On the Pseudorandomness of Parry–Bertrand Automatic Sequences

Pierre Popoli[1,†], Manon Stipulanti[1,†]

[1]Departement of Mathematics, ULiège, Liège, Belgium

Abstract

The correlation measure is a testimony of the pseudorandomness of a sequence $\mathbf{s}$ and provides information about the independence of some parts of $\mathbf{s}$ and their shifts. Combined with the well-distribution measure, a sequence possesses good pseudorandomness properties if both measures are relatively small. In combinatorics on words, the famous b-automatic sequences are quite far from being pseudorandom, as they have small factor complexity on the one hand and large well-distribution and correlation measures on the other. This paper investigates the pseudorandomness of a specific family of morphic sequences, including classical b-automatic sequences. In particular, we show that such sequences have large even-order correlation measures; hence, they are not pseudorandom. We also show that even- and odd-order correlation measures behave differently when considering some simple morphic sequences.

Keywords

Combinatorics on words, Correlation measures, Abstract numeration systems, Morphic sequences, Automatic sequences

1. Introduction

Pseudorandom sequences are generated by algorithms and behave similarly to truly random sequences. To study the pseudorandomness of a sequence, a wide variety of measures have been introduced, called *measures of complexity* or *measures of pseudorandomness*. For instance, from a purely theoretical point of view, the famous *Kolmogorov complexity* of a sequence gives the length of the shortest program that generates it [1]; see also [2, 3]. An infinite sequence is said to be *algorithmically random* if its Kolmogorov complexity grows like the length of the sequence itself. It is well known that Kolmogorov complexity is not a computable function; for example, see [4] for a formal proof and history behind.

As a consequence, many other measures of complexity can be used, especially for practical issues. The series of papers [5, 6], initiated by Mauduit and Sárközy in 1997, introduces and studies several such measures. Among the most important ones, the *well-distribution measure* and the *correlation measure* are considered in various contexts. Let $\mathbf{s} = (\mathbf{s}(n))_{n\geq 0}$ be a sequence over the alphabet $\{0, 1\}$. For $a \in \mathbb{Z}$ and $b, M \in \mathbb{N}$, write $U(\mathbf{s}, M, a, b) = \sum_{j=0}^{M-1}(-1)^{\mathbf{s}(a+jb)}$

ICTCS'24: Italian Conference on Theoretical Computer Science, September 11–13, 2024, Torino, Italy

[†]These authors contributed equally.

pierre.popoli@uliege.be (P. Popoli); m.stipulanti@uliege.be (M. Stipulanti)

0000-0002-4243-9180 (P. Popoli); 0000-0002-2805-2465 (M. Stipulanti)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3811/paper210.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

and for $k \in \mathbb{N}_{\geq 1}$, $D = (d_1, \ldots, d_k) \in \mathbb{N}^k$ such that $0 \leq d_1 < d_2 < \cdots < d_k$, write

$$V(\mathbf{s}, M, D) = \sum_{n=0}^{M-1} (-1)^{\mathbf{s}(n+d_1)+\mathbf{s}(n+d_2)\cdots+\mathbf{s}(n+d_k)}.$$

The *Nth well-distribution measure* of the sequence $\mathbf{s}$ is defined by

$$W(\mathbf{s}, N) = \max_{a,b,M} |U(\mathbf{s}, M, a, b)| = \max_{a,b,M} \left| \sum_{j=0}^{M-1} (-1)^{\mathbf{s}(a+jb)} \right|,$$

where the maximum is taken over all $a, b, M \in \mathbb{N}$ such that $0 \leq a \leq a + (M-1)b < N$, and the *$N$th correlation measure of order k* of $\mathbf{s}$ is given by

$$C_k(\mathbf{s}, N) = \max_{M,D} |V(\mathbf{s}, M, D)| = \max_{M,D} \left| \sum_{n=0}^{M-1} (-1)^{\mathbf{s}(n+d_1)+\mathbf{s}(n+d_2)\cdots+\mathbf{s}(n+d_k)} \right|,$$

where the maximum is taken over all vectors $D = (d_1, \ldots, d_k) \in \mathbb{N}^k$ and all integers M such that $0 \leq d_1 < d_2 < \cdots < d_k$ and $M + d_k \leq N$.

It is said that the sequence $\mathbf{s}$ possesses good properties of pseudorandomness if both its measures $W(\mathbf{s}, N)$ and $C_k(\mathbf{s}, N)$ are relatively small compared to N. Indeed, Alon et al. [7], improving a former result of Cassaigne, Mauduit, and Sárközy [8], prove that the expected order of $W(\mathbf{s}, N)$ and $C_k(\mathbf{s}, N)$ when $\mathbf{s}$ is a truly random binary sequence is $\sqrt{N}(\log N)^{O(1)}$. Furthermore, it is crucial to simultaneously consider the two measures, since the smallness of one does not necessarily imply that of the other; see [5] for more details. However, it is sufficient to show that the correlation measure of a sequence is large to tell it apart from a pseudorandom sequence. Notice that the combination of the well-distribution and the correlation measures gives birth to the *well-distribution-correlation measure*, which is however more complicated to study.

Usually, the methods to analyze correlation measures are borrowed from number theory; see [9, 10] and the references therein for general results. However, in this paper, we study a large family of sequences through the lens of their correlation measures by mixing tools from automata theory and general numeration systems. Such a family is that of *automatic sequences*; see Section 2 for precise definitions. In combinatorics on words, they are classical and have many beautiful properties. Among others, they are known to be bad pseudorandom sequences as explained hereafter. First, it is well known that their *factor complexity*, which counts the number of different blocks of each length appearing in the whole sequence, is bounded by a linear function and is, therefore, small. On the other hand, over an alphabet of size ℓ, the nth term of the factor complexity of a random sequence is of order ℓ^n; see [11, Chapter 10] for more details. Second, Mérai and Winterhof [12] prove that the order-2 correlation of an automatic sequence is large, as stated below.

