Jarkko Peltomäki

Introducing Privileged Words: Privileged Complexity of Sturmian Words

Turku Centre *for* Computer Science

TUCS Technical Report
No 1081, May 2013

Introducing Privileged Words: Privileged Complexity of Sturmian Words

Jarkko Peltomäki
Turku Centre for Computer Science
Joukahaisenkatu 3-5A, 20520 Turku, Finland

University of Turku, Department of Mathematics and Statistics
20014 Turku, Finland
jspelt@utu.fi

TUCS Technical Report
No 1081, May 2013

Abstract

In this paper we study the class of so-called *privileged words* which have been previously considered only a little. We develop the basic properties of privileged words, which turn out to share similar properties with palindromes. Privileged words are studied in relation to previously studied classes of words, *rich words*, *Sturmian words* and *episturmian words*. A new characterization of Sturmian words is given in terms of *privileged complexity*. The privileged complexity of the Thue-Morse word is also briefly studied.

Keywords: combinatorics on words, sturmian words, palindromes, privileged words, return words, rich words

TUCS Laboratory
FUNDIM, Fundamentals of Computing and Discrete Mathematics

1 Introduction

This work concerns a new class of words named privileged words which have previously been researched only a little. The motivation for defining these words comes from the research of so-called rich words [Gle+09] which are words having maximum number of distinct palindromes (thus the name, rich words are rich in palindromes). An important property of rich words is that a word is rich if and only if every complete first return to a palindrome is a palindrome. It's equivalent to say "every palindrome is a complete first return to a shorter palindrome". By a slight alteration of this condition we define privileged words: a word is privileged if it's a complete first return to a shorter privileged word. Moreover we need to define that the empty word and the letters of the alphabet are privileged. The effect of this modification is that every word is rich in privileged words, i.e. every word w has exactly $|w| + 1$ distinct privileged factors whereas a rich word w has exactly $|w| + 1$ distinct palindromes (there exist words which have strictly less palindromic factors). It turns out that privileged words and palindromes have some similar properties. This paper introduces the basic properties of privileged words, and questions regarding so-called privileged complexity of Sturmian words, episturmian words and the Thue-Morse word are studied.

After introducing the notations and definitions, in Section 3 privileged words and their basic properties are presented. These basic results emphasize the analogue between palindromes and privileged words. Moreover privileged words are studied in relation to rich words.

Section 4 studies the number of distinct privileged factors in finite words. There's also discussion how privileged words fit into a recent work of G. Fici and Z. Lipták [FL12].

Various complexity functions of infinite words have been previously considered. In Section 5 the notion of privileged complexity is defined. This section contains the main result of this paper: a characterization of Sturmian words using privileged complexity. As a by-product of the methods used in the proof of the main result, we obtain with little ex-

tra effort some previously known results, namely the fact that Sturmian words are rich, and partially a result of X. Droubay and G. Pirillo concerning the palindromic complexity of Sturmian words [DP99]. The section is concluded with a brief study of the privileged complexity of episturmian words.

The last section studies briefly the privileged complexity of the Thue-Morse word. It's proven that the Thue-Morse word doesn't contain a privileged factor of odd length greater than three. However the even case is left open. Some numerical data and a conjecture are provided.

2 Notation and Terminology

In this text, we denote by A a finite *alphabet*, which is a finite non-empty set of symbols. The elements of A are called *letters*. A (finite) *word* over A is a sequence of letters. To the empty sequence corresponds the *empty word*, denoted by ε. The set of all finite words over A is denoted by A^*. The set of non-empty words over A is the set $A^+ := A^* \setminus \{\varepsilon\}$. A natural operation of words is concatenation. Under this operation A^* is a free monoid over A. The letters occurring in the word w form the *alphabet of* w denoted by $\mathcal{A}lph(w)$. From now on we assume that binary words are over the alphabet $\{0, 1\}$. For binary words we define the *exchange operation*: $\hat{0} = 1$ and $\hat{1} = 0$. Given a finite word $w = a_1 a_2 \cdots a_n$ of n letters, we say that the *length* of w, denoted by $|w|$, is equal to n. By convention the length of the empty word is 0. We also denote by $|w|_a$ the number of occurrences of the letter a in w. The set of all words of length n over the alphabet A is denoted A^n.

An *infinite word* w over A is a function from the natural numbers to A. We consider such a function as a sequence indexed by the natural numbers with values in A. We write consicely $w = a_1 a_2 a_3 \cdots$ with $a_i \in A$. The set of infinite words is denoted by A^ω. The infinite word w is said to be *ultimately periodic* if it can be written in the form $w = uv^\omega = uvvv \cdots$ for some words $u, v \in A^*$, $v \neq \varepsilon$. If $u = \varepsilon$, then w is said to be *periodic*. An infinite word which is not ultimately periodic is said to be *aperiodic*.

A finite word u is a *factor* of the finite or infinite word w if it can be

written that $w = zuv$ for some $z \in A^*$ and $v \in A^* \cup A^\omega$. If $z = \varepsilon$, the factor u is called a *prefix* of w. If $v = \varepsilon$, then we say that u is a *suffix* of w. If word u is both a prefix and a suffix of w, then u is a *border* of w. The set of factors of w is denoted by $\mathcal{F}(w)$. The set $\mathcal{F}_n(w)$ is defined to contain all factors of w of length n. A set of words X is *factorial* if every factor of w is a member of X for all $w \in X$. If $w = a_1a_2 \cdots a_n$, then we denote $w[i,j] = a_i \cdots a_j$ whenever the choices of positions i and j make sense. This notion is extended to infinite words in a natural way. An *occurrence* of u in w is such a position i, that $w[i, i + |u| - 1] = u$. If such a position exists, we say that u *occurs* in w. If w has exactly one occurrence of u, then we say that u is *unioccurrent* in w. We say that a position i *introduces* a factor u if $w[i - |u| + 1, i] = u$, and u is unioccurrent in $w[1, i]$. A *complete first return* to the word u is a word starting and ending with u, and containing exactly two occurrences of u. A word which is a complete first return to some word is called a *complete return word*. A *complete return factor* is a factor of some word which is a complete return word.

