RELATIONS ON WORDS

MICHEL RIGO

Abstract. In the first part of this survey, we present classical notions arising in combinatorics on words: growth function of a language, complexity function of an infinite word, pattern avoidance, periodicity and uniform recurrence. Our presentation tries to set up a unified framework with respect to a given binary relation.

In the second part, we mainly focus on abelian equivalence, k-abelian equivalence, combinatorial coefficients and associated relations, Parikh matrices and M-equivalence. In particular, some new refinements of abelian equivalence are introduced.

1. Introduction

This paper follows and complements the talk I gave during the conference on *Automatic Sequences, Number Theory, and Aperiodic Order* held in Delft in October 2015. The aim is to survey various concepts arising in combinatorics on words and present them in a unified and general framework. In Section 3, *relatively to a given binary relation over* A^*, we define the growth function of a language, the complexity function of an infinite word and the notions of avoidable patterns, periodicity and uniform recurrence. These notions are usually first introduced for the restrictive case of equality of factors, e.g., the complexity function counts the number of factors of length n occurring in a given infinite word but we could count them up to rearrangement of the letters. In the second part of the paper, we review classical binary relations on words where these concepts may be applied. In Section 4, we consider abelian equivalence, then its extension to k-abelian equivalence is presented in Section 5. We pursue in Section 6 with binomial coefficients of words and various equivalence relations that can be associated with. In Section 7, we present Parikh matrices and related relations. In the last section, we briefly present partial words and their generalizations to similarity relations. The bibliography is not exhaustive (it is limited to 100 entries) but we hope that it could provide relevant entry points to the existing literature. We limit ourselves to the onedimensional case. Indeed, many of the presented concepts have counterparts in a multidimensional setting.

2. Basics

We give some basic definitions about words. For general references, see [11, 66, 85]. Let A be a finite alphabet, i.e., a finite set of elements called letters. A *finite word* w over A is a finite sequence of elements in A. So it is a map $w : \{1, \ldots, n\} \to A$ where $n \in \mathbb{N}$ is

Date: August 27, 2016.

dedicated to the memory of my grandmother Suzanne Wiame 1932–2016.

the length of the word w. In particular, the empty sequence is called the *empty word* and is denoted by ε. Its length is 0. Note that the indexing of finite words begins at position 1. The set of finite words over A is denoted by A^*. Endowed with the concatenation of words as product operation, A^* is a monoid with ε as neutral element. We write $|w|$ for the length of the word w and $|w|_a$ for the number of occurrences of the letter a in w. We directly have $|w| = \sum_{a\in A} |w|_a$. Let n be a non-negative integer. We write A^n to denote the set of words of length n over A. Let A, B be two finite alphabets. A map $f : A^* \to B^*$ is a *morphism* (of monoids) if $f(uv) = f(u)f(v)$ for all $u, v \in A^*$ and, in particular, we have $f(\varepsilon) = \varepsilon$. A morphism f is *non-erasing* if $f(u) \neq \varepsilon$ for all non-empty words $u \in A^*$. A morphism is characterized by the images of the letters of its domain. If the images of the letters all have length 1, the morphism is called a *coding* (i.e., a letter-to-letter morphism).

Let S be a set, A be an alphabet and ℓ be a non-negative integer. We let $S^{A\times A}$ denote the set of square matrices of size $\#A$ with entries in S and indexed by pairs in $A \times A$. Similarly, we let $S^{\ell\times\ell}$ denote the set of square matrices of size ℓ with entries in S and indexed by pairs of integers belonging to $\{1, \ldots, \ell\}$.

An *infinite word* over A is a map $\mathbf{w} : \mathbb{N} \to A$. Note that the indexing of infinite words begins at position 0 (which is quite convenient when dealing, for instance, with automatic sequences). A *factor* $u = u_1 \cdots u_m$ of length m occurring in a finite word $v = v_1 \cdots v_n$ of length n is a block of consecutive letters occurring in it, i.e., $m \leq n$ and there exists $r \leq n - m$ such that $u_j = v_{r+j}$ for $j \in \{1, \ldots, m\}$. In that case, we say that u occurs in v at position $r + 1$. A *factor* of an infinite word $\mathbf{w}$ is a factor occurring in a finite prefix of $\mathbf{w}$. The set of factors (resp. the set of factors of length n) occurring in $\mathbf{w}$ is denoted by $\mathrm{Fac}_{\mathbf{w}}$ (resp. $\mathrm{Fac}_{\mathbf{w}}(n) := \mathrm{Fac}_{\mathbf{w}} \cap A^n$). We denote similarly the set of factors of a finite word.

A morphism $f : A^* \to A^*$ is *prolongable* on the letter $a \in A$ if there exists a finite word u such that $f(a) = au$ and if $\lim_{n\to+\infty} |f^n(a)| = +\infty$. In that case, the sequence $(f^n(a))_{n\geq 0}$ converges to an infinite word denoted by $f^\omega(a)$ that is said to be a *pure morphic word*. The image under a coding of a pure morphic word is said to be *morphic*. Let $k \geq 2$ be an integer. If the morphism $f : A^* \to A^*$ verifies $|f(a)| = k$ for all $a \in A$, then every infinite word of the form $g(f^\omega(a))$, where g is a coding, is said to be k-automatic [5, 27].

3. General framework

Let $\sim$ be a reflexive and symmetric binary relation over A^*. In many cases discussed in this survey, $\sim$ will be an equivalence relation (or even a congruence with respect to the concatenation of words). A trivial but useful example is given by the equality relation, each equivalence class is restricted to a singleton.

Example 1. Let $k \geq 1$. Let u, v be two words. We write $u \sim_{\mathcal{H},\leq k} v$, if $|u| = |v|$ and the Hamming distance between u and v is at most k, i.e.,

$$d_{\mathcal{H}}(u, v) := \sum_{i=1}^{|u|} (1 - \delta_{u_i,v_i}) \leq k$$

where $\delta_{a,b} = 1$, if $a = b$; and 0, otherwise. This relation is reflexive and symmetric but is not an equivalence relation. We have $abba \sim_{\mathcal{H},\leq 1} abaa$, $abaa \sim_{\mathcal{H},\leq 1} aaaa$ but $abba \not\sim_{\mathcal{H},\leq 1} aaaa$.

A *language* over A is a subset of A^* (we only consider languages of finite words). The concatenation of words is naturally extended to the concatenation of languages: if L, M are languages, $LM = \{uv \mid u \in L, v \in M\}$. Hence, the set 2^{A^*} of languages over A equipped with concatenation is a monoid with $\{\varepsilon\}$ as neutral element.

Definition 2 (growth function)**.** Let $\sim$ be an equivalence relation over A^* and $L \subset A^*$ be a language. We may consider the quotient $A^*/\sim$ and therefore the *growth function* of L *with respect to* $\sim$ is defined as

$$\mathsf{g}_{\sim,L} : \mathbb{N} \to \mathbb{N},\ n \mapsto \#\big((L \cap A^n)/\sim\big).$$

If $\sim$ is the equality relation, $\mathsf{g}_{=,L}$ simply counts the number of words of length n occurring in L. If $L = A^*$, then $\mathsf{g}_{\sim,A^*}(n)$ counts the number of equivalence classes of $\sim$ partitioning A^n.

Question 3. Given an equivalence relation $\sim$ over A^*. One can be interested in questions such as the following ones.

- Q.1.1 Compute or estimate the growth rate of $\mathsf{g}_{\sim,A^*}(n)$.
- Q.1.2 Given a specific language L, compute or estimate the growth order of $\mathsf{g}_{\sim,L}(n)$. Trivial bounds, for all $n \geq 0$, are given by
$$1 \leq \mathsf{g}_{\sim,A^*}(n) \leq \#(L \cap A^n) \leq (\#A)^n.$$
- Q.1.3 For a class $\mathcal{F}$ of languages (e.g., the set of regular languages, the set of algebraic languages, the set of factors occurring in Sturmian words, etc.), does $\mathsf{g}_{\sim,L}$ have special properties for all $L \in \mathcal{F}$? Can it provide a characterization of $\mathcal{F}$?

Example 4. If L is a regular language (i.e., accepted by a finite automaton), then $(\mathsf{g}_{=,L}(n))_{n\geq 0}$ satisfies a linear recurrence equation with integer coefficients. This is a well-known consequence of Cayley–Hamilton theorem applied to the adjacency matrix of an automaton whose nth power counts walks of length n between every pair of states.

As a special case of the previous definition, we can consider the *language of an infinite word*, i.e., the set of factors occurring in it.

Definition 5 (complexity function)**.** Let $\sim$ be an equivalence relation over A^*. Since the quotient $A^*/\sim$ is well-defined, we thus define the *complexity function* of an infinite word $\mathbf{w}$ *with respect to* $\sim$ as

$$\mathsf{p}_{\sim,\mathbf{w}} : \mathbb{N} \to \mathbb{N},\ n \mapsto \#\big(\mathrm{Fac}_{\mathbf{w}}(n)/\sim\big).$$

If $\sim$ is the equality relation, then $\mathsf{p}_{=,\mathbf{w}}$ is the usual *factor complexity* counting the number of factors of length n occurring in $\mathbf{w}$ [39]. The latter measure also leads to defining the *topological entropy* of $\mathbf{w}$. For a comprehensive presentation, see Cassaigne and Nicolas' chapter [11, Chap. 4]. For instance, $\mathsf{p}_{=,\mathbf{w}}$ is in $\mathcal{O}(n)$ for every automatic sequence $\mathbf{w}$ [27, 5]. For a pure morphic word $\mathbf{w}$, a theorem of Pansiot [74, 75] shows that the growth order of $\mathsf{p}_{=,\mathbf{w}}$ can only take five forms $\mathcal{O}(1), \mathcal{O}(n), \mathcal{O}(n \log\log n), \mathcal{O}(n \log n), \mathcal{O}(n^2)$. See the survey [3].

Question 6. Given an equivalence relation $\sim$ over A^*. One can be interested in questions such as the following ones.

