Combinatorics on Words – A Tutorial *

J. Berstel[1] and J. Karhumäki[2**]

[1]Institut Gaspard-Monge, Université de Marne-la-Vallée, 77454 Marne-la-Vallée Cedex 2, France, email: jean.berstel@univ-mlv.fr
[2]Department of Mathematics and Turku Centre for Computer Science, University of Turku, 20014 Turku, Finland, email: karhumak@cs.utu.fi

Table of Contents

---

* To appear in the Bull. EATCS. The work of the second author was done while visiting Institut Gaspard-Monge Université de Marne-la-Vallée
** Supported by the Academy of Finland under the grant 44087

1 Introduction

During the last two decades research on combinatorial problems of words, i.e. on *Combinatorics on Words*, has grown enormously. Although there has been important contributions on words starting from the very beginning of last century, they were scattered and typically needed as tools to achieve some other goals in mathematics. A notable exception is combinatorial group theory, which studies combinatorial problems on words as representing group elements, see [LS77] and [MKS66]. Now, and in particularly after the appearance of Lothaire's book – Combinatorics on Words – in 1983 the topic has become a challenging research topic of its own. In the latest classification of Mathematical Reviews combinatorics on words constitutes its own section under the chapter discrete mathematics related to computer science. Although the applications of words are, by no means, only in computer science the classification catches the basic of the nature of combinatorics on words.

Recent developments of the field culminated in the second Lothaire's book – Algebraic Combinatorics on Words – which appeared in 2002. Its more than 500 pages witness the vital stage of the topic. The new book repeats basically nothing from the first one, and actually most of the results are discovered during the last twenty years. Biannual conference – referred to as WORDS – devoted entirely to combinatorics on words has also been created. The fourth event will be in Turku in 2003.

A word is a sequence of symbols, finite of infinite, taken from a finite alphabet. A natural environment of a finite word is a free monoid. Consequently, words can be seen as a discrete combinatorial objects or discrete algebraic objects in a noncommutative structure. These two facts – discreteness and noncommutativity – are the two fundamental features of words. At the same time they explain why many problems are so difficult.

Words are central objects of automata theory, and in fact in any standard model of computing. Even when computing on numbers computers operate on words, i.e. representations of numbers as words. Consequently, in one hand, it is natural to study algorithmic properties of words. On the other hand, the undecidability of problems is most easily stated in terms of words – Post Correspondence Problem being a splendid example. These both elements of words – algorithmic aspects and undecidability – are visible, often implicitly, throughout our presentation.

The goal of the tutorial is to discuss – without aiming to be exhaustive – several typical problems on words, as well as to try to point out several applications. With a few exceptions the proofs are not presented here, however, in some cases we use examples to illustrate the basic ideas. Open problems form an important part of our presentation.

The contents of this tutorial is as follows. At the end of this introductory section we discuss briefly the history of combinatorics on words, and fix some basic terminology. Then in Section 2 we consider connections to other fields of mathematics and computer science. Section 3 is devoted to the most fundamental notion of words, namely to periodicity. Dimension properties of words constitute Section 4, while Section 5 concentrates one of the most studied and most characteristic feature of words, namely unavoidable regularities. Words, indeed, are very suitable objects to formulate such fundamental properties. In Section 6 complexity issues of infinite words are studied from different points of view. An interesting phenomenon is that what is considered to be complicated in a classical sense, e.g. algebraicly, need not be so from the point of view of words. Finally, in Section 7 we discuss about some extensions of the theory to finite sets of words, and in Section 8 collect a list of important open problems, many of those being apparently very difficult. As we said everywhere above algorithmic and decidability issues are present.

We conclude by some bibliographic remarks. Combinatorics on words has now become a rich area, with many connections to algorithms, to number theory, to symbolic dynamics, and to applications in biology and text processing. Several books have appeared quite recently, or will appear in the next months, that emphasize these connections. What have to be mentioned are the book of Allouche and Shallit [AS03], where the emphasis is on relation to automata theory, and the book by Pytheas-Fogg [PF02] which is a *nom de plume* for the Marseille group. Quite recently the book by Crochemore and Rytter [CR02] appeared as a follow-up book to [CR94]. A detailed introduction into algorithms on words is given in the book [CHL01]. Algorithms on words are also described, from a more biological point of view, in Gusfield's book [Gus97]. Finally, we should point to algebraic applications of combinatorics on words, as they appear in the book of de Luca and Varricchio [dLV99]

1.1 History

History of combinatorics on words goes back to the beginning of the last century, when A. Thue started to work on repetition-free words. He proved, among other things, the existence of an infinite square-free word ovar a ternary alphabet. Interestingly, it seems that Thue had no outside motivation for his research on words. He published his results in two long papers [Th06] and [Th12], but unfortunately in a less known Norwegian journal, so that his results became known only much later, cf. [Be95]. Actually many of those were reproved several times.

The notion of word is, of course, so natural that it can be found even in several older mathematical works. Even Gauss considered a problem which was nothing but a problem on combinatorics on words, cf. [Ga00] and [KMPS92], and in 1850s Prouhet [Pr51] introduced the most famous infinite word redefined by Thue. However, Thue was clearly the first to study systematically problems on words, and moreover as problems of their own.

After Thue during the first half of the previous century there were only a few isolated works on words, such as those of [Ar37], [Mo21], [Mo38] and [MH38]. In these works, as it has been very typical for the whole field, properties of words were not that much the research topic of itself rather than tools for solving problems in other areas.

It took till the second half of the last century when the theory of words arose. This happened more or less simultaneously in France and Russia. In France it grew up from research of M.P. Schützenberger on theory of codes, see [Sch56]. The Russian school, in turn, developed from the seminal work of P.S. Novikov and S. Adian on Burnside Problem for groups, see [Ad79]. Especially, in Russia results on words, not to speak about the theory, were not so explicit, although their studies culminated rather soon to remarkable results, such as Makanin's algorithm for satisfiability of word equations, see [Ma77]. In France, the theory of words became an independent topic of its own rather soon, very much due to the stimulating paper [LS67] from 1967. Other stimulating early works were (hand written) book [Len72] and [Hm71].

Once the foundations of the theory were laid down it developed rapidly. One influencial paper should be mentioned here, from year 1979. In [BEM79] repetition-free words were studied very extensively, and an important notion of an avoidable pattern, as well as many open problems, were formulated. The $D0L$ systems, and particularly the $D0L$ problem [CF77], was an important source of many questions on combinatorics on words, including the Ehrenfeucht Compactness Property, see [Ka93].

In fifteen years or so, in 1983 the active research on words culminated to the first book of the topic, namely Lothaire's book Combinatorics on Words. The starting point of Lothaire's book was a mimeographed text of lectures given by M.P. Schützenberger at the University of Paris in 1966 and written down by J.F. Perrot. It had an enormous influence on the further development of the field. Results of this impact, including several jewels of theory, can be seen from the second Lothaire's book – Algebraic Combinatorics on Words – which appeared last year.

1.2 Notions and notations

We conclude this Introduction by fixing the terminology and a few notions needed in this presentation. For more detailed definitions we refer to [Lot02] or [CK97].

We denote by A a finite set of symbols referred to as an *alphabet*. Sequences, finite or infinite, of letters from A are called *words*. The empty sequence is called the *empty word* and is denoted by 1 or ε. The set of all finite words, in symbols A^*, is the *free monoid* generated by A under the operation of *product* or *concatenation* of words: $u \cdot v = uv$. The *free semigroup* generated by A, in symbols A^+, is $A^* \setminus \{1\}$. The set of one-way *infinite* words over A is denoted by A^w. Formally, such words are mappings from $\mathbf{N}$ into A.

For two words u and v we say that u is a *prefix* (resp. a *suffix* or a *factor*) of u if there exist a word x (or words x and y) such that $v = ux$ (resp., $v = xu$ or $v = xuy$). In the case of prefix we write $u = vx^{-1}$. Note that the mapping $(v,\ x) \mapsto vx^{-1}$ can be viewed as a partial mapping from $A^* \times A^*$ into A^*. These notions extend straightforwardly to subsets, i.e. to *languages*, of A^*. A very crucial notion in combinatorics on words is that of a *morphism* from the free monoid A^* into itself (or into another free monoid B^*), that is to say a mapping $h : A^* \to A^*$ satisfying $h(uv) = h(u)h(v)$ for all words u and v. Examples of important morphisms are

$$\mu : \begin{array}{l} a \mapsto ab \\ b \mapsto ba \end{array} \quad \text{and} \quad \varphi : \begin{array}{l} a \mapsto ab \\ b \mapsto a \end{array}.$$

The former, discovered by Thue, is so-called *Thue-Morse*, sometimes referred to *Prouhet-Thue-Morse*, *morphism.* It plays an important role in the study of repetition-free words. The other morphism is called *Fibonacci morphism.* These morphisms has a property that they map a to a word starting with a as a prefix. This implies that the limits

$$t = \lim_{i\to\infty} \mu^i(a) \quad \text{and} \quad f = \lim_{i\to\infty} \varphi^i(a)$$

exist. We say that f and t are obtained by *iterating morphisms* μ and φ at the word a. Clearly, f and t are the unique fixed-points of these morphisms. They are called *Thue-Morse* and *Fibonacci words*, respectively. We have, for example,

$$f = abaababaabaab\ldots$$

It is easy to see that alternatively

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

where

$$f_0 = a,\ f_1 = ab \quad \text{and} \quad f_{n+1} = f_n f_{n-1} \quad \text{for} \quad n \geq 1.$$

These formulas explain the name Fibonacci morphism.

In order to state some properties of these words, let us say that a word w is *k-free* if it does not contain as a factor any word of the form u^k. This notion extends, cf. [CK97], in a natural way to nonnegative rational numbers, and also to real numbers ζ, when the requirement is that w does not contain a factor of the form u^k with $k \in \mathbf{Q}$ and $k > \zeta$. If $w = xu^{k'}y$, with $k' \geq k$ and $u \neq 1$, we say that w contains a *repetition of order k*. By a k^+-free word we mean a word which is k'-free for any $k' > k$ (but not necessarily k-free).

What Thue proved was that the Thue-Morse word is 2^+-free, i.e. does not contain repetitions of higher order than 2. Indeed, it – like any binary word of length at least four – contains a repetition of order 2. For the Fibonacci word the repetitions which are avoided in it are exactly those being of order $> \varphi^2 + 1$ when φ is the number of the golden ratio, i.e. $\varphi = \frac{\sqrt{5}+1}{2}$. In other words, the Fibonacci word is $(\varphi^2+1)^+$-free, cf. [MP92]. This is just one of the many special properties of the Fibonacci word. In fact, it is almost a universal counterexample for conjectures or an example showing an optimality, an exception being a problem in [Cas97a].

2 Connections

In this section we point out some connections of combinatorics on words and other areas of mathematics and computing. Such connections are quite broad and has been fruitful in both directions. In fact, many important properties of words has been discovered when looking for tools to solve other completely different problems. More concretely, we discuss here three different connections, one to matrices, one to algebra and one to algorithms. These reflects, we hope, different aspects of such connections. Other connections to combinatorial group theory, to algebraic combinatorics, and to general combinatorics are sketched or described in [MKS66] and [LS77]. More specifically, let us just mention Lie Algebras, see [Re93], words as codings of combinatorial structures, and words as codings of permutations.

For the beginning, however, let us consider another typical and interesting relation between words and some classical mathematical notions. *Hilbert's space filling curve* has played an important role describing an intuitive anomaly in topology. From the point of view of topology it can be seen as quite a complicated object. However, from the point of view of words it is nothing but an infinite word over a four element alphabet which, moreover, is easy to define: it is a morphic image under a length preserving morphism, i.e. a coding, of the fixed-point of an iterated morphism.

2.1 To matrices

As the first connection we consider that of words and matrices, more precisely, that of multiplicative semigroups of matrices. Let us denote by $M_{n\times n}(S)$ the family of $n \times n$ matrices with entries in the semigroups S. It has been known since 1920s that free monoids can be embedded into the multiplicative semigroup of 2×2 matrices over $\mathbf{N}$, i.e. into $M_{2\times 2}(\mathbf{N})$. For $A = \{a,\ b\}$ such an embedding is given, for instance, by the mapping

$$(1) \qquad a \mapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad b \mapsto \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

In fact, this mapping is an isomorphism between $\{a,\ b\}^*$ and $SL_2(\mathbf{N})$, the set of matrices in $M_{2\times 2}(\mathbf{N})$ having a determinant equal to 1. Arbitrary, even countable, free semigroups can be embedded in $M_{2\times 2}(\mathbf{N})$ by employing an embedding from $\{a_i | i \in \mathbf{N}\}^*$ into A^* given by

$$a_i \mapsto ab^i \quad \text{for} \quad i \geq 0.$$

The above ideas become even more usable when we associate above with a morphism $h : A^* \to A^*$. In order to simplify the notation we set $A = \{1,\ 2\}$ and define the mapping

$$1 \mapsto \begin{pmatrix} k^{|h(1)|} & 0 \\ \nu(h(1)) & 1 \end{pmatrix} \quad \text{and} \quad 2 \mapsto \begin{pmatrix} k^{|h(2)|} & 0 \\ \nu(h(2)) & 1 \end{pmatrix},$$

where the vertical bars are used to denote the length of a word and $\nu : A^+ \to \mathbf{N}$ maps a word into the number it represents at base $k > 3$. It is straightforward to see that this mapping too is an embedding, that is to say, for any word $u = a_1 \cdots a_n$, we have

$$u = a_1 \cdots a_n \mapsto \begin{pmatrix} k^{|h(u)|} & 0 \\ \nu(h(u)) & 1 \end{pmatrix}.$$

Consequently, questions asking something about images of a morphism of A^* can be transformed into questions about matrices. More concretely; this allows to transfer the undecidability of the Post Correspondence Problem into undecidability results on matrices. Paterson [Pat70] was among the first to use this idea when showing (i) in the following theorem, for the other parts we refer to [HK97], where also further references can be found.

Theorem 2.1. *The following questions are undecidable:*

(i) Does a given finitely generated multiplicative subsemigroup of $M_{3\times3}(\mathbf{Z})$ contain the zero matrix?
(ii) Does a given finitely generated multiplicative subsemigroup of $M_{3\times3}(\mathbf{Z})$ contain a matrix having the zero in the right upper corner?
(iii) Is a given finitely generated multiplicative subsemigroup of $M_{3\times3}(\mathbf{N})$ free?

Part (iii) is undecidable even for upper triangular matrices.

All of the above problems are open in dimension $n = 2$. Another interesting open problem is the question of part (i) for the identity matrix.The embedding methods used in Problems (i)-(iii) does not seem to work here, see [CHK99]. In the dimension $n = 2$ this problem is decidable, see [CK02].

Embeddings like (1) are important not only to conclude the above undecidability results, but also to obtain results for words from those of matrices. The Ehrenfeucht Compactness Property, discussed in Section 4, is a splendid example of that.

2.2 To algebra

The second connection we consider is that to algebra. We want to give a concrete example rather than recalling that words are, after all, elements of free monoids or that representations of groups are relations on words. This example, due to [MH44], is an application of repetition-free words to solve *Burnside Problems for semigroups*. The problem asks whether the assumptions (i): the semigroup S is finitely generated; and (ii): each element of S is of a finite order, i.e. generates a finite cyclic subsemigroup, imply that the semigroup S itself is finite.

Theorem 2.2. *Burnside Problem for semigroups has a negative answer.*

The answer is achieved as follows, cf. [Lot83]. Let A be a three letter alphabet. Then by a result of Thue, there exists a square-free word over A, and consequently the set of finite square-free words is infinite. Now, adjoin to the free semigroup A^+ the zero element, and define a congruence $\approx$ on $A^+ \cup \{0\}$: each square-free word forms an equivalence class of its own, and the rest belongs to the class containing 0. Then the quotient semigroup

$$S = A^+ \cup \{0\} / \approx$$

is well defined and has the required properties. For finitely generated groups the Burnside Problem is much more complicated, cf. [Ad79].

2.3 To algorithmics

Finally, we discuss about connections to algorithmics. We consider very natural algorithmic problem, and point out how a certain property of words can be used to obtain an efficient solution to the problem. We ask

Question. How to decide efficiently whether a given word is primitive?

