Binary patterns in the Prouhet-Thue-Morse sequence

Jorge Almeida[1]* Ondřej Klíma[2†]

[1] *CMUP, Dep. Matemática, Faculdade de Ciências, Universidade do Porto, Portugal*
[2] *Dept. of Mathematics and Statistics, Masaryk University, Brno, Czech Republic*

received 15th May 2019, revised 27th July 2021, accepted 2nd Aug. 2021.

We show that, with the exception of the words a^2ba^2 and b^2ab^2, all (finite or infinite) binary patterns in the Prouhet-Thue-Morse sequence can actually be found in that sequence as segments (up to exchange of letters in the infinite case). This result was previously attributed to unpublished work by D. Guaiana and may also be derived from publications of A. Shur only available in Russian. We also identify the (finitely many) finite binary patterns that appear non trivially, in the sense that they are obtained by applying an endomorphism that does not map the set of all segments of the sequence into itself.

Keywords: Prouhet-Thue-Morse sequence, pattern, infinite word, special word

1 Introduction

Let μ be the endomorphism of the free semigroup $\{a, b\}^+$ defined by $\mu(a) = ab$ and $\mu(b) = ba$. Since a is a prefix of $\mu(a)$, $\mu^n(a)$ is also a prefix of $\mu^{n+1}(a)$. Hence, the sequence $(\mu^n(a))_n$ determines a sequence of letters, or infinite word, whose prefix of length 2^n is $\mu^n(a)$; we say that the infinite word $\mathbf{t}$ thus obtained is *generated* by μ. It is called the *Prouhet-Thue-Morse sequence* and it has been the object of extensive studies and applications. It was first considered by Prouhet (1851) in connection with a problem in number theory, five decades later by Thue (1906, 1912) to exhibit infinite words avoiding cubes and squares, and another two decades later by Morse (1921) as a discretized description of non-periodic recurrent geodesics in surfaces of negative curvature. See Allouche and Shallit (1999) for a survey on this topic, including several further connections with other branches of Mathematics. The first author and other collaborators have previously studied the sequence $\mathbf{t}$ in the framework of symbolic dynamics and its connections with free profinite semigroups (see Almeida and Costa (2013) and Almeida et al. (2020)). It was in fact an attempt to construct a profinite semigroup with certain properties that prompted this work, although no further references to profinite semigroups will be made in this paper.

*Work partially support by CMUP (UID/MAT/ 00144/2019), which is funded by FCT (Portugal) with national (MATT'S) and European structural funds (FEDER) under the partnership agreement PT2020. The work was carried out at Masaryk University, whose hospitality is gratefully acknowledged, with the support of the FCT sabbatical scholarship SFRH/BSAB/142872/2018.

†Work supported by Grant 19-12790S of the Grant Agency of the Czech Republic.

ISSN 1365–8050

This paper concerns the study of binary patterns of $\mathbf{t}$, that is, finite or infinite words w over the alphabet $\{a, b\}$ for which there exists an endomorphism φ of the semigroup $\{a, b\}^+$ (naturally extended to infinite words) such that the word $\varphi(w)$ can be found as a block of consecutive letters of $\mathbf{t}$ (which we call a *segment* of $\mathbf{t}$). Since we need to identify concrete finite segments of $\mathbf{t}$, a simple and efficient algorithm on how to compute them is presented in Section 2.

Characterizations of binary patterns of $\mathbf{t}$ are due to Shur (1996a) and D. Guaiana (unpublished work announced in Restivo and Salemi (2002b,a)). Our first main contribution is a proof of the characterization attributed to D. Guaiana (but also, independently, obtained by Shur (1997) in his thesis) using results from Shur (1996a): with the exception of a^2ba^2 and b^2ab^2, the binary patterns of $\mathbf{t}$ are its finite segments. Section 3 presents our proof of this result.

The endomorphism μ and the endomorphism ξ exchanging the letters a and b are easily seen to transform finite segments of $\mathbf{t}$ into other such segments. In Section 4, we consider the problem of determining which finite segments may only be transformed into other segments by endomorphisms of $\{a, b\}^+$ that may be obtained by composition of μ and ξ. Such words are said to be *typical* since we show that all but finitely many finite segments of $\mathbf{t}$ are typical. We further determine all atypical words. As an application of our results, we also determine all infinite binary patterns of $\mathbf{t}$.

We conclude the paper with Section 5, where we propose the investigation of the properties we established for $\mathbf{t}$ for arbitrary infinite words.

2 Segments of $\mathbf{t}$

By a *word* we always mean a finite sequence of letters of an alphabet A, that is a member of the free monoid A^*. A word u is a *factor* of a word v if there exist words x and y such that $v = xuy$. In spite of the terminology, an *infinite word* is not a word but rather an infinite sequence of letters.

Note that $\{ab, ba\}$ is a code, in the sense that it generates a free subsemigroup of $\{a, b\}^+$ and, therefore, μ is injective.

For an infinite word $w = a_1a_2\cdots$, by the *segments* of w we mean the words of the form $a_ka_{k+1}\cdots a_\ell$ with $k \leqslant \ell$ and the infinite words of the form $a_ka_{k+1}\cdots$. Note that, since $\mu^{n+1}(a) = \mu^n(a)\mu^n(b)$, all factors of the words $\mu^n(b)$ are segments of $\mathbf{t}$. It follows that a word $u \in \{a, b\}^+$ is a segment of $\mathbf{t}$ if and only if so is the word that is obtained from u by interchanging the letters a and b.

A word $w \in A^+$ is said to be *avoided* by $\mathbf{t}$ if there is no homomorphism $\varphi : A^+ \to \{a, b\}^+$ such that $\varphi(w)$ is a segment of $\mathbf{t}$. We also say that $w \in A^+$ is *unavoidable* in $\mathbf{t}$ if it is not avoided by $\mathbf{t}$; we then also say that w is a *pattern* of $\mathbf{t}$. For instance, it is well known that a^3 and $ababa$ are avoided by $\mathbf{t}$, which is also expressed by saying that $\mathbf{t}$ is, respectively, cube-free and overlap-free (Lothaire, 1983).

The preceding notions are extended to infinite words by saying how endomorphisms of $\{a, b\}^+$ are applied to infinite words. Given an infinite word $w = a_1a_2\cdots$ over the alphabet $\{a, b\}$ and an endomorphism φ of $\{a, b\}^+$, we let $\varphi(w)$ be infinite word obtained by concatenating the $\varphi(a_i)$: $\varphi(w) = \varphi(a_1)\varphi(a_2)\cdots$.