Q.2.1 Given a specific infinite word $\mathbf{w}$, compute or estimate the growth order of $\mathsf{p}_{\sim,\mathbf{w}}(n)$.

Q.2.2 For a class $\mathcal{F}$ of words (e.g., the set of Sturmian words [67], the set of Arnoux–Rauzy words [6], the set of (pure) morphic words, automatic words, etc.), does $\mathsf{p}_{\sim,\mathbf{w}}$ have special properties for all $\mathbf{w} \in \mathcal{F}$? Can it provide a characterization of $\mathcal{F}$?

Q.2.3 A special case of the previous question is to study the set of words $\mathbf{w}$ such as $(\mathsf{p}_{\sim,\mathbf{w}}(n))_{n\geq 0}$ is bounded. Is there a Morse–Hedlund type result relating boundedness of the sequence $(\mathsf{p}_{\sim,\mathbf{w}}(n))_{n\geq 0}$ to the (ultimate) periodicity of the word $\mathbf{w}$ (see Theorem 7 and Definition 17 for generalizations of the concept of periodicity).

Q.2.4 Does $(\mathsf{p}_{\sim,\mathbf{w}}(n))_{n\geq 0}$ have a geometrical or a dynamical interpretation if $\mathbf{w}$ is derived from a dynamical system such as a coding of rotation? One can also be interested in arithmetical or algebraic interpretations when $\mathbf{w}$ is the expansion of a real number in a specific numeration system.

Theorem 7 (Morse–Hedlund [72])**.** *An infinite word* $\mathbf{w}$ *is ultimately periodic, i.e.,* $\mathbf{w} = uvvv\cdots$ *for some finite words* u, v*, if and only if the sequence* $(\mathsf{p}_{=,\mathbf{w}}(n))_{n\geq 0}$ *is bounded (by a constant). Otherwise stated, either* $\mathbf{w}$ *is ultimately periodic, or* $\mathsf{p}_{=,\mathbf{w}}$ *is increasing.*

For a proof, for instance, see [11, Section 4.3] or [5, Thm. 10.2.6].

Example 8. Sturmian words have been extensively studied [67, 9] and several characterizations do exist. They can be defined as codings of particular rotations on the normalized interval $[0, 1)$ with irrational angle $\alpha < 1$ when the interval $[0, 1)$ is split into $[0, 1-\alpha)$ and $[1-\alpha, 1)$. An infinite word $\mathbf{w}$ is Sturmian if and only if $\mathsf{p}_{=,\mathbf{w}}(n) = n+1$ for all $n \geq 0$. For rotation words obtained with another partition of the interval $[0, 1)$, see [10, 33]. For the abelian equivalence $\sim_{\mathsf{ab}}$ discussed in the next section, Coven and Hedlund proved that an aperiodic word $\mathbf{w}$ is Sturmian if and only if $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{w}}(n) = 2$ for all $n \geq 1$ [29].

Remark 9. If the equivalence relation $\sim$ is a *congruence* over A^*, i.e., for all u_1, u_2, v_1, v_2, if $u_i \sim v_i$ for $i = 1, 2$, then $u_1u_2 \sim v_1v_2$, then the complexity function with respect to $\sim$ satisfies

$$\mathsf{p}_{\sim,\mathbf{w}}(m+n) \leq \mathsf{p}_{\sim,\mathbf{w}}(m).\mathsf{p}_{\sim,\mathbf{w}}(n).$$

Indeed, every factor of length $m+n$ is the concatenation of a factor of length m with a factor of length n but the converse does not necessarily holds. The concatenation of two factors is not always a factor occurring in $\mathbf{w}$.

Since the works of Thue, the study of repetitions and unavoidable patterns is one of the cornerstones in combinatorics on words [9, 95, 96]. An infinite word $\mathbf{w} \in A^{\mathbb{N}}$ *avoids* a set $S \subseteq A^*$, if $\mathrm{Fac}_{\mathbf{w}} \cap S = \emptyset$. If such a word $\mathbf{w}$ exists, we say that S is *avoidable* over A. A set $S \subseteq A^*$ is *unavoidable* over A whenever, for all $\mathbf{w} \in A^{\mathbb{N}}$, $\mathrm{Fac}_{\mathbf{w}} \cap S \neq \emptyset$. We now introduce the notion of $\sim$-unavoidable pattern. For a survey on repetitions and avoidance, see Rampersad and Shallit's chapter [12, Chap. 4].

Definition 10 (avoidance)**.** Let B be a finite alphabet. Any finite word over B will be called a *pattern*. Let $\sim$ be an equivalence relation over A^*. We now define a language-valued morphism that we will call a *substitution*[1]. Let $h : B \to 2^{A^*}$ be a map satisfying

(1) for all $b \in B$, $h(b)$ is a non-empty set and $\varepsilon \notin h(b)$;
(2) for all $b \in B$, if $u, v \in h(b)$, then $u \sim v$;

Note that the image of every letter $b \in B$ is a subset of an equivalence class for $\sim$. The map h is then extended to a morphism from B^* to 2^{A^*} by setting $h(\varepsilon) = \{\varepsilon\}$ and $h(PQ) = h(P)h(Q)$ for all $P, Q \in B^*$. We say that the morphism h is a $\sim$-*substitution*.

Let $P \in B^*$ be non-empty. The pattern P is $\sim$-*unavoidable* over A if the language

$$\mathcal{L}_\sim(P) := \bigcup_{\substack{h:B^*\to 2^{A^*} \\ h \text{ is a } \sim\text{−substitution}}} h(P) \subset A^*$$

is unavoidable over A. Otherwise, P is $\sim$-*avoidable* over A. If $\sim$ is the equality relation, we get back to the classical notion of avoidance. Every $=$-substitution is a non-erasing morphism and conversely. Note that it is enough to consider in the union defining $\mathcal{L}_\sim(P)$, the substitutions mapping letters of B to equivalence classes of $\sim$. The set $\mathcal{L}_\sim(P)$ is called the *pattern language* associated with P and $\sim$.

Definition 11. Let $\sim$ be an equivalence relation over A^*. A $\sim$-*square* (resp. a $\sim$-*cube*) is a word in $\mathcal{L}_\sim(XX)$ (resp. in $\mathcal{L}_\sim(XXX)$), $X \in B$. In general, a $\sim$-*nth-power* is a word in $\mathcal{L}_\sim(X^n)$, $X \in B$, $n \in \mathbb{N}$.

Question 12. Let $\sim$ be an equivalence relation over A^*.

Q.3.1 Given a pattern and an alphabet A of size k, is this pattern $\sim$-avoidable over A? In particular, are $\sim$-squares or $\sim$-cubes avoidable? As an example, the Thue–Morse word avoids $=$-cubes or even overlaps corresponding to the pattern $XYXYX$. For a proof, for instance, see [66].

Q.3.2 Given a pattern that is $\sim$-avoidable, what is the minimal size of the alphabet such that it can be avoided?

Q.3.3 Given a pattern P, an alphabet A of size k and an integer ℓ, does there exist an infinite word $\mathbf{w}$ over A such that

$$\#\big(\mathrm{Fac}_{\mathbf{w}} \cap \mathcal{L}_\sim(P)\big) \leq \ell.$$

Note that this is Q.3.1 when $\ell = 0$.

Q.3.4 A modification of the previous question is to ask whether it exists an infinite word $\mathbf{w}$ such that

$$\mathrm{Fac}_{\mathbf{w}} \cap \mathcal{L}_\sim(P) \subseteq A^{\leq n}$$

for some n. Otherwise stated, we only allow short occurrences of the pattern P.

[1]We here use the term 'substitution' to avoid any confusion with the term 'morphism'. In the literature, the word substitution is sometimes interchanged with morphism or non-erasing prolongable morphism.

Q.3.5 Let P be a pattern over B. A finite word $u \in A^*$ is *$\sim$-P-free*, if

$$\mathrm{Fac}_u \cap \mathcal{L}_\sim(P) = \emptyset.$$

A morphism $f : A^* \to A^*$ is *$\sim$-P-free* if, for all $\sim$-P-free words u, $f(u)$ also is $\sim$-P-free. Given a pattern P and an alphabet A, does there exist a non-trivial prolongable $\sim$-P-free morphism? If such a morphism exists, then P is $\sim$-avoidable over the alphabet A [19]. As an example, the Thue–Morse morphism $a \mapsto ab$, $b \mapsto ba$ is overlap-free [66].

Q.3.6 One can also be interested in enumeration questions such as counting the number of $\sim$-P-free finite words of length n. We give a few references where some interesting growth rates are exhibited [23, 63, 59, 31].

Remark 13. A variation of the notion of $\sim$-nth-power is defined as follows. Let $\sim$ be a binary relation over A^*. A word u is a *strongly $\sim$-nth power* if there exists a 'classical' nth power such that $u \sim v^n$. We will specialize this notion in Remark 40, when discussing strongly ℓ-abelian powers.

There is also a notion of *approximated squares* introduced in [73]. As an example, a word of the form uv with $u \sim_{\mathcal{H},\leq k} v$ can be considered as an approximated square, with the relation defined in Example 1.

Example 14. Related to questions Q.3.3 and Q.3.4, Fraenkel and Simpson have built an infinite word over a 2-letter alphabet with only 3 squares: 00, 11 and 0101 [42]. (It is easy to see that over a 2-letter alphabet, any word of length at least 4 contains a square.)

Properties 15. *Every infinite word over a 2-letter alphabet contains arbitrarily long squares and there exists an infinite word that avoids squares of the form uu with $|u| \geq 3$ [41].*

Remark 16. The reader may also think about pattern matching. In this survey, let us already mention that the topic will be considered, for two special cases only (ℓ-abelian equivalence and k-binomial equivalence), in Remarks 41 and 54.

The following definition is inspired by the definition given in [45] for similarity relations (see Section 8) and relational periods (in that case, the parameter ℓ is always equal to 1). For a survey, see [12, Chap. 6]. We will consider factorizations of an infinite word with words of a fixed length ℓ but one could relax this assumption. We are looking for a 'period' made of p words of length ℓ.

