Discrete Mathematics and Theoretical Computer Science **3**, 1999, 125–140

Applying a uniform marked morphism to a word †

Anna Frid

Novosibirsk State University, Pirogova st. 2, Novosibirsk, Russia
E-mail: frid@math.nsc.ru

received July 30, 1998, revised October 25, 1999, accepted October 29, 1999.

We describe the relationship between different parameters of the initial word and its image obtained by application of a uniform marked morphism. The functions described include the subword complexity, frequency of factors, and the recurrence function. The relations obtained for the image of a word can be used also for the image of a factorial language. Using induction, we give a full description of the involved functions of the fixed point of the morphism considered.

Keywords: D0L words, HD0L words, subword complexity, functions of a word

1 Introduction

Different languages and words generated by morphisms play a significant part in the formal language theory. Such languages include D0L languages (and infinite words) which are obtained by iterating a morphism, HD0L languages (and words) which are obtained from a D0L language (respectively, a D0L word) by application of another morphism, and others.

To emphasize the importance of such languages, the theory of avoidable patterns can be mentioned. Examples of infinite words on a given alphabet which avoid a pattern are usually constructed as D0L words or HD0L words. Only recently it has been proved [6] that there exists a pattern which is avoidable on the binary alphabet but is not avoidable in any binary D0L word. However, it is still unknown if every k-avoidable pattern is avoidable by some HD0L word on the k-letter alphabet.

We consider a general case of simple application of a morphism to a word. Imposing some restrictions to the morphism considered (it must be uniform, marked, and circular on the initial word), we find explicit formulas linking the quantitative characteristics of the initial word and its image. This approach is valid also for the application of a morphism to a language (see Subsection 7.1) and allows to give a full description of these characteristics of a D0L word (see Subsection 7.2).

This research can be considered as an extension and generalization of papers [10, 9] (devoted to the subword complexity) and [8] (devoted to the frequency of factors), where the properties of a D0L word

† Supported in part by RFBR (grant 96-01-01800) and Federal Aim Program "Integration" (grant 473).

have been considered. However, here a new approach is used based on the direct relations between sets of allowable words of given lengths. This approach can be applied to find many functions of a language; some of them are considered here.

2 Preliminaries

2.1 Uniform marked morphisms

We consider two alphabets $\Sigma_1 = \{a_1, \ldots, a_q\}$ and Σ_2 and a morphism $\varphi : \Sigma_1^* \to \Sigma_2^*$; clearly, φ can be extended to the set Σ_1^ω of all (right) infinite words on Σ_1.

We call images $\varphi(a_i)$ of letters *blocks*; the length of a finite word u is denoted by $|u|$. A morphism is called *non-erasing* if all its blocks are non-empty; i.e., if for all letters $a_i \in \Sigma_1$ we have $|\varphi(a_i)| > 0$. We suppose φ to be non-erasing.

The (finite or infinite) word we consider is denoted by w; its image is denoted by $\varphi(w)$. Note, that since φ is non-erasing, $\varphi(w)$ is finite if and only if w is finite.

We say that a word u is *allowable* in a word v if it occurs in v as a factor: $v = s_1 u s_2$ for some words s_1 and s_2. The set of words allowable in w is denoted by A, and the set of words allowable in $\varphi(w)$ is denoted by B. Our goal is to describe the relationship between A and B.

A morphism φ is called *uniform* if all its blocks are of the same length: $|\varphi(a_i)| = m$ for all $a_i \in \Sigma_1$; m is called the *block length* of φ. If φ is uniform, then the length of the image of a word u depends only on the length of u; so, there exists a direct relationship between lengths of elements of A and of elements of B. That is why in this paper we consider uniform morphisms.

A morphism is called *marked* if all the images of letters begin with distinct symbols and end with distinct symbols; note that a morphism $\varphi : \Sigma_1^* \to \Sigma_2^*$ can be marked only if $|\Sigma_1| \leq |\Sigma_2|$. If φ is marked, then every block together with its inverse image is uniquely determined by its first (or last) symbol.

2.2 Circularity on w

For a word $u = u_0 u_1 \ldots u_{n+1} \in \Sigma_1^+$, where $u_i \in \Sigma_1$, we define $\Psi_{jk}(u) \in \Sigma_2^+$ as the word obtained from $\varphi(u)$ by erasing j symbols from the left and k symbols from the right. We need only the case when we do not erase full blocks; i.e., when $0 \leq j < |\varphi(u_0)|$ and $0 \leq k < |\varphi(u_{n+1})|$.

The triplet $s = (u, j, k)$, where $u = u_0 u_1 \ldots u_{n+1} \in \Sigma_1^+$, and j and k are integers, is called an *interpretation* of a word $v \in \Sigma_2^+$ if $0 \leq j < |\varphi(u_0)|$, $0 \leq k < |\varphi(u_{n+1})|$, and $v = \Psi_{jk}(u)$. The word u is called the *ancestor* of the interpretation s and is denoted by $a(s)$.

If in addition $u \in A$, then s is called an interpretation of v *on* w. Note, that a word can admit several interpretations, and even several interpretations on w with the same ancestor. The set of all the interpretations of a word v is denoted by $I(v)$, and the set of its interpretations on w is denoted by $I_w(v)$.

The notions of an interpretation and its ancestor were introduced by F. Mignosi and P. Séébold in [11].

Remark 1 $I_w(v) \neq \emptyset$ if and only if $v \in B$.

Let $u \in B$. A pair of words (u_1, u_2) is called a *synchronization point* of u (for φ) *on* w if $u = u_1 u_2$ and

$$\forall t_1, t_2 \in \Sigma_2^* \; \forall v \in A \; \exists v_1, v_2 \in A :$$

$$[t_1 u t_2 = \varphi(v) \Longrightarrow (v = v_1 v_2, t_1 u_1 = \varphi(v_1), u_2 t_2 = \varphi(v_2))].$$

The notion of synchronization point was introduced by J. Cassaigne in [4]. Put very simply, a synchronization point notes the point in u where the demarcation between blocks necessarily takes place in an occurrence of u in $\varphi(w)$.

