Badkobeh, Golnaz and Ochem, Pascal. 2015. Characterization of some binary words with few squares. Theoretical Computer Science, 588, pp. 73-80. ISSN 0304-3975 [Article]

http://research.gold.ac.uk/28016/

The version presented here may differ from the published, performed or presented work. Please go to the persistent GRO record above for more information.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Goldsmiths, University of London via the following email address: gro@gold.ac.uk.

The item will be removed from the repository while any claim is being investigated. For more information, please contact the GRO team: gro@gold.ac.uk
Characterization of some binary words with few squares

Golnaz Badkobeh\textsuperscript{a}, Pascal Ochem\textsuperscript{b}

\textsuperscript{a}Department of Computer Science, University of Sheffield, UK
\textsuperscript{b}CNRS - LIRMM, Montpellier, France

Abstract
Thue proved that the factors occurring infinitely many times in square-free words over \{0, 1, 2\} avoiding the factors in \{010, 212\} are the factors of the fixed point of the morphism 0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1. He similarly characterized square-free words avoiding \{010, 020\} and \{121, 212\} as the factors of two morphic words. In this paper, we exhibit smaller morphisms to define these two square-free morphic words and we give such characterizations for six types of binary words containing few distinct squares.

1. Introduction

Let \(\Sigma_k\) denote the \(k\)-letter alphabet \{0, 1, \ldots, k-1\}. Let \(\varepsilon\) denote the empty word. A finite word is \emph{recurrent} in an infinite word \(w\) if it appears as a factor of \(w\) infinitely many times. An infinite word \(w\) is \emph{recurrent} if all its finite factors are recurrent in \(w\). If a morphism \(f\) is such that \(f(0)\) starts with 0, then the \emph{fixed point} of \(f\) is the unique word \(w = f^\infty(0)\) starting with 0 and satisfying \(w = f(w)\). An infinite word is \emph{pure morphic} if it is the fixed point of a morphism. An infinite word is \emph{morphic} if it is the image \(g(f^\infty(0))\) by a morphism \(g\) of a pure morphic word \(f^\infty(0)\). The \emph{factor complexity} of an infinite word or a language is the number of factors of length \(n\) of the infinite word or the language. A pattern \(P\) is a finite word of variables over the alphabet \{A, B, \ldots\}. A word \(w\) (finite or infinite) \emph{avoids} a pattern \(P\) if for every substitution \(\phi\) of the variables of \(P\) with non-empty words, \(\phi(P)\) is not a factor of \(w\). Given a finite alphabet \(\Sigma_k\), a finite set \(\mathcal{P}\) of patterns, and a finite set \(\mathcal{F}\) of factors over \(\Sigma_k\), we say that \(\mathcal{P} \cup \mathcal{F}\) \emph{characterizes} a morphic word \(w\) over \(\Sigma_k\) if \(w\) avoids \(\mathcal{P} \cup \mathcal{F}\) and every recurrent factor of an infinite word avoiding \(\mathcal{P} \cup \mathcal{F}\) is a factor of \(w\). In other words, \(\mathcal{P} \cup \mathcal{F}\) characterizes \(w\) if and only if every recurrent word over \(\Sigma_k\) avoiding \(\mathcal{P} \cup \mathcal{F}\) has the same set of factors as \(w\). In our results, we do not specify the alphabet size \(k\) since \(\Sigma_k\) corresponds to the set of letters appearing in \(\mathcal{F}\). A \emph{repetition} is a factor of the form \(r = u^n v\) where \(u\) is non-empty and \(v\) is a prefix of \(u\). Then \(|u|\) is the \emph{period} of the repetition \(r\) and its \emph{exponent} is \(|r|/|u|\). A \emph{square} is a repetition of exponent 2. Equivalently, it is an occurrence of the pattern \(AA\). An overlap is a repetition with exponent strictly greater than 2.

