A short walk across combinatorics of words

J.-P. Allouche
CNRS, IMJ-PRG, Sorbonne Université
4 Place Jussieu
F-75252 Paris Cedex 05 France
jean-paul.allouche@imj-prg.fr

Abstract

M. P. Schützenberger was extremely influential in theoretical computer science. We propose a general short survey on combinatorics of words and sequences, where we point to some of his contributions.

1 Introduction

Thue is often considered as the father of combinatorics of words. In two papers at the beginning of the 20th century Thue addressed the question of finding an infinite sequence on two symbols that is cube-free, i.e., that does not contain three consecutive identical blocks. May be Thue should be considered as a grandfather of combinatorics of words, since his work was later followed and largely extended by other researchers: among the fathers of combinatorics of words, Schützenberger certainly had a special and important rôle.

Combinatorics on words deal with finite or infinite sequences taking there values in a finite set. Finite such sequences are called words. Structures are added: concatenation of words (leading to the study of finitely generated free monoids), morphisms of monoids, topology on monoids, infinite sequences viewed as limits of words and their morphisms, periodicity and periodicity-like properties of infinite sequences... We will visit some of the concepts and results which go from Thue's discovery to recent papers. To keep this survey short, we will not give all proofs but either hints and sketches or precise references. We will only give a succinct bibliography. The interested readers can consult, e.g., the books [2, 5, 9, 10, 11, 12, 13, 14] and the references therein. Of course one should absolutely not forget [4] and its rich bibliography.

2 Basic definitions

We begin with basic definitions.

Definition 2.1

- A finite set A is called an *alphabet*. In case the cardinality of A is equal to 2, A is called a binary alphabet (e.g., $A = \{0, 1\}$). The elements of an alphabet are called *letters*.

- A *word* on the alphabet A is a finite sequence taking its values in A. If $w = (w_j)_{0 \leq j \leq n-1}$ is a word on A, it is written $w = w_0 w_1 \cdots w_{n-1}$, and n is called the *length* of w, and denoted by $|w|$. The word with no letter is called the *empty word* and denoted by $\emptyset$; its length is 0.

  The set of all words on A is denoted by A^*, and the set of non-empty words is denoted by A^+.

- A *language* on A is a subset of A^*.

- A sequence on A, i.e., an element $(a_n)_{n \geq 0} \in A^{\mathbb{N}}$, is sometimes called an *infinite word* on A. The set of all sequences on A is denoted by A^ω, and we let A^∞ denote $A^* \cup A^\omega$.

- If $v = v_0 v_1 \cdots v_{m-1}$ and $w = w_0 w_1 \cdots w_{n-1}$ are two words on the alphabet A, the *concatenation* of v and w, denoted by $v.w$ or vw, is defined by $vw = v_0 v_1 \cdots v_{m-1} w_0 w_1 \cdots w_{n-1}$. Note that the concatenation of a finite word and of an infinite word can be defined analogously.

- If t, v, w, z are four words on A and if $v = twz$, w is called a *factor* (or *subword*) of v, t is called a *prefix* of v and z is called a *suffix* of v. A prefix or a suffix of a word, different from the word itself, is called *proper*. Note that these definitions still make sense if z (hence v) is an infinite word.

- If w is a word on A and if k is an integer, then w^k is the concatenation of k copies of the word w (i.e., $w^k = w.w. \cdots .w$). Or, equivalently, w^k can be defined by induction: $w^0 = \emptyset$, and $w^{k+1} = w^k.w$ for all $k \geq 0$.

- A word (or a sequence) is called *square-free, cube-free, resp. k-free*, if it contains no prefix which is a square, a cube, resp. a k-th power.

- A word $w = w_0 w_1 \cdots w_{n-1}$ is called a *palindrome* if it has the property:

$$w = w_0 w_1 \cdots w_{n-1} = w_{n-1} w_{n-2} \cdots w_1 w_0.$$

Proposition 2.2 *The concatenation is associative in A^*. The empty word is the identity element. Hence $(A^*, .)$ is a monoid. (This is actually the* free monoid *generated by A.)*

Proof. Straightforward.

Remark 2.3

- A word $w = w_0 w_1 \cdots w_{n-1}$ is equal to the concatenation of its letters, namely:
$$w_0 w_1 \cdots w_{n-1} = w_0 \,.\, w_1 \,.\, \cdots \,.\, w_{n-1}$$
- The concatenation is not commutative in general if the cardinality of the alphabet A is at least 2, e.g., on the binary alphabet $\{0, 1\}$, we have $00.011 = 00011 \neq 01100 = 011.00$.

