Mathematisches Forschungsinstitut Oberwolfach

Report No. 37/2010

DOI: 10.4171/OWR/2010/37

Mini-Workshop: Combinatorics on Words

Organised by
Valerie Berthe, Montpellier
Juhani Karhumäki, Turku
Dirk Nowotka, Stuttgart
Jeffrey Shallit, Waterloo

August 22nd – August 28th, 2010

Abstract. The area of combinatorics on words is concerned with properties of sequences of symbols. It is characteristic to the field that questions arise from various mathematical problems, and hence, many fundamental results on words have been established in different areas. Over the last two decades the theory has developed into a quickly growing topic of its own. This workshop was dedicated to reflect on the current status of the field, discuss the impact of recent results, and provide new research challenges. This is a report on the meeting and presentation of extended abstracts of the lectures.

Mathematics Subject Classification (2000): 68R15 (main), 03D03, 03D40, 11N60, 20M05.

Introduction by the Organisers

Combinatorics on words studies properties of sequences of symbols, either finite or infinite, taken from a finite alphabet. The focus on words might be algebraic, combinatorial, or algorithmic. It is characteristic to the field that the motivation to study properties of sequences has arisen from very different mathematical problems. As a result many fundamental properties of words, e.g., on avoidable patterns, unavoidable regularities, and word equations have been established in various mathematical areas. Over the last two decades the theory has developed into a quickly growing topic of its own.

In this workshop the current status of the field was identified, the impact of recent breakthrough results like on unavoidable repetitions in infinite words, word equations, the structure of finite words of small index, and the transcendence of

certain morphic words was discussed, as well as, perspectives for research on this new and challenging field were created.

The talks of this workshop addressed a large portion of topics in combinatorics on words. They ranged from complexity questions (inconstancy, palindromic and factor complexity, rich words) over pattern avoidance, word equations, special factorizations and periods to topics touching on number theory and graph theory. One session was dedicated to the perspectives of combinatorics on words and open problems.

It has to be noted that one could experience a very productive atmosphere during the whole workshop. All talks were accompanied with interesting comments, proposed conjectures and (sometimes) solutions, connections with other fields were discovered, and lively discussions were held far beyond the time of the lectures. Just to name two examples of discussed topics: (1) The Halbeisen-Hungerbuehler-Pirillo-Varicchio problem asks for a word over some finite alphabet $\{0, 1, ..., k\}$ such that two consecutive factors of the same length never have the same sum. New approaches to this problem were raised and investigated during the workshop. (2) A new idea on the connection between solutions of word equations and semilinear sets was discussed and yields a fresh approach to the investigation of the cumulative defect effect that may eventually lead to new progress on the Culik-Karhumäki conjecture which has withstood a solution for about three decades.

The work of the participants at the workshop, as documented by the abstracts in this report, shows that combinatorics on words is an active field with many facets and surprising connections. We are grateful to all participants for their contributions to this successful workshop as well as to the staff of the MFO for their perfect service.

Mini-Workshop: Combinatorics on Words

Table of Contents

Jean-Paul Allouche (joint with Laurence Maillard-Teyssier)
Inconstancy and complexity of finite and infinite sequences 2199

Valérie Berthé
Word combinatorics, S-adic sequences and multidimensional continued fractions 2202

Julien Cassaigne
Words of very low factor complexity 2204

James D. Currie
Power-free Sequences: Topology, Reachability and Curling Numbers 2205

Aldo de Luca
On a palindromization map on free monoids 2207

Amy Glen
On (almost) rich words 2209

Tero Harju (joint with Volker Diekert, Dirk Nowotka)
Weinbaum Factorizations 2212

Štěpán Holub (joint with Dirk Nowotka)
Periods and Unbordered Factors: The Ehrenfeucht-Silberger Problem . . . 2213