Thue [3, 10, 11] gave the following characterization of overlap-free binary words: \{ABABA\} \cup \{000, 111\} characterizes the fixed point of the morphism
Concerning ternary square-free words, he proved that

- \( \{AA\} \cup \{010,212\} \) characterizes the fixed point of \( f : 0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1 \),
- \( \{AA\} \cup \{010,020\} \) characterizes the morphic word \( T_1(f^\infty(0)) \),
- \( \{AA\} \cup \{121,212\} \) characterizes the morphic word \( T_2(f^\infty(0)) \),

where the morphisms \( f_T, T_1, T_2 \) are given below.

\[
\begin{align*}
    f_T(0) &= 012, & T_1(0) &= 01210212, & T_2(0) &= 021012, \\
    f_T(1) &= 0432, & T_1(1) &= 012102012, & T_2(1) &= 02102012, \\
    f_T(2) &= 0134, & T_1(2) &= 0121021201, & T_2(2) &= 0210201, \\
    f_T(3) &= 013432, & T_1(3) &= 012102120212021, & T_2(3) &= 0210120120212, \\
    f_T(4) &= 0434, & T_1(4) &= 012102012, & T_2(4) &= 0210201.
\end{align*}
\]

To obtain the last two results, Thue first proved that \( f^\infty(0) \) is characterized by \( \{AA\} \cup \{02,03,10,14,21,23,24,30,31,41,42,040,1201,2012\} \).

In this paper, we prove such characterizations mostly for the binary words considered by the first author \[1\]. We also obtain smaller morphisms for Thue’s words avoiding \( \{AA\} \cup \{010,020\} \) and \( \{AA\} \cup \{121,212\} \) as well as a characterization for words avoiding the patterns \( AABBCC \) (i.e., three consecutive squares), \( ABCABC \) and a finite set of factors. The results are summarized in Table \( \ref{table} \). The first column shows the description of the considered language given in the literature. It is either given by forbidden sets of patterns and factors, or by the notation \((e,n,m)\), which means that we consider the binary words avoiding repetitions with exponent strictly greater than \( e \), containing exactly \( n \) distinct repetitions with exponent \( e \) as a factor, and containing the minimum number \( m \) of distinct squares. We use the notation \( SQ_t \) for the pattern corresponding to squares with period at least \( t \), that is, \( SQ_1 = AA, SQ_2 = ABAB, SQ_3 = ABCABC \), and so on. These languages actually have an equivalent definition with one forbidden pattern \( SQ_t \) and a finite set of forbidden factors. This standardized definition, given in the second column, is more suited for proving the characterization. The third column gives the corresponding morphic word. The fourth column indicates the section containing the corresponding set \( F_{xx} \) and morphism \( g_{xx} \).

To define a morphic word \( g(f^\infty(0)) \), we allow that \( g \) is an erasing morphism, i.e., that the \( g \)-image of a letter is empty. Notice that replacing \( g \) by \( h_c = g \circ f^c \) defines the same morphic word, and that \( h_c \) is non-erasing for some small constant \( c \).

The proofs are obtained by computer using the technique described in the next section. An example of proof by hand is given for Theorem \[3\]. The morphic words in Table \( \ref{table} \) are gathered according the pure morphic word they are built on. We introduce in Section \( \ref{section} \) a pure morphic word \( f_5^\infty(0) \) similar to Thue’s word \( f_T^\infty(0) \) and we characterize some of its morphic images. Section \( \ref{section} \) is devoted to characterizations of some morphic images of Thue’s ternary pure morphic word \( f_T^\infty(0) \).
2. Characterizing a morphic word

A morphism \( f : \Sigma_k^* \rightarrow \Sigma_k^* \) is primitive if there exists \( n \in \mathbb{N} \) such that \( f^n(a) \) contains \( b \) for every \( a, b \in \Sigma_k \). We are given a primitive morphism \( f : \Sigma_k^* \rightarrow \Sigma_k^* \), a morphism \( g : \Sigma_k^* \rightarrow \Sigma_k^* \), and a finite set of factors \( F_m \subset \Sigma_k^* \). We want to prove that \( g(f^\infty(0)) \) is characterized by \( \{SQ_1\} \cup F_m \).

We assume that \( g(f^\infty(0)) \) avoids \( \{SQ_1\} \cup F_m \). This can be checked using Cassaigne’s algorithm [5] that determines if a morphic word defined by circular morphisms avoids a given pattern with constants. We refer to Cassaigne [5] for the definitions of circular morphisms, synchronization point, and synchronization delay. We can use an online implementation [3] of this algorithm. We also assume that the pure morphic word \( f^\infty(0) \) is characterized by \( \{AA\} \cup F_p \) for some finite set of factors \( F_p \subset \Sigma_k^* \).

