Toeplitz Words, Generalized Periodicity and Periodically Iterated Morphisms

JULIEN CASSAIGNE AND JUHANI KARHUUMÄKI

We consider so-called Toeplitz words which can be viewed as generalizations of one-way infinite periodic words. We compute their subword complexity, and show that they can always be generated by iterating periodically a finite number of morphisms. Moreover, we define a structural classification of Toeplitz words which is reflected in the way in which they can be generated by iterated morphisms.

© 1996 Academic Press Limited

1. INTRODUCTION

Toeplitz introduced, in [20], an iterative construction to define almost periodic functions on the real line. In [12], Jacobs and Keane modified this construction to define infinite words. Their motivation, however, was on topological aspects of words; in particular, on ergodic theory.

Starting from [19], these words, now referred to as Toeplitz words, have been considered from the combinatorial point of view. Surprising connections of these words to different combinatorial problems are discussed in [2]. The combinatorial research of Toeplitz words seems to have concentrated on a special case of our later classification, corresponding to so-called paperfolding words (cf. [2, 6, 18]).

The Toeplitz words that we are considering here are defined as follows. Take an infinite periodic word \( w \) on the alphabet \( \Sigma \cup \{?\} \), where ? corresponds to a hole. Fill the holes iteratively by substituting the word itself into the remaining holes. At the limit, all holes are filled and an infinite word is defined.

The goals of this paper are as follows. We intend to demonstrate that the above Toeplitz words are natural, and not too large, generalizations of periodic words. As a part of that, we classify different types of Toeplitz words in terms of their subword complexity. We also point out connections between Toeplitz words and different types of iteration of morphisms (cf. [5, 16]).

More precisely, let us call the above Toeplitz word a \((p, q)\)-Toeplitz word if the length of the pattern, that is \( |w| \), is \( p \), and if it contains \( q \) holes. We show that:

(i) if \( q = 1 \), then the word can be generated by iterating a morphism;
(ii) if \( q \) divides \( p \), then it can be generated by iterating a morphism, and then mapping the result by another morphism (or, in fact, by a coding); and
(iii) otherwise, i.e. if \( q \) does not divide \( p \), it can be generated by iterating periodically \( q \) morphisms.

In each of the above cases the word can be periodic, as we shall see. However, we are able to characterize when this takes place and, moreover, we can show that if this is not the case, then the subword complexity of a word in case (i) or case (ii) is \( \Theta(n) \), and in case (iii) it is \( \Theta(n') \), where \( r = \log(p/d) / \log(p/q) \) with \( d \) equal to the greatest common divisor of \( p \) and \( q \). Consequently, the growth of the subword complexity does not depend on the pattern itself, that is on \( w \), but on \( p \) and \( q \) only, assuming that the infinite word is not periodic. This is evidence that Toeplitz words reflect the periodicity properties of words.

We want to finish this introduction with the following comment. As we said, Toeplitz words were introduced when studying properties of real functions. Their natural
environment, however, is that of words, and we believe that they provide most elegant results in this area. This is exactly the same as what happened to the well known periodicity lemma of Fine and Wilf (cf. [8]): it was introduced for real functions, but became really fundamental in connection with words (cf. [17])!

2. Basic Definitions

For a finite alphabet \( \Sigma \), let \( \Sigma^* \), \( \Sigma^+ \) and \( \Sigma^w \) denote the sets of all finite, finite non-empty and one-way infinite words over \( \Sigma \), respectively. We call an infinite word \( w \in \Sigma^w \) periodic or ultimately periodic, if it can be written in the form \( w = uv^w \) or \( w = uv^w \) for some finite words \( u \) and \( v \).

Let \( ? \) be a letter not in \( \Sigma \). For a word \( w \in \Sigma(\Sigma \cup \{?\})^* \), let

\[
T_0(w) = ?, \quad T_{i+1}(w) = F_i(T_i(w)),
\]

(1)

where \( F_i(u) \), defined for any \( u \in (\Sigma \cup \{?\})^w \), is the word obtained from \( w^w \) by replacing the sequence of all occurrences of \( ? \) by \( u \); in particular, \( F_i(u) = w^w \) if \( w \) contains no \( ? \).

Clearly,

\[
T(w) = \lim_{i \to \infty} T_i(w) \in \Sigma^w
\]

(2)
is well-defined, and it is referred to as the Toeplitz word determined by the pattern \( w \).

Let \( p = |w| \) and \( q = |w| \), be the length of \( w \) and the number of \( ?'s \) in \( w \), respectively. Then we call \( T(w) \) a \((p, q)\)-Toeplitz word.

The above definition emphasizes the iterative nature of Toeplitz words. There exists another equivalent definition, which points out the self-reading nature of these words. Let \( w \in \Sigma(\Sigma \cup \{?\})^* \) as above, and define a sequence of infinite words \( I_i(w) \) as follows: \( I_0(w) = w^w \) and \( I_i(w) \) is the word obtained from \( I_{i-1}(w) \) by replacing the first occurrence of \( ? \) in \( I_{i-1}(w) \) by the \( i \)th letter of \( I_{i-1}(w) \) (which is always different from \( ? \)).

Then, clearly,

\[
I(w) = \lim_{i \to \infty} I_i(w) \in \Sigma^w
\]

(3)
is well-defined, and it is a straightforward consequence of the definitions that

\[
I(w) = T(w).
\]

Note that in the process of defining \( I(w) \), holes are filled one by one by reading the prefix of \( I(w) \). Consequently, \( I(w) \) is self-reading in the same sense as the Kolakoski word (cf. [13, 5]).

We proceed with a few simple examples.

