Order conditions for languages

Sébastien Ferenczi, Pascal Hubert, Luca Zamboni

Order conditions for languages

Sébastien Ferenczi[1], Pascal Hubert[2], and Luca Q. Zamboni[3]

[1] Aix Marseille Université, CNRS, Centrale Marseille, Institut de Mathématiques de Marseille, I2M - UMR 7373, 13453 Marseille, (France) ssferenczi@gmail.com
[2] Aix Marseille Université, CNRS, Centrale Marseille, Institut de Mathématiques de Marseille, I2M - UMR 7373, 13453 Marseille, (France) hubert.pascal@gmail.com
[3] Institut Camille Jordan, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, F69622 Villeurbanne Cedex (France) zamboni@math.univ-lyon1.fr

Abstract. We define a condition on the resolution of bispecials in a language. A language satisfies this order condition if and only if it is the natural coding of a generalized interval exchange transformation, while the order condition plus some additional ones characterize the codings of various more classical interval exchange transformations. Also, a finite word clusters for the Burrows-Wheeler transform if and only if the language generated by its powers satisfies an order condition.

1 Languages satisfying an order condition

1.1 Usual definitions

Let $\mathcal{A}$ be a finite set called the *alphabet*, its elements being *letters*. A *word* w of *length* $n = |w|$ is $a_1a_2\cdots a_n$, with $a_i \in \mathcal{A}$. The *concatenation* of two words w and w' is denoted by ww'.

By a language L over $\mathcal{A}$ we mean a *factorial extendable language*: a collection of sets $(L_n)_{n\geq 0}$ where the only element of L_0 is the *empty word*, and where each L_n for $n \geq 1$ consists of words of length n, such that for each $v \in L_n$ there exist $a, b \in \mathcal{A}$ with $av, vb \in L_{n+1}$, and each $v \in L_{n+1}$ can be written in the form $v = au = u'b$ with $a, b \in \mathcal{A}$ and $u, u' \in L_n$.

The *complexity function* $p : \mathbb{N} \to \mathbb{N}$ is defined by $p(n) = \#L_n$.

A word $v = v_1...v_r$ is a *factor* of a word $w = w_1...w_s$ or an infinite sequence $w = w_1w_2...$ if for some $i \geq 1$ $v_1 = w_i$, $...v_r = w_{i+r-1}$.

The *reverse* of the word $v = v_1...v_r$ is the word $v_r...v_1$.

A word w is *primitive* if it is not equal to v^n for any word v and integer $n > 1$.

Let W be a family of words or (one- or two-sided) infinite sequences. Whenever the set L of all the factors of the words or sequences in W is a language (namely, is factorial and extendable), we say that L is *the language generated by* W and denote it by $L(W)$.

A language L is *recurrent* if for each $v \in L$ there exists a nonempty w such that vw is in L and ends with v.

A language L is *uniformly recurrent* if for each $v \in L$ there exists n such that v is a factor of each word $w \in L_n$.

A language L is *aperiodic* if for each nonempty word w in L, there exists n such that w^n is not in L.

For a word w in L, we call *arrival set of* w, denoted by $A(w)$, the set of all letters x such that xw is in L, and call *departure set of* w, denoted by $D(w)$, the set of all letters x such that wx is in L.

A word w in L is called *right special*, resp. *left special* if $\#D(w) > 1$, resp. $\#A(w) > 1$. If $w \in L$ is both right special and left special, then w is called *bispecial*. If $\#L_1 > 1$, the empty word ε is bispecial, with $A(\varepsilon) = D(\varepsilon) = L_1$.

A bispecial word w in L is a *weak bispecial* if $\#\{awb \in L, a \in A(w), b \in D(w)\} < \#A(w) + \#D(w) - 1$.

A bispecial word w in L is a *strong bispecial* if $\#\{awb \in L, a \in A(w), b \in D(w)\} > \#A(w) + \#D(w) - 1$.

To *resolve* a bispecial word w is to find all words in L of the form awb for letters a and b.

The *symbolic dynamical system* associated to a language L is the two-sided shift S acting on the subset X_L of $\mathcal{A}^{\mathbb{Z}}$ consisting of all bi-infinite sequences x such that $x_r \cdots x_{r+s-1} \in L_s$ for each r and s, defined by $(Sx)_n = x_{n+1}$ for all $n \in \mathbb{Z}$.

Thus in the present paper we use two-sided sequences $x \in X_L$, but also their (infinite) *suffixes* $(x_n, n \geq k)$ and (infinite) *prefixes* $(x_n, n \leq k)$.

1.2 Order conditions: first properties

The order condition can be traced to [15], but its first explicit mention appears in [11] in the particular case of one order and its reverse.