We compute the smallest integer \( c \) such that \( \min \{|g(f^c(a))|, a \in \Sigma_k\} \geq t \). This \( c \) exists because \( f \) is primitive. We can consider the morphism \( g' = g \circ f^c \) instead of \( g \) since we have \( g'(f^\infty(0)) = g(f^\infty(0)) \).

First, we check that \( g' \) is circular. Then, we compute the set \( S_l \) of words \( v \) such that there exists a word \( w \in \Sigma_k^* \) avoiding \( \{SQ_1\} \cup F_p \), where \( l = \max \{|u|, u \in F_p\} \times \max \{|g(f(a))|, a \in \Sigma_k\}, |v| = l \), and \( |p| = |s| = 4l \). To do this, we simply perform a depth-first exploration of the words of length \( 4l \) avoiding \( \{SQ_1\} \cup F_m \) for each of them, we put the central factor of length \( l \) in \( S_l \). The running time of this brute-force approach is not so prohibitive precisely because the characterization implies a polynomial factor complexity. Finally, we check that every word in \( S_l \) is a factor of \( g'(f^\infty(0)) \).

This implies that an infinite word over \( \Sigma_k \) avoiding \( \{SQ_1\} \cup F_m \) is the \( g' \)-image of an infinite word \( w \in \Sigma_k^* \). Now \( w \) is square-free, since otherwise \( g'(w) \) would contain a square of period at least \( t \). Also \( w \) does not contain a word \( y \in F_p \), because \( g'(y) \) is a word of length at most \( l \) that is not a factor of any word in \( S_l \). So \( w \) avoids \( \{AA\} \cup F_p \), and thus has the same set of factors as \( f^\infty(0) \). Thus, every infinite recurrent word over \( \Sigma_k \) avoiding \( \{SQ_1\} \cup F_m \) has the same set of factors as \( g'(f^\infty(0)) \).

The programs we used are available at http://www.lirmm.fr/~ochem/morphisms/characterization.htm.
3. A pure morphic word over $\Sigma_5$

We define the morphism $f_5$ from $\Sigma_5^*$ to $\Sigma_5^*$ as follows:

$$f_5(0) = 01, 
\quad f_5(1) = 23, 
\quad f_5(2) = 4, 
\quad f_5(3) = 21, 
\quad f_5(4) = 0.$$  

We also define the set

$$F_5 = \{02, 03, 13, 14, 20, 24, 31, 32, 40, 41, 43, 121, 212, 304, 3423, 4234\}.$$  

**Theorem 1.** $\{AA\} \cup F_5$ characterizes $f_5^\infty(0)$.

**Proof.** We adapt the method of the previous section for morphic words to the pure morphic word $f_5^\infty(0)$ by setting $g = g' = f_5$ and $F_m = F_p = F_5$. We set $l = \max \{|w|, w \in F_5\} \times \max \{|f_5(a)|, a \in \Sigma_k\} = 8$. We compute the set $S_l$ of words $v$ such that there exists a word $pvw \in \Sigma_5^*$ avoiding squares and $F_5$ with $|v| = l$ and $|p| = |s| = 4l$. Then we check that every word in $S_l$ is a factor of $f_5^\infty(0)$.

The morphism $f_5$ is circular with synchronization delay 1. Indeed, for every factor of length 1 of the $f_5$-image of some word, we can insert at least one synchronization point $|$ between letter images:

1. $0$ implies $0$,
2. $1$ implies $1$,
3. $2$ implies $2$,
4. $3$ implies $3$,
5. $4$ implies $4$.

This implies that every infinite recurrent word over $\Sigma_5$ avoiding $\{AA\} \cup F_5$ is the $f_5$-image of some infinite recurrent word $w$ over $\Sigma_5$. Notice that $w$ must be square-free, since otherwise $f_5(w)$ would not avoid squares. Now suppose that $w$ contains a factor $y \in F_5$. Then $f_5(y)$ must appear as a factor in $S_l$ since $|f_5(y)| \leq 8 = l$. Every word in $S_l$ is a factor of $f_5^\infty(0)$, so $f_5(y)$ should also be a factor of $f_5^\infty(0)$, which is a contradiction. So $w$ avoids squares and $F_5$, which implies by induction that it has the same set of factors as $f_5^\infty(0)$. Finally, we have that every infinite recurrent word over $\Sigma_5$ avoiding $\{AA\} \cup F_5$ is of the form $f_5(w)$ where $w$ has the same set of factors as $f_5^\infty(0)$, so that $f_5(w)$ also has the same set of factors as $f_5^\infty(0)$.