The problem has a brute force quadrative solution: divide the input into two parts and check whether the right part is a power of the left part. But how to get a faster solution. The answer comes from the property of primitive words: a word w is nonprimitive if and only if it is a factor of ${}^\bullet ww^\bullet$, i.e. w occurs properly as a factor in ww. For the definition of the operator $^\bullet$ see the next section. So the problem is reduced to a simple instance of the *string matching problem*, and hence doable in linear time, see e.g. [CR94].

Despite of the simplicity of the above example it is very illustrative: it shows how the correctness of an algorithm is finally based on a combinatorial property of words. This seems to be a common rule in efficient string algorithms. Or even more strongly, whenever a fundamental property of words is revealed, it has applications in improving algorithms on words, cf. e.g. [MRS95].

String matching and pattern matching are only two – although important – aspects of algorithmics on words, see [CR94], [CHL01] for expositions, and [Gus97]. Other algorithmic problems we do not consider here are: systematic generation of words (e.g. Dyck words), ranking, unranking, and random generation of words, see e.g. [Ru78], [BBG90] and [FZC94]. All of these are used in order to code combinatorial structures.

Another important topic on algorithmic combinatorics on words, which is neither considered in our tutorial, is the *satisfibility problem* for word equations, i.e. the decision question whether a give word equation with constants possesses a solution. The classical paper of Makanin [Ma77] answers affirmatively to this question. However, his algorithm is one of the most complicated algorithms ever presented, see Chapter

12 in [Lot02] for a detailed exposition. Rather recently W. Plandowski showed, by his completely new algorithm, that the problem is actually in *PSPACE*, cf. [Pla99]

3 Periodicity

In this section we consider one of the most fundamental notions of words, namely periodicity. We discuss three different topics, all of those being very fundamental.

3.1 Fine and Wilf's theorem

One of the oldest results in combinatorics on words concerns commutation of words. It is the following well-known statement. Typically this result, in one form or another, has been proved when needed, see e.g. [LS62].

Theorem 3.1. *Let x and y be nonempty words. The following properties are equivalent:*

(i) $xy = yx$,
(ii) the infinite words x^ω and y^ω are equal,
(iii) there exists a word z such that $x, y \in z^+$,
(iv) $\{x, y\}$ is not a code, i.e. satisfies a nontrivial relation.

There is another rather old result, due to Fine and Wilf, strongly related to this theorem. It uses the notion of a period. Let $w = a_1 \cdots a_n$ be a word, with $a_1, \ldots, a_n$ letters. An integer p is a *period* of w, if $1 \leq p \leq n$ and $a_i = a_{p+i}$ for $i = 1, \ldots, n-p$. Thus $a_1 \cdots a_{n-p} = a_{p+1} \cdots a_n$, and the word $w' = a_1 \cdots a_{n-p}$ is both a prefix and a suffix of w. Provided $p < n$, the word w' is called a *border* of w. Conversely, if z is a border of w, then the integer $n - |z|$ is a period of w.

Clearly, an integer p, for $1 \leq p \leq n$, is a period of w if and only if w is a prefix of the infinite word x^ω, where $x = a_1 \cdots a_p$.

A word of length n always has at least the period n. If w has two periods p and q, then w also has the period $p + q$ provided $p + q \leq n$. The set $\Pi(w)$ of all periods of a word w has been described in [GO81] (see also Chapter 8 of [Lot02]). The shortest integer in $\Pi(w)$ is frequently called *the period* of w.

Theorem 3.2. (Fine and Wilf's Theorem) *Let w be a word of length n. If w has two periods p and q and $n \geq p + q - \gcd(p, q)$, then also $\gcd(p, q)$ is a period of w.*

For the proof of Fine and Wilf's theorem, we consider a variation of Theorem 3.1. Given a word $w = a_1 \cdots a_n$, where $a_1, \ldots, a_n$ are letters, we set $w^\bullet = a_1 \cdots a_{n-1}$. In particular, $a^\bullet = \varepsilon$ for a letter a, and $\varepsilon^\bullet$ is undefined.

Lemma 3.3. *Let x and y be nonempty words. If $xy^\bullet = yx^\bullet$, then $xy = yx$.*

Proof. By induction on $|xy|$. If $|xy| = 2$, and more generally, if $|x| = |y|$, one gets $x = y$. Otherwise, one may assume $|x| > |y|$. Then y is a proper prefix of x, and $x = yz$ for some nonempty word z. It follows that $zy^\bullet = yz^\bullet$. By induction, $zy = yz$, and consequently $xy = yx$. $\square$

We now prove Fine and Wilf's theorem. First, we show that p and q may be assumed to be relatively prime. Indeed, assume $d = \gcd(p, q) > 1$, set $w = a_0 \cdots a_{n-1}$ and define d words $w^{[i]} = a_i a_{i+d} \cdots a_{i+n_i d}$, where $n_i = \lfloor (n-i-1)/d \rfloor$, for $i = 0, \ldots, d-1$. These words have periods p/d and q/d and length at least n/d. By the conclusion, each $w^{[i]}$ is a power of some letter. Thus w is the power of some word of length d, that is w has period d.

Suppose now that p and q are relatively prime. Let x and y be the prefixes of w of length p and q, respectively. Then w is a common prefix of x^ω and y^ω. Moreover, setting $w = xw'$, the word w' is a prefix of w, so w' is a prefix of y^ω and thus w is a prefix of xy^ω. Symmetrically, w is a prefix of yx^ω. Since the length of w is at least $p+q-1$, one gets $xy^\bullet = yx^\bullet$. By Lemma 3.3, one gets $xy = yx$. Thus x and y are in some z^+. But since $\gcd(p, q) = 1$, this implies that z is a letter. This completes the proof. $\square$

Fine and Wilf's original paper [FW65] contains three theorems. The first one is basically Theorem 3.2. It states indeed that if two infinite words u and v have periods p and q respectively, and they share a common prefix of length $p + q - \gcd(p, q)$, then they are equal. The two other theorems are of the same vein, but concern real continuous periodic functions.

The proof of Theorem 3.2 in Fine and Wilf's original paper [FW65] is quite different and deserves a short description. Any letter of the alphabet is considered as a number. Any infinite word $a_0a_1a_2 \cdots a_n \cdots$ corresponds to a formal series $a_0 + a_1t + a_2t^2 + \cdots + a_nt^n + \cdots$. Thus, the infinite periodic word x^ω, with $x = a_1 \cdots a_p$, corresponds to the formal series $F(t) = P(t)/(1 - t^p)$, with $P(t) = a_1 + \cdots + a_pt^{p-1}$, and similarly y^ω, with $y = b_1 \cdots b_q$, corresponds to $G(t) = Q(t)/(1-t^q)$, with $Q(t) = b_1 + \cdots + b_qt^{q-1}$. Now, a computation with rational fractions shows that

$$H(t) = F(t) - G(t) = \frac{1 - t^{\gcd(p,q)}}{(1-t^p)(1-t^q)} R(t)\,,$$

where $R(t)$ is a polynomial of degree at most $p + q - \gcd(p, q) - 1$. By assumption, $R(t) = 0$, and consequently $F = G$ and $x^\omega = y^\omega$.

Several other proofs of the theorem are known. Some are by induction on the length of the period (e.g. Chapter 8 of [Lot02] and [CHL01]). There are also proofs that argue directly on congruential properties of the indices [CK97]. Extension to more than two periods are given in [CMR99] and [Jus00]. Further related results are shown in [MSW01] and [BB99].

The bound in Fine and Wilf's theorem is sharp. A concrete example is the word *abaababaaba*. It has periods 5 and 8 and length $11 = 8+5-2$, and is not a power of a single letter. In fact, *all* words of length $p+q-2$, with periods p and q, for coprimes p, q, are known. They are all binary, and, more precisely, prefixes of infinite standard Sturmian words, see Chapter 2 of [Lot02] for a description and references.

3.2 Critical factorization theorem

The Critical Factorization Theorem stated below gives a relation between the global period of a word, and a notion of local period called local repetition, associated to a factorization of the word. As it will appear, global periods are always longer than local periods, but the remarkable fact is that for any word, there is always a factorization whose local period is equal to the global period. Such a factorization is called critical.

In order to formalize the above we say that two words x and y are *prefix comparable* (resp. *suffix comparable*) if one of the words is a prefix (resp. a suffix) of the other. Further, given a word w and a factorization $w = uv$ into nonempty words, a *repetition* at (u, v) is a nonempty word z such that z and u are suffix comparable, and z and v are prefix comparable. The (*local*) *period* of the factorization (u, v) is the length of the shortest repetition at (u, v). It is easy to see that any local period of a word is shorter that the period. A factorization (u, v) is *critical* if its period is equal to the period of w.

Consider for example the word $w = abaab$ which has the period 3. The factorization $(a, baab)$ has the repetition ba, so its period is 2. The factorization (aba, ab) has period 1. The factorizations (ab, aab) and $(abaa, b)$ both have period 3, and these are the critical factorizations of the word w. The following theorem shows that critical factorizations are unavoidable.

Theorem 3.4. (Critical Factorization Theorem.) *Every word of length at least* 2 *has a critical factorization.*

The first statements and proofs of the theorem are given in [CV78] and [Duv79]. A proof of the critical factorization theorem in its present form, and a discussion, is given in Chapter 8 of [Lot02]. A short proof is in [CP91]. Recent results related to this topic appear in [HN02].

We sketch now an interesting application of the Critical Factorization Theorem. Consider a finite set $X \subset A^+$ of nonempty words. Given a word $w \in A^+$, a sequence of nonempty words

$$(s, x_1, \ldots, x_m, p)$$

is an X-*interpretation* of w if $w = px_1 \cdots x_m p$, and p is a prefix of a word in X, s is a suffix of a word in X, and $x_1, \ldots, x_m \in X$. Two X-interpretations $(s, x_1, \ldots, x_m, p)$ and $(s', x'_1, \ldots, x'_{m'}, p')$ of a word w are *disjoint* if $sx_1 \cdots x_i \neq s'x'_1 \cdots x'_{i'}$ for $i = 1, \ldots, m$, $i' = 1, \ldots, m'$.

As an example, consider the set $X = \{a^3, b, aba, a^2ba^2\}$. The word $a^4ba^4ba^4b$ has the X-interpretation $(a, a^3, b, a^3, aba, a^3, b)$. The sequence $(a^3,\ aba,\ a^3,\ b,\ a^3,\ ab)$ is another X-interpretation of the word $a^4ba^4ba^4b$, and this interpretation is disjoint from the previous one.

Theorem 3.5. *Let X be a finite set of nonempty words, and let p be the maximum of the periods of the words in X. Every word w with the period strictly greater than p has at most* $\mathrm{Card}(X)$ *disjoint X-interpretations.*

As another example consider the set $X = \{a^p\}$, consisting of a single word which has period 1. Clearly, every word w that admits an X-interpretation is a power of a and so also has period 1. The number of its disjoint X-interpretations is bounded by p, and not by $\mathrm{Card}(X)$.

Proof. (Sketch) Assume that a word w has $m > n$ disjoint X-interpretations, where $n = \mathrm{Card}(X)$. Consider any factorization $w = uv$. There exist m distinct couples (y, z) of words such that $yz \in X$, u and y are suffix comparable, and v and z are prefix comparable. Since $m > n$, there are at least two couples (y_1, z_1) and (y_2, z_2) such that $y_1z_1 = y_2z_2$. Set $x = y_1z_1 = y_2z_2$. Assume $|y_1| > |y_2|$. Next y_1 and y_2 are prefix comparable, or z_1 and z_2 are suffix comparable. Except for some special cases, which we do not consider in this sketch, the number $d = |y_1| - |y_2| = |z_2| - |z_1|$ is a period of x. This implies that there is a repetition of length d at (u, v). Thus the local period at (u, v) is at most p.

Thus, all local periods are smaller than p. However, by the Critical Factorization Theorem, at least one local period is strictly greater than p. This yields a contradiction. □

A detailed proof is given in [Lot83]. An application to the order of the subgroups of the syntactic monoid of the set X^* was given in [Sch79]. There is a renewal of research on this topic now, see [PR02].

In [Ma02] examples were given showing that the bound for the number of disjoint X-interpretations is optimal in Theorem 3.5. For example, the set $\{a^iba^i \mid i = 1, \ldots, n-1\} \cup \{ba^nb\}$ is such a set.

3.3 Characterizations for ultimately periodic words

We conclude this section with a third fundamental periodicity result of words derived in [MRS95], see also [MRS98]. It characterizes one-way infinite ultimately periodic words in terms of local periodicity. Intuitively, it tells how much *local regularity*, i.e. periodicity, is needed to guarantee the *global regularity*, i.e. ultimate periodicity. Such results are the most basic goals 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 ρ (or equivalently – by definition – 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.

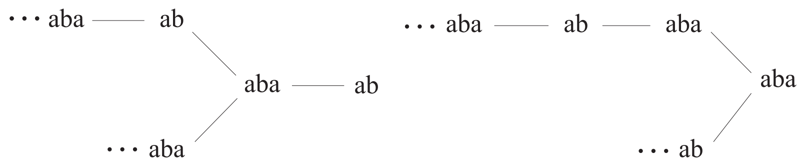
We describe the usefulness of these notions in the following two examples, which are illustrations of subsequent theorems.

Example 1. 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.

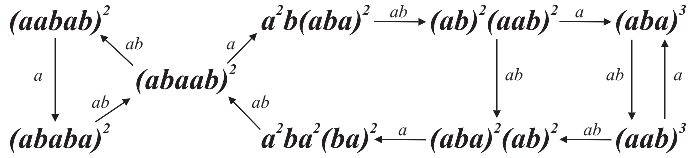
It is straightforward to see that f can be decomposed uniquely into blocks of ab and aba, such that ab does not occur twice and aba three times in a row. Consequently, suffixes of prefixes of f ending at one of the blocks are of the forms:



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.

Note that the word f is not ultimately periodic, see Section 6.3. □

Example 2. 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 not any more 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. $\square$

Above examples are special cases of much deeper results. If in Example 1 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 proved in [MRS95]:

Theorem 3.6. *(i) Each φ^2-legal word is periodic.*
(ii) The Fibonacci word is $(\varphi^2 - \varepsilon)$-legal for any $\varepsilon > 0$.

Example 2 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 3.7. *(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 suprisingly, 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].

As the second final comment we state another similar result, the proof of which is related to the Critical Factorization Theorem. Consider an infinite word $w = a_1a_2\cdots$ with $a_i \in A$. We say that w contains a square *centered at position* i if there exists a $t \in [1,\ i]$ such that $a_{i-t+1}\cdots a_i$ is a prefix of $a_{i+1}a_{i+2}\cdots$. Then we have:

Theorem 3.8. *An infinite word $w = a_1a_2\cdots$ is ultimately periodic if and only if, for any large enough i, there exists a square centered at position i.*

For a proof of this and related results we refer to Chapter 8 in [Lot02], where also the optimality of the result is shown: no smaller amount than a square, i.e. of order 2, of centered repetition guarantees the ultimate periodicity.

4 Dimension properties

In this section we consider properties of words which can be called *dimension properties*. We approach this by considering a finite set X as a solution of a constant free equation. Natural questions to be asked are what can be said about X if it satisfies a nontrivial equation, or several "different" equations? And how many different equations it can satisfy?

Here an *equation* over the variables $\Xi = \{z_1,\ \ldots,\ z_n\}$ is just a pair $(u,\ v)$, usually written as $u = v$, of words of Ξ, and $X = (x_1,\ \ldots,\ x_n) \subseteq (A^*)^n$ is a *solution* if x_is substituted for z_is makes the equation to be an equality in A^*. More formally, a solution of the equation $u = v$ is a morphism $\varphi : \Xi \to A^*$ such that $\varphi(u) = \varphi(v)$. Actually, for simplicity, we sometimes overlook the fact that X must be ordered, and consider it only as a set.

We say that two systems of equations are *equivalent* if they have exactly the same solutions, and a system S is *independent* if it is not equivalent to any of its proper subsystems. We use the independency to formalize the notion that "equations are different". Note also that we have defined only constant free equations.