Definition 1. A language L on an alphabet $\mathcal{A}$ satisfies a *local order condition* if, for each bispecial word w, there exist two total orders on $\mathcal{A}$, denoted by $<_{A,w}$ and $<_{D,w}$, such that whenever awc and bwd are in L with letters $a \neq b$ and $c \neq d$, then $a <_{A,w} b$ if and only if $c <_{D,w} d$.

A language L on an alphabet $\mathcal{A}$ satisfies an *order condition* if it satisfies a local order condition where the orders $<_{A,w}$ and $<_{D,w}$ are the same for all bispecial words.

The first notion in the following definition seems to be new, and its links with the order conditions will be studied below.

Definition 2. In a language L, a *locally strong bispecial word* is a bispecial word w such that there exist nonempty subsets $A' \subset A(w)$, $D' \subset D(w)$ such that $\#\{awb \in L, a \in A', b \in D'\} > \#A' + \#D' - 1$.

If a language L on an alphabet $\mathcal{A}$ satisfies a local order condition, a bispecial word w has a *connection* if there are letters $a <_{A,w} a'$, consecutive in the order $<_{A,w}$, letters $b <_{D,w} b'$, consecutive in the order $<_{D,w}$, such that awb and $a'wb'$ are in L, and neither awb' nor $a'wb$ is in L.

We recall a well-known result which can be deduced from [9] or [10].

Lemma 1. *A language L which has no strong bispecial word has a finite number of weak bispecial words, and the left (resp. right) special words are the prefixes (resp. suffixes) af a finite number of infinite suffixes (resp. prefixes) of sequences of X_L.*

Proof
See the proof of Lemmas 2 and 5 in [12].

We turn now to combinatorial consequences of order conditions.

Lemma 2. *A language L which satisfies a local order condition contains no locally strong bispecial word, and thus no strong bispecial word.*

Proof
See the proof of Lemma 1 in [12].

For languages where each word has at most two right (resp. left) extensions, the absence of strong bispecial words, the absence of locally strong bispecial words, and a local order condition are all equivalent. In the general case, it is easy to find bispecials which are locally strong but not strong (suppose for example that the possible xwy are awa, awb, bwa, bwb, cwc), and a local order condition is stricter than the absence of locally strong bispecials.

Example 1. Suppose L is a language whose words of length 2 are ac, ad, ba, bc, cb, cc, da. Then the empty word is not a locally strong bispecial, yet L does not satisfy any local order condition. We can then choose L_3 to be made with acc, ada, bac, bad, bcb, bcc, cba, cbc, ccb, dac, where each word has at most two left (resp. right) extensions, and continue by resolving the bispecials so that they are all neutral. We get a language without locally strong bispecials but not satisfying any local order condition.

A related notion is dendricity ([4] under the name of *tree sets*). This can be interpreted as the following.

Definition 3. A language is *dendric* if it has neither locally strong bispecial words nor weak bispecial words.

Thus, by Lemmas 1 and 2, a language satisfying a local order condition is *ultimately dendric*. But, because of Example 1 and the possibility of weak bispecials, there is no inclusion relation between dendric languages and languages satifying an order condition, local or not.

Lemma 3. *A language satisfying a local order condition has complexity* $p(n) = kn + l$ *for all* n *large enough and with* $0 \leq k \leq \#\mathcal{A} - 1$. *Moreover,* $k = \#\mathcal{A} - 1$ *if and only if* L *has no connection, and in that case* $l = 1$.

Proof
See the proof of Lemma 3 and Corollary 4 in [12].

1.3 Recurrence

We look at consequences of an order condition, or weaker properties, on the trajectories in X_L.

Definition 4. Let x be a bi-infinite sequence in $\mathcal{A}^{\mathbb{Z}}$, or a suffix of such a sequence. x is *right recurrent* if any factor of x is a factor of each suffix of x.

Let x be a bi-infinite sequence in $\mathcal{A}^{\mathbb{Z}}$, or a prefix of such a squence. x is *left recurrent* if any factor of x is a factor of each prefix of x.

A bi-infinite sequence x in $\mathcal{A}^{\mathbb{Z}}$ is *recurrent* if it is both left and right recurrent.

Lemma 4. *Suppose* L *has no strong bispecial word.*

If a suffix of a sequence in X_L *is right recurrent, it generates a uniformly recurrent language.*

If a prefix of a sequence in X_L *is left recurrent, it generates a uniformly recurrent language.*

If a sequence in X_L *is left or right recurrent, it is recurrent and generates a uniformly recurrent language.*

Proof
See the proof of Lemma 9 in [12].

