Combinatorics on Words

Joel Reyes Noche
jrnoche@adnu.edu.ph
Department of Mathematics and Natural Sciences
College of Arts and Sciences
Ateneo de Naga University

February 6, 2010
Abstract

Combinatorics on words is a mathematical discipline related to computer science. In this talk, I present through examples some terms, notation, and concepts used in the literature. Words such as the Thue-Morse, the Fibonacci, the Kolakoski, the Barbier, and the Champernowne words possess special properties, attract the attention of researchers, and find use in some applications.
Overview

- Mathematical disciplines
- Examples of words
- Alphabets, letters, and words
- Sets of words
- Factors
- Periodicity
- Subword complexity and Sturmian words
- Morphisms
- Thue-Morse word, Fibonacci word, Kolakoski word, Barbier word, Champernowne word
- Some applications
- Some problems and answers
- Some open problems
The Mathematics Subject Classification is used to categorize items covered by two reviewing databases, *Mathematical Reviews* and *Zentralblatt MATH*. The current system is MSC2000.

<table>
<thead>
<tr>
<th>00 General</th>
<th>18 Category theory, homological algebra</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 History &amp; biography</td>
<td>19 $K$-theory</td>
</tr>
<tr>
<td>03 Mathematical logic &amp; foundations</td>
<td>20 Group theory &amp; generalizations</td>
</tr>
<tr>
<td>04 Set theory [deleted, see 03]</td>
<td>22 Topological groups, Lie groups</td>
</tr>
<tr>
<td>05 Combinatorics</td>
<td>26 Real functions</td>
</tr>
<tr>
<td>06 Order, lattices, ordered algebraic structures</td>
<td>28 Measure &amp; integration</td>
</tr>
<tr>
<td>08 General algebraic systems</td>
<td>30 Functions of a complex variable</td>
</tr>
<tr>
<td>11 Number theory</td>
<td>31 Potential theory</td>
</tr>
<tr>
<td>12 Field theory &amp; polynomials</td>
<td>32 Several complex variables &amp; analytic spaces</td>
</tr>
<tr>
<td>13 Commutative rings &amp; algebras</td>
<td>33 Special functions</td>
</tr>
<tr>
<td>14 Algebraic geometry</td>
<td>34 Ordinary differential equations</td>
</tr>
<tr>
<td>15 Linear &amp; multilinear algebra; matrix theory</td>
<td>35 Partial differential equations</td>
</tr>
<tr>
<td>16 Associative rings &amp; algebras</td>
<td>37 Dynamical systems &amp; ergodic theory</td>
</tr>
<tr>
<td>17 Nonassociative rings &amp; algebras</td>
<td>(more on next slide)</td>
</tr>
</tbody>
</table>
39 Finite differences & functional equations
40 Sequences, series, summability
41 Approximations & expansions
42 Fourier analysis
43 Abstract harmonic analysis
44 Integral transforms, operational calculus
45 Integral equations
46 Functional analysis
47 Operator theory
49 Calculus of variations & optimal control
51 Geometry
52 Convex & discrete geometry
53 Differential geometry
54 General topology
55 Algebraic topology
57 Manifolds & cell complexes
58 Global analysis, analysis on manifolds
60 Probability theory & stochastic processes
62 Statistics
65 Numerical analysis
68 Computer science
70 Mechanics of particles and systems
73 Mechanics of solids
[deleted, see 74]
74 Mechanics of deformable solids
76 Fluid mechanics
78 Optics, electromagnetic theory
80 Classical thermodynamics, heat transfer
81 Quantum Theory
82 Statistical mechanics, structure of matter
83 Relativity and gravitational theory
85 Astronomy and astrophysics
86 Geophysics
90 Operations research, mathematical programming
91 Game theory, economics, social and behavioral sciences
92 Biology and other natural sciences
93 Systems theory; control
94 Information and communication, circuits
97 Mathematics education
68-xx Computer science

68-00 General reference works
68-01 Instructional exposition
68-02 Research exposition
68-03 Historical
68-04 Explicit machine computation and programs
68-06 Proceedings, conferences, collections, etc.
68Mxx Computer system organization
68Nxx Software
68Pxx Theory of data
68Qxx Theory of computing
68Rxx Discrete mathematics in relation to computer science
68Txx Artificial intelligence
68Uxx Computing methodologies and applications
68Wxx Algorithms

68Rxx Discrete mathematics in relation to computer science

68R01 General
68R05 Combinatorics
68R10 Graph theory
68R15 Combinatorics on words
68R99 None of the above, but in this section
Examples of words  (Sloane, 2008)

*Thue-Morse word:*  (A010060)
\[ t = 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ 1 \ ... \]

