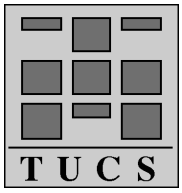
Combinatorics on Infinite Words[1]

J. Karhumäki
A. Lepistö

Department of Mathematics &
Turku Centre for Computer Science
University of Turku
20014 University of Turku, Finland
Email: karhumak@cs.utu.fi, alepisto@utu.fi



Turku Centre for Computer Science
TUCS Technical Report No 578
December 2003

ISBN 952-12-1282-9
ISSN 1239-1891

Abstract

We consider several problems of infinite words over a finite alphabet. In particular, we describe a few automata-theoretic methods to define infinite words. Properties of infinite words studied in more details are repetition-freeness, periodicity and different kinds of complexity issues. Examples are used to illustrate the power of infinite words in many applications, as well as illustrations of problems of different areas of mathematics.

A lecture in Ph.D. Program in Terragona.

[1]Supported by the Academy of Finland under the grant 44087

1 Introduction

Combinatorics on words is a relatively new area of discrete mathematics. Although it has connections to numerous branches of mathematics, and indeed much of the early theory is developed implicitly as tools to attack some quite different problems, its basic motivation and drive comes from computer science. This is seen, for instance, in the recent classification of the field in Mathematical Reviews. As illustrated below combinatorics on words is under the main section of Computer Science, but with the very strong mathematical emphasis.

The notion of a word, e.g. a sequence of symbols, is extremely natural in mathematics. Indeed, the representation of a natural number at any base is a word over a finite alphabet. And for a computer any algorithm on numbers is an algorithm on words!

Consequently, it is no surprise that the words has occurred - often implicitly - in mathematical considerations during several centuries. As an illustration we mention that already Gauss, see [Ga00], came to a problem on words, and that Prouhet, see [Pr51], discovered the infinite Thue-Morse word.

A systematic research on words, in fact on infinite words, was initiated by a Norwegian A. Thue almost hundred years ago. In 1906, see [Th06], he published his first results on repetition-free words, including a construction of an infinite binary word not containing any cubes, i.e. factors of the form u^3. Thue's results were for many decades unnoticed, and many of his results were subsequently rediscovered. Interestingly, Thue seemed to have no motivation - beyond scientific curiosity - for his research. Later very often results on words were obtained as byproducts when looking for tools to attack some other, often very unrelated-looking, problems. Papers to be mentioned after Thue are for example [MH38] and [Ar37].

As a theory Combinatorics on Words started in 1950's in two places simultaneously and independently. In Russia P.S. Novikov And S. Adian gathered a lot of knowledge when searching for a solution of Burnside Problem for groups, see [Ad79]. In their considerations much of the theory of words was implicit - as it has been throughout the history of the well developed combinatorial group theory, see [MKS66] and [LS77]. In France research on words was initiated by M.P. Schützenberger in connection with the theory of codes, see [Sc56].

In coming decades research on words extended rapidly, and geographically. Many remarkable results we revealed, such as the decidability of the satisfiability problem for word equations, see [Ma77] and [Pl99], and the compactness

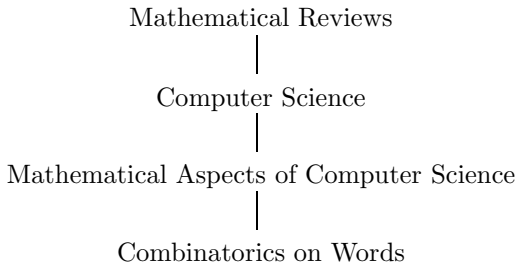


Figure 1: Position of Combinatorics on Words in Math. Rev.

property of word equations, see [AL85], [Gu86] or [CK97], just to mention only two jewels.

In 1983 the research, especially the part connected to France, culminated into the first monograph of the field, Combinatorics on Words, see [Lo83]. The book had a very inspiring effect: Since that the area has been growing steadily and very fast. A new monograph, Algebraic Combinatorics on Words, see [Lo02], appeared recently, and a biannual conference WORDS has been established. Papers [CK97] and [BK03] are two other recent survey articles.

In this presentation we consider a particular part of the theory of Combinatorics on Words, namely that of infinite words. Of course, the theory of infinite words is essentially more complicated than that of finite words. For example, the cardinality of the set of all infinite words (over a nonunary alphabet) is nondenumerable, while that of all finite words is only denumerable. However, many considerations are much neater for infinite words rather than for finite words, and moreover infinite words are often needed to describe certain phenomena.

Without trying to be exhaustive we concentrate on a few research topics of infinite words. We consider how infinite words can be defined, especially in terms of automata theory. Then we consider *avoidability theory of patterns*, i.e. whether there exist infinite words which do not contain a certain substructure, like a square, as a factor. Next certain *periodicity* aspects of infinite words are discussed, and then finally the notion of the *complexity* is considered from different perspectives. Examples play a crucial role in our presentation, while the complete proofs are omitted. These can be found from survey articles [Lo02], [CK97] and [BK03], or from references mentioned therein.

2 Preliminaries

This section is mainly to fix the terminology. For further background on words we refer to [Lo02] and [CK97]. For the results and more complete definitions on automata we refer to [Sa73], [Be79] and [HU79].