The word $u \in B$ is called *circular* on w if it has a synchronization point on w.

Remark 2 If φ is uniform, then in a circular word all the demarcations between blocks are determined.

Remark 3 If φ is marked, then circularity of a word $u \in B$ means that it admits a unique interpretation on w: as we know a synchronization point, we can reconstruct every block of this interpretation from its first (or last) symbol. The ancestor of this interpretation is called the ancestor of the word u and is denoted by $a(u)$. We call u a *descendant* of $a(u)$.

The morphism φ is called *circular on* w with *synchronization delay* equal to $L = L(\varphi, w)$ if every word allowable in $\varphi(w)$ of length at least L is circular. This notion was inspired by the papers of Mignosi and Séébold [11] and Cassaigne [4].

Let φ be a uniform marked morphism with block length equal to m, and let φ be circular on w with synchronization delay equal to L. In addition to L, we use another parameter of circularity called *structure ordering number*; it is denoted by $K = K(\varphi, w)$ and defined as the least integer satisfying the inequality $m(K-1)+1 \geq L$. In what follows, we mostly use the synchronization delay through the mediation of the structure ordering number.

2.3 Example: circularity of the Thue-Morse morphism

Let us consider the famous Thue-Morse morphism φ_{TM} on the two-letter alphabet $\Sigma = \{a, b\}$:

$$\begin{cases} \varphi_{TM}(a) = ab \\ \varphi_{TM}(b) = ba \end{cases}$$

Clearly, φ_{TM} is a uniform marked morphism. Our goal is to find the set C_{TM} of words such that φ_{TM} is circular on a word w if and only if $w \in C_{TM}$.

Let every power of both symbols occur in w; i. e., let for all n $a^n \in A$, $b^n \in A$. In this case, the word $(ab)^n$ is allowable in $\varphi(w)$ and non-circular for all n. Actually, it admits two interpretations on w: $(a^n, 0, 0)$ and $(b^{n+1}, 1, 1)$. So, there exist arbitrarily long non-circular words in B, and $w \notin C_{TM}$.

Conversely, if a^{n+1} is not allowable in w for some n, then $w \in C_{TM}$; more precisely, φ is circular on w with synchronization delay $L \leq 2n+1$.

Actually, we shall prove that a word $u \in B$ of length $2n+1$ is always circular on w. If two equal symbols occur successively in u, then there is a demarcation between blocks between them; so, two successive equal symbols determine a synchronization point, and u is circular.

If there are no successive equal symbols in u, then $u = (ab)^n a$ or $u = (ba)^n b$. In both cases, u admits two interpretations; the ancestor of one of them is a^{n+1}, and the ancestor of the other is b^{n+1}. Since a^{n+1} is not allowable in w, and u is allowable, u admits one interpretation on w (with ancestor b^{n+1}) and is circular.

So, a word $u \in B$ of length $2n+1$ is always circular, and φ is circular on w with synchronization delay $L \leq 2n+1$.

Analogously, φ is circular on w if for some n the word b^{n+1} is not allowable in w. So, we have described the set C_{TM}:

$$C_{TM} = \{w \mid \exists n : (a^n \notin A) \vee (b^n \notin A)\}.$$

3 Basic relations

3.1 On the number of occurrences

For words u and v, we denote the number of occurrences of u in v by $L_u(v)$. The following easy lemma is valid for an arbitrary morphism $\varphi : \Sigma_1^* \to \Sigma_2^*$ and is very useful in the subsequent discussion.

Lemma 1 *For all* $u \in \Sigma_1^*$ *and* $v \in \Sigma_2^*$

$$L_v(\varphi(u)) = \sum_{s \in I(v)} L_{a(s)}(u) \tag{1}$$

Proof. Every occurrence of v in $\varphi(u)$ corresponds to an occurrence of the ancestor of certain its interpretation in u, and vice versa.

Remark 4 As it was noticed in Remark 3, if φ is marked, and v is circular, then v admits a unique interpretation with the ancestor which can be denoted by $a(v)$. In this case, Equality (1) can be rewritten as

$$L_v(\varphi(u)) = L_{a(v)}(u). \tag{2}$$

3.2 Sets of allowable words of given lengths

Hence-forward, unless otherwise specified, φ is a uniform morphism with block length equal to m, and w is an arbitrary (finite or infinite) word.

In what follows, we obtain some relations linking quantitative characteristics of sets A and B of allowable words. We shall also strengthen these relations by assuming that φ is marked and circular on w.

To state the next results, we need to introduce some more notations.

First, recall that for a word $u = u_0 \dots u_n$, $u_i \in \Sigma_1$, the word $\Psi_{jk}(u)$ is obtained from $\varphi(u)$ by erasing j symbols from the left and k symbols from the right; here we need only the case when we do not erase complete blocks, i. e., when $0 \le j, k < m$.

For a set $M \subset \Sigma_1^*$ and for $j, k \in \{0, \dots, m-1\}$, we define the set $\Psi_{jk}(M)$ naturally:

$$\Psi_{jk}(M) = \{\Psi_{jk}(u) | u \in M\}.$$

Clearly, in general case $|\Psi_{jk}(M)| \le |M|$.

The set of words of length n allowable in w (in $\varphi(w)$) is denoted by $A(n)$ (respectively, by $B(n)$):

$$A(n) = \{u | |u| = n, u \in A\},$$

$$B(n) = \{u | |u| = n, u \in B\}.$$

Now we can formulate the main theorem. As its corollaries, we shall obtain a variety of relations among the quantitative characteristics of A and B.

