



J. Karhumäki

Combinatorics on Words: A New Challenging Topic

TURKU CENTRE *for* COMPUTER SCIENCE

TUCS Technical Report
No 645, December 2004



Combinatorics on Words: A New Challenging Topic¹

J. Karhumäki

Department of Mathematics and
Turku Centre for Computer Science
University of Turku
FIN-20014 Turku, Finland
email: karhumak@cs.utu.fi

Abstract

Combinatorics on Words is a relatively new research topic in discrete mathematics, mainly, but not only, motivated by computer science. In this paper we have three major goals. First, starting from very basic definitions we move via some typical proof techniques of the theory to a simply formulated open problems. Second, we discuss a few highlights of the theory, and finally we concentrate to connections between combinatorics on words and other branches of mathematics.

¹Supported by the Academy of Finland under the grant 44087.

Combinatorics on Words (CoW in short) is a relatively new rapidly growing area of discrete mathematics. Its main object is a word, i.e. a sequence – either finite or infinite – of symbols taken from a finite set, usually referred to as an alphabet. The natural environment of a word is that of a free monoid. Consequently, noncommutativity is a fundamental feature of words. Indeed, the words like “no” and “on” are not equal.

Although the theory of words has a plenty of connections to many areas of sciences there is no doubt that computer science is mainly responsible of the current expansion of research on words. Many aspects of computers and computing stand as real motivation to study words. Indeed, modern computers operate on sequences of bits - even when they carry out numerical computations. Consequently, algorithmic number theory, for example, can be viewed as research on words.

A fascinating feature of words is that many problems are very easy to formulate, while their solutions are very difficult. In other words, Combinatorics on Words provides very challenging problems, as we shall see.

The goals of this paper are manifold. In one hand, we want to introduce the field to a larger audience. To achieve this we recall the basic definitions and show a few typical simple proof techniques. On the other hand, we want to emphasize the very challenging nature of the topic. For this we recall a few highlights of the theory, as well as formulate several open problems. A crucial goal is to point out connections between combinatorics on words and other areas of mathematics.

In more details after giving a short history of the field in Section 1 and analyzing the position of combinatorics on words within mathematics in Section 2, we move to more technical parts.

In Section 3 we recall the basic terminology and give a few examples of simple proofs.

In Section 4 we deal with one of the remarkable topics of the theory, namely repetition free, or more generally avoidable words.

In Section 5 we formulate a few highlights of the theory, and in Section 6 we consider more closely connections of words and other topics. This allows to state several simply formulated open problems.

Finally, in Section 7 we conclude with another fundamental open problem.

1 A brief history

At the moment we are approaching a centennial anniversary of A. Thue’s papers on repetition free words, which are generally considered as a starting point of mathematical research on words. Indeed in 1906 Thue published his now classical first paper on repetition free words, see [Be95] and [Th06]. He published his papers as publications of Oslo University (called Christiania at that time). Consequently, his results remained unknown for many decades.

After that there were scattered research papers of the field – in many cases rediscovering some of the original results of Thue. We could mention

e.g. the papers of Morse and Hedlund from the year 1944 and that of Aršon from the year 1937, see [MH44] and [Ar37], respectively.

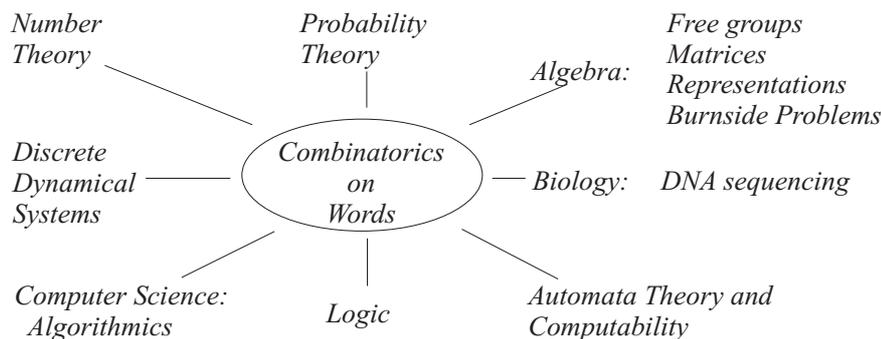
The systematic research on words finally started in 1950s – fifty years after Thue. This happened independently in Paris and in Moscow. In France the research was initiated by M. P. Schützenberger as systematic research on *theory of codes*, see [BP85]. So his motivation came from information theory and computing. In Russia the theory of words was more implicit. It was mainly a tool for P. S. Novikov and S. Adian in developing their fundamental solution to *Burnside problem for groups*, see [Ad79].

A real inspiration for research on words was the first book – *Combinatorics on Words* by M. Lothaire – which was published in 1983, see [Lo83]. It covered in its approximately 200 pages much of the research done before that time. The second volume – *Algebraic Combinatorics on Words* in 2002 – was not any more able to do that in its almost 500 pages, see [Lo02]. Also a biennial conference WORDS has been established, the last one taking place in Turku in 2003.

The above lines aimed to describe very briefly the main lines in the history of CoW. It has to be emphasized that in addition to the above words has played – typically as tools – an important role in some other occasions, in particular in combinatorial group theory.

2 CoW and other sciences

As we already hinted Combinatorics on Words is connected to many other topics, not only in mathematics, but also in computer science, physics and biology. In fact, many fundamental results of the theory have been discovered, or rediscovered, when using words as tools for other sciences. The following diagram describes the main connections of CoW and other sciences.



