Theoretical Computer Science 410 (2009) 3893–3900

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

A powerful abelian square-free substitution over 4 letters

Veikko Keränen *

Rovaniemi University of Applied Sciences (RAMK), School of Technology, Jokiväylä 11, 96300 Rovaniemi, Finland

ARTICLE INFO

Article history:
Received 8 January 2009
Received in revised form 16 May 2009
Accepted 24 May 2009
Communicated by A.K. Salomaa

Keywords:
Abelian square
Square-free
Repetition avoidance
Exponential growth

ABSTRACT

In 1961, Paul Erdös posed the question whether abelian squares can be avoided in arbitrarily long words over a finite alphabet. An abelian square is a non-empty word uv, where u and v are permutations (anagrams) of each other. The case of the four letter alphabet $\Sigma_4 = \{a, b, c, d\}$ turned out to be the most challenging and remained open until 1992 when the author presented an abelian square-free (a-2-free) endomorphism g_{85} of Σ_4^*. The size of this g_{85}, i.e., $|g_{85}(abcd)|$, is equal to 4×85 (uniform modulus). Until recently, all known methods for constructing arbitrarily long a-2-free words on Σ_4 have been based on the structure of g_{85} and on the endomorphism g_{98} of Σ_4^* found in 2002.

In this paper, a great many new a-2-free endomorphisms of Σ_4^* are reported. The sizes of these endomorphisms range from 4×102 to 4×115. Importantly, twelve of the new a-2-free endomorphisms, of modulus $m = 109$, can be used to construct an a-2-free (commutatively functional) substitution σ_{109} of Σ_4^* with 12 image words for each letter.

The properties of σ_{109} lead to a considerable improvement for the lower bound of the exponential growth of c_n, i.e., of the number of a-2-free words over 4 letters of length n. It is obtained that $c_n > \beta^{-50}\beta^n$ with $\beta = 12^{1/m} \simeq 1.02306$. Originally, in 1998, Carpi established the exponential growth of c_n by showing that $c_n > \beta^{-t}\beta^n$ with $\beta = 2^{19/t} = 2^{19/(85^3-85)} \simeq 1.000021$, where $t = 85^3 - 85$ is the modulus of the substitution that he constructs starting from g_{85}.

1. Introduction

The systematic study of word structures, i.e., combinatorics on words, was initiated by Axel Thue (1863–1922) at the beginning of the 20th century in [1]. One of his discoveries was that the consecutive repetitions of non-empty factors (squares) can be avoided in arbitrarily long words over a three letter alphabet. As a simple example of the square concept, consider the words *abacaba* and *ab cd cd ab*. The first word does not contain any square, i.e., it is square-free, whereas the second word contains the underlined square *cd cd* as a factor.

The above-mentioned square-freeness property of words is not trivial to prove. The tool, which Thue invented for constructing square-free, and other repetition-free, words, namely the concept of a repetition-free morphism, is still today a basic device in the study of avoidable patterns in words. Repetition-free morphisms (respectively substitutions) are mappings between free monoids (respectively into monoid of subsets) that preserve the repetition-freeness of words. The iteration of a non-trivial repetition-free endomorphism or a substitution produces repetition-free words of any length. Dealing with substitutions somewhat later, we point out that repetition-free morphisms have been sharply characterised in [2–9]. Those results concern different types of repetitions (k-repetitions for a given integer $k \geq 2$) and alphabet sizes. Informally speaking, most of the characterisational results for morphisms mean that in order to test the repetition-freeness

* Tel.: +358 207985338.
E-mail addresses: veikko.keranen@ramk.fi, veikko.keranen@gmail.com.
URL: http://south.rotol.ramk.fi.

0304-3975/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2009.05.027

of a given morphism, one only has to check whether the image words of short repetition-free words are also repetition-free. A general survey of these and related results, achieved before 1984, is given in [10]. For a short survey of Thue's results concerning repetition-free words and their applications, see [11]. Fundamental topics are discussed in [12,13]. For a connection to modern bioinformatics, see for example Mirkin [14]. Indeed, it has been noted that short repetitions in DNA often lead to hypohelix structures (loops) that in turn may be the cause of many diseases. For the case of abelian square-free words, a number of visualisations of the structures and of the related processes can be found in [15].

In a paper from 1961, see [16, p. 240], Paul Erdös (1913–1996) raised the question whether abelian squares can be avoided in infinitely long words, i.e., whether there exist infinitely many abelian square-free words over a given alphabet. Here, an abelian square means a non-empty word uv, where u and v are permutations (anagrams) of each other. For example, $\underline{abc}\,\underline{acb}$ is an abelian square. A word is called abelian square-free, if it does not contain any abelian square as a factor. For example, the word *abacaba* is abelian square-free, while *ab* $\underline{cabdc}$ $\underline{bcacd}$ *ac* is not.

In 1970, Pleasants [17] proved that there exists an infinite abelian square-free word over five letters. Finally, in 1992, see Keränen [18], the author managed to show that the same holds true also in the case of four letters. It is easily seen that abelian squares cannot be avoided over a three letter alphabet. Indeed, in this alphabet, each word of length 8 contains an abelian square. Entringer et al. [19] proved that every infinite word over a binary alphabet contains arbitrarily long abelian squares. Dekking [20] in turn proved that abelian repetitions to the fourth power can be avoided in infinite words over two letters, and abelian repetitions to the third power (cubes) can be avoided in infinite words over three letters. Currie and Aberkane [21] exhibited the smallest cyclic binary morphism avoiding abelian fourth powers. At any rate, their morphism is far larger than the non-cyclic morphism presented by Dekking in [20]. For a generalisation of abelian squares, see Avgustinovich and Frid [22]. Abelian fractional powers were studied by Cassaigne and Currie [23]. It was proved by Currie [24] that the number of binary words avoiding abelian fourth powers grows exponentially, and Aberkane, Currie, and Rampersad [25] proved that the number of ternary words avoiding abelian cubes grows exponentially as well.