Though the concatenation is not commutative, it may happen that two words commute, for example: $01.0101 = 0101.01$. Here the two words are 01 and $0101 = (01)^2$. A theorem of Lyndon and Schützenberger states that this is "essentially" the only case where this happens. More precisely:

Theorem 2.4 (Lyndon, Schützenberger) *Let v and w be two non-empty words on A. Then $vw = wv$ if and only if there exist a word z and two positive integers i and j such hat $v = z^i$ and $w = z^j$.*

Proof. (sketch) One direction is clear. The other direction uses induction on the length of vw.

3 Morphisms

Given a structure (here monoids generated by a finite set), a natural question is to look at maps which preserve this structure. This motivates the following definition.

Definition 3.1

- Let A and B be two alphabets. Let $(A^*, .)$ and $(B^*, .)$ be the two monoids generated by A, resp. B (the concatenation in the two monoids is denoted by the same symbol for simplicity). A map h from A^* to B^* is called a *morphism* (or a *morphism of monoids*), if for all words $v \in A^*$ and $w \in B^*$ one has $h(v.w) = h(v).h(w)$.
- A morphism h from $(A^*, .)$ to $(B^*, .)$ is called *non-erasing* if for each $a \in A$, one has $h(a) \neq \emptyset$.
- Let $\ell \geq 1$ a positive integer. A morphism h from $(A^*, .)$ to $(B^*, .)$ is called *ℓ-uniform* if for each letter $a \in A$ one has $|h(a)| = \ell$. A morphism is called *uniform* (or a *morphism of constant length*) if it is ℓ-uniform for some $\ell \geq 1$.

Remark 3.2 Since a word is the concatenation of its letters (see above), a morphism on A^* is totally determined by its values on A.

4 Topology on infinite sequences on a finite set

Let A be a finite set. The set of infinite sequences with values in A, denoted by A^ω can be equipped with a topology. Intuitively two sequences are "close" if they have a "long" common prefix. More precisely A^ω is the product of infinitely many copies of A; it can thus be equipped with the product topology, where each A is equipped with the discrete topology. The set A^ω is then a compact set (by Tychonoff's theorem). It is not difficult to see that this topology is metrizable, see the following definition and remark.

Definition 4.1 Let A be a finite set. If $\mathbf{x} = (x_n)_{n\geq 0}$ and $\mathbf{y} = (y_n)_{n\geq 0}$ are two infinite sequences on A, i.e., two elements of A^ω, then we define

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 \text{ if } \mathbf{x} = \mathbf{y}, \\ 2^{-\min\{n\geq 0,\ x_n \neq y_n\}}. \end{cases}$$

Remark 4.2

- We leave it to the reader to prove that d is a distance and that it defines the product topology on A^ω.
- The set A^* of all finite words on A can be considered as a subset of $(A \cup \{\diamond\})^\omega$ (where $\diamond$ is a new letter not in A) by identifying each word $v = v_0 v_1 \cdots v_{n-1}$ with the infinite sequence $v_0 v_1 \cdots v_{n-1} \diamond\diamond \cdots \diamond \cdots$. Since A^ω can also be considered a subset of $(A \cup \{\diamond\})^\omega$, and since this last set can be equipped with a metrizable topology similar to the one above, one sees that $A^\infty = A^* \cup A^\omega$ is a subset of the metrizable set $(A \cup \{\diamond\})^\omega$.
- Given two alphabets A and B, a morphism f from A^* to B^* can be extended to A^∞ by defining for any infinite sequence with values in A, say $u = u_0 u_1 \cdots u_n \cdots$, its image $f(u) := f(u_0) f(u_1) \cdots f(u_n) \cdots$, where the $f(u_i)$'s are first concatenated, then rewritten as finite sequences of letters: for example, if $f(0) = 01$ and $f(1) = 0$, then $f(01001100011100001111 \cdots) = 0100101000101010000101010000 \cdots$; the reader will have noted that this is exactly extending f by continuity.

5 Complements

Before speaking of particular infinite sequences, e.g., morphic, automatic or Sturmian sequences, we will mention the paper of Cori and Perrin [7], which gives some of the contributions of Schützenberger and some of his influential ideas. As indicated in [7], Schützenberger had more than 300 publications appeared between 1943 et 2000. They have been collected and analyzed, and can be accessed at `http://www-igm.univ-mlv.fr/~berstel/Mps/index.html`. The numerous contributions of Schützenberger made him a pioneer in the

development of mathematics that were necessary for computer science, from combinatorial algorithms to automata and to formal grammars, from semigroups to codes, from formal languages to algebraic combinatorics, from the *repetition threshold* of an alphabet to *Lyndon words* and to *factorizations* of free monoids. As written by Lichnerovicz (cited in [7]): *Schützenberger avait le génie de la "préscience" – du génie tout court : il a su dégager très tôt les fondements algébriques et combinatoires qui doivent sous-tendre tout le champ de l'informatique.*

6 Complexity of an infinite sequence

A "measure" of how complicated a sequence on a **finite alphabet** is consists of counting the number of distinct factors of each length: a sequence where all possible words of all possible lengths occur (e.g., a "random" sequence) is intuitively more complicated than the sequence $000\cdots$ which has exactly one word of each length. This suggests the following definition.