*Fibonacci word:*  (A003849)
\[ f = 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ ... \]

*Kolakoski word:*  (A000002)
\[ k = 1 \ 2 \ 2 \ 1 \ 1 \ 2 \ 1 \ 2 \ 2 \ 1 \ 2 \ 2 \ 1 \ 1 \ 2 \ 1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 2 \ 2 \ 1 \ 1 \ ... \]

*Barbier word:*  (A007376)
\[ b = 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 1 \ 0 \ 1 \ 1 \ 1 \ 2 \ 1 \ 3 \ 1 \ 4 \ 1 \ 5 \ 1 \ 6 \ 1 \ 7 \ 1 \ 8 \ 1 \ 9 \ ... \]

*Champernowne word:*  (A030190)
\[ c = 0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ ... \]
An *alphabet* $A$ is a non-empty set with elements called *letters*.

**Example:** $T = \{0, 1\}$, $F = \{0, 1\}$, $K = \{1, 2\}$,

$B = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $C = \{0, 1\}$

Sequences of letters are called *words*.

The *empty word* $\epsilon$ has a length of zero.

A *binary word* has an alphabet of two letters.

A *ternary word* has an alphabet of three letters.

A *finite word* has a finite number of letters.

**Example:** $s = 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 1$ (Stein, p. 144)

An *infinite word* is a map from $\mathbb{N} = \{1, 2, 3, ...\}$ to the alphabet.

**Example:** all the examples in the previous slide

A *doubly-infinite word* is a map from $\mathbb{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}$ to the alphabet.

**Example:** $r = \ldots 2\ 1\ 1\ 2\ 1\ 2\ $ (Robbins, p. 413)

We limit our discussion to finite and infinite words with finite alphabets.
Sets of words  (Rampersad, 2004)

The set of all finite words over $A$ is denoted by $A^*$.  
Example: $T^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$

The set of all finite, non-empty words over $A$ is denoted by $A^+$.  
Example: $T^+ = T^* - \{\epsilon\}$

The set of all infinite words over $A$ is denoted by $A^\omega$.  
Example: $T^* \not\subseteq T^\omega$

The set of all words over $A$ is denoted by $A^\infty$.  
Example: $T^\infty = T^* \cup T^\omega$

The length of $w \in A^*$ is denoted by $|w|$.  
Example: $|s| = 19$, $|\epsilon| = 0$

The number of occurrences of $a \in A$ in $w \in A^*$ is denoted by $|w|_a$.  
Example: $|s|_0 = 8$, $|s|_1 = 11$
Factors  (Lothaire, 1997; Rampersad, 2004; Karhumäki, 2004)

Concatenation of words is associative but not commutative.
Example: \( x = 0, y = 1\ 0, xy = 0\ 1\ 0, yx = 1\ 0\ 0, \epsilon x = x\epsilon = x \)

A word \( w' \) is called a factor of \( w \in A^\infty \) if \( w = uw'v \) for some \( u \in A^* \) and \( v \in A^\infty \).
If \( u = \epsilon \), then \( w' \) is called a left factor of \( w \).
If \( v = \epsilon \), then \( w' \) is called a right factor of \( w \).
Example: \( 0\ 1\ 0 \) is a left factor of \( f \), but not of \( t \).
\( 1\ 1\ 1 \) is a factor, a left factor, and a right factor of \( s \).

A square is a word of the form \( xx \) where \( x \in A^+ \).
A cube is a word of the form \( xxx \) where \( x \in A^+ \).
A word is square-free (it avoids squares) if it has no square factors.
A word is cube-free if it has no cube factors.
Example: \( 0\ 0 \) and \( 0\ 1\ 0\ 0\ 1\ 0 \) are squares.
\( 0\ 0\ 0 \) and \( 1\ 1\ 1 \) are the only cubes that \( s \) has.
More examples: All square-free words are cube-free. Not all cube-free words are square-free.

The set of all binary words of length 4 is

\{ 0 0 0 0 , 0 0 0 1 , 0 0 1 0 , 0 0 1 1 ,
    0 1 0 0 , 0 1 0 1 , 0 1 1 0 , 0 1 1 1 ,
    1 0 0 0 , 1 0 0 1 , 1 0 1 0 , 1 0 1 1 ,
    1 1 0 0 , 1 1 0 1 , 1 1 1 0 , 1 1 1 1 \}.

s is the shortest word that has all the words above as factors. All of the above words have squares. There are no square-free binary words of length at least 4.