An application of Dekking's result was given by Justin et al. [26], who showed that a finitely generated semigroup is uniformly repetitive if and only if it is finite. Pirillo et al. [27] used similar kind of reasoning when proving, among other results, that the additive semigroup $\mathbb{N}^+$ is not uniformly 4-repetitive. It seems to be an open problem whether $\mathbb{N}^+$ is uniformly 2-repetitive or 3-repetitive. In all these considerations, the use of van der Waerden's theorem has been central. In Lothaire [12, pp. 55–62], van der Waerden's theorem was used to show that every morphism from a free semigroup A^+, where A is finite, to $\mathbb{N}^+$ is repetitive. This means that every long enough sequence on a finite set of integers contains two adjacent segments (not necessarily of the same length) that have the same sum.

The original problem concerning abelian squares has attracted attention also in the study of free partially commutative monoids, see for instance [28,29]. Moreover, abelian square-free words have aroused interest in algorithmic music, see e.g. Laakso [30], and quite recently in cryptography, see Rivest [31] and Andreeva et al. [32].

In 1993, Carpi [2] gave sufficient conditions for morphisms to preserve abelian kth power-freeness of words. A conjecture is that these conditions yield an effective characterisation also for abelian square-free endomorphisms on a four letter alphabet $\Sigma_4 = \{a, b, c, d\}$. However, new examples of relatively short abelian square-free endomorphisms g of Σ_4^* have turned out to be extremely hard to find — and the same difficulty applies to every systematic attempt for constructing long abelian square-free words over 4 letters. Before the results of this article, there was no evidence that it would be possible to find more examples of abelian square-free endomorphisms — not to speak of proper substitutions of Σ_4^*. Thus far, after 1992, when the author [18] presented g_{85}, the only new kind of abelian square-free endomorphisms and substitutions had been found by Carpi [33], cf. also [34, pp. 80–81]. However, his mappings are all based on the structure of g_{85}. Moreover, the image words of these endomorphisms and substitutions are very long indeed. By using his substitutions, Carpi proved that the number of abelian square-free words of each length grows exponentially, and that the monoid of (uniform) abelian square-free endomorphisms of Σ_4^* is not finitely generated.

The newest abelian square-free endomorphisms, 200 in number, that were recently found for a starting point regarding our substitutions, have the following property: the image words $g(x)$, $x \in \Sigma_4$, are all obtained by cyclically permutating the letters in $g(a)$. The sizes of these endomorphisms range from 4×102 to 4×115 and the image words $g(a)$ can be viewed and copied from [35]. The same cyclic permutation property is true for g_{85} as well, and this method was already used by Pleasants [17] in connection with five letters. Consequently, all of these endomorphisms have a uniform modulus and the generated words are growing uniformly. The size of Pleasants' endomorphism is $5 \times 15 = 75$. In the four letter case, the author has checked, with computer, that the size $4 \times 85 = 340$ of g_{85}, in spite of its largeness, is actually minimal, at least as far as cyclic permutation method is used. Except for Carpi's very large endomorphisms and substitutions, the earlier searches for the other kind of abelian square-free endomorphisms of Σ_4^* (not possessing the cyclic permutation property) have not been successful, even the experiments have been quite extensive. However, in Section 4, the a-2-free endomorphisms, 20724 in number, indeed possess a different structure. Moreover, the author [36,37] found in 2002 a nice endomorphism g_{98} of Σ_4^* that can be used in iterations, and also together with g_{85} to produce infinite abelian square-free DT0L-languages (i.e., languages obtained by using compositions of morphisms). This g_{98}, in itself, is not an abelian square-free endomorphism, as it does not preserve abelian square-freeness for all words (starting already from the length 7). The structure of the image words of g_{98} also partly differs from the structure of g_{85} and of the above-mentioned 200 new cyclic endomorphisms.

Quite recently, we have also gained insight into why these abelian square-free structures are so rare over four letters. The author in [38] (partially) explains this rareness of long words avoiding abelian squares by using the concept of an

unfavourable factor. Working in a fixed alphabet, we take an abelian square-free word, and with computer try to extend it in abelian square-free fashion to the right and to the left with all possible ways up to a given upper bound for the total length. At each step, the length of the word increases only by a given fixed length. We extend alternately to right and left, and backtrack if necessary. If the given upper bounds are reached then the original word is a *so-far-favourable* one (it may still turn out to be unfavourable on later experiments). If there is no way to reach the upper bounds, then the original word is classified, without any doubt, to be unfavourable. Thus there are three kinds of words: *unfavourable* (*bad*), *so-far-favourable* (*so-far-so-good*), and *favourable* (*good*). It is a remarkable phenomenon that already relatively short so-far-favourable words turn out to be unfavourable factors after being 'safely' extendable (to right and left) for quite a long distance and sometimes with a really huge number of branches. One might have expected the quite long buffers to guarantee the further growth. It is conjectured that the majority of abelian square-free words over four letters cannot occur as proper factors in the middle of very long (infinite) abelian square-free words. In a way, this also explains why it has been so difficult to find abelian square-free endomorphisms over four letters. At present, it is known, for example, that about 60 % of the abelian square-free words of length 24 are indeed unfavourable in the four letter case. It will be interesting to study in a similar way also the three letter case, for which an exciting open problem was posed by Mäkelä [39], who allows repetitions *xx* and *xxx*, for a letter *x*, but no other abelian squares (or cubes).

Ochem and Reix in [40] presented efficient matrix methods for finding upper bounds for the growth of the number of repetition-free words. Their approach is directly applicable also for the abelian square-freeness case by using the above-mentioned so-far-favourable words. However, the involved computations are still going on, and therefore this topic is not elaborated here.

2. Preliminaries

In this section, pertinent notation and terminology is presented. The reader might also consult this section later, if needed.