Inside mathematics the position of CoW is illustrated in the classification of *Mathematical Reviews* (where the topic has had its code 68R15 since 2000). It is in the chapter of Computer Science under the title Discrete Mathematics in Relation to Computer Science.

3 Terminology and simple examples

In this section we recall some basic terminology of CoW, as well as show some proof techniques. In particular, we want to show how starting from scratch we can easily formulate natural open questions. We refer to [Lo83], [Lo02] and [CK97] as general texts on words and to [BK03] as a recent tutorial.

Alphabet A is a finite set. Elements of A are called *letters*, while sequences of letters are called *word*. A word can be either finite or infinite. The sets of all finite (resp. infinite) words are denoted by A^* (resp. A^ω). The sequence of zero letters is called the *empty* word, denoted by 1 . We set $A^+ = A^* \setminus \{1\}$. The *length* of a word u is denoted by $|u|$. We say that a word u is a *prefix* (resp. *suffix* or *factor*) of a word v if there exists a word t (resp. t or t and s) such that $ut = v$ (resp. $tu = v$ or $tus = v$). We write $u \leq v$ or $u = vt^{-1}$ if $ut = v$. For two words u and v we define their *product* or *catenation* as the word uv . A *power* of a word u is a word of the form u^k for some $k \in \mathbf{N}$. If $w = u_1 \dots u_t$ with each $u_i \in U \subseteq A^*$, we say that w possesses a *U -factorization*. Of course, U -factorizations of w need not be unique, for example if $U = \{a, ab, ba\}$, then the word aba has the U -factorizations $a.ba$ and $ab.a$. The above notions extend in a natural way to infinite words.

We call subsets of A^* *languages*. Clearly, A^* (resp. A^+) is the *free monoid*, often referred to as the *word monoid*, resp. the *free semigroup* generated by A . This means that each word has the unique A -factorization. More generally, for each $X \subseteq A^+$, X generates the monoid X^* (resp. semigroup X^+), which, however, need not be free, an example being the language $\{a, ab, ba\}$. Languages which are free generating sets of the semigroups they generate are called *codes*. It is straightforward to see that finite codes can be identified with images of an alphabet under an injective morphism from I^* into A^* . Here, of course, a morphism h is an operation preserving mapping, i.e. $h(uv) = h(u)h(v)$ for all u and v in I^* . Similarly, we say that a morphism $h : I^* \rightarrow A^*$ is an ω -code if it is injective on I^ω , i.e. each infinite word ω has at most one $h(I)$ -factorization.

As an example, the set $X = \{a, ab, bb\}$ is a code, but not an ω -code. Indeed, X is a suffix set meaning that none of the words of X is a suffix of another (and hence no finite word has two X -factorizations). On the other hand, X is not an ω -code since

$$a.bb.bb.bb\dots = ab^\omega = ab.bb.bb.bb\dots$$

The above definitions are enough for the majority of our subsequent considerations. Additional notions needed are defined in the proper places.

We start with a very simple and natural question:

When do two words commute?

Clearly, this is the case if the words are powers of a common word. But actually this is a characterization:

Theorem 1 *For $u, v \in A^*$ the following conditions are equivalent:*

- (i) u and v commute, i.e. $uv = vu$;

- (ii) u and v are powers of a common word, say $u, v \in t^*$ for some $t \in A^*$;
 (iii) u and v satisfy a nontrivial relation (e.g. equation).

Proof. Consider the equivalence of (i) and (ii). Clearly, we are done if one of the words is empty. Similarly the implication “(ii) \Rightarrow (i)” is obvious. So assume that $u, v \in A^+$ and that $uv = vu$. Now, since uv and vu are the same sequence of symbols either $u = v$ (when we are done) or there exists a $t \in A^+$ such that $ut = v$ (or symmetrically $u = vt$). But then

$$uut = u(ut) = uv = vu = (ut)u = utu,$$

so that

$$ut = tu.$$

This means that u and t commute. Consequently, since $|ut| < |uv|$, we can apply induction: u and t are powers of the same word. But then so are u and v , too.

To prove the equivalence of (ii) and (iii) we first note, by above, that (ii) implies (iii). Moreover, (iii) implies (ii) exactly in the same way as above: If u and v satisfy a nontrivial relation we can write

$$u\alpha = v\beta \quad \text{with } \alpha \text{ and } \beta \text{ being sequences of } u\text{'s and } v\text{'s.}$$

Then, by applying the transformation $u = vt$, we obtain

$$(1) \quad t\alpha' = \beta',$$

where α' and β' are sequences of t 's and v 's, and, moreover, the first element in β' is u (and not t). So (1) is *another* nontrivial identity, now on t and u , so that induction applies.

As we saw *two* words satisfy a nontrivial relation if and only if they are powers of a same word. How about if *three* words satisfy two “different” nontrivial relations. Or if they satisfy three such equations, and so on.

Let us continue with a special case. Assume that three *nonempty* words u, v and w satisfy the following relations:

$$(2) \quad \left. \begin{array}{l} u\alpha = v\beta \\ u\gamma = w\delta \end{array} \right\} \text{ with } \alpha, \beta, \gamma, \delta \in \{u, v, w\}^+.$$

A crucial point here is that the equations start differently, that is the first pairs of the words are (u, v) and (u, w) . Clearly, (2) is the general case under this assumption.

