Axel Thue's work on repetitions in words*

J. Berstel
L.I.T.P.
Université Pierre et Marie Curie
Institut Blaise Pascal
Paris, France

Abstract

The purpose of this survey is to present, in comtemporary terminology, the fundamental contributions of Axel Thue to the study of combinatorial properties of sequences of symbols, insofar as repetitions are concerned. The present state of the art is also sketched.

1 Introduction

In a series of four papers which appeared during the period 1906–1914, Axel Thue considered several combinatorial problems which arise in the study of sequences of symbols. Two of these papers [44, 46] deal with word problems for finitely presented semigroups (these papers contain the definition of what is now called a "Thue system"). He was able to solve the word problem in special cases. It was only in 1947 that the general case was shown to be unsolvable independently by E. L. Post [28] and A. A. Markov [24].

The other two papers [43, 45] deal with repetitions in finite and infinite words. Perhaps because these papers were published in a journal with restricted availability (this is guessed by G. A. Hedlund [20]), this work of Thue was widely ignored during a long time, and consequently some of his results have been rediscovered again and again. Axel Thues papers on sequences are now more easily accessible since they are included in the "Selected Papers" [47] which were edited in 1977.

It is the purpose of the present paper to give an account of Axel Thue's work on repetitions in sequences, both in more recent terminology and in relation with new results and directions of research. It appears that there is a noticeable difference, both in style and in amount of results, between the 1906 paper (22 pages) and the 1912 paper (67 pages). The first of these papers mainly contains the

*Partially supported by the PRC "Mathématique et Informatique"

construction of an infinite square-free word over three letters. Thue gives also an infinite square-free word over four letters obtained by what is now called an iterated morphism, whilst the three letter word is constructed in a slighly more complicated way (a uniform tag-system, in the terminology of Cobham [12]).

The second paper attacks the more general problem of what Thue calls *irreducible* words. He devotes special attention to the case of two and three letters. In particular, he introduces what is now called the *Thue-Morse sequence*, and shows that all twosided infinite overlap-free words are derived from this sequence. There are several aspects he did not consider: first, many combinatorial properties of the Thue-Morse sequence (such as the number of factors, the recurrence index, and so on) were only investigated by M. Morse [25] or later; next, the characterization of all onesided infinite overlap-free words — which is much more difficult than that of twosided words — was only given later by Fife [15]. However, Thue gives a complete description of circular overlap-free words. We will also mention the problem of counting the number of overlap-free words over two letters.

Axel Thue's investigation of square-free words over three letters is even more detailed. He gives, in this paper, another construction of an infinite square-free word, by iterated morphism, and then initiates, in a 30 pages development, a tentative to describe all square-free words over three letters. He observes that every infinite square-free word is an infinite product of words chosen in a set of six words, and classifies those infinite square-free words that are products of four among these six words. His classification, he observes, is similar both in statement and in proof technique to what is found in diophantine equations: the solutions are parametrized by some variables which are easier to manage.

The paper is organized as follows: after some preliminary definitions, we introduce the so-called Thue-Morse sequence. We next describe Thue's results on this word, and give a short account of other developments about overlap-free words. The next section contains a presentation of Thue's constructions of square-free words, and a comparison with other methods. Then, Thue's classification — which has been ignored for large parts — is described. We end with a short description of *avoidable* patterns, which is the main stream of actual research.

An *alphabet* is a finite set (of *symbols* or *letters*). A *word* over some alphabet A is a (finite) sequence of elements in A. The length of a word w is denoted by $|w|$. The *empty word* of length 0 is denoted by ε. An *infinite* word is a mapping form into A, and a twosided infinite word is a mapping from into A. A *circular* word or *necklace* is the equivalence class of a finite word under circular permutation. It can also be considered as a mapping of $/n$ into A for some positive integer n.

A *factor* of a word w is any word u that occurs in w, i. e. such that there exist word x, y with $w = xuy$. A *square* is a nonempty word of the form uu. A word is *square-free* if none of its factors is a square. Similarly, an *overlap* is a word of the form $xuxux$, where x is nonempty. The terminology is justified by

the fact that xux has two occurrences in $xuxux$, one as a *prefix* (initial factor) one as a *suffix* (final factor) and that thes occurrences have a common part (the central x). As before, a word is *overlap-free* if none of its factors is an overlap.

The set of words over A is the free monoid generated by A. It is denoted by A^*. A function $h : A^* \to B^*$ is a *morphism* if $h(uv) = h(u)h(v)$ for all words u, v. If there is a letter a such that $h(a)$ starts with the letter a, then $h^n(a)$ starts with the letter a for all $n \geq 0$. If the set words $\{h^n(a)) \mid n \geq 0\}$ is infinite, the morphism is *prolongeable* in a and defines a unique infinite word say $\mathbf{x}$ by the requirement that all $h^n(a)$ are prefixes of $\mathbf{x}$. The word $\mathbf{x}$ is said to be obtained by iterating h on a, and $\mathbf{x}$ is also denoted by $h^\omega(a)$. Clearly, $\mathbf{x}$ is a fixed point of h. This construction is frequently used by Axel Thue.