An *alphabet* Σ is a finite non-empty set of abstract symbols called *letters*. A *word* (*string*) over Σ is a finite (unless otherwise indicated) string, or sequence, of letters belonging to Σ. The set of all words [non-empty words] over Σ is denoted by $\Sigma^*[\Sigma^+]$. On the Σ^*, the associative binary operation of *catenation* is defined. For words u and v, it is the juxtaposition uv. The *empty word*, which is the neutral element of catenation, is denoted by λ. The algebraic structures Σ^* and Σ^+ are called, respectively, the free monoid and the free semigroup generated by Σ. For a word w and a natural number n, the notation w^n is defined by $w^0 = \lambda$ and $w^{n+1} = ww^n$.

Let $w = x_1 \cdots x_m$, $x_i \in \Sigma$. The *length* of the word w, denoted by $|w|$, is the number of occurrences of letters in w, i.e., $|w| = m$. Let $\Sigma = \{a_1, \ldots, a_n\}$. The number of occurrences of one letter $x \in \Sigma$ in w is denoted by $|w|_x$, or simply by $|w|_i$, if $x = a_i$. The notation $\psi_\Sigma(w)$ stands for the *Parikh vector* of w, i.e., $\psi_\Sigma(w) = (|w|_1, \ldots, |w|_n)$. Usually we will omit the subscript Σ and write simply ψ instead of ψ_Σ.

A word u is called a *factor* of a word w, if $w = pus$ for some words p and s. The notation FACT(w) stands for the set of all factors of w. If $p = \lambda$ [$s = \lambda$], then u is called a *prefix* [*suffix*] of w. By PREF(w) [SUFF(w)] we mean the set of all prefixes [suffixes] of w. The notation pref(w) [suff(w)] denotes an element in PREF(w) [SUFF(w)]. In the case $w \neq \lambda$, we write first(w) [last(w)] to denote the first [the last] letter of w.

Let $k \geq 2$ be a given integer. A *k-repetition* is a non-empty word of the form R^k. An *abelian k-repetition* is a non-empty word of the form $P_1 \cdots P_k$, where $\psi(P_\mu) = \psi(P_\nu)$ for all $1 \leq \mu < \nu \leq k$, i.e., P_i:s are *commutatively equivalent*, that is, they are permutations, or anagrams, of each other. Instead of [abelian] 2- and 3-repetitions, the terms [*abelian*] *squares* and *cubes* are often used. A word, or an ω-word (defined below), is called *k-repetition free*, or *k-free* for short, if it does not contain any k-repetition as a factor. A word sequence or a word set is *k-free*, if all words in it are k-free. Abelian analogs of these terms and definitions also exist and are formed in a natural way by preceding any term with the word abelian, i.e., abelian square, abelian cube, abelian k-repetition free, etc. The abelian analog of the short term, k-free is a-k-free. If, for a fixed k, it is possible to construct arbitrarily long (infinite) a-k-free (or other pattern-free) words over a given alphabet Σ, then we say that abelian k-repetitions (or those patterns) are *avoidable* over Σ.

Subsets of Σ^* are called *languages* over Σ. The *catenation* of languages L and L_1, denoted by LL_1, is the language $\{uv | u \in L, v \in L_1\}$. The notation L^i, where L is a language and $i \in \mathbb{N}$, is defined as follows: $L^0 = \{\lambda\}$ and $L^{i+1} = L^iL$. Let $L^* = \cup_{i=0}^{\infty} L^i$ and $L^+ = \cup_{i=1}^{\infty} L^i$. If a language contains only one word, say w, then we sometimes write w instead of $\{w\}$; especially, $w^* = \{w^i | i \geq 0\}$ and $w^+ = \{w^i | i \geq 1\}$.

For a language L we write $Q(L)$, where Q is FACT, PREF or SUFF, to denote the set $\cup_{w \in L} Q(w)$. Moreover, for $n \in \mathbb{N}$, we write $Q_n(L)$ to denote the words in $Q(L)$ of length $= n$. The notation $\text{pref}_n(w)$ [$\text{suff}_n(w)$] is used for an element in $\text{PREF}_n(w)$ [$\text{SUFF}_n(w)$].

A *morphism* h is a mapping between free monoids Σ^* and Δ^* satisfying $h(uv) = h(u)\,h(v)$ for every u and v in Σ^*. Especially, $h(\lambda) = \lambda$. A morphism $h : \Sigma^* \to \Delta^*$, being compatible with the catenation of words, is uniquely defined, if the word $h(x) \in \Delta^*$ is (effectively) given for each $x \in \Sigma$. If $\Delta = \Sigma$, we call h an *endomorphism* (and usually write g instead of h). For a morphism h and a language L we define $h(L) = \{h(w) \mid w \in L\}$. A morphism h is called *uniformly growing*, or, is said to have a *uniform modulus*, if $|h(x)| = |h(y)| \geq 2$ for every x and $y \in \Sigma$.

A substitution $\sigma : \Sigma^* \to \Delta^*$ is monoid morphism of Σ^* into a subset monoid of Δ^*. The substitution σ is finite, if $\sigma(\Sigma)$ is a finite subset of Δ^*. For a morphism $h : \Sigma^* \to \Delta^*$, it holds that $\text{Card}(h(\Sigma)) \leq \text{Card}(\Sigma)$, and thus a morphism is a special case of a finite substitution. Following the terminology of Carpi [34], a substitution $\sigma : \Sigma^* \to \Delta^*$ is called

commutatively functional, if $dom(\sigma) = \Sigma^*$, and, for all $x \in \Sigma$, $v, v' \in \sigma(x)$, it holds that $\psi(v) = \psi(v')$ (this is also written as $v \sim v'$). In other words, a substitution is termed commutatively functional, if the image words of a fixed letter are all commutatively equivalent. Moreover, for a commutatively functional substitution σ and any word w in Σ^*, all the words $\sigma(w)$ are commutatively equivalent.

