Repetitions in words

Recent results and open problems

Jean Berstel

http://www-igm.univ-mlv.fr/~berstel

Institut Gaspard-Monge
Université de Marne-la-Vallée
France
Outline

- Subword complexity of finite words
- Construction of infinite words
- Power-free words and power-free morphisms
- Counting power-free words
- Test sets and test words for power-free morphisms
- Open problems
Subword complexity of finite words

\[ p_x(n) = \text{number of distinct factors of } x \text{ of length } n. \]

Example:
1) \( x = 0011001 \)

\[
\begin{array}{c|cccccccc}
  n & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
  \hline
  p_x(n) & 1 & 2 & 4 & 4 & 4 & 3 & 2 & 1 \\
\end{array}
\]

\( M_x = 2, \ H_x = 4 \)

\[ M_x = \min \{ i \mid p_x(i) \text{ is maximal} \} \]
\[ H_x = \max \{ i \mid p_x(i) \text{ is maximal} \} \]
2) de Bruijn words \(=\) product of Lyndon words of length dividing \(n\) (Fredricksen Maiorana):

\[
\begin{align*}
n &= 3 & x &= 00010111 & w &= 0001011100 \\
001 & & 011 & & 1
\end{align*}
\]

\[
\begin{array}{cccccccccccc}
 n & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
p_w(n) & 1 & 2 & 4 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
\]

\[
M_w = H_w = 4
\]
The suffix tree of the word $w = abccbabcab$

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_w(n)$</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

$M_w = 3, H_w = 4$
A factor is *unrepeated* if it appears only once in $w$.

Recall the definition $H_x = \max\{i \mid p_x(i) \text{ is maximal}\}$

**Theorem** (Carpi & de Luca) $H_w$ is the smallest $n$ such that any factor of length $\geq n$ is unrepeated.
Subword complexity of finite words (cont’d)

**Theorem** (Carpi & de Luca, Levé & Séébold) *For any word* \(w\), there is an integer \(M_w\) such that:

1. \(p_w(i) < p_w(i + 1)\) for \(0 \leq i < M_w\),
2. \(p_w(i) = p_w(i + 1)\) for \(M_w \leq i < H_w\),
3. \(p_w(i) = p_w(i + 1) + 1\) for \(H_w \leq i < |w|\).

2. is by definition, 3. is easy, 1. is more difficult.

**Theorem** (Carpi & de Luca) *A word* \(w\) *is determined by its factors of length at most* \(H_w + 1\) *and even by the poset of these factors*.

**Example** The word \(w = abccbabcbab\) of length \(10\) is entirely defined by its factors of length \(5\).
**Theorem** (Morse & Hedlund, Coven & Hedlund).

Let $x$ be an infinite word over $k$ letters. The following are equivalent:

1. $x$ is ultimately periodic,
2. $p_x(n) = p_x(n + 1)$ for some $n$,
3. $p_x(n) < n + k - 1$ for some $n \geq 1$,
4. $p_x(n)$ is bounded.

Thus, either $p_x(n)$ is ultimately constant or $p_x(n) \geq n + 1$ for all $n$. A word is **Sturmian** if $p_x(n) = n + 1$ for all $n \geq 0$. A Sturmian word is binary because $p_x(1) = 2$. 
Construction of infinite words

Characteristic words are representations of sets of integers.

a) Squares $0, 1, 4, 9, \ldots$

$$11001000010000001000 \ldots$$

b) The spectrum of a number $\tau$ is the set $S_\tau = \{ \lfloor n\tau \rfloor : n \geq 1 \}$. For $\tau = \frac{1+\sqrt{5}}{2}$, it is $1, 3, 4, 6, 8, \ldots$. The infinite binary word $f$ is defined by

$$f_n = \begin{cases} a & \text{if } n + 1 \in S_\tau \\ b & \text{otherwise} \end{cases}$$

$$f = abaababaabaababaababaababaabaababa \ldots$$
Explicit description

c) Thue-Morse word

\[ t = 01101001100101101001011001101001 \cdots \]

defined by

\[ t_n = \text{the number of 1's in the binary expansion } \text{bin}(n) \text{ of } n \text{ modulo 2}. \]

d) More generally, by a finite automaton working on binary expansion:

\[ t_n = 1 \text{ iff } \text{bin}(n) \text{ is accepted by the automaton.} \]