For a nonempty word u, let $t_1(u)$ denote its last letter.

The computation of the segments of $\mathbf{t}$ may be carried out easily in view of the following proposition. The first part is an improved version of Shur (1996b, Corollary 1), although the same conclusion is in fact already established in the proof of the cited statement. We present a proof for the sake of completeness.

Proposition 2.1 *A word w is a segment of $\mathbf{t}$ if and only if it is a factor of $\mu^n(a)$, where $n = 1$ if $|w| = 1$, $n = 3$ if $|w| = 2$, and $n = 2 + \lceil \log_2(|w| - 1) \rceil$ otherwise. Moreover, for every integer $k \geqslant 3$, the value $n = 2 + \lceil \log_2(k - 1) \rceil$ is minimum for $\mu^n(a)$ to admit as factors all segments of $\mathbf{t}$ of length k.*

Proof: Since $\mathbf{t}$ is cube-free, the cases where $|w| \leqslant 3$ are easily verified by inspection. Suppose that w is a segment of $\mathbf{t}$ which we may assume to be of length at least 4. Then, w is a factor of $\mu^n(a)$ for some positive integer n. Take n to be minimum with that property. If m is the minimum positive integer such that w is a factor of $\mu^m(x)$ for some letter x, then either only $x = b$ can play that role and $n = m + 1$, or $x = a$ may play it and $n = m$. We need to show, respectively, that $m \leqslant 1 + \lceil \log_2(|w| - 1) \rceil$ or $m \leqslant 2 + \lceil \log_2(|w| - 1) \rceil$. Since $|w| \geqslant 4$, we may assume that $m \geqslant 4$ for, otherwise, the inequality $m \leqslant 1 + \lceil \log_2(|w| - 1) \rceil$ holds trivially.

Let $\mu(x) = xy$. As $\mu^m(x) = \mu^{m-1}(xy)$ and m is minimum, there must be a nontrivial factorization $w = w_1 w_2$ with w_1 a suffix of $\mu^{m-1}(x)$ and w_2 a prefix of $\mu^{m-1}(y)$.

If one of the factors w_1 or w_2 has length greater than 2^{m-2}, then we must have $|w| > 2^{m-2} + 1$, which implies that $m \leqslant 1 + \lceil \log_2(|w| - 1) \rceil$ and fulfills our aim. Thus, we may assume that both w_1 and w_2 have length at most 2^{m-2}. Now, we have $\mu^m(x) = \mu^{m-2}(xyyx)$ and w is a factor of $\mu^{m-2}(yy) = \mu^{m-3}(yxyx)$. If w is a factor of either $\mu^{m-3}(xyx)$ or $\mu^{m-3}(yxy)$, then it is also a factor of $\mu^{m-2}(xx) = \mu^{m-3}(xyxy)$ and, therefore, also of $\mu^m(y)$, so that we are in the case $n = m$. On the other hand, by the minimality of m the word w cannot be a factor of $\mu^{m-3}(xy) = \mu^{m-2}(x)$ or $\mu^{m-3}(yx) = \mu^{m-2}(y)$, and so we have $|w| \geqslant 2^{m-3} + 2$, which implies that $n = m \leqslant 2 + \lceil \log_2(|w| - 1) \rceil$, as claimed. It remains to consider the case where w is a factor of neither $\mu^{m-3}(xyx)$ nor $\mu^{m-3}(yxy)$. Then there must be a factorization $w = s\mu^{m-3}(xy)z$ with s and z nontrivial words, so that $|w| \geqslant 2^{m-2} + 2$, which yields $m \leqslant 1 + \lceil \log_2(|w| - 1) \rceil$. This completes the proof of the first part of the proposition.

To prove the last part of the proposition, first note that, for $k \geqslant 3$, the value of $f(k) = 2 + \lceil \log_2(k - 1) \rceil$ is at least 3. We claim that, for $n \geqslant 3$, there is a word of length $2^{n-3} + 2$ that is a segment of $\mathbf{t}$ but not a factor of $\mu^{n-1}(a)$. Noting that $f(2^{n-3} + 2) = n$, the result follows.

To establish the claim, consider the word $w = t_1(\mu^{n-3}(b))\mu^{n-3}(a)b$. It is a segment of $\mathbf{t}$, in fact a factor of $\mu^n(a) = \mu^{n-2}(a)\mu^{n-3}(b) \cdot \mu^{n-3}(a) \cdot \mu^{n-1}(b)$ since b is the first letter of $\mu^{n-1}(b)$. It remains to show that w is not a factor of $\mu^{n-1}(a) = \mu^{n-3}(a)\mu^{n-3}(b)\mu^{n-3}(b)\mu^{n-3}(a)$. Otherwise, since there are no overlaps in $\mathbf{t}$, w must be a factor of $\mu^{n-3}(bb)$. For $n = 3$, this is clearly impossible since not even $\mu^{n-3}(a) = a$ is a factor of $\mu^{n-3}(bb) = bb$. For $n > 3$, we have

$$\mu^{n-3}(a) = \mu^{n-4}(a)\mu^{n-4}(b) \tag{1}$$

$$\mu^{n-3}(bb) = \mu^{n-4}(b)\mu^{n-4}(a)\mu^{n-4}(b)\mu^{n-4}(a). \tag{2}$$

Since there are no overlaps in $\mu^{n-3}(bb)$, the only place where $\mu^{n-3}(a)$ is found as a factor of $\mu^{n-3}(bb)$ is as the product of the two middle factors in the factorization given in (2). Hence, the word $w = t_1(\mu^{n-3}(b))\mu^{n-3}(a)b$ is not a factor of $\mu^{n-3}(bb)$ since, for instance, b is not the first letter of $\mu^{n-4}(a)$. □

For example, the segments of lengths 4 and 5 of the infinite word $\mathbf{t}$ are the factors of those lengths of $\mu^4(a) = abbabaabbaababba$. But, for instance, $aabb$ is a segment of $\mathbf{t}$ but not a factor of $\mu^3(a) = abbabaab$; it is precisely the segment considered in the last part of the proof of Proposition 2.1.

Throughout the remainder of the paper, when we need to check whether a concrete finite word is a segment of $\mathbf{t}$, without any further reference we simply apply the algorithm given by Proposition 2.1, which is linear in the length of the given word. We proceed similarly when we need to compute all the segments of $\mathbf{t}$ of a given length.