Theorem 1 *Let* $N = m(n-1) + \Delta + 1$ *for some* $n > 0$ *and some* $\Delta \in \{0, \dots, m\}$. *Then*

$$B(N) = \left(\bigcup_{i=1}^{m-\Delta} \Psi_{i-1, m-i-\Delta}(A(n)) \right) \bigcup \left(\bigcup_{i=1}^{\Delta} \Psi_{m-\Delta+i-1, m-i}(A(n+1)) \right). \tag{3}$$

If in addition φ is marked and circular on w with the structure ordering number equal to K, and $n \geq K$, then all the maps Ψ_{jk} above are one-to-one on the sets $A(n)$ (and, respectively, $A(n+1)$), and all the mentioned sets $\Psi_{jk}(A(n+l))$, $l \in \{0,1\}$, are mutually disjoint.

Proof. Let us consider a word $v \in B(N)$. Since it is allowable in $\varphi(w)$, it occurs as a factor in some word $\varphi(u)$, where $u \in A$. We choose u as short as possible, that is, v must contain parts of the first and the last blocks of $\varphi(u)$. In other terms, we choose such $u \in A$ that $v = \Psi_{jk}(u)$ for some $j,k \in \{0,\dots,m-1\}$. It can be easily seen that the length of u must be equal to n or to $n+1$. More exactly, either $u \in A(n)$, and then $v \in \Psi_{jk}(A(n))$, where $j+k = m-\Delta-1$ (this is possible only if $\Delta < m$), or $u \in A(n+1)$, and then $v \in \Psi_{jk}(A(n+1))$, where $j+k = 2m-\Delta-1$ (this is possible only if $\Delta > 0$).

On the other hand, if a word v is an element of $\Psi_{i-1,m-i-\Delta}(A(n))$ for some $i \in \{1,\dots,m-\Delta\}$ (or an element of $\Psi_{m-\Delta+i-1,m-i}(A(n+1))$ for $i \in \{1,\dots,\Delta\}$), then v is obviously allowable in $\varphi(w)$, and $|v| = m(n-1)+\Delta+1 = N$. So, $v \in B(m(n-1)+\Delta+1)$, and the equality is proved.

Now let φ be marked and circular on w with structure ordering number K, and let $n \geq K$.

Sets $\Psi_{jk}(A(n+l))$, $l \in \{0,1\}$, mentioned in the equation (3) can intersect only because of some of their elements having several interpretations; by the same reason, it may happen that a map Ψ_{jk} is not one-to-one on $A(n+l)$. However, since $|v| = N = m(n-1)+\Delta+1 \geq m(K-1)+1 \geq L$, every word $v \in B(N)$ is circular. Since φ is marked, v admits a unique interpretation. Thus, the theorem is proved.

Remark 5 *Here and in Theorems 2 and 3 some redundancy exists in the range of values of Δ. It would be sufficient to consider only $\Delta \in \{0,\dots m-1\}$ or $\Delta \in \{1,\dots,m\}$. The latter version will be more useful for us, but we do not want to loose the case of $N = m(K-1)+1$.*

In what follows, we derive some corollaries of Equalities (1) and (2) and Theorem 1 concerning functions of w and $\varphi(w)$.

4 Subword complexity

In this section, we study the subword complexity functions of w and $\varphi(w)$. The *subword complexity* $f_s(n)$ of a word s counts the number of distinct words of length n which are allowable in s. This function of infinite words has been studied in numerous papers; see, for instance, the survey [1] by J.-P. Allouche.

In our terms, the subword complexity functions of w and $\varphi(w)$ are $f_w(n) = |A(n)|$ and $f_{\varphi(w)}(n) = |B(n)|$. Here we derive a corollary of Theorem 1 which expresses the subword complexity of $\varphi(w)$ in terms of the subword complexity of w.

Theorem 2 *Let φ be a uniform morphism with block length equal to m. Let $N = m(n-1)+\Delta+1$ for some $n > 0$ and some $\Delta \in \{0,\dots,m\}$. Then*

$$f_{\varphi(w)}(N) \leq (m-\Delta)f_w(n) + \Delta f_w(n+1).$$

If in addition φ is circular on w and marked, and $n \geq K$, then equality holds:

$$f_{\varphi(w)}(N) = (m-\Delta)f_w(n) + \Delta f_w(n+1). \tag{4}$$

Proof. Using Equality (3), we obtain that

$$f_{\varphi(w)}(N) = |B(N)| \leq$$

$$\sum_{i=1}^{m-\Delta} |\Psi_{i-1,m-i-\Delta}(A(n))| + \sum_{i=1}^{\Delta} |\Psi_{m-\Delta+i-1,m-i}(A(n+1))| \leq$$
$$(m-\Delta)|A(n)| + \Delta|A(n+1)| = (m-\Delta)f_w(n) + \Delta f_w(n+1).$$

Clearly, if the maps Ψ_{jk} above are one-to-one on the respective sets $A(n)$ and $A(n+1)$, and the images $\Psi_{jk}(A(n+l))$, $l \in \{0,1\}$, are mutually disjoint, then all the inequalities of the latter derivation become equalities. Thus, it follows from Theorem 1 that if φ is marked and circular on w with structure ordering number equal to K, and $n \geq K$, then Equality (4) holds.

Remark 6 Note that the subword complexity function of $\varphi(w)$ does not depend on φ itself or even on the alphabet Σ_2. The only necessary parameters of φ are its block length and the structure ordering number of φ on w.

4.1 Example: the subword complexity of images of Sturmian words

An infinite word w is called *Sturmian* if its subword complexity is $f_w(n) = n+1$; as $f_w(1) = 2$, the alphabet of a Sturmian word is binary: $\Sigma_1 = \{a, b\}$. It can be easily seen that the subword complexity function of Sturmian words is the least possible for a non-ultimately periodic word.

Sturmian words were introduced in the classical paper [12] and have been extensively studied; see, for example, the survey [3]. There exists a variety of equivalent definitions of Sturmian words; however, we are interested just in the subword complexity.