For Axel Thue, a word w over an alphabet of size n is *irreducible* if any two occurrences of the same word as a factor in w are always separated by at least $n-2$ letters. This means that an irreducible word over two letters is *overlap-free* and that an irreducible word over three letters is *square-free*.

2 The Thue-Morse sequence

In this section, we recall some basic properties concerning the Thue-Morse sequence. Other properties and proofs can be found in Lothaire [22] and Salomaa [34].

Let $A = \{a, b\}$ be a two letter alphabet. Consider the morphism μ from the free monoid A^* into itself defined by

$$\mu(a) = ab, \qquad \mu(b) = ba$$

Setting, for $n \geq 0$,

$$u_n = \mu^n(a), \qquad v_n = \mu^n(b)$$

one gets

$$\begin{array}{ll} u_0 = a & v_0 = b \\ u_1 = ab & v_1 = ba \\ u_2 = abba & v_2 = baab \\ u_3 = abbabaab & v_3 = baababba \\ \cdots & \end{array}$$

and more generally

$$u_{n+1} = u_n v_n, \qquad v_{n+1} = v_n u_n$$

and

$$u_n = \overline{v}_n, \qquad v_n = \overline{u}_n$$

where $\overline{w}$ is obtained from w by exchanging a and b. Words u_n and v_n are frequently called *Morse blocs*. It is easily seen that u_{2n} and v_{2n} are palindroms,

and that $u_{2n+1} = v_{2n+1}^{\sim}$, where $w^{\sim}$ is the reversal of w. The morphism μ can be extended to infinite words; it has two fixed points

$$\mathbf{t} = abbabaabbaababbabaab\cdots = \mu(\mathbf{t})$$

$$\overline{\mathbf{t}} = baababbaabbabaababba\cdots = \mu(\overline{\mathbf{t}})$$

and u_n (resp. v_n) is the prefix of length 2^n of $\mathbf{t}$ (resp. of $\overline{\mathbf{t}}$). It is equivalent to say that $\mathbf{t}$ is the limit of the sequence $(u_n)_{n\geq 0}$ (for the usual topology on finite and infinite words), obtained by iterating the morphism μ.

The *Thue-Morse sequence* is the word $\mathbf{t}$. There are several other characterizations of this word. Let t_n be the n-th symbol in $\mathbf{t}$, starting with $n = 0$. Then it is easily shown by induction that

$$t_n = \begin{cases} a & \text{if } d_1(n) \equiv 0 \pmod 2 \\ b & \text{if } d_1(n) \equiv 1 \pmod 2 \end{cases}$$

where $d_1(n)$ is the number of bits equal to 1 in the binary expansion of n. For instance, $bin(19) = 10011$, consequently $d_1(19) = 3$, and indeed $t_{19} = a$.

As a consequence, there is a finite automaton computing the values t_n as a function of $bin(n)$. This automaton has two states 0 and 1. It reads the string $bin(n)$ from left to right, starting in state 0. At the end, the state reached is 0 or 1 according to $t_n = b$ or $t_n = a$. In fact, the automaton computes $d_1(n)$ modulo 2. Another description is given by Christol, Kamae, Mendès-France, Rauzy in [11]. There are many generalizations of the Thue-Morse sequence, motivated by its simplicity, and by its numerous properties. The first definition of the sequence, by iterating a given morphism, is of course strongly related to Lindenmayer systems (see e.g. [32]). In the case where the morphism is *uniform*, that is when the lengths of the images of the letters are equal, a general theorem of Cobham [12] shows that the sequence $\mathbf{x}$ obtained by iterating the morphism can also be generated by a finite automaton working on expansions of natural integers in some base k. An equivalent way to state this is to say that there are only finitely many distinct subsequences $(x_{k^r n+s})_{n\geq 0}$ for $r \geq 0$ and $0 \leq s \leq k^r - 1$. Let us call such a sequence *automatic* (more precisely k-automatic). Another extension is by arithmetics. Consider a k letter alphabet $\{0, 1, \ldots, k-1\}$ and define an infinite word $\mathbf{x}$ by taking x_n to be the sum, modulo k, of all the digits in the expression for n in base k. The Thue-Morse sequence is then just the case $k = 2$. Since there is an automaton for computing x_n from the k-ary expansion of n, there is also a uniform morphism generating $\mathbf{x}$. For instance, if $k = 3$, the morphism, say μ_3, is given by $0 \to 012$, $1 \to 120$, $2 \to 201$ (this general definition was in fact given already given by Prouhet, in 1851. Several authors, such as Adler, Li [1] and Brlek [6], discuss the fact that Prouhet was the first to mention what perhaps should be called the Prouhet sequence).