Definition 17 (periodicity)**.** Let $\sim$ be a reflexive and symmetric binary relation over A^*. Let $\mathbf{w}$ be an infinite word over A. Let $p, \ell \geq 1$ be integers.

(1) The word $\mathbf{w}$ has (p, ℓ) as *global $\sim$-period* if there exists a sequence $(u_i)_{i\geq 0}$ of words of length ℓ such that $\mathbf{w} = u_0u_1u_2\cdots$ and, for all $i, j \in \mathbb{N}$,

$$i \equiv j \pmod{p} \Rightarrow u_i \sim u_j.$$

(2) The word $\mathbf{w}$ has (p, ℓ) as *external $\sim$-period* if there exist p words $v_0, \ldots, v_{p-1}$ and a sequence $(u_i)_{i\geq 0}$ of words of length ℓ such that $\mathbf{w} = u_0u_1u_2\cdots$ and, for all $n \in \mathbb{N}$ and all $r \in \{0, \ldots, p-1\}$, $u_{np+r} \sim v_r$.

(3) The word $\mathbf{w}$ has (p, ℓ) as *local* $\sim$*-period* if there exists a sequence $(u_i)_{i\geq 0}$ of words of length ℓ such that $\mathbf{w} = u_0u_1u_2\cdots$ and, for all $i \geq 0$, $u_i \sim u_{i+p}$.

If such a pair (p, ℓ) exists, we say that $\mathbf{w}$ is *globally* (resp. *externally*, *locally*) $\sim$*-periodic* and (p, ℓ) is a *global* (resp. *external*, *local*) $\sim$*-period*.

Example 18. Let $u \sim_{\mathcal{H},\leq 1} v$ be the relation defined in Example 1. Consider the generalized Thue–Morse word $\mathbf{t}_{3,3}$ (OEIS A004128)[2] over $\{0, 1, 2\}$

$$0121202011202010122010121201200122 01\cdots$$

and apply the morphism $0 \mapsto aaaa$, $1 \mapsto abaa$ and $2 \mapsto abba$ to get the word

$$\mathbf{w} = aaaaabaaabbaabaaabbaaaaaabbaaaaaabaaabaaabbaaaaaabbaaaaaabaa\cdots$$

It has external period $(1, 4)$, for all $n \geq 0$, $w_{4n}w_{4n+1}w_{4n+2}w_{4n+3} \sim_{\mathcal{H},\leq 1} abaa$. Actually, the proposed morphism maps any word over a 3-letter alphabet to a word with external period $(1, 4)$.

Remark 19. In the previous definition, if $\sim$ is also transitive, i.e., $\sim$ is an equivalence relation, then the three notions of global, external and local $\sim$-periods coincide. In that case, we simply say that a word is $\sim$-periodic or *ultimately* $\sim$-periodic if it has a $\sim$-periodic suffix.

Lemma 20. *Let $\sim$ be a congruence over A^*. If $\mathbf{w}$ has the pair (p, ℓ) as $\sim$-period, then $\mathbf{w}$ has $(1, p\ell)$ as $\sim$-period*

Proof. The exists a sequence $(u_i)_{i\geq 0}$ of words of length ℓ such that $\mathbf{w} = u_0u_1u_2\cdots$ and, for all $n \in \mathbb{N}$ and $r \in \{0, \ldots, p-1\}$, $u_{np+r} \sim u_r$. Since $\sim$ is a congruence, for all $n \in \mathbb{N}$, $u_{np}u_{np+1}\cdots u_{np+p-1} \sim u_0\cdots u_{p-1}$. Thus we can consider the sequence $(u_{np}u_{np+1}\cdots u_{np+p-1})_{n\geq 0}$ of words of length $p\ell$ showing that $\mathbf{w}$ is $(1, p\ell)$-periodic. □

Question 21. Given a (pure) morphic $\mathbf{w}$ (or a word given with a finite description) and a relation $\sim$, is it decidable whether or not $\mathbf{w}$ is of the form $u\mathbf{x}$ where u is a finite word and $\mathbf{x}$ is globally (resp. externally, locally) $\sim$-periodic? See, for instance, [48, 76, 37, 47].

Remark 22. Other periodicity-related topics such as variants of Fine–Wilf theorem [8, 16, 56, 18, 17, 46, 28] or codes and defect effect [45, 60] may be considered.

We introduce the last concept of this part of the paper. A subset $X = \{x_0 < x_1 < x_2 < \cdots\} \subseteq \mathbb{N}$ is *syndetic* (or, *with bounded gaps*) if there exists a constant C such that $x_{i+1} - x_i < C$ for all $i \geq 0$. In the last part of this section, we assume that if $u \sim v$, then $|u| = |v|$.

Definition 23 (uniform recurrence)**.** Let $\sim$ be a reflexive and symmetric binary relation over A^*. For every $u \in \mathrm{Fac}_{\mathbf{w}}$, consider the set of positions where occurs a factor in relation with u

$$\mathsf{Occ}_{\sim,u}(\mathbf{w}) := \{i \geq 0 \mid v_i\cdots v_{i+|u|-1} \sim u\}$$

[2]Let $m \geq 2$ and $k \geq 2$ be integers. The infinite word $\mathbf{t}_{k,m} := (s_k(n) \mod m)_{n\geq 0}$ over the alphabet $\{0, \ldots, m-1\}$, where $s_k(n)$ is the sum-of-digits of the base-k expansion of n, is overlap-free if and only if $k \leq m$. It is also known that $\mathbf{t}_{k,m}$ contains arbitrarily long squares [4].

If for all $u \in \mathrm{Fac}_{\mathbf{w}}$, the set $\mathsf{Occ}_{\sim,u}(\mathbf{w})$ is infinite (resp. infinite and syndetic), then we say that $\mathbf{w}$ is $\sim$-*recurrent* (resp. $\sim$-*uniformly recurrent*).

Definition 24. If $\mathbf{w}$ is $\sim$-uniformly recurrent, then we can factorize the word $\mathbf{w}$ using the set of positions $\mathsf{Occ}_{\sim,u}(\mathbf{w}) = \{i_1 < i_2 < \cdots\}$:

$$\mathbf{w} = (w_0 \cdots w_{i_1-1})(w_{i_1} \cdots w_{i_2-1})(w_{i_2} \cdots w_{i_3-1}) \cdots$$

Observe that uniform recurrence implies that the set of words $\{w_{i_j} \cdots w_{i_{j+1}-1} \mid j \geq 1\}$ is finite. These words are called the $\sim$-*return words* to u. Each such word shares a common prefix with a word in relation with u for $\sim$. If it is longer than u then it has u' as a prefix for some $u' \sim u$. This notion is similar to the first return map in dynamical systems theory.

Example 25. Consider the Thue–Morse word (OEIS A010060) and the abelian equivalence $\sim_{\mathsf{ab}}$ precisely defined in the next section. Two words are abelian equivalent if one is obtained by permuting the letters of the other. With the prefix 01101, we have marked all the occurrences of a factor of length 5 having precisely 3 ones, i.e., that is a rearrangement (or anagram) of this prefix:

$$|\underbrace{0}_{1}|\underbrace{110}_{2}|\underbrace{100}_{3}|110|0|\underbrace{1}_{4}|0|\underbrace{11010}_{5}|0|\underbrace{10}_{6}|1|10|0|110|100|110|0|10|1|10|0|1101001\cdots$$

One can prove that the only factors that occur are $\{0, 1, 10, 100, 110, 11010\}$ mapping this set onto $\{1, \ldots, 6\}$ (where the usual convention is that the index is given by the order of first appearance of the factor within the factorized word), we can code the previous factorization by

$$123214151646123216461\cdots$$

Such a sequence is called a *derived sequence* and is denoted by $\mathsf{D}_{\sim_{\mathsf{ab}},01101}(\mathbf{w})$

This concept of derived sequence (or descendant) was introduced independently by Durand [36] and Holton and Zamboni [49]. A morphism $f : A^* \to A^*$ is *primitive* if the matrix $M = (|f(a)|_b)_{a,b\in A} \in \mathbb{N}^{A\times A}$ is primitive, i.e., there exists n such that all entries of M^n are positive (we write $M^n > 0$).

Theorem 26. [36] *An infinite uniformly recurrent word* $\mathbf{w}$ *is of the form* $g(f^\omega(a))$ *where* $g : A^* \to B^*$ *is a coding and* $f : A^* \to A^*$ *is a primitive morphism prolongable on* a *if and only if the set* $\{\mathsf{D}_{=,p}(\mathbf{w}) \mid p$ *is a prefix of* $\mathbf{w}\}$ *is finite.*

Proposition 27. [36, Prop. 5.1] *Let* f *be a primitive morphism prolongable on the letter* a*. For every prefix* $p \neq \varepsilon$ *of* $f^\omega(a)$*, the sequence* $\mathsf{D}_{=,p}(f^\omega(a))$ *is also the fixed point of a primitive morphism.*

4. Abelian framework

Erdős raised the question whether abelian squares can be avoided by an infinite word over an alphabet of size 4. We refer to the paper [40] that can easily be accessed[3], the last

[3]It is common to refer to another Erdős'paper: Some unsolved problems, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221–254.

problem of the list of 28 problems is the following: "*Let $N(k)$ be the least number N with the property that each sequence $\{s_n\}_{n=1}^N$ of numbers taken from the set $\{1,\dots,k\}$ contains two adjacent blocks such that each is a rearrangement of the other. My earliest conjecture, that $N(k) = 2^k - 1$, has been disproved by Bruijn and myself. It is not even known whether $N(4) < \infty$.*" (Exhausting all the possible cases, it is an easy exercise to prove that any long enough finite word over an alphabet of size 3 contains an abelian square.)

Definition 28. Let $A = \{1 < \cdots < k\}$ be a finite alphabet that is assumed to be ordered. We define the *abelianization map* (also called *Parikh map*, see Theorem 35) denoted by $\Psi : A^* \to \mathbb{N}^k$. It is a morphism of monoids where $\Psi(u) = (|u|_1, \dots, |u|_k)^{\mathsf{T}}$ for all $u \in A^*$.