Theorem 1 ([12, Theorem 7]). *Let $b \geq 2$ be an integer and let $\mathbf{s}$ be a b-automatic binary sequence generated by a DFAO with m states. Then $C_2(\mathbf{s}, N) \geq \frac{N}{b(m+1)}$ for all $N \geq b(m+1)$.*

The focus on correlation measures of automatic sequences dates back to the previous century. In the late 1920s, Mahler [13] analyzes the sequence $((-1)^{\mathbf{s_2}(n)})_{n\geq 0}$, where $\mathbf{s_2}\colon \mathbb{N}\to\mathbb{N}$ is the *sum-of-digit function* in base 2. In particular, he considers, for any non-negative integer k, the quantity

$$c_M(k) = \frac{1}{M}V(\mathbf{s_2}, M, D) = \frac{1}{M}\sum_{n=0}^{M-1}(-1)^{\mathbf{s_2}(n)+\mathbf{s_2}(n+k)}$$

where $D=(0,k)$, and shows that the sequence $(c_M(k))_{M\geq 0}$ converges to some real number $c(k)$. It turns out that this quantity is related to the order-2 correlation measure of the 2-automatic Thue–Morse sequence $\mathbf{t}=(\mathbf{t}(n))_{n\geq 0}$ as it is defined by $\mathbf{t}(n)=\mathbf{s_2}(n) \bmod 2$ for all $n\geq 0$. We also refer to [14] for more historical background on $\mathbf{t}$ and specific values of the quantity $c(k)$. Later on, Mauduit and Sárközy [6] prove that $C_2(\mathbf{t},N)\geq \frac{1}{12}N$ for $N\geq 5$. However, when Theorem 1 is applied to $\mathbf{t}$, the lower bound is improved as $C_2(\mathbf{t},N)\geq \frac{1}{6}N$ for $N\geq 6$; see [12] for more details. In a similar context, Baake and Coons [15] recently study the correlations of $\mathbf{t}$ with specific values of the shifted vector D, as well as their means with renormalization techniques. Mazáč [16] carries out an analogous approach for the correlations of the Rudin–Shapiro sequence, another distinguished automatic sequence. We also mention the work of Grant, Shallit, and Stoll [17] (and papers following this trend), where correlations over alphabets of size more than 2 are considered.

It is known that automatic sequences are morphic; see [11, Chapter 6] and Section 2 for precise definitions. In this more general framework, the factor complexity is bounded by a quadratic function, which, in turn, again shows that they are far from being pseudorandom; see [11, Chapter 10]. Consequently, our study is driven by the natural question to know whether the order-2 correlation measure of any morphic sequence is large. As a first step towards answering this question, we presently consider a subfamily of morphic sequences based on Parry–Bertrand numeration systems (see Section 2 for precise definitions). In the theory of numeration systems, they have gained the attention of the scientific community for the nice framework they provide and have especially been used in [18, 19, 20, 21, 22].

Our main result is Theorem 2 below (see again Section 2 for precise definitions). It improves Theorem 1 in at least two ways. First, we consider a much larger class of sequences as classical automatic sequences are U_β-automatic for some integer β. Second, we deal with all even-order correlations, whereas only the case of order 2 is considered in [12]. Notice that we also discuss the case of odd-order correlations for which such a theorem fails to hold; see Section 3.1. Note that, for two maps f,g such that g takes only non-negative values, we write $f\gg g$ (resp., $f\ll g$) if there exists a positive constant c such that $|f(x)|\geq cg(x)$ (resp., $|f(x)|\leq cg(x)$) for all sufficiently large x.

Theorem 2. *Let $k\geq 1$ be an integer, $\beta\in\mathbb{R}_{>1}$ be a Parry number and U_β be the corresponding canonical Parry–Bertrand numeration. For all binary U_β-automatic sequences $\mathbf{s}$, we have $C_{2k}(\mathbf{s},N)\gg N$, where the constant depends on β, k, and the size of the automaton generating $\mathbf{s}$.*

The present paper is organized as follows. In Section 2, we recall the necessary background on combinatorics on words, automatic sequences, and numeration systems. Section 3 is divided into two parts. In Section 3.1, we deal with the case of linearly recurrent sequences. In particular,

we highlight the fact that even- and odd-order correlation measures behave differently, which will imply that there is no hope of extending Theorem 2 to odd-order correlations. In Section 3.2, we show with Theorem 2 that our specific morphic sequences have a large correlation measure of even orders. Finally, Section 4 presents possible future work.

2. Background

2.1. On Combinatorics on Words

As a general reference on words, we point out [23]. An *alphabet* is a finite set of elements called *letters*. A *word* on an alphabet A is a sequence of letters from A. It is either *finite* or *infinite*. In order to distinguish them, infinite words are written in bold. The *length* of a finite word, denoted between vertical bars, is the number of letters it is made of. The *empty word* is the only 0-length word. For all $n \geq 0$, we let A^n denote the set of all length-n words over A. We let A^* denote the set of finite words over A, including the empty word, and A^ω or $A^\mathbb{N}$ that of infinite words over A. The first is equipped with the concatenation of words. Let A and B be finite alphabets. A *morphism* $f\colon A^* \to B^*$ is a map satisfying $f(uv) = f(u)f(v)$ for all $u, v \in A^*$. In particular, $f(\varepsilon) = \varepsilon$. For an integer $\ell \geq 2$, a morphism is *ℓ-uniform* if it maps each letter to a length-ℓ word. A 1-uniform morphism is called a *coding*. An infinite word $\mathbf{x}$ is *morphic* if there exist a morphism $f\colon A^* \to A^*$, a coding $g\colon A^* \to B^*$, and a letter $a \in A$ such that $\mathbf{x} = g(f^\omega(a))$, where $f^\omega(a) = \lim_{n\to\infty} f^n(a)$.

2.2. On Generalized Automatic Sequences

Abstract numeration systems were introduced at the beginning of the century by Lecomte and Rigo [24]; see also [25, Chapter 3] for a general presentation. Such a numeration system is defined by a triple $S = (L, A, <)$ where A is an alphabet ordered by the total order $<$ and L is an infinite *regular* language over A, i.e., accepted by a deterministic finite automaton. We say that L is the *numeration language* of S. When we *genealogically* (i.e., first by length, then using the dictionary order) order the words of L, we obtain a one-to-one correspondence rep_S between $\mathbb{N}$ and L. Then, the *S-representation* of the non-negative integer n is the $(n+1)$st word of L, and the inverse map, called the *(e)valuation map*, is denoted by val_S.

Example 1. Consider the abstract numeration system S built on the language a^*b^* over the ordered alphabet $\{a, b\}$ with $a < b$. The first few words in the language are $\varepsilon, a, b, aa, ab, bb, aaa$. For instance, $\mathrm{rep}_S(5) = bb$ and $\mathrm{val}_S(aaa) = 6$.