Since many morphic words in this paper are obtained as the image of $f_5^\infty(0)$, let us state some of its properties. In $f_5^\infty(0)$, the letters 0, 1, and 2 have frequency $\sqrt{5} - 2$ and the letters 3 and 4 have frequency $(7 - 3\sqrt{5})/2$. Notice that $\{AA\} \cup F_5$, and thus the set of factors of $f_5^\infty(0)$, is invariant by the operation
consisting in reversing the word and exchanging 3 and 4. This is trivially true for squares. For a word in $F_5$, say 40, we obtain 04 by reversing the word and we obtain 03 by exchanging 3 and 4, then we have that $F_5$ contains indeed 03. The factor complexity of $f_5^\infty(0)$ seems to be $4n+1$ for every factor length $n \geq 0$.

3.1. Smaller morphisms for Thue’s words

Let $M_1$ and $M_2$ be the morphisms from $\Sigma_5^*$ to $\Sigma_3^*$ defined by

$M_1(0) = 012$,  
$M_1(1) = 1$,  
$M_1(2) = 02$,  
$M_1(3) = 12$,  
$M_1(4) = \varepsilon$.

$M_2(0) = 02$,  
$M_2(1) = 1$,  
$M_2(2) = 0$,  
$M_2(3) = 12$,  
$M_2(4) = \varepsilon$.

Theorem 2.

• $\{AA\} \cup \{010, 020\}$ characterizes the morphic word $M_1(f_5^\infty(0))$.

• $\{AA\} \cup \{121, 212\}$ characterizes the morphic word $M_2(f_5^\infty(0))$.

Thue noticed that every word avoiding $\{AA\} \cup \{121, 212\}$ can be obtained from a word avoiding $\{AA\} \cup \{010, 020\}$ by deleting the letter immediately after each occurrence of the letter 0. This property is easy to check by comparing $M_2$ to $M_1$ and it explains why the same pure morphic word is used for both types of words. The morphisms $M_1$ and $M_2$ are the smallest possible. However, the morphisms $M'_1 = M_1 \circ f_5$ and $M'_2 = M_2 \circ f_5$ given below provide additional insight.

$M'_1(0) = 0121$,  
$M'_1(1) = 0212$,  
$M'_1(2) = \varepsilon$,  
$M'_1(3) = 021$,  
$M'_1(4) = 012$.

$M'_2(0) = 0121$,  
$M'_2(1) = 0212$,  
$M'_2(2) = \varepsilon$,  
$M'_2(3) = 01$,  
$M'_2(4) = 02$.

The morphism $M'_1$ exhibits natural properties of words avoiding $\{AA\} \cup \{010, 020\}$ and of $M_1(f_5^\infty(0))$:

• The set \{0121, 0212, 012, 021\} is a code for words avoiding $\{AA\} \cup \{010, 020\}$.

• The asymptotic frequencies of the factors 121 and 212 are equal since the letters 1 and 2 are symmetrical for words avoiding $\{AA\} \cup \{010, 020\}$.

• Similarly, the asymptotic frequencies of 0120 and 0210 are equal.

• By applying the symmetry of the factors of $f_5^\infty(0)$ to $M'_1$, that is, reversing the $M'_1$-images of every letter and exchanging 3 and 4, we obtain the conjugate morphism of $M'_1$ such that the common prefix 0 becomes the common suffix.

Except for the last, similar remarks hold for $M'_2$. The factor complexity of $M_1(f_5^\infty(0))$ and $M_2(f_5^\infty(0))$ seems to be $4n - 2$ for every factor length $n \geq 2$. 