Now, if $u = v$ or $u = w$, then, by Theorem 1, all words are powers of a common one. If this is not the case, then we can write

$$u = vt \quad \text{or} \quad v = us \quad (\text{or symmetric cases}).$$

In these cases (2) can be written in the forms

$$\left\{ \begin{array}{l} t\alpha' = \beta' \\ vt\gamma' = w\delta' \end{array} \right. \quad \text{or} \quad \left\{ \begin{array}{l} \alpha'' = s\beta'' \\ u\gamma'' = w\delta'' \end{array} \right. ,$$

with $\alpha', \beta', \gamma', \delta' \in \{v, w, t\}^+$ and $\alpha'', \beta'', \gamma'', \delta'' \in \{u, w, s\}^+$. This means that we obtain new nontrivial pairs of relations of the original form on three words. But the total length of the new words over A are strictly shorter than that of the original ones. So induction applies. The conclusion is

Theorem 2 *If three nonempty words satisfy two relations starting differently, i.e. relations of form (2), then all three words are powers of a common word.*

The assumption “start differently” cannot be abandoned above:

Example 1. Consider the equations

$$(3) \quad xyz = zyx \quad \text{and} \quad xyzy = zyyx.$$

Clearly, the equations are “different”. In fact, the pair they define is even *independent*, meaning the set of all solutions of it is strictly smaller than that of the single equations. Indeed,

$$x = aba, \quad y = b, \quad z = a$$

is a solution of the former, but not of the latter, while

$$x = abba, \quad y = b, \quad z = a$$

is a solution of the latter, but not of the form. However, the pair (3) has a solution $x = z = \alpha$ and $y = \beta$, which need not be periodic. \square

By Example 1, Theorem 2 does not extend to all pairs of relations. How about if we take three relations, or even k for some natural number k . The answer is not known:

Open Problem I. *Can three nonempty words satisfy a system of three independent equations without being powers of a common word?*

The above problem is amazing. It is very simply formulated and motivated as we did. However, it does not seem to be easy.

Theorem 2 has a generalization. Consider a system S of equations with finite set X of unknowns. Here the equations are constant-free, that is pairs of words over X . A *solution* in A^* is a morphism $h : X^* \rightarrow A^*$ which identifies the left and right hand sides of the equations. Let $F \subseteq A^+$ be a solution of S , i.e. $F = h(X)$ for a morphism on X^+ . We associate F (or in fact the system of relations it satisfies) with the graph $G_F = (V, E)$, where

- the set of vertices V is X , and
- there exists an edge between x and y if and only if $h(x)h(X)^* \cap h(y)h(X)^* \neq \emptyset$.

The above graph is called the *dependency graph* of F . The graph becomes unique if S constitutes all relations satisfied by X . For our purposes this is not, however, important.

In Example 1 and Theorem 2 the dependency graphs are as follows:



A crucial difference of these is that the latter is connected, while the former is not. This is an important difference.

In order to formulate our result we define the *rank* of a finite set $F \subseteq A^+$ as the number

$$r(F) = \min\{\text{card}(L) \mid F \subseteq L^+\}.$$

This means that the rank of F tells the minimal cardinality of a set of words needed to construct all words of F (or F^+) as products of these words. The rank defined here is usually called *combinatorial rank* of F . The result is:

Theorem 3 *For each finite set $F \subseteq A^+$, we have*

$$\text{rank}(F) \leq c(G_F),$$

where $c(G_F)$ is the number of connected components of G_F .

Theorem 3 becomes particularly important in the case when G_F is connected. Then the set F is *periodic*, e.g. all of its words are powers of a common word. This is what happened in Theorem 2. The proof of Theorem 3 is similar to that of Theorem 2, see e.g. [HK04]. Note also that in Theorem 3 it is important that all words are nonempty.

All the results discussed so far are variants of so-called *defect theorem*, see e.g. [BPPR79]. In the simplest form it claims:

Defect Theorem. *If a set of n words satisfies a nontrivial relation, that is is not a code, then these words can be expressed as products of at most $n - 1$ words.*

As discussed in details in [HK04] there are many variants and aspects of defect theorems, including that they can be formulated for infinite relations as well.

One of the most interesting and natural notions of words is that of periodicity. We say that a word $w = a_1 \dots a_t$, $a_i \in A$, possesses a *period* p if $a_i = a_{i+p}$ for $i = 1, \dots, t - p$. The smallest period of w is called *the period* of w . The basic result on periodicity of words is the following lemma. Here we use the notion $u \wedge v$ for the *maximal common prefix* of words u and v .

Theorem 4 (*Periodicity Lemma of Fine and Wilf, 1956*). *Two words u and v are powers of a same word if and only if*

$$|u^\omega \wedge v^\omega| \geq |u| + |v| - \text{gcd}(|u|, |v|).$$

Intuitively, the above gives the exact borderline how much two periodic processes have to match in order to be the same processes, that is to have the common period. Interestingly, the original formulation of the theorem

was in terms of periodic real functions, see [FW65]. However, its natural environment is that of words, as stated in Theorem 4.

Words $F_4 = abaab$ and $F_5 = abaababa$ give an example of words showing the optimality of Theorem 4:

$$\text{abaababaaba} \left| \begin{array}{l} a^b a^b a^b \\ b^a a^b \end{array} \right.$$

Indeed, the words F_4^ω and F_5^ω differ from each other at the point equal to $|F_4| + |F_5| - 1 = 12$.