Example 1. The periodic word \((112)^w\) is obtained as a Toeplitz word in a number of different ways:

\[
(112)^w = T(112???????) = T(112?12?1?) = T(1?21?211?????).
\]

Example 2. As is easy to see, the only ultimately periodic Toeplitz words are the periodic ones. However, this is true only since we required that \( w \) start with a symbol in \( \Sigma \). Of course, if this is not required, and the definition (2) is used, then patterns of the form

\[
w = ?^p u ?^w
\]
determine all ultimately periodic words, where the ‘initial mess’ is arbitrary.
Example 3. The famous paperfolding sequence is the Toeplitz word determined by the pattern \(1?0\) (cf. [2]).

Example 4. The pattern \(12?\) determines a Toeplitz word which is the fixed point of the morphism \(h: 1 \mapsto 121, \ 2 \mapsto 122\), i.e. \(T(12?) = \lim_{i \to \infty} h^i(1)\). From this, it follows easily that this Toeplitz word is cube-free, although it contains arbitrarily long repetitions of the form \(uauau\) with \(a \in \{1, 2\}\).

We conclude this section by pointing out that Toeplitz words indeed possess certain kinds of periodicity properties. Consider a \((p, q)\)-Toeplitz word \(T(w)\). Hence the length of \(w\) is \(p\) and it contains \(q\) holes. Furthermore, let us call the word \(T_i(w)\) in (1) the \(i\)th iterate of \(w\). Of course, each \(T_i(w)\) is a word over \(\Sigma < h \ ? j\) and, moreover, it is periodic in this alphabet. We can easily compute a period of \(T_i(w)\), as well as the number of holes in it: if \(u\) has a period of length \(a\) containing \(b\) holes, then \(F_a(u)\) has a period of length \(pa\) containing \(qb\) holes. Repeating inductively we conclude, by the iterative definition of the Toeplitz word, that \(T_i(w)\) has a period of length \(p^i\) having exactly \(q^i\) holes.

It follows that, with the exception of \(q^i\) positions, \(T(w)\) is periodic with a period of length \(p^i\). Moreover, the ratio between the numbers of non-fixed and fixed positions in this period tends to zero very rapidly.

Consequently, Toeplitz words are really, in a sense, natural extensions of periodic words. Moreover, note that the above considerations do not depend on the word \(w\) itself, nor on its alphabet \(\Sigma\), but only on the numbers \(p\) and \(q\). Note also that, as is easy to see, if \(p\) and \(q\) are relative primes then the period that we computed for \(T_i(w)\) is its smallest period.

It follows from the above that Toeplitz words are uniformly recurrent in the sense of ergodic theory (cf. [10]), i.e. for each number \(k\) there exists a constant \(n(k)\) such that whenever \(u\) with length \(k\) occurs as a factor in \(T(w)\), then it occurs in any factor of \(T(w)\) of length \(n(k)\).

3. TOEPLITZ WORDS AND ITERATION OF MORPHISMS

In this section we show that all Toeplitz words can be obtained by iterating morphisms in a suitable way. Moreover, this allows us to classify Toeplitz words.

Let \(w\) be a \((p, q)\)-Toeplitz word. We define three different cases as follows:

(i) \(q = 1\), i.e. Toeplitz words of type \((p, 1)\);

(ii) \(q\) divides \(p\), i.e. Toeplitz words of type \((tq, q)\);

(iii) \(q\) does not divide \(p\).

Next, we define three different ways of iterating morphisms corresponding to the above classification. We recall that a \(DOL\) word \(\alpha \in \Sigma^\omega\) is a word which is a fixed point of a morphism, i.e.

\[\alpha = \lim_{i \to \infty} h^i(a)\]

for some morphism \(h: \Sigma^* \to \Sigma^*\) such that \(h(a) \in a \Sigma^*\), with \(a \in \Sigma\). A \(CDOL\) word \(\beta \in \Sigma^\omega\) is a morphic image of a \(DOL\) word under a letter-to-letter morphism (or a coding). Finally, a \(p\)-\(DOL\) word \(\gamma \in \Sigma^\omega\) is a fixed point of the periodic iteration of \(p\) morphisms \(h_1, \ldots, h_p: \Sigma^* \to \Sigma^*\), i.e.

\[\gamma = \lim_{i \to \infty} h^i(a),\]
where $h_1(a) \in a\Sigma^+$ and the mapping $h: \Sigma^* \to \Sigma^*$ is defined as follows: if
\[ u = a_1 \cdots a_p a_{p+1} \cdots a_{2p} \cdots a_{(t-1)p+1} \cdots a_p a_{p+1} \cdots a_{tp+j} \]
with $t \geq 0$ and $j \leq p$, then
\[ h(u) = \left( \prod_{k=0}^{t-1} h_1(a_{kp+1}) \cdots h_p(a_{tp+p}) \right) h_1(a_{tp+1}) \cdots h_1(a_{tp+j}). \]  
(4)
Clearly, the above mapping $h$ is a simple case of a deterministic GSM (cf. [11]).

It is easy to see that there are CDOL words which are neither DOL words nor $p$-DOL words (cf. [4]). On the other hand, to find a $p$-DOL word which is not a CDOL word is not that easy. The existence of such words was proved in [16], and we shall obtain a new proof for that as a corollary of Theorem 5.

We recall that a morphism $h: \Sigma^* \to \Sigma^*$ is uniform if $|h(a)| = |h(b)|$ for all $a, b \in \Sigma$.

Now we are ready for our results of this section.