Let A be a finite set referred to as an *alphabet*. A *word* over A is any sequence, finite or infinite, of elements of A. Consequently, a one-way infinite word w can be depicted as

$$w = a_0a_1a_2\ldots \qquad \text{with } a_i \in A$$

and can be formally defined as the mapping $\mathbb{N} \to A$. The *length* of a word w is the total number of its symbols and is denoted by $|w|$. A word u is a *factor* (resp. a *prefix* or *suffix*) of a word w if there exist words x and y such that $w = xuy$ (resp. $w = uy$ or $w = xu$). Recalling that a product of finite and infinite word (but not that of infinite word followed by a finite word) is well defined, we may allow w and/or u to be finite or infinite. Consequently, in the cases u is infinite y must be omitted. If u is a prefix of w we write $u \leq w$. The prefix (resp. suffix) of length k of a word w is denoted by $\mathrm{pref}_k w$ (resp. $\mathrm{suf}_k w$). Let A^* (resp. A^ω) denote the set of all finite (resp. infinite) words over A. Hence A^* is a monoid under the operation of *catenation* (or *product*) of words: $u \cdot v = uv$. Moreover, it is *free*, i.e. each word $w \in A^*$ has the unique representation as the product of letters, i.e. elements of A. Note that A^* contains also the sequence of zero symbols, so-called *empty* word, denoted by 1.

A mapping $h : A^* \to B^*$ is called a *morphism* if

$$h(uv) = h(u)h(v) \qquad \text{for all } u, v \in A^*.$$

It follows that any morphism satisfies that $h(1) = 1$ and that it is completely defined when the values $h(a)$ for $a \in A$ are given. We use the notation

$$h \ : \ a \mapsto \alpha$$

to denote that h maps a to α, i.e. $h(a) = \alpha$. Any morphism $h : A^* \to B^*$ extends to a mapping $h : A^\omega \to A^\omega$ satisfying

$$h(uv) = h(u)h(v) \qquad \text{for all } u \in A^*, v \in A^\omega.$$

We call a morphism

nonerasing	if $h(a) \neq 1$ for all $a \in A$,
uniform	if $\lvert h(a)\rvert = \lvert h(b)\rvert$ for all $a, b \in A$,
coding	if $\lvert h(a)\rvert = 1$ for all $a \in A$,
binary	if $\mathrm{card}(A) = 2$,
marked	if $\mathrm{pref}_1(h(a)) \neq \mathrm{pref}_1(h(b))$ for all $a \neq b$.

Other used notions are defined when needed.

3 Examples

In this section we give a few examples of infinite words. The words we have chosen are used to illustrate different aspects of the theory and applications, as well as problems.

Example 1. Consider the morphism $\mu : \{a, b\}^* \to \{a, b\}^*$ defined by

$$\mu \ : \ \begin{array}{ccc} a & \mapsto & ab \\ b & \mapsto & ba \end{array}.$$

The morphism is among the simplest ones; it is, for instance, binary, uniform and even marked. It is one of the morphisms Thue used in his considerations. Define

$$t_0 = a \text{ and } t_{i+1} = \mu(t_i) \quad \text{for } i \geq 0.$$

Then, clearly, $t_1 = t_0\alpha$, for some α, and consequently t_i is a prefix of t_{i+1} for all $i \geq 0$. Therefore we can write

$$t = \lim_{i\to\infty} t_i = \lim_{i\to\infty} \mu^i(a) = abbabaabbaababba \dots \ .$$

The infinite word t is referred to as *Thue-Morse word*. We can say that t is defined by *iterating a morphism*.

Example 2. As in Example 1 we can use the morphism

$$\varphi \ : \ \begin{array}{ccc} a & \mapsto & ab \\ b & \mapsto & a \end{array}.$$

to define the infinite word

$$f = \lim_{i\to\infty} \varphi^i(a) = abaababaabaab\ldots\ .$$

This word is called the *Fibonacci word*. It is a counterpart of Fibonacci numbers in the theory of words. Indeed, set

$$f_0 = a, \quad f_1 = ab, \quad f_{i+1} = f_i f_{i-1} \quad \text{for } i \geq 1.$$

Then the words f_i satisfy a similar recursion as the famous Fibonacci numbers, and in fact lengths of f_i's define exactly the sequence of Fibonacci numbers. Moreover

$$f = \lim_{i\to\infty} f_i.$$

The Fibonacci word possesses a lot of remarkable properties, as we shall see. It would not be surprise if the future would show it even more important than the sequence of Fibonacci numbers.

Example 3. We define an infinite work k by the rules:

(i) It constitutes of consecutive blocks of symbols 1 and 2.

(ii) The first block is 22.

(iii) The length of the ith block is equal to the value of ith digit of the whole word.

Consequently, the word k, referred to as *Kolakoski word*, starts as follows:

$$k = 2211212212211\ldots\ .$$

Amazingly little is known about this simply defined word. For example, it is only a conjecture, but not a fact, that the frequency of 1's is asymptotically 1/2. It is a n exersice to prove that it is not ultimately periodic.