Now, we take into account also the dual version of Proposition 2.1, where $\mu^n(b)$ is considered instead of $\mu^n(a)$, which is a direct consequence of Proposition 2.1 using the fact that the set of all segments of $\mathbf{t}$ of fixed length k is closed under taking images under ξ. Since every segment of $\mathbf{t}$ of length $2^{n+1}-1$ must contain the factor $\mu^n(a)$ or $\mu^n(b)$, it follows from Proposition 2.1 that, for $k \geqslant 3$, every segment of length k of $\mathbf{t}$ is a factor of every segment of $\mathbf{t}$ of some length ℓ which is at most

$$2^{2+\lceil\log_2(k-1)\rceil+1}-1 \leqslant 2^{4+\log_2(k-1)}-1 = 16k-17.$$

The existence of such an ℓ is the property known as *uniform recurrence* of $\mathbf{t}$ and holds for every sequence generated by iterating a primitive endomorphism of a free semigroup (Queffélec, 2010, Proposition 5.2). In the case of the Prouhet-Thue-Morse sequence, the optimum value of ℓ is presented in Allouche and Shallit (2003, Example 10.9.3): for $k \geqslant 3$, we have $\ell = 9 \cdot 2^r + k - 1$, where r is the integer determined by the inequalities $2^r + 2 \leqslant k \leqslant 2^{r+1}+1$. Note that using the first inequality determining r, one gets the upper bound $\ell \leqslant 10k-19$, which is better than our rough upper bound $\ell \leqslant 16k-17$.

3 Finite binary patterns

The following result plays a key role below.

Theorem 3.1 (Shur (1996a)) *The set of words of $\{a,b\}^+$ that are avoided by $\mathbf{t}$ is the fully invariant ideal generated by the set*

$$\{a^3, ababa, a^2ba^2b, ab^2ab^2, t_1(\mu^k(a))\mu^k(aba)a \ (k \geqslant 1), t_1(\mu^m(a))\mu^m(bab)a \ (m \geqslant 2)\}.$$

Moreover, the above is a minimal generating set for the fully invariant ideal of the words avoided by $\mathbf{t}$.

The generators corresponding to $k=1$ and $k=2$ are, respectively, bab^2a^2ba and $a\,ab^2a\,ba^2b\,ab^2a\,a$; the generator corresponding to $m=2$ is $a\,ba^2b\,ab^2a\,ba^2b\,a$ while, for $m=1$, the word given by $t_1(\mu^m(a))\mu^m(bab)a = b^2a^2b^2a^2$ is avoided by $\mathbf{t}$ but may be obtained for instance from the generator a^2ba^2b by mapping a to b and b to a^2.

Another useful ingredient in our arguments is the following “synchronization” result.

Lemma 3.2 (de Luca and Varricchio (1989, Lemma 3.9)) *Let $X = \{ab, ba\}$ and consider $s \in X^+$ with $|s| \geqslant 4$. If u and v are words such that $usv \in X^+$ and $|u|$ is odd then usv has an overlap.*

Since $\mathbf{t}$ has no overlaps, we conclude that $\mathbf{t} = usv$, with s as in Lemma 3.2 and u a finite word, then u has even length.

Corollary 3.3 *If there is a factorization $\mathbf{t} = u\mu^{n+1}(x)v$ where $u \in \{a,b\}^*$, $x \in \{a,b\}$, and $n \geqslant 0$, then $u = \mu^n(w)$ and $v = \mu^n(z)$ for some word w and infinite word z.*

Proof: We proceed by induction on n, the case $n=0$ being trivial as μ^0 is interpreted to be the identity function. Suppose that $n \geqslant 1$ and, by symmetry, assume that $x=a$. Since $\mu^{n+1}(a) = \mu^n(a)\mu^n(b) = \mu^{n-1}(abba)$, by the induction hypothesis we know that $u = \mu^{n-1}(w)$ and $v = \mu^{n-1}(z)$, for some finite word w and infinite word z, and so $\mathbf{t} = wabbaz$. Lemma 3.2 then implies that w and z belong to the image of μ. Hence, u and v belong to the image of μ^n. □

We say that a segment u of $\mathbf{t}$ is *special* if both ua and ub are segments of $\mathbf{t}$. The special segments of $\mathbf{t}$ have been investigated by de Luca and Varricchio (1989) with the purpose of counting the number of segments of each given length. For our purposes, it suffices to observe the following much simpler result.

Lemma 3.4 (de Luca and Varricchio (1989, Lemma 3.6)) *If the word w is special, then so is $\mu(w)$.*

We say that two words are *suffix comparable* if at least one of them is a suffix of the other. The following lemma is the core of our arguments.

Lemma 3.5 *Suppose that u is a finite word such that ua is unavoidable in $\mathbf{t}$ and ub is a segment of $\mathbf{t}$ but ua is not. If $n \geqslant 2$ and u is suffix comparable with $\mu^n(a)$, then u is also suffix comparable with $\mu^{n+1}(b)$.*

Proof: Since $\mu^n(a)$ is a suffix of $\mu^{n+1}(b) = \mu^n(b)\mu^n(a)$, we may assume that $\mu^n(a)$ is a suffix of u, say $u = u_1\mu^n(a)$. Consider a concrete occurrence of ub in $\mathbf{t}$: $\mathbf{t} = u_0ubv$, where $u_0 \in \{a,b\}^*$. Since $\mathbf{t}$ is recurrent, we may assume that $|u_0| \geqslant 2^n$. By Corollary 3.3, we know that $u_0u_1 = \mu^{n-1}(u')$ and $bv = \mu^{n-1}(v')$ for some word u' and infinite word v' (cf. Figure 1). Since $\mu^{n-1}(v')$ starts with b, so does v'.

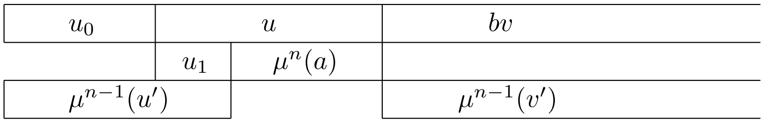

Fig. 1: Some segments of $\mathbf{t}$

Suppose first that u' ends with the letter b and $|u| > 3 \cdot 2^{n-1}$. If u' ends in ab then the word $t_1(\mu^{n-1}(a))\mu^{n-1}(bab)a$ is a suffix of ua which, in view of Theorem 3.1, contradicts the assumption that ua is unavoidable in $\mathbf{t}$. On the other hand, if u' ends with b^2 then, taking into account that bv starts with $\mu^{n-1}(b)$, we conclude that $\mathbf{t} = \mu^{n-1}(u''b^2ab^2v'')$ for some finite word u'' and infinite word v''. Since $\mathbf{t}$ is a fixed point of the injective endomorphism μ, it follows that b^2ab^2 is a segment of $\mathbf{t}$, which we know is not the case.

