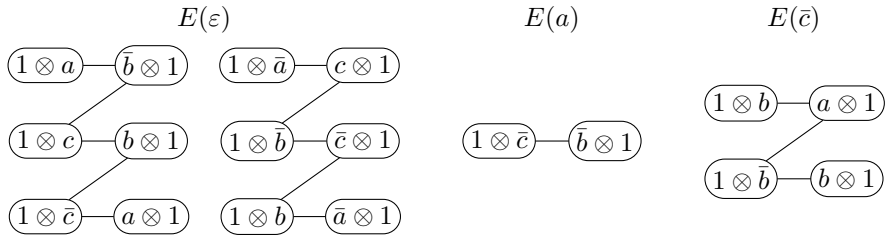

Figure 15: Some extension graphs.

References

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

[2] Jean Berstel, Clelia De Felice, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Bifix codes and Sturmian words. *J. Algebra*, 369:146–202, 2012.

[3] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2009.

[4] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Acyclic, connected and tree sets. *Monatsh. Math.*, 176(4):521–550, 2015.

[5] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Bifix codes and interval exchanges. *J. Pure Appl. Algebra*, 219(7):2781–2798, 2015.

[6] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. The finite index basis property. *J. Pure Appl. Algebra*, 219:2521–2537, 2015.

[7] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Maximal bifix decoding. *Discrete Math.*, 338:725–742, 2015.

[8] Valérie Berthé, Clelia De Felice, Vincent Delecroix, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Specular sets. In Florin Manea and Dirk Nowotka, editor, *Combinatorics on Words - 10th International Conference, WORDS 2015, Kiel, Germany, September 14-17, 2015, Proceedings.*, volume 9304 if *Lecture Notes in Computer Science*, pages 210–222. Springer, 2015.

[9] Valérie Berthé, Vincent Delecroix, Francesco Dolce, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Return words of linear involutions and fundamental groups. *Ergodic Th. Dyn. Syst.*, 2015. To appear, (http://arxiv.org/abs/1405.3529).

[10] Valérie Berthé and Michel Rigo. *Combinatorics, automata and number theory*, volume 135 of *Encyclopedia Math. Appl.* Cambridge Univ. Press, Cambridge, 2010.

[11] A. Blondin Massé, S. Brlek, S. Labbé, and L. Vuillon. Palindromic complexity of codings of rotations. *Theoret. Comput. Sci.*, 412(46):6455–6463, 2011.

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

[13] Francesco Dolce and Dominique Perrin. Enumeration formulæ in neutral sets. In Igor Potapov, editor, *Developments in Language Theory - 19th International Conference, DLT 2015, Liverpool, UK, July 27-30, 2015, Proceedings.*, volume 9168 of *Lecture Notes in Computer Science*, pages 215–227. Springer, 2015.

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

[15] Michael Keane. Interval exchange transformations. *Math. Z.*, 141:25–31, 1975.

[16] Arnaldo Nogueira, Benito Pires, and Serge Troubetzkoy. Orbit structure of interval exchange transformations with flip. *Nonlinearity*, 26(2):525–537, 2013.

[17] Laurent Vuillon. On the number of return words in infinite words constructed by interval exchange transformations. *Pure Math. Appl. (PU.M.A.)*, 18(3-4):345–355, 2007.