We start with the following simple example

Example 3. As an extension of the well known fact that two words satisfy a nontrivial relation if and only if they are powers of a common word, cf. Section 3, we consider the set $X = \{x,\ y,\ z\} \subseteq A^+$ of three nonempty words satisfying the equations

$$(1) \qquad x\alpha = y\beta \quad \text{and} \quad x\gamma = z\delta$$

with $\alpha,\ \beta,\ \gamma$ and δ in X^*. If two of the words are powers of a common word so is the third, by above. Consequently, assume that $x = yt$ for a nonempty word t. Now, substituting $x = yt$ to the equations in (1) we obtain two relations on words

$t,\ y,\ z \in A^+$. More specifically, the first equation comes into the form

$$u\alpha' = t\beta' \quad \text{with} \quad u \in \{y,\ z\} \quad \text{and} \quad \alpha',\ \beta' \in \{t,\ y,\ z\}^*,$$

and the second equation comes into the form

$$t\gamma' = z\delta' \quad \text{with} \quad \gamma',\ \delta' \in \{t,\ y,\ z\}^*.$$

Since $|tyz| < |xyz|$ induction applies showing that $t,\ x$ and y are powers of a common word, and so are $x,\ y$ and z. For the case $y = xt$ with $t \neq 1$, the reasoning is the same. Therefore, we have proved that nonempty words satisfying the pair (1) is necessarily *cyclic*, i.e. a subset of w^+ for some word w. □

4.1 Defect theorems

We continue by recalling so-called *defect theorem*: if a set of n words satisfies a nontrivial relation, then these words can be expressed as products of at most $n-1$ words. In other words, we can say that a nontrivial equation implies a *defect effect* of its solutions. Consequently, the defect theorem states a dimension type property for words.

Actually, there does not exist just one, but rather several, defect theorems witnessing the above defect effect. Namely, the set of $n-1$ words might be defined in different ways. In order to formalize those we recall that a submonoid M of A^* is called *right unitary* if its minimal generating set is a prefix code, cf. [BP85]. Then we define the *free hull* and the *prefix hull* of a finite set $X \subseteq A^+$ as

$$F(X) = \bigcap_{\substack{M \text{ is free monoid} \\ X \subseteq M}} M,$$

and

$$P(X) = \bigcap_{\substack{M \text{ is right unitary} \\ X \subseteq M}} M,$$

respectively. By the basic properties of these semigroups, $F(X)$ is free and $P(X)$ is right unitary. Hence, by definition, they are the smallest such semigroups containing X (and hence also X^*). The cardinalities of the minimal generating sets of $F(X)$ and $P(X)$ are called the *free rank* and the *prefix rank* of X, and those are denoted by $f(X)$ and $p(X)$, respectively. Before formulating several versions of the defect theorem we still define the *combinatorial rank* of X, in symbols $r(X)$, as the minimal cardinality of a set F such that $X \subseteq F^*$. Note that contrary to the above minimal generating sets F need not be unique. Now the defect theorem can be formulated, cf. e.g. [BPPR79], [Lot83] or [CK97]:

Theorem 4.1. *For any finite $X \subseteq A^+$ we have*

$$c(X) \leq p(X) \leq f(X) \leq \mathrm{card}(X),$$

and moreover the last inequality is proper if X satisfies a nontrivial equation.

The methods to compute these different ranks are discussed in [CK97]. Also the following example showing that all of these inequalities can be simultaneously proper is from [CK97].

Example 4. Consider the set $X = \{aa,\ aaba,\ bac,\ cbb,\ bbaa,\ baa\}$. It satisfies a nontrivial relation:

$$aa.\ bac.\ bbaa = aaba.\ cbb.\ aa$$

so that $f(X)$ should be < 6. By the well known method, see [BPPR79], one can compute that $X(f) = \{aa,\ ba,\ c,\ bb,\ baa\}$. But $(X(f))^+$ is not right unitary, so that, by methods in [CK97], one compute $X(p) = \{a,\ ba,\ c,\ bb\}$. Hence the cardinality of $X(p)$ is strictly smaller than that of $X(f)$. Finally, $X(p)$ is of combinatorial rank 3, yielding the following sequence of inequalities

$$3 = c(X) < p(X) < f(X) < \mathrm{card}(X) = 6.$$

□

The above shows that the notion of the defect effect is quite involved. In order to further emphasize this we recall that even sets generating isomorphic subsemigroups of A^+ might have nonisomorphic free hulls, cf. [HK86].

Example 4, and even more Example 3, might suggest that several different equations satisfied by X imply a *cumulative* defect effect for X. For example, two equalities would imply that $r(X)$ is at most $\mathrm{card}(X) - 2$. This indeed was the case in Example 3. However, words do not possess such a strong dimension property:

Example 5. The pair

$$xzy = yzx \quad \text{and} \quad xzzy = yzzx$$

has a solution $x = y = a$ and $z = b$ of (any) rank 2, and still the equations are independent. For example, $x = abba$, $y = a$ and $z = b$ is a solution of the latter but not of the former. □

We do not know whether there exist independent systems of equations with three unknowns, containing more than two equations and having a non-cyclic solution.

The above indicates that it is very difficult to impose a cumulative defect effect, and indeed there are very few results in that direction. Chapter 6 in [Lot02] gives one example showing that if X is a code, but not ω-code neither to the right nor to the left, then the rank of X is at most $\mathrm{card}(X)-2$. Here being not an ω-code means that some one-way infinite word can be decomposed in two different ways by X. Another complicated but still very special case is obtained in [KM02].

In the following lines, however, we show a simple but in many cases useful cumulative defect effect. In order to state it we need some terminology. We associate a finite set $X \subseteq A^+$ with a graph G_X as follows: the nodes of G_X are the elements of X, and there exists an edge between two nodes x and y if and only if

$$xX^+ \cap yX^+ \neq \emptyset.$$

G_X is called the *dependency graph* of X. Then the number of connected components of G_X, in symbols $c(G_X)$, gives an upper bound for the rank of X, cf. [HK86]:

Theorem 4.2. *Let $X \subseteq A^+$ be finite. Then we have*

$$c(X) \leq p(X) \leq c(G_X).$$

In particular, and this is typically the power of Theorem 4.2, if $c(G_X)$ is connected, then X is cyclic. Note that Example 3 is a simple special case of this result. Unlike in the usual defect theorems here it is crucial that all words of X are nonempty.

Defect theorems hold when X is not a code. A natural question is, can this assumption be weakened, for example, requiring only that X is not an ω-code, i.e. finite relations are replaced one-way infinite relations or even two-way infinite relations.

For one-way infinite relations the answer is easy: all results we stated, in particular Theorems 4.1 and 4.2, hold for one-way infinite relations as well. Even the proofs are basically the same. For two-way infinite words the situation is completely different. Strictly speaking no defect theorem holds, as shown by the set $X = \{ab,\ ba\}$ and two factorizations depicted as:

$$\cdots \ \text{ababababab} \ \cdots$$

However, one can prove the following defect theorem for two-way infinite words. In order to state it we need a few notions. Let $X \subseteq A^+$ be a finite set and w a two-way infinite word. Any decomposition of w into consecutive blocks of elements of X is an *X-factorization* of w, and two X-factorizations of w are *disjoint* if they do not match at any point of w. Finally, w is nonperiodic if it is not a two-way infinite repetition of a single word. We have, see [KMP02]:

Theorem 4.3. *Let $X \subseteq A^+$ be finite. If there exists a nonperiodic two-way infinite word with two disjoint X-factorizations, then*

$$c(X) < \mathrm{card}(X).$$

An interesting point here is that, unlike in all other known cases, the defect effect is witnessed only by the combinatorial rank, for a counterexample see [Ma02].

We conclude our discussion on defect effect by returning to Example 4. It shows that different notions of a rank of a finite set do not coincide, and this is unavoidable. Fortunately, however, the *rank* of an equation can be defined in the unique way. Let us define the rank of an equation as the maximal rank of its solutions. In theory, that would lead – in our considerations – to three different notions of the rank of an equation, namely the free rank, the prefix rank and the combinatorial rank. However, we have, see [CK97]:

Theorem 4.4. *Let $u = v$ be a constant-free equation over variables Ξ and A an alphabet such that* $\mathrm{card}(A) \geq \mathrm{card}(\Xi)$. *The following numbers coincide*

(i) the maximal of free ranks of solutions of $u = v$ in A^*,
(ii) the maximal of prefix ranks of solutions of $u = v$ in A^*,
(iii) the maximal of combinatorial ranks of solutions of $u = v$ in A^*.

The number specified in Theorem 4.4 is called the *rank* of an equation $u = v$.

4.2 Ehrenfeucht Compactness Property

In above we saw that words have certain dimension type properties, in fact even a rich theory in that direction. However, the dimension properties of words are rather weak. A natural question arises: How weak are they? Or more concretely, how large independent systems of equations can exist? These questions are partially answered in the following Ehrenfeucht Compactness Property:

Theorem 4.5. *Each system S of equations with a finite number of variables Ξ over free monoid A^* is equivalent to some of its finite subsets S_0.*

This compactness claim, conjectured by A. Ehrenfeucht – after studying the $D0L$ sequence equivalence problem, cf. [Ka93] – was proved simultaneously in [AL85] and [Gu86]. The proof is a marvelous example of the usefulness of the embedding (1) in Section 2.1. That allows to conclude the result of Theorem 4.5 from Hilbert's Basis Theorem for polynomials over commuting variables.

Interestingly, no bound for the cardinality of S_0, e.g. in the number of variables, is known. That leads to the following fundamental problem. Does there exist any such bound? Or would the bound 2^n, where n is the number of unknowns, be enough? The best known lower bounds for the maximal size of independent systems of equations are $\Theta(\mathrm{card}(\Xi)^3)$ and $\Theta(\mathrm{card}(\Xi)^4)$ in free semigroups and in free monoids, respectively, see [KP96]. It is interesting to note that there exist monoids, where Ehrenfeucht Compactness Property holds, but no bound for the maximal size of independent systems exists. An example of this is the variety of finitely generated commutative monoids, see again [KP96]. On the other hand, the property does not hold for all finitely generated monoids, see e.g. [HKP97] and [HK97].

5 Unavoidable regularities

In this section we consider several properties of words related to repetitions. It appears that some repetitions, or more generally some other regularities, are unavoidable, others are avoidable. For instance, it is easily checked that every binary word of length 4 contains a square. So squares are unavoidable over two letters. One may ask whether there exist arbitrarily long cube-free words over two letters. The answer is positive. Thus cubes are avoidable over two letters.

5.1 Power-free words

The *Thue-Morse* word defined earlier is cube-free. As we already mentioned in Section 1.2, the Thue-Morse word is even 2^+-free. This means that it is *overlap-free*, that is it does not contain any factor of the form $axaxa$, where a is a letter and x is a word.

As another example, we consider the infinite word z defined below, and we prove that it is cube-free. We choose to consider this infinite word rather than the Thue-Morse word because the proof is quite simple, although the argument of the proof is rather the same as for the Thue-Morse word. The cube-freeness of z is optimal in the sense that there are repetitions in z that are cubes up to one letter. One may say that z is 3^--free.

Consider the morphism

$$\zeta : \begin{array}{l} a \mapsto aba \\ b \mapsto abb \end{array}$$

Starting with a, it generates the infinite word

$$z = \zeta^\omega(a) = abaabbabaabaabbabbabaabbabaabaabbabaabaabbabbabaabbabb\cdots$$

Inspection shows that there are many factors of the form $uuu^\bullet$, that is cubes up to a final letter. For example, aab, bba, $abaabaabb$, $abbabbaba$ are almost cubes, and in fact all words $\zeta^n(aab)$ and $\zeta^n(bba)$ for $n \geq 1$ are almost cube factors of the infinite word z.

Fact 5.1. *The infinite word z is cube-free.*

Proof. We prove that ζ is a *cube-free* morphism in the sense that it preserves cube-free words: if w is cube-free, then $\zeta(w)$ is cube-free. This suffices to show that z is cube-free.

Assume the contrary, and consider a finite cube-free word w of minimal length n such that $\zeta(w)$ contains a cube uuu. The word $\zeta(w)$ has length $3n$, and it is a product of n blocks, each being either aba or abb.

Observe first that $|u|$ must be a multiple of 3. Indeed, the initial letter of u appears in w at three positions i, $j = i + |u|$, $k = i + 2|u|$ for some i, and if $|u| \not\equiv 0 \bmod 3$, then i, j, k take all possible values (mod 3). This means that the initial letter of u must appear in the images $\zeta(a) = aba$ and $\zeta(b) = abb$ at the first, the second and the third letter. However, b does not appear as an initial letter, and a does not appear as a middle letter. This proves the claim.

Next, observe that ζ is one-to-one. Indeed, the argument can be retrieved from the image of a letter because it is just the last letter of the image. This means that if $\zeta(w) = xuuuy$ for some x, y, we may assume that $|x| \equiv 0 \bmod 3$. We may even assume that x is the empty word because w was chosen to be minimal. But then, there is a unique word v such that $u = \zeta(v)$ and w starts with v^3. $\square$

As already mentioned earlier, the arguments in the proof are rather typical. Even the construction is typical: iterating a morphism f is a general tool for the construction of infinite words with predictable properties. Next, the property to be proved by induction (or by contradiction) uses the fact that one can infer a property for a word w from a property of the image $f(w)$ (in the previous example, of $\zeta(w)$). This step may be involved, because the morphism f may not always have such a simple form as the morphism ζ. In the present case, we proved that ζ is a *cube-free* morphisms, that is a morphisms that preserves cube-free words. Another, rather old, example is given e.g. in [Hal64]. It is the morphism

$$\begin{aligned} a &\mapsto abc \\ b &\mapsto ac \\ c &\mapsto b \end{aligned}$$

This morphism does not preserve square-free words, because the image of aba is $abcacabc$. However, iterating the morphism yields an infinite square-free word.

The question whether a morphism is k-free, that is whether it preserves k-free words, is not yet completely solved. The simplest, but not the easiest case is that of square-free morphisms. The final answer to this case was given in [Cro82]. For a nonerasing morphism h on A, set

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

Then the following holds:

Theorem 5.2. *Let $h : A^* \to B^*$ be a nonerasing morphism. If h preserves square-free words of length $K(h) = \max\left(3, 1 + \lceil (M(h) - 3)/m(h) \rceil\right)$, then h is square-free.*

As a special case, any *uniform* morphism h, i.e. such that the images of all letters have the same length, is square-free as soon as it preserves square-free words of length 3. In the ternary case, one has the following theorem [Cro82]:

Theorem 5.3. *A ternary endomorphism h is square-free if h preserves square-free words of length* 5.

No general result is known for cube-free morphisms. In the case of binary morphisms, one has the following bound [Kar83]:

Theorem 5.4. *A binary morphism h is cube-free if h preserves cube-free words of length* 10.

This bound was improved to the length 7 in [Lec85b]. In the same direction, we mention the following striking result concerning *power-free* morphisms, i.e. k-frees morphisms for all k [Lec85a].

Theorem 5.5. *A morphism h is power-free if h preserves square-free words and if the words $h(a^2)$ for a a letter, are cube-free.*

Deciding whether a given morphism is k-free, for a fixed $k \geq 3$, or even cube-free, is an amazing open problem.

Abelian repetitions, i.e. repetitions of commutatively equal blocks, constitute another type of interesting repetitions of words. Of course, every repetition of words is also Abelian repetitions, but not vice versa. It was shown in [Ev68] that Abelian squares can be avoided in infinite words if the alphabet contains 25 letters. In [Ple70] the result was improved to five letter alphabets, and finally the four letter case – known as an Erdös Problem – was solved in a complicated paper in [Ke92]. The required infinite word is obtained by iterating a uniform morphism, where the images of letters are of length 85! This amazing result has been completed by Carpi [Car98] who shows that the number of Abelian square-free words over 4 letters of given length grows exponentially. He also proves that there are "many" Abelian square-free morphisms, i.e., morphisms that map Abelian square-free words on Abelian square-free words. He shows that the monoid of Abelian square-free morphisms is not finitely generated. Similarly in [Ju72] it was shown that Abelian fifth powers can be avoided in a binary alphabet, and later [Dek79] filled the gaps: Abelian fourth powers can be avoided in a binary alphabet and cubes in a ternary alphabet. Moreover, all these results are optimal, for instance all binary words avoiding Abelian cubes are finite, in fact, of length at most 7.