5
3.2. Words containing two $5/2$-repetitions and 8 squares

If an infinite binary word contains the repetitions 0101 and 10101 of exponent 5/2 and no other overlap, then it contains at least 8 distinct squares. Moreover, if it contains exactly 8 distinct squares, then these 8 squares are $0^2$, $1^4$, $(01)^2$, $(10)^2$, $(0110)^2$, $(1001)^2$, $(011001)^2$, $(100110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids $SQ_7$ and the set

$$F_8 = \{000, 111, 00100, 11011, 010010, 010101, 101010, 101101, 00110011, 11001100, 1011001011, 0100110100\}.$$ 

Let $g_8$ be the morphism from $\Sigma_5^*$ to $\Sigma_2^*$ defined by

$$g_8(0) = 011,$$
$$g_8(1) = 0,$$
$$g_8(2) = 01,$$
$$g_8(3) = \varepsilon,$$
$$g_8(4) = \varepsilon.$$

**Theorem 3.** $\{SQ_7\} \cup F_8$ characterizes $g_8(f_5^\infty(0))$.

**Proof.** We assume that $g_8(f_5^\infty(0))$ avoids $SQ_7$ and $F_8$ and we prove the other direction of Theorem 8. That is, we suppose that $G_8$ is an infinite recurrent word avoiding $\{SQ_7\} \cup F_8$ and we show that every factor of $G_8$ is a factor of $g_8(f_5^\infty(0))$. We consider the morphism $g_8 = g_8 \circ f_5^\infty$ given below instead of $g_8$ because we have $\min(|g_8(a)|, a \in \Sigma_5) = 9 \geq 7 = t$, as specified in the method.

$$g_8(0) = 011001010011010110011010,$$
$$g_8(1) = 0110010111001101,$$
$$g_8(2) = 011001010,$$
$$g_8(3) = 0110010110011010,$$
$$g_8(4) = 01100101001101101.$$

Let $p = 01100101$ be the common prefix of the factors $g_8(a)$ for $a \in \Sigma_5$. It is easy to check that every occurrence of $p$ in the $g_8'$-image of a word is the prefix of $g_8'$-image of a letter. So $g_8'$ has bounded synchronization delay. Moreover, a computer check shows that the factors of $G_8$ are factors of the $g_8'$-image of a word. Let $L \subset \Sigma_5^*$ denote the language of words whose $g_8'$-image is a factor of $G_8$. We show that $L$ is the set of factors of $f_5^\infty(0)$. Suppose that $L$ contains a square $uu$ for some $u \in \Sigma_5^*$. Then $G_8$ contains the square $g_8'(uu)$ with period $|g_8'(u)| \geq 9$. This is a contradiction since $G_8$ avoids $SQ_7$, so $L$ is square-free.

Now, for every $w \in F_5$, we suppose that $w \in L$ and obtain a contradiction:

- $w \in \{02, 32\}$: $g_8'(02)p$ and $g_8'(32)p$ both contain the square $1g_8'(2)p = (001100101)^2$ with period 9 as a suffix.

- $w = 03$: $g_8'(03)p$ contains the square $(1001101001100101)^2$ with period 16 as a suffix.
• \( w \in \{13, 41, 43\} \): A common suffix of \( g'_8(1) \) and \( g'_8(4) \) is 1. A common prefix of \( g'_8(1) \) and \( g'_8(3) \) is 0110010111. So, in every case, \( g'_8(w) \) contains the factor 1011001011 \( \in \mathcal{F}_8 \).

• \( w = 14 \): \( g'_8(14) \) contains the square \( (00110101100101)^2 \) with period 14 as a suffix.

• \( w \in \{20, 24\} \): \( g'_8(20) \) and \( g'_8(24) \) both contain the square \( g'_8(22) \) with period 9 as a prefix.

• \( w = 31 \): \( g'_8(31) \) contains the square \( g'_8(33) \) with period 16 as a prefix.

• \( w = 40 \): \( g'_8(40) \) contains the square \( g'_8(44) \) with period 14 as a prefix.

• \( w = 304 \): \( g'_8(304) = 0110(010110011010011001010011)^201 \) contains a square with period 24.

• \( w = 121 \): Since \( L \) is square-free and avoids \( \{13, 14\} \), \( L \) must contain 1210. However, \( g'_8(1210) \) contains the square \( g'_8(1212) \) with period 24 as a prefix.

• \( w = 212 \): Since \( L \) is square-free and avoids \( \{20, 24\} \), \( L \) must contain 2123. However, \( g'_8(2123) \) contains the square \( g'_8(2121) \) with period 24 as a prefix.