The *reversal* $\widetilde{w}$ of $w = a_1a_2 \cdots a_n$ is the word $\widetilde{w} = a_n \cdots a_2a_1$. If $\widetilde{w} = w$, then we say that w is a *palindrome*. By convention the empty word is a palindrome. The set of palindromes of w is denoted by $\mathcal{Pal}(w)$. Moreover we define $\mathcal{Pal}_n(w) = \mathcal{Pal}(w) \cap \mathcal{F}_n(w)$.

Let $w = au$ where $a \in A$ and $u \in A^*$. We define the *circular shift operation* T as follows: $T(w) = ua$. By applying this shift operation repeatedly we obtain at most $|w|$ distinct words, called the *conjugates* of w.

Let A and B be two alphabets. A *morphism* from A^* to B^* is a mapping $f : A^* \to B^*$ such that $f(uv) = f(u)f(v)$ for all words $u, v \in A^*$. Because of this morphic property, the morphism f is fully determined by its images on the letters. The morphism f is said to be *non-erasing* if for every $a \in A$, $f(a) \in A^+$. A non-erasing morphism naturally extends to infinite words: for an infinite word $w = a_1a_2a_3 \cdots$, $f(w) = f(a_1)f(a_2)f(a_3) \cdots$. The morphism f is *prolongable* if there exists a letter a such that $f(a) = aw$ for some $w \in A^+$. An infinite word w may be a *fixed point* of a morphism, i.e. $f(w) = w$. For a prolongable morphism f we have that $f^n(a)$ is a prefix of $f^{n+1}(a)$ for all $n \geq 0$. Thus we obtain a unique fixed point $f^\omega(a) :=$

$\lim_{n\to\infty} f^n(a)$.

Given an infinite word w over the alphabet A we say that a factor u of w is *right special* (resp. *left special*) if ua and ub (resp. au and bu) are both factors of w for some distinct letters a and b. A factor that is both right and left special is called *bispecial*.

A set of binary words X is *balanced* if for all $n \geq 0$ and every word u and v of X of length n it holds that $||u|_1 - |v|_1| \leq 1$. A binary (finite or infinite) word is said to be balanced if its set of factors is balanced.

3 Privileged Words

Privileged words are a less known class of words which were recently introduced in [KLS11]. We define the set $\mathcal{P}ri_A$, the set of *privileged words over* A, recursively as follows:

- $\varepsilon \in \mathcal{P}ri_A$,
- $a \in \mathcal{P}ri_A$ for every letter a in the alphabet,
- if $|w| \geq 2$, then $w \in \mathcal{P}ri_A$ if w is a complete first return to a shorter privileged word.

When the alphabet is known from context, we omit the subscript A. Given a word w, we denote

$$\mathcal{P}ri(w) = \{u \in \mathcal{F}(w) : u \text{ is privileged}\}.$$

The set $\mathcal{P}ri_n(w)$ is defined to contain all privileged factors of w of length n.

The first few binary privileged words are

$$\varepsilon, 0, 1, 00, 11, 000, 111, 010, 101.$$

Not every privileged word needs to be a palindrome, for example the words 00101100 and 0120 are privileged, but not palindromic. However privileged words and palindromes have some analogous properties, as we shall soon see.

Lemma 3.1. *Let w be a privileged word, and u its any privileged prefix (respectively suffix). Then u is a suffix (respectively prefix) of w.*

Proof. If $|w| \leq 1$ or $u = w$, then the claim is clear. Suppose that $|w| \geq 2$ and $|u| < |w|$. By definition w is a complete first return to a shorter privileged word v. If $|v| < |u|$, then by induction v is a suffix of u, and thus v would have at least three occurrences in w which is impossible. If $u = v$, then the claim is clear. Finally assume that $|v| > |u|$, then by induction u is a suffix of v, and thus a suffix of w. The proof in the case that the roles of prefix and suffix are reversed is symmetric. □

The above Lemma is the first analogue to palindromes: a palindromic prefix of a palindrome occurs also as a suffix.

Lemma 3.2. *Let w be a privileged word, and u its longest proper privileged prefix (suffix). Then w is a complete first return to u. In other words the longest proper privileged prefix (suffix) of w is its longest proper privileged border.*

Proof. If $|w| \leq 1$, then the claim is clear. Suppose that $|w| \geq 2$, and that w is a complete first return to privileged word v. Now if $|u| > |v|$, then v is a prefix of u, and thus by Lemma 3.1 also a suffix of u. Hence w has at least three occurrences of v, a contradiction. Therefore $|u| \leq |v|$, and by the maximality of u, $u = v$, which proves the claim. The proof in the case that the roles of prefix and suffix are reversed is symmetric. □

Lemma 3.3. *Let w be a privileged word, and suppose that it has border u. Then u is privileged.*

Proof. If $|w| \leq 1$, the claim is clear. Suppose that $|w| \geq 2$, and that w is a complete first return to privileged word v. Since v is the longest proper border of w, we may assume that $|u| < |v|$. Now u is a prefix of v, and since u is a suffix of w and v is a suffix of w, also u is a suffix of v. Thus u is a border of v, and by induction, a privileged word. □

Palindromes share this property too: every border of a palindrome is a palindrome.

The study of so-called rich words was initiated in [Gle+09]. Rich words are words having maximum number of distinct palindromic factors. In the following definition we count ε as a palindromic factor.

Definition 3.4. A word w is *rich* if it has exactly $|w| + 1$ distinct palindromic factors. An infinite word is rich if its every factor is rich.

Next we state a useful characterization of rich words proven in [Gle+09].

Theorem 3.5. *For any finite or infinite word w, the following properties are equivalent:*

(i) w is rich,

(ii) every factor of w which is a complete first return to a palindrome is itself a palindrome. □

The fact that the condition in the next Proposition is necessary was proved in [KLS11].