We conclude this section by looking for a connection between codes and periodic words. By an *ultimately periodic* word we mean an infinite word of the form

$$w = uvvv \dots = uv^\omega.$$

Accordingly, a *periodic* bi-infinite word is a word of the form

$$w = \dots vvv \dots = {}^\omega v^\omega.$$

We recall also that a *code* (resp. ω -*code*) is a morphism $h : I^* \rightarrow A^*$ which is injective on I^* (resp. on I^ω). We have the following result which connects the interesting notions of codes and periodic words.

Theorem 5 *A morphism $h : I^* \rightarrow A^*$ is a code if and only if only the images of ultimately periodic words under h are ultimately periodic.*

Proof. In one direction the implication is easy: if h is not a code, e.g. there are two different words w_1 and w_2 such that $h(w_1) = h(w_2)$, it is a simple task to construct a nonultimately periodic word the image of which equals to $h(w_1)^\omega$.

For the other direction, we prove here a slightly weaker result, namely that ω -codes satisfy the condition. The extension to all codes can be found in [De93]. So assume that $h : I^* \rightarrow A^*$ is an ω -code, and further that $h(w) = uv^\omega$ for some finite words u and v . We can assume that the length of v is minimal, that is v is *primitive*, in this representation. Then by the pigeon hole principle there exists a factorization

$$w = psq$$

such that $h(s)$ is a power of a conjugate of v (see Section 5). This means that $h(q)$ is an ω -power of the same word implying that

$$h(w) = h(ps^\omega).$$

But since h is an ω -code, necessarily $w = ps^\omega$, that is w is ultimately periodic, which was to be proved.

The above result can be modified also for bi-infinite words:

Theorem 6 *A morphism $h : I^* \rightarrow A^*$ is a code if and only if it maps only periodic bi-infinite words into periodic bi-infinite words.*

Note that in Theorem 6 we have to use ultimately periodic words since the code

$$h(a) = b \quad \text{and} \quad h(b) = ab$$

maps the word ab^ω to the fully periodic word $(ba)^\omega$.

4 Fixed points of morphisms

In this section we consider a typical research topic of CoW, namely the theory of *avoidable words*, as well as one of the central tools used therein, that is *fixed points of morphisms*.

Let $h : A^+ \rightarrow A^+$ be a morphism, and moreover assume that for some letter $a \in A$ the word a is a proper prefix of $h(a)$, that is

$$h(a) = au \quad \text{for } u \in A^+.$$

Then, for any $i \geq 1$, we have

$$h^i(a) = h^{i-1}(h(a)) = h^{i-1}(au) = h^{i-1}(a)h^{i-1}(u),$$

so that $h^{i-1}(a)$ is also a proper prefix of $h^i(a)$. This implies that the limit

$$\alpha_h = \lim_{i \rightarrow \infty} h^i(a)$$

exists. (More precisely, the limit is taken under the metric defined by $d(u, v) = 2^{-|u \wedge v|}$). Clearly, α_h is a *fixed point* of h , that is $h(\alpha_h) = \alpha_h$. We say that α_h is defined by *iterating a morphism h at point a* . Let us take two examples.

Example 2. (*Fibonacci Word*) Consider the morphism

$$f : \begin{array}{l} a \mapsto ab \\ b \mapsto a \end{array}$$

and its unique fixed point

$$\alpha_f = \lim_{i \rightarrow \infty} f^i(a) = abaababaabaab \dots$$

Denote by $F_n = f^{n-1}(a)$ so that $F_1 = a$ and $F_2 = ab$. Then by the definition the sequence $(|F_i|)_{i \geq 0}$ (with $|F_0| = 1$) forms the well known *Fibonacci sequence of numbers*. Hence, this sequence of words is referred to as the sequence of Fibonacci words, and α_f is called the (infinite) *Fibonacci word*. It follows easily that

$$F_{i+1} = F_i F_{i-1} \quad \text{for } i \geq 1.$$

Therefore the word sequence is defined exactly as the number sequence, the operation being the product of words.

The Fibonacci word has a number of remarkable properties. Indeed, it is feasible to believe that its importance to the theory of words is matching to that of Fibonacci numbers to the theory of numbers. Already now it is known to be the most frequently used counterexample to show optimality properties of words.

Here we mention just a few important properties of the Fibonacci word. First, for each $n \geq 1$, the number of factors of length n in α_f is $n+1$. In other words its (*subword*) *complexity* $p(n)$ is $n+1$. This means that of the factors of length n only one can be extended to the right by the both two letters staying in α_f . The complexity $p(n) = n+1$ is the lowest possible complexity among the nonultimately periodic words (for ultimately periodic words the complexity is bounded, as is relatively easy to see). For more details see e.g. [BK03].

As we hinted the Fibonacci word is not ultimately periodic. This is a nice simple exercise. However, it is *locally regular* in the sense that any long enough prefix of it ends with a repetition, e.g. with a suffix of the form uu . In fact, the length of u is at most five. So the local regularity “ending with a square” does not imply the global regularity “being ultimately periodic”. For the cubes instead of the squares the global regularity is forced! More about these so-to-say chaotic vs. predictable properties can be found in [MRS98], [KLP02] and [Le02]. The Fibonacci word and the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ play a fundamental role here. \square

Example 3. (*Thue-Morse Word*) The morphism

$$t : \begin{array}{l} a \mapsto ab \\ b \mapsto ba \end{array}$$

defines the infinite word

$$\alpha_t = abbabaabbaabba \dots$$

This is the word discovered by Thue in 1906, and later rediscovered by Morse in 1940s. In fact, it was considered already in the middle of 19th century by Prouhet. Consequently, it is called *Thue-Morse word* (or sometimes *Thue word* or *Prouhet-Thue-Morse word*). Its remarkable property is that it does not contain any factor of the form $u\text{first}(u)$, where $\text{first}(u)$ denotes the first symbol of u . In particular, it is *cube-free*. \square

A major application area of fixed points of morphisms is the theory of repetition free words, or more generally avoidable words. We say that a word w is a *square*, if it is of the form $w = uu$, or in general w is a *kth power* if $w = u^k$. Here k is allowed to be a rational number. For example, *abaabaa* is the $2\frac{1}{3}$ power of *aba*. Further we say that w is *k-free* or *avoids kth powers* if w does not contain as a factor any power of order $\geq k$. This definition allows to consider also *k-free* words for an irrational k .