For a given integer $k \geq 2$, a substitution (or a morphism) $\sigma : \Sigma^* \to \Delta^*$ is called *k-free* [*a-k-free*], if all the words $\sigma(w)$ are (or the word $\sigma(w)$ is) k-free [a-k-free] for every k-free [a-k-free] word $w \in \Sigma^*$.

With regard to *L-systems* (Aristid Lindenmayer 1925–1989), we specify the following concepts. A *D0L-system* is a triple $G = (\Sigma, g, \alpha_0)$, where Σ is an alphabet, $g : \Sigma^* \to \Sigma^*$ is an endomorphism, and α_0, called the *axiom*, is a word over Σ. The (*word*) *sequence* $S(G)$ generated by G consists of the words

$$\alpha_0 = g^0(\alpha_0), g^1(\alpha_0), g^2(\alpha_0), g^3(\alpha_0), \ldots,$$

where $g^i(\alpha_0) = g(g^{i-1}(\alpha_0))$ for $i \geq 1$. The *language* of G is defined by $L(G) = \{g^i(\alpha_0) | \, i \geq 0\}$. Languages [sequences] defined by a D0L-system are referred to as *D0L-languages* [*D0L-sequences*]. D0L-systems provide a very convenient way for defining languages and infinite words. Furthermore, if g and α_0 are k-free [a-k-free], then the iteration of g will yield a k-free [a-k-free] D0L-sequence. An *HD0L-system* is a 5-tuple $G_1 = (\Sigma, \Delta, g, h, \alpha_0)$, where (Σ, g, α_0) is a D0L-system, called the underlying D0L-system of G_1, Δ is an alphabet, and $h : \Sigma^* \to \Delta^*$ is a morphism. The *HD0L-sequence* $S(G_1)$ generated by G_1 consists of the words

$$h(\alpha_0) = h(g^0(\alpha_0)), h(g^1(\alpha_0)), h(g^2(\alpha_0)), h(g^3(\alpha_0)), \ldots,$$

and the *HD0L-language* of G_1 is the set $L(G_1) = \{h(g^i(\alpha_0)) | \, i \geq 0\}$. A DT0L-*system* is a triple $G_2 = (\Sigma, H, \alpha_0)$, where H is a finite non-empty set of morphisms (called tables) and (Σ, h, α_0) is a D0L-system for every $h \in H$. The *DT0L-language* of G_2 is the set $L(G_2) = \{w | \, w = \alpha_0 \text{ or } w = h_k \cdots h_1(\alpha_0)\}$, where the compositions $h_k \cdots h_1$ of morphisms are constructed from $h_1, \ldots, h_k \in H$. Obviously, a DT0L-system can be regarded as a D0L-system, when H contains only one endomorphism. For a thorough discussion of various L-systems the reader is referred to Rozenberg and Salomaa [41].

An *ω-word* is a right infinite sequence, of letters of an alphabet Σ. Thus an ω-word can be identified with a mapping of $\mathbb{N}_+$ into Σ. One can construct an ω-word, for example, by iterating an endomorphism $g : \Sigma^* \to \Sigma^*$ such that $\lambda \notin g(\Sigma)$ and $g(x) = xw$ for some $x \in \Sigma$, $w \in \Sigma^+$. Such a morphism g is called *prefix preserving* for the reason that $g^i(x)$ is a proper prefix of $g^{i+1}(x)$ whenever $i \geq 0$. An ω-word is obtained as the "limit" of the sequence $g^i(a)$; $i = 0, 1, 2, \ldots$.

3. Carpi's characterisations of a-2-free morphisms and substitutions

In this section, firstly, a result from [2], which characterises a-n-free morphisms, is restated. This result is then reformulated in such a way that the a-2-freeness can be tested for a commutatively functional substitution.

Proposition 1 (*Carpi [2]*). *Given an integer $n \geq 2$, two alphabets Σ and Δ, and a morphism $h : \Sigma^* \to \Delta^*$, let us denote by G_h the subgroup of $\mathbb{Z}^\Delta$ generated by $\psi(h(\Sigma))$. Then, the morphism h is abelian nth power-free (a-n-free) provided that the following conditions are satisfied:*

(i) *$h(w)$ is a-n-free for every a-n-free word $w \in \Sigma^*$ of length 2,*
(ii) *h is commutatively injective, i.e., the set $\psi(h(\Sigma))$ is linearly independent,*
(iii) *for all $a_i \in \Sigma$, $p_i \in PREF(h(a_i)) \setminus \{h(a_i)\}$; $0 \leq i \leq n$; such that*

$$\psi(p_{j+1}) - 2\psi(p_j) + \psi(p_{j-1}) \in G_h, \quad j = 1, 2, \ldots, n-1,$$

there exist integers $\delta_i \in \{0, 1\}$ such that

$$\psi(p_{j+1}) - 2\psi(p_j) + \psi(p_{j-1}) = \delta_{j+1}\psi(h(a_{j+1})) - 2\delta_j\psi(h(a_j)) + \delta_{j-1}\psi(h(a_{j-1})), \quad j = 1, 2, \ldots, n-1.$$

If $Card(\Sigma) \geq 6$, then an a-n-free morphism h always satisfies conditions (i)–(iii). □

Informally speaking, the condition (iii) above means that if $h(w)$ contains an abelian repetition as a factor, then there is also an abelian repetition in the preimage word w. In this paper, we are interested in applying this characterisation when $n = 2$ (and $j = 1$).

Independently of the author, Carpi in [2] verified with computer that the endomorphism g_{85}, presented in author's earlier publication [18], indeed satisfies the conditions (i)–(iii) of Proposition 1, for $n = 2$, and is therefore a-2-free.

Carpi characterised commutatively functional abelian square-free substitutions basically in a similar fashion in [34]. However, for our purpose, the notation in [34] is perhaps too plentiful and technical, cf. Proposition 7, Corollary 3 of Proposition 8, and Proposition 9 therein. For a visualisation of the structures involved the reader is referred to [33, p. 159].