Definition 6.1 Let A be an alphabet and $u = (u_n)_{n\geq 0}$ be a sequence in A^ω. The *(block-)complexity* of the sequence u is the function $(p(k))_{k\geq 0}$ defined by: for each $k \geq 0$, $p_u(k) = p(k)$ is the number of factors of u of length k. [In particular $p(0) = 1$ since the empty word is the only factor of u of length 0.]

A first result on the complexity of a sequence was given by Morse and Hedlund.

Proposition 6.2 *Let $u = (u_n)_{n\geq 0}$ be a sequence with values in the alphabet A of cardinality ≥ 2. Let $(p(k))_{k\geq 0}$ be the complexity of u. Then the following conditions are equivalent:*

(i) The function p is increasing, i.e., $p(k+1) > p(k)$ for all integers k.

(ii) The function p is unbounded.

(iii) For all integers k we have $p(k) > k$.

(iv) The sequence u is not ultimately periodic.

Proof. See, e.g., [2, Theorem 10.2.6, p. 302].

Remark 6.3 The penultimate item of Proposition 6.2 above shows that if $p_u(k) \leq k$ for some k, then the sequence u is ultimately periodic. In other words the complexity p_u of any non-ultimately periodic u satisfies $p(k) \geq k+1$.

Remark 6.4 Other complexities can be defined where, instead of counting the number of all factors of each length, one restricts to factors of a special type, or to classes of factors for some equivalence relation. For example, the

palindrome complexity counts the number of palindrome factors of each length: it can be proved that, e.g., the palindrome complexity of the Golay-Shapiro sequence is zero for large values of length (more precisely, the Golay-Shapiro sequence does not contain palindromes of length ≥ 15). Another example is the *abelian complexity* which counts the number of factors of each length without taking into account the order of letters: in other words it counts the number of equivalence classes for the relation where two words are equivalent if one can be obtained from the other by permuting its letters.

7 Morphic sequences

Consider the morphism h defined on $\{0,1\}$ (hence on $\{0,1\}^*$) by $h(0) = 01$, $h(1) = 0$. Let us compute the first iterates of h at 0:

$$\begin{array}{lcl} h(0) & = & 0 \\ h^2(0) & = & h(h(0)) = h(01) = h(0)h(1) = 010 \\ h^3(0) & = & h(h^2(0)) = h(010) = h(0)h(1)h(0) = 01001 \\ & \cdots & \end{array}$$

Looking at the sequence of words $(h(0), h^2(0), h^3(0), \ldots) = 01, 010, 01001, \ldots$, we see (and it can be proved) that each word $h^k(0)$ is a prefix of $h^{k+1}(0)$, hence that this sequence of words converges (in the topology of $(A \cup \{\diamond\})^\omega$ indicated above) to an infinite binary sequence $0\ 1\ 0\ 0\ 1\ 0\ 1\ \cdots$, called the *(binary) Fibonacci sequence* (the readers will certainly have noticed –and they can prove– that the sequence of lengths of the words $h(0), h^2(0), h^3(0), \ldots$, i.e., $1, 3, 5, \cdots$ is precisely the usual Fibonacci sequence of integers).

Definition 7.1 Let A be an alphabet. A non-erasing morphism from A^* to A^* is said to be *prolongable* on some letter $a \in A$, if there exists a non-empty word $w \in A^*$ such that $h(a) = aw$.

Proposition 7.2 *Let A be an alphabet and a be a letter in A. Let h be a non-erasing morphism, prolongable on a, then the sequence of words $(h^k(a))_{k\geq 1}$ converges to an infinite sequence denoted by $h^\infty(a)$ which is a fixed point of the (extension to A^ω of the) morphism h. The sequence $h^\infty(a)$ is called an* iterative fixed point *of the morphism h.*

Proof. (sketch) Suppose that $h(a) = aw$. It is easy to see that one has $h^k(a) = awh(w)h^2(w)\cdots h^{k-1}(w)$. So $h^k(a)$ is a prefix of $h^{k+1}(a)$ and $|h^k(a)|$ tends to infinity, which proves the claim.

Definition 7.3

- A sequence with values in the alphabet A is said to be *pure morphic* if it is an iterative fixed point of a non-erasing prolongable morphism from A^* to A^*.