Proposition 3.6. *Let w be a word. Then w is rich if and only if $\mathcal{P}ri(w) = \mathcal{P}al(w)$.*

Proof. ($\Longrightarrow$) Suppose that the word w is rich. The claim is clear for factors u of length $|u| \leq 1$. Assume first that u, $|u| > 1$, is privileged. By definition u is a complete first return to a shorter privileged word v. By induction v is a palindrome, and hence u is a complete first return to a palindrome, and is by Theorem 3.5 itself a palindrome.

Suppose then that u is a palindrome. Let v be the longest proper palindromic prefix of u. Now u is a complete first return to v. Otherwise u would have a proper prefix which is a complete first return to u, and by Theorem 3.5 this prefix would be a longer proper palindromic prefix of u than v. By induction it follows that v is privileged, and thus u too is a privileged word.

($\Longleftarrow$) Suppose now that $\mathcal{P}ri(w) = \mathcal{P}al(w)$. Now let q be a complete first return to a palindrome p in w. By assumption p is privileged, and thus q

too is privileged. Again by assumption q is a palindrome, and the claim follows from Theorem 3.5. □

As noted in [KLS11], privileged words are a "maximal generalization" of palindromes in the sense that every word is rich in privileged words, as is seen in Corollary 4.3.

4 Privileged Words and Complete Return Factors

Privileged words are special kind of complete return words. In this section we will prove that every word w has $|w| + 1$ distinct privileged factors. We will also state a characterization of those words whose all complete return factors are privileged. This characterization has already been done in [FL12], but it seems that the authors missed the concept of privileged words, so we will briefly show the connection between privileged words and their work.

The authors of [FL12] called complete return factors *closed factors*, but here we stick with more conventional vocabulary. In this section we count the empty word and the letters of the alphabet (as complete returns to the empty word) as complete return factors.

Lemma 4.1. *Let $w \in A^+$. Then every position of w introduces at least one complete return factor of w.*

Proof. Consider the position i of the word w, and the longest complete return factor v ending in i. Factor v exists since letters are complete return factors. We prove that v is unioccurrent in $w[1, i]$, which proves the claim. Now if v had been introduced earlier, say at position j, then the factor $w[j - |v| + 1, i]$ would be a complete first return to v contradicting the maximality of v. □

Corollary 4.2. [FL12] *Every word w has at least $|w| + 1$ complete return factors.* □

A word w might have more than $|w| + 1$ complete return factors (and most words do). Consider for instance the word $w = 1^k 0 1^k 0$. Every position in w except the last introduces exactly one new complete return factor, but the last position introduces $k + 1$ new complete return factors, yielding a total of $3k + 2 = |w| + \frac{1}{2}|w| - 1$ complete return factors in w.

Corollary 4.3. *Every word $w \in A^*$ has exactly $|w| + 1$ distinct privileged factors, i.e. every word is rich in privileged words.*

Proof. If we replace "longest complete return factor" with "longest privileged factor" in the proof of Lemma 4.1, we obtain that every position of w introduces at least one new privileged factor. Now if some position i would introduce two privileged factors, say u and v, with $|u| < |v|$, then by Lemma 3.1 u would also be a prefix of v, i.e. it wouldn't be unioccurrent in $w[1, i]$. This is a contradiction, and thus every position of w introduces exactly one new privileged factor. □

From the proof we obtain the following facts:

Corollary 4.4. *Let w be a word. If some position in w introduces exactly one complete return factor, then this factor is privileged. The word w has exactly $|w| + 1$ complete return factors if and only if its every complete return factor is privileged.* □

In the article [FL12] words having the minimum number of complete return factors were considered. The authors called such words *C-poor words*. Since the minimum number of complete return factors in a word w is $|w| + 1$, we have that a word is C-poor if and only if its every complete return factor is privileged. We are ready to state a characterization of C-poor words.

Proposition 4.5. [FL12] *Let w be a word. Then the following are equivalent:*

(i) w is C-poor,

(ii) every complete return factor of w is privileged,

(iii) *w doesn't contain as a factor a complete first return to xy for distinct letters x and y.* □

It's worth noting that by part *(iii)* C-poor words avoid all squares except squares of letters. If the word w is binary, then it can be said more:

Proposition 4.6. [FL12] *Let w be a binary word. Then the following are equivalent:*

(i) *w is C-poor,*

(ii) *every complete return factor of w is a palindrome,*

(iii) *w is a conjugate of a word in 0^*1^*.* □

Now part *(ii)* of the above Proposition actually says that for a binary C-poor word w, $\mathcal{Pri}(w) \subseteq \mathcal{Pal}(w)$. We obtain that $\mathcal{Pri}(w) = \mathcal{Pal}(w)$, since $|w| + 1 = |\mathcal{Pri}(w)| \leq |\mathcal{Pal}(w)| \leq |w| + 1$. Thus by Proposition 3.6 the word w must be rich (as was observed in [FL12]). If the alphabet is larger than two letters, then not every C-poor word is rich: for instance the word 0120 is C-poor, but not rich.

5 Privileged Complexity, Sturmian Words and Epistur-mian Words

In the study of infinite words many different so-called complexity functions have been considered. It's clearly of interest to try to count the number of distinct privileged words of length n occurring in a finite or infinite word w, that is, to figure out the privileged complexity of words.

Definition 5.1. Let w be a finite or infinite word. The *privileged complexity function* which counts the number of distinct privileged factors of length n in w is defined as

$$\mathcal{A}_n(w) = |\mathcal{Pri}_n(w)|$$

for all $n \geq 0$.

5.1 Sturmian Words

In this section we prove some basic results about the privileged complexity function function, and a characterization of Sturmian words using privileged complexity. First we need to discuss some related complexity functions.

The factor complexity function $\mathcal{C}_n(w)$ of the word w counts the number of distinct factors of w of length n, i.e. $\mathcal{C}_n(w) = |\mathcal{F}_n(w)|$. We state the following well-known Theorem (for a proof see Theorem 1.3.13 of [Lot02]).