**Theorem 1.** Every Toeplitz word of type $(p, 1)$ is a DOL word generated by a uniform morphism.

**Proof.** Let the pattern be $w = a_1 \cdots a_{i-1} a_{i+1} \cdots a_p$, for $2 \leq i \leq p$. Now, define the morphism $h: \Sigma^* \to \Sigma^*$ by the condition
\[ h(a) = a_1 \cdots a_{i-1} a a_{i+1} \cdots a_p. \]
It follows from the self-reading definition of Toeplitz words that
\[ I(w) = \lim_{i \to \infty} h^i(a_i). \]
As $|h(a)| = p$ for all $a \in \Sigma$, the morphism $h$ is clearly uniform.

**Theorem 2.** Every Toeplitz word of type $(tq, q)$ is a CDOL word. Moreover, the underlying DOL word is generated by a uniform morphism.

**Proof.** Let the pattern be $w = w_1 \cdots w_i$, with $|w_1| = |w_2| = \cdots = |w_i| = q$. Furthermore, for each $i = 1, \ldots, t$, let $\bar{w}_i$ be the word obtained from $w_i$ by filling its holes in the generation of $T(w)$. Finally, let
\[ T(w) = \bar{w}_1 \bar{w}_2 \bar{w}_3 \cdots, \]
with $|\bar{w}_i| = tq$ for $i \geq 1$. Then it follows that, for $i = 1, \ldots, t$, $\bar{w}_i$ is obtained from $w$ by filling its holes with the letters of $\bar{w}_i$. Or, more generally, consecutive blocks of $T(w)$ of length $q$ determine consecutive blocks of $T(w)$ of length $tq$.

Therefore we are guided to define a morphism $h: (\Sigma^q)^* \to (\Sigma^q)^*$ by the condition: $h(a_1 \cdots a_q)$ is the word obtained from $w$ by filling its holes on letters $a_1, \ldots, a_q$ in this order, viewed as a word of length $t$ on the alphabet $\Sigma^q$.

It follows that
\[ T(w) = f \left( \lim_{i \to \infty} h^i(\bar{w}_i) \right), \]
where $f: (\Sigma^q)^* \to \Sigma^*$ is the morphism satisfying
\[ f(a_1 \cdots a_q) = a_1 \cdots a_q. \]
Consequently, we have proved that $T(w)$ can be generated by iterating a uniform morphism, and by mapping the result by another uniform morphism. To obtain a CDOL system, we have to replace the final morphism by a letter-to-letter one. It is a
general result that this can always be done, but in this particular case a simpler direct construction is possible.

Let \( \Sigma' = \Sigma^q \times \{1, \ldots, q\} \) be the new intermediate alphabet, and define three morphisms \( \varphi: (\Sigma')^* \to \Sigma^* \), \( \overline{f}: \Sigma^* \to \Sigma^* \) and \( \overline{h}: \Sigma^* \to \Sigma^* \) by \( \varphi(x) = (x, 1)(x, 2) \cdots (x, q) \), \( \overline{f}(x, j) = a_j \), and \( \overline{h}(x, j) \) is the \( j \)th block of length \( t \) of \( \varphi(h(x)) \), for \( x = a_1a_2 \cdots a_q \in \Sigma^q \) and \( j \in \{1, \ldots, q\} \). Then the following diagram

\[
\begin{array}{ccc}
(\Sigma^q)^* & \xrightarrow{h} & (\Sigma^q)^* \\
\downarrow \varphi & & \downarrow \varphi \\
\Sigma^* & \xrightarrow{f} & \Sigma^*
\end{array}
\]

is commutative, and we clearly have \( f(h'(w_i)) = \overline{f}(\overline{h}(\varphi(w_i))) \). Therefore

\[
T(w) = \overline{f}\left(\lim_{i \to \infty} \overline{h}(\varphi(w_i))\right),
\]

where \( \overline{h} \) is uniform and \( \overline{f} \) is letter-to-letter; cf. also Example 3 below.

**Example 3 (continued).** The construction of Theorem 2 yields the morphism

\[
\begin{align*}
11 & \mapsto 11 \cdot 01 \\
01 & \mapsto 10 \cdot 01 \\
h: & \mapsto 10 \mapsto 11 \cdot 00 \\
00 & \mapsto 10 \cdot 00
\end{align*}
\]

and the morphism \( f: \{00, 01, 10, 11\}^* \to \{0, 1\}^* \) mapping each block of length two into itself.

Now, a CDOL presentation for the word \( T(w) \) is obtained by defining the two morphisms \( \overline{h}: \{0, 1, \bar{0}, \bar{1}\}^* \to \{0, 1, \bar{0}, \bar{1}\}^* \) and \( \overline{f}: \{0, 1, \bar{0}, \bar{1}\}^* \to \{0, 1\}^* \) such that

\[
\begin{align*}
1 & \mapsto \bar{1} \\
\bar{1} & \mapsto 0\bar{1} \\
\bar{0} & \mapsto 0\bar{0}
\end{align*}
\]

and \( \overline{f} \) just erases the bars. Then, clearly,

\[
f\left(\lim_{i \to \infty} h'(11)\right) = \overline{f}\left(\lim_{i \to \infty} \overline{h}(11)\right).
\]

**Example 5.** Toeplitz words of Theorem 2 need not be DOL words. Such an example is the word \( w = T(1???) = 11121122112122 \cdots \).

Assume, on the contrary, that \( w \) is a fixed point of a morphism \( h \). Clearly, \( h(1) \) is a prefix of \( w \), and it is different from 1, 11 and 111. Therefore \( h(1) = 111u \) for some non-empty word \( u \), so that 111u111u is a prefix of \( w \), and the form of \( w \) implies that the length of 111u is divisible by 4. Consequently, by the self-reading definition of \( w \), both the first and second half of 111u define the same word, namely 111u. Hence we can write 111u = 111u111u. Now, we can proceed by induction and conclude that \( w \) contains a prefix 11121112, a contradiction.
Actually, we believe that, typically, Toeplitz words of Theorem 2 are not DOL words.