Let $S = (L, A, <)$ be an abstract numeration system. An infinite word $\mathbf{x}$ is *S-automatic* if there exists a deterministic finite automaton with output (DFAO) $\mathcal{A}$ such that, for all $n \geq 0$, the nth term $\mathbf{x}(n)$ of $\mathbf{x}$ is given by the output $\mathcal{A}(\mathrm{rep}_S(n))$ of $\mathcal{A}$. See [11, 26, 27] for general references.

2.3. Representation of Real Numbers

We now recall several definitions and results about representations of real numbers; see, for instance, [25, 28]. Let $\beta \in \mathbb{R}_{>1}$ and let $A_\beta = \{0, 1, \ldots, \lceil\beta\rceil - 1\}$. Every real number $x \in [0, 1)$

can be written as a series $x = \sum_{j=1}^{+\infty} c_j \beta^{-j}$, where $c_j \in A_\beta$ for all $j \geq 1$. We let $d_\beta(x)$ denote the *β-expansion* of x obtained greedily. In an analogous way, the *β-expansion* $d_\beta(1)$ of 1 is $d_\beta(1) = \lim_{x \to 1} d_\beta(x)$. We define the *quasi-greedy β-expansion* $d^*_\beta(1)$ of 1 as follows. Write $d_\beta(1) = (t(n))_{n \geq 1}$. If $d_\beta(1) = t(1) \cdots t(m) 0^\omega$ with $t(m) \neq 0$, then $d^*_\beta(1) = (t(1) \cdots t(m-1)(t(m)-1))^\omega$; otherwise $d^*_\beta(1) = d_\beta(1)$. We define the *$\beta$-shift* S_β as the topological closure of the set $D_\beta = \{d_\beta(x) \mid x \in [0,1)\}$. A real number $\beta > 1$ is a *Parry number* if $d_\beta(1)$ is ultimately periodic.

Example 2. Any integer $\beta > 1$ is a Parry number. The golden ratio φ is a Parry number for which $d_\varphi(1) = 11$ and $d^*_\varphi(1) = (10)^\omega$ (since $1 = 1/\varphi + 1/\varphi^2$). The square φ^2 of the golden ratio is also a Parry number with $d_{\varphi^2}(1) = 21^\omega = d^*_{\varphi^2}(1)$.

Parry's theorem [29] notably provides a purely combinatorial condition to check if an infinite sequence is in S_β by comparing this sequence to $d^*_\beta(1)$ using the lexicographic order.

Proposition 1 ([28, Corollaries 2.71 and 2.72])**.** *Let $\beta \in \mathbb{R}_{>1}$ be a Parry number. Write $d_\beta(1) = t(1) \cdots t(m)(t(m+1) \cdots t(m+k))^\omega$ (where $m, k \geq 0$ are taken to be minimal). Then an infinite word belongs to S_β if and only if it is the label of a path in the automaton $\mathcal{A}_\beta = (Q, a_0, A_\beta, \delta, Q)$, where $Q = \{a_0, \ldots, a_{m+k-1}\}$ and where the transition function δ is defined as*

$$\begin{aligned}
&\delta(a_{i-1}, t) = a_0, \text{ for all } 1 \leq i < m+k \text{ and all } 0 \leq t < t(i);\\
&\delta(a_{i-1}, t(i)) = a_i, \text{ for all } 1 \leq i < m+k;\\
&\delta(a_{m+k-1}, t(m+k)) = a_m.
\end{aligned}$$

Every Parry number is canonically associated with a linear positional numeration system. We now expose this framework. Let $U = (U(n))_{n \geq 0}$ be an increasing sequence of integers with $U(0) = 1$. Any integer n can be decomposed uniquely in a greedy way as $n = \sum_{i=0}^{t} c_i U(i)$ with non-negative integer coefficients c_i. The finite word $\mathrm{rep}_U(n) = c_t \cdots c_0$ is called the *(greedy) U-representation* of n. By convention, the representation of 0 is the empty word ε. The finiteness of the *digit-set* A_U for greedy representations is implied by the boundedness of the sequence $\sup_{i \geq 0} U(i+1)/U(i)$. A sequence U satisfying all the above conditions defines a *positional numeration systems*. The *numeration language* is the set $L_U = \{\mathrm{rep}_U(n) \mid n \geq 0\}$. For any $c_t \cdots c_0 \in \mathbb{N}^*$, we let $\mathrm{val}_U(c_t \cdots c_0)$ denote the integer $\sum_{i=0}^{t} c_i U(i)$. In addition, a positional numeration system is *linear* if U satisfies a linear recurrence relation.

Example 3. The Fibonacci sequence $F = (F(n))_{n \geq 0}$ with initial conditions $F(0) = 1$ and $F(1) = 2$ and $F(n+2) = F(n+1) + F(n)$ for all $n \geq 0$ gives rise to the famous *Zeckendorf* or *Fibonacci* linear positional numeration system [30]. The numeration language is $L_F = \{\varepsilon\} \cup 1\{0, 01\}^*$.

Let $\beta \in \mathbb{R}_{>1}$ be a Parry number. We define a particular linear positional numeration system $U_\beta = (U_\beta(n))_{n \geq 0}$ associated with β as follows. Write $d_\beta(1) = t(1) \cdots t(m)(t(m+1) \cdots t(m+k))^\omega$ (m, k are minimal and might be zero). We define $U_\beta(0) = 1$, $U_\beta(i) = t(1)U_\beta(i-1) + \cdots + t(i)U_\beta(0) + 1$ for all $i \in \{1, \ldots, m+k-1\}$ and

$$U_\beta(n) = t(1)U_\beta(n-1) + \cdots + t(m+k)U_\beta(n-m-k) + U_\beta(n-k)$$

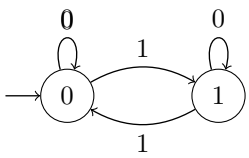

Figure 1: A DFAO producing the sum-of-digit function $\mathbf{s}_F$ modulo 2 in the Zeckendorf numeration system F.

$$-t(1)U_\beta(n-k-1)-\cdots-t(m)U_\beta(n-m-k)$$

for all $n \geq m+k$. We set $L_\beta = L_{U_\beta}$ for the sake of readability.