The above definitions were extended in [BEM79] to avoidable words. Let X be an other alphabet, the alphabet of patterns. A *pattern* is any word

over X . We say that a word w *avoids* a pattern p if, for any morphism $h : X^+ \rightarrow A^+$, the word w avoids $h(p)$. Consequently, for example 2-free words are exactly those which avoid the pattern $p = xx$. It is also easy to see that the words not containing as a factor a word of the form $u\text{first}(u)$, are exactly those avoiding $xyxyx$. These words are also referred to as 2^+ -free or *overlap free words*.

A fundamental question is which patterns are avoidable in infinite words. Such patterns are called *avoidable*, or even n -*avoidable* if the infinite word is over an n -letter alphabet. The following fact, based on the famous König's Lemma, relates infinite sets of finite words and infinite words in this context.

Fact. *A pattern p is avoided by an infinite language over A if and only if it is avoided by an infinite word over A . \square*

What Thue proved, among other things, in 1906 is as follows.

Theorem 7 (Thue, 1906) (i) *There exists an infinite 2^+ -free binary word.*
(ii) *There exists an infinite 3-free word over a ternary alphabet.*

Such words are obtained by iterating the morphisms

$$t : \begin{array}{l} a \mapsto ab \\ b \mapsto ba \end{array} \quad \text{and} \quad \varphi : \begin{array}{l} a \mapsto abc \\ b \mapsto ac \\ c \mapsto a \end{array} ,$$

respectively. Actually, we also have

$$\alpha_\varphi = h^{-1}(\alpha_t),$$

where $h : \{a, b, c\}^* \rightarrow \{a, b\}^*$ is the morphism

$$h : \begin{array}{l} a \mapsto abb \\ b \mapsto ab \\ c \mapsto a \end{array} .$$

In terms of avoidability the above means that the pattern xx is 3-avoidable (but not 2-avoidable since every word longer than 3 contains a square), while the pattern $xyxy$ is even 2-avoidable. Hence we have an example of a pattern separating binary and ternary alphabets. As a much more complicated result a pattern is also known which separates 3- and 4-letter alphabets, see [BMT89]. But we do not know any pattern which would separate 4- and 5-letter alphabets! However, the following is proved, see [Ca93]:

Theorem 8 *The characterization of binary patterns which are avoidable in the binary alphabet is known.*

Let us denote by A_n an n -letter alphabet. We set

$$k\text{-}F_l(n) = \{w \in A_n^l \mid w \text{ is } k\text{-free}\}$$

and

$$k^+-F_l(n) = \{w \in A_n^l \mid w \text{ is } k^+\text{-free}\}.$$

Here k^+ -free means that no repetitions of order strictly larger than k are allowed as a factor. So $2^+-F_l(2)$ denotes the set of 2^+ -free binary words of length l . Obviously, l above can be infinity as well.

In order to formulate our next results we call $L \subseteq A^*$ *exponential* (resp. *polynomial*) if, there exist constants $a, b > 0$ and $\alpha, \beta > 1$ such that, for all $n \geq 1$

$$a\alpha^n \leq \text{card}(L \cap A^n) \leq b\beta^n$$

$$\text{(resp. } \text{card}(L \cap A^n) \leq bn^\beta \text{)}.$$

As an extension of Thue's result we have, see [Br83]:

Theorem 9 (i) *The sets of 2-free ternary words and 3-free binary words are exponential.* (ii) *The sets of 2-free ternary infinite words and 3-free binary infinite words are nondenumerable.*

Surprisingly, the results are not the same for 2^+ -free words, see [RS85] and [Lo83]:

Theorem 10 (i) *The set of 2^+ -free binary words is polynomial.*

(ii) *The set of 2^+ -free binary infinite words is nondenumerable.*

Recently, the exact borderline between above exponential and polynomial cases was obtained in [KS04]:

Theorem 11 (Karhumäki-Shallit, 2004) *The set of binary $2\frac{1}{3}$ -free words is polynomial while that of $2\frac{1}{3}^+$ -free words is exponential.*

The above three theorems show that repetition free words are after all not that rare. However, these results are more existential than tools to construct such words. Most of the concrete examples are obtained by iterating a morphism, or applying another morphism to a word obtained by this method. Interestingly, this method can produce only a numerable portion of all of these words.

5 Two highlights of CoW

In this section we discuss about two major results on CoW. We believe that they are fundamental not only from the point of view of words, but also for the development of mathematics as a whole. Both results are related to word equations.