These are *automatic sequences*.

e) By induction:

\[ t_{2n} = t_n \]
\[ t_{2n+1} = 1 - t_n \]

This gives the Thue-Morse sequence if \( t_0 = 0 \) and its opposite if \( t_0 = 1 \).
Infinite products

Any infinite product $x_0 x_1 \cdots x_n \cdots$ of nonempty words has a limit.

\[ c = 0110111001011101111000 \cdots \]

The *Champernowne* word is the product of the words $\text{bin}(n)$ (binary representation of $n$).

- Every word is factor of $c$: $p_c(n) = 2^n$.
- $c$ is recurrent: every factor that appears in $c$ appears infinitely many times.
- It is not uniformly recurrent: the gap between consecutive occurrences of a given factor is not bounded.
Morphisms

A morphism \( h : A^* \to B^* \) is a function satisfying

\[
h(xy) = h(x)h(y)
\]

for all words \( x \) and \( y \). It is defined by the values on the letters.

**Example.** The *Thue-Morse* morphism \( \mu : \begin{align*} 0 & \mapsto 01 \\ 1 & \mapsto 10 \end{align*} \). For instance

\[
\mu(1100) = 10100101.
\]

For \( t = 01101001100101101001011011010101 \ldots \)

one gets

\[
\mu(t) = t
\]

that is \( t \) is a fixed point for \( \mu \).
A morphism $h : A^* \rightarrow A^*$ is \textit{prolongable} in the letter $a$ if

$$h(a) = ax$$

for some word $x$ with $h^n(x) \neq \varepsilon$ for all $n \geq 0$. Then

$$h^2(a) = axh(x)$$
$$h^3(a) = axh(x)h^2(x)$$
$$h^n(a) = axh(x)h^2(x) \cdots h^{n-1}(x)$$

and the sequence $(h^n(a))$ converges to

$$h^\omega(a) = axh(x)h^2(x) \cdots h^n(x) \cdots$$
Words generated by a morphism (cont’d)

\[ h : \begin{align*}
a & \mapsto aba \\
b & \mapsto abb
\end{align*} \]

Then
\[ h^3(a) = a ba abbaba abaabbababaabbaba \]

Of course, \( h^n(a) \) is always a prefix of \( h^{n+1}(a) \). For \( z = h^\omega(a) \), one has
\[ z = h(z) \]
that is \( z \) is a fixed point of \( h \).
Recurrence relations between words associated to morphisms.

\[
h : \begin{array}{c}
a \mapsto aba \\
b \mapsto abb
\end{array}
\]

Set \( u_n = h^n(a) \) and \( v_n = h^n(b) \). Then

\[
\begin{align*}
    u_{n+1} &= u_n v_n u_n \\
v_{n+1} &= u_n u_n v_n .
\end{align*}
\]

Thus, a (binary) morphism gives a system of recurrence relations for the words \( u_n \) and \( v_n \).
Substitution

\[ f : B^* \to B^* \] a morphism prolongable in the letter \( b \).
\[ g : B^* \to A^* \] be a letter-to-letter morphism.
The pair \((f, g)\) is a substitution. It generates the word \( g(f^\omega(b)) \).