Other sequences related to the Thue-Morse sequence are obtained by counting factors in the binary expansion, instead of bits. The *Rudin-Shapiro* [33, 38]

sequence is the infinite word $\mathbf{x}$ over $\{a, b\}$ defined by

$$x_n = \begin{cases} a & \text{if } d_{11}(n) \equiv 0 \pmod 2 \\ b & \text{if } d_{11}(n) \equiv 1 \pmod 2 \end{cases}$$

where $d_{11}(n)$ is the number of factors 11 in the binary expansion of n. Similarily, in the sequence of Baum and Sweet [3], the n-th symbol is a or b according to whether ther exists a factor of odd length containing only the bit 0 in the binary expansion of n. Again, this sequence is automatic. Many number-theoretic results have been given for automatic sequences. Let us just mention the following, due to Loxton and van der Poorten [23]:

THEOREM *For any automatic infinite word* $\mathbf{x}$ *over the alphabet* $\{0, \ldots, p-1\}$, *the real number*

$$\sum x_n p^{-n}$$

is transcendental.

As an example, the real number whose binary expansion is $0.011010011\cdots$ (associated to the Thue-Morse sequence) is trancendental (this was already known before).

3 Overlap-free words

As already mentioned, the Thue-Morse sequence is overlap-free. Indeed, A. Thue proved[1]

THEOREM (Satz 6) *The sequence* $\mathbf{t}$ *is overlap-free.*

What Thue actually show, is that a word w over the two letter alphabet $A = \{a, b\}$ is overlap-free iff $\mu(w)$ is overlap-free. Thue observes that the same result holds for circular words. More precisely, he gives the following complete characterization of circular overlap-free words:

THEOREM (Satz 13) *Every circular overlap-free word over the two letter alphabet* $A = \{a, b\}$ *is of the form* $\mu^n(ab)$, $\mu^n(aab)$ *or* $\mu^n(abb)$ *for some* $n \geq 0$.

As a consequence, a circular overlap-free word has length 2^n or $3 \cdot 2^n$ for some $n \geq 0$. These results are interesting because they are related to overlap-free squares. It is indeed easy to show that a circular word w is overlap-free iff the (ordinary) word ww is overlap-free. Thus, Thue characterizes overlap-free squares, a result that was discovered later also by [42]. T. Harju [19] gives a result which is similar, but different.

THEOREM (Satz 9) *For every twosided infinite overlap-free word* $\mathbf{x}$, *there exists a unique infinite overlap-free word* $\mathbf{y}$ *such that* $\mathbf{x} = \mu(\mathbf{y})$.

[1] The mention Satz n refers to theorem n in [45]

This gives, in some sense, a complete description of the set of overlap-free twosided infinite words; indeed, it means that this set is a *minimal* set. More precisely, recall that a *dynamical system* is a set X of infinite words that is closed for the shift operator, defined by $T(\mathbf{x})(n) = \mathbf{x}(n+1)$, and that is closed for the usual topology on infinite words. It is not difficult to show that $\mathbf{x}$ is in X iff $\mathrm{Fact}(\mathbf{x}) \subset \mathrm{Fact}(X)$, where $\mathrm{Fact}(X)$ is the set of finite words that are factors of some element in X. A dynamical system X is minimal if it does not contain strictly any other dynamical system. This means that X is equal to the dynamical system generated by any of its elements, and also that $\mathrm{Fact}(\mathbf{x}) = \mathrm{Fact}(X)$ for any $\mathbf{x} \in X$.

The property that the dynamical system generated by the (twosided) Thue-Morse sequence is minimal was explicitly proved by Gottschalk and Hedlund [16]. As a consequence, every factor appears with bounded gaps (is *recurrent*, in the terminology of M. Morse [25]). Axel Thue (Satz 11) only mentions that every factor appears infinitely many often.

The structure of onesided infinite overlap-free words is more complicated. Axel Thue was interested in the tree of infinite overlap-free words and tried to to characterize those overlap-free words which can be extended into infinite overlap-free words. His main result in this direction is

THEOREM (Satz 15) *Let w be an overlap-free word of w length n such that there exist words u and v of length $8n$ with the property that uwv is still overlap-free. Then any overlap-free word x of length $26n$ contains w as a factor.*

In the proof of this result, he shows that the word x contains a Morse bloc which contains w, and he concludes that w is indefinitely extensible in both directions. An explicit description of the tree of infinite overlap-free words by means of a finite automaton was given by E. D. Fife and deserves a mention.