Our last result shows that all of the Toeplitz words that we have been considering are $p$-DOL words for some $p$. This extension is, however, necessary since as an application of the subword complexity results of Section 5 we see that there are Toeplitz words which are not CDOL words.

**Theorem 3.** Every $(p, q)$-Toeplitz word is a $q$-DOL word.

**Proof.** Consider a $(p, q)$-Toeplitz word $T(w)$ defined by the pattern

$$w = w_0w_1 \cdots w_q,$$

with $w_0 \in \Sigma^*$ and $w_i \in \Sigma^*$ for $i = 1, \ldots, q$. Define the morphisms $h_i : \Sigma^* \to \Sigma^*$, for $i = 1, \ldots, q$, as follows. For $i < q$, set

$$h_i(a) = w_{i+1}a$$

for $a \in \Sigma$,

and, for $i = q$ set

$$h_q(a) = w_qaw_q$$

for $a \in \Sigma$.

It follows directly from the self-reading definition of Toeplitz words that

$$T(w) = \lim_{i \to \infty} h'(a_i),$$

where $h$ is defined as in (4) and $a_1$ denotes the first letter of $w_0$. □

4. Periodic Toeplitz Words

As we have said, in each case of our classification of Toeplitz words we can obtain periodic words; cf. Example 1. In the case of one hole this is possible only if the alphabet $\Sigma$ is unary. The goal of this section is to characterize when a Toeplitz word is periodic. Theorem 4 states that it is sufficient to examine a short prefix of a Toeplitz word in order to check if it is periodic.

**Theorem 4.** Let $T(w)$ be a $(p, q)$-Toeplitz word, and let $d = \gcd(p, q)$. Then $T(w)$ is periodic if and only if its prefix of length $p$ is $d$-periodic. In this case, the minimal period of $T(w)$ is a divisor of $d$.

Note that this means that the size of the alphabet $\Sigma$ of the pattern $w$ cannot be larger than $d$ whenever $T(w)$ is periodic. In particular, when $p$ and $q$ are coprime, $d = 1$ and $T(w)$ is not periodic as soon as the pattern contains two different letters.

The proof of Theorem 4 is divided into three lemmas.

**Lemma 1.** Let $T(w)$ be a $(p, q)$-Toeplitz word, $d = \gcd(p, q)$, and $p' = p/d$. Suppose that the prefix of length $p$ of $T(w)$ is $v^{p'}$, where $v$ is a word of length $d$. Then $T(w) = v^n$.

**Proof.** We prove, by induction on $n$, that $T(w)$ and $v^n$ have the same prefix of length $n$ for every $n \in \mathbb{N}$.

(i) The property holds for $n = 1$, since $T(w)$ has the same prefix of length $p$ as $v^n$.

(ii) Suppose that the prefixes of length $n - 1$ of $T(w)$ and $v^n$ are the same, with $n > p$. Let $n = kp + i$, $1 \leq i \leq p$. If the $i$th letter of $w$ is not a hole, then the $n$th letter of
T(w) is equal to this letter and also to the $i$th letter of $T(w)$. Otherwise, suppose that the $i$th letter of $w$ is the $j$th hole, with $1 \leq j \leq q$. Then the $n$th letter of $T(w)$ is equal to the $(kq + j)$th letter of $T(w)$, which in turn is equal to the $j$th letter of $T(w)$, since $kq + j \leq n - 1$ and $T(w)$ is $d$-periodic up to this point. Now this letter fills the $j$th hole in $w^\omega$, i.e. is equal to the $i$th letter of $T(w)$. In both cases, we conclude that the $n$th letter of $T(w)$ is the same as the $i$th and $n$th letters of $w^\omega$.  

We now need some notation.

Let us denote by $u_n$ the $n$th letter of the $(p, q)$-Toeplitz word $T(w)$, and by $p'$ and $q'$ the quotients $p' = p/d$ and $q' = q/d$. Let $\tilde{w}$ be the word in $(\Sigma \cup \{h_1, \ldots, h_q\})^\ast$ obtained from $w$ by replacing the $j$th hole by the symbol $h_j$. Let $w_i$ be the $i$th letter in $\tilde{w}$.

We define a directed graph $G(w)$ with $d + \#\Sigma$ vertices labelled by the numbers $1, \ldots, d$ and the letters of $\Sigma$, and at most $p$ edges defined as follows: there is an edge from the number $i$ to the letter $x$ whenever there is an integer $k$ such that $w_{kd+i} = x$, and there is an edge from the number $i$ to the number $j$ whenever there are integers $k$ and $l$ such that $w_{kd+i} = h_{ld+j}$. Let $\Sigma_i(w)$, for $1 \leq i \leq d$, be the set of letters $x$ in $\Sigma$ such that there is a directed path from $i$ to $x$ in $G(w)$. Note that $\Sigma_i(w)$ is never empty.

In the rest of this section, since there is only one pattern involved, we will use the shorter notation $G = G(w)$ and $\Sigma_i = \Sigma_i(w)$.

**Example 6.** The pattern $w = a?bbaa?ba?b$ defines a $(12, 4)$-Toeplitz word. To construct the graph $G$, we compare $i \mod d$ (here $d = 4$) and $w_i$ for $1 \leq i \leq p$:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$i \mod d$</th>
<th>$w_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$a$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$h_1$</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>$b$</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$b$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>$a$</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>$a$</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>$h_2$</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>$b$</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>$a$</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>$h_3$</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>$b$</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>$h_4$</td>
</tr>
</tbody>
</table>

and we obtain the following graph:

![Diagram](https://example.com/diagram.png)
with $\Sigma_1 = \{a\}$, $\Sigma_2 = \Sigma_3 = \{a, b\}$ and $\Sigma_4 = \{b\}$. As we shall see below, the fact that $\Sigma_2$ and $\Sigma_3$ have two elements imply that $T(w)$ is not periodic. In this example we took, for clarity, $\Sigma = \{a, b\}$ instead of $\Sigma = \{1, 2\}$.