Above Kolakoski word was defined via a precise *selfreading* rule. There are also at least two other ways how it can be defined. First, consider the morphisms

$$h_o : \begin{array}{ccc} 1 & \mapsto & 2 \\ 2 & \mapsto & 22 \end{array} \qquad h_e : \begin{array}{ccc} 1 & \mapsto & 1 \\ 2 & \mapsto & 11 \end{array},$$

and let $h : \{1,2\}^* \to \{1,2\}^*$ be a mapping satisfying

$$a_1a_2a_3\ldots a_{2t}a_{2t+1} \mapsto h_o(a_1)h_e(a_2)h_o(a_3)\ldots h_e(a_{2t})h_o(a_{2t+1}),$$

i.e. every letter at odd position is mapped by h_o and every letter at even position is mapped by h_e. Then, as in the previous examples h, when iterated, defines the unique infinite word, indeed

$$k = \lim_{i\to\infty} h^i(2).$$

We can say that k is defined by *iterating two morphisms periodically.*

In the particular case of Kolakoski word it can be obtained also as a result of *iterating a deterministic gsm.* Consider the dgsm shown in Fig. 2. Now, the machine M maps 2 into 22, that into 2211, that into 221121 and so on. More generally

$$k = \lim_{i\to\infty} \mathcal{M}^i(2),$$

i.e. k is obtained by iterating a dgsm.

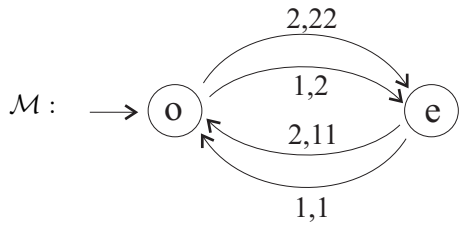


Figure 2: Kolakoski machine M.

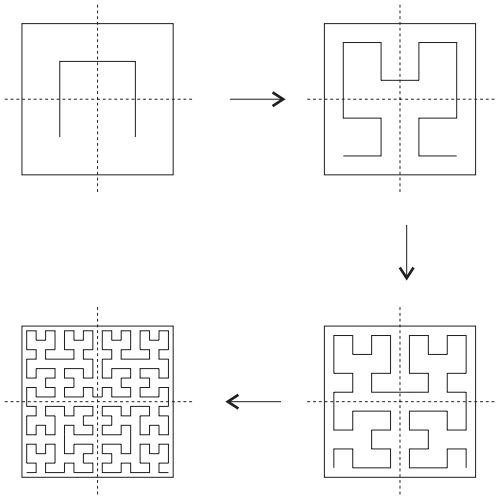


Figure 3: Four first steps in Hilbert's construction.

Example 4. We show how so-called *Hilbert's space filling curve* can be defined easily as an infinite word. We start by illustrating the construction as shown in Fig. 3. Now, the curve is obtained as the limit of the above procedure. As is seen from Fig. 3, step i as a word of the four letter alphabet $\{u(p), d(own), r(ight), l(eft)\}$ is a prefix of the word in step $i+2$. Consequently, the sequence of the words described in every second step defines the unique infinite word, a representation of Hilbert's curve.

Interestingly, this word is easy to describe. It can be obtained by iterating a uniform morphism, and then mapping the result by a coding. We leave it as an exercise to construct the actual morphism, cf. [Sh89] or [KM03].

What is interesting above is that the phenomenon, which is considered as an anomaly in topology, can be defined in a very simple way in tools of combinatorics on words.

4 How to define infinite words

In this section we describe some algorithmic methods to define infinite words. However, as we already noted, the cardinality of all infinite words is nondenu-

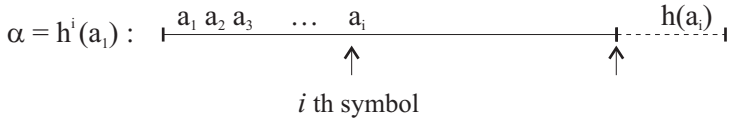


Figure 4: Iteration of a morphism as D0L TAG system

merable, so that all algorithmic methods can capture only very very few of all words.

Iterating morphisms. The most frequently used method is that of iterating a morphism. We already used this method in Examples 1 and 2.

To recall the method let $h : A^* \to A^*$ be a nonerasing morphism satisfying

$$h(a) = a\alpha \qquad \text{for some } \alpha \in A^+.$$

Then $a \leq h(a)$, and consequently $h^i(a) \leq h^{i+1}(a)$ for any $i \geq 0$. Therefore we can write

$$w = \lim_{i \to \infty} h^i(a),$$

i.e. the limit exists. It follows that $h(w) = w$, meaning that w is a *fixed-point* of h. We say that w is obtained as the fixed-point of iterated morphism h at point a.

The above method can be seen as a self-reading procedure described in Example 3. This is illustrated in Fig. 4. There we have one tape and two heads on it. The leftmost head reads a symbol and the rightmost head adds $h(a)$ to the end of the word. At the same time the left head moves to the next symbol and the right head to the end of the word. It is obvious that this mechanism realizes exactly the iteration of a morphism.

The model resembles TAG systems of Post, see [Mi67]. Therefore, and more to name its generalizations property, we refer this mechanism also to *D0L TAG system*, see [CK94], and call such words *D0L words*.

A step further is to introduce a coding (or another morphism) which is used to map a D0L word into a new word. Such words are called *CD0L words*.

Periodic iteration of morphisms. As we explained in Example 3 iteration of morphism can be extended in a natural way to the method of periodic iteration of morphisms. This method fits very well to the illustration of Fig. 4. In this method we have p morphisms $h_1, \ldots, h_p$, and in each rewriting step the leftmost head scans a symbol a_i and remembers $i(\mathrm{mod}p)$ in order to write $h_{i(\mathrm{mod}p)}(a_i)$ to the right end of the word.