- A sequence $(x_n)_{n\geq 0}$ with values in the alphabet A is said to be *morphic* if there exist an alphabet B, a pure morphic sequence $(z_n)_{n\geq 0}$ with values in B, and a map ψ from B to A such that $x_n = \psi(z_n)$ for all $n \geq 0$. The sequence $(x_n)_{n\geq 0}$ is also called the *pointwise image* of the sequence $(z_n)_{n\geq 0}$ under the map ψ.

Remark 7.4

- A pure morphic sequence is in particular morphic.
- The pointwise image of a sequence under a map ψ can also be seen as the image of this sequence under the 1-uniform morphism defined by ψ (see Definition 3.1).

Example 7.5 The binary Fibonacci sequence defined at the beginning on this section is the iterative fixed point of the morphism $h(0) = 01$, $h(1) = 0$. It is pure morphic.

8 Automatic sequences

In this section we will focus on morphic sequences corresponding to morphisms of constant length (see Definition 3.1).

8.1 Definitions

Definition 8.1 Let ℓ be an integer ≥ 2. A sequence is said to be *ℓ-automatic*, if it is an iterative fixed point of a non-erasing prolongable morphism of constant length ℓ. A sequence is called *automatic* (or *uniformly morphic*) if it is ℓ-automatic for some $\ell \geq 2$.

Remark 8.2 The terminology "automatic" comes from an equivalent definition (that we will not give here) through *finite automata with output function.*

A somehow simpler (but equivalent) definition is given in the following proposition.

Proposition 8.3 *Let ℓ be an integer ≥ 2. A sequence $(x_n)_{n\geq 0}$ is ℓ-automatic if and only if the set of subsequences $\{(x_{\ell^i n+a})_{n\geq 0},\ i \geq 0,\ a \in [0, \ell^i - 1]\}$ is finite.*

Proof. See, e.g., [2, p. 185–186].

Remark 8.4

- The set of subsequences $\{(x_{\ell^i n+j})_{n\geq 0},\ i \geq 0,\ a \in [0, \ell^i - 1]\}$ of the sequence $(x_n)_{n\geq 0}$ is called the *ℓ-kernel* of the sequence $(x_n)_{n\geq 0}$.

- To have a graphical view of Proposition 8.3 above, take $\ell = 2$. Then construct a tree starting from the sequence $(x_n)_{n\geq 0}$. Then draw two branches, say a left branch going from $(x_n)_{n\geq 0}$ to the sequence $(x_{2n})_{n\geq 0}$ and a right branch going from $(x_n)_{n\geq 0}$ to the sequence $(x_{2n+1})_{n\geq 0}$, i.e., the two branches go to the subsequences of $(x_n)_{n\geq 0}$ of even and of odd indices. "Split" in the same way the sequences $(x_{2n})_{n\geq 0}$ and $(x_{2n+1})_{n\geq 0}$, thus drawing branches, going respectively from $(x_{2n})_{n\geq 0}$ to $(x_{4n})_{n\geq 0}$ and to $(x_{4n+2})_{n\geq 0}$, and from $(x_{2n+1})_{n\geq 0}$ to $(x_{4n+1})_{n\geq 0}$ and to $(x_{4n+3})_{n\geq 0}$. If $(x_n)_{n\geq 0}$ is *any* sequence one obtains in general an infinite tree, in which it may happen that some leafs are equal: if two leaves are equal, suppress the youngest one and the whole subtree that it generates; the resulting simplified tree has all its leaves distinct, but it might still be infinite, or even equal to the initial tree. Proposition 8.3 characterizes the 2-automatic sequences as sequences having a finite simplified tree.

8.2 Examples

The first "trivial" examples are given in Proposition 8.5 below.

Proposition 8.5 *All ultimately periodic sequences are ℓ-automatic for any integer $\ell \geq 2$.*

Proof. Left to the reader.

Now we give a few famous examples of 2-automatic sequences.

8.2.1 The Prouhet-Thue-Morse sequence

The *Prouhet-Thue-Morse sequence*, also called the *Thue-Morse sequence*, is the most famous automatic sequence. It was introduced by Thue, as an example of a sequence on two symbols that has no cubes (i.e., no block composed of three identical consecutive blocks): the two papers of Thue in 1906 and 1912 seem to be the first papers addressing a question in combinatorics on words. This sequence was rediscovered several times, in particular by Morse in 1921. Its first occurrence in the literature actually goes back to Prouhet in 1851 (for a survey on this sequence, see [1]).