**Lemma 2.** The subsequence $(u_{nd+i})_{n \geq 0}$ of $T(w)$ has exactly alphabet $\Sigma_i$, for $1 \leq i \leq d$. Moreover, for every letter $x \in \Sigma_i$, there exists an integer $k$, with $0 < k < p'$, such that

$$\forall n \geq 0, \quad u_{nd+p'j+kj} = x,$$

where $l \leq d$ is the length of one of the shortest paths from $i$ to $x$ in $G$.

**Proof.** First we prove by induction on $n$ that, for all $n$, $u_n \in \Sigma_i$, where $i = n \pmod{d}$, i.e. the alphabet of the subsequence $(u_{nd+i})_{n \geq 0}$ is a subset of $\Sigma_i$.

(i) It holds when the $n$th letter $u_n$ in $w^*$ is not a hole, since there is then a path of length 1 from $i$ to $u_n$ in $G$. This is true in particular when $n = 1$.

(ii) Suppose that it holds for every $n$ less than $m$, and suppose that the $m$th letter in $w^*$ is a hole. Then the $m$th letter in $w^*$ is some $h_{kd+j}$, with $1 \leq j \leq d$, and there is an edge in $G$ from $i$ to $j$, where $i = m \pmod{d}$. Then $u_m = u_{n/p}q' + kd + j < m$, so that by the induction hypothesis this letter is in $\Sigma_j$, and hence also in $\Sigma_i$ because of an edge from $i$ to $j$.

Next, we prove the last part of Lemma 2, which implies that every letter in $\Sigma_i$ actually occurs in the subsequence $(u_{nd+i})_{n \geq 0}$. We prove, by induction on $l$, that if there is a path of length $l$ from $i$ to $x$, then there is some $k$, with $0 < k < p'$, such that $\forall n \geq 0$, $u_{nd+p'j+kj} = x$.

(i) It holds for $l = 1$, with $k < p'$ such that $w_{kd+i} = x$. Such a $k$ exists since there is an edge from $i$ to $x$ in the graph $G$.

(ii) Suppose that it holds for $l - 1$, and that there is a path of length $l$ from $i$ to $x$, starting with an edge from $i$ to $j$, followed by a path of length $l - 1$ from $j$ to $x$. Then, by the induction hypothesis, there is $k' < p'^{l-1}$ such that, for all $n$, $u_{nd+p'j+k'j} = x$. Let $k'' < p'$ be such that $w_{k''j+kd+i} = h_{kd+j}$. Then, for all $n$, $u_{np+kd+i} = u_{np+kd+j}$. By Bezout's theorem, there are two integers $\alpha$ and $\beta$ such that $ap'^{l-1} + k' = \beta q'' + k''$, since $p'$ and $q''$ are coprime, and it is always possible to choose these integers so that $0 \leq \beta < p'^{l-1}$. Let $k = \beta p' + k''$, which satisfies $0 \leq k < p'$. Then, for all $n$, $u_{nd+p'j+kj} = u_{np+kd+i} = u_{np+kd+j} = x$.

As a corollary of Lemma 2, we obtain a useful criterion for testing whether or not $T(w)$ is $d$-periodic.

**Lemma 3 (periodicity criterion).** The Toeplitz word $T(w)$ is $d$-periodic if and only if $\Sigma_i$ has exactly one element for all $i$.

**Proof.** $T(w)$ is $d$-periodic if and only if the subsequences $(u_{nd+i})_{n \geq 0}$ are constant for all $1 \leq i \leq d$. According to the first part of Lemma 2, this is equivalent to the fact that $\Sigma_i$ has exactly one element.

Now we can proceed with the proof of Theorem 4.

**Proof of Theorem 4.** One implication is easy: if the prefix of length $p$ of $T(w)$ is $d$-periodic then, by Lemma 1, the Toeplitz word $T(w)$ is $d$-periodic as well.

Let us now prove the other direction of the theorem. Suppose that $T(w)$ is periodic, with the minimal period $t$. Then lcm$(t, p)$ is also a period of $T(w)$, and since $T(w)$ is fixed by the transformation that substitutes $T(w)$ into the holes of $w^*$, then
(q/p)\text{lcm}(t,p)\) is also a period of \(T(w)\). Therefore \(\gcd(t, (q/p)\text{lcm}(t,p)) = (1/p)\text{lcm}(t,p)\gcd(q,(t,d), q) = t \gcd(t,d)/\gcd(t,p)\) is also a period of \(T(w)\). Since \(\gcd(t,d)/\gcd(t,p) = 1\) and \(t\) is minimal, this is possible only if \(\gcd(t,p) = \gcd(t,d)\). Let \(t' = t/\gcd(t,d)\), so that \(t'd\) is a period and \(\gcd(t', p') = 1\).

Now let \(x\) and \(y\) be two elements of \(\Sigma\). According to Lemma 2, there are integers \(l\) and \(k\) such that, for all \(n\), \(u_{nlp^{l}+kd+i} = x\) and \(u_{nlp^{l}+kd+i} = y\). By Bezout’s theorem, there are two integers \(\alpha\) and \(\beta\), which can be made positive, such that \(\alpha t' + k = \beta p' + k'\), since \(t'\) and \(p'\) are coprime. Then \(x = u_{kd+i} = u_{\alpha t'+k+d+i} = u_{\beta p'+k'+d+i} = y\), which shows that \(\Sigma\) has only one element. By the above periodicity criterion, \(T(w)\) is therefore \(d\)-periodic.