The word of squares

\[ s = 1100100001000000100 \cdots \]

is generated by

\[ a \mapsto a1 \quad a \mapsto 1 \]
\[ f : 1 \mapsto 001 \quad g : 1 \mapsto 1 \]
\[ 0 \mapsto 0 \quad 0 \mapsto 0 \]

Indeed

\[ f^\omega(a) = a100100001 \cdots \]
A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921): *is it decidable whether, for a word* \(w\) *on the tape, the reading head can reach the writing head?*

Kolakoski sequence  \[
\begin{array}{cccccccccc}
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1 & \cdots \\
\end{array}
\]
A *Tag machine* is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921): *is it decidable whether, for a word* $w$ *on the tape, the reading head can reach the writing head?*

Kolakoski sequence  

\[
\begin{array}{ccccccccc}
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1 & \cdots \\
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1 & \\
\end{array}
\]

\[
\begin{array}{llllll}
2 & 2 & 1 & 1 & 2 & 2 \\
2 & 2 & 1 & 1 & 2 & 2 \\
\end{array}
\]

\[
\begin{array}{llllll}
2 & 2 & 1 & 1 & 2 & 2 \\
2 & 2 & 1 & 1 & 2 & 2 \\
\end{array}
\]
A **Tag machine** is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921): *is it decidable whether, for a word $w$ on the tape, the reading head can reach the writing head?*

**Kolakoski sequence**

\[
\begin{array}{cccccccc}
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1 & \cdots \\
\end{array}
\]

![Diagram of Kolakoski sequence with arrows indicating the movement of the heads.](image-url)
A **Tag machine** is a machine with one read-only head and one write-only head. Both move on the same tape, from left to right only. The output depends on the state of the machine and on the input.

The question asked originally by Post (1921): *is it decidable whether, for a word \( w \) on the tape, the reading head can reach the writing head?*

**Kolakoski sequence**

\[
\begin{array}{cccccccccccc}
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1 \\
2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 1
\end{array}
\]
Iterating sequential functions

A sequential function is a morphism with states. For the Kolakoski sequence:

The sequential machine can be viewed as a special case of a Tag machine, when reading and writing on the same tape.
**Toeplitz words**

\[ x = x_0 ? x_1 ? x_2 ? \cdots \] with \( x_n \) words and ? a placeholder.

\[ y = a_0 a_1 a_2 \cdots \] with \( a_n \) letters.

The **Toeplitz** product is

\[ x \tau y = x_0 a_0 x_1 a_1 x_2 a_2 \cdots \]

**Example** Consider \( x = ab?ab?ab? \cdots = (ab?)^\omega \). Then

\[ x \tau x = abaabbab?abaabbab? \cdots = (abaabbab?)^\omega \]

\[ x \tau x \tau x = abaabbabaabaabbababbabaababababaababab \cdots \]

Under mild conditions, the limit exists, and is of course a fixed point:

\[ y = x \tau y \]
We consider words \( x = w^\omega \) of type \((p, q)\), that is \(|w| = p\) and \(w\) contains \(q\) placeholders. (E.g. \( w = aa?b? \) has type \((5, 2)\).)

**Theorem** (Cassaigne & Karhumäki) Let \( y \) be generated by a word of type \((p, q)\).

- if \( q = 1 \), then \( y \) can be obtained by iterating a morphism;
- if \( q \) divides \( p \), then \( y \) can be obtained by a substitution;
- otherwise \( y \) can be obtained by iterating periodically \( q \) morphisms.

The word obtained by \( x = ab?ab?ab? \cdots = (ab?)^\omega \) is generated by

\[
\begin{align*}
a &\mapsto aba \\
b &\mapsto abb
\end{align*}
\]
Palindromic closure

The (right) palindromic closure $w\pi$ of a word $w$ is the shortest palindrome word that starts with $w$.

$$(01001)\pi = 010010 \quad (01001010)\pi = 01001010010$$

Given a word $d = a_0a_1 \cdots a_n \cdots$, the word $d\pi$ directed by $d$ is the limit of the sequence $u_0 = \varepsilon$ and

$$u_{n+1} = (u_na_n)\pi$$

For $d = 010101 \cdots$ one gets

\[
\begin{array}{llllllll}
0 & 0 \\
1 & 010 \\
0 & 010010 \\
1 & 01001010010
\end{array}
\]