Proposition 1. *Suppose* L *has no strong bispecial word. Then there are at most a finite number of* orbits $\{S^n x, n \in \mathbb{Z}\}$, $x \in X_L$, *with* x *not recurrent.*

Proof
See the proof of Proposition 17 in [12].

Proposition 2. *Let* L *be a recurrent language satisfying a local order condition. Then* L *is a finite union of uniformly recurrent languages.*

Proof
See the proof of Proposition 12 in [12].

Lemma 5. *Given a language* L, *we denote by* L' *the sublanguage of* L *generated by all recurrent sequences in* X_L. *Then* L' *is a recurrent language over an alphabet* $\mathcal{A}' \subset \mathcal{A}$. *Moreover, if* L *satisfies an order condition, so does* L'.

Proof
See the proof of Lemma 18 in [12].

2 Interval exchange transformations

2.1 Definitions

Generalized interval exchange transformations are defined in [1] and [26] and do generalize the well-known classical, or standard, interval exchange transformations of [29] [22].

All intervals are open on the right, closed on the left.

Definition 5. A *generalized interval exchange transformation* is a map T defined on $[0,1)$ partitioned by intervals I_e, $e \in \mathcal{A}$, continuous and (strictly) increasing on each I_e, and such that the TI_e, $e \in \mathcal{A}$, are intervals partitioning $[0,1)$.

The I_e, indexed in $\mathcal{A}$, are called the *defining intervals* of T.

If the restriction of T to each I_e is an affine map, T is an *affine* interval exchange transformation.

If the restriction of T to each I_e is an affine map of slope 1, T is a *standard* interval exchange transformation.

The endpoints of the I_e, resp. TI_e, excluding 0 and 1, will be denoted by γ_i, resp. β_j, for i, resp. j, taking $\#\mathcal{A} - 1$ values.

Definition 6. T is *minimal* if every orbit is dense in $[0,1)$.

T satisfies the *i.d.o.c. condition* if there is at least one point γ_i (of Definition 5), and there is no i, j, $k \geq 0$, such that $T^k\beta_i = \gamma_j$.

A *wandering interval* is an interval J for which $T^n J$ is disjoint from J for all $n > 0$.

Note that our i.d.o.c. condition is a modified version of the one introduced by Keane [22]; our condition depends on the defining intervals, and is not intrinsic to T as the original Keane's condition.

Definition 7. For a generalized interval exchange transformation T, its *natural coding* is the language $L(T)$ generated by all the *trajectories*, namely the sequences $(x_n, n \in \mathbb{Z}) \in \mathcal{A}^{\mathbb{Z}}$ where $x_n = e$ if $T^n x$ falls into I_e, $e \in \mathcal{A}$.

Thus we can look at the symbolic system associated to $L(T)$. Note that the set $X_{L(T)}$ is the closure in $\mathcal{A}^{\mathbb{Z}}$ of the set of trajectories, for the product topology defined by the discrete topology on $\mathcal{A}$.

Example 2. A *Sturmian language* [27] is the natural coding of the standard interval exchange transformation T sending $[0, 1-\alpha)$ to $[\alpha, 1)$ and $[1-\alpha, 1)$ to $[0, \alpha)$ for α irrational; T is conjugate to a rotation of angle α on the 1-torus.

We shall also consider slightly more general codings, by merging into intervals $\tilde{I}_e$ some adjacent intervals I_e whose images by T are also adjacent. This is equivalent to taking the natural coding of another interval exchange transformation $\tilde{T}$, but when T is affine, if we define $\tilde{T}$ by the intervals $\tilde{I}_e$ it will not necessarily be affine by our definition, as the slope is not constant on its defining intervals, see Example 4 below. Thus we define

Definition 8. A language L is a *grouped coding* of an affine interval exchange transformation T if there exist intervals $\tilde{I}_e$, $e \in \tilde{\mathcal{A}}$ such that

- each $\tilde{I}_e$ is an interval, and a disjoint union of defining intervals of T,
- T is a continuous monotone map on each $\tilde{I}_e$,
- L is the coding of T by the $\tilde{I}_e$, that is the language generated by the trajectories $(x_n, n \in \mathbb{Z}) \in \tilde{\mathcal{A}}^{\mathbb{Z}}$ where $x_n = e$ if $T^n x$ falls into $\tilde{I}_e$, $e \in \tilde{\mathcal{A}}$.

2.2 Interval exchanges satisfy order conditions

Definition 9. A generalized interval exchange transformation defines two orders on $\mathcal{A}$:

- $e <_D f$ whenever the interval I_e is strictly to the left of the interval I_f,
- $e <_A f$ whenever the interval TI_e is strictly to the left of the interval TI_f.