If $|u| \leqslant 3 \cdot 2^{n-1}$, then u is suffix of $\mu^{n-1}(xab)$ for a letter x. By Lemma 3.4, as it is easy to check that xab is special, so is $\mu^{n-1}(xab)$. Hence, u is special, contradicting the assumption that ua is not a segment of $\mathbf{t}$.

Thus, u' must end with ba, so that u_0u ends with $\mu^{n+1}(b)$. Since both u and $\mu^{n+1}(b)$ are suffixes of u_0u, they must be suffix comparable, thereby concluding the proof of the lemma. □

Similarly, one can prove the following lemma.

Lemma 3.6 *Suppose that u is a finite word such that ua is unavoidable in $\mathbf{t}$ and ub is a segment of $\mathbf{t}$ but ua is not. If $n \geqslant 2$ and u is suffix comparable with $\mu^n(b)$ then u is also suffix comparable with $\mu^{n+1}(a)$.*

Shur also observed in Shur (1996a) that the word a^2ba^2 (and, therefore, also b^2ab^2) is unavoidable in $\mathbf{t}$ but it is not a segment of $\mathbf{t}$. Our first main result is that there are no other such examples, thus providing an alternative characterization of two-letter words unavoidable in $\mathbf{t}$.

According to Restivo and Salemi (2002a, Theorem 2) and Restivo and Salemi (2002b, Theorem 3), the following theorem, which is considered to be surprising, was first proved by D. Guaiana in 1996 but, through private communication with A. Restivo, we learned that the proof was never published and the manuscript appears to be lost. On the other hand, we later learned from A. M. Shur that the next theorem also appears in his Ph.D. thesis (1997), which has never been published other than as a document in the Russian State Library. Moreover, Shur observed that the result can also easily be drawn from Theorem 3.1 using a characterization of the finite words on the alphabet $\{a, b\}$ that are not segments of $\mathbf{t}$, which is given in Shur (2005, Statement 1), a paper also in Russian. Since all the proofs seem to be either lost or somewhat inaccessible in the Russian literature, again we present our own proof for the sake of completeness.

Theorem 3.7 *A word $w \in \{a, b\}^+$ is unavoidable in $\mathbf{t}$ if and only if it is one of the words a^2ba^2 and b^2ab^2, or it is a segment of $\mathbf{t}$.*

Proof: We proceed by induction on the length of w. In view of Theorem 3.1, it is easy to check that the theorem holds for words of length at most 5. Assuming inductively that the result holds for words of length n, let w be a word of length $n + 1 \geqslant 6$ that is unavoidable in $\mathbf{t}$. Since interchanging the letters a and b does not affect either of the properties of being unavoidable in $\mathbf{t}$ and being a segment of $\mathbf{t}$, we may as well assume that a is the last letter of w.

Let $w = ua$. Since w is unavoidable in $\mathbf{t}$, so is u. Hence, by induction hypothesis, u may be found somewhere as a segment of $\mathbf{t}$. Take such an occurrence of u in $\mathbf{t}$ and let x be the letter immediately after it. We wish to show that there is such an occurrence of u in $\mathbf{t}$ with $x = a$. Aiming at a contradiction, we may assume that there is no such occurrence, that is, we always have $x = b$. Since $\mathbf{t}$ is recurrent, the segment ux may be found in $\mathbf{t}$ as far as we wish, so that we may continue prolonging it on the left as much as may be convenient. Thus, we are assuming that ua is unavoidable in $\mathbf{t}$ and that ub is a segment of $\mathbf{t}$ as long as desired but ua is not a segment of $\mathbf{t}$. However, we have to be careful because there is in principle no assurance that such an extension of u to the left retains the property that ua is unavoidable in $\mathbf{t}$.

Since a^3 is avoided by $\mathbf{t}$ and we are assuming that $x = b$, u cannot end with b^2. We distinguish several cases according to the termination of the word u.

If u ends with b then, by the above, it ends with ab. Suppose, more precisely, that u ends with bab. Since w ends with $baba$ and $ababa$ is avoided by $\mathbf{t}$, in fact u must end with b^2ab. This situation is impossible since we know that the suffix b^2ab^2 of ub is not a segment of $\mathbf{t}$.

Alternatively, assuming that u ends with b, it must end with $ba^2b = \mu^2(b)$. We may then apply successively Lemmas 3.5 and 3.6 to deduce that there is some $n \geqslant 2$ such that u is a suffix of $\mu^n(a)$. By Lemma 3.4, it follows that u is special, which contradicts the assumption that ua is not a segment of $\mathbf{t}$.

The next case we consider is that where u ends with aba. Note that u cannot end with a^2ba for, otherwise, $w = ua$ ends with ba^2ba^2 and, therefore, it cannot be unavoidable in $\mathbf{t}$. Also, u cannot end with $baba$ since $babab$ is not a segment of $\mathbf{t}$. Hence, aba is not a suffix of u.

Thus, assuming that u ends with ba, it must end with $ab^2a = \mu^2(a)$. We are then again led to a contradiction as above using Lemmas 3.5, 3.6 and 3.4. □

4 Typical finite binary patterns

Recall that the endomorphism of $\{a, b\}^+$ switching the letters a and b is denoted ξ. Since the finite segments of $\mathbf{t}$ are the finite words that are factors of $\mu^n(a)$ for all sufficiently large n and $\mu^n(a)$ is a factor of $\mu^{n+1}(b)$, we may replace a by b in that characterization of the segments of $\mathbf{t}$. Moreover, as ξ commutes with μ, we conclude that the set of finite segments of $\mathbf{t}$ is closed under applying the substitutions ξ and μ. By induction on n, the words $\mu^{2n}(a)$ are palindromic, in the sense that they coincide with the words read in the reverse order; this entails the well known fact that the set of segments of $\mathbf{t}$ is closed under reversal.

We say that a word $w \in \{a,b\}^+$ is *atypical* if it is a segment of $\mathbf{t}$ and there is an endomorphism φ of $\{a,b\}^+$ such that $\varphi(w)$ is also a segment of $\mathbf{t}$ and φ is not of one of the forms μ^n or $\xi \circ \mu^n$ with $n \geqslant 0$. Segments of $\mathbf{t}$ that are not atypical are said to be *typical*.