Is there an infinite square-free ternary word?
Is there an infinite cube-free binary word?
Is there an infinite binary word with no squares \( xx \) with \( |x| \geq 3 \)?
Periodicity  (Berstel & Karhumäki, 2003)

A (finite) natural number $p$ is a period of $w = a_1 a_2 ... a_n$ if $p \leq n$ and $a_i = a_{p+i}$ for $i = 1, ..., n - p$.

A finite non-empty word of length $n$ has at least one period: $n$.

Example: 0 has a period of 1.

1 1 1 has periods of 1, 2, and 3.

0 1 0 0 1 0 1 0 0 1 0 has periods of 5, 8, and 11.

$s$ has a period of 19.

A word is periodic if it has a period.

An infinite word is ultimately periodic if and only if it has a periodic right factor.

An infinite word is nonperiodic if it is not ultimately periodic.

Example: All finite words are periodic and not nonperiodic.

$p = 0 0 0 ...$ (the other letters are 0's) is periodic & ultimately periodic.

$u = 1 0 0 ...$ (the other letters are 0's) is neither periodic nor nonperiodic but is ultimately periodic.
Subword complexity and Sturmian words
(Berstel, 1996; Berstel & Karhumäki, 2003)

The *subword complexity* of an infinite word $w$ is the function $P_w$ where $P_w(n)$ is the number of factors of length $n$ of $w$.

**Example:**
- $P_p(0) = 1$, $P_p(1) = 1$, $P_p(2) = 1$, $P_p(3) = 1$, ...
- $P_u(0) = 1$, $P_u(1) = 2$, $P_u(2) = 2$, $P_u(3) = 2$, ...
- $P_t(0) = 1$, $P_t(1) = 2$, $P_t(2) = 4$, $P_t(3) = 6$, ...
- $P_f(0) = 1$, $P_f(1) = 2$, $P_f(2) = 3$, $P_f(3) = 4$, ...

If $P_w(n) \leq n$ for some $n \geq 0$, then $w$ is ultimately periodic.
If $P_w(n) \geq n + 1$ for all $n \geq 0$, then $w$ is nonperiodic.
If $P_w(n) = n + 1$ for all $n \geq 0$, then $w$ is a *Sturmian* word.

A Sturmian word is a nonperiodic word of minimal complexity.

**Example:**
- $P_p(n) \leq n$ for $n \geq 1$
- $P_u(n) \leq n$ for $n \geq 2$
- $P_t(n) \geq n + 1$ for all $n \geq 0$
- $P_f(n) = n + 1$ for all $n \geq 0$
Morphisms  (Rampersad, 2004)

A map $m : A^* \rightarrow B^*$ is called a morphism if $m$ satisfies
\[ m(xy) = m(x)m(y) \]
for all $x, y \in A^*$.

A morphism may be specified by providing the image words $m(a)$
for all $a \in A$.

Example: Let $A = \{0, 1\}$ and $m_1 : A^* \rightarrow A^*$ be defined by
\[ m_1(0) = 1, \quad m_1(1) = 10. \]

Thus, $m_1(m_1(1)) = m_1(10) = m_1(1)m_1(0) = 101$

We denote $m(m(a))$ by $m^2(a)$, $m(m(m(a)))$ by $m^3(a)$ and so on.

A morphism $m : A^* \rightarrow A^*$ such that $m(a) = ax$ for some $a \in A$
and $x \in A^*$ is said to be prolongable on $a$.

Example: $m_1$ is prolongable on 1 but not on 0.
Morphisms (continued)  (Rampersad, 2004)

If $m$ is prolongable on $a$, then repeatedly iterating $m$ on $a$ yields the fixed point $m^\omega(a)$.

$m(a) = ax$
$m^2(a) = m(ax) = axm(x)$
$m^3(a) = m(axm(x)) = axm(x)m^2(x)$
\vdots
$m^\omega(a) = axm(x)m^2(x)m^3(x)m^4(x)\cdots$

Example: $m^\omega_1(1) = 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0\ \cdots$
Thue-Morse word
(Allouche & Shallit, 1999)

t = 0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 0 1 1 0 1 ...

It is generated by \( t : T^* \to T^* \) with \( t(0) = 0 1 \), \( t(1) = 1 0 \), then taking \( t^\omega(0) \).

It is an infinite cube-free (and thus nonperiodic) binary word.

Comment: Let \( v_n \) be the number of 1’s between the \( n \)th and the \( (n + 1) \)st occurrence of 0 in \( t \). Let \( v = v_1 \ v_2 \ v_3 \ v_4 \ldots \)
\( v = 2 \ 1 \ 0 \ 2 \ 0 \ 1 \ 2 \ 1 \ 0 \ 1 \ 2 \ 0 \ 2 \ldots \) is an infinite square-free ternary word.
**Fibonacci word**  
(Berstel & Karhumäki, 2003)

\[
f = 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 0 ...\]

It is generated by \( f : F^* \rightarrow F^* \) with \( f(0) = 0 1, \ f(1) = 0 \), then taking \( f^\omega(0) \).