These orders correspond to the two permutations used by Kerckhoff [23] to define standard interval exchange transformations: the unit interval is partitioned into semi-open intervals which are numbered from 1 to k, ordered according to a permutation π_0 and then rearranged according to another permutation π_1; in more classical definitions, there is only one permutation π, which corresponds to π_1 while $\pi_0 = Id$; note that in some papers the orderings are by π_0^{-1} and π_1^{-1}.

Proposition 3. *Let T be a generalized interval exchange transformation. Then its natural coding $L(T)$ satisfies an order condition.*

Proof
See the proof of Proposition 8 in [12].

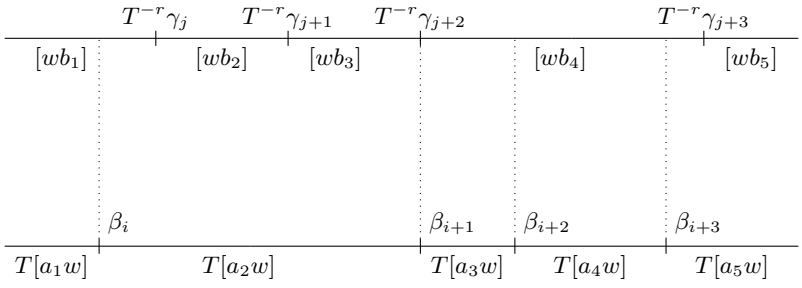

Fig. 1. A bispecial interval

Remark 1. The notion of interval exchange has been extended to *interval exchanges with flips* [28] in the standard case, allowing the slope of T on some defining intervals to be -1. It can be further extended to generalized interval exchanges with flips, allowing T to be decreasing on some defining intervals. The natural codings of such transformations satisfy a *flipped* order condition, which is a local order condition where the order $<_{D,w}$ is always the same order $<_D$, while $<_{A,w}$ is allowed to be either an order $<_A$ or its reverse, according to the number of letters in w corresponding to flipped intervals. Indeed, all the results in Section 2.3 hold for (generalized) interval exchanges with flips, mutatis mutandis.

Remark 2. The language of an *interval translation mapping* [5] does not necessarily satisfy a local order condition: it is possible that TI_a intersects TI_b for $a \neq b$, and, if $TI_a \cap TI_b$ intersects both I_c and I_d, $c \neq d$, then the empty word is a locally strong bispecial. For this family, not much is known; it is an open question, asked by Boshernitzan, whether all interval translation mappings have linear complexity. It is generalized to *piecewise isometries* [18] for which even less is known.

2.3 The converse

Theorem 1. *A language L on at least two letters is the language of a standard interval exchange transformation satisfying the i.d.o.c. condition if and only if it satisfies an order condition, is aperiodic and uniformly recurrent, and has no connection.*

Proof
See the proof of Theorem 15 in [12].

Theorem 2. *A language L is the language of a generalized, or equivalently of a standard, minimal interval exchange transformation if and only if it satisfies an order condition, is aperiodic and uniformly recurrent.*

Proof
See the proof of Theorem 14 in [12].

Theorem 1 is proved in [16], and Theorem 2 uses the same method. Some results similar both to Theorem 2 and Theorem 3 below are proved in [21], using a description of the evolution of *Rauzy graphs* which is somewhat cumbersome to state, but where an order condition seems to be hidden.

Theorem 3. *For a language L on an alphabet $\mathcal{A}$, the following are equivalent:*

- *(i) L satisfies an order condition and is recurrent;*
- *(ii) L is the language of a standard interval exchange transformation;*
- *(iii) L is the language of a generalized interval exchange transformation without wandering intervals.*

Proof
See the proof of Theorem 13 in [12].

We want now to get rid of extra conditions besides the order condition; as the following chain of counter-examples shows, this obliges us to weaken the classical notion of standard interval exchange, thus the successive generalizations are indeed relevant.

Example 3 (Fake Sturmian). Let L be generated by the bi-infinite sequence $\dots 111222 \dots$. Note that it is of complexity $n+1$ but not uniformly recurrent, and in the founding paper [27] it is not included in Sturmian languges, hence we call it a fake Sturmian language.

It satisfies the order condition with $1 <_D 2$, $1 <_A 2$, but (unsurprisingly as it is not recurrent) is *not the language of a standard interval exchange transformation* as that could only be the identity on two disjoint open intervals I_1 and I_2, and the only possible words are 1^n and 2^m. However, L is *the natural coding of an affine interval exchange transformation:* L_2 is the language of length 2 of any affine 2-interval exchange transformation, with the same orders, such that TI_1 is strictly longer than I_1, and, as L is determined by L_2 because there is no bispecial word except the empty one, L is indeed the natural coding of any of these affine interval exchange transformations.