Fife defines three operators on words, say α, β, γ, and he shows that every overlap-free infinite words is the "value" of some infinite word $\mathbf{f}$ in the three operators, provided the word $\mathbf{f}$ is in some rational set he gives explicitely. To be more precise, let $X_n = \{u_n, v_n\}$ be the set of Morse blocs of index n and let $X = \bigcup_{n \geq 0} X_n$. Any word $w \in A^*X_1$ admits a *canonical decomposition* $(z, y, \bar{y})$ where y is the longest word in X such that $w = zy\bar{y}$. It is equivalent to say that $(z, y, \bar{y})$ is the canonical decomposition of w if $\bar{y}y$ is not a suffix of z. As an example, the canonical decomposition of *aabaabbabaab* is

$$(aaba, abba, baab)$$

and the decomposition of *abaabbaababbaabbabaab* is

$$(abaab, baababba, abbabaab)$$

The three functions $\alpha, \beta, \gamma : A^*X_1 \to A^*X_1$, acting on the right, are defined as follows for a word $w \in A^*X_1$ with canonical decomposition $(z, y, \bar{y})$:

$$w \cdot \alpha \quad = zy\bar{y} \cdot \alpha = zy\bar{y}yy\bar{y} = wyy\bar{y}$$

$$w \cdot \beta = zy\bar{y} \cdot \beta = zy\bar{y}y\bar{y}\bar{y}y = wy\bar{y}\bar{y}y$$
$$w \cdot \gamma = zy\bar{y} \cdot \gamma = zy\bar{y}\bar{y}y = w\bar{y}y$$

Since w is a prefix of $w \cdot \alpha$, $w \cdot \beta$, and of $w \cdot \gamma$, it makes sense to define $w \cdot f$ by induction for all "words" f in B^*, with $B = \{\alpha, \beta, \gamma\}$. By continuity, $w \cdot \mathbf{f}$ is definde also for infinite words $\mathbf{f}$. Here are some examples:

$$ab \cdot \alpha = abaab$$
$$ab \cdot \beta = ababba$$
$$ab \cdot \gamma = abba$$
$$ab \cdot \gamma^\omega = \mathbf{t}$$
$$aab \cdot \alpha = aabaab = a(ab \cdot \alpha)$$
$$ab \cdot \alpha\beta\gamma = abaababbabaababbaabbabaab$$

Observe that the last word contains an overlap. Note also that, for $w \in A^*X_1$ and $f \in B^*$, one has $\mu(w \cdot f) = \mu(w) \cdot f = w \cdot \gamma f$. A *description* of an infinite word $\mathbf{x}$ starting with ab or aab is an infinite word $\mathbf{f}$ over B such that $\mathbf{x} = ab \cdot \mathbf{f}$ or $\mathbf{x} = aab \cdot \mathbf{f}$, according to $\mathbf{x}$ starts with ab or aab.

PROPOSITION *Every infinite overlap-free word starting with the letter a admits a unique description.*

Let

$$F = B^\omega - B^*IB^\omega$$

be the (rational) set of infinite words over B having no factor in the set

$$I = \{\alpha, \beta\}(\gamma^2)^*\{\beta\alpha, \gamma\beta, \alpha\gamma\}$$

and let G bet the set of words $\mathbf{f}$ such that $\beta\mathbf{f}$ is in F. Then:

THEOREM (Fife's Theorem) *Let $\mathbf{x}$ be an infinite word over $A = \{a, b\}$.*

(*i*) *if $\mathbf{x}$ starts with ab, then $\mathbf{x}$ is overlap-free iff its description is in F;*

(*ii*) *if $\mathbf{x}$ starts with aab, then $\mathbf{x}$ is overlap-free iff its description is in G.*

A direct consequence is the following

COROLLARY *An overlap-free word w is the prefix of an infinite overlap-free word iff w is a prefix of a word $ab \cdot f$ with $f \in W$ or of a word $aab \cdot f$ with $\beta f \in W$, where $W = B^* - B^*IB^*$.*

This implies in particular a result of Restivo et Salemi [30], namely that it is decidable whether an overlap-free word is extensible into an infinite overlap-free word. Another consequence of Fife's description is the following

COROLLARY *The Thue-Morse word $\mathbf{t}$ is the greatest infinite overlap-free word, in lexicographical order, that start with the letter a.*

Indeed, the choice of the letters α, β, et γ implies that if $\mathbf{f} \leq \mathbf{f}'$, then $ab \cdot \mathbf{f} \leq ab \cdot \mathbf{f}'$. The greatest word in F is γ^ω, and this shows the corollary. A. Carpi [8]

has developed a description for finite overlap-free words by means of a finite automaton. Unfortunately, his automaton is rather big (more than 300 states).

There is another property that singles out the Thue-Morse word (and which was rediscovered and generalized by P. Séébold [37]). Call a morphism *overlap-free* if the image of an overlap-free word is always overlap-free.