Indeed, $\Psi(uv) = \Psi(u) + \Psi(v)$ for all $u, v \in A^*$. In particular, if Ψ can be extended to languages and $\Psi^{-1}(\Psi(L))$ is the *commutative closure* of the language L.

Definition 29. The notion of abelian square introducing this section is a special case of Definition 11 when considering the *abelian equivalence* $\sim_{\mathsf{ab}}$ over A^* defined by

$$u \sim_{\mathsf{ab}} v \Leftrightarrow \Psi(u) = \Psi(v).$$

Otherwise stated, u is obtained by applying a permutation to the letters of v. The relation $\sim_{\mathsf{ab}}$ is clearly a congruence.

About Definition 5, one introduces the notion of *abelian complexity* $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{w}}$ where factors occurring in $\mathbf{w}$ are counted up to abelian equivalence. In contrast with the usual factor complexity function $\mathsf{p}_{=,\mathbf{w}}$ which is non-decreasing, this property no longer holds for $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{w}}$: it is possible that $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{w}}(n) > \mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{w}}(n+1)$ for some n. For instance, for the Tribonacci word $\mathbf{t}$ (OEIS A000073) $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{t}}(7) = 4$ but $\mathsf{p}_{\sim_{\mathsf{ab}},\mathbf{t}}(8) = 3$. A few references are [83, 84, 15, 97, 98] and [68] where the abelian complexity of the paper-folding word is shown to be 2-regular (in the sense of Allouche and Shallit), see, for instance, [7]. In particular, bounded abelian complexity is related to balance properties and existence of frequencies [2].

Theorem 30. [84] *An infinite word has a bounded abelian complexity if and only if it is C-balanced for some $C > 0$, i.e., for all $u, v \in \mathrm{Fac}_{\mathbf{w}}(n)$, $n \geq 1$, we have $\big|\, |u|_a - |v|_a \,\big| \leq C$ for every letter a in the alphabet.*

Properties 31. *Keränen has built a pure morphic word over a 4-letter alphabet that avoids abelian squares* [61, 20]*. Dekking has obtained an infinite word over a 3-letter alphabet that avoids abelian cubes, and an infinite word over a 2-letter alphabet that avoids abelian 4-powers* [32]*. (Note that in all these results, the size of the alphabet is optimal.)*

About *abelian power-free morphisms*, see [21, 30]. See also [25].

About enumeration results like counting the number of finite words of length n avoiding abelian cubes, see [1, 22].

On the characterization of classes of words with respect to abelian equivalence and in particular, Sturmian words. Let us mention the following results. Extending a result of Vuillon in [100] to $\sim_{\mathsf{ab}}$. (recall Definition 24 of return words.)

Theorem 32. [79] *A recurrent infinite word is Sturmian if and only if each of its factors has two or three $\sim_{\mathbf{ab}}$-return words.*

Properties 33. [87]

(1) *Let* $\mathbf{w}$ *be a recurrent word. The set $\sim_{\mathbf{ab}}$-return words is finite if and only if* $\mathbf{w}$ *is periodic.*
(2) *Let* $\mathbf{w}$ *be a Sturmian word (we assume that the notion of intercept is understood). The set of $\sim_{\mathbf{ab}}$-return words to the prefixes is finite if and only if* $\mathbf{w}$ *has a non-zero intercept.*

The latter result can be extended to rotation words [80].

Remark 34. Closely related to abelian equivalence, one can also consider an *additive relation* where two words u, v (one can add the extra assumption that $|u| = |v|$) over a finite alphabet of integers are *additively equivalent*, if $\sum u_i = \sum v_i$. For instance, 134233 is an additive square. The paper [24] shows the existence of an infinite word over $\{0, 1, 3, 4\}$ avoiding additive cubes (OEIS `A191818`). Also see [81] where subsets of $\mathbb{N}$ of size 3 are considered.

5. k-abelian equivalence

We now present a first generalization of the concept of abelian equivalence stemming from a classical result in formal language theory: Parikh's theorem. See any standard textbook on formal language theory, e.g. [94], in particular for the definition of a context-free language. A set $M \subseteq \mathbb{N}^d$ is said to be *linear*, if there exist $x \in \mathbb{N}^d$ and a finite set (possibly empty) $V = \{v_1, \ldots, v_k\} \subset \mathbb{N}^d$ such that

$$M = \left\{ x + \sum_{i=1}^{k} \lambda_i \, v_i \mid \lambda_1, \ldots, \lambda_k \in \mathbb{N} \right\}.$$

A finite union of linear sets is a *semi-linear* set.

Theorem 35 (Parikh's theorem [78])**.** *If L is a context-free language over a k-letter alphabet, then $\Psi(L)$ is a semi-linear set of $\mathbb{N}^k$.*

Let $\ell \geq 1$. Trying to strengthen Parikh's theorem, instead of counting occurrences of letters, we could count occurrences of factors of length at most ℓ [55]. In that setting, assuming that $A = \{1 < \cdots < k\}$ is ordered, we get extra information on the structure of the word given by an extended abelianization map, also called *generalized Parikh mapping*,

$$\Psi_\ell : A^* \to \mathbb{N}^{k+k^2+\cdots+k^\ell}$$

where, for all $u \in A^*$,

$$\Psi_\ell(u) = (|u|_1, \ldots, |u|_k, |u|_{11}, \ldots, |u|_{kk}, \ldots, |u|_{1^\ell}, \ldots, |u|_{k^\ell}) \tag{1}$$

and $|u|_v$ denotes the number of occurrences of the factor v in u. Note that the size of $\Psi_\ell(u)$ grows exponentially with ℓ: it is a vector of size $k(k^\ell - 1)/(k - 1)$. As an example, $|0110100|_{10} = 2$ and $|01110|_{11} = 2$ (overlaps are allowed). The following relation was introduced in [57].

Definition 36. Let $\ell \geq 1$ be an integer. Two finite words u and v are *ℓ-abelian equivalent*, if $\Psi_\ell(u) = \Psi_\ell(v)$. We write $u \sim_{\ell-\mathsf{ab}} v$. Otherwise stated, if, for all words $x \in A^{\leq \ell}$, $|u|_x = |v|_x$. Clearly, for $\ell = 1$ we are back to the usual abelian equivalence.

Note that, for all $n \leq |u|$,

$$|u| = \sum_{x \in A^n} |u|_x + n - 1$$

and $\Psi_\ell(A^*)$ is a strict subset of $\mathbb{N}^{k(k^\ell-1)/(k-1)}$.

Example 37. The words $u = 010110$ and $v = 011010$ are 3-abelian equivalent. We have $|u|_0 = 3 = |v|_0$, $|u|_1 = 3 = |v|_1$, $|u|_{00} = 0 = |v|_{00}$, $|u|_{01} = 2 = |v|_{01}$, $|u|_{10} = 2 = |v|_{10}$, $|u|_{11} = 1 = |v|_{11}$. Finally, $|u|_{010} = 1 = |v|_{010}$, $|u|_{101} = 1 = |v|_{101}$, $|u|_{011} = 1 = |v|_{011}$ $|u|_{110} = 1 = |v|_{110}$. But the two words u and v are not 4-abelian equivalent: the factor 1010 occurs in v but not in u. The relation $\sim_{(\ell+1)-\mathsf{ab}}$ is a refinement of $\sim_{\ell-\mathsf{ab}}$ (see the lattice in Figure 1).

Remark 38. In terms of rational series (we refer the reader to [7] for definitions), since the characteristic series of A^* denoted by $\underline{A}^*$ is rational, we deduce that the formal series in $\mathbb{N}\langle\langle A \rangle\rangle$

$$\underline{A}^* u \underline{A}^* = \sum_{w \in A^*} |w|_u \, w$$

is rational.

It is not difficult to see that two words u and v of length at least $\ell - 1$ are ℓ-abelian equivalent if and only if they share respectively the same prefix and the same suffix of length $\ell - 1$ and if $|u|_x = |v|_x$ for all words x of length ℓ. This property implies that $\sim_{\ell-\mathsf{ab}}$ is again a congruence. In [57], the growth of $\mathsf{g}_{\sim_{\ell-\mathsf{ab}}}$ is estimated. Ultimately periodic words and Sturmian words can be characterized by the ℓ-abelian complexity function.

Theorem 39. [57] *Let $\ell \geq 1$. An infinite aperiodic word $\mathbf{w}$ is Sturmian if and only if*

$$\mathsf{p}_{\sim_{\ell-\mathsf{ab}},\mathbf{w}}(n) = \begin{cases} n+1, & \text{if } n < 2\ell; \\ 2\ell, & \text{if } n \geq 2\ell. \end{cases}$$

About the fluctuations of $\mathsf{p}_{\sim_{\ell-\mathsf{ab}},\mathbf{w}}$, see the papers [26, 58]. The 2-abelian complexity of the Thue-Morse word is shown to be 2-regular in [77] and, independently, in [44].

Many results on avoidance are available. In [81], Rao provides morphic words avoiding ℓ-abelian powers: an infinite word over a 2-letter alphabet avoiding 2-abelian cubes and an infinite word over a 3-letter alphabet avoiding 3-abelian squares. The paper also deals with bounds on enumeration results in that context of avoidance. About other avoidance results, also see [50, 52]

Remark 40. A variant of the notion of repetition is considered in [51], a word is a *strongly ℓ-abelian nth power*, if it is ℓ-abelian equivalent to a 'classical' nth power. As an example, the word $aabb$ is not an abelian square because $aa \not\sim_{\mathsf{ab}} bb$ but it is a strongly abelian square because $aabb \sim_{\mathsf{ab}} (ab)(ab)$.

Remark 41. [ℓ-abelian pattern matching] Pattern matching has many applications, here we concentrate on 'approximate' pattern matching problems (that can be considered with respect to a given equivalence relation). In [38], making use of suffix arrays, the following problems are positively answered.