5.2 Test sets and test words

A convenient framework is to define test sets for morphisms. A finite set $T \subset A^*$ is a *test set* for overlap-free (resp. square-free, cube-free) morphisms if every morphism $f : A^* \to B^*$ is overlap-free (resp. square-free, cube-free) as soon as $f(w)$ is overlap-free (resp. square-free, cube-free) for all $w \in T$. Theorem 5.4 can be rephrased as follows: the set of binary cube-free words of length 10 is a test set for binary cube-free morphisms. Test sets for overlap-free morphisms are characterized in [RS99]. Test sets for k-free morphisms with $k \geq 3$ are studied in [RW02b]. It has been shown that no test sets exist for the set of square-free morphisms over a 4-letter alphabet.

Another interesting notion is that of test word. A word $w \in A^*$ is a *test word* for overlap-free (resp. square-free, cube-free) morphisms if every morphism $f : A^* \to B^*$ is overlap-free (resp. square-free, cube-free) as soon as the word $f(w)$ is overlap-free (resp. square-free, cube-free). Thus w is a test word if $\{w\}$ is a test set. It has been shown [BS93] that *abbabaab* is a test word for overlap-free morphisms. We refer to [RS99] and [RW02b] for a detailed study of test words in various cases.

Overlap-free morphisms were already characterized by Thue [Th12]. The following is Satz 16 of his 1912 paper; the morphism μ was given in the introduction.

Theorem 5.6. *Let h be an overlap-free binary endomorphism. Then there is an integer n such that $h = \mu^n$ or $h = \pi \circ \mu^n$, where π is the endomorphism that exchanges the two letters of the alphabet.*

Clearly, this means that the monoid of binary overlap-free endomorphisms is finitely generated. The same does not hold for larger alphabets, neither for overlap-free morphisms, nor for k-power free morphisms: all these monoids of endomorphisms are not finitely generated [Ric02b]. Many other monoids of endomorphisms are finitely generated neither, such as the monoid of primitive morphisms (that is preserving primitive words) or of Lyndon morphisms (that is preserving Lyndon words), see [Ric02a].

5.3 Repetition threshold

More general repetitions can be considered as well. Thue himself called a word on n letters *irreducible* if two distinct occurrences of a nonempty factor are always separated by at least $n-2$ letters. Thus, irreducible means overlap-free if $n=2$, and square-free if $n=3$. A more general concept, first considered by F. Dejean [Dej72], is to require that the length of the word y separating two occurrences of x is bounded from below by the length of x times some factor. This is precisely what we called earlier a repetition: a word xyx, where x is nonempty, is a repetition of order k, where $k = |xyx|/|xy|$.

We have seen that every binary word of length 4 contains a square, and that there exist infinite binary overlap-free words (such as the Thue-Morse word): these words are 2^+-free.

A similar property holds for ternary words: every ternary word of length 39 contains a repetition of order 7/4, and there exists [Dej72] an infinite ternary word that has no repetition of order $> 7/4$. More precisely, we have:

Theorem 5.7. *The word generated by the endomorphism*

$$\begin{aligned} a &\mapsto abcacbcabcbacbcacba \\ b &\mapsto bcabacabcacbacabacb \\ c &\mapsto cabcbabcabacbabcbac \end{aligned}$$

has no repetitions of order $> 7/4$. So it is $(\frac{7}{4})^+$-free.

Call *repetition threshold* the smallest number $s(k)$ such that there exists an infinite word over k letters that has only repetitions of order less than or equal to $s(k)$. We know that $s(2) = 2$, $s(3) = 7/4$. It is conjectured in [Dej72] that $s(4) = 7/5$ and $s(k) = k/k-1$ for $k \geq 5$. We know that this conjecture is true up to 11, see [Pan84b] and [MO92], but the general case is still open.

5.4 Unavoidable patterns

Another topic on unavoidable regularities concerns *unavoidable patterns*. This is a generalization of the notion of square-freeness and k-power-freeness. Interestingly, this notion has a rather deep relation to universal algebra, see [BMT89]. We consider two alphabets, the first denoted by A as usual, and the second one, by E, called the *pattern alphabet*. Given a pattern $p \in E^*$, the *pattern language* associated to p on A is the set of all words $h(p)$, where h is a non-erasing morphism from E^* to A^*. A word w is said to *avoid* the pattern p, if no factor of w is in the pattern language of p. For example, consider the pattern $p = \alpha\alpha\beta\beta\alpha$, where α and β are letters. The word $1(011)(011)(0)(0)(011)1$ does not avoid p. On the contrary, 0000100010111 avoids p. A pattern is *avoidable* on A if there exists an infinite word on A that avoids p, otherwise it is unavoidable. A pattern is k-avoidable if it is avoidable on a k letter alphabet.

For instance, since there exist infinite square-free words over three letters, the pattern $\alpha\alpha$ is 3-avoidable. Also, the Thue-Morse infinite word avoids the patterns $\alpha\alpha\alpha$ and $\alpha\beta\alpha\beta\alpha$, the latter one corresponding to overlaps.

Clearly, the pattern $\alpha\beta\alpha$ is unavoidable. More generally, define the *Zimin words* Z_n as follows. Let α_n, for $n \geq 0$, be distinct letters in E. Set $Z_0 = \varepsilon$ and $Z_{n+1} = Z_n\alpha_n Z_n$ for $n \geq 1$.

Fact 5.8. *The Zimin words Z_n are all unavoidable.*

In some sense, these words are all unavoidable words. Indeed, say that a pattern p *divides* a pattern p' if p', viewed as an ordinary word, has a factor in the pattern language of p.

Theorem 5.9. *A pattern p is unavoidable if and only if it divides some Zimin word.*

There exists an algorithm to decide whether a pattern is avoidable. This has been given independently in almost the same terms by [BEM79] and [Zim82]. Let us just sketch the construction.

Let $p \in E^*$ be a pattern, and let P the set of variables occurring in p. The *adjacency graph* of p is the bipartite graph $G(p)$ with two copies of P as vertices, denoted P^L and P^R, and with an edge between λ^L and μ^R if and only if $\lambda\mu$ is a factor of p. For example, the adjacency graph of the pattern $p = \alpha\beta\alpha\gamma\beta\alpha$ has six vertices and four edges.

A subset F of P is *free* if there is no path in $G(p)$ from a vertex λ^L to a vertex μ^R with λ, μ in F. In our example, the free sets are $\{\alpha\}$ and $\{\beta\}$. Given a pattern p and a free set F for p, we *reduce* p to q by deleting in p all occurrences of the letters of F. In our example, $p = \alpha\beta\alpha\gamma\beta\alpha$ reduces to $q = \beta\gamma\beta$ by the free set $\{\alpha\}$, and q itself reduces to γ which itself reduces to the empty word. A pattern is *reducible* if it can be reduced to the empty word in a finite number of steps. The remarkable theorem that yields the algorithm is:

Theorem 5.10. *A pattern p is avoidable if and only if it is reducible.*

As shown by the pattern $\alpha\alpha$, an avoidable pattern is not necessarily avoidable on two letters. In other terms, the same pattern may be 2-unavoidable, but k-avoidable for some larger k. The *avoidability index* $\mu(p)$ of a pattern p is the smallest integer k such that p is k-avoidable, or ∞ if p is unavoidable. Contrary to the previous theorem, there is no known algorithm to compute the avoidability index of a given pattern. Even for short patterns, the exact value of $\mu(p)$ may be unknown. For instance, it is not known whether the value of $\mu(\alpha\alpha\beta\beta\gamma\gamma)$ is 2 or 3, although there is some experimental evidence that the index is 2, see Chapter 3 in [Lot02]. However, the proof of Theorem 5.10, if analyzed carefully, provides an upper bound on the avoidability index of a pattern.

Theorem 5.11. *Let p be a pattern on k symbols. If p is avoidable, then $\mu(p) \leq 2k+4$.*

The above bound is probably far from being optimal. In fact, it is quite difficult to find patterns with high avoidability index. A pattern which has index 4 is

$$p = \alpha\beta\zeta_1\beta\gamma\zeta_2\gamma\alpha\zeta_3\beta\alpha\zeta_4\alpha\gamma$$

over a pattern alphabet of 7 symbols. No pattern of index 5 is known, and one may ask whether such a pattern exists. Many results on avoidable and unavoidable patterns are reported in Chapter 3 of [Lot02], while [Cur93] is a source of several open problems. Recent results concerning complexity of the reduction algorithm appear in [Hei02a], [Hei02b] and [Hei02c].

5.5 Shirshov's theorem

An unavoidable regularity is a property of words that can be observed on any word, provided it is sufficiently long. Several of these unavoidable regularities exist. The most famous are Ramsey's theorem, van der Waerden's theorem and Shirshov's theorem. We just state here Ramsey's and van der Waerden's theorems in the perspective of words, and then discuss Shirshov's theorem and some of its consequences.

Let A be a finite alphabet, S a set and $k \geq 2$ an integer. A map $f : A^* \to S$ is called *k-ramseyan* if there exists an integer $L(f, k)$ such that any word w of length at least $L(f, k)$ admits a factor u of the form $u = w_1 \cdots w_k$, with $w_1, \ldots, w_k$ nonempty words, such that

$$f(w_i \cdots w_j) = f(w_{i'} \cdots w_{j'})$$

for all $1 \leq i \leq j \leq k$, $1 \leq i' \leq j' \leq k$. Further f is *ramseyan* if it is k-ramseyan for some k. Ramsey's theorem, see [GRS90], can be stated in a number of combinatorial structures. On words it can be stated as the following unavoidable regularity:

Theorem 5.12. (Ramsey's theorem) *Every map $f : A^* \to S$ into a finite set S is ramseyan.*

Given a word $w = a_1 \cdots a_n$, where $a_1, \ldots, a_n$ are letters, a *cadence* of w of order r is a sequence $0 < t_1 < \cdots < t_r \leq n$ such that $a_{t_1} = a_{t_2} = \cdots = a_{t_r}$. A cadence is *arithmetic* if the integers are in arithmetic progression, that is if $t_{i+1} - t_i = t_i - t_{i-1}$ for $1 < i < r$. Now,we can formulate the second unavoidable regularity of words:

Theorem 5.13. (van der Waerden's theorem) *Let A be an alphabet with k letters. For any positive integer n, there exists an integer $W(k, n)$ such that any word w over A of length at least $W(k, n)$ contains an arithmetic cadence of order n.*

Ramsey's and van der Waerden's theorems are rather well known. However, the evaluation of the integers $W(k, n)$, as well as the function L, is very difficult and not at all completely known, see [GRS90]. We refer to [Lot83] and to [dLV97] for proofs of our above formulations and further discussions.

We now turn to Shirshov's theorem, which is much less known. It was proved by Shirshov in connection with so-called polynomial identities (see [Lot83] for a discussion of this issue). The unavoidable regularity concerned by this result has also other far reaching applications.

Let A be a totally ordered alphabet, and let $<$ denote the *lexicographic* order on A^* induced by the order on the alphabet. Denote by S_n the symmetric group on n elements. A sequence $(u_1, u_2, \ldots, u_n)$ of nonempty words is called an *n-division* of the word $u = u_1 u_2 \cdots u_n$ if, for any nontrivial permutation σ of S_n, one has

$$u_1 u_2 \cdots u_n > u_{\sigma(1)} u_{\sigma(2)} \cdots u_{\sigma(n)}$$

Example 6. Consider the binary alphabet $A = \{a, b\}$ ordered by setting $a < b$. The word $w = ababbaba$ is 3-divided by the sequence $(ababb, ab, a)$. One can also verify that this is the only 3-division of w and that there is no 4-division. □

Now we can formulate our third unavoidable regularity:

Theorem 5.14. (Shirshov's theorem) *Let A be an alphabet with k letters. For any positive integers n, k, there exists an integer $N(k, p, n)$ such that any word w over A of length at least $N(k, p, n)$ contains as a factor an n-divided word or a pth power.*

Again, the integers $N(k, p, n)$ are quite difficult to compute. It is also quite interesting that this result admits an extension to infinite words, in contrast to van der Waerden's theorem. An infinite word s is called *ω-divided* if it can be factorized into an infinite product $s = s_1 s_2 \cdots s_n \cdots$ of nonempty words such that $(s_1, \ldots, s_n)$ is an n-division for all $n > 0$. Now, by denoting by $F(s)$ the set of all (finite) factors of an infinite word s we can formulate:

Theorem 5.15. *For any infinite word t over A, there exists an infinite word s such that $F(s) \subset F(t)$ and s is ultimately periodic or s is ω-divided.*

This is, in fact, a consequence of another structural result of infinite words. In order to state this, we recall that a *Lyndon word* is a word w that is primitive and that is smaller (for the lexicographic order) than all of its conjugates, i.e., such that whenever $w = xy$ with x, y nonempty, then $xy < yx$. Then we have:

Theorem 5.16. *For any infinite word t over A, there exists an infinite word s such that $F(s) \subset F(t)$ and s is a product of an infinite sequence of non increasing Lyndon words: $s = \ell_1\ell_2 \cdots \ell_n \cdots$, with $\ell_n \geq \ell_{n+1}$ and ℓ_n is a Lyndon word for all n.*

As we already said Shirshov's theorem has quite interesting consequences. We consider here one of those. A semigroup S is called *periodic* if every subsemigroup of S generated by a single element is finite. A sequence $(s_1, \ldots, s_n)$ of n elements of S is called *permutable* if there exists a nontrivial permutation σ of S_n such that $s_1 \cdots s_n = s_{\sigma(1)} \cdots s_{\sigma(n)}$. The semigroup is called *n-permutable* if any sequence of n elements of S is permutable, and S is called *permutable* if S is n-permutable for some n. A language L is called *periodic* (resp. *permutable*) if its syntactic semigroup is periodic (resp. permutable). A striking characterization of rational languages can now be stated:

Theorem 5.17. *A language L is rational if and only if it is periodic and permutable.*

The proof is based on Shirshov's theorem, see [RR85], [dLV97]. Also the proofs of Theorems 5.15 and 5.16, as well as more related results can be found in [dLV97].

5.6 Unavoidable sets of words

We conclude this section with still another notion of an unavoidability, namely the notion of *unavoidable set* of words. A set X of words over A is unavoidable if any long enough word over A has a factor in X, that is if $A^* - A^*XA^*$ is finite.

For example, the set $\{a, bb\}$ is unavoidable over $\{a, b\}$. Indeed, any word of length 2 either contains an a or contains bb. Since a superset of an unavoidable set is again unavoidable, it is natural to consider *minimal* unavoidable sets. Every minimal unavoidable set is finite. Indeed, if X is unavoidable, then let d be the maximal length of the words in the finite set $A^* - A^*XA^*$. Let Z be the set of words in X of length at most $d+1$. Every word of length $d+1$ has a factor in X which actually is in Z. Thus Z is unavoidable and finite.

In fact, there exists an algorithm to decide whether a finite set X of words in unavoidable, and the structure of unavoidable sets of words is rather well understood, see Chapter 1 of [Lot02] for a general exposition and for references. A recent result concerns unavoidable sets of words of the same length, i.e. *uniform* unavoidable sets. For $k, q \geq 1$, let $c(k, q)$ be the number of conjugacy classes of words of length k on q letters. An unavoidable set of words of length k on q symbols clearly has at least $c(k, q)$ elements. It is quite remarkable that the converse holds as well, see [CHP02]:

Theorem 5.18. *For any $k, q \geq 1$, there exists an unavoidable set of words of length k on q letters having exactly $c(k, q)$ elements.*

As an illustration assume that $k = 3$ and $q = 2$, so that $c(k, q) = 4$. Now the sets $\{aaa, aab, aba, bbb\}$ and $\{aaa, aab, bba, bbb\}$, for example, are avoidable, the former because it does not intersect all the conjugacy classes and the latter since it avoids

$(ab)^\omega$. On the other hand, the set $X = \{aaa, aab, bab, bbb\}$ is unavoidable. Indeed, infinite words containing no letter a or the factor aa clearly contains a factor from X, and all the other words contain either bab or bbb.

6 Complexity

Given a set X of words over an alphabet A the complexity function of X is the function p_X defined by $p_X(n) = \mathrm{Card}(X \cap A^n)$. We are interested here in the (*subword*) *complexity* (one should say *factor* complexity!) of a finite or an infinite word u. It is the function p_u which is the complexity function for the set $F(u)$ of factors of u, thus $p_u(n) = \mathrm{Card}(F(u) \cap A^n)$ is just the number of factors of length n in u. The study of infinite words of given complexity has revealed a large set of surprising results. Moreover, there are strong relations to number theory and symbolic dynamics. Subword complexity was studied, in relation with languages generated by D0L-systems, already in [ELR75].