A famous example of a Sturmian word is the Fibonacci word

$$w_F = abaababaabaababaababa\ldots.$$

Let w be a Sturmian word and φ be a circular on w uniform marked morphism; let its block length be equal to m and the structure ordering number be equal to $K(\varphi, w) = K$. It follows from Theorem 2 that for every $n \geq K$ and for every $\Delta \in \{0, \ldots, m-1\}$ the following equality holds:

$$f_{\varphi(w)}(m(n-1)+\Delta+1) = (m-\Delta)(n+1) + \Delta(n+2);$$

in other terms, for every $n \geq m(K-1)+1$

$$f_{\varphi(w)}(n) = n + 2m - 1.$$

This equality conforms Proposition 8 in [5] which states that a morphic image of a Sturmian word is quasi-Sturmian, i.e., has subword complexity $f(n) = n + c$ for some c and all $n \geq n_0$.

In particular, let φ be the Thue-Morse morphism φ_{TM} (see Example 2.3); its block length is $m = 2$. Without loss of generality we can state that the word *bb* never occurs in a Sturmian word (since in a non-ultimately periodic binary word *ab* and *ba* must occur, and $f_w(2) = 3$). So, it follows from Example 2.3 that the Thue-Morse morphism is circular on a Sturmian word with the synchronization delay $L \leq 3$ (in fact, $L = 3$). Structure ordering number is $K + 2$, and for $n \geq 3 = m(K-1)+1$ we have

$$f_{\varphi(w)}(n) = n + 3.$$

In particular, this is the subword complexity of the word

$$\varphi_{TM}(w_F) = abbaababbaabbaababbaababba\ldots.$$

5 Frequency of allowable words and frequency tables

In this section, we deal with the frequency of factors in an infinite word s on an alphabet Σ. Let $s(n)$ denote the prefix of length n of s; for every word $u \in \Sigma^+$ its frequency $\mu_s(u)$ in s is denoted as a limit

$$\mu_s(u) = \lim_{n\to\infty} \frac{L_{s(n)}(u)}{n} \tag{5}$$

Of course, this definition is valid only if this limit exists.

The set of frequencies μ_s determines a probability measure on the Borel subsets of Σ^w: the frequency $\mu_s(u)$ can be considered as the measure of the cylinder set $[u] = \{s | s = us',\ s' \in \Sigma^w\}$.

5.1 *Frequency of a word in* $\varphi(w)$

Our goal is to relate the frequencies of words in $\varphi(w)$ to the frequencies of words in w. Here we assume that w is an infinite word on Σ, and φ is a uniform morphism with block length equal to m.

Lemma 2 *If for every word $u \in \Sigma_1^+$ its frequency $\mu_w(u)$ exists, then for every $v \in \Sigma_2^+$ its frequency in $\varphi(w)$ exists and is equal to*

$$\mu_{\varphi(w)}(v) = \frac{1}{m} \sum_{s\in I_w(v)} \mu_w(a(s)). \tag{6}$$

Proof. Since ancestors $s \in I(v) \setminus I_w(v)$ do not occur in w, and $\varphi(w(n)) = \varphi(w)(mn)$, it follows from Lemma 1 that for every $n > 0$

$$L_{\varphi(w)(mn)}(v) = \sum_{s\in I_w(v)} L_{w(n)}(a(s)).$$

Dividing this equality by mn and passing to the limit $n \to \infty$, we obtain that

$$\begin{aligned} \mu_{\varphi(w)}(v) \quad &= \quad \lim_{n\to\infty} \frac{L_{\varphi(w)(mn)}(v)}{mn} = \lim_{n\to\infty} \frac{\sum_{s\in I_w(v)} L_{w(n)}(a(s))}{mn} = \\ &\frac{1}{m} \sum_{s\in I_w(v)} \lim_{n\to\infty} \frac{L_{w(n)}(a(s))}{n} = \frac{1}{m} \sum_{s\in I_w(v)} \mu_w(a(s)). \end{aligned}$$

A similar formula for a fixed point of a non-erasing morphism was proved in [8].

Using Formula (6), one can compute the frequency of a word in $\varphi(w)$ from the frequency in w of the ancestors of its interpretations. Note, that if φ is marked and circular on w, and the length of v exceeds the synchronization delay, then v has the only ancestor $a(v)$ (see Remark 3), and Formula (6) can be rewritten as

$$\mu_{\varphi(w)}(v) = \frac{1}{m}\mu_w(a(v)) \tag{7}$$

In what follows, we use Theorem 1 and Equation (7) to describe the set of frequencies of words of given length in $\varphi(w)$ in terms of frequencies of words in w. To do it, we introduce the notion of *frequency tables*.

5.2 Frequency tables

Let s be an infinite word with the subword complexity equal to $f(n)$. A *frequency table* on s is a table

$$F = \begin{array}{|c|c|c|c|}\hline k_1 & k_2 & \dots & k_l \\ \hline \mu_1 & \mu_2 & \dots & \mu_l \\ \hline \end{array},$$

where $k_i \geq 0$, $0 \leq \mu_i \leq 1$, $\mu_i \neq \mu_j$ for $i \neq j$, and a column $\begin{array}{|c|}\hline k_i \\ \hline \mu_i \\ \hline \end{array}$ means k_i distinct words allowable in s and having frequency μ_i.

The table F is called the *frequency table of s for* the length n and is denoted by $F_s(n)$ if $\sum_{i=1}^{l} k_i \mu_i = 1$, $\sum_{i=1}^{l} k_i = f(n)$, and F describes the frequencies of allowable words of length n in s.

It can be easily seen that adding columns of the kind $\begin{array}{|c|}\hline 0 \\ \hline \mu \\ \hline \end{array}$ (*zero columns*) and permuting columns do not change the meaning of a frequency table. We do not distinguish tables which differ by the order of columns or by zero columns.