The Thue-Morse sequence on the alphabet $\{0, 1\}$ is the iterative fixed point beginning with 0 of the morphism h defined by $h(0) = 01$, $h(1) = 10$. One sees that

$$\begin{aligned} h(0) &= 01 \\ h^2(0) &= h(h(0)) = h(01) = 0110 \\ h^3(0) &= 01101001 \\ h^4(0) &= 0110100110010110\ldots \end{aligned}$$

so that the Thue-Morse sequence on $\{0,1\}$ is the sequence

$$h^\infty(0) = 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ \cdots$$

This sequence is not only cube-free but it is also *overlap-free*, i.e., it does not contain any factor wwa where $w \in \{0,1\}^*$ and $a \in \{0,1\}$.

8.2.2 The Golay-Shapiro sequence

The *Golay-Shapiro sequence*, also called the *Rudin-Shapiro sequence*, was discovered independently by Golay and Shapiro in 1951, and studied by Rudin in 1959 (see, e.g., [3, Remark 1, p. 206]). One remarkable property of this sequence is the following. If $(a_n)_{n\geq 0}$ is a binary sequence with values in $\{0,1\}$, then

$$\sqrt{N} \leq \sup_{\theta\in[0,1]} \left| \sum_{0\leq n\leq N-1} (-1)^{a_n} e^{2i\pi n\theta} \right| \leq N$$

(while the upper bound is trivial, the lower bound is obtained by comparing the L^∞ and L^2 norms of the sum). Furthermore, for (Lebesgue) almost all sequences $(a_n)_{n\geq 0}$, the upper bound can be replaced by $\mathcal{O}(\sqrt{N\log N})$. But if $(r_n)_{n\geq 0}$ is the Golay-Shapiro sequence, then

$$\sqrt{N} \leq \sup_{\theta\in[0,1]} \left| \sum_{0\leq n\leq N-1} (-1)^{r_n} e^{2i\pi n\theta} \right| \leq C\sqrt{N}$$

where $C = 2+\sqrt{2}$. In other words this sequence almost behaves like almost all sequences, i.e., like a "random" sequence. But $(r_n)_{n\geq 0}$ is 2-automatic (hence quite non-random), namely it can be defined as follows. Let $A = \{a,b,c,d\}$, define the morphism σ on A by:

$$\sigma(a) = ab,\ \sigma(b) = ac,\ \sigma(c) = db,\ \sigma(d) = dc$$

and let ψ be the map: $\psi(a) = \psi(b) = 0$, $\psi(c) = \psi(d) = 1$. Then define the sequence $v = (v_n)_{n\geq 0}$ by $v = \sigma^\infty(a)$. The Golay-Shapiro sequence is the pointwise image of the sequence v by the map ψ (i.e., $r_n = \psi(v_n)$ for all $n \geq 0$).

8.2.3 The (regular) paperfolding sequence

Repeatedly folding a strip of paper yields, when the paper is unfolded, a sequence of folds consisting of "peaks" and "valleys". This sequence turned into a binary sequence on $\{0,1\}$ is 2-automatic. It can be also constructed as follows. Let $A = \{a,b,c,d\}$, and let τ be the morphism defined on A by:

$$\tau(a) = ab,\ \tau(b) = cb,\ \tau(c) = ad,\ \tau(d) = cd.$$

Let $x = (x_n)_{n\geq 0}$ be the sequence $x = \tau^\infty(a)$ and let φ be the map defined by: $\varphi(a) = \varphi(b) = 0$, $\varphi(c) = \varphi(d) = 1$. Then the (regular) paperfolding sequence is the sequence $p = (p_n)_{n\geq 0}$ defined as the pointwise image of the sequence x by the map φ.

8.2.4 The Baum-Sweet sequence

While it is well known that the continued fraction expansion of a real number quadratic over the rationals is (ultimately) periodic (Lagrange's theorem), nothing is known for algebraic numbers of degree ≥ 3: no example is known with bounded partial quotients, nor with unbounded partial quotients. If the field of rationals is replaced with the 2-element field and the real numbers with the field of Laurent power series in $1/X$ over the 2-element field, there is also a theory of continued fractions, where the partial quotients are polynomials, and where "bounded partial quotients" means "partial quotients of bounded degree" (i.e., there are only finitely many distinct partial quotients). In this setting Lagrange's theorem still holds. But more is known for the case of algebraic Laurent series of large degree: Baum and Sweet gave in 1976 the first example of a Laurent formal power series, algebraic of degree 3, whose sequence of partial quotients is bounded. The sequence of its coefficients, say $(b_n)_{n\geq 0}$, is 2-automatic. It is defined as follows.
Let $A = \{a, b, c, d\}$. Let ρ be the morphism on A^* such that

$$\rho(a) = ab,\ \rho(b) = cb,\ \rho(c) = bd,\ \rho(d) = dd.$$

Let ς be the morphism defined on A by: $\varsigma(a) = \varsigma(b) = 1$, $\varsigma(c) = \varsigma(d) = 0$. Define the sequence $z = (z_n)_{n\geq 0}$ by $z := \rho^\infty(a)$. The Baum-Sweet sequence is the pointwise image of the sequence z by the map ς.