Theorem 5.2. *An infinite word w is aperiodic if and only if $\mathcal{C}_n(w) \geq n+1$ for all $n \geq 0$, i.e w is aperiodic if and only if it has at least one right special factor of each length.* □

Sturmian words are characterized by the fact that they are the simplest infinite aperiodic words in terms of the complexity function. They are defined as follows:

Definition 5.3. An infinite word w is *Sturmian* if $\mathcal{C}_n(w) = n+1$ for all $n \geq 0$.

For more information about Sturmian words see the Chapter 2 of [Lot02]. The next two Propositions are well-known (for proofs see Propositions 2.1.2 and 2.1.3 of [Lot02]).

Proposition 5.4. *Let X be a factorial set of words. If X is balanced, then $|X \cap \{0,1\}^n| \leq n+1$ for all $n \geq 0$.* □

Proposition 5.5. *Let X be a factorial set of words. If X is unbalanced, then there exists a unique minimal unbalanced pair of the form $(0x0, 1x1)$ in X where the word x is a palindrome.* □

Sturmian words are also characterized as follows (see Theorem 2.1.5 of [Lot02]):

Theorem 5.6. *An infinite binary word is Sturmian if and only if it's aperiodic and balanced.* □

Sturmian words have numerous other characterizations. The characterization of interest here is the characterization in terms of the palindromic complexity function, due to X. Droubay and G. Pirillo [DP99]. *The palindromic complexity function* $\mathcal{P}_n(w)$ is defined as $\mathcal{P}_n(w) = |\mathcal{P}al_n(w)|$, it counts the number of distinct palindromes of length n in w.

Theorem 5.7. [DP99] *An infinite word w is Sturmian if and only if it has palindromic complexity*

$$\mathcal{P}_n(w) = \begin{cases} 1, & \text{if } n \text{ is even} \\ 2, & \text{if } n \text{ is odd} \end{cases}$$

for all $n \geq 0$. □

We say that the word w has the property $\mathcal{P}_{Pal}(w)$ if the word w satisfies the palindromic complexity of the above Theorem. We shall prove that the condition of this Theorem is sufficient after we have established a (similar) proof of Theorem 5.8.

It's natural to ask what is the privileged complexity of Sturmian words, and if the answer to this question characterizes Sturmian words. This indeed is the case, and it's the main result of this section.

Theorem 5.8. *An infinite word w is Sturmian if and only if it has privileged complexity*

$$\mathcal{A}_n(w) = \begin{cases} 1, & \text{if } n \text{ is even} \\ 2, & \text{if } n \text{ is odd} \end{cases}$$

for all $n \geq 0$.

The proof of this theorem is based on two Lemmas 5.20 and 5.22. To simplify notations, we say that the word w has the property $\mathcal{ISP}(w)$ if the word w satisfies the privileged complexity of the above Theorem.

We will first prove that an infinite word w having the property $\mathcal{ISP}(w)$ must be Sturmian. For this purpose we introduce the concepts of Q-property and Q-factors of words.

Definition 5.9. A *Q-property* of words is defined to satisfy the following conditions:

- $Q(\varepsilon)$ and $Q(a)$ hold for all letters $a \in A$,
- for every position i in every word there exists a factor with property Q ending at position i,
- every position in every word introduces at most one factor with property Q.

Factors with property Q are called *Q-factors*.

Definition 5.10. The *Q-complexity function* of a word w is defined as

$$\mathcal{H}_n^Q(w) = |\{u \in \mathcal{F}_n(w) : Q(u)\}|.$$

For a finite word w we define $\mathcal{H}^Q(w) = |\{u \in \mathcal{F}(w) : Q(u)\}|$.

The following Lemma follows easily from the definition.

Lemma 5.11. *Every word has at most* $|w| + 1$ *distinct Q-factors, i.e.* $\mathcal{H}^Q(w) \leq |w| + 1$. □

It's well-known and easy to see that every position in every word introduces at most one new palindrome (see Proposition 2 of [DJP01]). Hence "being a palindrome" is a Q-property. From the proof of Corollary 4.3 we see that every position in every word introduces exactly one new privileged factor, and thus "being a privileged word" is a Q-property. Third possible Q-property could be "being a power of a letter".

Lemma 5.12. *Let* $w \in A^+$ *be a finite word with* $|\mathcal{A}lph(w)| \geq 2$. *Then* $\mathcal{H}_n^Q(w) = 0$ *for some* $2 \leq n \leq |w|$.

Proof. We may assume that w has exactly $|w| + 1$ distinct Q-factors, since otherwise clearly $\mathcal{H}_n^Q(w) = 0$ for some $2 \leq n \leq |w|$. Now if $\mathcal{H}_n^Q(w) > 0$ for all $1 \leq n \leq |w|$, then since we have $|w|$ positions in w, every position needs to introduce a distinct Q-factor of different length. This however is impossible since by assumption $|\mathcal{A}lph(w)| \geq 2$ positions introduce a Q-factor of length one. □

Definition 5.13. If $\mathcal{H}_n^Q(w) = 0$ for some integer n, then we say that n is a *vanishing index of* w.

When it's said that n is a vanishing index in the last y-block of y^m in the next Lemma, we mean that n is a vanishing index of y^m, and that $(m-1)|y| < n \leq m|y|$.