The limits of binary words are the Sturmian words. A word $d\pi$ is a fixed point of a morphisms if and only if $d$ is periodic (de Luca, Justin & Pirillo).
Repetitions

- A *power* of rational exponent $r$ is a word of the form $u^n u'$

where $u'$ is a prefix of $u$, and

$$
\begin{align*}
  r &= \frac{|u^n u'|}{|u|} = n + \frac{|u'|}{|u|}.
\end{align*}
$$

- For example, *aaabaaabaa* is a power of exponent $5/2$.
- It is convenient to write $u^r$.
Power-free words

Several types of power-free words

- A *square-free* word is a word that contains no factor that is a square.
- Let $r > 1$ be a real number. A word is $r$-free if it contains no factor of the form $u^k$ for $k \geq r$, $k$ rational.
- A word is $k^+$-free if it is $r$-free for all $r > k$ (not necessarily for $k$).
- A word is $k^-$-free if it is $k$-free but not $r$-free for $r < k$.

Examples. Consider the morphisms

\[
\begin{align*}
    a & \mapsto aba & a & \mapsto ab \\
    b & \mapsto abb & b & \mapsto ba
\end{align*}
\]

The word generated by the first morphism is $3^-$-free, the word generated by the second (Thue-Morse) is $2^+$-free (= overlap-free).
Proof

The word generated by iterating the morphism $f : \begin{align*} a &\mapsto aba \\ b &\mapsto abb \end{align*}$ is

$$z = abaabbabaabaabbababaababaabaababbababaabababbbababbabab \cdots$$

The words $aab, f(aab) = aba aba abb$, and in fact all words $f^n(aab)$ are cubes except for their last letter.

Assume that $f(w)$ contains a cube $uuu$.

a) $|u|$ is a multiple of 3.

The initial letter of $u$ appears at positions $i, i + |u|, i + 2|u|$. If $|u| \not\equiv 0 \mod 3$, the initial letter of $u$ appears in $f(a)$ or $f(b)$ at the first, the second and the third position. This is not the case.

b) $f$ is injective, and the preimages are the last letters of the images.

c) One may assume $i \equiv 0 \mod 3$. Thus $w$ also contains a cube.
Morphisms

A simple way to generate infinite power-free words: use a morphism.

Even better: use a power-free morphism: A morphism $f : A^* \rightarrow B^*$ is $k$-free if, for each $k$-free word $w$ over $A^*$, the word $f(w)$ is $k$-free over $B^*$.

1)

\[
\begin{align*}
a & \mapsto ab \alpha \\
b & \mapsto ab \beta \\
\end{align*}
\]

is a cube-free morphism.

2)

\[
\begin{align*}
a & \mapsto ab \gamma \\
b & \mapsto ac \gamma \\
c & \mapsto b \\
\end{align*}
\]

generates an infinite square-free word (used by M. Hall). It is not a square-free morphism ($aba \mapsto ab \gamma ac \gamma ac \gamma ac \gamma ac \gamma ac$).
Morphisms

3) \[
\begin{align*}
a & \mapsto ab \\
b & \mapsto ba
\end{align*}
\]
(Thue-Morse) is an overlap-free morphism.