6.1 Subword complexity of finite words

The subword complexity of a finite word w is simply the function p_w such that $p_w(n)$ is the number of factors of length n of w. Of course, $p_w(0) = p_w(|w|) = 1$, and $p_w(n) = 0$ for $n > |w|$. The description of the shape of the function p_w uses the parameter G_w defined as follows: G_w is the maximal length of a *repeated* factor, that is a factor that appears at least twice in w. For instance, let $w = abccacbccabaab$. The word $bcca$ is a repeated factor of maximal length, so $G_w = 4$. The shape of the function p_w is given by the following result, see [dL99], [CdL01b], [CdL01a] and [LS01]:

Theorem 6.1. *Let w be a word over at least two letters. There is an integer m_w such that the function $p_w(n)$ behaves as follows:*

(i) it is strictly increasing for $0 \le n \le m_w$,
(ii) it is constant for $m_w \le n \le G_w + 1$,
(iii) it is decreasing for $G_w+1 \le n \le |w|$, and more precisely $p_w(n+1) = p_w(n)-1$ for $n = G_w + 1, \ldots, |w|$.

Consider again the word $w = abccacbccabaab$. One gets $p_w(1) = 3$, $p_w(2) = 8$, $p_w(3) = 10$, $p_w(4) = 11$, and this is the maximum because $G_w = 4$. Then it decreases by 1 at each step: $p_w(5) = 10 \ldots$. In our example, the integer m_w is equal to G_w.

A precise description of the parameter m_w is missing. On the contrary, the parameter G_w is strongly related to other structural parameters of the word w. Denote by R_w the minimal number such that there is no factor x of w of length R_w that has two right extensions in w, i.e., such that xa and xb are factors of w for distinct letters a and b. Next, let K_w be the length of the shortest suffix of w which is an unrepeated factor of w. Then it can be shown, see [CdL01b], that $G_w + 1 = \max(R_w, K_w)$. The parameter G_w has the following interesting property, proved in [CdL01b]:

Theorem 6.2. *A word w is completely determined by its set of factors of length at most $G_w + 2$.*

For example, the word $w = abccacbccabaab$, of length 14, is entirely determined by its factors of length 6. Many combinatorial facts, and properties of distribution of these parameters, are given in [CdL02a] and [CdL02b].

Related to Theorem 6.2 one can ask several questions when a given word is uniquely determined by its factors or *sparse subwords*. Besides these two variants one can also ask the same when multiplicities are taken into account. This leads to four different problems, see [Ma00] or Chapter 6 in [Lot83]. Among those is a question, sometimes referred to as Milner's Problem, asking what is the minimal length of sparse subwords with multiplicities which defines the word uniquely.

6.2 Subword complexity of infinite words

Let us start with two examples. For the infinite word $01(10)^\omega$, the complexity function is easily computed. One gets $p(0) = 1$, $p(1) = 2$, $p(2) = 3$, $p(n) = 4$ for $n \geq 4$. More generally, it is easily seen that an ultimately periodic infinite word has a complexity function that is bounded, and therefore is ultimately constant because a complexity function is never decreasing. On the contrary, consider the infinite binary word c known as the Champarnowne word:

$$c = 0110111001011101111000\cdots$$

This is obtained by concatenating the binary expansions of all positive integers in order. Clearly, every binary word is a factor of c and thus $p_c(n) = 2^n$ for all $n \geq 0$. These examples show that both extreme behaviours are possible for complexity functions. However, there are gaps in the growth of complexity functions. The first such result was given in [MH40,CH73]:

Theorem 6.3. *Let x be an infinite word. The following are equivalent:*

(i) x is ultimately periodic,
(ii) $p_x(n) = p_x(n+1)$ for some n,
(iii) $p_x(n) < n + k - 1$ for some $n \geq 1$, where k is the number of letters appearing in x,
(iv) $p_x(n)$ is bounded.

Proof. The implications (1) $\Rightarrow$ (4) $\Rightarrow$ (3) $\Rightarrow$ (2) are easy. For the remaining implication (2) $\Rightarrow$ (1) we observe that each factor of length n in x can be extended in exactly one manner to a factor of length $n + 1$. This means that each occurrence of a given factor of length n is always followed by the same letter in x. Consider any factor u of length n that appears twice in x, and denote by y the word that separates the two occurrences, so that uyu is a factor of x. Since the letters that follow u are determined by u, this means that in fact $uyuy$ is a factor of x, and that $(uy)^\omega$ is a suffix of x. $\square$

This result shows a "gap" for complexity functions: either a function p is ultimately constant, or $p(n) \geq n+1$ for all n. So an infinite aperiodic word, that is a word that is not ultimately periodic, cannot have a complexity function bounded by n. It appears that aperiodic infinite words of complexity $p(n) = n+1$ indeed exist. These words are called *Sturmian words*.

For another gap of complexity functions see [Cas97b].

6.3 Sturmian words

A Sturmian word x is always a binary word because $p_x(1) = 2$ and $p_x(1)$ is the number of letters appearing in x. Every factor of length n can be extended to the right into a factor of length $n+1$. Since $p_x(n+1) = 1 + p_x(n)$, this extension is unique for all factors up to one, and this last factor has two extensions. A factor with two extension is called *right special*. More precisely, call the (*right*) *degree* of a word u the number of distinct letters a such that ua is a factor of x. Then, in the binary case, a right special factor is a word u of degree 2. A Sturmian word has exactly one special factor of each length. We give an example:

Fact 6.4. *The Fibonacci word* $f = 0100101001001010010100 \cdots$ *is Sturmian.*

To check this, we observe that f is the fixed-point of the morphism φ given in Section 1.3. Thus f is a product of words 0 and 01. In particular, 11 is not a factor of f and $p_f(2) = 3$. Also, the word 000 is not a factor of f since otherwise it is a factor of the image of a factor of f that must contain 11.

Next, we show that for all finite words x, neither $0x0$ nor $1x1$ is a factor of f. This is clear if x is the empty word or is a single letter. Arguing by induction, assume that both $0x0$ or $1x1$ are factors of f. Then x starts and ends with 0, so $x = 0y0$ for some word y. Since $10y01$ is a factor of f, there exists a factor z of f such that $\varphi(z) = 0y$. Moreover, $00y00 = \varphi(1z1)$ and $010y01 = \varphi(0z0)$ are factors of f, showing that $0z0$ and $1z1$ are factors, a contradiction.

We now prove that f has at most one special factor of each length. Assume that u and v are right special factors of f of the same length, and let x be the longest common suffix of u and v. Then the four words $0x0$, $0x1$, $1x0$, $1x1$ are factors of f which contradicts our previous observation.

We finally prove that f has at least one special factor of each length. For this, it suffices to prove that f is aperiodic. Recall that f is the limit of the sequence of finite Fibonacci words defined by $f_0 = a$, $f_1 = ab$, and $f_{n+2} = f_{n+1}f_n$. It is easily shown that the

$$\frac{|f_n|_a}{|f_n|} \to \frac{1}{\tau}, \qquad n \to \infty,$$

with $\tau = (1+\sqrt{5})/2$, whereas in a ultimate periodic word, this limit is a rational number. □

We now give two other descriptions of Sturmian words, namely as balanced words and as mechanical words. Given two binary words u and v of the same length, the *balance* of u and v is the number $b(u,v) = ||u|_1 - |v|_1|$, that is the absolute value of the difference of the number of occurrences of the letter 1 in the words u and v. Since u and v have the same length, one could also have defined this number by taking the number of 0's instead of the number of 1's. As an example, for 01001 and 11001, the balance is 1. An infinite word x is *balanced* if the balance of any two factors of x of the same length is at most 1. Intuitively, a balanced word cannot have big differences in factors. In particular, a balanced word cannot contain simultaneously the factors $0u0$ and $1u1$. Moreover, it can be shown that a balanced word x has a *slope*, that is that the limit $\lim_{n\to\infty} b_n/n$ exists, where b_n is the number of 1's in the prefix of length n. As an example, the slope of the Fibonacci word is $1/\varphi^2$, where $\varphi = (1+\sqrt{5})/2$.

Another notion strongly related to Sturmian words is of more arithmetical nature. Given two real numbers α and ρ with $0 \le \alpha \le 1$, we define two infinite words $s_{\alpha,\rho} = s_{\alpha,\rho}(0)s_{\alpha,\rho}(1)\cdots$ and $s'_{\alpha,\rho} = s'_{\alpha,\rho}(0)s'_{\alpha,\rho}(1)\cdots$ by

$$\begin{aligned} s_{\alpha,\rho}(n) &= \lfloor \alpha(n+1)+\rho \rfloor - \lfloor \alpha n + \rho \rfloor \\ s'_{\alpha,\rho}(n) &= \lceil \alpha(n+1)+\rho \rceil - \lceil \alpha n + \rho \rceil \end{aligned} \qquad \text{for } n > 0.$$

The word $s_{\alpha,\rho}$ is the *lower mechanical word* and $s'_{\alpha,\rho}$ is the *upper mechanical word* with *slope* α and *intercept* ρ. It is clear that we may assume $0 \le \rho \le 1$. If α is irrational, $s_{\alpha,\rho}$ and $s'_{\alpha,\rho}$ differ by at most one factor of length 2.

The terminology stems from the following graphical interpretation. Consider the straight line defined by the equation $y = \alpha x + \rho$. The points with integer coordinates just below this line are $P_n = (n, \lfloor \alpha n + \rho \rfloor)$. Two consecutive points P_n and P_{n+1} are joined by a straight line segment that is horizontal if $s_{\alpha,\rho}(n) = 0$ and diagonal if $s_{\alpha,\rho}(n) = 1$. The same observation holds for the points located just above the line.

A special case deserves to be considered separately, namely when $0 < \alpha < 1$ and $\rho = 0$. In this case, $s_{\alpha,0}(0) = 0$, $s'_{\alpha,0}(0) = 1$, and if α is irrational

$$s_{\alpha,0} = 0c_\alpha, \quad s'_{\alpha,0} = 1c_\alpha,$$

where the infinite word c_α is called the *characteristic* word of α.

Mechanical words can be interpreted in several other ways. One is as cutting sequences, and is as follows. Consider again a straight line $y = \beta x + \rho$, for some $\beta > 0$ not restricted to be less than 1, and ρ not restricted to be positive. Consider the intersections of this line with the lines of the grid with nonnegative integer coordinates. We get a sequence of intersection points. Writing a 0 for each vertical intersection point and a 1 for each horizontal intersection point, we obtain an infinite word $K_{\beta,\rho}$ that is called the *cutting sequence*. Then

$$K_{\beta,\rho} = s_{\beta/(1+\beta),\rho/(1+\beta)}$$

Indeed, the transformation $(x, y) \mapsto (x+y, x)$ of the plane maps the line $y = \beta x + \rho$ to $y = \beta/(1+\beta)x + \rho/(1+\beta)$. Thus, cutting sequences are just another formulation of mechanical words, see also [CMPS93] for a more detailed discussion.

Mechanical words can also be generated by rotations. Let $0 < \alpha < 1$. The *rotation* of angle α is the mapping $R = R_\alpha$ from $[0, 1[$ into itself defined by

$$R(z) = \{z + \alpha\}$$

Iterating R, one gets $R^n(\rho) = \{n\alpha + \rho\}$. Thus, defining a partition of $[0, 1[$ by

$$I_0 = [0, 1-\alpha[, \quad I_1 = [1-\alpha, 1[\,,$$

one gets

$$s_{\alpha,\rho}(n) = \begin{cases} 0 & \text{if } R^n(\rho) \in I_0 \\ 1 & \text{if } R^n(\rho) \in I_1 \end{cases}.$$

The three properties are related by the following theorem [MH40]:

Theorem 6.5. *Let s be an infinite word. The following are equivalent:*

(i) s is Sturmian;
(ii) s is balanced and aperiodic;
(iii) s is mechanical with an irrational slope.

A proof can be found in Chapter 2 of [Lot02]. As a example, the Fibonacci word f is indeed the lower mechanical word with slope and intercept equal to $1/\varphi^2$. There is a special class of Sturmian words called *characteristic* Sturmian words. These are the words where the intercept equals the slope. Each characteristic Sturmian word s has a description as the limit of a sequence s_n of finite words, quite as the Fibonacci word f is the limit of the finite words f_n. The recurrence relation is slightly more complicated. It has the form

$$s_n = s_{n-1}^{d_n} s_{n-2} \qquad \text{for } n \geq 1$$

with $s_{-1} = 1$, $s_0 = 0$, and where $d_1, d_2, \ldots$ is a sequence of integers with $d_1 \geq 0$ and $d_n > 0$ for $n > 1$. This sequence is related to the slope α of s by the fact that $[0, 1+d_1, d_2, \ldots]$ is the continued fraction expansion of α. In the case of the Fibonacci word, the continued fraction expansion of $1/\varphi^2$ is indeed $[0, 2, 1, 1, \ldots]$ and the d_n are all equal to 1. The sequence s_n is called the standard sequence, and every word that appears in a standard sequence is a standard word.

A beautiful theorem, proved in [CMPS93], describes those irrational numbers α for which the standard Sturmian word can be generated by iterating a morphism. We give here the description that follows from a complement given in [All98]

Theorem 6.6. *Let $0 < \alpha < 1$ be an irrational number. The characteristic Sturmian word of slope α is a fixed-point of some nontrivial morphism if and only if α is a quadratic irrational number such that $1/\bar{\alpha} < 1$*

In this theorem, $\bar{\alpha}$ denotes the *conjugate* of the number α, that is the other root of its minimal polynomial. This result shows a relation between a combinatorial property of words and some arithmetical counterpart.

Sturmian words have a tremendous amount of combinatorial or arithemetic properties. An account can be found in Chapter 2 of [Lot02] and in [PF02].

6.4 Episturmian words

There have been several attempts to extend the notion of Sturmian words. By the definition, Sturmian words are binary, and so relaxations of the constraints on the complexity function were looked for. It appears that a very good extension is rather related to what is called *Arnoux-Rauzy*, or more generally *episturmian words*. We start with a typical word of this family called *Tribonacci* word. This word is defined, like the Fibonacci word, as the limit of a sequence of words defined by a recurrence relation:

$$t_{n+3} = t_{n+2}t_{n+1}t_n, \qquad t_0 = 0,\ t_1 = 01,\ t_2 = 0102$$

So the Tribonacci word is

$$t = 0102010010201010201\cdots$$

It is also defined as the fixed-point of the mophism

$$\begin{aligned} 0 &\mapsto 01 \\ 1 &\mapsto 02 \\ 2 &\mapsto 0 \end{aligned}$$

It is not very difficult to see that $p_t(n) = 2n + 1$ for $n \geq 0$. This means that for each n, there must be 2 additional extensions to the right of factors of length n. To do this, there are two possibilities: either there are 2 distinct right special factors, each of which has degree 2, or there is just one right special factor which has degree 3. It can be checked that the second property holds for the Tribonacci word. The right special factors are ε, 0, 10, 010, 2010,... . As in the case of the Fibonacci words, the right special factors are the reversal of the prefixes of t.

Infinite words that have exactly one right special factor of each length, and each having degree 3, where introduced by Arnoux and Rauzy in [AR91], and are therefore called Arnoux-Rauzy words, or AR-words for short. This terminology has been extended to words over k-letters, see e.g. [WZ01], by requiring that for each length n, there is a unique right special factor with degree k. This definition is relaxed in [DJP01] and [JP02] to allow right special factors of degree at most k. Infinite words with this property are called *episturmian* words, and AR-words are called strict episturmian [JP02]. Strict episturmian binary words are Sturmian, whereas episturmian binary words may be ultimately periodic, so they are the mechanical words.