5. Complexity of Toeplitz Words

In this section we compute the subword complexity of Toeplitz words. We recall that the subword complexity, or briefly the complexity, of an infinite word \(u\) is the function \(p_{u}: \mathbb{N} \to \mathbb{N}\) such that

\[
p_{u}(n) = \text{the number of factors of } u \text{ of length } n.
\]

Recently, quite a lot of research has been done in this field, (cf. [1, 15]).

It follows from our Theorems 1 and 2, by a result in [7], that in our first two cases of Toeplitz words, i.e., when the number of holes divides the length of the pattern, the subword complexity is linear. In this section we show that arbitrary Toeplitz words have polynomial complexity, or more precisely:

**Theorem 5.** Let \(T(w)\) be a non-periodic \((p,q)\)-Toeplitz word, and define \(d = \gcd(p,q)\), \(p' = p/d\), \(q' = q/d\). Then the complexity \(p(n)\) of \(T(w)\) satisfies \(p(n) = \Theta(n')\) with \(r = (\log p')/(\log p' - \log q')\), i.e., there exist two positive constants \(C_{1}\) and \(C_{2}\) such that \(C_{1}n' < p(n) < C_{2}n'\) for every \(n > 0\).

In order to simplify the proof of Theorem 5, we first show that it is sufficient to study Toeplitz words with two additional properties, involving the sets \(\Sigma_{i}(w)\) defined in the previous section.

**Definition 1.** A pattern \(w\) is said to be reduced if it satisfies the following two properties:

(i) each of the sets \(\Sigma_{i}(w)\) associated to the pattern \(w\) has at least two elements;

(ii) \(w\) is a primitive word.

**Lemma 4.** Let \(T(w)\) be a non-periodic Toeplitz word. Then there exist a reduced pattern \(w''\) with \(|w''| = |w|\) and \(|w''| = |w|/|w|\), and a second pattern \(z\) such that \(T(w) = E_{z}(T(w''))\). Moreover, the complexities \(p(n)\) and \(p''(n)\) of \(T(w)\) and \(T(w'')\) are linked: \(p(n) = \Theta(n')\) if and only if \(p''(n) = \Theta(n')\).

**Proof.** We first construct a pattern \(w'\) satisfying (i), then we reduce it to a primitive word \(w''\).

1. Consider the integers \(p'\), \(q'\) and \(d\) and the sets \(\Sigma_{i}(w)\) associated in the previous section to the pattern \(w\), and let \(I\) be the set of indices \(i\) such that \(#\Sigma_{i}(w) = 1\). Note that \(I \neq \{1, \ldots, d\}\) since \(T(w)\) is non-periodic. In the word \(\tilde{w}\), replace \(h_{j}\) by the unique letter of \(\Sigma_{i}(w)\) whenever \(j = i (mod d)\) and \(i \in I\). The \((kd + i)\)th letter of this word, for \(i \in I\), is the element of \(\Sigma_{i}(w)\). Delete all these letters, yielding the new word \(\tilde{w}'\) of length \(p'd'\) and containing \(q'd'\) holes, with \(d' = d - \#I\). Let \(w'\) be the pattern obtained by replacing any \(h_{j}\) by \(?\) in \(\tilde{w}'\). This pattern satisfies (i), since the sets \(\Sigma_{i}(w')\) are exactly...
the sets $\Sigma_j(w)$ with $j \not\in I$, and we have $|w'| = p'd' \leq p'd = |w|$ and $|w'|/|w'| = p'/q' = |w|/|w'|$. Let $z$ be the pattern of length $d$ defined by the conditions: $z_i \in \Sigma_i(w)$ if $i \in I$, and $z_i = ?$ otherwise. Then $T(w)$ is obtained from $T(w')$ by inserting it into the holes of $z''$, i.e. $T(w) = F_e(T(w'))$.

2. Let $w'$ be the primitive root of $w'$, and let $w' = w^{dk}$. Then, clearly, $T(w') = T(w''')$, so that $T(w) = F_e(T(w'''))$, and $T(w')$ is a $(p'd', q'd')$-Toeplitz word with $d'' = d'/k$; hence $|w''| \leq |w|$ and $|w''|/|w''| = |w|/|w'|$. Note that the sets $\Sigma_j(w')$ associated to $w''$ are given by

$$\Sigma_j(w') = \sum_{i=0}^{k-1} \Sigma_{j+ik}(w'),$$

and have thus at least two elements.

Let $p(n)$ and $p'(n) = p''(n)$ be the complexities of $T(w)$ and $T(w') = T(w'')$, respectively. The factor of length $kd$ of $T(w)$ beginning at position $md + 1$ is obtained by inserting the factor of length $kd'$ of $T(w')$ beginning at position $md' + 1$ into the holes of $z^k$: therefore the number of factors of length $kd$ of $T(w)$ beginning at position $1$ (mod $d$) is at most $p'(kd')$. A factor of length $kd$ of $T(w)$ beginning at any other position can be extended into a factor of length $(k + 1)d$ beginning at position $1$ (mod $d$), so that there are at most $p'((k + 1)d')$ such factors. Consequently, summing these numbers for every possible position modulo $d$, we obtain an upper bound:

$$p(kd) \leq p'(kd') + (d - 1)p'((k + 1)d') \leq dp'((k + 1)d').$$

Conversely, a factor of length $kd'$ of $T(w')$ beginning at position $1$ (mod $d'$) can be obtained from a factor of length $kd$ of $T(w)$ by keeping only letters corresponding to the holes of $z'k$. Extending other factors as above, we obtain the relation $p(kd) \geq (1/d')p'((k - 1)d')$.