As we said the Kolakoski word can be obtained by iterating two morphisms periodically. It is not known whether it is a CD0L word, but can be shown that it is not a D0L word. However, it is known that words obtained by periodic iteration of morphisms need not be CD0L words, see [Le93]. The other inclusion is open, see [CK94].

Iterating dgsm's. A natural extension of the above method is the iteration of a deterministic generalized sequential machine, dgsm for short. This was already illustrated in Example 3. From the above we get this model when the left head, instead of remembering $i(\mathrm{mod}p)$, remembers the next state of the

Figure 5: DGSM TAG system.

machine M. Consequently, the illustration of the model is as in Fig. 5. Here, as earlier, the current positions of heads are denoted by arrows, and the new positions by spotted arrows. Moreover, the current and new states are shown, as well as rewritten word $\delta(a_i, p)$. Such a model is called *DGSM TAG system*, and words obtained are called *DGSM TAG words*.

Example 5. Let τ be a deterministic Turing machine and

$$w_0, w_1, w_2, \ldots$$

a sequence of its configurations. It is easy to find a dgsm which computes

$$w_i \mapsto w_{i+1}.$$

Consequently, the word

$$Sw_0\#w_1\#w_2\ldots$$

is a DGSM TAG word. □

The above example explains why many problems on DGSM TAG systems are undecidable. Moreover, [DM03] translates a classical problem of the complexity theory to a problem on DGSM TAG words.

Iterating morphisms on two tapes. In all previous models we had only one tape in use. Or to be precise the second tape remembering either $i(\mathrm{mod}p)$ or the current state of a dgsm was very restricted. In our last model we have two unrestricted tapes, the first one to generate the word and the second one to generate the control tape. The model is depicted in Fig. 6. Rewriting rules are of the form

$$\begin{pmatrix} a \\ b \end{pmatrix} \to \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

That is, the symbols read by the left head determine the continuations of the both tapes. Further the left head move one step to the right on both tapes and the right one to the ends of the both tapes. In this method we have two interrelated D0L TAG systems. Hence we call it *double D0L TAG system*.

As pointed out in [CK94] there are several other variants of similar models generating infinite words. More examples on words obtained by these models can be found in [CK94]. Further in [DM03] concrete words which are not double D0L TAG words are shown.

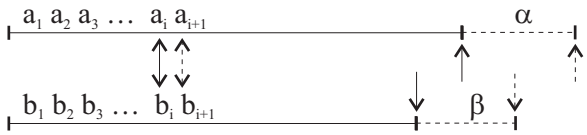


Figure 6: Double D0L TAG system

5 Unavoidability

One of the most studied combinatorial properties of words is the unavoidability. Indeed, major results of Thue show that there exist an infinite binary (resp. ternary) word, not containing any cube (resp. square) as a factor.

We need some terminology. A *power of order* k is a word of the form u^k. Here $k > 1$ can be also rational, say $k = p/q$. Then when u^k means the prefix of u^ω of length $|u|\frac{p}{q}$. For example *abaaabaaaba* is of order $2\ ^3/_4$ or $(1\ ^{10}/_{11})$. Further w is *k-free* if it does not contain any factors of order k or larger. Note that "or larger" is essential here. Similarly, w is k^+-free if it does not contain any factors of order strictly larger than k. For example *abbabaab* is 2^+-free but not 2-free.

The k-freeness is a special case of avoidability introduced in [BEM79]. Let Ξ be a finite set of unknowns disjoint with A. A pattern over Ξ is a finite word in Ξ^+. We say that a word w *avoids* a pattern p if none of the images $h(p)$, where $h : \Xi^+ \to A^+$ is a nonerasing morphism, is a factor of w. Clearly the pattern xx corresponds to the square-freeness.

We proceed with a simple fact.

Lemma 1. *Let p be a pattern. It is avoided by an infinite language if and only if it is avoided by an infinite word.*

Proof. The implication to the left is clear. If an infinite word w avoids p, so do all of its prefixes which constitutes an infinite language.

For the reverse implication let L be an infinite language, so that all of its words avoid p. Now choose a letter $a_1 \in A$ such that L contains infinitely many words starting with a_1, and define

$$L_i = \{w \in L \mid \mathrm{pref}_1(w) = a_1\}.$$

Since A is finite the required a exists. Set $\alpha_1 = a_1$ and repeat the procedure for $a_1^{-1}L_1$. This yields a letter a_2 and we set $\alpha_2 = \alpha_1 a_2$. Repeating the process we define infinite word $\alpha = a_1 a_2 \ldots$. It avoids p since any of its prefixes is a prefix of a word in L, and hence avoids p. □

The above lemma, based on the finiteness of A, holds for any property P which is preserved under taking factors of words and under taking limits of finite words (in an ordinary topological sense).

Above also motivates to study avoidability in infinite words. We start with Thue's result.

Theorem 1. (i) *There exists an infinite 2^+-free binary word.*

(ii) *There exists an infinite cube-free ternary word.*

Instead of proving Theorem 1, we consider in our next example a weaker result which, however, reveals the structure of the proof of Theorem 1 and related results.