4) \[
\begin{align*}
a & \mapsto aababb \\
b & \mapsto aabbab \\
c & \mapsto abbaab
\end{align*}
\]
is a cube-free morphism from 3 letters to 2 letters (Bean, Ehrenfeucht, McNulty).
a) Necessary conditions

Let $h : A^* \rightarrow B^*$ be a morphism. If $h$ is square-free, then

- $h(a) \neq h(b)$ for letters $a \neq b$ since otherwise $h(ab)$ is a square.
- $h(a) \neq \varepsilon$ since otherwise $h(bab)$ is a square.
- $h(a)$ is not a prefix (suffix) of $h(b)$ for $a \neq b$ since otherwise $h(ab)$ starts with a square ($h(ba)$ ends with a square).
Looking for square-free morphisms (cont’d)

b) Sufficient conditions

**Theorem** Let $h : A^* \rightarrow B^*$ be a morphism. If

- no word $h(a)$ is a factor of a word $h(b)$ for letters $a \neq b$.
- if $h(a) = xy$, $h(b) = zy$, $(c) = zt$ for letters $a, b, c$ and words $x, y, z, t$, then $a = b$ or $b = c$.

then $h$ is square-free.

5)

\[
\begin{align*}
a &\mapsto abcab \\
b &\mapsto acabcb \\
c &\mapsto acbcacb
\end{align*}
\]

is a square-free morphism (proved by Thue 1912). Its *length* is 18. There is no square-free endomorphism over 3 letters with smaller length (Carpi).
Another example

6)

\begin{align*}
  a &\mapsto abacabc acbabc bacbc \\
  b &\mapsto abacabc acbaca bacbc \\
  c &\mapsto abacabc acbcab cbabc \\
  d &\mapsto abacabc bacab acbabc \\
  d &\mapsto abacabc bacbc acbabc
\end{align*}

is a square-free morphism from 5 letters to 3 letters (Brandenburg).
Characterization of square-free morphisms

Usually, the monoid is not finitely generated.

Let $h : A^* \rightarrow B^*$ be a nonerasing morphism, and set

$$M(h) = \max_{a \in A} |h(a)|, \quad m(h) = \min_{a \in A} |h(a)|.$$

**Theorem.** (Crochemore) The morphism $h$ is square-free if and only if $h$ preserves square-free words of length

$$K(h) = \max(3, 1 + \lceil (M(h) - 3)/m(h) \rceil).$$

For Thue's morphism, check the images of square-free words of length 3.

$$a \mapsto abcab$$
$$b \mapsto acabcb$$
$$c \mapsto acbcacb$$
Theorem. *There are only polynomially many overlap-free binary words.* Contributions by Restivo and Salemi, Kfoury, Kobayashi, Lepistö, Cassaigne.

Theorem. (Brandenburg) *There are exponentially many cube-free binary words.*

A good question, by Kobayashi (1988) : where is the frontier between polynomial and exponential.

Answer, by Karhumäki and Shallit (2003) : the frontier is $7/3$ !

Theorem. *There are polynomially many $7/3$–free words, and there are exponentially many $7/3^+$–free words.*
There is a magic morphism $f$

\[
0 \mapsto 0110\ 1001\ 10010\ 0110\ 1001 \\
1 \mapsto 1001\ 0110\ 01001\ 1001\ 0110
\]

This morphism is $(7/3)^+$–power-free (Kolpakov, Kucherov and Tarannikov, perhaps known earlier). Note that $f(0)$ contains $1001001$, a power of exponent $7/3$. Note that $f(1) = f(0)^\sim$.

Karhumäki and Shallit use the morphism $h$ defined by

\[
0 \mapsto 0110\ 1001\ 10010\ 0110\ 1001 = f(0) \\
1 \mapsto 1001\ 0110\ 01001\ 1001\ 0110 = f(0)^\sim \\
2 \mapsto 1001\ 0110\ 01101\ 1001\ 0110 = f(0) \\
3 \mapsto 0110\ 1001\ 10110\ 0110\ 1001 = f(0)^\sim
\]

from $\{0, 1, 2, 3\}$ to $\{0, 1\}$ to produce many $(7/3)^+$–power-free words.
A set $T \subset A^*$ is a test set for $k$-free morphisms $A^* \to B^*$ if, for any $f : A^* \to B^*$,

$$f \text{ is } k\text{-free if and only if } f \text{ is } k\text{-free on } T.$$  

A test set that is a singleton is a test word.