8.2.5 The Hanoi sequence

It was proved that the classical puzzle called "the Tower of Hanoi" can be solved by using an infinite sequence on six symbols (corresponding to all possible directed moves that take a ring from the top of a peg to place it on the top of another peg). This sequence called *the Hanoi sequence* has the property that its prefixes of length $2^n - 1$ give the moves that take n rings from the first to the second peg is n is odd, and from the first to the third peg if n is even. This sequence, say $(H_n)_{n\geq 0}$ can be defined as follows. Let B be the alphabet $B = \{a, b, c, \overline{a}, \overline{b}, \overline{c}\}$ and let ε be the morphism defined on B by

$$\varepsilon(a) = a\overline{c},\ \varepsilon(b) = c\overline{b},\ \varepsilon(c) = b\overline{a},\ \varepsilon(\overline{a}) = ac,\ \varepsilon(\overline{b}) = cb,\ \varepsilon(\overline{c}) = ba.$$

Then the Hanoi sequence is the sequence $(H_n)_{n\geq 0} = \varepsilon^\infty(a)$.

Remark 8.6 The reader can prove that all examples above are 2-automatic sequences by proving that their respective 2-kernels are finite.

8.3 First properties of automatic sequences

Let us begin with basic properties of automatic sequences.

Proposition 8.7

- *If a sequence is ℓ-automatic for some $\ell \geq 2$, the sequences obtained respectively by inserting a finite number of terms, deleting a finite number of terms, or changing a finite number of terms, are still ℓ-automatic.*
- *If a sequence is ℓ-automatic, any pointwise image of this sequence is still ℓ-automatic.*
- *Let $\ell \geq 2$ and $m \geq 1$ be two integers. A sequence is ℓ-automatic if and only if it is ℓ^m-automatic.*
- *Let $\ell \geq 2$ be an integer. Let a, b be two nonnegative integers. If a sequence $(x_n)_{n\geq 0}$ is ℓ-automatic, then the sequence $(x_{an+b})_{n\geq 0}$ is also ℓ-automatic.*
- *Let $\ell \geq 2$ be an integer. Let $a \geq 1$ be a positive integer. Suppose that the sequences $(x_{an+b})_{n\geq 0}$ are ℓ-automatic for all $b \in [0, a-1]$, then the sequence $(x_n)_{n\geq 0}$ is ℓ-automatic.*
- *Let $\ell \geq 2$ and $k \geq 1$ be two integers. The image of an ℓ-automatic by a uniform morphism of length k is ℓ-automatic.*
- *Let A be an alphabet equipped with an operation $*$ such that $(A, *)$ is a monoid. If the sequences $(v_n)_{n\geq 0}$ and $(w_n)_{n\geq 0}$ are ℓ-automatic for some integer $\ell \geq 2$, so are the sequences $(x_n)_{n\geq 0}$ and $(y_n)_{n\geq 0}$, where $x_n = v_n * w_n$ and $y_n = v_0 * v_1 * \cdots * v_n$ for all $n \geq 0$.*
- *Let $u = (u_n)_{n\geq 0}$ be an automatic sequence. Then its complexity $(p_u(k))_{k\geq 0}$ satisfies $p(k) = \mathcal{O}(k)$.*

8.4 The theorem of Christol

In two papers Christol (1979) and Christol, Kamae, Mendès France and Rauzy (1980) established an unexpected link between p^r-automaticity (p a prime number), and algebraicity. We give a slightly simplified version of their result.

Theorem 8.8 (Christol / Christol, Kamae, Mendès France, Rauzy) *Let p be a prime integer, let $r \geq 1$ be a positive integer, and let $q = p^r$. Let $\mathbb{F}_q$ be the q-element field and let $(v_n)_{n\geq 0}$ be a sequence with values in $\mathbb{F}_q$. Then the sequence $(v_n)_{n\geq 0}$ is q-automatic if and only if the formal power series $\sum_{n\geq 0} v_n X^n$ is algebraic over the field of rational functions $\mathbb{F}_q(X)$.*

Proof. See, e.g., [2, Theorem 12.2.5, p. 356].