We say that a word is a *variant* of another word w if it may be obtained from w by applying reversal or ξ or both. Note that the set of atypical words is closed under taking factors and, by the above discussion, it is also closed under taking variants.

The following result appears explicitly as Brlek (1989, Proposition 3.3) but may already be extracted from Thue (1912) (see Berstel (1995, Chapter 3, Proposition 2.13)).

Proposition 4.1 *If u^2 is a segment of $\mathbf{t}$ then u is one of the words $\mu^n(a)$, $\mu^n(b)$, $\mu^n(aba)$ or $\mu^n(bab)$ for some $n \geqslant 0$.*

Yet another property of the Prouhet-Thue-Morse infinite word is the following result which explains the above terminology.

Theorem 4.2 *Let $w \in \{a,b\}^+$ be a segment of $\mathbf{t}$ containing at least one of the segments aba and bab along with all other segments of $\mathbf{t}$ of length 3. Then w is typical.*

Proof: Suppose φ is an endomorphism of $\{a,b\}^+$ such that $\varphi(w)$ is a segment of $\mathbf{t}$. By Proposition 4.1, since $\varphi(a^2)$ and $\varphi(b^2)$ are square segments of $\mathbf{t}$, each of the words $\varphi(a)$ and $\varphi(b)$ must be obtained by applying a power of μ to one of the words a, b, aba, bab. Let then $\varphi(a) = \mu^k(u)$ and $\varphi(b) = \mu^\ell(v)$, where $u, v \in \{a, b, aba, bab\}$. We may assume that $k \leqslant \ell$ since, otherwise, we would consider the pair $(\xi(w), \varphi\circ\xi)$ instead of (w, φ). Then, we have the factorization $\varphi = \mu^k \circ \psi$, where ψ is the endomorphism of $\{a,b\}^+$ defined by $\psi(a) = u$ and $\psi(b) = \mu^{\ell-k}(v)$. Since μ is injective and $\mathbf{t}$ is a fixed point of μ, from the fact that $\varphi(w)$ is a segment of $\mathbf{t}$, we conclude that so is $\psi(w)$. On the other hand, since ξ and μ commute, if ψ is a product of μ and ξ then so is φ. Hence, we may assume that $k = 0$.

The mapping $\xi\circ\varphi$ has also the property that $(\xi\circ\varphi)(w)$ is a segment of $\mathbf{t}$. Since $(\xi\circ\varphi)(a) = \xi(\mu^k(u)) = \mu^k(\xi(u))$, we may further assume that u is one of the words a or aba. Since aab is a factor of w and neither a^3 nor a^2ba^2 is a factor of $\mathbf{t}$, then v cannot start with a and, therefore, it must be either b or bab.

Consider first the case where $u = a$. Since baa is a factor of w but a^3 is not a factor of $\mathbf{t}$, $\varphi(b)$ cannot end in a, which implies that ℓ is even. If $\ell \geqslant 2$, then $\varphi(b)$ starts with $\mu^2(b) = ba^2b$. But, since a^2b is a factor of w, this implies a^2ba^2 is a factor of $\varphi(w)$ and, therefore, a segment of $\mathbf{t}$, which we know is not the case. Hence, we must have $\ell = 0$. It remains to rule out the case $\varphi(b) = bab$, which results from noting that in that case, from the assumption that either aba or bab is a factor of w, it follows that either $ababa$ or $babab ab$ is a factor of $\varphi(w)$ while we know that $ababa$ is not a segment of $\mathbf{t}$.

Next, consider the case where $u = aba$. Since a^2b is a factor of w and $ababa$ is not a segment of $\mathbf{t}$, $\varphi(b)$ cannot start with ba. As $\mu^\ell(b)$ is a prefix of $\varphi(b)$, it follows that $\ell = 0$ and $\varphi(b) = b$. This leads to a similar situation as that considered at the end of the preceding paragraph, with the letters a and b interchanged which is, therefore, excluded. This completes the proof of the theorem. □

The assumption of Theorem 4.2 that a segment of $\mathbf{t}$ contains as factors at least one of the words aba and bab along with all other segments of length 3 of $\mathbf{t}$ holds for all segments of $\mathbf{t}$ of length 10, as may be easily checked by examining all segments of that length. Hence, by Theorem 4.2 there are only finitely many atypical words. Since there are 5 different segments of length 3 of $\mathbf{t}$ that are supposed to appear in the word, no word with length shorter than 7 satisfies the criterion of Theorem 4.2 and ab^2a^2ba is a word of length 7 that does satisfy it. On the other hand, the segment a^2bab^2aba, of length 9, fails to have the segment ba^2 as a factor.

The following result completes the above observations by giving the full identification of atypical words.

Theorem 4.3 *Up to taking variants, the atypical words are the factors of the words $aabab$, $abaaba$, and $aabbaab$.*

Proof: To check that all relevant words have been duly considered, the reader may wish to refer to the diagram in Figure 2 later in the paper, where all atypical words are represented.

The following is the complete list of segments of $\mathbf{t}$ of length 5:

$$aabab, aabba, abaab, ababb, abbaa, abbab,$$
$$baaba, baabb, babaa, babba, bbaab, bbaba.$$

Note that all these words are variants of factors of at least one of the three words in the statement of the theorem. Hence, by showing that those three words are atypical, we obtain that so are all words of length up to 5.

We next indicate for each of the words in the statement of the theorem an endomorphism φ of $\{a,b\}^+$ not of the forms μ^n and $\xi \circ \mu^n$ that maps it to a segment of $\mathbf{t}$:

- $aabab$: $\varphi(a) = a$, $\varphi(b) = b^2aba^2b$;
- $abaaba$: $\varphi(a) = a$, $\varphi(b) = b^2$;
- $aabbaab$: $\varphi(a) = a$, $\varphi(b) = bab$.

The verification of all these statements amounts to routine calculations.

Showing that there are no other atypical words requires more work. Note that a word is typical if it has a typical factor. Hence, also excluding variants and words that satisfy the criterion of Theorem 4.2, we obtain the following reduced list of words remaining to be treated:

$$aababb, aabbab, abbaabba, ababba. \tag{3}$$

We proceed to show that each word w in the list (3) is typical. For that purpose, assume that φ is an endomorphism of $\{a,b\}^+$ such that $\varphi(w)$ is a segment of $\mathbf{t}$.