THEOREM (Satz 16) *Let h be an overlap-free morphism. Then there is an integer n such that $h = \mu^n$ or $h = \pi \circ \mu^n$, where π is the morphism that exchanges the two letters of the alphabet.*

Thus, the infinite words $\mathbf{t}$ and $\bar{\mathbf{t}}$ are the only infinite overlap-free words generated by iterated morphisms.

Since overlap-free words have a strong structure, it seems natural to count them. The first result is due to Restivo and Salemi [30]. They prove that the number γ_n of overlap-free words over two letters grows polynomially in n (in fact slower than n^4). Kobayashi [21] has used Fife's theorem to derive the lower of the more precise bounds for γ_n :

THEOREM *There are constants C_1 and C_2 such that*

$$C_1 n^{\alpha} < \gamma_n < C_2 n^{\beta}$$

where $\alpha = 1.155\ldots$ and $\beta = 1.5866\ldots$.

One might ask what is the "real" limit. In fact, a recent and surprising result by J. Cassaigne [10] shows that there is no limit. More precisely, set

$$\alpha' = \sup\{r \mid \exists C > 0, \forall n, \gamma_n \geq C n^r\}$$

and

$$\beta' = \sup\{r \mid \exists C > 0, \forall n, \gamma_n \leq C n^r\}$$

Then

THEOREM *One has $1.155 < \alpha' < 1.276 < 1.332 < \beta' < 1.587$.*

This is to be compared with the situation for square-free words. Indeed, Brandenburg [5] proved that for the number $c(n)$ of square-free words of length n over three letters, there are constants $c_1 \geq 1.032$ and $c_2 \leq 1.38$ such that $6c_1^n < c(n) < 6c_2^n$. Brandenburg also proves that the number of cube-free words over two letters grows exponentially.

4 Square-free words

4.1 First examples

It is easily seen that the only square-free words over two letters are a, b, ab, ba, aba, bab. However, there exist arbitrarily long square-free words over three letters, and by a simple argument, there exist infinite square-free words over

three letters. Historically the first infinite square-free word was given by Thue in his 1906 paper. It is over four letters, and it is obtained by iterating the following morphism h, starting with the letter a:

$$\begin{aligned} a &\mapsto adbcb \\ b &\mapsto abdcb \\ c &\mapsto abcdb \\ c &\mapsto abcbd \end{aligned}$$

Thue explain his construction as follows : take a square-free word over three letters, here $abcb$, and interleave it with the letter d. This gives the morphism. The proof is not very difficult.

In the same paper, Thue gives another infinite square-free word, over three letters. The word is by iterating the following construction: given a square-free word w over $A = \{a, b, c\}$, build $\alpha(w)$ by replacing each letter a by $abac$, each b by $babc$, and each c either by $bcac$ or by $acbc$, according to the letter preceding c in w is a or b. Starting with a, one gets an infinite word

$$abacbabcabacbcacbabcabacbabcacbcabacbabc\cdots$$

which he shows to be square-free. Although the definition is not by a morphism, the construction is very close to it. There exist several ways to formulate it differently: in fact, one has a fourth letter hidden in the description, which appears when we note differently a letter c precede by an a and a letter c preceded by a b. The four letter word thus obtained is generated by a morphism, and at the end, the two variants of the letter c are identified.

In the 1912 paper, Axel Thue gives a morphism for generating an infinite square-free word over three letters. The morphism is the following (Satz 18):

$$\begin{aligned} a &\mapsto abcab \\ b &\mapsto acabcb \\ c &\mapsto acbcacb \end{aligned}$$

This morphism seems to be rather complicated. Its *size*, i. e. the sum of the length of the images, is 18. It has been shown by A. Carpi [7] that this is the best bound : every morphism over three letters that preserves square-free words has size at least 18. (See also the discussion in [4].) However, there is a simpler morphis that generates a square-free word (starting with a) given e.g. by Hall [18], namely

$$\begin{aligned} a &\mapsto abc \\ b &\mapsto ac \\ c &\mapsto b \end{aligned}$$

This mophism does not preserve square-free words, because the image of aba is $abcacabc$.

4.2 A Classification

Since every twosided infinite square-free word $\mathbf{x}$ over three letters a, b and c is some product of the six words in the set

$$X = \{ab, abc, abcb, ac, acb, acbc\}$$

Thue studies a classification according to words of X that appear in $\mathbf{x}$. It is quite remarkable that he achieves a classification of those square-free infinite words that contain exactly four of the six words in X. After some discussion, he reduces the 15 cases (two words lacking among six) to the following three cases :

$$aca \text{ and } bcb \qquad \text{I}$$

$$aba \text{ and } aca \qquad \text{II}$$

$$aba \text{ and } bab \qquad \text{III}$$

are missing in the infinite word under consideration. In order to describe these three families, he gives some "parametrization", and as we will see, reduces them to minimal dynamical systems.