Indeed, Proposition 1 can be reformulated in a straightforward way for testing the a-2-freeness of a commutatively functional substitution $\sigma : \Sigma^* \to \Delta^*$. In this case, by definition, for all $x \in \Sigma$, $v, v' \in \sigma(x)$, it holds that $\psi(v) = \psi(v')$. Consequently, for each letter x, the set $\psi(\sigma(x))$ simply consists of one single element, and thus $\psi(h(\Sigma))$ of Proposition 1 can be identified with $\psi(\{\sigma(x) | \, x \in \Sigma\})$. This leads to

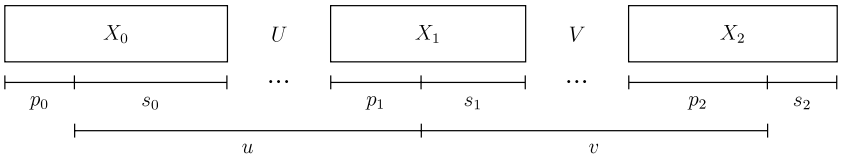

Fig. 1. Depicting the idea of condition (iii) of Proposition 2.

Proposition 2. *Given two alphabets Σ and Δ, and a commutatively functional substitution $\sigma : \Sigma^* \to \Delta^*$, let us denote by G_σ the subgroup of $\mathbb{Z}^\Delta$ generated by $\psi(\{\sigma(x) | x \in \Sigma\})$. Then, the substitution σ is abelian square-free provided that the following conditions are satisfied:*

(i) *$\sigma(w)$ is an a-2-free set for every a-2-free word $w \in \Sigma^*$ of length 2,*
(ii) *σ is commutatively injective, i.e., the set $\psi(\{\sigma(x) | x \in \Sigma\})$ is linearly independent,*
(iii) *for all $a_i \in \Sigma$, $p_i \in PREF(\sigma(a_i))$; $i = 0, 1, 2$; such that*

$$\psi(p_0) - 2\psi(p_1) + \psi(p_2) \in G_\sigma,$$

there exist integers $\delta_i \in \{0, 1\}$ such that

$$\psi(p_0) - 2\psi(p_1) + \psi(p_2) = \delta_0\psi(X_0) - 2\delta_1\psi(X_1) + \delta_2\psi(X_2),$$

where $\psi(X_i)$ is the unique element in $\psi(\sigma(a_i))$ for all $i = 0, 1, 2$. □

The idea of condition (iii) of Proposition 2 (and Proposition 1) can be explained as follows. Consider an image word $X_0UX_1VX_2 \in \sigma(w)$ of the word $w = a_0u_1a_1v_1a_2 \in \Sigma^*$, where $X_i \in \sigma(a_i)$, $a_i \in \Sigma$, $U \in \sigma(u_1)$, $V \in \sigma(v_1)$, $u_1, v_1 \in \Sigma^*$. Suppose there is an abelian square uv as a factor of $X_0UX_1VX_2$. Without loss of generality, one can restrict the study to just the cases depicted in Fig. 1, where $X_0 = p_0s_0 \in \sigma(a_0), X_1 = p_1s_1 \in \sigma(a_1), X_2 = p_2s_2 \in \sigma(a_2)$, with some of the p_i or s_i being possibly the empty word, and $\psi(v) - \psi(u) = \mathbf{0}$.

Now it is derived that

$$\begin{aligned}
\mathbf{0} &= \psi(v) - \psi(u)\\
&= \psi(s_1Vp_2) - \psi(s_0Up_1)\\
&= \psi(V) - \psi(U) + \psi(s_1) + \psi(p_2) - \psi(s_0) - \psi(p_1)\\
&= \psi(V) - \psi(U) + (\psi(X_1) - \psi(p_1)) + \psi(p_2) - (\psi(X_0) - \psi(p_0)) - \psi(p_1)\\
&= \psi(V) - \psi(U) + \psi(X_1) - \psi(X_0) + (\psi(p_0) - 2\psi(p_1) + \psi(p_2)),
\end{aligned}$$

that is, $\psi(p_0) - 2\psi(p_1) + \psi(p_2) = \psi(U) - \psi(V) + \psi(X_0) - \psi(X_1) \in G_\sigma$. In case there exist integers $\delta_i \in \{0, 1\}$ such that $\psi(p_0) - 2\psi(p_1) + \psi(p_2) = \delta_0\,\psi(X_0) - 2\delta_1\psi(X_1) + \delta_2\psi(X_2)$, one obtains

$$\psi(U) - \psi(V) = \delta_0\psi(X_0) - 2\,\delta_1\psi(X_1) + \delta_2\psi(X_2) + \psi(X_1) - \psi(X_0), \quad \text{for } \delta_i \in \{0, 1\},$$

and listing all the possible combinations for $\delta_0, \delta_1, \delta_2 \in \{0, 1\}$, it is found that

$$\begin{aligned}
\psi(U) - \psi(V) \in \{&\psi(X_1) - \psi(X_0), \psi(X_1X_2) - \psi(X_0), -\psi(X_0X_1), \psi(X_2) - \psi(X_0X_1), \psi(X_1), \psi(X_1X_2),\\
&-\psi(X_1), \psi(X_2) - \psi(X_1)\}
\end{aligned}$$

Remembering the a-2-freeness in condition (i), and the linear independence of the set $\psi(\{\sigma(x) | x \in \Sigma\})$ in condition (ii), it is established that whenever conditions (i)–(iii) are satisfied, the existence of an abelian square uv as a factor of $X_0UX_1VX_2 \in \sigma(w)$ implies the existence of an abelian-square in the preimage $w = a_0u_1a_1v_1a_2$.