The sum $F + F'$ of two frequency tables

$$F = \begin{array}{|c|c|c|c|}\hline k_1 & k_2 & \dots & k_l \\ \hline \mu_1 & \mu_2 & \dots & \mu_l \\ \hline \end{array} \text{ and } F' = \begin{array}{|c|c|c|c|}\hline k'_1 & k'_2 & \dots & k'_{l'} \\ \hline \mu'_1 & \mu'_2 & \dots & \mu'_{l'} \\ \hline \end{array}$$

is defined as the table obtained from

$$\begin{array}{|c|c|c|c|c|c|c|c|}\hline k_1 & k_2 & \dots & k_l & k'_1 & k'_2 & \dots & k'_{l'} \\ \hline \mu_1 & \mu_2 & \dots & \mu_l & \mu'_1 & \mu'_2 & \dots & \mu'_{l'} \\ \hline \end{array}$$

by junction of columns corresponding to the same frequency: instead of two columns $\begin{array}{|c|}\hline k_i \\ \hline \mu \\ \hline \end{array}$ and $\begin{array}{|c|}\hline k'_j \\ \hline \mu \\ \hline \end{array}$, the column $\begin{array}{|c|}\hline k_i + k'_j \\ \hline \mu \\ \hline \end{array}$ should be written.

For a frequency table

$$F = \begin{array}{|c|c|c|c|}\hline k_1 & k_2 & \dots & k_l \\ \hline \mu_1 & \mu_2 & \dots & \mu_l \\ \hline \end{array},$$

and for $p \geq 0$, $r > 0$, we define a table $T_r^p(F)$ as follows:

$$T_r^p(F) = \begin{array}{|c|c|c|c|}\hline pk_1 & pk_2 & \dots & pk_l \\ \hline \frac{1}{r}\mu_1 & \frac{1}{r}\mu_2 & \dots & \frac{1}{r}\mu_l \\ \hline \end{array}.$$

Theorem 3 *Let the frequency μ_w be defined for every word on Σ_1, and let φ be a uniform marked morphism circular on w. If n exceeds the structure ordering number of φ on w, $\Delta \in \{0, \dots, m\}$, and $N = m(n-1) + \Delta + 1$, then*

$$F_{\varphi(w)}(N) = T_m^{m-\Delta}(F_w(n)) + T_m^{\Delta}(F_w(n+1)). \tag{8}$$

Proof. The statement of the theorem follows immediately from Theorem 1 and Formula (7): each word $u' \in A$ of length n with frequency μ' in w is an ancestor of $m - \Delta$ words of length N allowable in φ with frequency $\frac{1}{m}\mu'$: they are words $\Psi_{i-1,m-i-\Delta}(u')$, where $i \in \{1, \ldots, m - \Delta\}$. Analogously, each allowable in w word of length $n + 1$ and of frequency μ'' is an ancestor of Δ words of length N allowable in $\varphi(w)$; each of them has frequency $\frac{1}{m}\mu''$. Since every allowable in $\varphi(w)$ word of length N has an ancestor of length n or $n + 1$ allowable in w, the theorem is proved.

5.3 Example: frequency of factors in the Thue-Morse image of the Fibonacci word

Let us compute the set of frequencies of factors in the word $\varphi_{TM}(w_F)$ (see Examples 2.3 and 4.1) using frequency of factors in the Fibonacci word w_F described by M. Dekking in [7].

Let P_l denote the lth Fibonacci number: $P_0 = 0$, $P_1 = 1$, $P_{k+1} = P_k + P_{k-1}$ for $k \geq 2$; let $p = \dfrac{\sqrt{5}-1}{2}$. Every length n can be uniquely decomposed as $n = P_l + r + 1$, where $r \in \{0, \ldots, P_{l-1} - 1\}$.

In terms of frequency tables, the frequency of factors of length n in w_F described in [7] is

$$F_{w_F}(n) = \begin{array}{|c|c|c|}\hline P_{l-1} - r & P_{l-2} + r & r \\ \hline p^{l-2} & p^{l-1} & p^l \\ \hline \end{array}.$$

This table means that $P_{l-1} - r$ of $n + 1$ allowable words of length n have frequency p^{l-2}, $P_{l-2} + r$ words have frequency p^{l-1}, and r words have frequency p^l.

Note, that $F_{w_F}(n + 1)$ looks analogously even if r is maximal ($r = P_{l-1} - 1$):

$$F_{w_F}(n+1) = \begin{array}{|c|c|c|}\hline P_{l-1} - r - 1 & P_{l-2} + r + 1 & r + 1 \\ \hline p^{l-2} & p^{l-1} & p^l \\ \hline \end{array}.$$

Now, let us apply Theorem 3 to find frequencies of factors in the Thue-Morse image $\varphi_{TM}(w_F)$ of the Fibonacci word. Let $N = 2n - 1 + \Delta$ for $\Delta \in \{0, 1\}$; then $N = 2P_l + r' + 1$, where $r' = 2r + \Delta \in \{0, \ldots, 2P_{l-1} - 1\}$. Equality (8) applied to N is

$$F_{\varphi_{TM}(w_F)}(N) = T_2^{2-\Delta}(F_{w(F)}(n)) + T_2^{\Delta}(F_{w_F}(n+1)) =$$

$$\begin{array}{|c|c|c|}\hline (2-\Delta)(P_{l-1} - r) & (2-\Delta)(P_{l-2} + r) & (2-\Delta)r \\ \hline \frac{1}{2}p^{l-2} & \frac{1}{2}p^{l-1} & \frac{1}{2}p^l \\ \hline \end{array} +$$

$$+ \begin{array}{|c|c|c|}\hline \Delta(P_{l-1} - r - 1) & \Delta(P_{l-2} + r + 1) & \Delta(r + 1) \\ \hline \frac{1}{2}p^{l-2} & \frac{1}{2}p^{l-1} & \frac{1}{2}p^l \\ \hline \end{array} =$$

$$\begin{array}{|c|c|c|}\hline 2P_{l-1} - r' & 2P_{l-2} + r' & r' \\ \hline \frac{1}{2}p^{l-2} & \frac{1}{2}p^{l-1} & \frac{1}{2}p^l \\ \hline \end{array}.$$

We have found the set of frequencies of factors of given length in the Thue-Morse image of the Fibonacci word.