Consider first square-free words of type (I), i. e. without occurrences of aca or bcb. Define a morphism h from $A = \{a, b, c\}$ into $B = \{\alpha, \beta\}$ by

$$\begin{aligned} a &\mapsto \alpha \\ b &\mapsto \alpha\beta\beta \\ c &\mapsto \alpha\beta \end{aligned}$$

Then the following holds

THEOREM (Satz 20,21) *If $\mathbf{x}$ is a square-free infinite word of type* (I)*, then $h(\mathbf{x})$ is overlap-free. Conversely, for every overlap-free word $\mathbf{y}$, there exists a unique word $\mathbf{x}$ such that $h(\mathbf{x}) = \mathbf{y}$, and $\mathbf{x}$ is square-free of type* (I).

Thus, the square-free words of type (I) are describe by the (minimal) set of overlap-free words over two letters. For the two other types, the situation is slightly more involved (and the proofs are more difficult). First, Thue observes that the cases (II) and (III) reduce one to each other. Any word $\mathbf{x}$ of type (II) is uniquely decomposable as a product of words in the set $\{ca, cb, cab, cba\}$. Let s be the substitution defined by

$$\begin{aligned} ca &\mapsto abc \\ cb &\mapsto acb \\ cab &\mapsto abcb \\ cba &\mapsto acbc \end{aligned}$$

For a word $\mathbf{x}$ of type (II), the word $y = s(\mathbf{x})$ is of type (III), and conversely, every word $\mathbf{y}$ of type (III) is of this form. Thus, it suffices to describe square-free words of type (II). For this, Thue introduces a new, five letter alphabet $\{A, B, C, D, E\}$, and a morphism $h : \{A, B, C, D, E\}^* \to \{a, b, c\}^*$ defined by

$$\begin{aligned} A &\mapsto abcbacbcacbabcacbc \\ B &\mapsto abcbacbc \\ C &\mapsto abcbacbcacb \\ D &\mapsto abcbabcacbc \\ E &\mapsto abcbabcacb \end{aligned}$$

Finally, he defines a set of words

$$\begin{aligned} W = \{&AB, AD, BA, BC, CA, CD, CE, DB, DE, EC, ED, \\ &BEB, EBE, DAC, DCBD, CBDC\} \end{aligned}$$

In order to state simply the next theorem, let us denote by $\mathcal{Y}$ the set of twosided infinite square-free words over the five letter alphabet $\{A, B, C, D, E\}$ that have no factors in W. Then Thue proves

THEOREM (Satz 26) *The set of twosided infinite square-free words of type (II) is the set of words of the form $h(\mathbf{y})$ for $\mathbf{y}$ in $\mathcal{Y}$.*

This theorem seems to be a little disappointing, since a rather simple description of three letter square-free words is replaced by a cumbersome and complicated family $\mathcal{Y}$ of words over five letters. However, this family has an important property: let α be the morphism from $\{A, B, C, D, E\}^*$ into itself defined by

$$\begin{aligned} A &\mapsto BDAEAC \\ B &\mapsto BDC \\ C &\mapsto BDAE \\ D &\mapsto BEAC \\ E &\mapsto BEAE \end{aligned}$$

THEOREM (Satz 23,24) *The morphism α is a bijection of the set $\mathcal{Y}$ onto itself.*

Thus, as before, the set $\mathcal{Y}$ is a minimal dynamical system.

A full description of the tree of square-free words, like Fife's description for overlap-free words, is not yet available. Shelton and Soni have investigated this tree [39, 40, 41]. They have shown in particular that the set of infinite square-free words over three letters is perfect. Roughly speaking, this means if $\mathbf{x}$ is any square-free infinite word, then for any prefix p of $\mathbf{x}$, there are infinitely many infinite square-free words that have p as prefix. They show also that it is

decidable whether a square-free word p of length n is a prefix of some infinite square-free word, and their procedure is "uniform": There is a constant K such that if there exists a word q of length $n + Kn^{3/2}$ such that pq is square-free, then p is the prefix of some infinite square-free word.