From the above two inequalities we conclude that $p(n) = \Theta(n')$ if and only if $p'(n) = \Theta(n')$.

EXAMPLE 6 (continued). The pattern $w = 1?2211?21??$ can be reduced as $\Sigma_i(w)$ and $\Sigma_j(w)$ have only one element. The resulting reduced pattern is $w' = 121?22$, and $T(w) = F_e(T(w'))$ with $z = 1??2$.

We now suppose that the pattern $w$ is reduced, and establish the following lemma.

LEMMA 5. Let $T(w)$ be a non-periodic $(p, q)$-Toeplitz word, with $w$ reduced. Define $n_0 = dp''$, and let $v$ be any factor of $T(w)$ such that $|v| > n_0$. Then the positions at which $v$ occurs in $T(w)$ are all equal modulo $p$.

PROOF. Let $m$ be one of the positions at which $v$ occurs, i.e.

$$v = u_m u_{m+1} \cdots u_{m+|v|-1},$$

and let $1 \leq i \leq p$. If the $(m + i - 1)$th letter of $v$ is not a hole, then the $(tp + i)$th letter of $v$ is equal to this letter of $y$ for every $t$ such that $1 \leq tp + i \leq |v|$. On the other hand, if the $(m + i - 1)$th letter in $\hat{w}^w$ is a hole $h_{tp+i}$, then the $(tp + i)$th letter of $v$ is equal to the $(t + t')q + ld + j$th letter of $T(w)$, where $t'$ is some constant. Now using Lemma 2 and the fact that $q'$ and $p''$ are coprime, we see that for each $x \in \Sigma_i(w)$ there is an integer $t$, with $0 \leq t < p''$, such that

$$\forall m \geq 0, \quad u_{(tq + t'q + ld + j)} = x.$$
are not all equal. Looking at these letters, we can therefore decide whether or not \( y \) is a hole, and so by determining in this way \( p \) consecutive letters in \( w^o \) at positions \( m, m+1, \ldots, m+p-1 \), we can compute \( m \) modulo \( p \), since \( w \) is a primitive word. So the lemma is proved. \( \square \)

We can now prove the theorem.

**Proof of Theorem 5.** According to Lemma 4, it is sufficient to consider the case in which \( w \) is reduced. Indeed, the reduction in Lemma 4 does not change the values of \( p' \) and \( q' \) and hence neither the exponent \( r \). Using Lemma 5, we can decompose \( p(n) \), when \( n \geq n_0 \), as

\[
p(n) = \sum_{i=1}^{d} p_i(n),
\]

where \( p_i(n) \) is the number of factors of length \( n \) occurring at a position equal to \( i \) modulo \( d \) in \( T(w) \).

For any \( i \), let \( j(i) \) be such that \( h_{j(i)} \) is the first hole in \( \tilde{w}^o \) after or at the position \( i \), and let \( l(i, n) \) be the total number of holes in the factor of length \( n \) of \( w^o \) starting at position \( i \). Clearly,

\[
q \lfloor n/p \rfloor \leq l(i, n) \leq q \lceil n/p \rceil.
\]

For a word \( v \) of length \( n \) occurring at position \( i \) in \( T(w) \), let \( \tilde{v} \) be the word of length \( l(i, n) \) obtained from \( v \) by keeping only letters which were originally holes in \( \tilde{w}^o \). If \( n \geq n_0 \), the word \( \tilde{v} \) is uniquely determined from \( v \) according to Lemma 5. The word \( \tilde{v} \) is also a factor of \( T(w) \), and occurs at position \( j \) with \( j = j(i) \) (mod \( q \)). Conversely, if \( n \) and \( i \) are given and \( v' \) is a factor of length \( l(i, n) \) occurring at a position \( j' = j(i) \) (mod \( d \)), then it certainly also occurs at a position \( j = j(i) \) (mod \( q \)), and so there is a word \( v \) occurring at position \( i' \), with \( i' = i \) (mod \( p \)), such that \( v' = \tilde{v} \). Therefore, the number of factors occurring at position \( i \) modulo \( p \) is exactly \( p_{j(i)}(l(i, n)) \). We thus obtain the recurrence relation, valid for all \( i \) and \( n \geq n_0 \):

\[
p_i(n) = \sum_{k=0}^{p' - 1} p_{j(kd+i)}(l(kd+i, n)).
\]

This relation is enough to compute \( p(n) \).

Let \( r = (\log p')/(\log p' - \log q') \), and choose two integers \( n_1 \) and \( n_2 \) such that

\[
pq/(p - q) < n_1 \leq q\lfloor n_2/p \rfloor \leq n_2
\]

and \( n_2 \geq n_0 \). Next we choose constants \( C_1 \) and \( C_2 \) such that, for all \( i \) and for \( n_1 \leq n < n_2 \),

\[
C_1\left(n + \frac{pq}{p - q}\right)^r \leq p_i(n) \leq C_2\left(n - \frac{pq}{p - q}\right)^r.
\]

Now we prove by induction on \( n \) that the same inequality holds for any \( n \geq n_1 \).

(i) By the construction, the inequality holds for \( n_1 \leq n < n_2 \).

(ii) Suppose that it holds for \( n_1 \leq n < m \), with \( m \geq n_2 \), and let \( n = m \). According to (8), we then have

\[
p' C_1\left(q^\left\lfloor n/p \right\rfloor + \frac{pq}{p - q}\right)^r \leq p_i(n) \leq p' C_2\left(q^\left\lceil n/p \right\rceil - \frac{pq}{p - q}\right)^r
\]
as \( n_1 \leq q \lfloor n/p \rfloor \leq l(kd + i, n) \leq q \lceil n/p \rceil < m \) which, in turn, follows from (7) and (9). Then
\[
p_c(n) \geq p'C_1\left(\frac{q}{p}n - q + \frac{pq}{p - q}\right) = C_1\left(n + \frac{pq}{p - q}\right)^r
\]
since \( p' = \left(\frac{p}{q}\right)^r \) and, similarly, \( p_c(n) \leq C_2(n - pq/(p + q))^r \).