6 The recurrence function and its relatives

Let $A_s(n)$ be the set of all words of length n which are allowable in a word s. The *recurrence function* $R_s(n)$ of s can be defined as the least length satisfying the following condition: all the words of $A_s(n)$ are allowable in *every* word $u \in A_s(R_s(n))$. In other terms, $R_s(n)$ is the size of the smallest window which contains all the elements of $A_s(n)$ whatever its position in s.

The recurrence function is the classical tool associated with infinite words. Recently, two more functions $R'_s(n)$ and $R''_s(n)$ slightly different from $R_s(n)$ have been defined. The function $R'_s(n)$ introduced by J.-P. Allouche and M. Bousquet-Mélou in [2] is the length of the shortest *prefix* of s containing all the words of $A_s(n)$; and the function $R''_s(n)$ introduced by J. Cassaigne in [5] is the length of the *shortest* allowable word in which every word from $A_s(n)$ is allowable.

Clearly, $R''_s(n) \leq R'_s(n) \leq R_s(n)$. The recurrence function is finite only if s is *uniformly recurrent*, i. e., if for every word u allowable in s the distance between two successive occurrences of u in s is bounded; if it is not for some u of length n, $R_s(n)$ is defined to be $+\infty$. The other two functions are defined everywhere.

Unlike R'_s, the functions R_s and R''_s depend in fact only on the set of words allowable in s.

Let the word w be infinite. Here we use Equality (2) and Theorem 1 to relate these functions of $\varphi(w)$ to the corresponding functions of w; we assume here, that φ is a uniform marked morphism circular on w.

Theorem 4 *Let w be an infinite word, and φ be a uniform marked morphism circular on w. Let n exceed the structure ordering number of φ on w. Then for every $\Delta \in \{1, \ldots, m\}$ and for $N = m(n-1)+\Delta+1$*

$$R_{\varphi(w)}(N) = mR_w(n+1) - m + \Delta.$$

Proof. Let $R_w(n+1)$ be finite.

It can be easily seen that the recurrence function $R(n)$ is directly connected with the maximal distance between two adjacent occurrences of a word of length n. The latter function is equal to $R(n) - n + 1$ and is non-decreasing.

Note that Δ here is chosen to be larger than 0. Thus, it follows from Theorem 1 that every word of A of length $n+1$ has a descendant of B of length N. Since φ is uniform, and due to circularity, if the length between two occurrences of a word in w is l, then the length between corresponding occurrences of its descendants in $\varphi(w)$ is ml. In terms of the recurrence function it means that

$$R_{\varphi(w)}(N) - N + 1 = m(R_w(n+1) - n).$$

Substituting to this equality the expression for N, we obtain the statement of the theorem.

Now let $R_w(n+1)$ be infinite. It means that there exist as long as is wished words of A in which not all the words of $A(n+1)$ occur. But due to Theorem 1, their images cannot contain all the words of $B(N)$; thus, $R_{\varphi(w)}(N)$ is also infinite.

Theorem 5 *Under conditions of Theorem 4,*

$$R'_{\varphi(w)}(N) = mR'_w(n+1) - m + \Delta.$$

Proof. Let u be the prefix of w of length $R'_w(n+1)$: $u = w(R'_w(n+1))$. By the definition of the function R'_w, all the elements of $A(n+1)$ (and, consequently, of $A(n)$) are allowable in u, and the suffix v_1 of u of length $n+1$ does not occur earlier in it. However, since the suffix v_0 of u of length n is the prefix of some allowable word of length $n+1$, it occurs somewhere else in u.

Since N is larger than the synchronization delay of φ on w, every word allowable in $\varphi(w)$ of length N is a descendant of some word of length n or $n+1$. Since $\Delta \geq 1$, it follows from Equality (3) that every word allowable in w of length $n+1$ (and in particular v_1) is an ancestor of some word of length N allowable in $\varphi(w)$.

It follows from Theorem 1 applied to u that $\varphi(u)$ contains as factors all the elements of $B(N)$; so, $R'_{\varphi(w)}(N) \leq mR'_w(n+1)$. On the other hand, unlike the descendants of v_0 which occur somewhere earlier in $\varphi(w)$, each of the descendants of v_1 (including $\Psi_{m-1,m-\Delta}(v_1)$) occurs in $\varphi(w)$ only once. Thus, the last word from $B(N)$ which has a unique occurrence in $\varphi(u)$ is $\Psi_{m-1,m-\Delta}(v_1)$, and so $R'_{\varphi(w)}(N) = mR'_w(n+1) - m + \Delta$.

As for the function R'', to find the relationship between that of w and of $\varphi(w)$, we need some additional properties of w (or, more generally, of the language A) to be satisfied.

Namely, a language $L \subset \Sigma^*$ is called *prolongable* if for every word $u \in L$ there exist symbols $a, b \in \Sigma$ such that $au \in L$, $ub \in L$.

A word $u \in A$ is called *(right) special* if it can be prolonged to the right to an element of A by at least two different letters: $ua, ub \in A$ for some $a, b \in \Sigma_1$, $a \neq b$. Note that a sequence w is non-ultimately periodic if and only if for each n there exists a special word of length n.

Theorem 6 *Under conditions of Theorem 4, let the language A be prolongable and w be non-ultimately periodic. Then*

$$R''_{\varphi(w)}(N) = mR''_w(n+1) - 2m + 2\Delta.$$

Proof. Let us call a word $u \in A(R''_w(n+1))$ *trivial* if it begins and ends with the same word $s \in A(n)$ which does not occur anywhere else in u: $u = su' = u''s$, but $u \neq u_1su_2$ for $u_1, u_2 \in \Sigma_1^+$.