6.5 Hierarchies of complexities

One of the first papers on subword (factor) complexity is [ELR75]. In this paper, it is shown that the subword complexity of a D0L language is bounded by cn^2 (resp. $cn \log n$, cn) if the morphism that generates the languages is arbitrary (resp. growing, uniform). This result was extended in [Pan84a]:

Theorem 6.7. *The subword complexity of an infinite word generated by iterating a morphism is of one of the following types: $\Theta(n)$, $\Theta(n \log n)$, $\Theta(n \log n \log n)$, $\Theta(n^2)$, or $\Theta(1)$.*

Each of the complexity classes corresponds to a class of morphisms. We just give some examples. The morphism

$$\begin{aligned} a &\mapsto ab \\ b &\mapsto bc \\ c &\mapsto c \end{aligned}$$

generates the infinite word

$$abbcbc^2bc^3 \cdots bc^n \cdots$$

having quadratic complexity. Consider next the morphism

$$\begin{aligned} a &\mapsto abc \\ b &\mapsto bb \\ c &\mapsto ccc \end{aligned}$$

It generates the infinite word

$$abcb^2c^3b^4c^9 \cdots b^{2^n}c^{3^n} \cdots$$

which has complexity $\Theta(n \log n)$. Consider finally the morphism

$$\begin{aligned} a &\mapsto abab \\ b &\mapsto bb \end{aligned}$$

Starting in a, one gets the infinite word

$$abab^3abab^7ababbbabab^{15} \cdots$$

which has complexity $\Theta(n \log \log n)$.

6.6 Subword complexity and transcendence

Consider the Thue-Morse word t and the Fibonacci word f written over the alphabet $\{0, 1\}$:

$$\begin{aligned} t &= 0110100110010110100101100110 1001 \cdots \\ f &= 0100101001001010 01010 \cdots \end{aligned}$$

One may consider the infinite words as binary expansions of real numbers in the interval $[0,1]$, and ask whether these numbers are algebraic or transcendental. We address this and related questions in this section.

It is well known that the expansion in an integral base $b \geq 2$ of a rational number is ultimately periodic. This means that the subword complexity of a rational number is very low. On the other hand, it has been conjectured by Borel, see [All00] for a more detailed discussion, that the infinite word representing the expansion in base b of an algebraic number is "normal", which means that all words appear as factors, and that all factors of length n appear with the same frequency $1/b^n$. This conjecture is presently far from being proved, and even the following, much weaker conjecture, seems to be very difficult: let x be the expansion in base b of a real number r. If $p_x(n) < b^n$ for all n, then r is either a rational or a transcendental number, again see [All00].

However, there exist several results in number theory related to the fact that algebraic numbers cannot be well approximated by rational numbers. In other terms, if a number is too well approximated, then it is either rational or transcendental. Two of these results are the famous theorem of Roth and a refinement of this, which is the theorem of Ridout, see [All00] for references. It is remarkable that these results can be translated into combinatorial properties of infinite words [FM97]:

Theorem 6.8. *Let x be the expansion in base $b \geq 2$ of a real number $r \in [0,1]$. If there exist real numbers $\alpha > 0$ and $\varepsilon > 0$ such that the word x has infinitely many prefixes of the form $uv^{2+\varepsilon}$, with $|u| \leq \alpha|v|$, then r is either a rational or a transcendental number.*

The first condition means that x must contain arbitrarily long powers of exponent strictly greater than 2, and the second condition requires that these powers appear "not too far" from the beginning of the word. In fact, the present statement is the translation of Ridout's theorem, and when u is assumed to be the empty word, that is when it is required that x starts with infinitely many $2+\varepsilon$ powers, this is Roth's theorem. If the infinite word x has the prefix $uv^{2+\varepsilon}$, then $x = uv^{2+\varepsilon}y$ for some infinite word y, and x is "close" to the ultimately periodic word uv^{ω}, so the real r is "close" to the rational number represented by uv^{ω}. For the complete proof of Theorem 6.8, however, some further work is needed.

Theorem 6.8 gives a powerful tool to prove that real numbers are transcendental. As the first example, we consider the number whose binary expansion is the Fibonacci word f. Recall that f is the limit of the sequence of finite words f_n defined by $f_0 = 0$, $f_1 = 01$, $f_{n+2} = f_{n+1}f_n$. It is easily checked that f starts with infinitely many $(2 + 1/\varphi^2 - \varepsilon)$-powers. Indeed, for $n \geq 4$, the infinite word f starts with $f_n f_n g_{n-2}$, where g_n is equal to f_n up to the last two letters. For instance, f starts with (01001010)(01001010)(01010). Thus the number whose binary expansion is the Fibonacci word is indeed transcendental.

The Fibonacci word is a special case of a general result of [FM97]:

Theorem 6.9. *Let x be the expansion in base 2 of a real number $r \in [0,1]$. If x is a Sturmian word, then r is transcendental.*

It is shown also in [FM97] that the same result holds for above mentioned Arnoux-Rauzy words.

We know that the Thue-Morse word is overlap-free, so it does not contain $2+\varepsilon$-powers. Thus Theorem 6.9 does not apply. However, the transcendence of the number represented by the Thue-Morse word has been proved directly by Mahler and Dekking, see again [All00] for references.

Recently, the case of infinite words that are fixed-points of morphisms has been considered. An endomorphism h of A^* is called *primitive*, if there exists an integer m such that each word $h^m(a)$, for $a \in A$, contains at least one occurrence of every letter in A. For instance, the Fibonacci morphism is primitive. The result is the following, see [AZ98] or already [FM97] for primitive morphisms:

Theorem 6.10. *If the binary expansion of a real number is the fixed-point of a non-trivial morphism that is either primitive or uniform, then this number is either rational or transcendental.*

We mention that it is decidable whether the fixed-point of a non-trivial morphism is an ultimately periodic word, see [Pan86] and [HL86], or [Séé88] for the complete characterization of the binary case. Thus, it is decidable whether the number in Theorem 6.10 is rational or transcendental.

6.7 Descriptive and computational complexity

In addition to subword complexity the classification of infinite words can be based on many other measures reflecting different aspects of complexity. Here we consider briefly two such directions, namely *descriptive* and *computational* complexities. The former one measures how complicated mechanisms are needed to generate infinite words. In general, this is not obvious to formalize, however, concrete examples can be given to illustrate it, see [CK94] for different such mechanisms. Computational complexity, in turn, asks how much computational resources are needed to generate a particular infinite word, say by a Turing machine, see e.g. [HKL94].

The most frequently used method of defining an infinite word is, as we have already seen, that of iterating a morphism. More complicated methods are, for example, iterating two or more morphisms periodically, iterating a sequential transducer and so on. The famous *Kolakoski word* can be obtained by iterating two morphisms periodically, see Section 8 and [CKL92], but it is not known whether it can be obtained as a morphic image of an infinite word obtained by iterating a morphism. By iterating sequential transducers one can define even much more complicated words as shown below.

Example 7. For a natural number n let $\mathrm{bin}_r(n)$ denote the reverse binary representation of n. Then, clearly, a sequential transducer can compute

$$\mathrm{bin}_r(n) \;\mapsto\; \mathrm{bin}_r(n+1),$$

and consequently the infinite word

$$S\,\mathrm{bin}_r(1)\#\mathrm{bin}_r(2)\#\cdots\#\mathrm{bin}_r(n)\cdots$$

can be generated by iterating a sequential transducer. □

Example 8. Let $\mathcal{M}$ be a Turing machine and

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

the sequence of configurations of its computation on an input word w. Then, similarly to above, the word

$$Sw_0\#w_1\#w_2\cdots$$

can be generated by iterating a sequential transducer. Consequently, problems on such words are typically undecidable. □

Neither of the above words is, in general, definable by iterating a morphism. This follows, for example, the subword complexity considerations, cf. Theorem 6.7.

Infinite words definable by iterating a morphism are simple, not only descriptively, but also computationally. In order to formalize this we consider Turing machines as generators of infinite words: The machine has one one-way write only tape for the generation and several two-way working tapes for the computations. At the beginning all the tapes are empty, and the *space complexity* of the machine is the number of cells used in working tapes for printing the nth letter of the word. It follows from standard considerations in the complexity theory that the first complexity class after the trivial class $\mathcal{O}(1)$ is that of $\mathcal{O}(\log n)$, for details see [HKL94]. The former one corresponds to ultimately periodic infinite words.

Now, we can state:

Theorem 6.11. *Each infinite word obtained by iterating a morphismis is of space complexity* $\mathcal{O}(\log n)$.

Of course, the class of $\mathcal{O}(\log n)$ is much larger than that of infinite words obtained by iterated morphisms. Indeed, the question whether or not all words obtained by iterating sequential transducers are in this class is equivalent to a classical open problem in the complexity theory, see [DM02].

7 From words to finite sets of words

So far we have been interested in properties of single words, finite or infinite, and not sets of words. Even in the defect theorems, although those deal with finite sets of words, the crucial point is that those sets are considered as solutions of single (or several) word equations. Similarly, avoidable sets of words avoided a single word. A natural question is to ask whether at least some of the basic properties of words can be extended to finite sets of words. Among the most natural such question is the one asking what can be said about the commutation of (finite) sets of words. This and some related problems have been considered in several recent papers, cf. [Ka01] for a survey.

7.1 Conway's Problem

More than 30 years ago Conway asked in his book [Con71] whether the maximal set commuting with a given rational set R is rational as well. Such a maximal set, which clearly exists – it is the union of all sets commuting with R – is called the *centralizer* of R, in symbols $\mathcal{C}(R)$. It is straightforward to see that the centralizer is a semigroup or a monoid depending on whether the considerations are in A^+ or A^*. Apart from some trivial observations the problems seems to be equally hard, but not known to be related, in these two cases. We concentrate to the semigroup case here.

Conway's Problem asks a very natural and simple looking question on rational languages. Surprisingly, however, the answer is not known even for finite languages, and in fact even a much simpler looking question, namely whether the centralizer of a finite X is recursive, or equivalently recursively enumerable, is unanswered. Here, the equivalence comes from the fact that the complement of the centralizer of even a recursive set, is recursively enumerable, as is rather is to see. The following simple example from [CKO02] reveals something of the complexity of the commutation of finite languages.

Example 9. Consider the four element set $X = \{a,\ ab,\ ba,\ bb\}$. As is straightforward to compute it commutes with the sets

$$Y_1 = X \cup X^2 \cup \{bab,\ bbb\}$$

and

$$Y_2 = \{a,\ b\}^+\{b\},$$

the latter one being the centralizer of X (in A^+). Note that Y_2 is not only rational but also finitely generated, since $Y_2 = (X \cup \{bab,\ bbb\})^+$. □

Obviously, X^+ is always a subset of the centralizer, or more generally, any semigroup generated by a *root* of X, i.e. a set Y such that $X = Y^i$ for some $i \geq 1$, is a subset of the centralizer of X. Let us call a root of X *minimal* if it is not a proper root of any set, and denote it by $\rho_m(X)$. A minimal root of a set need not be unique, for example, even a unary set like $\{a^i \mid 0 \leq i \leq 30, i \neq 1, 8, 11, 23\}$ can possess two different (minimal) square roots, see [CK00]. If X possesses just one minimal root we call it the *primitive* root and denote it by $\rho(X)$. In the case of codes or nonperiodic two or three element sets the primitive root exist. We call a set *primitive* if $\rho(X) = X$.

We can state the following simple estimates for the centralizer: for any $X \subseteq A^+$, we have

$$\rho_m(X)^+ \subseteq \mathcal{C}(X) \subseteq \mathrm{Pref}(X^+) \cap \mathrm{Suf}(X^+),$$

where $\rho_m(X)$ is any minimal root of X. The left inclusion is obvious and the right one is easy to conclude. We say that $\mathcal{C}(X)$ is *trivial* if it coincides with $\rho_m(X)^+$ for some minimal root of X. Being trivial does not mean that the triviality is easy to check. Indeed, we do not know whether the question "Is $\mathcal{C}(X) = X^+$?" is decidable even for finite sets X.

An affirmative answer of Conway's Problem has been shown in a number of special cases, see e.g. [CKO02], [KP02] and [Pe02] for a general treatment. The first nontrivial result was that of [Ra89] solving the problem for prefix sets. This together with a recent extension of the above special cases, see [KLP03], are summarized as follows:

Theorem 7.1. *(i) For each three element set X the centralizer of X is either w^+, for a primitive word w, or X^+ depending on whether X is periodic or not.*

(ii) For each prefix set X the centralizer of X is $\rho(X)^+$.

In particular, in both of these cases Conway's Problem has an affirmative answer.

It is a challenge to find a short proof for Part (ii) in Theorem 7.1. In what follows we give a short proof, see [KP01], for a special case of part (i) in Theorem 7.1, namely for the case when X is binary. This is based on so-called branching point approach. For $X \subset A^+$, we say that $w \in \mathrm{Pref}(X^+)$ is a *branching* point if w can be extended in $\mathrm{Pref}(X^+)$ at least by two different letters

Theorem 7.2. *Let $X = \{x,\ y\} \subseteq A^+$. Then the centralizer of X is either w^+, for a primitive word w, or X^+ depending on whether or not $xy = yx$.*

Proof. The case when x and y commute is a simple application of the defect theorem for two element sets.

So assume that $xy \neq yx$. Let z be the maximal common prefix of xy and yx, in symbols $z = xy \wedge yx$. Hence, as seen in Lemma 3.3, $|z| \leq |xy| - 2$. Further let us call a two element set *marked* if the first symbols of the words are different.

Claim I. For a marked Z we have $\mathcal{C}(Z) = Z^+$.

Claim I follows directly from the following three facts: First, the centralizer is a subset of $\mathrm{Pref}(Z^+)$. Second the set Z has branching points only in Z^+. This, indeed, is true since Z is marked. Third, $\mathcal{C}(Z)$ is a semigroup, so that if $z' \in \mathcal{C}(Z) \subseteq \mathrm{Pref}(Z^+)$ so do $z'x$ and $z'y$. Consequently, $\mathcal{C}(Z)$ can contain only branching points of Z.

Now, we are done if X is marked. If not, say $X \subseteq aA^*$ with $a \in A$, we set $X' = a^{-1}Xa$, and show

Claim II. $\mathcal{C}(X) = a\mathcal{C}(X')a^{-1}$.

In order to prove Claim II, we first write $\mathcal{C}(X) = aY$, with $y \subseteq A^*$. This is possible since $X\mathcal{C}(X) = \mathcal{C}(X)X$. Recalling that $aX' = Xa$ we can compute

$$X\mathcal{C}(X)a = XaYa = aX'Ya$$

and

$$\mathcal{C}(X)Xa = aYXa = aYaX'.$$

Since the left hand sides are equal so are the right ones.This means that Ya commutes with X' and hence, by the maximality of the centralizer, we obtain that $Ya \subseteq \mathcal{C}(X')$. This, however, can be written in the form $\mathcal{C}(X) = aY \subseteq aC(X')a^{-1}$.

Starting from the fact that $\mathcal{C}(X') = Y'a$ for some $Y' \subseteq A^+$, we conclude similarly that $aY' \subseteq \mathcal{C}(X)$ or, equivalently, that $\mathcal{C}(X') \subseteq a^{-1}\mathcal{C}(X)a$. This can be rewritten in the form $aC(X')a^{-1} \subseteq \mathcal{C}(X)$. Hence, Claim II follows.

Now, the existence of z and Claims I and II yield the theorem. □

Above theorems deserve a few remarks. First, the equality of two sets in Claim II can not be done in a usual way showing that an element from the left is also on the right hand side, and vice versa. Instead sets of words has to be considered. Second, although in three element case the reduction of the proof of Theorem 7.2 does not lead to a marked instance, the branching point approach is a cornerstone in the proof of Part (i) in Theorem 7.1 as well. Third, as shown by Example 4, no similar result for the centralizer of a four element set is possible. Finally, the above results lead to a nice characterization of all sets commuting with a given set of the considered types, as seen in the next section.

7.2 Characterization of commuting sets

In order to formulate results of this section, we say that a set $X \subseteq A^+$ possesses *BTC-property* if the following implication holds:

$$\forall Y \subseteq A^+ : \text{ if } \ XY = YX, \ \text{ then there exist } \ I, \ J \subseteq \mathbf{N} \ \text{ and } \ V \subseteq A^+$$

such that

$$X = \cup_{i \in I} V^i \quad \text{and} \quad Y = \cup_{j \in J} V^j. \tag{1}$$

Note that in the case of nonperiodic two or three element sets the condition (1) reduces to

$$Y = \cup_{i \in I} X^i. \tag{2}$$

The above property is similar to that characterizing commutation of polynomials and formal power series over noncommuting variables, cf. [Ber69], [Coh78] and Chapter 9 in [Lot02]. The abbreviation BTC comes from there: Bergman type of characterization. We could also say that sets satisfying BTC are of *word type*.