It now follows from (6) that, for \( n \geq n_2 \), \( dC_1n^r \leq p(n) \leq dC_2n^r \); that is, \( p(n) = \Theta(n^r) \).

\[\square\]

**Example 7.** The pattern \( w = 123 \ldots \) generates the \((5, 3)\)-Toeplitz word
\[
T(w) = 121211211222112112122112121222112121121212222112112 \ldots
\]
According to Theorem 5, the complexity of this word satisfies \( p(n) = \Theta(n^r) \) with \( r = (\log 5)/(\log 5 - \log 3) = 3.15066 \). It was proved in [7] that CDOL words have a complexity \( p(n) = O(n^2) \). Therefore \( T(w) \) is not a CDOL word.

As \( d = \gcd(5, 3) = 1 \), the recurrence relation (8) can be written in a simpler form:
\[
\begin{align*}
p(5n) &= 5p(3n) \\
p(5n + 1) &= 2p(3n) + 3p(3n + 1) \\
p(5n + 2) &= p(3n) + 2p(3n + 1) + 2p(3n + 2) \\
p(5n + 3) &= 2p(3n + 1) + 2p(3n + 2) + p(3n + 3) \\
p(5n + 4) &= 3p(3n + 2) + 2p(3n + 3),
\end{align*}
\]
valid for \( n \geq n_0/5 = 5 \) (in fact, also for \( n = 4 \)), with the initial values:

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p(n) )</th>
<th>( n )</th>
<th>( p(n) )</th>
<th>( n )</th>
<th>( p(n) )</th>
<th>( n )</th>
<th>( p(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
<td>30</td>
<td>10</td>
<td>240</td>
<td>15</td>
<td>878</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>52</td>
<td>11</td>
<td>340</td>
<td>16</td>
<td>1071</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>86</td>
<td>12</td>
<td>466</td>
<td>17</td>
<td>1336</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>128</td>
<td>13</td>
<td>598</td>
<td>18</td>
<td>1626</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>9</td>
<td>176</td>
<td>14</td>
<td>733</td>
<td>19</td>
<td>1952</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To obtain a better estimation of \( p(n) \), it is useful to study the quotient \( p(n)n^{-r} \). This quotient is bounded, but experimentation suggests that it has no limit when \( n \to \infty \). Indeed, it can be shown for this example that \( p(n)n^{-r} = F(\log n) + O(1/n) \), where \( F \) is a periodic and continuous function, with period \( \log \frac{5}{3} \). This is not very surprising, as similar recurrences are known to produce this kind of asymptotic behaviour (cf. [9]).

We conclude this section with a few remarks. First, by choosing \( p = 2k + 1 \) and \( q = 2k - 1 \), we can define the pattern \( w_k \) such that \( p_{T(w_k)} = \Theta(n^r) \), with \( \lim_{k \to +\infty} r_k = \infty \). Hence, the complexities of Toeplitz words are not bounded by polynomials of any fixed degree. Consequently, the same property holds for the more general case of complexities of words obtained by iterating periodically several morphisms. On the
other hand, the complexity of a given Toeplitz word is always polynomial, but we do not know whether other \( q \)-DOL words can have exponential complexity.

6. CONCLUDING REMARKS

We have considered one-way infinite Toeplitz words over a finite alphabet. We recalled the fact that they provide a natural generalization of periodicity, and also pointed out new evidence on that. Namely, by Theorem 5, the complexity of a Toeplitz word defined by a pattern of a certain length \( p \) and with a certain number \( q \) of holes is independent of this pattern, but depends only on \( p \) and \( q \), provided that the Toeplitz word is not periodic.

Our main concern was, on the one hand, to classify different types of Toeplitz words in terms of possibilities of defining them by iterated morphism and, on the other hand, to compute the complexity of Toeplitz words.

We noted that the relative sizes of the above \( p \) and \( q \) play an important role here. In general, Toeplitz words can be generated by iterating periodically a finite number of morphisms. If there is only one hole, that is \( q = 1 \), then one morphism is enough. An interesting in-between case occurs when \( p \) divides \( q \). Then one morphism is enough if, after the iteration, another one is allowed to translate the result; but, as pointed out in Example 5, this other morphism is necessary.

Concerning the complexity of Toeplitz words, \( p \) and \( q \) are important too. When computing the complexity in Theorem 5 we at the same time introduced a new possibility for the complexity of infinite words (cf. [1]). Moreover, Theorem 5 allows us to re-prove a result of Lepistö in [16]: there exist \( q \)-DOL words that are not CDOL. Namely, taking \( p = 5 \) and \( q = 3 \) as in Example 7, we obtain a Toeplitz word having a complexity that is more than quadratic, and hence, by a result in [7], it cannot be generated by iterating a single morphism and then applying another one to the result.

REFERENCES


Received 13 November 1995 and accepted 16 July 1996

**Julien Cassaigne**

*Institut de Mathématiques de Luminy,*

*Case 930,*

*163, avenue de Luminy,*

*F-13288 Marseille Cedex 9, France*

*E-mail: cassaign@iml.univ-mrs.fr*

**Juhani Karhumäki**

*Department of Mathematics,*

*University of Turku,*

*SF-20500 Turku, Finland*

*E-mail: karhumak@cs.utu.fi*