Example 6. Consider the morphism

$$h : \begin{array}{ccc} a & \mapsto & aba \\ b & \mapsto & abb \end{array}$$

and the infinite word obtained by iterating it:

$$z = abaabbabaabaabb\ldots\ .$$

We claim that z is a cube-free word and, moreover, that for any $n \in \mathbb{N}$ and $\varepsilon > 0$ it contains a repetition of order larger than $3-\varepsilon$ of a word longer than n.

The latter condition is obvious: by the form of h a word of the form $uuu\,\mathrm{suf}_1(u)^{-1}$ is mapped by h into the word of the same form, and aab is a factor of z of this form. The former condition is proved by a contradiction. Assume that z contains a cube, say uuu. We show that then it contains also a cube $u'u'u'$ with $|u'| < |u|$. This yields a contradiction. Now, we analyze how uuu is covered. There are four cases depending on the prefix of length two of u. The most complicated one is when $u = abu'$. We leave the other cases as an exercise. In this case we can cover the beginning of u by $h(a)$ and $h(b)$ in two different ways:

If the two first occurrences of u are covered in the same way, so is the third, by the length considerations. Moreover, since h is even marked (from left to right) all the occurrences of u are covered in the same way, showing that we can shift a cube by one or two letters to left and make u to match with an image under h. Hence a shorter cube occurs in z.

If the two occurrences are covered by different ways, then the third one would still be covered in different positions, by length considerations, but this is impossible. □

Theorem 1 deserves several further comments.

First, the order of repetition in part (i) is optimal, since every 2-free word over a binary alphabet is of length at most three. In part (ii) the value $k = 3$, in turn, is not optimal. As shown in [De72] the optimal value is $^7/_4$, that is to say every $^7/_4$-free ternary word is finite (in fact of the length at most 38), while there exists an infinite $^7/_4{}^+$-free ternary word.

Second, Theorem 1 can be extended to, see [Br83] and [Lo83]:

Theorem 2. *There exist nondenumerably many infinite* 2^+*-free binary words. The same holds for infinite* 2*-free ternary words.*

If only finite k-free words are considered, then in the binary case the number of 2^+-free words of length n is polynomial while that of cube-free words is exponential. Recently, the exact borderline between polynomial and exponential growth was shown in [KS03]: the value is $2\ ^1/_3$, that is the number of $2\ ^1/_3$-free words is polynomial while that of $2\ ^1/_3{}^+$-free words is exponential.

Our third remark deals with the generation of k-free words. A solution to part (i) in Theorem 1 is given by Thue-Morse word t. In other words, a word obtained by iterating of morphisms. This, indeed, is very typical in the field: most of the known repetition-free words are defined either by iterating a morphism or in addition by mapping a word obtained in this way by another morphism. For example, a solution to part (i) in Theorem 1 is obtained by iterating the morphism

$$\begin{array}{lcl} a & \mapsto & abc \\ b & \mapsto & ac \\ c & \mapsto & b \end{array} .$$

Above motivates a natural question: How to decide whether a given morphism is *k-free*, i.e. maps all k-free words into k-free words. This is a remarkable problem! While the case $k = 2$ is known, and there exist efficient methods to decide 2-freeness, see e.g. [BK03], the decidability status for general integers k or even $k = 3$ is open.

Repetition-free words has a lot of applications. One of the most beautiful ones is a negative solution to the Burnside problem for semigroups, see [Lo83].

6 Periodicity

Periodicity is one of the most fundamental properties of words. In this section we consider periodicity in a particular setting of infinite words. We say that an infinite word w is *globally periodic* if it is *ultimately periodic*, i.e. of the form $w = uv^\omega$ for some finite words u and v. Further we define *local* regularity by the requirement that all of its long enough prefixes ends with a repetition of certain order. Our goal is to analyze the relationships when a local periodicity implies the global one. Such results are among the most fundamental ones in mathematics.

In order to continue let us fix a few notions. Let $\rho \geq 1$ be a real and $p \geq 1$ a natural number. We say that a finite word is *ρ-legal* if it contains as a suffix a repetition of order larger than or equal to ρ, and that it is *$(\rho,\ p)$-legal* if it contains as a suffix a repetition of order ρ of a word of length at most p. Similarly, an infinite word w is ρ-legal or $(\rho,\ p)$-legal if it so for all of its long enough prefixes. Note that $(\rho,\ \infty)$-legality can be interpreted as simply ρ-legality.

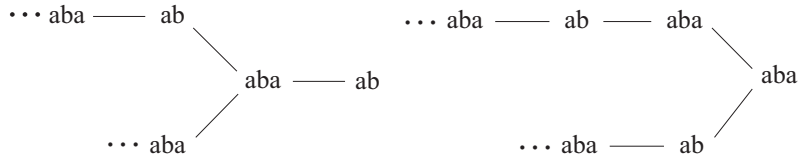
We describe the usefulness of these notions in the following two examples, which are illustrations of subsequent theorems.

Example 7. We claim that the Fibonacci word

$$f = \lim_{i \to \infty} f_i = abaababaabaab \cdots$$

where $f_0 = a,\ f_1 = ab$ and $f_{n+1} = f_n f_{n-1}$ for $n \geq 1$, is $(2,\ 5)$-legal, as first observed by J. Shallit, personal communication.