Example 4. We resume Example 2. For the golden ratio, we obtain $m = 2$, so $U_\varphi(0) = 1$, $U_\varphi(1) = 1 \cdot U_\varphi(0) + 1 = 2$ and $U_\varphi(n) = 1 \cdot U_\varphi(n-1) + 1 \cdot U_\varphi(n-2)$ for all $n \geq 2$. We find back the Zeckendorf numeration system of Example 3. For its square, we have $m = 1 = k$, so $U_{\varphi^2}(0) = 1$, $U_{\varphi^2}(1) = 2 \cdot U_{\varphi^2}(0) + 1 = 3$ and $U_{\varphi^2}(n) = 3 \cdot U_{\varphi^2}(n-1) - 1 \cdot U_{\varphi^2}(n-2)$ for all $n \geq 2$. Its first few elements are thus $1, 3, 8, 21, 55$, the even-indexed Fibonacci numbers.

The numeration system U_β has two useful properties. First, it possesses the *Bertrand property*, i.e., for all $w \in A_{U_\beta}^+$, $w \in L_\beta \Leftrightarrow w0 \in L_\beta$; see [31] for the original version and [18] for the recently corrected one. In that case, any word w in the set 0^*L_β of all U_β-representations with leading zeroes is the label of a path in the automaton $\mathcal{A}_\beta$ from Proposition 1 [18, Theorem 2]. Second, it satisfies the *dominant root condition*, i.e., there is some complex number c such that

$$\lim_{n\to+\infty} \frac{U_\beta(n)}{\beta^n} = c; \tag{1}$$

see, for instance, [18, Theorem 2]. In the following, we will consider *Parry–Bertrand automatic* sequences, which are defined as S-automatic for the abstract numeration system built on the language L_β of U_β for some Parry number β. For the sake of conciseness, we call them *U_β-automatic*. Notice that when β is an integer b, the corresponding sequences are usually called *b-automatic*.

Example 5. In the Zeckendorf numeration system F from Example 3, if we write $\mathrm{rep}_F(n) = n_\ell \cdots n_0$ for each non-negative integer n, then the *sum-of-digit function* $\mathbf{s}_F \colon \mathbb{N} \to \mathbb{N}$ is defined by $\mathbf{s}_F(n) = \sum_{i=0}^{\ell} n_i$. When modded out by 2, it is F-automatic (or Fibonacci-automatic) as the DFAO in Fig. 1 produces it. Note that it is [32, A095076].

3. Main Results

As stated in the introduction, the main purpose of this paper resides in Theorem 2, which extends Theorem 1 to a larger family of morphic sequences by considering all even-order correlations. Such a result will also imply that this family of morphic sequences cannot be qualified as pseudorandom, such as classical automatic sequences, from the point of view of correlations.

3.1. Linearly Recurrent Sequences

Before developing the arguments for Theorem 2, we first deal with so-called linearly recurrent sequences. A sequence $\mathbf{s}$ is *recurrent* if any factor of $\mathbf{s}$ occurs infinitely often in $\mathbf{s}$. It is moreover *uniformly recurrent* if the distance between two consecutive occurrences of any factor x in $\mathbf{s}$ is bounded by a constant that depends only on x. For instance, the Thue-Morse sequence is uniformly recurrent. We say that $\mathbf{s}$ is *linearly recurrent* if the gap between two consecutive occurrences of any length-n factor x in $\mathbf{s}$ is bounded by $C \cdot n$, where C is a constant. See [33] for more examples.

Theorem 3. *Let $\mathbf{s}$ be a linearly recurrent sequence. For all $k \geq 1$, we have $C_{2k}(\mathbf{s}, N) \gg N$.*

Proof. Let $C > 0$ be such that two consecutive occurrences of any factor x of $\mathbf{s}$ being at positions i, j satisfies $|j - i| < C|x|$. Let $N \geq 1$ be sufficiently large and consider the prefix $x = s(0)s(1)\dots s(\lfloor \frac{N}{C+1} \rfloor - 1)$ of $\mathbf{s}$. By assumption, there exists $d \leq C\lfloor \frac{N}{C+1} \rfloor$ such that $s(i) = s(d+i)$ for all $i \in \{0, \dots, \lfloor \frac{N}{C+1} \rfloor - 1\}$ and $d + \lfloor \frac{N}{C+1} \rfloor - 1 \leq N$. Therefore, for $D = (0, d)$ and $M = \lfloor \frac{N}{C+1} \rfloor$, we have $\sum_{n=0}^{M-1} (-1)^{s(n)+s(n+d)} = \sum_{n=0}^{M-1} (-1)^{2s(n)} = M$. This proves that $C_2(\mathbf{s}, N) \gg N$.

The case of even-order of correlations can be treated similarly by considering $2k$ consecutive occurrences of the prefix $x = s(0)s(1)\dots s(\frac{1}{2k}\lfloor \frac{N}{C+1} \rfloor - 1)$, $d \leq \frac{C}{2k}\lfloor \frac{N}{C+1} \rfloor$ such that $s(i) = s(jd+i)$ for all $j \geq 0$ and $i \in \{0, \dots, \lfloor \frac{N}{C+1} \rfloor - 1\}$, and $D = (0, d, \dots, (2k-1)d)$. □

Remark 1. Surprisingly, the case of odd-order correlations cannot be treated similarly. To highlight this, we consider the case of *periodic* sequences, defined as the infinite repetition $u^\omega = uuu\cdots$ of a (finite) pattern u and which are automatic in any numeration system. More specifically, consider the simple periodic sequence $\mathbf{s} = (01)^\omega$. For all $k \geq 1$ and all $N \geq 1$, we have $C_{2k+1}(\mathbf{s}, N) = 1$ since, for $D = (d_1, \dots, d_k)$, $|V(\mathbf{s}, M, D)|$ equals 1 if M is odd, 0 otherwise. Actually, this simple observation explains why the statement of our main result does not hold for odd-order correlations.

Note that there exist automatic sequences that are not uniformly recurrent (and therefore not linearly recurrent). For instance, consider the 3-automatic sequence $\mathbf{ca} = ababbbababbbbbbbbbabab\cdots$ defined as the fixed point of the uniform morphism $a \mapsto aba, b \mapsto bbb$ (sometimes referred to as the *Cantor sequence*; see, for instance, [34, Section 2.4.6]) and for which consecutive occurrences of the factor aba are separated by larger and larger gaps. Therefore, Theorem 2 covers a larger class of sequences than Theorem 3.

3.2. Proof of Theorem 2

Now we attack the proof of Theorem 2 for which we need some additional preparation. First, we count the number of words of each length in the numeration language with leading zeroes, since this language will appear later on.