In the first three cases, since $\varphi(a^2)$ and $\varphi(b^2)$ are factors of $\varphi(w)$, we may start the argument using Proposition 4.1 as in the proof of Theorem 4.2, assuming that $\varphi(a)$ is either a or aba.

Consider first the case $\varphi(a) = a$. Since in the three cases, aab is a factor of w but $\mathbf{t}$ is cube-free, $\varphi(b)$ must start with b. Therefore, we may assume that $\varphi(b) = bv$ with $v \in \{a,b\}^+$. In all three cases, since abb is a factor of w, we get that $abvbv$ is a factor of $\varphi(w)$ and this would provide an overlap in $\mathbf{t}$ if v ends with a. Hence v ends with b. Since $\varphi(b)$ is of the form $\mu^n(x)$ for some $x \in \{a, aba, b, bab\}$, we conclude that either $\varphi(b)$ starts with $baab$ or it is bab. The first case is excluded since $aabaa$ is then a factor of $\varphi(aab)$, whence also of $\varphi(w)$, while it is not a segment of $\mathbf{t}$. The case $\varphi(b) = bab$ is also excluded if w is either $aababb$ or $aabbab$ since it leads to the overlap $babab$ in the factor $\varphi(bab)$ of $\varphi(w)$. In case $w = abbaabba$, one can simply check directly that $\varphi(w)$ is not a segment of $\mathbf{t}$.

Still treating for the moment only the first three of the words in the list (3), suppose next that $\varphi(a) = aba$. Again, as aab is a factor of w and $aabaa$ cannot be a factor of $\varphi(w)$, $\varphi(b)$ must start with b. If it ends with a, then $a\varphi(bb)$ would be an overlap in $\varphi(w)$ since abb is a factor of w. Hence, $\varphi(b)$ starts and ends with b. This is impossible in case w has the factor bab since it would lead to the overlap $babab$ in $\varphi(w)$. This excludes the cases where w is the first or the second word in the list (3). So, we have $w = abbaabba$. Then $\varphi(w)$ is a square segment of $\mathbf{t}$. By Proposition 4.1, $\varphi(abba)$ is one of the words $\mu^n(x)$ with $x \in \{a, aba, b, bab\}$. Since $n \leqslant 1$ gives a word that is too short to be $\varphi(abba)$, we must have $n \geqslant 2$, in which case a simple calculation shows that $\mu^n(x)$ cannot start with aba. This ends the verification that the first three words in the list (3) are typical.

It remains to consider the word $w = ababba$. Here, we have two square factors of $\varphi(w)$, namely the squares of $\varphi(b)$ and $\varphi(ab)$. By Proposition 4.1 we know that there are words $x, y \in \{a, aba, b, bab\}$ and non negative integers m, n such that $\varphi(b) = \mu^m(x)$ and $\varphi(ab) = \mu^n(y)$. In case $m > n$, comparing the lengths of the word $\varphi(ab)$ and its factor $\varphi(b)$, we obtain the inequality $2^n|y| > 2^m|x|$, so that $3 \geqslant |y| > 2^{m-n}|x| \geqslant 2^{m-n}$. It follows that $|x| = 1$, $|y| = 3$ and $m = n + 1$. From the equalities $\mu^n(y) = \varphi(ab) = \varphi(a)\mu^{n+1}(x)$ we then deduce that $\varphi(a) = \mu^n(y_1)$, where y_1 is the first letter of y, and $\mu(x) = xy_1$. Since abb is a factor of w, $\varphi(abb) = \mu^n(y_1xy_1xy_1)$ is a segment of $\mathbf{t}$, which contradicts $\mathbf{t}$ being overlap free. Thus, we must have $n \geqslant m$. Then $\varphi(a)$ must be of the form $\mu^m(z)$ where z is a prefix of $\mu^{n-m}(y)$. It follows that, as in the proof of Theorem 4.2, we may then assume that $m = 0$ and that $\varphi(b)$ is either b or bab. Consider first the case where $\varphi(b) = b$. Since $abba$ is a factor of w but b^3 is not a factor of $\varphi(w)$, the word $\varphi(a)$ must start and end with the letter a and we may assume that it is not reduced to a. Since $\varphi(a) = z$, we conclude that $\varphi(a)$ must start with $abba$. Since bba is a factor of w, this yields the factor $bbabb$ of $\varphi(w)$, which is not possible since $\varphi(w)$ is a segment of $\mathbf{t}$. Finally, the case $\varphi(b) = bab$ is excluded since $\varphi(ab) = \mu^n(y)$ cannot end with bab. This concludes the proof of the theorem. □

To facilitate the visualization of the set of atypical words, we give a semigroup theoretical formulation. Although we do not go deep into it, the reader unfamiliar with semigroup theory may prefer to skip these considerations or refer to a standard textbook in the area such as Clifford and Preston (1961); Howie (1995).

Let S be the set of atypical words. We may define a multiplication on the set $S^0 = S \cup \{0\}$ as follows: for $u, v \in S$, $u \cdot v$ is uv if uv is atypical and 0 otherwise; for all $s \in S^0$, $s \cdot 0 = 0 \cdot s = 0$. Note that S^0 is the Rees quotient of $\{a, b\}^+$ by the ideal consisting of the typical words together with the words that are not segments of $\mathbf{t}$.

The diagram in Figure 2 represents S^0 as a partially ordered set for the Green $\mathcal{J}$-order, in which an element u lies above v if and only if u is a factor of v. The words in bold are the lexicographic minima among their variants; note that those that are atoms (which are underlined) are precisely the words that were shown directly to be atypical in Theorem 4.3.

We conclude this section with another application of Theorem 4.2, this one concerning infinite patterns of $\mathbf{t}$.