We have:

Theorem 7.3. *Any two element set possesses BTC. Consequently, any set commuting with $X = \{x, \ y\}$, where $xy \neq yx$, is of the form $\cup_{i \in I} X^i$ for some $I \subseteq \mathbf{N}$.*

Proof. It is enough to show that, if $XY = YX$ and $w \in Y \cap X^n$, then $X^n \subseteq Y$. This, however, follows directly from the fact that X is a code: Indeed, assume that $y = x_1 \cdots x_n$, with $x_i \in X$. Then, for any $x_{x+1} \in X$, we can write

$$x_1 \cdots x_n x_{n+1} = x_1' w' \quad \text{with} \quad w' \in Y \subseteq \mathcal{C}(X) = X^+.$$

Consequently, $x_1 = x_1'$ and $w' = x_n \cdots x_{n+1} \in Y$. Repeating the argument we obtain that any word in X^n is in Y. □

The similar reasoning can be used to obtain corresponding result for all three element sets. In the this case the considerations are, however, much more involved – due to the fact that such a set need not be a code.

Theorem 7.4. *(i) Any three element set possesses BTC. In particular, any set commuting with a nonperiodic three element set is a union of powers of X.*

(ii) Any prefix set possesses BTC. Consequently, any set commuting with a prefix set X is a union of powers of the primitive root of X.

We recall that Part (i) is optimal in the sense that the BTC property does not hold for four element sets – as shown again by Example 9.

As a summary of the above we can say that the commutation of languages is well understood in above special – but still very interesting – cases. Common to all of those is that the centralizer is always trivial, although to show that is not always easy. In general, tools to attack Conway's Problem seems to be completely lacking. Results of the next section might explain, at least intuitively, why the problem looks so difficult. In Section 8 several open problems on commutation are formulated.

7.3 Undecidability results for finite sets

In this section we show that some simple questions on finite sets of words are actually undecidable.

In [HIKS02] it was proved, based on Theorem 7.3:

Theorem 7.5. *It is undecidable whether a given two element set and a given context-free language commute.*

In order to formulate another undecidability result on finite sets of words we need some terminology. We call a morphism from A^* into the monoid of finite languages a *finite substitution.* Following [CS78] we say that two mappings f, g defined on A^* *are equivalent* on language $L \subseteq A^*$ if they map all the words of L into the same element. In the case of finite substitutions the requirement is that, for any $w \in L$, the finite languages $f(w)$ and $h(w)$ coincide. In [KL02] the following undecidability result was proved.

Theorem 7.6. *It is undecidable whether two given finite substitutions f and g are equivalent on language ab^*c, i.e. whether or not*

$$f(ab^ic) = g(ab^ic) \quad \textit{for all} \quad i \geq 0.$$

Despite of being an interesting, and maybe also surprising, undecidability result as such Theorem 7.6 has also a few nice consequences. We formulate here just one of those, for the other see [KL02].

We say that two finite sets X and Y of words are *conjugates* if there exists a set Z such that

$$XZ = ZY.$$

Actually, the above splits into two notions, since the set Z might be required to be finite or allowed to be arbitrary. Problems on conjugacy was considered in [CKM01]. However, the following natural question was not answered: is it decidable whether two given finite sets are conjugates? Even the case where X and Y are biprefix sets seems not to be trivial, although decidable, cf. [CK03]. Note that the above question can be viewed as a very special question on equations over the monoid of (finite) languages: the equation contains only one unknown and two constants. The equation

$$\{a,\ ab,\ abb,\ ba,\ babb\}Z = Z\{a,\ ba,\ bba,\ bbba\}$$

is an instance of such an equation. It happens to have a solution $Z = \{a,\ ab\}$.

The above motivates to propose a general problem. Is it decidable whether a given equation with constants has a solution in the monoid of finite languages?

Very little seem to be known about this important problem. One thing separating the finite set case from the word case is a consequence of our Theorem 7.6. In order to formulate it we need some terminology. We consider infinite systems of equations with a finite number of unknowns Ξ and constants C. We say that a system $\{u_i = v_i | i \in \mathbf{N}\}$ is *rational* if the set $\{(u_i,\ v_i) | i \in \mathbf{N}\}$ is a rational subset of $(\Xi \cup C)^2$. In other words there is a finite transducer mapping each left hand side of the equations to the corresponding right hand side. Now, the corollary is:

Theorem 7.7. *It is undecidable whether a given rational system of equations over the monoid of finite sets of words has a solution.*

Interestingly, the satisfiability problem for rational systems of word equations is decidable. Indeed, such a system is computably equivalent to one of its finite subsystems, and each finite system of equations can be encoded into a single equation. The first reduction here is an effective variant of Ehrenfeucht Compactness Property, that is Theorem 4.5, cf. [HK97], and the second reduction is easy, see e.g. [Hm71]. So the satisfiability problem for rational word equations is reduced to that of a single equation, i.e. to the seminal result of Makanin.

8 Open Problems

In this final section we formulate several open problems, many of those being already discussed in previous sections. We start with two decidability questions of matrices.

Problem 1. Given a finite set of $n \times n$ matrices over integers with $n \geq 3$, is it decidable whether or not the identity matrice is obtainable as a product of these matrices.

Problem 2. Given a finite set of 2×2 matrices over natural numbers is it decidable whether the multiplicative semigroup generated by these matrices is free.

Of course, the decidability of Problem 1 may depend on the parameter n, for $n = 2$ the problem is decidable, see [CK02]. As an evidence of an intriguing nature of Problem 2 we recall from [CHK99] that we do not even know whether the concrete matrices

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \quad B = \begin{pmatrix} 3 & 5 \\ 0 & 5 \end{pmatrix}$$

generate a free semigroup. Accordingly the problem is open even in the case $n = 2$.

As a consequence of the critical factorization theorem we know that two-way infinite noncyclic word w cannot have $n + 1$ pairwise disjoint factorizations of a set X of n words. On the other hand, Theorem 4.3 shows that if such a word w has two disjoint X-factorizations then the combinatorial rank of X is at most $\mathrm{card}(X) - 1$. This motivates the following:

Problem 3. Let X be a finite set of words, and w a two-way noncyclic infinite word. What is the relation between the number of pairwise disjoint X-factorizations of w, in symbols $df(X)$, and the combinatorial rank of X, in symbols $r_c(X)$? More concretely, is it true that $r_c(X) \leq \mathrm{card}(X) - df(X) + 1$?

A result in [KM02] proves the inequality of Problem 3 in a special case when X is a prefix set and $df(X) = 3$.

Two other open problems on dimension properties concern independent systems of equations.

Problem 4. Does there exist an independent system of three constant-free equations with three unknowns possessing a noncyclic solution?

Problem 5. (i) Does there exist a function $f : \mathbf{N} \to \mathbf{N}$ such that any independent system of constant-free equations with n unknowns contains at most $f(n)$ equations?

(ii) Does the requirement of part (i) hold for the function $f(n) = 2^n$?

As shown in [KP96] the Ehrenfeucht Compactness Property might hold in some monoids, like in Abelian monoids, without giving any bound asked in Problem 5. We recall that some lower bounds for the function $f(n)$ are also given in [KP96]

Next we formulate a few problems on the avoidability

Problem 6. Is it decidable whether a given morphism $h : A^* \to B^*$ is k-free for a fixed integer $k \geq 3$?

We recall that repetition threshold on an n letter alphabet is the smallest number $s(n)$ such that infinite words on n letters can contain only repetitions of order less than or equal to $s(n)$.

Problem 7. Is the repetition threshold for $n \geq 12$ equal to $n/n-1$?

Another crucial questions on avoidable patterns are:

Problem 8. Does there exist a pattern which is avoidable in a five letter alphabet but unavoidable in a four letter alphabet?

Problem 9. Is the size of the smallest alphabet where a give pattern is avoidable, i.e. its avoidability index, algorithmically computable?

In order to formulate our next problem we say that u is a *sparse subword* of a word w if it is obtained from w by deleting some occurrences of letters, i.e. there exist words $u_1, \ldots, u_n$ and $v_0, \ldots, v_n$ such that $u = u_1 \cdots u_n$ and $w = v_0 u_1 v_1 \cdots u_n v_n$. Then, for $n > 0$, *n spectrum* of a word w is the set of all sparse subwords of w with multiplicities and of length at most n. Hence, the n spectrum is a formal polynomial of degree n. For example, the 2 spectrum of the word $aabba$ is $3a + 2b + 2aa + 4ab + 2ba + bb$. In algebra these polynomials are often defined as results of Magnus Transform of a word. Now we formulate, for more see [Ma00]:

Problem 10. What is the minimal n such that the n spectrum of a word w determines w uniquely?

One of the amazing sequences of words is that of *Kolakoski word k*. It can be defined as a self-reading sequence by the following rule: it consists of blocks of 1s and 2s each block being of length either 1 or 2, and the length of the ith block is equal to the ith element of the word. Consequently, the word k starts as

$$k = 2211212212211\ldots .$$

Descriptionally the Kolakoski word is quite easy: it can be defined by iterating a sequential transducer, or even by iterating periodically two morphisms, namely

$$h_1 : \begin{cases} 1 \mapsto 2, \\ 2 \mapsto 22 \end{cases} \qquad \text{and} \qquad h_2 : \begin{cases} 1 \mapsto 1, \\ 2 \mapsto 11 \end{cases} .$$

There is a larger literature on the Kolakoski sequence, e.g. [Dek95], [Car93], [BL03], [Ch93] and [Le94].

However, very little is known about the Kolakoski word. For example, it is not known whether the numbers of 1s and 2s are asymptotically equal. Connected to this tutorial we can state:

Problem 11. Is the complexity of Kolakoski word at most quadratic?

Problem 12. Is Kolakoski word obtainable as a morphic image of a fixed-point of an iterated morphism?

Clearly, a negative answer to Problem 11 would give that for Problem 12, due to Theorem 6.7. It is also relatively easy to show that k is not obtainable as a fixed-point of an iterated morphism, see e.g. [CKL92].

We conclude with open problems on finite (and rational) sets of words, in particular, on commutation of those. We formulate several variants of Conway's Problem:

Problem 13. Is the centralizer of a given rational set a) rational, b) recursive?

Problem 14. Is the centralizer of a given finite set a) finitely generated, b) rational, c) recursive?

As other problems on finite sets of words we state

Problem 15. Is it decidable whether two finite sets X and $Y \subseteq A^+$ are conjugates, i.e. there exists a set Z such that $XZ = ZY$?

Problem 16. Is it decidable whether a given equation with constants has a solution in the monoid of finite languages?

Note that Problem 15 contains actually two variants depending on whether Z is allowed to be arbitrary or finite. In Problem 16 the constants are, of course, elements of the monoid, i.e. finite languages.

Acknowledgements

Many thanks to J.-P. Allouche, J. Cassaigne, A. Lepistö, G. Richomme and P. Séébold for their careful reading of parts of a preliminary version of the paper. In addition the authors are grateful to A. Lepistö for his help to complete the presentation.

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.

[All98] C. Allauzen, Une caractérisation simple des nombres de Sturm, *J. Théo. Nombres Bordeaux* 10, 237–241, 1998.

[All00] J.-P. Allouche, Nouveaux résultats de transcendance de réels à développement non aléatoire, *Gazette des Mathématiciens* 84, 19–34, 2000.

[AS03] J.-P. Allouche and J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge University Press, 2003.

[AZ98] J.-P. Allouche and L. Q. Zamboni, Algebraic irrational binary numbers cannot be fixed points of non-trivial constant length or primitive morphisms, *J. Number Theory* 69, 119–124, 1998.

[AR91] P. Arnoux and G. Rauzy, Représentation géométrique de suites de complexité $2n+1$, *Bull. Soc. Math. France* 119, 199–215, 1991.

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

[BMT89] K. A. Baker, G. F. McNulty, and W. Taylor, Growth problems for avoidable words, *Theoret. Comput. Sci.* 69, 319–345, 1989.

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

[Ber69] G. Bergman, Centralizers in free associative algebras, *Transactions of the American Mathematical Society* 137, 327–344, 1969.

[Be95] J. Berstel, Axel Thue's papers on repetition in words: a translation, *Publications de Laboratoire de Combinatoire et d'Informatique Mathématique*, Université du Québec á Montréal 20, 1995.

[BB99] J. Berstel and L. Boasson, Partial words and a theorem of Fine and Wilf, WORDS (Rouen, 1997), *Theoret. Comput. Sci.* 218, 135–141, 1999.

[BP85] J. Berstel and D. Perrin, *Theory of Codes*, Academic Press, 1985.

[BPPR79] J. Berstel, D. Perrin, J.-F. Perrot, and A. Restivo, Sur le théorème du défaut, *J. Algebra* 60, 169–180, 1979.

[BS93] J. Berstel and P. Séébold, A characterization of overlap-free morphisms, *Discr. Appl. Math.* 46, 275–281, 1993.

[BBG90] A. Bertoni, D. Bruschi, and M. Goldwurm, Ranking and formal power series, In: *Algorithms and complexity (Rome, 1990)*, 159–171, World Sci. Publishing, Teaneck, NJ, 1990.

[BL03] S. Brlek and A. Ladouceur, A note on differentiable palindromes, *Theoret. Comput. Sci.* (to appear).

[Car93] A. Carpi, Repetitions in the Kolakovski sequence, *Bull. EATCS* 50, 194–196, 1993.

[Car98] A. Carpi, On the number of Abelian square-free words on four letters, *Discr. Appl. Math.* 81, 155–167, 1998.

[CdL01a] A. Carpi and A. de Luca, Periodic-like words, periodicity, and boxes, *Acta Inform.* 37, 597–618, 2001.

[CdL01b] A. Carpi and A. de Luca, Words and special factors, *Theoret. Comput. Sci.* 259, 145–182, 2001.

[CdL02a] A. Carpi and A. de Luca, On the distribution of characteristic parameters of words, *Theoret. Inform. Appl.* 36, 67–96, 2002.

[CdL02b] A. Carpi and A. de Luca, On the distribution of characteristic parameters of words II, *Theoret. Inform. Appl.* 36, 97–127, 2002.

[Cas97a] J. Cassaigne, On a conjecture of J. Shallit, In: P. Degano, R. Gorrieri, and A. Marchetti-Spaccamela (Eds.), *Automata, Languages and Programming*, LNCS 1256, 693–704, Springer-Verlag, 1997.

[Cas97b] J. Cassaigne, Complexité et facteurs spéciaux, *Bull. Belg. Math. Soc.* 4, 76–88, 1997.

[CHK99] J. Cassaigne, T. Harju, and J. Karhumäki, On the undecidability of freeness of matrix semigroups, *Intern. J. Alg. & Comp.* 9, 295–305, 1999.

[CK03] J. Cassaigne and J. Karhumäki, manuscript in preparation, 2003.

[CKM01] J. Cassaigne, J. Karhumäki, and J. Manuch, On conjugacy of languages, *Theor. Inform. Appl.* 35, 535–550, 2001.

[CMR99] M. G. Castelli, F. Mignosi, and A. Restivo, Fine and Wilf's theorem for three periods and a generalization of Sturmian words, *Theoret. Comput. Sci.* 218, 83–94, 1999.

[CV78] Y. Césari and M. Vincent, Une caractérisation des mots périodiques, *C. R. Acad. Sci. Paris (Série A)* 286, 1175–1177, 1978.

[CHP02] J.-M. Champarnaud, G. Hansel, and D. Perrin, Unavoidable sets of constant length, Technical Report 06, Institut Gaspard Monge, Université de Marne-la-Vallée, 2002.

[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.

[CK00] C. Choffrut and J. Karhumäki, On Fatou properties of rational languages, In: C. Martin-Vide and V. Mitrana (eds.), *Where Mathematics, Computer Science, Linguistics and Biology Meet*, Kluwer, Dordrecht 2000.

[CK02] C. Choffrut and J. Karhumäki, Some decision problems on integer matrices, manuscript 2002.