Lemma 1. *Let $\beta \in \mathbb{R}_{>1}$ be a Parry number and consider the canonical Parry–Bertrand numeration, U_β whose numeration language is L_β. For all $n \geq 0$, the number of length-n words in $0^* L_\beta$ is $U_\beta(n)$.*

Proof. First, for all $m \geq 1$, the length-m words in L_β correspond to the integers in $[U_\beta(m-1), U_\beta(m))$, so there are $U_\beta(m) - U_\beta(m-1)$ of them. The number of length-n words in 0^*L_β is thus equal to

$$\left(\sum_{m=1}^{n-1} U_\beta(m) - U_\beta(m-1)\right) + 1 = U_\beta(n).$$

Therefore, the lemma is proved. □

Since automata producing Parry–Bertrand automatic sequences recognize the language L_β, we build up one from that of Proposition 1.

Lemma 2. *Let $\beta \in \mathbb{R}_{>1}$ be a Parry number. Then there exists a deterministic finite automaton recognizing the language L_β.*

Proof. Let $\mathcal{A}_\beta = (Q, a_0, A_\beta, \delta, Q)$ be the DFA recognizing the language 0^*L_β from Proposition 1 with state-set $Q = \{a_0, \ldots, a_{m+k-1}\}$. The automaton $\mathcal{A}'_\beta = (Q', a'_0, A_\beta, \delta', Q')$ defined as follows recognizes L_β. We set $Q' = \{a'_0\} \cup Q$ where a'_0 is a new state. We replace the initial state a_0 of $\mathcal{A}_\beta$ by a'_0. The new transition function δ' is defined as $\delta'(a'_0, c) = \delta(a_0, c)$ for all non-zero letter c, and $\delta'(a_i, c) = \delta(a_i, c)$ for all $i \in \{0, \ldots, m+k-1\}$ and all letter c.

□

It is well known due to Cobham that b-automatic sequences are characterized by b-uniform morphisms; see [35] or [11, Theorem 6.3.2]. We have the following generalization for S-automatic sequences.

Theorem 4 ([26])**.** *A word is morphic if and only if it is S-automatic for some abstract numeration system S.*

The proof of the above theorem is constructive: given the morphisms producing the word $\mathbf{s}$, one can build an abstract numeration system S and a DFAO generating $\mathbf{s}$, and vice versa. As our proof of Theorem 2 relies on this construction, we develop it now.

Fix a Parry number $\beta \in \mathbb{R}_{>1}$ and consider the canonical Parry–Bertrand numeration U_β. For the sake of conciseness, let L be the numeration language L_β and let A denote the (minimal) alphabet over which L is defined. Order the letters of A as $0 < 1 < \cdots < m$. Consider also a binary U_β-automatic sequence $\mathbf{s}$. Let $\mathcal{A} = (Q, a_0, A, \delta_\mathcal{A}, F)$ be the DFA accepting L as in Lemma 2 and let $\mathcal{B} = (R, r_0, A, \delta_\mathcal{B}, \tau\colon R \to \{0,1\})$ be a DFAO generating the sequence $\mathbf{s}$.

We explicitly give a pair of morphisms producing $\mathbf{s}$. Define the Cartesian product automaton $\mathcal{P} = \mathcal{A} \times \mathcal{B}$, which imitates the behavior of $\mathcal{A}$ on the first component and $\mathcal{B}$ on the second. Its set of states is $Q \times R$, the initial state is (a_0, r_0), and the alphabet is A. Its transition function $\Delta\colon (Q \times R) \times A^* \to Q \times R$ is defined by $\Delta((q,r), w) = (\delta_\mathcal{A}(q,w), \delta_\mathcal{B}(r,w))$ for all $q \in Q$, $r \in R$ and $w \in A^*$. In particular, $\Delta((q,r), w) \in F \times R$ if and only if w belongs to L. Furthermore, if $\Delta((a_0, r_0), \mathrm{rep}_{U_\beta}(n)) = (q, r)$, then $\mathbf{s}(n) = \tau(r)$. Now let α be a symbol not belonging to $Q \times R$. Define the morphism $\varphi_\mathcal{P}\colon (Q \times R \cup \{\alpha\})^* \to (Q \times R \cup \{\alpha\})^*$ by $\varphi_\mathcal{P}(\alpha) = \alpha(a_0, r_0)$ and, for all $q \in Q$ and $r \in R$,

$$\varphi_\mathcal{P}((q,r)) = \Delta((q,r), 0) \cdots \Delta((q,r), m) = \prod_{i=0}^{m} \Delta((q,r), i),$$

where, if $\Delta((q,r),i)$ is not defined for some i, then it is replaced by ε (note that the product has to be understood as the concatenation). Observe that $\varphi_{\mathcal{P}}$ is not necessarily uniform.

Order the words in L genealogically as $w_0 < w_1 < \cdots$. Define $\mathbf{u}_{\mathcal{P}} = \varphi_{\mathcal{P}}^{\omega}(\alpha)$. Then, by [28, Lemma 2.25], the shifted sequence of $\mathbf{u}_{\mathcal{P}}$ is the sequence of states reached in $\mathcal{P}$ by the words in L ordered genealogically, i.e., for all $n \in \mathbb{N}$, the $(n+1)$st letter of $\mathbf{u}_{\mathcal{P}}$ is equal to $\Delta((a_0,r_0),w_n)$. Now define the morphism $\nu_{\mathcal{P}}\colon (Q \times R \cup \{\alpha\})^* \to \{0,1\}^*$ by $\nu_{\mathcal{P}}(\alpha) = \varepsilon$ and, for all $q \in Q$ and $r \in R$, $\nu_{\mathcal{P}}((q,r)) = \tau(r)$. Note that $\nu_{\mathcal{P}}$ is non-erasing on $Q \times R$. By construction, we have $\mathbf{s} = \nu_{\mathcal{P}}(\mathbf{u}_{\mathcal{P}}) = \nu_{\mathcal{P}}(\varphi_{\mathcal{P}}^{\omega}(\alpha))$. To help the reader with the construction, we illustrate it in Example 6 below.

We now give the proof of Theorem 2. In particular, one side feature of the proof is a characterization of the distance between the repeating part in the sequence under study.