4. The new a-2-free substitution σ_{109} over 4 letters

Let $\Sigma_4 = \{a, b, c, d\}$. Define the substitution $\sigma_{109} : \Sigma_4^* \to \Sigma_4^*$ as follows. First, let the 12 image words of $\sigma_{109}(a)$, say $\{A_1, A_2, \ldots, A_{12}\}$, have the form

$$\begin{aligned}
A_i &= p_{16}w_4u_{27}w_3s_{59}\\
&= abcacdcbcdcadcdb\ w_4\ badacdadbdcdbdabdbcbabcbdcb\ w_3\\
&\quad bdcdadcdbcbabcbdcbcacdcacbadabcbdcbcadbabcbabdbcdbdadbdcbca,
\end{aligned} \tag{1}$$

with 12 different factor pairs (w_4, w_3), taken in the natural lexicographical order form $\{abcd, abdc, adbc, dabc\} \times \{acd, adc, cad\}$. The subscripts of the factors $p_{16}, w_4, u_{27}, w_3, s_{59}$ indicate their lengths. Note that all the words in $\{abcd, abdc, adbc, dabc\}$, and respectively in $\{acd, adc, cad\}$, are commutatively equivalent and differ only by the movement of the letter d or c that creates the needed delicate mutations for the words of $\sigma_{109}(a)$.

To complete the definition of σ_{109}, let $\sigma_{109}(\phi(x)) = \phi(\sigma_{109}(x))$ for all $x \in \{a, b, c, d\}$, where $\phi : \Sigma_4^* \to \Sigma_4^*$ is the circular letter-to-letter endomorphism defined by $\phi(a) = b$, $\phi(b) = c$, $\phi(c) = d$, $\phi(d) = a$. Thus, informally, the set of image words for b, c, d are obtained, at a time, by cyclic permutation of letters of all the words in $\{A_1, A_2, \ldots, A_{12}\}$. Obviously, σ_{109} is a commutatively functional substitution of Σ_4^*. The Parikh vectors for the image words of letters are the rows of the matrix below:

$$\begin{pmatrix} \psi(A) \\ \psi(B) \\ \psi(C) \\ \psi(D) \end{pmatrix} = \begin{pmatrix} 21 & 31 & 29 & 28 \\ 28 & 21 & 31 & 29 \\ 29 & 28 & 21 & 31 \\ 31 & 29 & 28 & 21 \end{pmatrix},$$

whenever

$$A \in \sigma_{109}(a), B \in \sigma_{109}(b), C \in \sigma_{109}(c), D \in \sigma_{109}(d).$$

The author has checked the a-2-freeness of σ_{109} with computer in two albeit not completely different ways. One way was a direct but long method similar to that the author used in [18]. This method allowed us to make most of the computational steps visible, and provided a means of rechecking the computations. The other method in turn is an application of Proposition 2. In fact, Proposition 1 can also be utilised. One possible procedure regarding this latter method is explained here. However, for rechecking the properties of σ_{109}, one might like to use a slightly different approach than we describe. Indeed, it can be desirable to use just one endomorphism, say $g_{109,1}$ in (2) below, as a starting point, and then proceed to check all the cases in one run.

Let us recall, once again, that σ_{109} is a commutatively functional substitution. In what follows, Proposition 1 will be used firstly to test the a-2-freeness of endomorphisms in (2) and (3). One may start by checking that the conditions (i)–(iii) of Proposition 1 are satisfied for the 12 endomorphisms $g_{109,i}$ of Σ_4^*, defined by

$$\begin{aligned} g_{109,i}(a) &= A_i, \quad g_{109,i}(b) = B_i = \phi(A_i), \quad g_{109,i}(c) = C_i = \phi(B_i), \\ g_{109,i}(d) &= D_i = \phi(C_i), \quad i = 1, \ldots, 12. \end{aligned} \tag{2}$$

Indeed, the necessary computations can be accomplished quite quickly, because these morphisms have the same uniform modulus, and, for each morphism, the image words are obtained by cyclic permutation of letters. This is noticeable especially for condition (iii) of Proposition 1. These 12 endomorphisms $g_{109,i}$ were found in extensive computer searches and they are almost the same, apart from the slight mutations. Next, one checks that the conditions (i) and (iii) of Proposition 1 are satisfied for all the remaining combinations of the $12^4 - 12 = 20724$ endomorphisms $g_{109,ijkl}$ of Σ_4^*, defined by

$$\begin{aligned} g_{109,ijkl}(a) &= A_i, \quad g_{109,ijkl}(b) = B_j, \quad g_{109,ijkl}(c) = C_k, \\ g_{109,ijkl}(d) &= D_l, \quad i, j, k, l = 1, \ldots, 12, \end{aligned} \tag{3}$$

where the case $i = j = k = l$, due to (2), can be excluded.

Here the computations do not take a long time, because one can restrict testing of prefixes in condition (iii) of Proposition 1 to the lengths 17, 18, 19, 48, 49, i.e., one only needs to study those prefixes the endpoints of which are inside the (possibly) changing occurrences of w_4 and w_3, and their cyclic permutations (the occurrences of p_{16}, u_{27}, s_{59} and their cyclic permutations are the same in all the cases). These affirmative checkings guarantee that there exists at least $12^4 = 20736$ structurally different abelian square-free endomorphisms of Σ_4^* of uniform modulus 109. One still has to use condition (iii) of Proposition 2, and to check separately those cases of the prefixes (of length 17, 18, 19, 48, 49), in which two or three of the letters a_0, a_1, a_2 (in (iii) of Proposition 2) are the same but the image words are different. Note that these combinations cannot appear for the endomorphisms in (2) and (3) alone. However, these cases pose no extra difficulties. Now that all the possible prefix combinations have been checked, it is derived that if an abelian square is a factor of any of the $\sigma_{109}(w) \in \Sigma_4^*$, then the preimage w contains an abelian square as well.