• \( w = 3423 \): Since \( L \) is square-free and avoids \( \{03, 13, 43\} \), \( L \) must contain 23423. Since \( L \) is square-free and avoids \( \{31, 32\} \), \( L \) must contain 234230. However, \( g'_8(234230) \) contains the square \( g'_8(234234) \) with period 39 as a prefix.

• \( w = 4234 \): Since \( L \) is square-free and avoids \( \{40, 41, 43\} \), \( L \) must contain 42342. Since \( L \) is square-free and avoids \( \{20, 24\} \), \( L \) must contain 423421. However, \( g'_8(423421) \) contains the square \( g'_8(423423) \) with period 39 as a prefix.

Therefore \( L \) is square-free and does not contain a factor in \( \mathcal{F}_5 \), thus \( L \) is the set of factors as \( f_5^\infty(0) \) by Theorem [1].

Notice that the last part of the proof above (when we prove that every word in \( \mathcal{F}_5 \) is a forbidden factor in \( L \)) differs from the computer check described in Section [2]. The proof by hand exhibits witness forbidden factors in \( \{SQ_1\} \cup \mathcal{F}_m \). The algorithm does the contrapositive: It lists all words avoiding \( SQ_1 \) and \( \mathcal{F}_m \) of some sufficient length and checks that they are \( g' \)-images of some word. The proof by hand exhibits witness forbidden factors in \( \{SQ_1\} \cup \mathcal{F}_m \). The algorithm does the contrapositive: It lists all words avoiding \( \{SQ_1\} \cup \mathcal{F}_m \) of some sufficient length and checks that they are images of some word avoiding \( \{AA\} \cup \mathcal{F}_p \).

The factor complexity of \( g_8(f_5^\infty(0)) \) seems to be \( 4n - 6 \) for every factor length \( n \geq 3 \).
3.3. Words containing two $7/3$-repetitions and 12 squares

If an infinite binary word contains the repetitions 0110110 and 1001001 of exponent $7/3$ and no other overlap, then it contains at least 12 distinct squares. Moreover, if it contains exactly 12 distinct squares, then these 12 squares are $0^2$, $1^2$, $(01)^2$, $(10)^2$, $(001)^2$, $(010)^2$, $(011)^2$, $(100)^2$, $(101)^2$, $(110)^2$, $(01101001)^2$, $(10010110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids $SQ_9$ and the set

$$F_{12} = \{000, 111, 01010, 10101, 001100, 110011, 0010010, 0100100, 1011011, 1101101, 0011010011, 0101100101, 1010011010, 1100101100, 0100101100101\}.$$ 

Let $g_{12}$ be the morphism from $\Sigma_5^*$ to $\Sigma_2^*$ defined by

$$
g_{12}(0) = 01, \\
g_{12}(1) = 0, \\
g_{12}(2) = 011, \\
g_{12}(3) = \varepsilon, \\
g_{12}(4) = \varepsilon.
$$

**Theorem 4.** $\{SQ_9\} \cup F_{12}$ characterizes $g_{12}(f_5^\infty(0))$.

The factor complexity of $g_{12}(f_5^\infty(0))$ seems to be $4n - 6$ for every factor length $n \geq 3$.

3.4. Words containing one $7/3$-repetition and 14 squares

If an infinite binary word contains the repetition 1001001 of exponent $7/3$ and no other overlap, then it contains at least 14 distinct squares. Moreover, if it contains exactly 14 distinct squares, then these 14 squares are $0^2$, $1^2$, $(01)^2$, $(10)^2$, $(001)^2$, $(010)^2$, $(011)^2$, $(100)^2$, $(101)^2$, $(110)^2$, $(01101001)^2$, $(10010110)^2$, $(0110010)^2$, and $(10010110)^2$. Equivalently, a recurrent binary word containing these overlaps and squares avoids $SQ_9$ and the set

$$F_{14} = \{000, 111, 11011, 010101, 101010, 00110010, 0100100, 00110011, 11001100, 0100100101, 1010011010, 1100101100, 00100100100, 010011010011, 001100100110, 1011001010110011\}.$$ 

Let $g_{14}$ be the morphism from $\Sigma_5^*$ to $\Sigma_2^*$ defined by

$$
g_{14}(0) = 01, \\
g_{14}(1) = 00110, \\
g_{14}(2) = 1, \\
g_{14}(3) = 0010110, \\
g_{14}(4) = 0110.
$$

**Theorem 5.** $\{SQ_9\} \cup F_{14}$ characterizes $g_{14}(f_5^\infty(0))$.

The factor complexity of $g_{14}(f_5^\infty(0))$ seems to be $4n - 1$ for every factor length $n \geq 11$. 