Corollary 4.4 *Let w be an infinite word and suppose that there is an endomorphism φ of $\{a, b\}^+$ such that $\varphi(w)$ is a suffix of either $\mathbf{t}$ or $\xi(\mathbf{t})$. Then w is itself a suffix of either $\mathbf{t}$ or $\xi(\mathbf{t})$.*

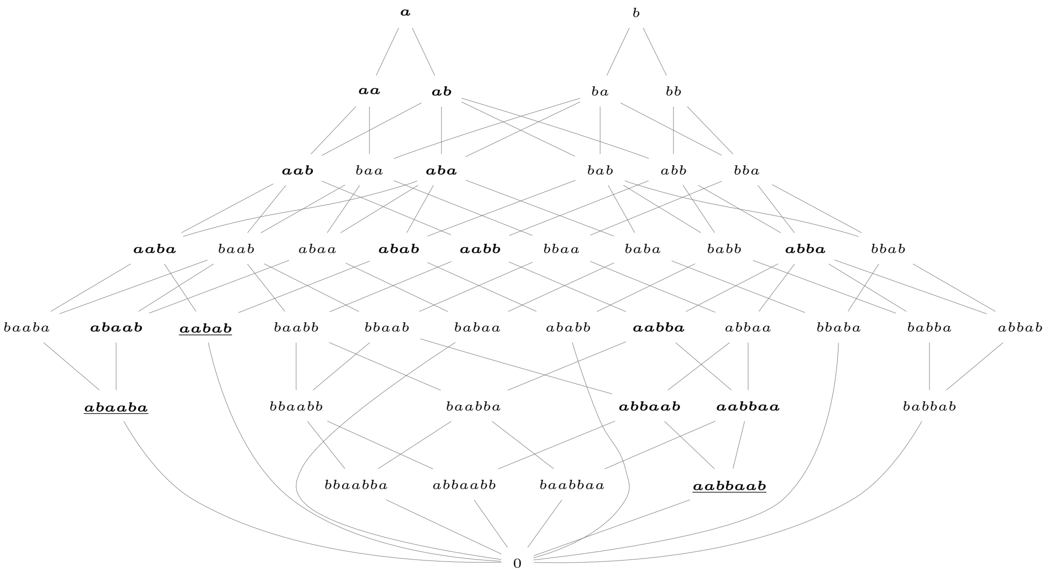

Fig. 2: The semigroup S^0

Proof: Since all segments of w are unavoidable in $\mathbf{t}$ and they are all extendable on the right, by Theorem 3.7 they are segments of $\mathbf{t}$. Since the language of the segments of $\mathbf{t}$ defines a minimal subshift (Queffélec, 2010, Proposition 5.2), it follows that w and $\mathbf{t}$ have the same segments. In particular, the word $a^2b^2a^2bab$ is a segment of w and it satisfies the assumption of Theorem 4.2. It follows that there is $n \geqslant 0$ such that $\varphi = \mu^n$ or $\varphi = \xi \circ \mu^n$. Again, since μ is injective and both $\mathbf{t}$ and $\xi(\mathbf{t})$ are fixed by μ, the result follows. □

The somewhat different formulation for finite and infinite segments (compare Theorem 3.7 with Corollary 4.4) is fully justified by the following result, which entails that the infinite words $\mathbf{t}$ and $\xi(\mathbf{t})$ have no common suffix.

Proposition 4.5 *If* $\mathbf{s}$ *is an infinite word over* $\{a, b\}$ *and* $\mathbf{w}$ *is a common infinite suffix of* $\mathbf{s}$ *and* $\xi(\mathbf{s})$*, then* $\mathbf{w}$ *is periodic.*

Proof: By assumption, there are finite words x and y such that $\mathbf{s} = x\mathbf{w}$, $\xi(\mathbf{s}) = y\mathbf{w}$. Since $\mathbf{w}$ and $\xi(\mathbf{w})$ start with different letters, the words x and y have different lengths. Replacing $\mathbf{s}$ by $\xi(\mathbf{s})$, if needed, we may assume that x is shorter than y. As $\xi(x)$ is a prefix of $\xi(\mathbf{s}) = y\mathbf{w}$, it follows that $y = \xi(x)z$ for some word z. From $\xi(\mathbf{s}) = \xi(x)z\mathbf{w}$, we deduce that $\xi(\mathbf{w}) = z\mathbf{w}$ and so $\mathbf{w} = \xi^2(\mathbf{w}) = \xi(z)z\mathbf{w}$, thereby showing that $\mathbf{w}$ is periodic. □

5 Final remarks and problems

For an infinite word $\mathbf{w}$ over a finite alphabet A, let $L(\mathbf{w})$ be the language consisting of its finite segments. Note that the automorphisms of the semigroup A^+ permute the letters of A; we call them *letter exchanges*.

The language obtained from $L(\mathbf{w})$ by applying all possible letter exchanges is denoted $\bar{L}(\mathbf{w})$. Let $\mathcal{E}(\mathbf{w})$ denote the set of all endomorphisms φ of A^+ such that $\varphi(L(\mathbf{w})) \subseteq L(\mathbf{w})$. The set $\bar{\mathcal{E}}(\mathbf{w})$ is similarly defined using $\bar{L}(\mathbf{w})$ instead of $L(\mathbf{w})$. Note that both $\mathcal{E}(\mathbf{w})$ and $\bar{\mathcal{E}}(\mathbf{w})$ are submonoids of the monoid $\mathrm{End}(A^+)$ of all endomorphisms of the semigroup A^+.

The following is an immediate consequence of Theorem 4.2.

Corollary 5.1 *The monoid $\mathcal{E}(\mathbf{t}) = \bar{\mathcal{E}}(\mathbf{t})$ is generated by the set $\{\xi, \mu\}$. In particular, it is finitely generated.* □

Corollary 5.1 is intimately related with a result of Thue (see Berstel (1995, Chapter 3, Theorem 2.16)) that characterizes the set of the so-called *overlap-free morphisms*, that is, endomorphisms of $\{a, b\}^+$ that map the set of all overlap-free words into itself, namely as the monoid generated by $\{\xi, \mu\}$. In fact, in view of another result of Thue (see Berstel (1995, Chapter 3, Theorem 2.15)), all (overlap-free) words that can be arbitrarily prolonged in both directions to overlap-free words are segments of $\mathbf{t}$. It follows that overlap-free morphisms belong to $\mathcal{E}(\mathbf{t})$ and so Corollary 5.1 immediately yields Thue's necessary condition for overlap-free morphisms. That the condition is also sufficient is given by another result of Thue (see Berstel (1995, Chapter 3, Lemma 2.2)). It does not appear to be immediately obvious how to deduce Corollary 5.1 from Thue's results.

Corollary 5.1 is also related with a result of Pansiot (1981) characterizing the endomorphisms of $\{a, b\}$ that generate some infinite word obtained from $\mathbf{t}$ by dropping a finite prefix as precisely the powers of μ. Since $\mathbf{t}$ is recurrent, all such infinite words $\mathbf{w}$ have the same language $L(\mathbf{w}) = L(\mathbf{t})$. Hence, the endomorphisms φ considered by Pansiot belong to $\mathcal{E}(\mathbf{t})$, whence they are products of ξ and μ. Since ξ and μ commute, it follows from Corollary 5.1 that φ is either μ^k or $\xi\mu^k$ for some $k \geqslant 0$, the latter possibility being excluded because $\mathbf{w}$ is assumed to be a fixed point of φ. This gives Pansiot's result. Again, it is not clear how to deduce Corollary 5.1 from Pansiot's results.