Example 4 (Skew Sturmian). Let L be the language generated by the bi-infinite sequence $\dots 1112111 \dots$, which is a skew Sturmian language as defined in [27].

It satisfies the order condition with $1 <_D 2$, $2 <_A 1$, but is *not the natural coding of any affine interval exchange transformation* T: indeed, the sequence $\dots 1111 \dots$ in X_L would define a fixed point x for T, in the interior of I_1, and, if $0 < y < x$ is the right endpoint of TI_2, T would have to send $[0, x)$ to $[y, x)$ and $[x, 1-y)$ to $[x, 1)$, thus having a slope < 1 on a part of I_1 and a slope > 1 on another part. However, if $\tilde{L}$ is the language generated by the bi-infinite sequence $\dots 3332111 \dots$, as in Example 3 $\tilde{L}$ is the natural coding of any affine interval exchange transformation T sending $I_1 = [0, x)$ to $[y, x)$, $I_3 = [x, 1-y)$ to $[x, 1)$, $I_2 = [1-y, 1)$ to $[0, y)$, with $0 < y < x < 1-y$. If we now code T by the intervals $\tilde{I}_1 = I_1 \cup I_3$ and $\tilde{I}_2 = I_2$, we see that L is *a grouped coding of an affine interval exchange transformation* as in Definition 8.

Example 5 (Episkew). Let L' be the Sturmian language which is the natural coding of the unflipped standard interval exchange transformation T' sending $I_1 = [0, 1-\alpha)$ to $[\alpha, 1)$ and $I_2 = [1-\alpha, 1)$ to $[0, \alpha)$ for an irrational $\alpha < 1/2$. Let $y_n = i$ whenever $T^n\alpha$ is in I_i, $n \geq 0$, and $y'_n = i$ whenever $T^n(1-2\alpha)$ is in I_i, $n \leq 0$; when $\alpha = \frac{3-\sqrt{5}}{2}$, y is the so-called Fibonacci sequence on 1 and 2, and y' is y written backwards. Let L be the language generated by the infinite sequence $\dots y'_{-2}y'_{-1}y'_0 3 y_0 y_1 y_2 \dots$ Extending to languages the definition in [3], we can call it an episkew language. It satisfies the order condition with $1 <_D 3 <_D 2$, $2 <_A 3 <_A 1$ (note that no other order is possible, because of the way the empty bispecial is resolved).

L is *the natural coding of a generalized interval exchange transformation*, by Theorem 4 below, but it is *not the natural or grouped coding of any affine interval exchange transformation:* this will be a straightforward consequence of either one of two independent results we show below, Theorems 6 and 7.

And finally

Theorem 4. *A language L is a natural coding of a generalized interval exchange transformation if and only if L satisfies an order condition.*

Proof
See the proof of Theorem 19 in [12].

2.4 Examples and questions

We do not have a complete characterization of the codings of affine interval exchanges. The best we can do is

Theorem 5. *If L is a natural coding of an affine interval exchange transformation for which the absolute value of the slope is $\exp\theta_e$ on the defining interval I_e, then L satisfies an order condition and for each non recurrent sequence z in X_L, $\sum_{n\geq 0}\exp\left(\sum_{j=0}^{n}\theta_{z_j}\right)<+\infty$, and $\sum_{n>0}\exp\left(-\sum_{j=-n}^{-1}\theta_{z_j}\right)<+\infty$.*

If L satisfies an order condition and there exist real numbers $\theta_e, e\in\mathcal{A}$, such that for each non recurrent sequence z in L, $\sum_{n\geq 0}\exp\left(\sum_{j=0}^{n}\theta_{z_j}\right)<+\infty$, and $\sum_{n>0}\exp\left(-\sum_{j=-n}^{-1}\theta_{z_j}\right)<+\infty$, then L is a group coding of an affine interval exchange transformation.

Proof
See the proof of Theorem 20 in [12].

The generalizations of standard interval exchanges have seen a recent surge in activity (see [25] [26] [19] and others) primarily centered on the conjugacy problem between these different classes of maps; in this context, standard and generalized interval exchange transformations are the extreme cases while affine interval exchange transformations constitute a fundamental middle step. The following questions and conjectures can be considered as related to this problem.

Conjecture 1. The conditions in Theorem 5 are necessary and sufficient for L to be a natural coding of an affine interval exchange transformation.

Question 1. Does there exist an aperiodic language which is a grouped coding of an affine interval exchange transformation, but not a natural coding of any affine interval exchange transformation?

Conjecture 1 and Question 1 suggest what we dare not call a conjecture.