Proof of Theorem 2. Let L denote the numeration language L_β and let A be the (minimal) alphabet over which L is defined. Let $\mathcal{A} = (Q, a_0, A, \delta_{\mathcal{A}}, F)$ be the DFA accepting L as in Lemma 2, $\mathcal{B} = (R, r_0, A, \delta_{\mathcal{B}}, \tau\colon R \to \{0,1\})$ be a DFAO generating the sequence $\mathbf{s}$, $\mathcal{P} = \mathcal{A} \times \mathcal{B}$ be the Cartesian product automaton, with transition function $\Delta\colon (Q \times R) \times A^* \to Q \times R$, and α be a symbol not belonging to $Q \times R$. Consider the morphisms $\varphi_{\mathcal{P}}\colon (Q \times R \cup \{\alpha\})^* \to (Q \times R \cup \{\alpha\})^*$ and $\nu_{\mathcal{P}}\colon (Q \times R \cup \{\alpha\})^* \to \{0,1\}^*$ as defined above and write $\mathbf{u}_{\mathcal{P}} = \varphi_{\mathcal{P}}^{\omega}(\alpha)$.

Let $N, M \geq 1$ where M is chosen later accordingly to N. Consider two distinct words u, v such that $\Delta((a_0,r_0),u) = \Delta((a_0,r_0),v)$. Then, for all length-$M$ words $w \in 0^*L$, the states $\Delta((a_0,r_0),uw)$ and $\Delta((a_0,r_0),vw)$ are equal. The word $\varphi_{\mathcal{P}}^M(\Delta((a_0,r_0),u)) = \varphi_{\mathcal{P}}^M(\Delta((a_0,r_0),v))$ is made of the sequence of such states. Indeed, for $x \in \{u,v\}$, we have

$$\varphi_{\mathcal{P}}^M(\Delta((a_0,r_0),x)) = \prod_{\substack{w \in 0^*L \\ |w|=M}} \Delta((a_0,r_0),xw).$$

Applying the morphism $\nu_{\mathcal{P}}$ then gives back the original sequence $\mathbf{s}$. As $\nu_{\mathcal{P}}$ is non-erasing on $Q \times R$, $|\varphi_{\mathcal{P}}^M(\Delta((a_0,r_0),x))| = U_\beta(M)$ by Lemma 1.

Order 2. Firstly, we study the case of the order-2 correlation, for the sake of simplicity and since the case of even-order correlations will follow the same lines. Consider two distinct words $u, v \in L$ such that $\Delta((a_0,r_0),u) = \Delta((a_0,r_0),v)$. By the pigeonhole principle, we may choose u, v among the first $|Q| \cdot |R|$ words of L. Without loss of generality, we may assume that u is lexicographically smaller than v and that u is non-empty (recall Lemma 2). Then in the sequence $\mathbf{s}$ we have two identical blocks of length $U_\beta(M)$ starting at positions p_u, p_v respectively associated with the word $u0^M$ and $v0^M$ in L, i.e., $p_x = \mathrm{val}_{U_\beta}(x0^M)$ for $x \in \{u,v\}$. The situation is depicted in Fig. 2. Therefore, we have

$$\mathbf{s}(p_u + n) = \mathbf{s}(p_v + n) \tag{2}$$

for all $0 \leq n < U_\beta(M)$. We now make several observations. First, notice that the largest index in Eq. (2) is smaller than $p_v + U_\beta(M)$ by assumption on u, v. Then, note that, for $x \in \{u,v\}$, we have $p_x = \mathrm{val}_{U_\beta}(x0^M)$ and $\mathrm{val}_{U_\beta}(x) \leq |Q| \cdot |R|$ by choice of x.

Let $N \geq 1$ be sufficiently large. Choose $M \geq 1$ such that $p_v + U_\beta(M) \leq N < p_v + U_\beta(M+1)$

and set $D = (p_u, p_v)$. Therefore, using Eq. (2), we have

$$C_2(\mathbf{s}, N) \geq |V(\mathbf{s}, N, U_\beta(M), D)| = \sum_{n=0}^{U_\beta(M)-1} (-1)^{\mathbf{s}(p_u+n)+\mathbf{s}(p_v+n)} = U_\beta(M). \tag{3}$$

On the other hand, we have $U_\beta(M) = \Theta(\beta^M)$ by Eq. (1), which yields

$$p_v + U_\beta(M+1) \leq \lceil \beta \rceil^{M+1} \cdot |Q| \cdot |R| + U_\beta(M+1) \ll c\beta^{M+1} \tag{4}$$

for some non-zero constant c depending on β and the sequence $\mathbf{s}$. Then Eqs. (3) and (4) lead to

$$C_2(\mathbf{s}, N) \geq U_\beta(M) \gg \beta^M \gg \frac{p_v + U_\beta(M+1)}{c\beta} \gg N$$

by choice of M. This concludes the case of the order-2 correlation.

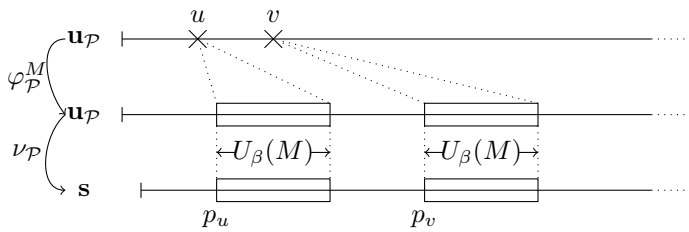

Figure 2: A representation of the situation depicted in the proof of Theorem 2.

Even orders. The general case of even-order correlations can be treated similarly. Indeed, let $k \geq 1$. Consider $2k$ pairwise distinct words $u_1, \ldots, u_{2k} \in L$ such that $\Delta((a_0, r_0), u_i) = \Delta((a_0, r_0), u_j)$ for all $i, j \in \{1, \ldots, 2k\}$. By the pigeonhole principle, we may choose them among the first $(2k-1) \cdot |Q| \cdot |R|$ words of L. Without loss of generality, we may assume that they are increasingly nested lexicographically speaking (and non-empty). Then in the sequence $\mathbf{s}$ we have $2k$ identical blocks of length $U_\beta(M)$ starting at positions $p_{u_i} = \mathrm{val}_{U_\beta}(u_i 0^M)$ for $i \in \{1, \ldots, 2k\}$. Therefore, we have $\mathbf{s}(p_{u_i} + n) = \mathbf{s}(p_{u_j} + n)$ for all $1 \leq i \leq j \leq 2k$ and $0 \leq n < U_\beta(M)$. We now finish up the proof by following the same argument as in the previous case.