Lemma 3 *There exists a non-trivial word in $A(R''_w(n+1))$ containing all the words of $A(n+1)$.*

Proof. Let us choose a word $v \in A(R''_w(n+1))$ containing all the words of $A(n+1)$: by the definition of R''_w, such a word exists. If it is non-trivial, then the lemma is proved; otherwise, $u = su' = u''s$ for some $s \in A(n)$. Since s does not occur anywhere in the middle of u, it can be prolonged to the right only by the first letter of u' (which is denoted by a). Thus, s is not special, and neither is u: the only word of $u\Sigma_1 \cap A$ is ua.

Let $s = bs'$ for $b \in \Sigma_1$. Consider the word $s'u'a \in A(R''_w(n+1))$. It contains all the words of $A(n+1)$: all of them except sa are factors of $s'u'$, and sa is a suffix of $s'u'a$. If $s'u'a$ is non-trivial, then the lemma is proved; otherwise, it is not special, and we can repeat the procedure and obtain from $s'u'a$ a new word by erasing the first letter and adding the allowable letter to the right.

However, if we could repeat this procedure infinitely many times, if would mean the existence of an infinite suffix of w not containing special words of length $R''_w(n+1)$. This is possible only if w is ultimately periodic which contradicts with the conditions of the theorem.

Since $m > 0$, it follows from Theorem 1 that a word cannot contain the elements of $B(N)$ if its ancestor does not contain all the words of $A(n+1)$. We must look for shortest words containing all the elements of $B(N)$ among descendants of words of $A(R''_w(n+1))$ containing all the elements of $A(n+1)$.

Let a word $u_0 \in A(R''_w(n+1))$ contain all the words of $A(n+1)$ and be non-trivial. By the definition of R''_w, its prefix and suffix of length $n+1$ do not occur anywhere else in it, but since A is a prolongable language, the prefix and the suffix of u_0 of length n occur in it somewhere else; since u_0 is non-trivial, each of them occur in the middle of u_0 (even if they are equal). So, to obtain the shortest descendant v_0 of u_0 containing all the words of $B(N)$, we can erase all the symbols from the left (and from the right) of $\varphi(u_0)$ which occur only in the descendants of length N of these prefix and suffix of length n. These symbols are the first and the last $m - \Delta$ symbols of $\varphi(u_0)$; so, $v_0 = \Psi_{m-\Delta,m-\Delta}(u_0)$.

On the other hand, if $u_1 \in A(R''_w(n+1))$ contains all the words of $A(n+1)$ and is trivial, we must be more cautious with erasing letters from the left and from the right of its φ-image, and the descendants of u_1 containing all the words of $B(N)$ are longer than v_0. Thus,

$$R''_{\varphi(w)}(N) = |v_0| = mR''_w(n+1) - 2m + 2\Delta.$$

6.1 Example: On words with ultimately grouped factors

In [5], J. Cassaigne introduced a notion of *words with ultimately grouped factors*: a word w if said to have ultimately grouped factors if $R''_w(n) = f_w(n) + n - 1$ for all $n \geq n_0(w)$. He proved also that every quasi-Sturmian word (i.e., a word whose subword complexity is $f_w(n) = n + c$ for some c and for all $n \geq n_1$) has ultimately grouped factors.

Let w be a non-ultimately periodic word with ultimately grouped factors; let the language of factors of w be prolongable, and a uniform marked morphism φ be circular on w. If n exceeds the structure ordering number of φ on w and $n_0(w)$, and $N = m(n-1) + \Delta + 1$ for $\Delta \in \{1, \dots, m\}$, then it follows from Theorem 6 and Formula (4) that

$$R''_{\varphi(w)}(N) = m(F_w(n+1) + n) - 2m + 2\Delta =$$

$$f_{\varphi(w)}(N) + N - 1 + (m - \Delta)(f_w(n+1) - f_w(n) - 1).$$

Thus, $\varphi(w)$ has ultimately grouped factors if and only if $f_w(n+1) = f_w(n) + 1$ for all large enough n; i. e., if and only if w is a quasi-Sturmian word.

7 Some generalizations

7.1 A generalization to factorial languages

It can be easily seen that sets of allowable words of given length in fact do not depend on the words w and $\varphi(w)$, but rather on the sets A and B. The same can be said on the subword complexity, the recurrence function, and its relative R''. But the initial language A can be defined not only as the set of words which are allowable in a word w. The only condition it must satisfy is factually closure under taking a factor: if a word u is an element of A, then all its factors are also elements of A. In other terms, A must be a *factorial* language on Σ_1.

For a factorial language A, we define its *factorial image* $B \subset \Sigma_2^*$ as the least factorial language containing the set $\varphi(A) = \{\varphi(u) | u \in A\}$. Properties of A and B are similar to those of sets of allowable words of w and $\varphi(w)$ respectively.

Replacing everywhere in Subsection 2.2 w by A and $\varphi(w)$ by B, we obtain the necessary definitions concerning circularity on a factorial language A (see also Subsection 3.2 in [4]).

The subword complexity function $f_L(n)$ of a factorial language L is defined naturally as the number of its distinct elements of length n.

It can be easily seen that after replacing circularity on w by circularity on A, we can extend Theorem 1 to an arbitrary factorial language A. The same can be done with Theorem 2 which is a direct corollary of Theorem 1; in the relations between subword complexity values, f_w should be replaced by f_A and $f_{\varphi(w)}$ should be replaced by f_B.

As it was mentioned in Section 6, the recurrence function R and its relative R'' depend in fact not on the words w and $\varphi(w)$ themselves, but on the sets of allowable words. So, similar functions can be defined on an arbitrary factorial language L: $R_L(n)$ is the least length such that *every* element of L of length $R_L(n)$ contains as factors all the elements of L of length n, and $R''_L(n)$ is the length of the *shortest* element of L which contains all the words from L of length n. On a factorial language, both these functions can be not everywhere defined; if the length from a definition does not exist, the corresponding function is supposed to be infinite.