8
3.5. Words avoiding AABBCC

The second author proved that the pattern AABBCC, i.e., three consecutive squares, can be avoided over the binary alphabet \(\mathbb{B}\). More precisely, there exist exponentially many binary words avoiding both AABBCC and SQ_3. However, if we forbid also the factors in

\[F_{cs}' = \{0001100101100, 0110100111000, 1001011000111, 110001101001\},\]

we obtain a characterization of the morphic word \(g_{cs}(f_5^\infty(0))\), where \(g_{cs}\) is the morphism from \(\Sigma^*_5\) to \(\Sigma^*_2\) defined by

\[
\begin{align*}
g_{cs}(0) &= 0010110011010, \\
g_{cs}(1) &= 0111, \\
g_{cs}(2) &= 0010111010, \\
g_{cs}(3) &= 011100011010, \\
g_{cs}(4) &= 001011000111.
\end{align*}
\]

The word \(g_{cs}(f_5^\infty(0))\) avoids SQ_3 and the set

\[F_{cs} = \{0000, 1111, 01010, 01011, 01100, 110010, 100110, 00011010110, 0110100111000, 110001101001, 1110001101001\}\]

**Theorem 6.** \(\{AABBCC, SQ_3\} \cup F_{cs}'\) and \(\{SQ_3\} \cup F_{cs}\) both characterize \(g_{cs}(f_5^\infty(0))\).

The factor complexity of \(g_{cs}(f_5^\infty(0))\) seems to be \(4n + 4\) for every factor length \(n \geq 6\).

4. Thue’s ternary pure morphic word

Thue [3, 10, 11] proved that \(\{AA\} \cup \{010, 212\}\) characterizes the fixed point of \(f_3\). In this section, we give characterizations of three words that are morphic images of \(f_3^\infty(0)\). It is not surprising that \(f_3^\infty(0)\) appears in the context of characterizations: as soon as a morphism \(m\) is such that \(m(0) = axb\) and \(m(1) = ab\), the \(m\)-image of words of the form \(0u1u0, u \in \Sigma^*_3\), contains a large square: \(m(0u1u0) = axbm(u)abm(u)axb\) contains \((bm(u)a)^2\). Moreover, a ternary square-free word avoids factors of the form \(0u1u0\) with \(u \in \Sigma^*_3\) if and only if it avoids \(\{010, 212\}\). So, the set of factors of a factorial language containing only square-free factors in \(\{m(0), m(1), m(2)\}\) such that \(m(0) = axb\) and \(m(1) = ab\) is the set of factors of \(m(f_3^\infty(0))\). It is also easy to check that \(\{AA\} \cup \{010, 212\}\) characterizes the same ternary word as \(\{AA\} \cup \{1021, 1201\}\).

4.1. Words containing one 5/2-repetition and 11 squares

If an infinite binary word contains the repetition 10101 of exponent 5/2 and no other overlap, then it contains at least 11 distinct squares. Moreover, if it contains exactly 11 distinct squares, then these 11 squares are \(0^2, 1^2, (01)^2,\)
(10)^2, (001)^2, (010)^2, (011)^2, (100)^2, (101)^2, (110)^2, (01100110)^2. Equivalently, a recurrent binary word containing these overlaps and squares avoids $SQ_7$ and the set

$$F_{11} = \{000, 111, 01010, 001100, 00010010, 0100100, 1011011, 1101101 \}.$$  

Let $g_{11}$ be the morphism from $\Sigma_3^*$ to $\Sigma_2^*$ defined by

$$g_{11}(0) = 10010011010110011010010110010011011000110100110100110100101100110101$$  
$$g_{11}(1) = 100100110100101$$  
$$g_{11}(2) = 10010011011001011010011010010110011010101.$$  