Let X be a finite set of *unknowns* and A , as usual, our word alphabet. An *equation* over A^* with X as the set of unknowns is a pair $(u, v) \in (X \cup A)^+ \times (X \cup A)^+$ usually written as $u = v$. If $u, v \in X^+$ then the equation is *constant free*. A *solution* of equation $u = v$ is a morphism

$h : (X \cup A)^* \rightarrow A^*$ identifying u and v and being identity on letters of A , e.g. satisfying

$$h(u) = h(v) \quad \text{and} \quad h(a) = a \quad \text{for} \quad a \in A.$$

Clearly, the above notions extend to arbitrary systems of equations. We call two systems S_1 and S_2 of equations *equivalent* if they have exactly the same solutions. A system S is *independent* if it is not equivalent to any of its proper subsystems.

Example 4. Theorem 1, interpreted in terms of equations, says that the general solution of the *commutation equation* $xy = yx$ is

$$\{(\alpha^i, \alpha^j) \mid i, j \geq 0, \alpha \in A^+\},$$

that is morphism $h : \{x, y\}^* \rightarrow A^*$, $x \mapsto \alpha^i$ and $y \mapsto \alpha^j$. □

Example 5. The *conjugacy equation* $xz = zy$ is known to have the general solution, see e.g. [Lo83]:

$$\left. \begin{array}{l} x \mapsto \alpha\beta \\ y \mapsto \beta\alpha \\ z \mapsto \alpha(\beta\alpha)^i \end{array} \right\} \text{ for } i \geq 0 \text{ and } \alpha, \beta \in A^*.$$

We recall that two words are *conjugates* if they are like x and y above, i.e. one is obtained from the other by moving a prefix to the end. □

In the above examples general solutions were expressed via *parametric words*, where parameters were of two types: word parameters and numerical parameters. In general, the solutions are quite complicated. Indeed, in [Hm71], for a simple proof see [Pe04], it was shown that the general solution of the equation $xyz = zvx$ can not be expressed as a finite formula of parametric words. For three unknown constant free equations this, however, is possible, see again [Hm71].

An amazing thing is that it is easy to give examples of equations which are very difficult to solve, in fact it is not known whether they have solutions (of certain types). As an illustration we do not know whether the equation $x^2y^3x^2 = u^2v^3u^2$ has any nonperiodic solution (e.g. a solution where all components are not powers of a same word).

Despite of above we have the fundamental result of Makanin, see [Ma77].

Theorem 12 (*Makanin, 1976*) *It is decidable whether a given equation over A^* possesses a solution.*

The Makanin's algorithm is one of the most complicated ever constructed. Accordingly its computational complexity is terrible. Recently Plandowski introduced a completely new algorithm for Makanin's problem, which is usually referred to as the *satisfiability problem for word equations*, see [Pl04].

Theorem 13 (*Plandowski, 2004*) *Satisfiability problem for word equations is in PSPACE.*

Recall that PSPACE denotes the class of problems which can be solved by an algorithm using only polynomial size memory. Consequently, Plandowski's algorithm is not practical either, but on the other hand, it is not known to be more complex than the best algorithms for very natural and simple problems, like the equivalence problem for two finite nondeterministic automata.

As another highlight we state a fundamental compactness property of word equations, usually referred to as the *Ehrenfeucht Compactness Property*. It was conjectured by A. Ehrenfeucht when studying a problem on iterated morphisms in mid 1970s and shown to be true in [AL85] and [Gu86].

Theorem 14 (*Ehrenfeucht Compactness Property*) *Any system of equations over word monoids with only a finite number of unknowns is equivalent to one of its finite subsystems.*

The result does not hold for all finitely generated monoids but does hold true for abelian monoids, see [KP96] or [HKP02]. Interestingly both the abelian case and the free case are consequences of a fundamental result in polynomial algebras, namely Hilbert Bases Theorem.

An interesting open question is to look for bounds for the size of a finite set in Theorem 14. No upper bound depending on the number of unknowns is known, while the best lower bounds are of order $\Omega(n^4)$, see [KP96]. Consequently we have open problems:

Open Problem II (i) *Does there exist any function $f : \mathbf{N} \rightarrow \mathbf{N}$ such that any independent set of equations over A^* having n unknowns is of size at most $f(n)$.*

(ii) *Would any polynomial work in (i)?*

6 CoW as a tool

Combinatorics on words has had a lot of applications to other research topics, many of those having quite a different flavour. Fundamental basic results, such as Fine and Wilf's Theorem or Thue's results, have immediate applications. The former are used in algorithmic problems, like in pattern matching, to conclude the optimality of certain algorithms, cf. e.g. [CR94]. The latter provides an elegant solution to Burnside Problem for semigroups, see [Lo83].

Among the most useful impacts of CoW to other areas is achieved, however, when words and their properties have been used as tools to solve problems of other areas. And this impact is by no means one-way: many results on words have been discovered via such considerations.

We concentrate here in two matters. First, we give a few examples showing that something which is *complex* or *complicated* in a classical sense need not be so in terms of words. Then we consider in more details a technical connection between words and integer valued matrices.

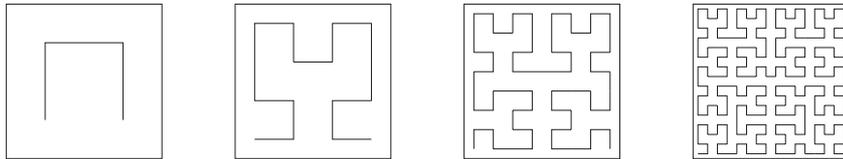
Example 6. *Cantor's Dust* (CD for short) has played an important role in measure theory. It is a set of points in the *unit interval* I which is

obtained as the limit when the following rule is used infinitely many times starting from I : Divide all intervals what are at the i th step into three equal parts and delete all the middle ones. A question is what remains at the limit. As is well known the remaining set is counterintuitive: its size (measure) is zero, but still it is nondenumerable, that is of the same cardinality as the whole interval. In terms of words Cantor's Dust has a simple and natural representation in ternary notation. It consists of all infinite sequences of words over $\{0, 1, 2\}$ which do not contain the digit 1 at all. Consequently, we have

$$I \leftrightarrow \{0, 1, 2\}^\omega \quad \text{and} \quad CD \leftrightarrow \{0, 2\}^\omega.$$

In particular, the nondenumerability is clear.