Lemma 5.14. *Let $w = y^\omega$ be a periodic infinite word. Then either $\mathcal{H}_n^Q(w) = 0$ for infinitely many n or there exists such k that $\mathcal{H}_n^Q(w) = 1$ for all $n \geq k$.*

Proof. Let $r = |y|$. If there are infinitely many vanishing indices, then the claim is clear. Assume that there are only finitely many vanishing indices. Let m be such an integer that the last vanishing index n is in the last y-block of y^m. If no such integer n exists, we set $m = 0$ and $n = -1$. Now concatenating y to y^m introduces at most r new Q-factors of different length since every position can introduce at most one new Q-factor. Now there might be some vanishing indices in the last y-block of y^{m+1}. However when the next y is concatenated to y^{m+1} it must be that $\mathcal{H}_i^Q(y^{m+2}) > 0$ for all $n+1 \leq i \leq (m+1)r$ since if $\mathcal{H}_j^Q(y^{m+2}) = 0$ for some $n+1 \leq j \leq (m+1)r$, then since adding more y's to the end of y^{m+2} doesn't introduce any new factors of length j, it would be that $\mathcal{H}_j^Q(w) = 0$ contradicting the maximality of n.

Now if there are s vanishing indices in the last y-block of y^{m+1}, then there are at least s vanishing indices in the last y-block of y^{m+2}. Otherwise concatenating y to y^{m+1} would have introduced more than r Q-factors of different length which is impossible. It could be that the number of vanishing indices in the last y-block of y^{m+2} increases, but such a phenomenom can occur at most r times. Hence there exists an integer m' such that the last y-block of $y^{m'+d}$ has s' vanishing indices for all $d \geq 0$. We claim that $\mathcal{H}_i^Q(w) = 1$ for all $i > m'r$, which proves the claim. Clearly by the maximality of n $\mathcal{H}_i^Q(w) \geq 1$ for all $i > m'r$. Now concatenating y to $y^{m''}$ for $m'' \geq m'$ must introduce r new Q-factors of different length, since otherwise the number of vanishing indices in the last y-block of $y^{m''+1}$ would increase. Now if $\mathcal{H}_j^Q(w) \geq 2$ for some $m'r < j$, then concatenating y to $y^{m''}$ for some $m'' \geq m'$ would introduce at least two new Q-factors of the

same length or it would introduce at least one Q-factor of some length already introduced. This is impossible, since such a concatenation would introduce less than r Q-factors of different length. □

Corollary 5.15. *Let w be an ultimately periodic infinite word. Then either $\mathcal{H}_n^Q(w) = 0$ for infinitely many n or there exists such k that $\mathcal{H}_n^Q(w) = 1$ for all $n \geq k$.*

Proof. Let $w = xy^\omega$. Adding the prefix x to y^ω introduces only finitely many new Q-factors. Hence the claim follows from Lemma 5.14. □

As a consequence of the above Corollary and the discussion after Lemma 5.11 we have the following two Corollaries.

Corollary 5.16. *Infinite word w having the property $\mathcal{ISP}(w)$ is aperiodic.* □

Corollary 5.17. *Infinite word w having the property $\mathcal{P}_{Pal}(w)$ is aperiodic.* □

However there exist aperiodic infinite words w having $\mathcal{Pri}_n(w) = 0$ for infinitely many n. One example is the Thue-Morse word. See Proposition 6.3.

To simplify notations, we define for a word w the following properties:

- $\mathcal{Rch}_n(w) \iff \mathcal{Pri}_n(w) = \mathcal{Pal}_n(w)$,
- $\mathcal{Spe}_n(w) \iff$ there exists a unique right special factor of length $n-1$,
- $\mathcal{Bal}_n(w) \iff$ all factors of w of length n are balanced,
- $\mathcal{Rev}_n(w) \iff$ for each factor u of length n, also $\widetilde{u}$ is a factor of w.

Lemma 5.18. *Let w be an infinite binary word for which $\mathcal{Bal}_m(w)$ holds for all $0 \leq m \leq n$. Then any right special factor of w with length strictly less than n has at most two complete return factors in w.*

Proof. First we reason that there exists at most one right special factor of length $m < n$. Suppose on the contrary that there are two right special factors of length m, say u and v. Let z be the longest common suffix of u

and v, i.e. $u = u'az$ and $v = v'\hat{a}z$ for some letter a. Since u and v are right special, $aza, \hat{a}z\hat{a} \in \mathcal{F}(w)$, contradicting $\mathcal{B}al_m(w)$ for some $m \leq n$.

Let then u be a right special factor of w with length $|u| < n$. Suppose that v_1 and v_2 are two distinct complete first returns to u in w. Let x be the longest common prefix of v_1 and v_2. We claim that $x = u$. Assume on the contrary that $|x| > |u|$. Then $v_1 = xav_1'$ and $v_2 = x\hat{a}v_2'$ for some letter a, and hence x is a right special factor. Now the suffix of x of length $|u|$ is right special, so by the reasoning in the beginning of the proof, x has u as a suffix. This however contradicts the fact that v_1 and v_2 are complete returns to x. Now if there was a third complete return to u in w, then it would have a common prefix of length $|u| + 1$ with either v_1 or v_2, which is not possible by the above. □

Proposition 5.19. *Let w be an infinite binary word. If $\mathcal{B}al_m(w)$ holds for all $0 \leq m \leq n$, then $\mathcal{R}ch_m(w)$ holds for all $0 \leq m \leq n+1$. Specifically an infinite balanced binary word is rich.*

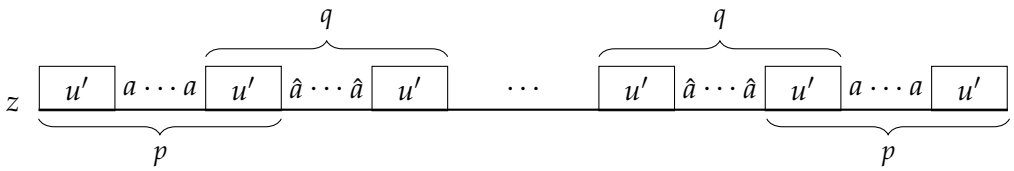
Proof. The claim clearly holds for $m \leq 1$. Assume then that $\mathcal{B}al_m(w)$ holds for all $0 \leq m \leq k \leq n$. We will show that then $\mathcal{R}ch_{k+1}(w)$ holds.