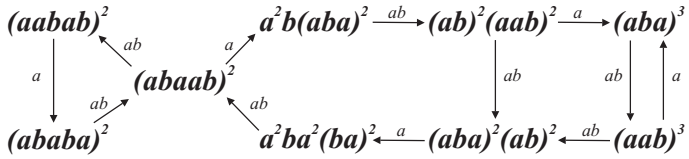
It is straightforward to see that f can be decomposed uniquely into blocks of ab and aba, such that ab does not occur twice and aba three times in a row. Consequently, suffixes of prefixes of f ending at one of the blocks are of the forms:



Now consider the suffixes ending at the rightmost ab in the left graph, i.e. ending either at a or b there. In the former case there is a suffix aa, a square. In the latter case there is necessarily either square $aabaab$ or square $abaababaab$, i.e., also a square of a word of length at most five. The similar argumentation applies for the right graph.

The fact that the word f is not ultimately periodic is left as an exercise. □

Example 8. In this example we point out a striking difference of $(2,\ 5)$-legal and $(2+\varepsilon,\ 5)$-legal infinite words for any $\varepsilon > 0$. Let us search for $(2,\ 5)$-legal infinite words containing a factor $(abaab)^2$. Now, we try to extend a suffix ending to the above mentioned factor exhaustively symbol by symbol preserving the $(2,\ 5)$-legality, and not reporting the extensions leading only to ultimately periodic words. We obtain the graph:



Here the labels tell the extensions, and the nodes correspond the suffixes obtained at particular moments. In the suffixes a short square is always shown (proving the legality) together with a sufficient amount of other letters needed in further steps. In some nodes some continues are not shown – in these cases only ultimately periodic words would be $(2,\ 5)$-legal; for instance from $(abaab)^2$ by b we would obtain a $(2,\ 5)$-legal word, which, however, could be continued only by bs preserving the legality.

It follows from the construction that all words spelled from this graph are $(2,\ 5)$-legal. In particular, there exist nondenumerably many such infinite words, since the graph contains intersecting loops, labeled by noncommuting words. One can also show that actually this graph gives all $(2,\ 5)$-legal nonultimately periodic words. We did the exhaustive search for a particular square, the other squares do not give any other nonultimately periodic $(2,\ 5)$-legal infinite words.

Now, an interesting observation is that if instead of the $(2,\ 5)$-legality the $(2+\varepsilon,\ 5)$-legality is considered, then the node $(aba)^2(ab)^2$ is no longer legal. Indeed, independently of x the word $xabaabaabab$ does not contain at the end a repetition of order strictly larger than 2 of a word of length at most 5. Consequently, intersecting loops are lost, meaning that any $(2+\varepsilon,\ 5)$-legal infinite word is necessarily ultimately periodic.

Constructing graphs similar to the above one for $(2,\ 4)$-legal words one can conclude that all $(2,\ 4)$-legal words are ultimately periodic. □

Above examples are special cases of much deeper results. If in Example 7 cubes instead of squares were asked our approach would not work. Indeed, all 3-legal infinite words are ultimately periodic, or even much strongly ρ-legal infinite words are necessarily ultimately periodic if and only if $\rho \geq \varphi^2 = \varphi + 1 = 2.6\ldots$ where φ is the number of golden ratio $\frac{1+\sqrt{5}}{2}$. This is a remarkable theorem, conjectured by J. Shallit in 1994, and proved in [MRS98] by F. Mignosi, A. Restivo and S. Salemi in 1998:

Theorem 3. *(i) Each φ^2-legal word is periodic.*

(ii) The Fibonacci word is $(\varphi^2 - \varepsilon)$-legal for any $\varepsilon > 0$.

Example 8 considers a similar phenomena is a simple setting yielding the following result, cf. [KLP02]. As outlined in the example the optimality is with respect to both of the parameters.

Theorem 4. *(i) Each $(2,\ 4)$-legal infinite word is ultimately periodic.*

(ii) For any $\varepsilon > 0$, each $(2+\varepsilon,\ 5)$-legal infinite word is ultimately periodic.

(iii) There exists nondenumerably many $(2,\ 5)$-legal infinite words, including the Fibonacci word.

In [Le02] the similar optimal value of ρ is found for any finite length n of the period. For example, and interestingly, the optimal ρ for $n = 5,\ 6,\ \ldots,\ 11$ is the same, namely 2, while for $n = 12$ it is $2\frac{1}{12}$. Further, after some anomaly in small values of n, the behaviour of such optimal ρs is regular, but amazing: there exists just one jump in between two values of consecutive Fibonacci numbers, except that every sixth jump is missing. Also surprisingly, it is not exactly the Fibonacci word, but very related one, which determines these jumps.

We conclude this section with a few remarks. First the above results are beautiful examples, not only in combinatorics on words, but in a much broader perspective, where a local regularity implies the global one, and, in fact, in an optimal way. In other words, they can be seen as results strictly separating a *predictable*, i.e., ultimately periodic, behaviour from a *chaotic* one, i.e., allowing nondenumerably choices. This is more discussed in [KLP02].

7 Complexity

Complexity is a fundamental topic in any domain of science. In the case of infinite words complexity can be defined in a number of very different ways. We consider a few such examples briefly.