- Given $\ell \geq 1$ and two words u, v of length n, decide, in polynomial time with respect to n and ℓ, whether or not $u \sim_{\ell-\mathsf{ab}} v$.
- Given $\ell \geq 1$ and two words w, x, find, in polynomial time, all occurrences of factors of w which are ℓ-abelian equivalent to x.
- Given two u, v of length n, find the largest ℓ such that $u \sim_{\ell-\mathsf{ab}} v$.

6. Binomial coefficients

The notion of a binomial coefficient of words is classical in combinatorics on words. See, for instance, Sakarovitch and Simon's chapter in [66]. Let $w, x \in A^*$. The integer denoted by

$$\binom{w}{x}$$

counts the number of times x appears as a (scattered) subword[4] of w, i.e., x occurs as a subsequence of w. Otherwise stated, we count the number of increasing maps $\varphi : \{1, \ldots, |x|\} \to \{1, \ldots, |w|\}$ such that

$$\varphi(1) < \cdots < \varphi(|x|) \quad \text{and} \quad w_{\varphi(1)} \cdots w_{\varphi(|x|)} = x.$$

As an example, we have $\binom{aabbab}{ab} = 7$. It generalizes the usual binomial coefficients of integers because, over a 1-letter alphabet,

$$\binom{a^m}{a^n} = \binom{m}{n}, \quad m, n \in \mathbb{N}.$$

These coefficients can easily be computed from the relations

$$\binom{w}{\varepsilon} = 1, \qquad \binom{w}{x} = 0, \quad \text{if } |w| < |x|$$

and

$$\forall u, v \in A^*, a, b \in A, \quad \binom{ua}{vb} = \binom{u}{vb} + \delta_{a,b}\binom{u}{v}.$$

Remark 42. We have an observation similar to Remark 38. Let $u = u_1 \cdots u_n$. In terms of rational series (we again refer to [7]), since the characteristic series of A^* is rational, we deduce that the formal series in $\mathbb{N}\langle\langle A \rangle\rangle$

$$\underline{A}^* u_1 \underline{A}^* u_2 \underline{A}^* \cdots \underline{A}^* u_n \underline{A}^* = \sum_{w \in A^*} \binom{w}{u} w$$

is rational.

[4]This is the reason why we make a distinction between factors made of consecutive letters and subwords. Be aware that in the literature these two terms are sometimes used with the same meaning.

It is not difficult to prove the following result.

Proposition 43. *Let s, t, w be three words of A^*. Then we have*

$$\binom{sw}{t} = \sum_{uv=t} \binom{s}{u}\binom{w}{v}.$$

Let us mention the so-called *Cauchy inequality*. Several proofs of this result exist, see [88].

Theorem 44. *For all words $w, x, y, z \in A^*$, we have*

$$\binom{w}{y}\binom{w}{xyz} \leq \binom{w}{xy}\binom{w}{yz}.$$

A general question is to 'reconstruct' a word from some of its binomial coefficients: What numbers $\binom{w}{u}$ suffice to determine the word w uniquely? See, for instance, [89]. Schützenberger and Simon proved that two words of length n with the same subwords of length up to $\lfloor n/2 \rfloor + 1$ are identical. In [64], it is shown that any word of length n is uniquely determined by all its subwords of length k, if $k \geq \lfloor 16\sqrt{n}/7 \rfloor + 5$. The authors relate this problem to well-known vertex reconstruction problems in graph theory and trace the origin of the problem back to [53]. For algorithmic considerations, see, for instance, [34]. About generalized Pascal triangle and Sierpiński gasket, see [65].

Similarly to ℓ-abelian equivalence, these binomial coefficients allows us to define an independent refinement of abelian equivalence.

Definition 45. Let $k \geq 1$. Two words u, v are *k-binomially equivalent*, and we write $u \sim_{k-\mathsf{bin}} v$, if and only if, for all $x \in A^{\leq k}$,

$$\binom{u}{x} = \binom{v}{x}.$$

In particular, since $\binom{w}{a} = |w|_a$, for $a \in A$, if $k = 1$, then we have the usual equivalence relation $\sim_{\mathsf{ab}}$. The fact that $\sim_{k-\mathsf{bin}}$ is a congruence is a consequence of Proposition 43. Similarly to (1), assuming that $A = \{1 < \cdots < t\}$ is ordered, one could introduce the map

$$\Psi'_k(u) := \left(\binom{u}{1}, \ldots, \binom{u}{t}, \binom{u}{11}, \ldots, \binom{u}{tt}, \ldots, \binom{u}{1^k}, \ldots, \binom{u}{t^k}\right)$$

and u, v are k-binomially equivalent if and only if $\Psi'_k(u) = \Psi'_k(v)$. In [91], the 2-binomial equivalence was called *binary equivalence*.

As observed in [35], if $|u| \geq k \geq |x|$, then

$$\binom{|u| - |x|}{k - |x|}\binom{u}{x} = \sum_{t \in A^k} \binom{u}{t}\binom{t}{x}.$$

Indeed, on the right hand side, a fixed occurrence of the subword x in u is counted as many times as it appears in any bigger subword. Thus, if the positions of the letters of x are fixed, we can build a bigger subword made of k symbols and containing that particular

occurrence of x by selecting $k-|x|$ positions amongst the $|u|-|x|$ remaining ones in u. Consequently, we deduce the following result.

Lemma 46. *If u, v are words of length at least k, then $u \sim_{k-\mathsf{bin}} v$, if and only if $\binom{u}{t} = \binom{v}{t}$ for all words t of length k.*

Example 47. The four words *ababbba*, *abbabab*, *baabbab* and *babaabb* are 2-binomially equivalent. For any w amongst these words, we have the following coefficients

$$\binom{w}{aa} = 3,\ \binom{w}{ab} = 7,\ \binom{w}{ba} = 5,\ \binom{w}{bb} = 6.$$

But one can check that they are not 3-binomially equivalent, as an example, take the word $w_1w_2\cdots w_7 = ababbba$. We have

$$\binom{ababbba}{aab} = 3 \text{ but } \binom{abbabab}{aab} = 4$$

indeed, for this last binomial coefficient, *aab* appears as subwords $w_1w_4w_5$, $w_1w_4w_7$, $w_1w_6w_7$ and $w_4w_6w_7$. The k-abelian equivalence and the k-binomial equivalence relations are incomparable. Considering again the first two words, we find $|ababbba|_{ab} = 2$ and $|abbabab|_{ab} = 3$, showing that these two words are not 2-abelian equivalent. Conversely, the words *abbaba* and *ababba* are 2-abelian equivalent but are not 2-binomially equivalent:

$$\binom{abbaba}{ab} = 4 \text{ but } \binom{ababba}{ab} = 5.$$

Remark 48. Since the relation $\sim_{(k+1)-\mathsf{bin}}$ is a refinement of $\sim_{k-\mathsf{bin}}$, we have a lattice of relations over A^* as depicted in Figure 1. The coarsest relation is abelian equivalence and the finest relation is equality. To provide a single figure, we have already represented on the rightmost branch in this lattice the relations $\sim_{\psi_{w_0\cdots w_k}}$ that will be introduced in Definition 62.

In the literature, one also finds the notion of *k-spectrum* of a word u which is the (formal) polynomial (we refer to [7] for definitions) in $\mathbb{N}\langle A^*\rangle$ of degree k

$$\mathsf{Spec}_{u,k} := \sum_{w\in A^{\le k}} \binom{u}{w} w.$$

The 2-spectrum of the word $u = abbab$ is

$$\mathsf{Spec}_{u,2} = 1\varepsilon + 2a + 3b + aa + 4ab + 2ba + 3bb.$$

If we replace a with 0 and b with 1 and if every word is preceded by a leading 1, every word w over $\{a, b\}$ corresponds to a unique integer (there is no leading 0) written in base-2, $\mathrm{val}_2(1w)$, so this spectrum can also be represented as a univariate polynomial where the word w is replaced with $X^{\mathrm{val}_2(1w)}$. With the same word u, we have

$$1 + 2X^2 + 3X^3 + X^4 + 4X^5 + 2X^6 + 3X^7. \tag{2}$$

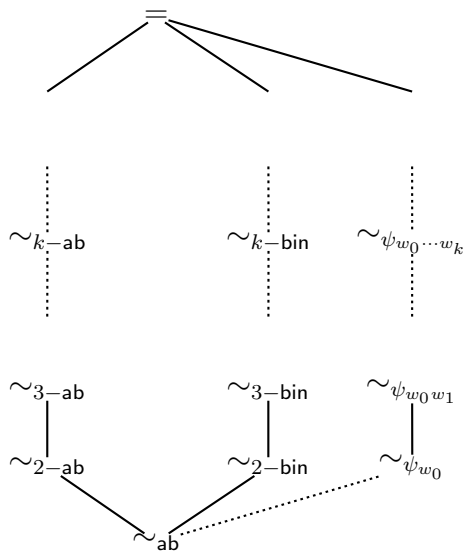

FIGURE 1. A lattice of congruences, the finest is $=$, the coarsest is $\sim_{\mathsf{ab}}$.

The 3-spectrum of the same word u is

$$\mathsf{Spec}_{u,3} = \mathsf{Spec}_{u,2} + aab + 2aba + 3abb + 2bab + bba + bbb.$$

Note that, just like Ψ'_k, the k-spectrum grows exponentially with k, it contains $(\#A^{k+1} - 1)/(\#A - 1)$ (possibly zero) coefficients. Let us quote Salomaa: "*a notion often mentioned but not much investigated in the literature,* [9, 69, 71, 88]*, is that of a t-spectrum.*". In particular, in terms of 'reconstruction' what is the relation between n and k such that if two words u, v of length n have the same k-spectrum, then $u = v$?

Remark 49. Two words are k-binomially equivalent if and only if they have the same k-spectrum.