Remark 8.9 In order to "see" the meaning of the Christol theorem, let us look at an explicit example. We take the Thue-Morse sequence $t = (t_n)_{n\geq 0}$,

which (Section 8.2.1) satisfies $t = h^\infty(0)$, where h is the morphism defined on $\{0,1\}$ by $h(0) = 01$, $h(1) = 10$. Since the sequence t is a fixed point of (the extension to $\{0,1\}^\omega$) of this morphism, we have $t_{2n} = t_n$ and $t_{2n+1} = 1 - t_n$: namely, the image of a sequence $(z_n)_{n\geq 0}$ by h is the sequence

$$z_0\ (1-z_0)\ z_1\ (1-z_1)\cdots = (z'_n)_{n\geq 0}$$

where $z'_{2n} = z_n$ and $z'_{2n+1} = 1 - z'_n$. Let F be the formal power series defined by $F = F(X) = \sum_{n\geq 0} t_n X^n$. We have

$$\begin{aligned}
F(X) &= \sum_{n\geq 0} t_{2n} X^{2n} + \sum_{n\geq 0} t_{2n+1} X^{2n+1} \\
&\qquad \text{(by splitting the sum between even and odd indices)} \\
&= \sum_{n\geq 0} t_n X^{2n} + \sum_{n\geq 0} (1-t_n) X^{2n+1} \\
&\qquad \text{(by using the properties of the sequence } (t_n)_{n\geq 0}) \\
&= F(X^2) + \sum_{n\geq 0} X^{2n+1} - XF(X^2) \\
&= (1-X)F(X^2) + \frac{X}{1-X^2}
\end{aligned}$$

Now, identifying $\{0,1\}$ with the 2-element field $\mathbb{F}_2$, recall that $+1 = -1 \bmod 2$, and note that we have $F(X^2) = F(X)^2$ and $(1-X^2) = (1+X)^2$. So that the previous equality can be written

$$\begin{aligned}
F(X) &= (1+X)F(X)^2 + \frac{X}{(1+X)^2} \\
&= -(1+X)F(X)^2 - \frac{X}{(1+X)^2}.
\end{aligned}$$

In other words

$$(1+X)^3 F^2 + (1+X)^2 F + X = 0$$

which proves that F is algebraic over $\mathbb{F}_2(X)$ of degree at most 2. But the fact that $(t_n)_{n\geq 0}$ has no cube implies that it cannot be ultimately periodic, so that F is quadratic over $\mathbb{F}_2(X)$.

Remark 8.10 Of course the Christol theorem has nothing to do with the Chomsky-Schützenberger theorem on context-free languages [6] which states: *If L is a context-free language admitting an unambiguous context-free grammar, and if $a(k)$ is the number of words of length k in L, then the formal power series $\sum_{k=0}^{\infty} a(k)X^k$ is algebraic over the function field $\mathbb{Q}(X)$.*

8.5 More on automatic sequences?

The Christol theorem shows a deep link between automatic sequences and number theory. There are many other number-theoretical results for these sequences (automatic Dirichlet series, transcendence properties of reals given by their base-b expansion or by their continued fraction expansion, distribution modulo 1, etc.). Automatic sequences also occur in the theory of dynamical systems, in physics, and even in music. We will not develop these points here; the reader can consult, e.g., [2] and the references given therein.

9 Sturmian sequences and beyond

9.1 Sturmian sequences

Another famous class of infinite sequences studied by the word-combinatorists is the class of Sturmian sequences. Recall from Remark 6.3 that the complexity of any non-ultimately periodic sequence u satisfies $p_u(k) \geq k+1$. The "simplest" (in the sense of "having a minimal complexity") sequences would satisfy $p(k) = k+1$ for all k. Such sequences should be defined on a two-letter alphabet (because $p(1) = 2$). It was proved by Morse and Hedlund, and Coven that such sequences do exist.

Theorem 9.1 *Let $u = (u_n)_{n\geq 1}$ be a sequence defined on the binary alphabet $\{0,1\}$. Then the following conditions are equivalent.*

(i) The complexity of u satisfies $p_u(k) = k+1$ for all $k \geq 0$.

(ii) There exists an irrational $\alpha \in (0,1)$ and a real $\theta \in [0,1)$ such that

$$\begin{cases} \text{either} & u_n = \lfloor \alpha(n+1)+\theta \rfloor - \lfloor \alpha n + \theta \rfloor, \text{ for all } n \geq 1, \\ \text{or} & u_n = \lceil \alpha(n+1)+\theta \rceil - \lceil \alpha n + \theta \rceil, \text{ for all } n \geq 1. \end{cases}$$

Definition 9.2 If a sequence u satisfies the conditions in Theorem 9.1 above, it is called a *Sturmian sequence* of *slope* α and *intercept* θ. Furthermore, if $\theta = 0$, the sequence is called a *characteristic sequence*.

Remark 9.3 It can be proved that the binary Fibonacci sequence previously described is (the possibly simplest example of) a Sturmian sequence.

A nice survey on Sturmian sequences is [12, Ch. 2] by Berstel and Séébold.