To start with we want to emphasize one particular phenomena. Many things which are complex, i.e. complicated, in classical sense of mathematics need not be so from the point of view of combinatorics on words. We already saw in Example 3 that Hilbert's Curve which is analytically very difficult - a kind of topologival anomaly - is very simple in terms of infinite words. Another such situation is encountered when considering, for example, the Fibonacci word as the decimal number it represents. The number is trancendential, although the word is among the most simple nonultimately periodic words, see [FM97] or [BK03] and references given therein.

In what follows we consider three types of complexity in infinite words.

Subword complexity. Let w be an infinite word. Its *subword complexity* (or simply *complexity* if there is no danger of confusion) is the function $p_w : \mathbb{N} \to \mathbb{N}$ such that

$$p_w(n) = \text{card}\{w \in A^n \mid w \in F(w)\},$$

where $F(w)$ is used to denote the set of all factors of w. We note immediately that

- p_w is a nondecreasing function, and
- p_w is bounded for ultimately periodic words.

Also the reverse of the latter condition is true as first noticed in [MH40].

Lemma 2. *The following conditions are equivalent:*

(i) *w is ultimately periodic;*

(ii) *there exists an n such that $p_w(n+1) = p_w(n)$;*

(iii) *p_w is bounded.*

The nontrivial part of Lemma 2 is the implication (ii) $\Rightarrow$ (i). We leave it as an exercise.

It follows that if the complexity is not ultimately constant then $p_w(n) \geq n+1$ for any $n \geq 0$. And the corresponding words are nonperiodic. As we will now show there really exist words such that $p_w(n) = n+1$ for all n. Such words are called *Sturmian words*, the Fibonacci word being an example of those.

Lemma 3. *The Fibonacci word is Sturmian.*

Proof. We recall that the Fibonacci word $f = abaababaabaab\ldots$ is defined by iterating the morphism $\varphi : a \mapsto ab$ and $b \mapsto a$. Consequently, $bb \notin F(f)$ and hence $p_f(2) = 3$. Similarly, aaa is not a factor of f.

Next we show that for any word x both axa and bxb are not factors of f. For $|x| = 0$ we already concluded this. We argue inductively that both axa and bxb are factors of f. Necessarily, x both starts and ends with a so that $x = aya$. Now, consider the factor $bayab$. By the form of the morphism, there exists z such that $\varphi(z) = ay$. Now, both $aayaab$ and $abayab$ are factors of f. Since, necessarily $aayaab = \varphi(bzba)$ and $abayab = \varphi(aza)$ implying that both bzb and aza are factors of f, a contradiction.

By above we can conclude that f has at most one *special* factor of each length n, meaning that only one word of that length can be extended to the right by both the symbols such that the word remains as a factor of f. Indeed, assume that there are two special factors, say u and v, of the same length. Let s be the longest common suffix of these words. Then all the words asa, asb, bsa and bsb are factors of f, a contradiction with our previous considerations.

It follows that $p_w(n) \leq n+1$ for all $n \geq 0$. So, to conclude, it is enough, by Lemma 2, to prove that f is not ultimately periodic, see Example 7. □

Quite a lot of research has been done on subword complexity of infinite words. We mention here only one nice result, see [ELR75] and [Pa84], and otherwise refer to [BK03] and references given therein.

Theorem 5. *The subword complexity of a word obtained by iterating a morphism is always in one of the following five classes* $\Theta(1)$, $\Theta(n)$, $\Theta(n \log n)$, $\Theta(n \log\log n)$ *and* $\Theta(n^2)$.

We conclude by noting that it is not known whether the complexity of the Kolakoski word is in $\mathcal{O}(n^2)$.

Descriptive complexity. This notion is considered implicitly in Section 4. Here it is asked how complicated mechanisms are needed to generate a particular infinite word. The iteration of morphisms and dgsm's are two examples of different devices, and hence also derive to two complexity classes of words. Several problems on this area were already mentioned in Section 4.

Computational complexity is the fundamental part of formal language theory. Since infinite words are closely related to infinite languages, as shown in Lemma 1, it is natural to study the complexity of generation of infinite words. Here the complexity is measured as the amount of resources, for example time or space, needed to print the nth symbol of the word. Most natural model is that of off-line Turing machine generating an infinite word, for definitions see [HKL94]. Quite a lot of traditional complexity theory can be translated into this formalism. We, however, state here only two simple results.

Theorem 6. (i) *The family of infinite words generated in space* $\Theta(1)$ *coincides with the family of ultimately periodic words.*

(ii) *All infinite words in the space class* $o(\log n)$ *are ultimately periodic.*

Next result relates the descriptional and computational complexities of infinite words.

Theorem 7. *Each word defined by iterating morphism can be generated in space* $\mathcal{O}(\log n)$.

It is an open problem, and, in fact, a reformulation of a classical problem in complexity theory as shown in [DM03], whether Theorem 7 extends to all words obtained by iterating a deterministic gsm.

References

[Ad79] S. I. Adian, *The Burnside Problem and Identities in Groups*, Springer-Verlag, 1979.

[AL85] M. H. Albert and J. Lawrence, A proof of Ehrenfeucht's conjecture, *Theoret. Comput. Sci.* 41, 121–123, 1985.

[Ar37] S. E. Aršon, Proof of the existence on n-valued infinite asymmetric sequences, *Mat. Sb.* 2(44), 769–779, 1937.