Properties 50. *About avoidance of k-binomial repetitions, see* [82]*: 2-binomial squares (resp. cubes) are avoidable over a 3-letter (resp. 2-letter) alphabet. The sizes of the respective alphabets are optimal.*

Theorem 51. [86, Thm. 7] *If $\mathbf{x}$ is a Sturmian word, then $\mathsf{p}_{\sim_{2-\mathsf{bin},\mathbf{x}}}(n) = n+1$ for all $n \geq 0$. Moreover, we also have $\mathsf{p}_{\sim_{k-\mathsf{bin},\mathbf{x}}}(n) = n + 1$ for all $n \geq 0$ and $k \geq 2$.*

Properties 52. [86, Thm. 17] *If an infinite recurrent word has bounded 2-binomial complexity, then the frequency of each symbol exists and is rational.*

Lemma 53. *Let $k \geq 1$ and $p, q, r, x, y \in A^*$. If $x \sim_{k-\mathsf{bin}} y$, then $pxqyr \sim_{(k+1)-\mathsf{bin}} pyqxr$.*

Proof. If we have to count the number of occurrences of a subword z of length at most $k + 1$, either this subword occurs completely inside one of the factors p, q, r, x, y, or it is obtained from several shorter subwords occurring in at least two of these factors. To get

the conclusion, observe that, by assumption, x, y share the exactly the same subwords of length at most k. □

But it is not clear that a form of converse for this result exists (see Proposition 65 where we have a characterization of equivalent words in terms of this kind of transformations but only for a 2-letter alphabet). Over a 3-letter alphabet: $2100221 \sim_{2-\mathsf{bin}} 0221102$ but 2100221 cannot be factorized into $pxqyr$ with $x \sim_{\mathsf{ab}} y$ and $x \neq y$.

Remark 54. [k-binomial pattern matching] In their very nice paper [43], Freydenberger *et al.* answer positively to the following questions (similar to Remark 41):

- Given $k \geq 1$ and two words u, v of length n, decide, in polynomial time with respect to n and k, whether or not $u \sim_{k-\mathsf{bin}} v$.
- Given $k \geq 1$ and two words w, x, find, in polynomial time, all occurrences of factors of w which are k-binomially equivalent to x.
- Given two u, v of length n, find the largest k such that $u \sim_{k-\mathsf{bin}} v$.

One answer is given by building a non deterministic automaton accepting a language with multiplicities (one counts the number of accepting paths for a given input word) associated with a word w and an integer k. This automaton accepts exactly the subwords of w of length at most k and the number of accepting paths of a subword x is precisely $\binom{w}{x}$. The number of states of this automaton is proportional to $|w| \cdot k$. There exist polynomial time procedures to test the equivalence of two such automata [62, 92, 99] that were initially considered by Schützenberger for the minimization of weighted automata. Another clever answer is a randomized one based on the evaluation of a polynomial, similar to (2), over a sufficiently large finite field equivalent to the k-spectrum (considering evaluation avoids the problem of considering the polynomial as an exponentially growing list of coefficients). Ideas are similar to those found in primality testing algorithms.

Other related relations exist.

Definition 55. The *Simon congruence* $\sim_S$ is defined as follows. We have $u \sim_S v$ if and only if the series $\sum_{w \in A^*} \binom{u}{w} w$ and $\sum_{w \in A^*} \binom{u}{w} w$ have the same support, i.e., they have the same non-zero binomial coefficients.

The latter congruence has applications to piecewise testable[5] languages. About Q.1.1 and counting the number of classes for $\sim_S$, see [54].

Remark 56. Let us mention an extra notion studied by Salomaa in [91]. Let $u = u_1 \cdots u_n$ be a finite word. The *sum of the position indices* for the letter $a \in A$ is defined by

$$S_a(u) := \sum_{i=1}^{|u|} i \, \delta_{u_i,a}.$$

For instance, $S_b(abacbcaba) = 2 + 5 + 8 = 15$. Similarly to binomial coefficients, this type of quantity provides information about the positions of occurrences of letters in a word.

[5]A regular language is *piecewise testable* if it is a finite Boolean combination of languages of the form $A^* a_1 A^* \cdots A^* a_t A^*$ where the a_i's are letters in A.

7. Parikh matrices

Let $k \geq 2$ be an integer. Also related to Theorem 35 and binomial coefficients, one can extend the abelianization map Ψ as follows. Let $\mathbb{N}^{\ell\times\ell}$ be the monoid of $\ell \times \ell$ matrices equipped with the multiplication of matrices. Let $A_k := \{a_1, \ldots, a_k\}$ be an ordered finite alphabet. The *Parikh matrix mapping*

$$\psi_k : A_k^* \to \mathbb{N}^{(k+1)\times(k+1)}$$

is the morphism of monoids defined by the condition: if $\psi_k(a_q) = (m_{i,j})_{1\leq i,j\leq k+1}$, then for each $i \in \{1, \ldots, k+1\}$,

$$m_{i,i} = 1, \quad m_{q,q+1} = 1,$$

all other elements of the matrix $\psi_k(a_q)$ being 0. There are many papers dealing with Parikh matrices, we only refer to a few of them [69, 71, 88, 90, 93].

Definition 57. Two words over A_k are *M-equivalent*, or *matrix equivalent*, if they have the same Parikh matrix. Again, this relation is clearly a congruence because ψ_k is a morphism. If the equivalence class of a word w is reduced to a single element, then w is said to be *M-unambiguous*.

Consider $A = \{a, b\}$ and $a < b$. We have

$$\psi_2(a) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ \psi_2(b) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \psi_2(abbab) = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

The next proposition can be easily deduced from elementary properties of binomial coefficients of words and matrix computations. It shows that Parikh matrices for an alphabet of cardinality k encode $k(k+1)/2$ of the binomial coefficients of a word w for subwords of length at most k. With the above example, the word *abbab* contains 2 a's, 3 b's and 4 occurrences of the subword ab.

Theorem 58. [70] *Let w be a finite word over A_k and $\psi_k(w) = (m_{i,j})_{1\leq i,j\leq k+1}$. Then*

$$m_{i,j+1} = \binom{w}{a_i \cdots a_j}$$

for all $1 \leq i \leq j \leq k$.

Generalized Parikh mappings ψ_u, for all words $u \in A^*$ can be defined as follows. Let $u = u_1 \cdots u_\ell$. If $\psi_u(a) = (m_{i,j})_{1\leq i,j\leq \ell+1}$, then for each $i \in \{1, \ldots, \ell+1\}$, $m_{i,i} = 1$, and for each $i \in \{1, \ldots, \ell\}$,

$$m_{i,i+1} = \delta_{a,u_i},$$

all other elements of the matrix $\psi_u(a)$ being 0.

Remark 59. We get back to the 'classical' Parikh matrices over A_k with $u = a_1 a_2 \cdots a_k$.

Theorem 58 has the following natural generalization.

Theorem 60. [93] *Let $u = u_1 \cdots u_\ell$ and w a word. Let $\psi_u(w) = (m_{i,j})_{1\leq i,j\leq \ell+1}$. Then, for all $1 \leq i \leq j \leq \ell$,*

$$m_{i,j+1} = \binom{w}{u_i \cdots u_j}.$$

In particular, the first row of $\psi_u(w)$ contains the coefficients corresponding to the prefixes of u: $\binom{w}{\varepsilon}$, $\binom{w}{u_1}$, $\binom{w}{u_1u_2}$, $\ldots$, $\binom{w}{u_1\cdots u_{\ell-1}}$, $\binom{w}{u}$. Similarly, the last column of $\psi_u(w)$ contains the coefficients corresponding to the suffixes: $\binom{w}{u}$, $\binom{w}{u_2\cdots u_\ell}$, $\ldots$, $\binom{w}{u_1}$, $\binom{w}{\varepsilon}$.

Example 61. Here is an illustration of the latter theorem:

$$\psi_{abba}(w) = \begin{pmatrix} 1 & \binom{w}{a} & \binom{w}{ab} & \binom{w}{abb} & \binom{w}{abba} \\ 0 & 1 & \binom{w}{b} & \binom{w}{bb} & \binom{w}{bba} \\ 0 & 0 & 1 & \binom{w}{b} & \binom{w}{ba} \\ 0 & 0 & 0 & 1 & \binom{w}{a} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

With ℓ-abelian equivalence and k-binomial equivalence, we had two infinite families of refinements. We can also introduce similar refinements, actually uncountably many families of refinements.

Definition 62. Let $\mathbf{w} = w_0w_1w_2\cdots$ be an infinite word. Considering the prefixes of $\mathbf{w}$, with this infinite word is associated a sequence of maps

$$(\psi_{w_0\cdots w_j})_{j\geq 0}.$$

We say that two finite words u, v are $(\mathbf{w}, j)$-*equivalent*, if

$$\psi_{w_0\cdots w_j}(u) = \psi_{w_0\cdots w_j}(v).$$

This means that u and v have the same binomial coefficients corresponding to the factors occurring in the prefix of length $j + 1$ of $\mathbf{w}$.

In the above definition, if $\mathbf{w}$ contains every letter of the alphabet, taking j large enough such that every letter of A_k appears in $w_0 \cdots w_j$, $(\mathbf{w}, j)$-equivalence is a refinement of the abelian equivalence. Note that $u \sim_{\ell-\mathsf{bin}} v$ trivially implies that u, v are $(\mathbf{w}, \ell-1)$-equivalent. Also, for every word $\mathbf{w}$, $(\mathbf{w}, j + 1)$-equivalence is a refinement of $(\mathbf{w}, j)$-equivalence (the matrix $\psi_{w_0\cdots w_j}(u)$ is the upper-left corner of $\psi_{w_0\cdots w_jw_{j+1}}(u)$). See Figure 1.

Example 63. Let us illustrate the relations existing between binomial equivalence and M-equivalence. Again, these equivalences are, in general, incomparable.

- The two words $u = abcbabcbabcbab$ and $v = bacabbcabbcbba$ are not 3-binomially equivalent: $\binom{u}{abb} = 34$ and $\binom{v}{abb} = 36$ but they share the same Parikh matrix $\psi_3(u) = \psi_3(v)$. This observation only reflects that Parikh matrices encode a fraction of the binomial coefficients. Nevertheless, for a well-chosen generalized Parikh matrix, the two words can of course be distinguished by $\psi_{abb}(u) \neq \psi_{abb}(v)$.