9.2 More on Sturmian sequences. Generalizations

We will not enter the numerous properties of Sturmian sequences which are linked to billiards, to continued fraction expansions, to digital straight lines, to rotations on the circle, to physics (quasicrystals)... A particular feature

of Sturmian sequences is that several properties that are equivalent to their definition can be generalized as tentative definitions of sequences on alphabets of cardinality larger than 2, but these various possible definitions are *not* all equivalent. The most studied generalization seems to be *episturmian sequences*, which were first introduced in [8].

10 Multidimensional sequences

Is it possible to define "words" in several dimensions? Let us restrict to dimension 2. One possiblity is to consider "rectangular" or even "square" words (more complicated patterns make the generalization of concatenation difficult). Square words naturally extend to $2D$-sequences, i.e., sequences $(u_{m,n})_{m,n\in\mathbb{N}^2}$. To give the flavor of word-combinatorial properties of such sequences, we give one of the definitions of 2-automatic $2D$-sequences. We also state a generalization of Christol's theorem.

Definition 10.1 Let $u = (u_{m,n})_{m,n\in\mathbb{N}^2}$ be a $2D$-sequence with values in some alphabet A. Let $\ell \geq 2$ be an integer. Then the sequence u is said to be *ℓ-automatic* if its ℓ-kernel is a finite set, where the ℓ-kernel of the sequence u is the set of subsequences

$$\mathrm{Ker}(u) = \{(u_{\ell^i m+a,\ell^i n+b})_{m,n\in\mathbb{N}^2},\ i \geq 0,\ a,b \in [0,\ell^i - 1]\}.$$

Theorem 10.2 (Salon) *Let p be a prime integer, let r be a positive integer, and let $q = p^r$. Let $\mathbb{F}_q$ be the q-element field and let $(v_{m,n})_{m,n\in\mathbb{N}^2}$ be a (double) sequences with values in $\mathbb{F}_q$. Then the sequence $(v_{m,n})_{m,n\in\mathbb{N}^2}$ is q-automatic if and only if the formal power series $\sum v_{m,n}X^mY^n$ is algebraic over the field of rational functions $\mathbb{F}_q(X,Y)$.*

Remark 10.3 To give both an idea of concatenation of $2D$-square words and of $2D$-uniform morphisms, we propose to look at the morphism

$$0 \to \begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \qquad 1 \to \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}.$$

Iterating this morphism as in the one-dimensional case gives:

$$0 \to \begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \to \begin{array}{cccc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array} \to \cdots$$

where the reader can see that in the last array the two 2×2 top squares, namely $\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}$ and $\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}$ are the images of the two elements of the first line of the penultimate array, and similarly that the two 2×2 bottom squares are the images of the two elements of the second line of the penultimate array.

References

[1] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, in *Sequences and their Applications (Singapore, 1998)*, 1–16, Springer Ser. Discrete Math. Theor. Comput. Sci., Springer, London, 1999.

[2] J.-P. Allouche, J. Shallit, *Automatic Sequences. Theory, Applications, Generalizations*, Cambridge University Press, Cambridge, 2003.

[3] J.-P. Allouche, On a Golay-Shapiro-like sequence, *Unif. Distrib. Theory* **11** (2016), 205–210.

[4] J. Berstel, D. Perrin, The origins of combinatorics on words, *European J. Combin.* **28** (2007), 996–1022.

[5] V. Berthé (ed.), M. Rigo (ed.), *Combinatorics, Words and Symbolic Dynamics*, Encyclopedia of Mathematics and its Applications **159**, Cambridge University Press, 2016.

[6] N. Chomsky and M. P. Schützenberger, *The algebraic theory of context-free languages*, in Computer Programming and Formal Languages, 118–161, North Holland, Amsterdam, 1963.

[7] R. Cori, D. Perrin, Une introduction à la contribution scientifique de Marcel-Paul Schützenberger, *1024 – Bull. Soc. Info. France* **9** (2016), 35–49.

[8] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, *Theoret. Comput. Sci.* **255** (2001), 539–553.

[9] N. Pytheas Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*, Lecture Notes in Mathematics **1794**, Springer-Verlag, Berlin, 2002.

[10] F. von Haeseler, *Automatic Sequences*, de Gruyter Expositions in Mathematics **36**, Walter de Gruyter & Co., Berlin, 2003.

[11] M. Lothaire, *Combinatorics on Words*, 2nd ed., Encyclopedia of Mathematics and its Applications **17**, 1997.

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

[13] M. Lothaire, *Applied Combinatorics on Words*, Encyclopedia of Mathematics and its Applications **105**, Cambridge University Press, 2005.

[14] M. Rigo, *Formal Languages, Automata and Numeration Systems, vol.1. Introduction to Combinatorics on Words*, ISTE, John Wiley & Sons, 2014.