[BEM79] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty, Avoidable patterns in strings of symbols, *Pacific J. Math.* 85, 261–294, 1979.

[Be79] J. Berstel, *Transductions and Context-Free Languages*, Teubner, 1979.

[BK03] J. Berstel and J. Karhumäki, Combiantorics on Words - A Tutorial, *Bull. EATCS* **79**, 178–228, 2003.

[Br83] F.-J. Brandenburg, Uniformly growing k-th power-free homomorphisms, *Theoret. Comput. Sci.* 23, 69–82, 1983.

[CK97] C. Choffrut and J. Karhumäki, Combinatorics of words, In: A. Salomaa and G. Rozenberg (eds.), *Handbook of Formal Languages, Vol. 1*, 329–438. Springer-Verlag, 1997.

[CK94] K. Culik II and J. Karhumäki, Iterative devices generating infinite words, *Int. J. Found. Comput. Sci.* 5, 69–97, 1994.

[De72] F. Dejean, Sur un théorème de Thue, *J. Combin. Th. A* 13, 90–99, 1972.

[DM03] P. Ďuriš and J. Manuch, On the computational complexity of infinite words, *Theoret. Comput. Sci.* **1-3**, 141–151, 2003.

[ELR75] A. Ehrenfeucht, K. P. Lee, and G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interaction, *Theoret. Comput. Sci.* 1, 59–75, 1975.

[FM97] S. Ferenczi and C. Mauduit, Transcendence of numbers with a low complexity expansion, *J. Number Theory* 67, 146–161, 1997.

[HU79] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, 1979.

[HKL94] J. Hromkovič, J. Karhumäki, and A. Lepistö, Comparing descriptional and computational complexity of infinite words, In: J. Karhumäki, H. Maurer, and G. Rozenberg (Eds.), *Results and Trends in Theoretical Computer Science*, LNCS 812, 169–182, Springer-Verlag, 1994.

[Ga00] C. F. Gauss, *Werke*, Teubner, Leibzig, 1900 (pp. 272 and 282–286).

[Gu86] V. S. Guba, The equivalence of infinite systems of equations in free groups and semigroups to their finite subsystems, *Math. Zametki* 40, 321–324, 1986.

[KLP02] J. Karhumäki, A. Lepistö, and W. Plandowski, Locally periodic infinite words and a chaotic behaviour, *J. Comb. Theor., Ser. A* 100, 250–264, 2002.

[KM03] J. Karhumäki and J. Manuch, in preparation.

[KS03] J. Karhumäki and J. Shallit, Polynomial versus exponential growth in repetition-free binary words, manuscript, 12pp, 2003.

[Le93] A. Lepistö, On the power of periodic iteration of morphisms, LNCS **700**, 496–506, 1993.

[Le02] A. Lepistö, On Relations between Local and Global Periodicity, Ph.D. Thesis, University of Turku, *TUCS Dissertations* 43, 2002.

[Lo83] M. Lothaire, *Combinatorics on Words*, *Encyclopedia of Mathematics* 17, Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, 1997.

[Lo02] M. Lothaire, *Algebraic Combinatorics on Words. Encyclopedia of Mathematics* 90, Cambridge University Press, 2002.

[LS77] R. C. Lyndon and P. E. Schupp, *Combinatorial Group Theory*, Springer-Verlag, 1977.

[MKS66] W. Magnus, A. Karrass and D. Solitar, *Combinatorial Group Theory*, Wiley, 1966.

[Ma77] G. S. Makanin, The problem of solvability of equations in a free semigroup, *Mat. Sb.* 103, 147–236, 1977 (English transl. in *Math. USSR Sb.* 32, 129–198).

[MRS98] F. Mignosi, A. Restivo, and S. Salemi, Periodicity and golden ratio, *Theoret. Comput. Sci.* 204, 153–167, 1998.

[Mi67] M. L. Minsky, *Computation: Finite and Infinite Machines*, Prentice-Hall, 1967.

[MH38] M. Morse and G. Hedlund, Symbolic dynamics, *Amer. J. Math.* 60, 815–866, 1938.

[MH40] M. Morse and G. A. Hedlund, Symbolic dynamics II: Sturmian trajectories. *Amer. J. Math.* 62, 1–42, 1940.

[Pa84] J.-J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphismes intérés, In: J. Paredaens (ed.), *Automata, Languages and Programming*, LNCS 172, 380–389, Springer-Verlag, 1984.

[Pl99] W. Plandowski, Satisfiability of word equations is in PSPACE, Proc. of FOCS, 495–500, 1999.

[Pr51] E. Prouhet, Mémoire sur quelques relations entre les preissances des numbres, *C. R. Acad. Sci. Paris* 33, Cahier 31, 225, 1851.

[Sa73] A. Salomaa, Formal Languages, Academic Press, 1973.

[Sc56] M. P. Schützenberger, Une théorie algébrique du codage, *Seminaire Dubreil-Pisot 1955–1956 Expose 15*, Institut H. Poincare, Paris, 1956.

[Sh89] J. Shallit, 2 methods for generating fractals, *Computer & Graphics* **13(2)**, 185–191, 1989.

[Th06] A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl., Christiana* 7, 1–22, 1906.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi

University of Turku
- **Department of Information Technology**
- **Department of Mathematics**

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

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**