Case 1. $\mathcal{P}ri_{k+1}(w) \subseteq \mathcal{P}al_{k+1}(w)$
Let $x \in \mathcal{P}ri_{k+1}(w)$. Then x is a complete first return to a shorter privileged word u. By the induction hypothesis u is a palindrome. If u overlaps with itself in x or $x = u^2$, then x must be a palindrome. Assume that this is not the case.

Now if $|u| = 1$, then $x = u\hat{u}^l u$ is a palindrome. Suppose that $|u| = 2$, so $u = aa$ for some letter a. Then $x = aa\lambda aa$ for some $\lambda \neq \varepsilon$. Hence there exists in x a complete first return to a of the form $a\hat{a}^l a$ for some $l \geq 1$. By Lemma 5.18 the words $a\hat{a}^l a$ and aa are the only complete first returns to a. Then as aa is not a factor of $a\lambda a$, it must be that $a\lambda a = (a\hat{a}^l)^t a$ for some $t \geq 1$. Thus x is a palindrome.

We may now assume that $|u| \geq 3$, Write $u = au'a$ for some letter a and $u' \neq \varepsilon$. Note that u' is a palindrome, and hence by the hypothesis privileged. Now $x = au'a\lambda au'a$ with $\lambda \neq \varepsilon$. Consider $z = u'a\lambda au'$, the center of

Figure 1: A picture clarifying the proof of Proposition 5.19. Note that not all occurrences of u' need to be non-overlapping.



x. We will prove that z is a palindrome. From this it follows that x too is palindromic. If z is a complete first return to u', then z is privileged, and by hypothesis a palindrome. Assume then that z contains at least three occurrences of u'. Now word z has as a proper prefix a complete return to u' which begins with $u'a$. Denote this prefix as p. As x is a complete first return to $u = au'a$, the word z doesn't have $au'a$ as a factor. Hence it now must have $au'\hat{a}$ as a factor. Therefore z contains as a factor a complete first return to u' beginning with $u'\hat{a}$. Denote this factor by q. For a better grasp of the situation see Figure 1. As u' is right special, by Lemma 5.18 words p and q are the only complete return factors of u' in w. Since $au'a$ and $\hat{a}u'\hat{a}$ are not factors of z (z is balanced), the occurrences of p and q must alternate in z. Since both p and q are palindromes as complete first returns to u' and z begins and ends with p, it follows that z is a palindrome.

Case 2. $\mathcal{P}al_{k+1}(w) \subseteq \mathcal{P}ri_{k+1}(w)$
Let then $x \in \mathcal{P}al_{k+1}(w)$ and u its longest proper border. Now u must be a palindrome, and hence by the induction hypothesis, a privileged word. The word x must be a complete first return to u, since otherwise there would be a privileged proper prefix v longer than u. That would be a contradiction with the maximality of u, since v would also be a palindrome by the induction hypothesis. Therefore x is privileged.

The last claim follows now from Proposition 3.6. □

We are now ready to prove the other direction of Theorem 5.8. The

proof is similar to the proof of Theorem 5.7 in [DP99].

Lemma 5.20. *An infinite word w having the property $\mathcal{JSP}(w)$ is Sturmian.*

Proof. First of all since $\mathcal{A}_1(w) = 2$, $\mathcal{P}ri_1(w) = \{0, 1\}$ and the word w is thus binary. Hence ε is the unique right special factor of length 0. Clearly $\mathcal{R}ch_1(w), \mathcal{S}pe_1(w), \mathcal{B}al_1(w)$ and $\mathcal{R}ev_1(w)$ hold. We will next assume that $\mathcal{R}ch_n(w), \mathcal{S}pe_n(w), \mathcal{B}al_n(w)$ and $\mathcal{R}ev_n(w)$ hold for $n \geq 1$, and prove that w satisfies all these properties for $n + 1$. This proves the claim.

Case 1. $\mathcal{R}ch_{n+1}(w)$
By the induction hypothesis $\mathcal{B}al_m(w)$ holds for all $m \leq n$. From Proposition 5.19 it follows that also $\mathcal{R}ch_{n+1}(w)$ holds.

Case 2. $\mathcal{B}al_{n+1}(w)$
Assume on the contrary that $\mathcal{B}al_{n+1}(w)$ doesn't hold. Then by Proposition 5.5 there exists a palindromic factor x such that $0x0, 1x1 \in \mathcal{F}_{n+1}(w)$. By Case 1 we know that $\mathcal{P}ri_{n+1}(w) = \mathcal{P}al_{n+1}(w)$, and since palindromes $0x0, 1x1 \in \mathcal{P}al_{n+1}(w)$, by assumption, $n + 1$ is odd, and hence also $n - 1$ is odd. Again by assumption $\mathcal{P}al_{n-1}(w) = \{x, t\}$ where $x \neq t$. Moreover x is a bispecial factor of length $n - 1$. Since $\mathcal{S}pe_n(w)$ holds, we conclude that t isn't right special. Now $ta \in \mathcal{F}_n(w)$ for some letter a. By the property $\mathcal{R}ev_n(w)$, also $at \in \mathcal{F}_n(w)$. Since t isn't right special, $ata \in \mathcal{F}_{n+1}(w)$. Thus $\{0x0, 1x1, ata\} \subseteq \mathcal{P}al_{n+1}(w) = \mathcal{P}ri_{n+1}(w)$, contradicting the fact that $\mathcal{A}_{n+1}(w) = 2$.

Case 3. $\mathcal{S}pe_{n+1}(w)$
Since by Corollary 5.16 the word w is aperiodic, we have that it has at least one right special factor of length n (Theorem 5.2). Arguing as in the first paragraph of the proof of Lemma 5.18 we see that it has at most one right special factor of length n.