4.3 Repetitions

As already mentioned, Thue calls a word on n letters *irreducible* if every factor xyx verifies $|y| \geq n - 2$. A more general concept, first considered by F. Dejean [14], is to require that the length of the word y separating the occurrences of x is bounded from below by the length of x (times some factor). More precisely, we call *repetition* a word xyx with x non empty, and *index* of this repetition the quotient $|y|/|x|$. We are looking for words where all repetitions have high index. F. Dejean has proved that there exists an infinite word over 3 letters that has only repetitions of index greater or equal to $1/3$, and she also shows that this bound is the best possible. Call *repetition threshold* the smallest number ρ_k such that there exists an infinite word over k letters that has only repetitions of index greater or equal to ρ_k. Thus, Dejean's result may be stated as : $\rho_3 = 1/3$. She conjectured that $\rho_4 = 3/2$, a result proved by Pansiot [27], and that $\rho_k = k - 2$ for $k \geq 5$. The conjecture was proved up to 9 by Moulin-Ollagnier [26]

5 Avoidable patterns

The overlap-freeness of the Thue-Morse sequence, and the square-freeness of the other words we have presented can be expressed in the more general framework of avoidable and unavoidable patterns in strings. This concept has been introduced in the context of equations defining algebras. Certain unavoidable words have been used e.g. in [35] to characterize those finite semigroups S that are inherently nonfinitely based, in the sense that S is not a member of any locally finite semigroup variety definable by finitely many equations. It may be noticed that Axel Thue replaces his research on repetitions in strings in an even slightly more general context, since he considers avoiding patterns with constants. However, he has not stated results in this specific framework.

Consider an alphabet E of "pattern symbols". A word e over E is a pattern. A pattern e is said to *occur* in some word $w \in A^*$ if ther is a nonerasing morphism $h : E^* \to A^*$ such that $h(e)$ is a factor of w. A pattern e is *avoidable* over k letters, or is k-avoidable, if there is an infinite word $\mathbf{x}$ over k letters such that e does not occur in $\mathbf{x}$. The Thue-Morse sequence shows that the patterns aaa and $ababa$ are 2-avoidable, and square-free infinite words show that aa is 3-avoidable (but not 2-avoidable). Avoidable and unavoidable patterns have been studied by several people (Zimin [48], Schmidt [36], Bean, Ehrenfeucht, McNulty [4], Roth [31], Cassaigne [9], Goralcik, Vanicek [17], Baker, McNulty,

Taylor [2], Crochemore, Goralcik [13]).

Problems which have been stated, and partially solved, include the following: given a pattern e, is it avoidable or not ? There is a nice algorithm in [4], and basically the same in [48], to decide whether a pattern is avoidable. The complexity of their algorithm is at least exponential. P. Roth (personal communication) recently has proved that the general problem is NP-complete.

For a pattern e, denote by $\alpha(e)$ the number of distinct letters occurring in e. Every pattern e such that $|e| \geq 2^{\alpha(e)}$ is in fact avoidable, and this is the best possible bound because there exists un unavoidable pattern of length $2^n - 1$ over an n letter pattern alphabet. The next problem is to determine, for some unavoidable pattern e, the size $\mu(e)$ of the smallest k such that e is k-avoidable. The first word that is 4-avoidable but not 3-avoidable has been given by [2]. Upper bounds of μ, as a function of α are also given there. Recently, Roth [31], Cassaigne [9], Goralcik, Vanicek [17] have solved the problem of determining all the 2-avoidable binary patterns.

References

[1] A. Adler, S. Li, Magic cubes and Prouhet sequences, *American Math. Monthly* (1977), 618–627.

[2] K. A. Baker, G. F. McNulty, W. Taylor, Growth problems for avoidable words, *Theoret. Comput. Sci.* **69** (1989), 319–345.

[3] L. Baum, M. Sweet, Continued fractions of algebraic power series in characteristic 2, *Ann. Math.* **103** (1976), 593–610.

[4] D. R. Bean, A. Ehrenfeucht, G. F. McNulty, Avoidable patterns in strings of symbols, *Pacific J. of Math.* **85** (1979), 261–294.

[5] F. J. Brandenburg, Uniformely growing k-th powerfree homomorphisms, *Theoret. Comput. Sci.* **23** (1983), 69–82.

[6] S. Brlek, Enumeration of factors in the Thue-Morse word, *Discr. Appl. Math.* **24**, (1989), 83–96.

[7] A. Carpi, On the size of a squarefree morphism on a three letter alphabet, *Inform. Proc. Letters* **16** (1983), 231–236.

[8] A. Carpi, Overlap-free words and finite automata, manuscrit, 1990.

[9] J. Cassaigne, Unavoidable binary patterns, manuscrit Ecole normale supérieure, Paris, 1991, submitted.

[10] J. Cassaigne, Counting overlap-free binary words, manuscrit Ecole normale supérieure, Paris, 1992.

[11] G. Christol, T. Kamae, M. Mendès France, G. Rauzy, Suites algébriques, automatset substitutions, *Bull. Soc. Math. France* **108** (1980), 401–419.

[12] A. Cobham, Uniform tag sequences, *Math. Systems Theory* **6** (1972), 164–192.

[13] M. Crochemore, P. Goralcik, Mutually avoiding ternary words of small exponents, manuscript, 1991, submitted.