Question 2. Is it true that L is a group coding of an affine interval exchange transformation if and only if L satisfies an order condition and there exist real numbers $\theta_e, e \in \mathcal{A}$, such that the two following conditions hold?

- For each non recurrent sequence z in L which is not ultimately periodic to the left,
  $\sum_{n\geq 0} \exp\left(\sum_{j=0}^{n} \theta_{z_j}\right) < +\infty$.
- For each non recurrent sequence z in L which is not ultimately periodic to the right,
  $\sum_{n>0} \exp\left(-\sum_{j=-n}^{-1} \theta_{z_j}\right) < +\infty$.

There are many examples of codings of affine interval exchange transformations which are not natural codings of standard ones; they can be built by using the methods of [8] [6] [25] and others. But codings of generalized interval exchange transformations which are not codings of affine ones seem to be completely new, and we know two combinatorial ways of building them, expressed in the two following theorems.

Theorem 6. *Let L be non recurrent, and a natural coding of a generalized interval exchange transformation T. Suppose the language L' of Lemma 5 is aperiodic, uniformly recurrent, and its arrival and departure orders are conjugate by a circular permutation. Then T cannot be* of class P*, class P [20] meaning that, except on a countable set of points, its derivative DT exists and $DT = h$ where h is a function with bounded variation, and $|h|$ is bounded from below by a strictly positive number.*

Proof
See the proof of Theorem 23 in [12].

Theorem 7. *Let L' be a natural coding of a non purely periodic standard interval exchange transformation. Let $w_n = aw'_n b$, $a \in \mathcal{A}$, $b \in \mathcal{A}$, be an infinite sequence of bispecial words in L' . Let u be the left-sided infinite sequence ending with w_n for all n, and v the right-sided infinite sequence beginning with w_n for all n. Let ω be a symbol which is not a letter of L', and L be the language generated by the union of all words in L' and the bi-infinite word $u\omega v$.*

Then L is a natural coding of a generalized interval exchange transformation, but not a grouped coding of any affine interval exchange transformation.

Proof
See the proof of Theorem 24 in [12].

3 Order conditions and the Burrows-Wheeler transform

Let $\mathcal{A} = \{a_1 < a_2 < \cdots < a_r\}$ be an ordered alphabet. For a permutation π on $\mathcal{A}$, we define the order $<_\pi$ by $x <_\pi y$ if $\pi^{-1}x < \pi^{-1}y$.

Definition 10. The *(cyclic) conjugates* of $w = w_1 \cdots w_n$ are the words $w_i \cdots w_n w_1 \cdots w_{i-1}$, $1 \leq i \leq n$. If w is primitive, w has precisely n cyclic conjugates. Let $w_{i,1} \cdots w_{i,n}$ denote the i-th conjugate of w where the n conjugates of w are ordered by ascending lexicographical order. Then the *Burrows-Wheeler transform* [7] of w, denoted by $B(w)$, is the word $w_{1,n}w_{2,n} \cdots w_{n,n}$. It depends on the given order $<$ on $\mathcal{A}$.
We say w is *clustering for the order* $<$ *and the permutation* π [16] if $B(w) = (\pi a_1)^{n_{\pi a_1}} \cdots (\pi a_r)^{n_{\pi a_r}}$, where π is a permutation on $\mathcal{A}$ and n_a is the number of occurrences of a in w (we allow some of the n_a to be 0, thus, given the order and w, there may be several possible π). We say w is *perfectly* clustering if it is clustering for the *symmetric* permutation $\pi a_i = a_{r+1-i}$, $1 \leq i \leq r$ ([30] though it is not named).

Non-primitive words. As remarked in [16], the Burrows-Wheeler transform can be extended to a non-primitive word $w_1 \cdots w_n$, by ordering its n (non necessarily different) cyclic conjugates by non-strictly increasing lexicographical order and taking the word made by their last letters. Then $B(v^m)$ is deduced from $B(v)$ by replacing each of its letters x_i by x_i^m, and v^m is clustering for π if and only if v is clustering for π.

Theorem 8. *For a given order $<$ on the alphabet, a primitive word w is clustering for the order $<$ and the permutation π if and only if every bispecial word v in the language L_w generated by w^n, $n \in \mathbb{N}$, satisfies the order condition where the order $<_D$ is the order $<$ and the order $<_A$ is $<_\pi$.*
All bispecial words in L_w are factors of length at most $|w| - 2$ of ww.