[CKO02] C. Choffrut, J. Karhumäki, and N. Ollinger, The commutation of finite sets: a challenging problem, *Theoret. Comput. Sci.* 273, 69–79, 2002.

[Ch93] Vašek Chvátal, Notes on the Kolakoski sequence, *DIMACS technical report* 93-84, 1993.

[Coh78] P. M. Cohn, Centralisateurs dans les corps libres, In: J. Berstel (ed.), *Series Formelles*, Paris, 45–54, 1978.

[Con71] J. H. Conway, *Regular Algebra and Finite Machines*, Chapman Hall, 1971.

[CH73] E. M. Coven and G. A. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* 7, 138–153, 1973.

[CMPS93] D. Crisp, W. Moran, A. Pollington, and P. Shiue, Substitution invariant cutting sequences, *J. Théor. Nombres Bordeaux* 5, 123–137, 1993.

[Cro82] M. Crochemore, Sharp characterizations of square-free morphisms, *Theoret. Comput. Sci.* 18, 221–226, 1982.

[CHL01] M. Crochemore, C. Hancart, and T. Lecroq, *Algorithmique du texte*, Vuibert, 2001.

[CP91] M. Crochemore and D. Perrin, Two-way string matching, *J. Assoc. Comput. Mach.* 38(3), 651–675, 1991.

[CR94] M. Crochemore and W. Rytter, *Text algorithms*, Oxford University Press, 1994.

[CR02] M. Crochemore and W. Rytter, *Jewels in Stringology*, World Scientific, 2002.

[CF77] K. Culik II and I. Fris, The decidability of the equivalence problem for $D0L$-systems, *Information and Control* 35, 20–39, 1977.

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

[CKL92] K. Culik II, J. Karhumäki, and A. Lepistö, Alternating iteration of morphisms and Kolakovski sequence, In: G. Rozenberg and A. Salomaa (eds.), *Lindermayer Systems, Impacts on Theoretical Computer Science, Computer Graphics and Developmental Biology*, 93–106, Springer-Verlag, 1992.

[CS78] K. Culik II and A. Salomaa, On the decidability of homomorphism equivalence for languages, *J. Comput. Syst. Sci.* 17, 163–175, 1978.

[Cur93] J. D. Currie, Open problems in pattern avoidance, *Amer. Math. Monthly* 100, 790–793, 1993.

[Dej72] F. Dejean, Sur un théorème de Thue, *J. Combin. Th. A* 13, 90–99, 1972.

[Dek79] F. M. Dekking, Strongly non-repetitive sequences and progression-free set, *J. Combin. Th. A* 27, 181–185, 1979.

[Dek95] F. M. Dekking, What is the long range order in the Kolakoski sequence?, In: R. V. Moody and J. Patera (eds.), *The Mathematics of Long-Range Aperiodic Order (Waterloo, ON, 1995)*, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 489, 115–125, Kluwer Acad. Publ., Dordrecht, 1997.

[dL99] A. de Luca, On the combinatorics of finite words, *Theoret. Comput. Sci.* 218, 13–39, 1999.

[dLV97] A. de Luca and S. Varricchio, Regularity and Finiteness conditions, In: G. Rozenberg and A. Salomaa (eds.), *Handbook of Formal Languages I*, Springer-Verlag, 737–810, 1997.

[dLV99] A. de Luca and S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages*, Springer-Verlag, 1999.

[DJP01] X. Droubay, J. Justin, and G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, *Theoret. Comput. Sci.* 255, 539–553, 2001.

[DM02] P. Ďuriš and J. Manuch, On the computational complexity of infinite words, *Theoret. Comput. Sci.* (to appear).

[Duv79] J.-P. Duval, Périodes et répétitions des mots du monoïde libre, *Theoret. Comput. Sci.* 9(1), 17–26, 1979.

[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.

[Ev68] A. A. Evdokimov, Strongly asymmetric sequences generated by a finite number of symbols, *Dokl. Akad. Nauk. SSSR* 179, 1268–1271, 1968 (English transl. *Soviet. Math. Dokl.* 9, 536–539, 1968).

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

[FW65] N. J. Fine and H. S. Wilf, Uniqueness theorem for periodic functions, *Proc. Amer. Math. Soc.* 16, 109–114, 1965.

[FZC94] P. Flajolet, P. Zimmerman, and B. Van Cutsem, A calculus for the random generation of labelled combinatorial structures, *Theoret. Comput. Sci.* 132, 1–35, 1994.

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

[GRS90] R. L. Graham, B. L. Rothschild, and J. H. Spencer, *Ramsay Theory*, Second Edition, John Wiley & Sons, 1990.

[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.

[GO81] J. L. Guibas and A. Odlyzko, Periods in strings, *J. Combin. Th. A* 30, 19–42, 1981.

[Gus97] D. Gusfield, *Algorithms on strings, trees and sequences*, Cambridge University Press, 1997.

[Hal64] M. Hall, Generators and relations in groups – the Burnside problem, In: T. L. Saaty (ed.), *Lectures on Modern Mathematics*, vol. 2, 42–92. Wiley, 1964.

[HIKS02] T. Harju, O. Ibarra, J. Karhumäki, and A. Salomaa, Some decision problems concerning semilinearity and commutation, *J. Comput. Syst. Sci.* 65, 278–294, 2002.

[HK86] T. Harju and J. Karhumäki, On the defect theorem and simplifiability, *Semigroups Forum* 33, 199–217, 1986.

[HK97] T. Harju and J. Karhumäki, Morphisms, In: G. Rozenberg and A. Salomaa (eds.), *Handbook of Formal Languages*, 439–510, Springer-Verlag, 1997.

[HKP97] T. Harju, J. Karhumäki, and W. Plandowski, Compactness of systems of equations in semigroups, *Intern. J. Alg. & Comp.*, 457–470, 1997.

[HL86] T. Harju and M. Linna, On the periodicity of morphisms in free monoids, *Theoret. Inform. Appl.* 20, 47–54, 1986.

[HN02] T. Harju and D. Nowotka, Density of critical factorizations, *Theoret. Inform. Appl.*, 2002.

[Hei02a] C. E. Heitsch, Exact distribution of deletion sizes for unavoidable strings, *Fundamenta Informaticae*, 2002.

[Hei02b] C. E. Heitsch, Insufficiency of the four known necessary conditions on string avoidability, *J. Algorithms*, 2002.

[Hei02c] C. E. Heitsch, Intractability of the reductive decision procedure for unavoidability testing, a special case of generalized pattern matching, *J. Algorithms*, 2002.

[Hm71] Y. I. Hmelevskii, Equations in free semigroups, *Proc. Stoklov Inst. Math.* 107, 1971 (English transl. *Amer. Math. Soc. Translations*, 1976).

[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.

[Ju72] J. Justin, Characterization of the repetitive commutative semigroups, *J. Algebra* 21, 87–90, 1972.

[Jus00] J. Justin, On a paper by Castelli, Mignosi, Restivo, *Theoret. Inform. Appl.* 34, 373–377, 2000.

[JP02] J. Justin and G. Pirillo, Episturmian words and episturmian morphisms, *Theoret. Comput. Sci.* 276, 281–313, 2002.

[Kar83] J. Karhumäki, On cube-free ω-words generated by binary morphisms, *Discr. Appl. Math.* 5, 279–297, 1983.

[Ka93] J. Karhumäki, The impact of the $D0L$ problem, In: G. Rozenberg and A. Salomaa (eds.), *Current Trends in Theoretical Computer Science, Essays and Tutorials*, 586–594, World Scientific, 1993.

[Ka01] J. Karhumäki, Combinatorial and computational problems on finite sets of words, In: M. Margenstern and Y. Rogozhin (Eds.), *Machines, Computations, and Universality*, LNCS 2055, 69–81, Springer-Verlag, 2001.

[KLP03] J. Karhumäki, A. Latteux, and I. Petre, The commutation with codes and ternary sets of words, Proceedings of STACS'03, LNCS (to appear).

[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.

[KL02] J. Karhumäki and L. P. Lisovik, The equivalence problem of finite substitutions an ab^*c, with applications, In: P. Widmayer, F. Triguero,

R. Morales, M. Hennessy, S. Eidenbenz, and R. Conejo (Eds.), *Automata, Languages and Programming*, LNCS 2380, 812–820, Springer-Verlag, 2002.

[KM02] J. Karhumäki and J. Manuch, Multiple factorizations of words and defect effect, *Theoret. Comput. Sci.* 273, 81–97, 2002.

[KMP02] J. Karhumäki, J. Manuch, and W. Plandowski, A defect theorem for two-way infinite words, *Theoret. Comput. Sci.* (to appear).

[KP01] J. Karhumäki and I. Petre, Conway's Problem and the commutation of languages, *Bull. of EACTS* 74, 171–177, 2001.

[KP02] J. Karhumäki and I. Petre, Conway's Problem for three word sets, *Theoret. Comput. Sci.* 289, 705–725, 2002.

[KP96] J. Karhumäki and W. Plandowski, On the size of independent systems of equations in semigroups, *Theoret. Comput. Sci.* 168, 105–119, 1996.

[KMPS92] L. Kari, S. Marcus, Gh. Paun, and A. Salomaa, In the prehistory of formal language theory: Gauss languages, *Bull. EATCS* 46, 124–139, 1992.

[Ke92] V. Keränen, Abelian squares are avoidable on 4 letters, In: W. Kuich (ed.), *Automata, Languages and Programming*, LNCS 623, 41–52, Springer-Verlag, 1992.

[Lec85a] M. Leconte, A characterization of power-free morphisms, *Theor. Comput. Sci.* 38, 117–122, 1985.

[Lec85b] M. Leconte, *Codes sans répétitions*, Thèse 3ième cycle, Univ. Paris VII, Technical Report LITP 85-56, 1985.

[Len72] A. Lentin, *Equations dans les monoïdes libres*, Gouthiers-Villars, Paris, 1972.

[LS67] A. Lentin and M. P. Schützenberger, A combinatorial problem in the theory of free monoids, In: R. C. Bose and T. E. Dowling (eds.), *Combinatorial Mathematics*, North Carolina Press, 112–144, 1967.

[Le94] A. Lepistö, Repetitions in Kolakoski sequence, In: G. Rozenberg and A. Salomaa (eds.), *Developments in Language Theory*, Proceedings of DLT'93, World Scientific, 130–143, 1994.

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

[LS01] F. Levé and P. Séébold, Proof of a conjecture on word complexity, *Bull. Belg. Math. Soc.* 8, 277–291, 2001.

[Lot83] M. Lothaire, *Combinatorics on Words*, *Encyclopedia of Mathematics* 17, Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, 1997.

[Lot02] 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.

[LS62] R. C. Lyndon and M. P. Schützenberger, The equation $a^m = b^n c^p$ in a free group, *Michigan Math J.* 9, 289–298, 1962.

[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).

[Ma00] J. Manuch, Characterization of a word by its subwords, *Proc. of DLT*, 210–219, World Scientific, 2000.

[Ma02] J. Manuch, Defect Theorems and Infinite Words, Ph.D. Thesis, University of Turku, *TUCS Dissertations* 41, 2002.

[MP92] F. Mignosi and G. Pirillo, Repetitions in the Fibonacci infinite word, *RAIRO Theor. Inform. Appl.* 26, 199–204, 1999.

[MRS95] F. Mignosi, A. Restivo, and S. Salemi, A periodicity theorem on words and applications, In: J. Wiedermann and P. Hajek (Eds.), *Mathematical Foundations of Computer Science 1995*, LNCS 969, 337–348, Springer-Verlag, 1995.

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

[MSW01] F. Mignosi, J. Shallit, and M.-W. Wang, Variations on a theorem of Fine & Wilf, In: J. Sgall, A. Pultr, and P. Kolman (Eds.), *Mathematical Foundations of Computer Science 2001*, LNCS 2136, 512–523, Springer-Verlag, 2001.

[Mo21] M. Morse, Reccurent geodesics on a surface of negative curvature, *Trans. Am. Math. Soc.* 22, 84–100, 1921.

[Mo38] M. Morse, A solution of the problem of infinite play in chess, *Bull. Amer. Math. Soc.* 44, 632, 1938.

[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 sequences. *Amer. J. Math.* 61, 1–42, 1940.

[MH44] M. Morse and G. Hedlund, Unending chess, symbolic dynamics and a problem in semigroups, *Duke Math. J.* 11, 1–7, 1944.

[MO92] J. Moulin-Ollagnier, Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10 and 11 letters, *Theoret. Comput. Sci.* 95, 187–205, 1992.

[Pan84a] 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.

[Pan84b] J.-J. Pansiot, A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Discr. Appl. Math.* 7, 297–311, 1984.

[Pan86] J.-J. Pansiot, Decidability of periodicity for infinite words, *Theoret. Inform. Appl.* 20, 43–46, 1986.

[Pat70] M. S. Paterson, Unsolvability in 3×3-matrices, *Studies in Appl. Math.* 49, 105–107, 1970.

[PR02] D. Perrin and G. Rindone, On renewal systems, Technical report, Institut Gaspard Monge, Université de Marne-la-Vallée, 2002.

[Pe02] I. Petre, Commutation Problems on Set of Words and Formal Power Series, Ph.D. Thesis, University of Turku, *TUCS Dissertations* 38, 2002.

[Pla99] W. Plandowski, Satisfiability of word equations is in PSPACE, Proc. of FOCS, 495–500, 1999.

[Ple70] P. A. B. Pleasants, Non-Repetitive Sequences, *Math. Proc. Cambridge Philos. Soc.* 68, 267–274, 1970.

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

[PF02] N. Pytheas-Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*, *Lecture Notes in Mathematics* 1794, Springer-Verlag, 2002.

[Ra89] B. Ratoandramanana, Codes et motifs, *RAIRO Theor. Inform. Appl.* 23, 425–444, 1989.

[RR85] A. Restivo and C. Reutenauer, Rational languages and the Burnside problem, *Theoret. Comput. Sci.* 40, 13–30, 1985.

[Re93] C. Reutenauer, *Free Lie Algebras*, London Mathematical Monographs New Series No. 7, Claredon Press, 1993.

[Ric02a] G. Richomme, Lyndon morphisms, Technical report LaRia 2002-15, Laboratoire de Recherche en Informatique d'Amiens (Laria), Université d'Amiens, 2002.

[Ric02b] G. Richomme, Some non finitely generated monoids of repetition-free endomorphisms, *Inform. Proc. Letters*, 2002.

[RS99] G. Richomme and P. Séébold, Characterization of test-sets for overlap-free morphisms, *Discr. Appl. Math.*, 98:151–157, 1999.

[RW00] G. Richomme and F. Wlazinski, About cube-free morphisms, In: H. Reichel and S. Tison (eds.), *STACS '2000*, LNCS 1770, 99–109. Springer-Verlag, 2000.

[RW02a] G. Richomme and F. Wlazinski, Finite test-sets for overlap-free morphisms, In: K. Diks and W. Rytter (eds.), *Mathematical Foundations of Computer Science 2002*, LNCS 2420, 605–614. Springer-Verlag, 2002.

[RW02b] G. Richomme and F. Wlazinski, Some results on k-power-free morphisms, *Theoret. Comput. Sci.* 273, 119–142, 2002.

[Ru78] F. Ruskey, Generating t-ary trees lexicographically, *SIAM J. Comput.* 7, 424–439, 1978.

[Sch56] M. P. Schützenberger, Une théorie algébrique du codage, *Seminaire Dubreil-Pisot 1955–1956 Expose 15*, Institut H. Poincare, Paris, 1956.

[Sch79] M. P. Schützenberger, A property of finitely generated submonoids of free monoids, In: G. Pollak (ed.), *Algebraic Theory of Semigroups*, Proc. Sixth Algebraic Conf., Szeged, 1976, 545–576. North-Holland, 1979.

[Séé88] P. Séébold, An effective solution to the $D0L$-periodicity problem in the binary case, *Bull. EACTS* 36, 137–151, 1988.

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

[Th12] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl., Christiana* 10, 1–67, 1912.

[WZ01] N. Wozni and L. Q. Zamboni, Frequencies of factors in Arnoux-Rauzy sequences, *Acta Arith.* 96, 261–278, 2001.

[Zim82] A. I. Zimin, Blocking sets of terms, *Matem. Sbornik* 119, 363–375, 1982. English translation: *Math. USSR Sbornik* 47, 353–364, 1984.