□

4. Discussion and Future Work

We can actually obtain a lower bound on correlations using the factor complexity, i.e., the map counting the number of different factors of each length within a given sequence. Consider a sequence $\mathbf{s}$ that is not ultimately periodic. Let $p_\mathbf{s} \colon \mathbb{N} \to \mathbb{N}$ denote its factor complexity. Let k, N be integers with $k \geq 1$. Then, using Morse and Hedlund's famous result [36], one can show that, within the prefix of length $kp_\mathbf{s}(N)$ of $\mathbf{s}$, there exists a length-N factor that occurs at least k times. Therefore, we obtain the lower bound $C_{2k}(\mathbf{s}, 2kp_\mathbf{s}(N)) \geq N$. In particular, for Parry–Bertrand automatic sequences, [20, Theorem 3.4] allows to obtain the behavior emphasized in Theorem 2.

However, the aforementioned lower bound does not reflect the expected behavior of C_k when the studied sequence is random, or in some other simple cases. For instance, for the binary Champernowne sequence $\mathbf{ch} = 011011100101110111\cdots$, we have $C_2(\mathbf{ch}, N) > \frac{1}{48}N$ (see [6, Theorem 1]), whereas $p_{\mathbf{ch}}(N) = 2^N$ for all sufficiently large N.

Furthermore, Theorem 2 extends to numeration systems that are not Parry-Bertrand. In fact, scrutinizing its proof, we need a good control on the proportion of letters erased under the morphism $\nu_{\mathcal{P}}$. Indeed, the proportion of non-erased letters is precisely the length of the block that is represented in Fig. 2. So, in turn, a good knowledge on the growth rate of the numeration numeration language, i.e., the number of words of each length within the language, is required. In the next example, using the method presented in the proof of Theorem 2, we exhibit an automatic sequence having a quadratic factor complexity but a linear order-2 correlation.

Example 6. Consider the linear positional numeration system $U = (U(n))_{n\geq 0}$ defined by $U(0) = 1$ and $U(n+1) = 3U(n) + 1$ for all $n \geq 0$. It has several useful properties: its numeration language is given by the DFA in Fig. 3 and U is Bertrand but not Parry; see [20, Example 2 and Lemma 2.5], as well as [37, page 131]. We also have $U(n) = \Theta(\beta^n)$ where $\beta = \frac{1}{2}(3+\sqrt{13}) \approx 3.30278$. Consider the fixed point $\mathbf{x}$ of the morphism $a \mapsto aaab$, $b \mapsto b$ and the coding $\tau\colon a \mapsto 0, b \mapsto 1$. The sequence $\tau(\mathbf{x})$ is produced by the DFAO in Fig. 3. It is binary and U-automatic. In addition, its factor complexity is known to be quadratic (see [20, Theorem 3.3]).

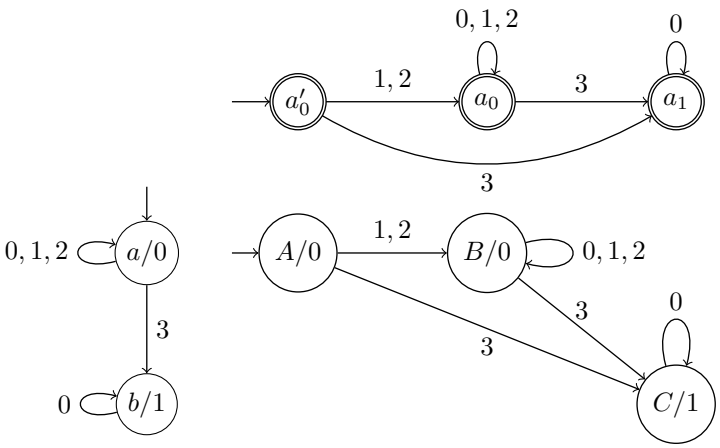

Figure 3: The product $\mathcal{P}$ between a DFA for L_U (top) and a DFAO for the morphic sequence $\tau(\mathbf{x})$ (right) where U, $\mathbf{x}$, and τ are all defined in Example 6.

Using the notation of Section 3.2, the morphism $\varphi_{\mathcal{P}}$ is defined by $\alpha \mapsto \alpha A$, $A \mapsto BBC$, $B \mapsto BBBC$, $C \mapsto C$. Note that it is not uniform. We get $\varphi_{\mathcal{P}}^{\omega}(\alpha) = \alpha ABBCBBBCBBBCCBBBC\cdots$, whose shift indeed corresponds to the sequence $A, B, B, C, B, B, B, C, \ldots$ of states reached by the words of $L_U = \{\varepsilon, 1, 2, 3, 10, 11, 12, 13, 20, 21, 22, 23, 30, 100, 101, 102, \ldots\}$ in the product automaton $\mathcal{P}$ of Fig. 3. The morphism $\nu_{\mathcal{P}}$ is defined by $\alpha \mapsto \varepsilon$, $A, B \mapsto 0$, and $C \mapsto 1$. Applying $\nu_{\mathcal{P}}$ to $\varphi_{\mathcal{P}}^{\omega}(\alpha)$ gives the sequence $\tau(\mathbf{x})$. To illustrate the method of the proof of Theorem 2, we can

consider $u = 10$ and $v = 100$ for which $\delta(A, u) = B = \delta(A, v)$. Now we also conclude that $C_2(\tau(\mathbf{x}), N) \gg N$.

As a consequence, we raise the following more general and natural question: Is any morphic sequence not pseudorandom as far as the correlation measures is concerned, i.e., is the correlation of any morphic sequence large or small? In the near future, we hope to revisit the method of the proof of Theorem 2 to extend it to a larger class of morphic sequences. In particular, as a first step, we will need to find the appropriate conditions to describe this particular class. On the other hand, as shown in [12], a lower bound on the state complexity of binary sequences in terms of the order-2 correlation measure can be straightforwardly derived from Theorem 1. In our setting, can a refinement of the proof of Theorem 2, involving only the number of states of some automata, lead to similar bounds?

Acknowledgments

We thank the reviewers for providing useful remarks that improved our initial work, especially for thinking of the alternative proof mentioned at the beginning of Section 4, which encourages us to look for extensions of our Theorem 2 to a broader family of sequences. We also thank France Gheeraert and Michel Rigo for insightful discussions.

Pierre Popoli's research is supported by ULiège's Special Funds for Research, IPD-STEMA Program. Manon Stipulanti is an FNRS Research Associate supported by the Research grant 1.C.104.24F.

References

[1] A. N. Kolmogorov, Three approaches to the quantitative definition of information, Internat. J. Comput. Math. 2 (1968) 157–168. doi:10.1080/00207166808803030.

[2] G. J. Chaitin, On the length of programs for computing finite binary sequences, J. Assoc. Comput. Mach. 13 (1966) 547–569. doi:10.1145/321356.321363.