Case 4. $\mathcal{R}ev_{n+1}(w)$
Denote $\widetilde{\mathcal{F}}_n(w) = \{\widetilde{w} : w \in \mathcal{F}_n(w)\}$. Let us consider the set

$$X = \bigcup_{i=0}^{n+1} \mathcal{F}_i(w) \cup \widetilde{\mathcal{F}}_i(w).$$

This set is balanced since otherwise by Proposition 5.5 there would exists palindromes $0x0, 1x1 \in X$, and hence (since these words are palindromes) $0x0, 1x1 \in \mathcal{F}_m(w)$ for some $m \leq n+1$. This is a contradiction with the induction hypothesis or the Case 2. Thus by Proposition 5.4 $|X \cap \{0,1\}^{n+1}| \leq n+2$. On the other hand by Theorem 5.2 $|X \cap \{0,1\}^{n+1}| \geq C_{n+1}(w) = n+2$. Thus $|X \cap \{0,1\}^{n+1}| = n+2$, and it must therefore be that $\mathcal{F}_{n+1}(w) = \widetilde{\mathcal{F}}_{n+1}(w)$ which means that $\mathcal{R}ev_{n+1}(w)$ is satisfied. □

For the converse of Lemma 5.20 we state the immediate Corollary of Proposition 5.19. For another proof see Corollary 4 of [DJP01].

Corollary 5.21. *Sturmian words are rich.* □

From this we easily deduce the converse result:

Lemma 5.22. *Sturmian word w has the property $\mathcal{JSP}(w)$.*

Proof. By Corollary 5.21 the Sturmian word w is rich. Next, Proposition 3.6 says that $\mathcal{P}ri(w) = \mathcal{P}al(w)$, and hence by Theorem 5.7 the word w has the property $\mathcal{JSP}(w)$. □

Lemmas 5.20 and 5.22 establish Theorem 5.8. The proof of Lemma 5.20 with minor modifications proves too that an infinite word w having the property $\mathcal{P}_{Pal}(w)$ is Sturmian: the Case 1 is omitted, and the Case 2 needs to be slightly adjusted. Otherwise the proof can be kept intact, Corollary 5.17 ensures that an infinite word having property $\mathcal{P}_{Pal}(w)$ must be aperiodic.

Note that not every Q-complexity function characterizes Sturmian words. Take Q to be the property "being a power of a letter". Now for

a Sturmian word w either $00 \notin \mathcal{F}(w)$ or $11 \notin \mathcal{F}(w)$. By symmetry assume that $11 \notin \mathcal{F}(w)$. It's also well-known that there exists a k such that $0^k \in \mathcal{F}(w)$, but $0^{k+1} \notin \mathcal{F}(w)$. Hence the Q-complexity of w would be

$$\mathcal{H}_n^Q(w) = \begin{cases} 1, & \text{if } n = 0 \text{ or } 2 \leq n \leq k \\ 2, & \text{if } n = 1 \\ 0, & \text{otherwise.} \end{cases}$$

However for instance the non-Sturmian word $(0^k1)^\omega$ has the same Q-complexity function.

5.2 Episturmian Words

We conclude Section 5 by considering briefly the privileged complexity of so-called episturmian words, which are a generalization of Sturmian words to arbitrary alphabet. Episturmian words have a rich theory, for more about these intriguing words, see the foundational paper [DJP01] and the survey [GJ09].

Definition 5.23. An infinite word w over alphabet A ($|A| \geq 2$) is *episturmian* if w has the property $\mathcal{R}ev_n(w)$ and has at most one right special factor of length n for all $n \geq 0$.

Definition 5.24. Episturmian word w is *A-strict* if for each $n \geq 0$ there exists a unique right special factor u of length n and $ua \in \mathcal{F}(w)$ for all $a \in A$.

Actually A-strict episturmian words over the alphabet A are exactly the so-called *Arnoux-Rauzy* words over A. Note that A-strict episturmian words are aperiodic, and if $A = \{0, 1\}$, then these words are exactly the Sturmian words.

In the Corollary 2 of [DJP01] it was proved that episturmian words are rich. Therefore we may proceed as we did with Sturmian words: the palindromic complexity of episturmian words gives us their privileged complexity by Proposition 3.6.

The following is Theorem 4.4 in [JP02].

Theorem 5.25. [JP02] *Let w be a A-strict episturmian word over the alphabet A. Then w has palindromic complexity*

$$\mathcal{P}_n(w) = \begin{cases} 1, & \text{if } n \text{ is even} \\ |A|, & \text{if } n \text{ is odd} \end{cases}$$

for all $n \geq 0$. □

Thus we have the following result:

Theorem 5.26. *Let w be a A-strict episturmian word over the alphabet A. Then w has privileged complexity*

$$\mathcal{A}_n(w) = \begin{cases} 1, & \text{if } n \text{ is even} \\ |A|, & \text{if } n \text{ is odd} \end{cases}$$

for all $n \geq 0$. □

However the privileged complexity of A-strict episturmian words doesn't characterize them when $|A| > 2$. Words coding r-interval exchange transformations are a class of rich words which satisfy the palindromic complexity of Theorem 5.25 [BMP07]. Being rich they also have the same privileged complexity as episturmian words. However words coding r-interval exchange transformations are not episturmian when $r > 2$ (here r is the number of letters). One example of such a word is the fixed point of the following morphism

$$\alpha : \begin{array}{l} a \mapsto c, \\ b \mapsto ca, \\ c \mapsto caba. \end{array}$$

The fixed point isn't episturmian since both letters a and c are right special, but it satisfies the complexities of Theorems 5.25 and 5.26.

Actually not even both factor complexity and privileged complexity of A-strict episturmian words characterizes them since words coding r-interval exchange transformations have the same factor complexity as episturmian words [BMP07].