**Theorem 7.** $\{SQ_5\} \cup F_{11}$ characterizes $g_{11}(f_3^\infty(0))$. 

4.2. Words containing 3 squares

It is known that there exist exponentially many binary words containing only 3 distinct squares [7, 8]. Without loss of generality, we assume that these 3 squares are 00, 11, and 1010. To obtain a characterization, we forbid also the factors in $F'_3 = \{01000110, 10011101, 1001101000, 1110100110\}$. If $w$ is a recurrent binary word avoiding $F'_3$ and squares distinct from 00, 11, and 1010, then $w$ avoids $SQ_3$ and the set

$$F_3 = \{0000, 0101, 1111, 01000110, 10011101, 1001101000, 1110100110\}.$$  

Let $g_3$ be the morphism from $\Sigma_3^*$ to $\Sigma_2^*$ defined by

$$g_3(0) = 000111,$$  
$$g_3(1) = 0011,$$  
$$g_3(2) = 01001110001101.$$  

**Theorem 8.** $\{SQ_3\} \cup F_3$ characterizes $g_3(f_3^\infty(0))$. 

4.3. Words avoiding AABBCABBA

Another characterization has been obtained by the second author [9]:  

$$\{AABBCABBA\} \cup \{0011, 1100\}$$ characterizes $g_q(f_3^\infty(0))$, where $g_q$ is given below.

$$g_q(0) = 0010110111011101001,$$  
$$g_q(1) = 00101101101001,$$  
$$g_q(2) = 00010.$$  

Equivalently, $g_q(f_3^\infty(0))$ is characterized by $\{SQ_5\} \cup F_q$ where

$$F_q = \{0000, 0011, 1100, 1111, 01010, 10101, 010111, 101000, 0001001, 1110110, 00100100, 01011010, 10100101, 11011011, 0110111010, 1001000101\}.$$  

10
5. Concluding remarks

We have seen in Section 4 why \( f_3^\infty(0) \) appears often in the context of characterization. Also, we have seen in Section 3.1 why Thue’s words avoiding \( \{AA\} \cup \{010, 020\} \) and \( \{AA\} \cup \{121, 212\} \) use the same pure morphic word \( f_3^\infty(0) \). However, we do not see why \( f_5^\infty(0) \) is used in other “natural” languages. It would be interesting to investigate its properties, in particular to prove that its factor complexity is \( 4n + 1 \) and that its critical exponent is \( (5 + \sqrt{5})/4 \).

The fixed point of \( 0 \mapsto 01, 1 \mapsto 0 \), known as the Fibonacci word, seems to have the same set of factors as \( g_{fib}(f_5^\infty(0)) \), where \( g_{fib} \) is given below. Moreover, the Rote-Fibonacci word studied in [6] seems to have the same set of factors as \( g_{rt}(f_5^\infty(0)) \), where \( g_{rt} \) is given below.

\[
\begin{align*}
g_{fib}(0) &= 01, & g_{fib}(1) &= 01, \\
g_{fib}(1) &= 0, & g_{fib}(1) &= 10, \\
g_{fib}(2) &= 1, & g_{fib}(2) &= 0, \\
g_{fib}(3) &= 0, & g_{fib}(3) &= 11, \\
g_{fib}(4) &= 0. & g_{fib}(4) &= 00.
\end{align*}
\]

The method discussed in this paper is not able to prove such equivalences because the languages are not defined by avoiding large squares and a finite set of factors. Maybe it can be proven by the method used in [6] to recover many known results about the Fibonacci word.

Baker, McNulty, and Taylor [2] obtained that \( ABXBAYACZCAWBC \cup \{02\} \) characterizes the fixed point of \( 0 \mapsto 01, 1 \mapsto 21, 2 \mapsto 03, 3 \mapsto 23 \) over \( \Sigma_4 \). Notice that the forbidden factor 02 is not crucial here, its only role is to distinguish one out of three symmetric versions obtained by permutation of the alphabet letters. So, characterizations are known for the patterns \( AA, ABABA, ABCABC, AABBCCC, AABBCBABA, \) and \( ABXBAYACZCAWBC \). An interesting open question is the following: Suppose that \( P \) is an avoidable pattern with avoidability index \( \lambda(P) = k \). Is it possible to find a finite set \( P \) of patterns and a finite set \( F \) of factors such that \( P \in P \) and \( P \cup F \) characterizes a morphic word over \( \Sigma_k \)? This would be a strengthening of Cassaigne’s conjecture stating that there exists a morphic word avoiding \( P \) over \( \Sigma_k \).

References