In conclusion, we obtain the following

Proposition 3. *The substitution $\sigma_{109} : \Sigma_4^* \to \Sigma_4^*$ defined above is abelian square-free.*

Most likely, also new abelian square-free substitutions of Σ_4^* can be constructed from the other (than $g_{109,i}$, $g_{109,ijkl}$) a-2-free endomorphisms that were recently found. The image word $g(a)$ for all the new (original) 200 a-2-free endomorphisms, the sizes of which range from 4×102 to 4×115, g_{85} (found in 1990), and g_{98} (found in 2002), can be viewed and copied from [35].

5. A new lower bound for the exponential growth of a-2-free words over 4 letters

The properties of σ_{109} lead to a considerably sharper lower bound for the exponential growth of c_n, i.e., of the number of a-2-free words over 4 letters of length n. It is obtained that $c_n > \beta^{-50}\beta^n$ with $\beta = 12^{1/m} \simeq 1.02306$. Originally, the exponential growth of c_n was proved by Carpi [4], who showed that $c_n > \beta^{-t}\beta^n$ with $\beta = 2^{19/t} = 2^{19/(85^3-85)} \simeq 1.000021$, where $t = 85^3 - 85$ is the modulus of his substitution constructed from g_{85}.

The explanation for the new lower bound is as follows. Let $m = 109$, $\beta = 12^{1/m} \simeq 1.02306$, and an integer $n \geq 50$ be given. Note that in (1) we have $|p_{16}w_4u_{27}w_3| = 50$. Furthermore, let q be the quotient of $n - 50$ by m, i.e., $n - 50 = qm + r$, $0 \leq r < m$. There are at least $(\beta^m)^{q+1} = \beta^{qm+m}$ a-2-free words of length $qm + 50$, which are prefixes of some of the (infinite) a-2-free words $X_1X_2 \cdots X_i \cdots$, where $X_i \in \sigma_{109}(x_i)$, $x_i \in \Sigma_4$, for all $i \geq 1$. Thus there are at least $\beta^{qm+m} = \beta^{(n-50-r)+m} = \beta^{m-r}\beta^{n-50} > \beta^{n-50} = \beta^{-50}\beta^n$ words of length $n \geq 50$ that are prefixes of an a-2-free word $X_1X_2 \cdots X_i$ with $i > q$.

The number of all a-2-free words over 4 letters up to the length 60 can be found from [35].

Acknowledgements

Originally, the author in [42] presented the main results of this paper at WORDS'07, the 6th International Conference on Combinatorics on Words, Institut de Mathématiques de Luminy, Marseille.

The author thanks Gautam Dasgupta, Professor of Columbia University, for his continued encouragement.

The participation of several students of RAMK over the course of this study from 1990 to 2008 is also gratefully acknowledged. These students were responsible for coding a number of computer programs for searching strings with desirable properties, and set up and worked with the distributed computing environments at RAMK. The participating students, with their starting year given in parentheses are: Kari Tuovinen (1990), Minna Iivonen, Anja Keskinarkaus, Marko Manninen (1993), Abdeljalil Chabani, Tomi Laakso (1994), Mika Moilanen, Juha Särestöniemi (1996), Juho Alfthan (1999), Olli-Pentti Saira (2000), Marja Kenttä, Ville Mattila (2001), Lauri Autio, Marianna Mölläri (2002), Antti Eskola (2003), Antti Karhu, Veli-Matti Lahtela, Olli-Pekka Siivola (2004), Esa Nyrhinen, Sami Vuolli (2005), Esa Taskila, Mikhail Kalkov, Antti Oja, Viet Pham Hoang (2006), Alena Mekhnina, Shijing Zhang, Jing Lin, Irina Sekushina (2007), and Igor Melnikov (2008).

For the recent findings, the code written by Kari Tuovinen, Ville Mattila, Mikhail Kalkov, and Viet Pham Hoang has been used. Viet Pham Hoang was also responsible for building the final distributed computing environment and for executing the crucial search for candidate morphisms.