[14] F. Dejean, Sur un théorème de Thue, *J. Combin. Th. A* **13** (1972), 90–99.

[15] E. D. Fife, Binary sequences which contain no BBb, *Trans. Amer. Math. Soc.* **261** (1980), 115–136.

[16] W.H. Gottschalk, G.A. Hedlund, A characterization of the Morse minimal set, *Proc. Amer. Math. Soc.* **15** (1964), 70–74.

[17] P. Goralcik, T. Vanicek, Binary patterns in binary words, manuscript, 1991, submitted.

[18] M. Hall, Generators and relations in groups – the Burnside problem, in T. L. Saaty (ed) *Lectures on Modern Mathematics* **2**, Wiley, 1964, 42–92.

[19] T. Harju, On cyclically overlap-free words in binary alphabets, *The Book of L*, Springer-Verlag, 1986, 123–130.

[20] G.A. Hedlund, Remarks on the work of Axel Thue, *Nordisk Mat. Tidskr.* **15** (1967), 148–150.

[21] Y. Kobayashi, Enumeration of irreductible binary words, *Discrete Appl. Math.***20** (1988), 221–232.

[22] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, 1983.

[23] J. H. Loxton, A. J. van der Poorten, Arithmetic properties of the solutions of a class of functional equations, *J. reine angew. Math.* **330** (1982), 159–172.

[24] A. A. Markov, Imossibility of certain algorithms in the theory of associative systems *Dokl. Akad. Nauk. SSSR* **55** (1941), 587–590.

[25] M. Morse, Recurrent geodesics on a surface of negative curvature, *Transactions Amer. Math. Soc.* **22** (1921), 84–100.

[26] J. Moulin-Ollagnier, Preuve de la conjecture de Dejean pour des alphabets à 5, 6, 7, 8, 9 lettres, Prépublication du département de mathématique et informatique, Université Paris-Nord, Nr. 89–4, 1989.

[27] J.-J. Pansiot, A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Discrete Appl. Math.***7** (1984), 297–311.

[28] E. L. Post, Recursive unsolvability of a problem of Thue, *J. Symbolic Logic* **11** (1947), 1–11.

[29] M. E. Prouhet, Mémoire sur quelques relations entre les puissances des nombres, *C. R. Acad. Sci. Paris.* **33** (1851), 31.

[30] A. Restivo, S. Salemi, Overlap-free words on two symbols, in: *Automata on infinite words*, Nivat, Perrin (eds), Lect. Notes Comp. Sci.,**192**, Springer-Verlag, 1985, 198–206.

[31] P. Roth, Every binary pattern of length six is avoidable on the two-letter alphabet, *Acta Informatica* (1992).

[32] G. Rozenberg, A. Salomaa, *The Mathematical Theory of L-Systems*, Academic Press, 1980.

[33] W. Rudin, Some theorems on Fourier coefficients, *Proc. Amer. Math. Soc.* **10** (1959), 855–859.

[34] A. Salomaa, *Jewels of Formal Language Theory*, Computer Science Press, 1981.

[35] M. Sapir, Inherently nonfinitely based finite semigroups, *Mat. Sb.* **133** (1987), 154–166.

[36] U. Schmidt, Avoidable patterns on two letters, *Theoret. Comput. Sci.* **63** (1989), 1–17.

[37] P. Séébold, Sequences generated by infinitely iterated morphisms, *Discrete Appl. Math.***11**, (1985), 255–264.

[38] H. S. Shapiro, Extremal problems for polynomials and power series, Thesis, M.I.T., 1951.

[39] R. Shelton, Aperiodic words on three symbols I, *J. Reine Angew. Math.* **321** (1981), 195–209.

[40] R. Shelton, R. Soni, Aperiodic words on three symbols II, *J. Reine Angew. Math.* **327** (1981), 1–11.

[41] R. Shelton, R. Soni, Aperiodic words on three symbols III, *J. Reine Angew. Math.* **330** (1982), 44–52.

[42] R. Shelton, R. Soni, Chains and fixing blocks in irreducible sequences, *Discrete Math.***54** (1985), 93–99.

[43] A. THUE, Über unendliche Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1906, Nr. 7.

[44] A. THUE, Die Lösung eines Spezialfalles eines generellen logischen Problems, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1910, Nr. 8.

[45] A. THUE, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1912, Nr. 10.

[46] A. THUE, Probleme über Veränderungen von Zeichenreihen nach gegebenen Regeln, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.*, Christiana 1914, Nr. 7.

[47] A. THUE, *Selected Mathematical Papers*, edited by T. Nagell, A. Selberg, S. Selberg, K. Thalberg, Universitetsforlaget, Oslo 1977.

[48] A. I. ZIMIN, Blocking sets of terms, *Math. USSR Sb.* **47**, (1984), 353–364.