- Erasing the c's in the previous two words, we get two words $u' = abbabbabbab$ and $v' = baabbabbbba$ that are not 3-binomially equivalent: we again have $\binom{u'}{abb} = 34$ and $\binom{v'}{abb} = 36$. But they have the same Parikh matrix, i.e., $\psi_2(u) = \psi_2(v)$ (and from the next proposition, we also have that the words are 2-binomially equivalent). Indeed, 3-binomial equivalence is a strict refinement of the 2-binomial equivalence.
- Finally, the two words $u = bccaa$ and $v = cacab$ are not 2-binomially equivalent: $\binom{u}{ca} = 4$ and $\binom{v}{ca} = 3$, but they share the same Parikh matrix $\psi_3(u) = \psi_3(v)$.

Also, ℓ-abelian equivalence and M-equivalence are incomparable. Take the same two words as in Example 47: *abbaba* and *ababba* are 2-abelian equivalent, but they are not M-equivalent.

Over a 2-letter alphabet, the situation is usually simpler.

Proposition 64. *Over a 2-letter alphabet, two words are 2-binomially equivalent if and only if they have the same Parikh matrix, i.e., are M-equivalent.*

Proof. One direction is obvious. Let the alphabet be $\{a, b\}$. Assume that $\psi_2(u) = \psi_2(v)$. We have

$$\binom{u}{aa} = \binom{|u|_a}{2} = \binom{|v|_a}{2} = \binom{v}{aa}.$$

The same holds for the subword bb. We only have to check that $\binom{u}{ba} = \binom{v}{ba}$. This follows from the fact that, for all words w,

$$\sum_{x \in A^2} \binom{w}{x} = \binom{|w|}{2}.$$

□

In particular, the next result completely characterizes the equivalence classes for 2-binomial equivalence over a 2-letter alphabet.

Theorem 65. [90] *Over a 2-letter alphabet A, two words are M-equivalent if and only if one can be obtain from the other by a finite sequence of transformations of the form $xabybaz \to xbayabz$ where $a, b \in A$ and $x, y, z \in A^*$.*

As a consequence of this result, a word over a 2-letter alphabet is M-unambiguous (there is no other word with the same Parikh matrix) if and only if it belongs to $a^*b^* + b^*a^* + a^*ba^* + b^*ab^* + a^*bab^* + b^*aba^*$.

8. Other relations

To conclude with this survey, let us mention a few other relations that can be encountered in combinatorics on words and formal language theory. Berstel and Boasson initiated the study of *partial words* containing a 'do not know' symbol $\diamond$ serving as a wild card [8]. Two words such as

$$\begin{array}{cccccc} & a & \diamond & b & b & \diamond \\ \text{and} & \diamond & a & b & \diamond & \diamond \end{array}$$

are *compatible* because one can replace the symbols $\diamond$ in such a way that both words match the word *aabba* (or *aabbb*). This relation '*being compatible*' is reflexive and symmetric (but clearly not transitive). Also see, [13, 14, 16, 17, 18]. One can generalize this to a *similarity relation* associated with a binary relation over an alphabet, see the survey chapter by Halava, Harju and Kärki in [12, Chap. 6].

Definition 66. Consider a reflexive and symmetric relation R over an alphabet A. Two words $u_1 \cdots u_n$ and $v_1 \cdots v_n$ are *R-similar*, if $(u_i, v_i) \in R$ for all i. We write $u \sim_R v$. Partial words corresponds to the special case where the relation R is defined by $(a, \diamond) \in R$, for all $a \in A$.

Example 67. Assume that $R = \{(a, b), (b, c), (c, d), (d, a)\}$ and take its symmetric and reflexive closure. As an example, we have the following relations:

$$abcd \sim_R bcbd,\ bcbd \sim_R ccac,\ abcd \not\sim_R ccac.$$

One can also think about relations derived from languages or automata. We just give an example. Let L be a language over A. The *syntactic congruence* is defined as follows. The *context* of a word u is the set of pairs of words (x, y) such that xuy belongs to L. Two words are syntactically congruent if they have the same context. For instance, see [7].

References

[1] A. Aberkane, J.D. Currie, N. Rampersad, The number of ternary words avoiding abelian cubes grows exponentially, *J. Integer Seq.* **7.2** (2004), Art. 04.2.7.

[2] B. Adamczewski, Balances for fixed points of primitive substitutions, *Theoret. Comput. Sci.* **307** (2003), 47–75.

[3] J.-P. Allouche, Sur la complexité des suites infinies, *Bull. Belg. Math. Soc.* **1** (1994), 133–143.

[4] J.-P. Allouche, J. Shallit, Sums of digits, overlaps, and palindromes, *Discrete Math. Theor. Comput. Sci.* **4** (2000), 1–10.

[5] J.-P. Allouche, J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge Univ. Press (2003).

[6] P. Arnoux, G. Rauzy, Représentation géométrique de suites de complexité $2n + 1$, *Bull. Soc. Math. France* **119** (1991), 199–215.

[7] J. Berstel, C. Reutenauer, *Noncommutative rational series with applications*, Encycl. of Math. and its Appl. **137**, Cambridge Univ. Press (2011).

[8] J. Berstel, L. Boasson, Partial words and a theorem of Fine and Wilf, *Theoret. Comput. Sci.* **218** (1999), 135–141.

[9] J. Berstel, J. Karhumäki, Combinatorics on words — a tutorial, *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **79** (2003), 178–228.

[10] J. Berstel, L. Vuillon, Coding rotations on intervals, *Theoret. Comput. Sci.* **281** (2002), 99–107.

[11] V. Berthé, M. Rigo (Eds.), *Combinatorics, Automata and Number Theory*, Encycl. of Math. and its Appl. **135**, Cambridge Univ. Press (2010).

[12] V. Berthé, M. Rigo (Eds.), *Combinatorics, Words and Symbolic Dynamics*, Encycl. of Math. and its Appl. **159**, Cambridge Univ. Press (2016).

[13] F. Blanchet-Sadri, Codes, orderings and partial words, *Theoret. Comput. Sci.* **239** (2004), 177–202.

[14] F. Blanchet-Sadri, Periodicity on partial words, *Comput. Math. Appl.* **47** (2004), 71–82.

[15] F. Blanchet-Sadri, J. D. Currie, N. Rampersad, N. Fox, Abelian complexity of fixed point of morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$, Integers **14** (2014), paper A11.

[16] F. Blanchet-Sadri, R. A. Hegstrom, Partial words and a theorem of Fine and Wilf revisited, *Theoret. Comput. Sci.* **270** (2002), 401–419.
[17] F. Blanchet-Sadri, T. Oey, T. Rankin, Fine and Wilf's theorem for partial words with arbitrarily many weak periods, *Internat. J. Found. Comput. Sci.* **21** (2010), 705–722.
[18] F. Blanchet-Sadri, S. Simmons, A. Tebbe, A. Veprauskas, Abelian periods, partial words, and an extension of a theorem of Fine and Wilf, *RAIRO Theor. Inform. Appl.* **47** (2013), 215–234.
[19] F.-J. Brandenburg, Uniformly growing k-th power-free homomorphisms, *Theoret. Comput. Sci.* **23** (1983), 69–82.
[20] T. C. Brown, Is there a sequence on four symbols in which no two adjacent segments are permutations of one another? *Amer. Math. Monthly* **78** (1971), 886–888.
[21] A. Carpi, On Abelian Power-Free Morphisms, *Int. J. Algebra Comput.* **3** (1993), 151–167.
[22] A. Carpi, On the number of Abelian square-free words on four letters, *Discrete Appl. Math.* **81** (1998), 155–167.
[23] J. Cassaigne, Counting overlap-free binary words, *Lect. Notes in Comp. Sci.* **665** (1993), 216–225.
[24] J. Cassaigne, J.D. Currie, L. Schaeffer, J. Shallit, Avoiding three consecutive blocks of the same size and same sum, *J. ACM* **61** (2014), Art. 10.
[25] J. Cassaigne, G. Richomme, K. Saari, L. Q. Zamboni, Avoiding Abelian powers in binary words with bounded Abelian complexity, *Int. J. Found. Comp. Sci.* **22** (2011), 905–920.
[26] J. Cassaigne, J. Karhumäki, A. Saarela, On growth and fluctuation of k-abelian complexity, *Lect. Notes in Comput. Sci.* **9139** (2015), 109–122.
[27] A. Cobham, Uniform tag sequences, *Math. Systems Theory* **6** (1972), 164–192.
[28] S. Constantinescu, L. Ilie, Fine and Wilf's theorem for abelian periods, *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **89** (2006), 167–170.
[29] E. M. Coven and G. A. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973), 138–153.
[30] J. D. Currie, N. Rampersad, Fixed points avoiding Abelian k-powers, *J. Combin. Theory Ser. A* **119** (2012), 942–948.
[31] J. Currie, N. Rampersad, Growth rate of binary words avoiding xxx^R, *Theoret. Comput. Sci.* **609** (2016), 456–468.
[32] F. M. Dekking, Strongly nonrepetitive sequences and progression-free sets, *J. Combin. Theory Ser. A* **27** (1979), 181–185.
[33] G. Didier, Combinatoire des codages de rotations, *Acta Arith.* **85** (1998), 157–177.
[34] A. W. M. Dress, P. L. Erdős, Reconstructing words from subwords in linear time, *Annals of Combinatorics* **8** (2004), 457–462.
[35] M. Dudik, L. J. Schulman, Reconstruction from subsequences, *J. Combin. Theory, Ser. A* **103** (2003), 337–348.
[36] F. Durand, A characterization of substitutive sequences using return words,*Disc. Math.* **179** (1998), 89–101.
[37] F. Durand, Decidability of the HD0L ultimate periodicity problem, *RAIRO - Theoret. Inf. and Appl.* **47** (2013), 201–214.
[38] T. Ehlers, F. Manea, R. Mercas, D. Nowotka, k-abelian pattern matching, *Lect. Notes Comput. Sci.* **8633** (2014), 178–190.
[39] A. Ehrenfeucht, K. P. Lee, G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interaction, *Theoret. Comput. Sci.* **1** (1975), 59–75.
[40] P. Erdős, Some unsolved problems, *Michigan Math. J.* **4** (1957), 291–300.
[41] R.C Entringer, D.E Jackson, J.A Schatz, On nonrepetitive sequences, *J. Combin. Theory Ser. A* **16**, (1974), 159–164.
[42] A. S. Fraenkel, R. J. Simpson, How Many Squares Must a Binary Sequence Contain?, *The Electronic Journal of Combinatorics*, **2**, (1995).