References

[1] A. Thue, Über unendliche Zeichenreihe, Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania 7 (1906) 1–22.
[2] A. Carpi, On abelian power-free morphisms, Internat. J. Algebra Comput. 3 (1993) 151–167.
[3] M. Crochemore, Sharp characterizations of square-free morphisms, Theoret. Comput. Sci. 18 (1982) 221–226.
[4] V. Keränen, On the k-freeness of morphisms on free monoids, in: Annales Academiæ Scientiarum Fennicæ, Ser. A. I. Mathematica Dissertationes, vol. 61, Finnish Science Academy, 1986, 55 pages.
[5] V. Keränen, On the k-freeness of morphisms on free monoids, in: F. Brandenburg, G. Vidal-Naquet, M. Wirsing (Eds.), Proc. STACS'87, in: Lecture Notes in Comp. Sci., vol. 274, Springer-Verlag, Berlin, 1987, pp. 180–188.
[6] M. Leconte, A characterization of power-free morphisms, Theoret. Comput. Sci. 38 (1985) 117–122.
[7] M. Leconte, Kth power-free codes, in: M. Nivat, D. Perrin (Eds.), Proc. Automata on Infinite Words'84, in: Lecture Notes in Comp. Sci., vol. 192, Springer-Verlag, Berlin, 1985, pp. 172–187.
[8] G. Richomme, F. Wlazinski, Some results on k-power-free morphisms, Theoret. Comput. Sci. 273 (2002) 119–142.
[9] F. Wlazinski, Ensembles de test et morphismes sans répétition, Thèse. Université de Picardie Jules Verne – LaRIA, 2002.
[10] J. Berstel, Some recent results on square-free words, in: M. Fontet, K. Melhorn (Eds.), Proc. STACS'84, in: Lecture Notes in Comp. Sci., vol. 166, Springer-Verlag, Berlin, 1984, pp. 14–25.
[11] G. Hedlund, Remarks on the work of Axel Thue on sequences, Nordisk Mat. Tidskr. 15 (1967) 148–150.
[12] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, Massachusetts, 1983.
[13] A. Salomaa, Jewels of Formal Language Theory, Computer Science Press, Rockville, Maryland, 1981.
[14] S. Mirkin, Expandable DNA repeats and human disease, Nature 447 (2007) 932–940.
[15] V. Keränen, Mathematica in word pattern avoidance research, in: A. Mylläri, V. Edneral, N. Ourusoff (Eds.), CADE 2007, Computer Algebra and Differential Equations, in: Acta Academiae Aboensis, Ser. B, Mathematica et physica, vol. 67, Åbo Akademi University Press, 2007, pp. 12–27. Available online at http://urn.fi/URN:ISBN:978-951-765-403-6.
[16] P. Erdös, Some unsolved problems, Magyar Tud. Kutató Int. Közl. 6 (1961) 221–254.
[17] P. Pleasants, Non-repetitive sequences, Proc. Camb. Phil. Soc. 68 (1970) 267–274.
[18] V. Keränen, Abelian squares are avoidable on 4 letters, in: W. Kuich (Ed.), Proc. ICALP'92, in: Lecture Notes in Comp. Sci, vol. 623, Springer-Verlag, Berlin, 1992, pp. 41–52.
[19] R. Entringer, D. Jackson, J. Schatz, On non-repetitive sequences, J. Combin. Theory Ser. A 16 (1974) 159–164.
[20] F. Dekking, Strongly non-repetitive sequences and progression-free sets, J. Combin. Theory Ser. A 27 (1979) 181–185.
[21] J. Currie, A. Aberkane, A cyclic binary morphism avoiding abelian fourth powers, Theoret. Comput. Sci. 410 (2009) 44–52.
[22] S. Avgustinovich, A. Frid, Words avoiding abelian inclusions, J. Autom. Lang. Comb. 7 (2002) 3–9.
[23] J. Cassaigne, J. Currie, Words strongly avoiding fractional powers, European J. Combin. 20 (1999) 725–737.
[24] J. Currie, The number of binary words avoiding abelian fourth powers grows exponentially, Theoret. Comput. Sci. 319 (2004) 441–446.
[25] A. Aberkane, J. Currie, N. Rampersad, The number of ternary words avoiding abelian cubes grows exponentially, J. Integer Seq. 7 (2004) 13 pages. Article 04.2.7.
[26] J. Justin, G. Pirillo, S. Varricchio, Unavoidable regularities and finiteness conditions for semigroups, in: A. Bertoni, C. Bohm, P. Miglioli (Eds.), Proc. 3rd Italian Conf. on Theoret. Comp. Sci. '89, World Scientific, Singapore, 1989, pp. 350–355.
[27] G. Pirillo, S. Varricchio, On uniformly repetitive semigroups, Semigroup Forum 49 (1994) 125–129.
[28] R. Cori, M. Formisano, Partially abelian square-free words, RAIRO Inform. Théor. Appl. 24 (1990) 509–520.
[29] V. Dickert, Research topics in the theory of free partially commutative monoids, Bull. Eur. Assoc. Theor. Comput. Sci. 40 (1990) 479–491.
[30] T. Laakso, Musical rendering of an infinite repetition-free string, in: C. Gefwert, P. Orponen, J. Seppänen (Eds.), Logic, Mathematics and the Computer, in: Proc. Finnish Artificial Intelligence Society, vol. 14, Hakapaino, Helsinki, 1996, pp. 292–297.
[31] R. Rivest, Abelian square-free dithering for iterated hash functions, MIT, 2005. Available online at http://people.csail.mit.edu/rivest/publications.html.
[32] E. Andreeva, C. Bouillaguet, P. Fouque, J. Hoch, J. Kelsey, A. Shamir, S. Zimmer, Second preimage attacks on dithered hash functions, in: N. Smart (Ed.), Advances in Cryptology – EUROCRYPT 2008, in: Lecture Notes in Comp. Sci., vol. 4965, Springer, Berlin / Heidelberg, 2008, pp. 270–288.
[33] A. Carpi, On the number of abelian square-free words on four letters, Discrete Appl. Math. 81 (1998) 155–167.

[34] A. Carpi, On abelian squares and substitutions, Theoret. Comput. Sci. 218 (1999) 61–81.
[35] V. Keränen, A-2-free endomorphisms and substitutions over 4 letters, 2007. Available online at http://south.rotol.ramk.fi/keranen/words2007/a2f.html.
[36] V. Keränen, New abelian square-free DT0L-languages over 4 letters, in: V. Demidov, V. Keränen (Eds.), Proc. IAS 2002, Murmansk State Pedagogical Institute and Rovaniemi Polytechnic, 2002. Available online at http://south.rotol.ramk.fi/keranen/ias2002/ias2002papers.html.
[37] V. Keränen, On abelian square-free DT0L-languages over 4 letters, in: T. Harju, J. Karhumäki (Eds.), Proc. WORDS'03, 4th International Conference on Combinatorics on Words, TUCS General Publication, Turku, 2003, pp. 95–109.
[38] V. Keränen, Suppression of unfavourable factors in pattern avoidance, in: B. Autin, Y. Papegay (Eds.), eProc. (CD) IMS'06, 8th International Mathematica Symposium, 2006, TMJ (in press).
[39] S. Mäkelä, Patterns in words, M.Sc. Thesis. Univ. Turku, 2002 (in Finnish).
[40] P. Ochem, T. Reix., Upper bound on the number of ternary square-free words, in: Proc. Workshop on Words and Automata, St Petersburg Department of Steklov Institute of Mathematics, 2006. Available online at http://www.lri.fr/~ochem/publi.htm.
[41] G. Rozenberg, A. Salomaa, The Mathematical Theory of L-systems, Academic Press, New York, London, Toronto, Sydney, San Fransisco, 1980.
[42] V. Keränen, New abelian square-free endomorphisms and a powerful substitution over 4 letters, in: P. Arnoux, N. Bédaride, J. Cassaigne (Eds.), Proc. WORDS'07, 6th International Conference on Combinatorics on Words, Institut de Mathématiques de Luminy, Marseille, 2007, pp. 189–200.