Example 7. *Hilbert Space Filling Curve* is another classical anomaly in classical analysis and topology. It describes a continuous bijective mapping from the unit interval I into its square $I \times I$. It is typically described as the limit of the process depicted as:



Now consider four letters $u(p)$, $d(own)$, $l(ef)t$ and $r(igh)t$. Using this we can describe the above approximations of the Hilbert Curve as words over the alphabet $\{u, d, l, r\}$. Moreover, the i th word is an extension of the $(i - 2)$ nd word. Consequently, the Hilbert Curve seems to be related to infinite words obtained by iterating a morphism. More precise, what is true is that the Hilbert Curve is of the form $c(\alpha_h)$, where α_h is a fixed point of a morphism over 12-letter alphabet and c is a renaming of the letters back to the four letter alphabet. Without going into the details of the construction, which we leave to the reader as an exercise, we want to emphasize that in terms of CoW the Hilbert Curve is, actually, among the most simple and natural objects. \square

Example 8. The third connection is to the theory of numbers. Having only a very coarse classification we can say that the complexity of numbers increases when we move from rational numbers, to algebraic ones and further to transcendental ones. Now, we think infinite words as k -ary representations of numbers. Is the above growth of complexity reflected to their representations, i.e. to corresponding words? Of course, ultimately periodic words, which are the simplest infinite words, correspond to rational numbers. Among the second simplest infinite words are those obtained as fixed points of morphisms. However, the Fibonacci word for example, represents not an algebraic but transcendental number, see [FM97]. \square

We conclude this section by considering a particular technical connection known already from 1920s. Namely that word monoids, even denumerably

generated ones, can be embedded into the multiplicative semigroup of (2×2) -matrices over \mathbf{N} , that is to say we have

$$\{a_i | i \in \mathbf{N}\}^* \hookrightarrow \{a, b\}^* \hookrightarrow M_{2 \times 2}(\mathbf{N}).$$

The former embedding is given by the mapping

$$a_i \mapsto ab^i \quad \text{for } i \in \mathbf{N},$$

and the latter, for example, by

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

These embeddings allow to translate problems of words to those of matrices, and vice versa. Both directions have turned out very fruitful. We give just two examples, for more we refer to [HK97].

The fundamental Ehrenfeucht Compactness Property is an example of a result of words which is proved via the known properties of matrices (or related matters). In the other directions undecidability results for matrices are neat examples. Indeed, undecidability is most natural for words (which are basic objects of Turing type computing). Two concrete results are as follows:

Theorem 15 (i) *It is undecidable whether the multiplicative semigroup generated by a finite set of triangular (3×3) -matrices over \mathbf{N} is free.*

(ii) *It is undecidable whether the multiplicative semigroup generated by a finite set of (3×3) integer matrices contains the zero matrix.*

For the proofs of these results we refer to [CHK99] and [Pa70], see also [HK97].

Note that part (ii) of the above theorem has a nice interpretation to elementary linear algebra. Indeed, it implies that the question whether some composition of a given finite set of linear transformations in three dimensional Euclidean space to the zero mapping is undecidable!

Theorem 15 motivates amazing open problems on matrices.

Open Problem III. *Given a finite set S of 2×2 matrices with nonnegative integer values. Is it decidable whether S under the product of matrices is free?*

Actually, the problem remains even if S consists of two matrices only. We believe that the above problem is not easy. As an indication we mention that we do not know whether the concrete matrices

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

generate the free semigroup, see [CHK99].

7 Conclusions

We have discussed on quite a general level on many aspects of Combinatorics on Words. We hope we were able to convince the reader of many challenges of the topic, as well as to point out some fundamental results achieved so far. In particular, we would like to repeat our view that CoW is a field where natural open problems are extremely easy to state.

Throughout the presentation we have formulated several open problems. We do not repeat those here, instead we mention one more fundamental problem. It asks whether Makanin's result for the satisfiability problem of word equations extends from words to finite sets of words.

Open Problem IV. *Is it decidable whether a given equation over finite sets of words has a solution?*

In this problem the underlying monoid is that of finite languages, which is not free, contrary to word monoids. Of course, constants (e.g. finite sets) are allowed here in the equation. If decidable, the proof of that is likely to be very difficult. An evidence of that is a result in [KL03] which says that, unlike in the case of word monoids, the problem becomes undecidable if the single equation is replaced by a system of so-called *rational* equations. Another evidence is that even the one unknown case is unanswered:

Open Problem V. *Is it decidable whether for two finite sets A and B the equation $Az = zB$ has a solution?*

The equation considered here is the *conjugacy equation* (see Section 5). So the problem asks to decide whether two finite sets are conjugates.