[43] D. D. Freydenberger, P. Gawrychowski, J. Karhumäki, F. Manea, W. Rytter, Testing k-binomial equivalence, `arXiv:22600.9051`.
[44] F. Greinecker, On the 2-abelian complexity of the Thue-Morse word, *Theoret. Comput. Sci.* **593** (2015), 88–105.
[45] V. Halava, T. Harju, T. Kärki, Relational codes of words, *Theoret. Comput. Sci.* **389** (2007), 237–249.
[46] V. Halava, T. Harju, T. Kärki, The theorem of fine and Wilf for relational periods, *Theor. Inform. Appl.* **43** (2009), 209–220.
[47] V. Halava, T. Harju, T. Kärki, M. Rigo, On the periodicity of morphic words, *Lect. Notes in Comput. Sci.* **6224** (2010), 209–217.
[48] T. Harju, M. Linna, On the periodicity of morphisms on free monoids, *RAIRO Inform. Théor. Appl.* **20** (1986), 47–54.
[49] C. Holton, L. Q. Zamboni, Descendants of primitive substitutions, *Theory Comput. Systems* **32** (1999), 133–157.
[50] M. Huova, Existence of an infinite ternary 64-abelian square-free word, *RAIRO - Theoretical Informatics and Applications* **48** (2014), 307–314.
[51] M. Huova, A. Saarela, Strongly k-abelian repetitions, *Lect. Notes in Comput. Sci.* **8079**, Springer, (2013).
[52] M. Huova, J. Karhumäki, On unavoidability of k-abelian squares in pure morphic words, *J. Integer Seq.* **16** (2013), no. 2, Art. 13.2.9.
[53] L. I. Kalashnik, The reconstruction of a word from fragments, in "Numerical Mathematics and Computer Technology," pp. 56–57, Akad. Nauk Ukrain. SSR Inst. Mat., Preprint IV, (1973).
[54] P. Karandikar, M. Kufleitner, Ph. Schnoebelen. On the index of Simon's congruence for piecewise testability, *Inform. Processing Let.* **15** (2015), 515–519.
[55] J. Karhumäki, Generalized Parikh mappings and homomorphisms, *Inform. and Control* **47** (1980), 155–165.
[56] J. Karhumäki, S. Puzynina, A. Saarela, Fine and Wilf's theorem for k-abelian periods, *Internat. J. Found. Comput. Sci.* **24** (2013), 1135–1152.
[57] J. Karhumäki, A. Saarela, L. Q. Zamboni, On a generalization of Abelian equivalence and complexity of infinite words, *J. Combin. Theory Ser. A* **120** (2013), 2189–2206.
[58] J. Karhumäki, A. Saarela, L. Q. Zamboni, Variations of the Morse-Hedlund theorem for k-abelian equivalence, *Lect. Notes in Comput. Sci.* **8633** (2014), 203–214.
[59] J. Karhumäki, J. Shallit, Polynomial versus Exponential Growth in Repetition-Free Binary Words, *J. Combin. Theory Ser. A* **105** (2004), 335–347.
[60] T. Kärki, Compatibility relations on codes and free monoids, *Theor. Inform. Appl.* **42** (2008), 539–552.
[61] V. Keränen, Abelian squares are avoidable on 4 letters, *Lecture Notes in Comput. Sci.* **623** (1992), 41–52.
[62] S. Kiefer, A. S. Murawski, J. Ouaknine, B. Wachter, J. Worrell, On the complexity of the equivalence problem for probabilistic automata, *Lect. Notes in Comput. Sci.* **7213** (2012), 467–481.
[63] Y. Kobayashi, Enumeration of irreducible binary words, *Disc. Appl. Math.* **20** (1988), 221–232.
[64] I. Krasikov, Y. Roditty, On a Reconstruction Problem for Sequences, *J. Combin. Theory, Ser. A* **77** (1997), 344-348.
[65] J. Leroy, M. Rigo, M. Stipulanti, Generalized Pascal triangle for binomial coefficients of words, *Adv. Appl. Math.* **80** (2016), 24–47.
[66] M. Lothaire, *Combinatorics on Words*, Cambridge Mathematical Library, Cambridge Univ. Press (1997).
[67] M. Lothaire, *Algebraic Combinatorics on Words*, Encycl. of Math. and its Applic. **90**, Cambridge Univ. Press (2002).
[68] B. Madill, N. Rampersad, The abelian complexity of the paperfolding word, *Discrete Math.* **313** (2013), 831–838.

[69] J. Maňuch, Characterization of a word by its subwords, in: G. Rozenberg, W. Thomas (Eds.), Developments in Language Theory, World Scientific Publ. Co., Singapore, 2000, pp. 210–219.
[70] A. Mateescu, A. Salomaa, K. Salomaa, S. Yu, A Sharpening of the Parikh Mapping, *RAIRO-Theoretical Informatics and Applications* **35** (2001), 551–564.
[71] A. Mateescu, A. Salomaa, S. Yu, Subword histories and Parikh matrices, *J. Comput. Systems Sci.* **68** (2004), 1–21.
[72] M. Morse,G. A. Hedlund, Symbolic Dynamics, *Amer. J. Math.* **60** (1938), 815–866.
[73] P. Ochem, N. Rampersad, J. Shallit, Avoiding approximate squares, *Internat. J. Found. Comput. Sci.* **19** (2008), 633–648.
[74] J.-J. Pansiot, Bornes inférieures sur la complexité des facteurs des mots infinis engendrés par morphismes itérés, *Lect. Notes in Comput. Sci.* **166** (1984), 230–240.
[75] J.-J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphismes itérés, *Lect. Notes in Comput. Sci.* **172** (1984), 380–389.
[76] J.-J. Pansiot, Decidability of periodicity for infinite words, *RAIRO Inform. Théor. Appl.* **20** (1986), 43–46.
[77] A. Parreau, M. Rigo, E. Rowland, É. Vandomme, A new approach to the 2-regularity of the ℓ-abelian complexity of 2-automatic sequences, *Electron. J. Combin.* **22** (2015), paper 1.27.
[78] R. Parikh, On Context-Free Languages, *J. of the ACM* **13** (1966).
[79] S. Puzynina, L. Q. Zamboni, Abelian returns in Sturmian words, *J. Combin. Theory Ser. A* **120** (2013), 390–408.
[80] N. Rampersad, M. Rigo, P. Salimov, A note on abelian returns in rotation words, *Theoret. Comput. Sci.* **528** (2014), 101–107.
[81] M. Rao, On some generalizations of abelian power avoidability, *Theoret. Comput. Sci.* **601** (2015), 39–46.
[82] M. Rao, M. Rigo, P. Salimov, Avoiding 2-binomial squares and cubes, *Theoret. Comput. Sci.* **572** (2015), 83–91.
[83] G. Richomme, K. Saari, L.Q. Zamboni, Balance and abelian complexity of the Tribonacci word, *Adv. in Appl. Math.* **45** (2010), 212–231.
[84] G. Richomme, K. Saari, L.Q. Zamboni, Abelian complexity of minimal subshifts, *J. Lond. Math. Soc.* **83** (2011), 79–95.
[85] M. Rigo, *Formal Languages, Automata and Numeration Systems: Introduction to Combinatorics on Words*, ISTE-Wiley (2014).
[86] M. Rigo, P. Salimov, Another generalization of abelian equivalence: binomial complexity of infinite words, *Theoret. Comput. Sci.* **601** (2015), 47–57.
[87] M. Rigo, P. Salimov, E. Vandomme, Some properties of abelian return words, *J. Integer Seq.* **16** (2013), Art. 13.2.5.
[88] A. Salomaa, Counting (scattered) subwords, *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **81** (2003), 165–179.
[89] A. Salomaa, Connections between subwords and certain matrix mappings, *Theoret. Comput. Sci.* **340** (2005), 188–203.
[90] A. Salomaa, Criteria for the matrix equivalence of words, *Theoret. Comput. Sci.* **411** (2010), 1818–1827.
[91] A. Salomaa, Subword balance, position indices and power sums, *J. Comput. Systems Sci.* **76** (2010), 861–871.
[92] M.-P. Schützenberger, On the definition of a family of automata, *Inf. and Control* (1961), 245–270.
[93] T.-F. Şerbănuţă, Extending Parikh matrices, *Theoret. Comput. Sci.* **310** (2004), 23–246.
[94] T. A. Sudkamp, *Languages and Machines: An Introduction to the Theory of Computer Science*, Addison-Wesley Pub. (1997).
[95] A. Thue, Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22.

[96] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67.
[97] O. Turek, Abelian complexity function of the Tribonacci word, *J. Integer Seq.* **18** (2015), Art. 15.3.4.
[98] O. Turek, Abelian complexity and abelian co-decomposition, *Theoret. Comput. Sci.* **469** (2013), 77–91.
[99] W. Tzeng, A polynomial-time algorithm for the equivalence of probabilistic automata, *SIAM J. Comput.* **21** (1992), 216–227.
[100] L. Vuillon, A characterization of Sturmian words by return words, *European J. Combin.* **22** (2001), 263–275.

University of Liege, Dept. of Math., Allée de la découverte 12 (B37), B-4000 Liège, Belgium.
E-mail address: M.Rigo@ulg.ac.be