Proof
We begin by the last assertion. Suppose v is a bispecial of L_w. Then v must occur at two different positions in some word w^k. If $|w| = n$ and $|v| \geq |w| - 1$, this implies in particular $w_i ... w_n w_1 ... w_{i-2} = w_j ... w_n w_1 ... w_{j-2}$ for $1 < j - i < n$, and we notice that each w_i is in at least one member of the equality, thus we get that w is a power of a word whose length is the GCD of n and $j - i$, which contradicts the primitivity. Thus the length of v is at most $|w| - 2$, and v occurs in ww.

We prove now that our order condition is equivalent to the following *modified order condition*: whenever $z = z_1 ... z_n$ and $z' = z'_1 ... z'_n$ are two different cyclic conjugates of w, $z < z'$ (lexicographically) if and only if $z_k <_\pi z'_k$ for the largest $k \leq n$ such that $z_k \neq z'_k$. Indeed, by definition $z < z'$ if and only if $z_j < z'_j$ for the smallest $j \geq 1$ such that $z_j \neq z'_j$. If w satifies the order condition, we apply it to the bispecial word $z_{k+1} ... z_n z_1 ... z_{j-1}$, with k and j as defined, and get the modified order condition. Let v be a bispecial word in L_w; by the first paragraph of this proof it can be written as $z_1 ... z_{k-1}$ for some $1 \leq k \leq n$, with the convention that $k = 1$ whenever v is empty, and at least two different cyclic conjugates z of w, and its possible extensions are the corresponding $z_n z_1 ... z_k$, thus, if the modified order condition is satisfied, v does satisfy the requirement of the order condition.

The modified order condition implies clustering, as then if two cyclic conjugates of w satisfy $z < z'$, their last letters z_n and z'_n satisfy either $z_n = z'_n$ or $z_n <_\pi z'_n$. Suppose $w = w_1 \cdots w_n$ is clustering for π. Suppose two cyclic conjugates of w are such that $z_k \neq z'_k$, $z_j = z'_j$ for $k+1 \leq j \leq n$. Then $z < z'$ is (by definition of the lexicographical order) equivalent to $z_{k+1}...z_n z_1..z_k < z'_{k+1}...z'_n z'_1..z'_k$, and, as these two words have different last letters, because of the clustering this is equivalent to $z_k <_\pi z'_k$, thus the modified order condition is satisfied. □

Theorem 8 remains valid if $w = v^m$ is non-primitive (it can be slightly improved as there are less bispecial words to be considered, it is enough to look at factors of vv of length at most $|v| - 2$). An immediate consequence is the following, which seems to be new.

Proposition 4. *If w clusters for the order $<$ and the permutation π, its reverse clusters for the π-order, and the permutation π^{-1}.*

Proof
This follows in a straightforward way from Theorem 8. □

The order condition can be applied to get the clustering properties of classical families of words. Thus it can be used to reprove the result of [24]: a Sturmian language contains infinitely many clustering words.

Theorem 9. *The natural coding of a standard k-interval exchange in the* hyperelliptic class *(this consists in all the symmetric ones, i.e. those for which the order $<_A$ is the reverse of $<_D$, and all those which can be obtained from the symmetric ones by an induction process, see [14]) satisfying the i.d.o.c. condition, or of any standard* 3- *or* 4*-interval exchange satisfying the i.d.o.c. condition, contains infinitely many clustering words.*

Proof
By Theorem 8 and Proposition 3 (or by Theorem 4 of [16]), if L is the language of a standard interval exchange satisfying the i.d.o.c. condition, w clusters if ww is in L. The fact that L contains infinitely many squares is proved in [14] for the hyperelliptic class, [13] for the other cases mentioned above. □

Note that Theorem 9 has not yet been generalized to wider classes of interval exchanges. Also, in the forthcoming [17], we shall prove that an Arnoux-Rauzy language [2] contains finitely many clustering words, while an episturmian language [3] may contain finitely or infinitely many clustering words, with a full characterization of each case.

References

1. P. ARNOUX: Un invariant pour les échanges d'intervalles et les flots sur les surfaces, (in French) *Thèse de 3e cycle*:, Reims, 1981.
2. P. ARNOUX, G. RAUZY: Représentation géométrique de suites de complexité $2n+1$, (in French) *Bull. Soc. Math. France* 119 (1991), 199–215.