Let A be a factorial language and B be its factorial image. After replacing the occurrences of w by those of A, and the occurrences of $\varphi(w)$ by those of B, Theorem 4 can be extended to a factorial language A and its factorial image B. To extend Theorem 6, we must just add the conditions for A to be prolongable and to have infinitely many special elements (the latter condition serves instead of non-periodicity of w).

7.2 A generalization to D0L words

Let the alphabets Σ_1 and Σ_2 be equal; i.e., let φ be a morphism $\varphi : \Sigma^* \to \Sigma^*$. If the block $\varphi(a)$ begins with a for some symbol $a \in \Sigma$, then φ admits a *fixed point* w which can be obtained as a limit

$$w = \lim_{i\to\infty} \varphi^i(a);$$

if φ is non-erasing and $\varphi(a) \neq a$, then w is an infinite word satisfying the equality $w = \varphi(w)$.

Fixed points of morphisms are called also D0L words. They have been studied in numerous papers and contain many famous examples, including the Fibonacci word (which is the fixed point of the morphism φ_F defined by $\varphi_F(a) = ab$, $\varphi_F(b) = a$) and the Thue-Morse word which is a fixed point of the Thue-Morse morphism φ_{TM} (see Example 2.3).

In [13], the probability measure defined by the set of frequencies of factors in a D0L word has been studied, and a sufficient condition of the existence of such a measure was given.

A fixed point w of a morphism φ is called *circular* if φ is circular on w. An easy-to-check criterion of circularity of a fixed point of a uniform marked morphism was proved in [9].

If w is a circular fixed point of a uniform marked morphism φ, then the formulas obtained above can be applied to w many times, until the length n considered is less than the structure ordering number of φ on w.

More precisely, let K be the structure ordering number of φ on w; let m be the block length of φ. For every $n \geq K$ there exists a unique triplet of *decomposition parameters* $(p(n), k(n), \Delta(n))$ such that

$$\begin{array}{l} p(n) \geq 0 \\ k(n) \in \{K, \dots, m(K-1)\} \\ \Delta(n) \in \{1, \dots, m^{p(n)}\}, \text{ and} \end{array}$$

$$n = m^{p(n)}(k(n) - 1) + \Delta(n).$$

Theorem 7 *For all $n \geq K$*

$$
\begin{aligned}
f_w(n+1) &= (m^{p(n)} - \Delta(n)) f_w(k(n)) + \Delta(n) f_w(k(n)+1), && (9)\\
R_w(n+1) &= m^{p(n)}(R_w(k(n)+1)-1) + \Delta(n), && (10)\\
R'_w(n+1) &= m^{p(n)}(R'_w(k(n)+1)-1) + \Delta(n), && (11)\\
R''_w(n+1) &= m^{p(n)}(R''_w(k(n)+1)-2) + 2\Delta(n); && (12)
\end{aligned}
$$

if the frequency is defined for every factor of w, then

$$F_w(n+1) = T_{m^{p(n)}}^{m^{p(n)}-\Delta(n)}(F_w(k(n))) + T_{m^{p(n)}}^{\Delta(n)}(F_w(k(n)+1)). \tag{13}$$

Proof. The proof is carried out by induction on $p(n)$. If $p(n) = 0$, then Formulas (9)–(13) are obviously correct. The induction step is given by Theorems 2–6.

As it follows from Lemma 3 of [11], a circular fixed point of a uniform morphism cannot be ultimately periodic. On the other hand, it is not difficult to show that the language of factors of a fixed point of marked morphism is prolongable. That is why we do not need any additional conditions for Formula (12).

Equality (9) was proved in detail in [9]. Equality (13) was proved in [8].

Acknowledgements

I am grateful to the anonymous referee who mentioned flaws in what is now Theorems 4 and 6, and showed the way to correct them.

References

[1] J.-P. Allouche (1994). Sur la complexité des suites infinies. *Bull. Belg. Math. Soc.* **1,** 133–143.

[2] J.-P. Allouche and M. Bousquet-Mélou (1995). On the conjectures of Rauzy and Shallit for infinite words. *Comment. Math. Univ. Carolinae* **36**, 705–711.

[3] J. Berstel (1996). Recent results on Sturmian words, *Developments in Language Theory II*, World Scientific, 13–24.

[4] J. Cassaigne (1994). An algorithm to test if a given circular HD0L-language avoids a pattern, in: *IFIP World Computer Congress'94* **1**, pp. 459–474, Elsevier (North-Holland).

[5] J. Cassaigne (1998). Sequences with grouped factors, in: *Developments in Language Theory III*, Publications of Aristotle University of Thessaloniki, pp. 211–222, Thessaloniki (Greece).

[6] J. Cassaigne (1993). Unavoidable binary patterns. *Acta Inf.* **30**, 385-395.

[7] M. Dekking (1992). On the Thue-Morse measure. *Acta Univ. Carolinae Math. Phys.* **33**, no. 2, 35–40.

[8] A. Frid (1998). On the frequency of factors in a D0L word. *Journal of Automata, Languages and Combinatorics* **3**, no. 1, 29–41.

[9] A. Frid (1998). On Uniform DOL words, in: M. Morvan, C. Meinel, D. Krob (editors). STACS'98, Lect. Notes Comp. Sci. **1373**, pp. 544–554, Springer.

[10] A. Frid (1997). The subword complexity of fixed points of binary uniform morphisms, in: B. Chlebus, L. Czaja (editors). FCT'97, Lect. Notes Comp. Sci. **1279**, pp. 179–187, Springer.

[11] F. Mignosi and P. Séébold (1993). If a D0L language is k-power-free then it is circular, in: A. Lingas, R. Karlsson, S. Carlsson (editors). ICALP'93, Lect. Notes Comp. Sci. **700**, pp. 507–518, Springer.

[12] M. Morse and G. A. Hedlund (1940). Symbolic dynamics II: Sturmian trajectories. *Amer. J. Math.* **61**, 1–42.

[13] M. Queffélec (1987). Substitution dynamical systems — spectral analysis. Lect. Notes Math. **1294**, Springer.