Theorems 3.7 and 4.2, together with Corollary 5.1 may be regarded as three finiteness properties of the Prouhet-Thue-Morse sequence. It is natural to ask which infinite words possess such finiteness properties. More precisely, we propose the following problems.

Problem 1 *Which infinite words $\mathbf{w}$ have the property that, up to finitely many exceptions, the patterns of $\mathbf{w}$ on the same alphabet are obtained from its segments up to an exchange of letters?*

Problem 2 *For which infinite words $\mathbf{w}$ is the monoid $\mathcal{E}(\mathbf{w})$ finitely generated? Similar question for $\bar{\mathcal{E}}(\mathbf{w})$.*

We say that a finite segment u of $\mathbf{w}$ is $\mathbf{w}$-*atypical* if there is some endomorphism $\varphi \notin \bar{\mathcal{E}}(\mathbf{w})$ of A^+ such that $\varphi(u)$ is also a segment of $\mathbf{w}$.

Problem 3 *Which infinite words $\mathbf{w}$ have only finitely many $\mathbf{w}$-atypical segments?*

A negative example for Problem 1 is provided by the *Fibonacci infinite word*, which is the only fixed point $\mathbf{f}$ of the endomorphism ϕ of $\{a, b\}^+$ defined by $\phi(a) = ab$ and $\phi(b) = a$. That there are infinitely many finite binary patterns of $\mathbf{f}$ that are not segments of $\mathbf{f}$ was proved in Restivo and Salemi (2002a) (see also Restivo and Salemi (2002b)), where it is also shown that there are Sturmian infinite words that admit as patterns all segments of all Sturmian infinite words. Recall that an infinite word is *Sturmian* if it has exactly $n + 1$ segments of each length $n \geqslant 1$. We do not know whether $\mathcal{E}(\mathbf{f})$ is generated by φ and $\bar{\mathcal{E}}(\mathbf{f})$ is generated by φ and ξ. We also do not know whether the set of $\mathbf{f}$-atypical words is finite.

Problem 1 was raised in Restivo and Salemi (2002a) for binary infinite words that are either fixed points of endomorphisms or of linear complexity. In the same paper, it is observed that if $\mathbf{w}$ is an infinite word with all elements of A^+ as segments (which may be obtained for instance by concatenating all the words in a sequence enumerating the elements of A^+), then obviously $\mathbf{w}$ is a positive example for Problem 1. Note that $\mathcal{E}(\mathbf{w}) = \bar{\mathcal{E}}(\mathbf{w}) = \mathrm{End}(A^+)$ and it is easy to see that $\mathrm{End}(A^+)$ is not finitely generated: for the endomorphisms that maps each letter to itself, except for one letter a that is mapped to a^p, where p is prime, the only elements of $\mathrm{End}(\mathbf{w})$ that are factors of it are the letter exchanges and the factors of which it is also a factor. From the preceding observation it also follows that there are no $\mathbf{w}$-atypical words. Thus, $\mathbf{w}$ is a negative example for Problem 2 and a positive example for Problem 3.

Acknowledgments

Besides the connection with Pansiot's work, we thank the anonymous referee for comments that led to improved readability of this paper.

References

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

J.-P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, 2003.

J. Almeida and A. Costa. Presentations of Schützenberger groups of minimal subshifts. *Israel J. Math.*, 196:1–31, 2013.

J. Almeida, A. Costa, R. Kyriakoglou, and D. Perrin. *Profinite semigroups and symbolic dynamics*, volume 2274 of *Lect. Notes in Math.* Springer, Cham, 2020.

J. Berstel. Axel Thue's papers on repetitions in words: a translation. Technical report, Université du Québec à Montréal, 1995. Publications du LaCIM 20, available at `http://www-igm.univ-mlv.fr/~berstel/Articles/1994ThueTranslation.pdf`.

S. Brlek. Enumeration of factors in the Thue-Morse word. *Discrete Appl. Math.*, 24:83–96, 1989.

A. H. Clifford and G. B. Preston. *The Algebraic Theory of Semigroups*, volume I. Amer. Math. Soc., Providence, R.I., 1961.

A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theor. Comp. Sci.*, 63(3):333–348, 1989.

J. M. Howie. *Fundamentals of semigroup theory*, volume 12 of *London Mathematical Society Monographs. New Series*. The Clarendon Press, Oxford University Press, New York, 1995.

M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading, Mass., 1983.

H. M. Morse. Recurrent geodesics on a surface of negative curvature. *Trans. Amer. Math. Soc.*, 22: 84–100, 1921.

J.-J. Pansiot. The Morse sequence and iterated morphisms. *Inform. Process. Lett.*, 12(2):68–70, 1981.

E. Prouhet. Mémoire sur quelques relations entre les puissances des nombres. *C. R. Acad. Sci. Paris*, 33: 31, 1851.

M. Queffélec. *Substitution dynamical systems—spectral analysis*, volume 1294 of *Lect. Notes in Math.* Springer-Verlag, Berlin, second edition, 2010.

A. Restivo and S. Salemi. Words and patterns. In *Developments in language theory (Vienna, 2001)*, volume 2295 of *Lect. Notes in Comput. Sci.*, pages 117–129. Springer, Berlin, 2002a.

A. Restivo and S. Salemi. Binary patterns in infinite binary words. In *Formal and natural computing*, volume 2300 of *Lect. Notes in Comput. Sci.*, pages 107–116. Springer, Berlin, 2002b.

A. M. Shur. Binary words avoided by the Thue-Morse sequence. *Semigroup Forum*, 53:212–219, 1996a.

A. M. Shur. Overlap-free words and Thue-Morse sequences. *Int. J. Algebra Comput.*, 6:353–367, 1996b.

A. M. Shur. *Algebraic and combinatorial properties of rational languages*. PhD thesis, Ural State University, 1997. In Russian.

A. M. Shur. Combinatorial complexity of rational languages. *Diskretn. Anal. Issled. Oper.*, 12(2):78–99, 2005. In Russian.

A. Thue. Über unendlichen Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, (7):1–22, 1906.

A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter, I. Mat. Nat. Kl.*, (1):1–67, 1912.