[3] R. J. Solomonoff, A formal theory of inductive inference. I, Information and Control 7 (1964) 1–22.

[4] P. M. Vitányi, How incomputable is kolmogorov complexity?, Entropy 22 (2020). doi:10.3390/e22040408.

[5] C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol, Acta Arith. 82 (1997) 365–377. doi:10.4064/aa-82-4-365-377.

[6] C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences. II. The Champernowne, Rudin-Shapiro, and Thue-Morse sequences, a further construction, J. Number Theory 73 (1998) 256–276. doi:10.1006/jnth.1998.2286.

[7] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira, V. Rödl, Measures of pseudorandomness for finite sequences: typical values, Proc. Lond. Math. Soc. (3) 95 (2007) 778–812. doi:10.1112/plms/pdm027.

[8] J. Cassaigne, C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences. VII. The measures of pseudorandomness, Acta Arith. 103 (2002) 97–118. doi:10.4064/aa103-2-1.

[9] K. Gyarmati, On a pseudorandom property of binary sequences, The Ramanujan Journal 8 (2004) 289–302. doi:10.1007/s11139-004-0139-z.
[10] K. Gyarmati, C. Mauduit, On the correlation of binary sequences, ii, Discrete Mathematics 312 (2012) 811–818. doi:https://doi.org/10.1016/j.disc.2011.09.013.
[11] J.-P. Allouche, J. Shallit, Automatic sequences: Theory, applications, generalizations, Cambridge University Press, Cambridge, 2003.
[12] L. Mérai, A. Winterhof, On the pseudorandomness of automatic sequences, Cryptogr. Commun. 10 (2018) 1013–1022. doi:10.1007/s12095-017-0260-7.
[13] K. Mahler, The spectrum of an array and its application to the study of the translation properties of a simple class of arithmetical functions: Part two on the translation properties of a simple class of arithmetical functions, Journal of Mathematics and Physics 6 (1927) 158–163. doi:https://doi.org/10.1002/sapm192761158.
[14] C. Mauduit, Multiplicative properties of the Thue-Morse sequence, Period. Math. Hungar. 43 (2001) 137–153. doi:10.1023/A:1015241900975.
[15] M. Baake, M. Coons, Correlations of the Thue—Morse sequence, Indagationes Mathematicae (2023). doi:https://doi.org/10.1016/j.indag.2023.02.001.
[16] J. Mazáč, Correlation functions of the Rudin–Shapiro sequence, Indagationes Mathematicae (2023). doi:https://doi.org/10.1016/j.indag.2023.03.003.
[17] E. Grant, J. Shallit, T. Stoll, Bounds for the discrete correlation of infinite sequences on k symbols and generalized Rudin-Shapiro sequences, Acta Arith. 140 (2009) 345–368. doi:10.4064/aa140-4-5.
[18] E. Charlier, C. Cisternino, M. Stipulanti, Robustness of Pisot-regular sequences, Adv. in Appl. Math. 125 (2021) Paper No. 102151, 22. doi:10.1016/j.aam.2020.102151.
[19] E. Charlier, C. Cisternino, M. Stipulanti, A full characterization of Bertrand numeration systems, in: Developments in language theory, volume 13257 of *Lecture Notes in Comput. Sci.*, Springer, Cham, 2022, pp. 102–114. doi:10.1007/978-3-031-05578-2_8.
[20] A. Massuir, J. Peltomäki, M. Rigo, Automatic sequences based on Parry or Bertrand numeration systems, Adv. in Appl. Math. 108 (2019) 11–30. doi:10.1016/j.aam.2019.03.003.
[21] F. Point, On decidable extensions of Presburger arithmetic: from A. Bertrand numeration systems to Pisot numbers, J. Symbolic Logic 65 (2000) 1347–1374. doi:10.2307/2586704.
[22] M. Stipulanti, Convergence of Pascal-like triangles in Parry-Bertrand numeration systems, Theoret. Comput. Sci. 758 (2019) 42–60. doi:10.1016/j.tcs.2018.08.003.
[23] M. Lothaire, Combinatorics on words, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997. doi:10.1017/CBO9780511566097.
[24] P. B. A. Lecomte, M. Rigo, Numeration systems on a regular language, Theory Comput. Syst. 34 (2001) 27–44. doi:10.1007/s002240010014.
[25] V. Berthé, M. Rigo (Eds.), Combinatorics, Automata, and Number Theory, volume 135 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, 2010. doi:10.1017/CBO9780511777653.
[26] M. Rigo, A. Maes, More on generalized automatic sequences, Journal of Automata, Languages, and Combinatorics 7 (2002) 351–376. doi:10.25596/jalc-2002-351.
[27] J. Shallit, A generalization of automatic sequences, Theoret. Comput. Sci. 61 (1988) 1–16. doi:10.1016/0304-3975(88)90103-X.

[28] M. Rigo, Formal languages, automata and numeration systems. 2, Networks and Telecommunications Series, ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2014. Applications to recognizability and decidability.

[29] W. Parry, On the β-expansions of real numbers, Acta Math. Acad. Sci. Hungar. 11 (1960) 401–416. doi:10.1007/BF02020954.

[30] E. Zeckendorf, Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas, Bull. Soc. Roy. Sci. Liège 41 (1972) 179–182.

[31] A. Bertrand-Mathis, Comment écrire les nombres entiers dans une base qui n'est pas entière, Acta Math. Hungar. 54 (1989) 237–241. doi:10.1007/BF01952053.

[32] N. J. A. Sloane, et al., The On-Line Encyclopedia of Integer Sequences, 2024. URL: https://oeis.org.

[33] J.-P. Allouche, J. Cassaigne, J. Shallit, L. Q. Zamboni, A taxonomy of morphic sequences, 2019. Preprint available at https://arxiv.org/pdf/1711.10807.

[34] J. Shallit, The logical approach to automatic sequences—exploring combinatorics on words with `Walnut`, volume 482 of *London Mathematical Society Lecture Note Series*, Cambridge University Press, Cambridge, 2023.

[35] A. Cobham, Uniform tag seqences, Mathematical Systems Theory 6 (1972) 164–192. doi:10.1007/BF01706087.

[36] M. Morse, G. A. Hedlund, Symbolic Dynamics, Amer. J. Math. 60 (1938) 815–866.

[37] M. I. Hollander, Greedy numeration systems and regularity, Theory Comput. Syst. 31 (1998) 111–133. doi:10.1007/s002240000082.