3. J. BERSTEL: Sturmian and episturmian words (a survey of some recent results), in "Algebraic informatics", *Lecture Notes in Comput. Sci.* 4728, p. 23–47, Springer, Berlin, 2007.
4. V. BERTHÉ, C. DE FELICE, F. DOLCE, J. LEROY, D. PERRIN, C. REUTENAUER, G. RINDONE: Acyclic, connected and tree sets, *Monatsh. Math.* 176 (2015), p. 521–550.
5. M. BOSHERNITZAN, I. KORNFELD: Interval translation mappings, *Ergodic Theory Dynam. Systems* 15 (1995), p. 821–832.
6. X. BRESSAUD, P. HUBERT, A. MAASS: Persistence of wandering intervals in self-similar affine interval exchange transformations. *Ergod. Theory Dyn. Syst.* 30 (2010), p. 665–686.
7. M. BURROWS, D.J. WHEELER: A block-sorting lossless data compression algorithm, *Technical Report 124* (1994), Digital Equipment Corporation.
8. R. CAMELIER, C. GUTTIEREZ: Affine interval exchange transformations with wandering intervals, *Ergod. Theory Dyn. Syst.* 17 (1997), p. 1315–1338.
9. J. CASSAIGNE: Complexité et facteurs spéciaux, (in French) Journées Montoises (Mons, 1994), *Bull. Belg. Math. Soc. Simon Stevin* 4 (1997), p. 67–88.
10. J. CASSAIGNE, F. NICOLAS: Factor complexity, *Combinatorics, automata and number theory*, p. 163–247, Encyclopedia Math. Appl., 135, Cambridge Univ. Press, Cambridge, 2010.
11. A. DE LUCA, M. EDSON, L.Q. ZAMBONI: Extremal values of semi-regular continuants and codings of interval exchange transformations, *Mathematika* 69 (2023) p. 432–457.
12. S. FERENCZI, P. HUBERT, L.Q. ZAMBONI: Languages of general interval exchange transformations, arXiv: 2212.01024
13. S. FERENCZI: A generalization of the self-dual induction to every interval exchange transformation, *Ann. Inst. Fourier (Grenoble)* 64 (2014), p. 1947–2002.
14. S. FERENCZI, L.Q. ZAMBONI: Structure of K-interval exchange transformations: induction, trajectories, and distance theorems, *J. Anal. Math.* 112 (2010), p. 289–328.
15. S. FERENCZI, L.Q. ZAMBONI: Languages of k-interval exchange transformations, *Bull. Lond. Math. Soc.* 40 (2008), p. 705–714.
16. S. FERENCZI, L.Q. ZAMBONI: Clustering words and interval exchanges, *J. Integer Seq.* 16 (2013), Article 13.2.1, 9 pp.
17. S. FERENCZI, L.Q. ZAMBONI: Clustering of Arnoux-Rauzy words, *in preparation.*
18. D. GABORIAU, G. LEVITT, F. PAULIN: Pseudogroups of isometries of R and Rips' theorem on free actions on R-trees, *Israël J. Math.* 87 (1994), p. 403–428.
19. S. GHAZOUANI, C. ULCIGRAI: A priori bounds for GIETS, affine shadows and rigidity of foliations in genus two, arXiv:2106.03529.
20. M.-R. HERMAN: Sur la conjugaison différentiable des difféomorphsimes du cercle à des rotations, (French), *Inst. Hautes Études Sci. Publ. Math.* 49 (1979), p. 5–233.
21. A.Ya. KANEL-BELOV, A.L. CHERNYAT'EV: Describing the set of words generated by interval exchange transformations, *Comm. Algebra 38* (2010), p. 2588–2605.
22. M.S. KEANE: Interval exchange transformations, *Math. Zeitsch.* 141 (1975), p. 25–31.
23. S. KERCKHOFF: Simplicial systems for interval exchange maps and measured foliations, *Ergod. Theory Dyn. Syst.* 5 (1985), p. 257–271.
24. S. MANTACI, A. RESTIVO, M. SCIORTINO: Burrows-Wheeler transform and Sturmian words, *Inform. Process. Lett.* 86 (2003), p.241–246.

25. S. MARMI, P. MOUSSA, J.-C. YOCCOZ: Affine interval exchange maps with a wandering interval. Proc. Lond. Math. Soc. (3) 100 (2010), p. 639–669.
26. S. MARMI, P. MOUSSA, J.-C. YOCCOZ: Linearization of generalized interval exchange maps, *Ann. of Math. (2)* 176 (2012), p. 1583–1646.
27. M. MORSE, G.A. HEDLUND: Symbolic dynamics II. Sturmian trajectories, *Amer. J. Math.* 62 (1940), p. 1–42.
28. A. NOGUEIRA: Nonorientable recurrence of flows and interval exchange transformations. *J. Differential Equations* 70 (1987), p. 153–166.
29. V.I. OSELEDEC: The spectrum of ergodic automorphisms, (in Russian) *Dokl. Akad. Nauk. SSSR* 168 (1966), p. 1009–1011.
30. J. SIMPSON, S. J. PUGLISI: Words with simple Burrows-Wheeler transforms, *The Electronic Journal of Combinatorics* 15 (2008), Research Paper 83, 17 pp.