It is an infinite Sturmian (and thus nonperiodic) binary word.

**Comment:** \( f^{n+1}(0) = f^n(0)f^{n-1}(0) \) for \( n \geq 2 \)

**Illustration:** For \( n = 2 \), \( f(0) = 0 1 \), \( f^2(0) = 0 1 0 \), \( f^3(0) = 0 1 0 0 1 \)
Kolakoski word
(Kolakoski, 1965; Kolakoski & Ucoluk, 1966; Berstel & Karhumäki, 2003; Sloane, 2008)

\[ k = 1\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 2\ 1\ 1\ \ldots \]

It is generated by \( k_1 : K^* \to K^* \) with \( k_1(1) = 1, \ k_1(2) = 11 \)
and \( k_2 : K^* \to K^* \) with \( k_2(1) = 2, \ k_2(2) = 22, \)
then taking \( k_1(\kappa_1)k_2(\kappa_2)k_1(\kappa_3)k_2(\kappa_4)k_1(\kappa_5)\ldots \)
where \( \kappa_1 = 1, \ \kappa_2 = 2, \) and \( k = \kappa_1\kappa_2\kappa_3\kappa_4\kappa_5\ldots \)

Illustration: \( k = \kappa_1\kappa_2\ldots = 1\ 2\ldots \)
\( k = k_1(\kappa_1)k_2(\kappa_2)\ldots = 1\ 2\ 2\ldots = \kappa_1\kappa_2\kappa_3\ldots \)
\( k = k_1(\kappa_1)k_2(\kappa_2)k_1(\kappa_3)\ldots = 1\ 2\ 2\ 1\ 1\ldots = \kappa_1\kappa_2\kappa_3\kappa_4\kappa_5\ldots \)

It is an infinite cube-free (and thus nonperiodic) binary word.

Comment: \( k \) cannot be obtained as a fixed point of an iterated morphism.
Barbier word

\[ b = 1 2 3 4 5 6 7 8 9 1 0 1 1 1 2 1 3 1 4 1 5 1 6 1 7 1 8 1 9 \ldots \]

It is an infinite nonperiodic decimal word of maximal complexity. \( P_b(n) = 10^n \) for all \( n \geq 0 \). Every decimal word is a factor of \( b \).

Champernowne word
(Berstel & Karhumäki, 2003)

\[ c = 0 1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0 1 0 0 1 1 0 1 \ldots \]

It is an infinite nonperiodic binary word of maximal complexity. \( P_c(n) = 2^n \) for all \( n \geq 0 \). Every binary word is a factor of \( c \).
Some applications
(Allouche & Shallit, 1999; Karhumäki, 2004)

Number theory, probability theory, algebra (free groups, matrices, representations, Burnside problems), real analysis, differential geometry, computer science (algorithmics), logic, automata theory and computability, discrete dynamical systems, physics (quasi-crystals), biology (DNA sequencing)
The German rule in chess states that a draw occurs if the same sequence of moves occurs three times in succession. Using the cube-free property of the Thue-Morse sequence, under such a rule infinite games of chess are possible, and no draw occurs.
Some problems and answers
(Entringer, Jackson, & Schatz, as cited in Rampersad, 2004)

Is there an infinite square-free ternary word?
Yes, for example, \( v \).

Is there an infinite cube-free binary word?
Yes, for example, \( t \) and \( k \).

Is there an infinite binary word with no squares \( xx \) with \( |x| \geq 3 \)?
Yes. Choose an infinite square-free ternary word, say \( v \), with \( V = \{0, 1, 2\} \) and \( T = \{0, 1\} \). Let \( g : V^* \to T^* \) be defined by
\[
\begin{align*}
g(0) &= 1 0 1 0, \\
g(1) &= 1 1 0 0, \\
g(2) &= 0 1 1 1.
\end{align*}
\]
Then \( g(v) \) contains no squares \( xx \) with \( |x| \geq 3 \).
Some open problems
(Berstel & Karhumäki, 2003; Rampersad, 2004)

Is the number of occurrences of 1’s and 2’s in the Kolakoski word asymptotically equal?

An Abelian square is a nonempty word $xy$, where $y$ is a permutation of the letters of $x$.

Is there an infinite ternary word containing no Abelian squares $xy$, where $|x| = |y| \geq 2$?

Is there an infinite binary word containing no Abelian cubes $xyz$, where $|x| = |y| = |z| \geq 2$?
References