**Theorem** (Karhumäki, Leconte) Let $f$ be a binary morphism. Then $f$ is cube-free if and only if $f$ is cube-free on cube-free words of length at most 7.

A complicated way to state this result: The set of cube-free binary words of length at most 7 is a test set for cube-free binary morphisms.
The test sets for cube-free binary morphisms are the following.

**Theorem** (Richomme & Wlazinski) A set $T$ of cube-free words is a test set for cube-free binary morphisms if and only if

$$F(T) \supset X \cup X^\sim \cup E(X \cup X^\sim)$$

where $F(T)$ is the set of factors of $T$ and $X = \{ababba, ababba, aabba, ababa\}$.

**Theorem** (Richomme & Wlazinski) Let $f$ be a binary morphism. Then $f$ is cube-free if and only if

$$f(ababbaabaabbaababaababaabbaababaabababab)$$

is a cube-free word.

Thus, $ababbaabaabaabbaababaabbaababaabbaabababab$ is a test word.
**Theorem** (Crochemore) Let $f$ be a ternary morphism. Then $f$ is square-free if and only if $f$ is square-free on square-free words of length 5.

Bounds are known for special families of morphisms (uniform, infix).

**Theorem** (Crochemore) Testing whether a word $w$ of length $n$ is square-free can be done in time linear in $n$. 

Test sets: overlap-free morphisms

**Theorem** (Berstel & Séébold, Richomme & Séébold) Let $f$ be a binary endomorphism. Then the following are equivalent

- $f$ is overlap-free,
- the four words $f(aab)$, $f(aba)$, $f(bab)$, $f(abb)$ are overlap-free,
- the word $f(bbaba)$ is overlap-free.

Richomme & Séébold have characterized *all* test sets for overlap-free binary morphisms.

**Theorem** (Richomme & Wlazinski) A set $T$ of overlap-free words is a test set for overlap morphisms from $A^* \rightarrow B^*$ with $|A| = 2$, $|B| \geq 3$ if and only if $F(T)$ contains the four words $aba, bab, abba, baab$. The word $abbaba$ is a test word.
Open problems

**Problem** Prove or disprove that it is decidable whether a morphism is cube-free

Special cases are known (Leconte, Keränen). Since no algorithm is known, perhaps it is undecidable?
Open problems: repetition threshold

Every binary word of length 4 contains a square, and there exist infinite binary $2^+\text{-free words.}$

Every ternary word of length 39 contains a repetition of exponent $7/4$, and there exists (Dejean) an infinite ternary $(7/4)^+\text{-free word.}$

The repetition-threshold is the smallest number $s(k)$ such that there exists and infinite word over $k$ letters that has only repetitions of exponent less than or equal to $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\cdots$</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s(k)$</td>
<td>2</td>
<td>$7/4$</td>
<td>$7/5$</td>
<td>$5/4$</td>
<td>$\cdots$</td>
<td>$11/10$</td>
</tr>
</tbody>
</table>

**Problem** Is it true that the repetition threshold is always $k/(k − 1)$?
Open problems : avoidable pattern

- $E$ is a pattern alphabet, $A$ is a target alphabet.
- $\mathcal{M}(E, A)$ is the set of morphisms from $E^+$ to $A^+$.
- For $p$ over $E$, the pattern language of $p$ over $A$ is the set $H(p) = \{h(p) \mid h \in \mathcal{M}(E, A)\}$.
- A word $w$ over $A$ avoids $p$ if no factor of $w$ is in $H(p)$.

Examples

- A square-free word is a word that avoids the pattern $\alpha\alpha$.
- No word over $n$ letters of length $n + 1$ avoids the pattern $\alpha\beta\alpha$.

A pattern $p$ is $k$-avoidable if there exists an infinite word over $k$ letters that avoids $p$.

**Problem** Is there a pattern that is 4- unavoidable and 5-avoidable?
Literature