6 Privileged Complexity and the Thue-Morse Word

In this section we investigate briefly the privileged complexity of the Thue-Morse word. The infinite Thue-Morse word t is defined as the fixed point of the morphism φ:

$$\varphi : \begin{array}{l} 0 \mapsto 01, \\ 1 \mapsto 10. \end{array}$$

For more information about the Thue-Morse word, see Chapter 2 of [Lot83]. The word t has the following well-known property:

Theorem 6.1. *The Thue-Morse word t is overlap free.* □

Lemma 6.2. *Let w be a non-empty even length privileged factor of t. Then* 00 *or* 11 *is a factor of w.*

Proof. Using the fact that t is overlap free, it can be easily shown that no factor of t of length greater than four avoids factors 00 and 11. The only possible privileged factors of length two are 00 and 11. For privileged factors of length four, the possibilities are 0000, 1111, 0110 and 1001. Words 0000 and 1111 are not factors of t, but anyway the claim is proved. □

Proposition 6.3. *The infinite Thue-Morse word t doesn't have any privileged factors of length n, when n is odd and $n \geq 5$.*

Proof. Let w be an privileged factor of t of odd length, which is a complete first return to a privileged word u. Denote $x = 01$ and $y = 10$.

Assume first that $|u|$ is odd. Since 000 and 111 are not factors of t, it must be that $|u| > 1$. Moreover $|w| > 5$, since if $|w| = 5$, then the occurrences of u in w would need to overlap, and t is overlap free. We need to only prove that u can't be 010 or 101 (the privileged factors of t of length three). We prove that u can't be 010, the other case is symmetric. Assume first that the factorization of w over $\{x, y\}$ matches from the beginning of w. So we are looking for a factor of t starting and ending with xx, and containing no internal occurrences of $xx = 0101$ or $yy = 1010$. Using the fact that t is overlap free, one can by inspection deduce that the only

possibility is $xxyxx$. However $xxyxx$ is not a factor of t. Assume then that the factorization over $\{x, y\}$ doesn't match from the beginning. Now if $u = 010$, then u must be preceded by 1 in t. Thus we would have found a complete first return to 101, say v, of length $|w|$, and the factorization of v over $\{x, y\}$ would match from the beginning. Earlier it was proved that such a factor v can't exist.

Assume then that $|u|$ is even. By Lemma 6.2 u contains 00 or 11 as a factor, say it contains 00. Suppose that 00 occurs at an even position in the prefix v of w. Then since $|w|$ is odd, it must be that 00 occurs at an odd position in the suffix v. Therefore w doesn't match any factorization over $\{x, y\}$. If 00 occurs at an odd position in the prefix v, one arrives at a contradiction using a symmetric argument. □

Now it's also true that the Thue-Morse word doesn't contain any odd palindrome of length greater than three (for a proof see [BBL08]). However the number of even length privileged factors can't be calculated in the same way as we did earlier with Sturmian words, since the Thue-Morse word isn't rich. For instance the following factor of t is not rich: 11010011.

The case of even length privileged factors is more complicated. So far it's not known to the author how to evaluate the number of even length privileged factors in t. In the next table there are some values for $\mathcal{A}_n(t)$ for even n. The results are based on a computer search.

2-10	12-20	22-30	32-40	42-50	52-60	62-70
2	4	4	14	8	0	0
2	0	8	14	4	0	0
4	0	8	6	2	0	2
8	2	4	4	2	0	2
8	2	6	8	0	0	2

There are interesting gaps of zeros in $\mathcal{A}_n(t)$. For instance $\mathcal{A}_n(t) = 0$ for $81 \leq n \leq 85$, $113 \leq n \leq 117$, $145 \leq n \leq 149$, $177 \leq n \leq 181$ and $189 \leq n \leq 257$. Based on the computer searches and an educated guess, we state the following conjecture:

Conjecture. *There exist arbitrarily long (but not infinite) gaps of zeroes in the values of* $\mathcal{A}_n(t)$.

7 Acknowledgements

I thank my advisor Tero Harju for discussions on the practical matters of mathematics and for introducing me to combinatorics on words. I also thank my other advisor Luca Zamboni for suggesting the study of privileged words and for helping me during my research. Finally I thank Markus Whiteland for useful discussion sessions.

References

[BMP07] P. Baláži, Z. Masáková, and E. Pelantová. "Factor Versus Palindromic Complexity of Uniformly Recurrent Infinite Words". In: *Theoretical Computer Science* 380 (2007), pp. 266–275.

[BBL08] A. Blondin-Massé, S. Brlek, and S. Labbé. "Palindromic Lacunas of the Thue-Morse Word". In: *Proc. GASCom 2008, 6th International Conference on Random Generation of Combinatorial Structures (June 16th-20th, 2008, Arezzo, Italia)*. 2008, pp. 53–67.

[DJP01] X. Droubay, J. Justin, and G. Pirillo. "Episturmian Words and Some Constructions of de Luca and Rauzy". In: *Theoretical Computer Science* 225 (2001), pp. 539–553.

[DP99] X. Droubay and G. Pirillo. "Palindromes and Sturmian Words". In: *Theoretical Computer Science* 223 (1999), pp. 73–85.

[FL12] G. Fici and Z. Lipták. "Words with the Smallest Number of Closed Factors". In: *ICTCS '12, Proc. of the 13th Italian Conference on Theoretical Computer Science*. 2012.

[GJ09] A. Glen and J. Justin. "Episturmian Words: A Survey". In: *RAIRO - Theoretical Informatics and Applications* 43 (2009), pp. 403–442.

[Gle+09] A. Glen et al. “Palindromic Richness”. In: *European Journal of Combinatorics* 30 (2009), pp. 510–531.

[JP02] J. Justin and G. Pirillo. “Episturmian Words and Episturmian Morphisms”. In: *Theoretical Computer Science* 276 (2002), pp. 281–313.

[KLS11] J. Kellendonk, D. Lenz, and J. Savinien. “A Characterization of Subshifts with Bounded Powers”. Preprint arXiv:1111:1609. 2011.

[Lot02] M. Lothaire. *Algebraic Combinatorics on Words*. Vol. 90. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 2002.

[Lot83] M. Lothaire. *Combinatorics on Words*. Vol. 17. Encyclopedia of Mathematics and Its Applications. Addison-Wesley, 1983.

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 A, 20520 TURKU, Finland | www.tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics

Turku School of Economics
- Institute of Information Systems Sciences

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

ISBN 978-952-12-2904-6
ISSN 1239-1891