References

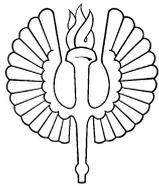
- [Ad79] S. I. Adian, *The Burnside Problem and Identities in Groups*, Springer-Verlag, 1979.
- [AL85] M. H. Albert and J. Lawrence, A proof of Ehrenfeucht's conjecture, *Theoret. Comput. Sci.* 41, 121–123, 1985.
- [Ar37] S. E. Aršon, Proof of the existence on n -valued infinite asymmetric sequences, *Mat. Sb.* 2(44), 769–779, 1937.
- [BEM79] D. R. Bean, A. Ehrenfeucht, and G. F. McNulty, Avoidable patterns in strings of symbols, *Pacific J. Math.* 85, 261–294, 1979.
- [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 à Montréal 20, 1995.
- [BK03] J. Berstel and J. Karhumäki, Combinatorics on Words - A tutorial, *Bull. EATCS* 79, 178–229, 2003; also in: *Current Trends in Theoretical Computer Science. The Challenge of the New Century*, Gh. Paun, G. Rozenberg, A. Salomaa (eds.), World Scientific, Singapore, 2004.

- [BMT89] K. A. Baker, G. F. McNulty and W. Taylor, Growth problems for avoidable words, *Theoret. Comput. Sci.* 69, 319–345, 1989.
- [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.
- [Br83] F.-J. Brandenburg, Uniformly growing k -th power-free homomorphisms, *Theoret. Comput. Sci.* 23, 69–82, 1983.
- [Ca93] J. Cassaigne, Unavoidable binary patterns, *Acta Informatica* 30, 385–395, 1993.
- [CHK99] J. Cassaigne, T. Harju, and J. Karhumäki, On the undecidability of freeness of matrix semigroups, *Intern. J. Alg. & Comp.* 9, 295–305, 1999.
- [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.
- [CR94] M. Crochemore and W. Rytter, *Text algorithms*, Oxford University Press, 1994.
- [De93] J. Devolder, Particular Codes and Periodic Biinfinite Words, *Information and Control* 107, 185–201, 1993.
- [FM97] S. Ferenczi and C. Mauduit, Transcendence of numbers with a low complexity expansions, *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.
- [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.
- [HK97] T. Harju and J. Karhumäki, Morphisms, In: G. Rozenberg and A. Salomaa (eds.), *Handbook of Formal Languages*, 439–510, Springer-Verlag, 1997.
- [HK04] T. Harju and J. Karhumäki, Many aspects of defect theorems, *Theoret. Comput. Sci.* 324, 35–54, 2004.
- [HKP02] T. Harju and J. Karhumäki, W. Plandowski, Independent system of equations, in: M. Lothaire (Ed.), *Algebraic Combinatorics on Words*, Cambridge University Press, Cambridge, 2002.
- [Hm71] Y. I. Hmelevskii, Equations in free semigroups, *Proc. Stoklov Inst. Math.* 107, 1971 (English transl. *Amer. Math. Soc. Translations*, 1976).

- [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.
- [KL03] J. Karhumäki and L. P. Lisovik, The equivalence problem for finite substitutions an ab^*c , with applications, *Intern. J. Found. Comput. Sci.* 14, 699–710, 2003.
- [KP96] J. Karhumäki and W. Plandowski, On the size of independent systems of equations in semigroups, *Theoret. Comput. Sci.* 168, 105–119, 1996.
- [KS04] J. Karhumäki and J. Shallit, Polynomial versus exponential growth in repetition-free binary words, *J. Comb. Theory Ser. A* 105, 335–347, 2004.
- [Le02] A. Lepistö, On Relations between Local and Global Periodicity, Ph.D. Thesis, University of Turku, *TUCS Dissertations* 43, 2002.
- [Lo83] M. Lothaire, *Combinatorics on Words*, *Encyclopedia of Mathematics* 17, Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, 1997.
- [Lo02] M. Lothaire, *Algebraic Combinatorics on Words*. *Encyclopedia of Mathematics* 90, Cambridge University Press, 2002.
- [Ma77] G. S. Makanin, The problem of solvability of equations in a free semigroup, *Mat. Sb.* 103, 147–236, 1977 (English transl. in *Math. USSR Sb.* 32, 129–198).
- [MRS98] F. Mignosi, A. Restivo, and S. Salemi, Periodicity and golden ratio, *Theoret. Comput. Sci.* 204, 153–167, 1998.
- [MH44] M. Morse and G. Hedlund, Unending chess, symbolic dynamics and a problem in semigroups, *Duke Math. J.* 11, 1–7, 1944.
- [Pa70] M. S. Paterson, Unsolvability in 3×3 -matrices, *Studies in Appl. Math.* 49, 105–107, 1970.
- [Pe04] E. Petre, An Elementary Proof for the Non-parametrizability of the Equation $xyz = zvx$, in: *Proceedings of MFCS2004*, LNCS 3153, 807–817, 2004.
- [Pl04] W. Plandowski, Satisfiability of word equations with constants is in PSPACE, *Journal of the ACM* 51(3), 483–496, 2004.
- [RS85] A. Restivo and S. Salemi, Overlap-free words on two symbols, In: M. Nivat and D. Perrin (eds.), *Automata on Infinite Words*, LNCS 192, 198–206, 1984.
- [Th06] A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl., Christiania* 7, 1–22, 1906.

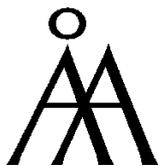
TURKU
CENTRE *for*
COMPUTER
SCIENCE

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi



University of Turku

- Department of Information Technology
- Department of Mathematical Sciences



Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research



Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 952-12-1469-4

ISSN 1239-1891