Juhani Karhumäki
Independent Systems of Word Equations and Related Topics 2215

Dirk Nowotka (joint with Bastian Bischoff)
Word periods under involution 2219

Pascal Ochem
Repetition-free colorings of trees 2222

Elena V. Pribavkina
On the Minimal Uncompletable Word Problem 2224

Eric Rowland (joint with Bobbe Cooper, Doron Zeilberger)
Ambiguity in a certain context-free grammar 2226

Kalle Saari
The enumeration of squares and runs in the Fibonacci words revisited . . 2229

Jeffrey Shallit
On Patterns and Pattern Avoidance 2230

Thomas Stoll
On the sum of digits of n and n^h 2237
Luca Q. Zamboni
A Note on Coloring Factors of Words 2240

Abstracts

Inconstancy and complexity of finite and infinite sequences

Jean-Paul Allouche
(joint work with Laurence Maillard-Teyssier)

This text is an extended abstract of a paper of the same authors [4].

1. Introduction

What is a "complicated" or "complex" (finite or infinite) sequence? The reader will probably agree that the sequence 0000... is very "simple", that the sequence 010101... is also simple though may be less simple, that an eventually periodic sequence with a very long preperiod might be very complicated, and that a "random" sequence must be complicated. One can imagine of several ways of defining complexity in mathematical terms. Classical approaches include

- algorithmic complexities: in particular *Kolmogorov-Solomonoff-Chaitin complexity* and its relation to compressibility: see, e.g., [16];
- combinatorial complexities: in particular *block – or factor – complexity*, see, e.g., [1, 11], *repetition complexity*, see, e.g., [15], *palindrome complexity*, see, e.g., [3], *arithmetical complexity*, see, e.g., [6], *maximal pattern complexity*, see, e.g., [12, 13]. Also see the paper [14], and the talk of J. Cassaigne at this mini-workshop. We also mention quasi-periodicity, recurrence and uniform recurrence, some definitions of pseudo-randomness, e.g., [17], the study of certain subsequences of the given sequence – in particular the *measure of automaticity* introduced by Shallit et al., see, e.g., Chapter 15 of [5];
- number-theoretical complexity: in relation with periodicity, but also with algebraicity properties of real numbers, formal power series or continued fractions associated with a given sequence, see, e.g., [2].

Now what is a *fluctuating* (finite or infinite) sequence? How is it possible to detect and define sequences that admit large variations or fluctuations? A classical criterion is the *residual variance* of a sequence: this is a measure of the "distance" between the piecewise affine curve associated with the sequence and its regression line. Residual variance does not discriminate between a sequence that oscillates wildly and a sequence that grows very rapidly. We thus propose to bring to light an old result of Cauchy and Crofton, in order to define what we call the *inconstancy* of a sequence. This definition is based upon the idea that a complicated curve is cut by a "random" straight line in many more points than a "quasi-affine" curve.

2. Cauchy-Crofton's theorem. Inconstancy of a sequence

Let Γ be a plane curve. Let $\ell(\Gamma)$ denote its length and let $\delta(\Gamma)$ denote the perimeter of its convex hull. Any straight line in the plane can be defined as the set of (x, y) such that $x\cos\theta + y\sin\theta - \rho = 0$, where θ belongs to $[0, \pi)$ and ρ

is a real number, and hence is completely determined by (ρ, θ). Letting μ denote the Lebesgue measure on the set $\{(\rho, \theta),\ \rho \geq 0,\ \theta \in [0, \pi)\}$, the *average* number of intersection points between the curve Γ and straight lines is defined to be the quantity

$$\int_{D \in \Omega(\Gamma)} \sharp(\Gamma \cap D) \frac{\mathrm{d}\rho\ \mathrm{d}\theta}{\mu(\Omega(\Gamma))}$$

where $\Omega(\Gamma)$ is the set of straight lines which intersect Γ.

The following result can be found in [10, p. 184–185], see also the papers of Cauchy [8, 9].

Theorem 1 (Cauchy-Crofton)**.** *The average number of intersection points between the curve* Γ *and the straight lines in* $\Omega(\Gamma)$ *satisfies the equality*

$$\int_{D \in \Omega(\Gamma)} \sharp(\Gamma \cap D) \frac{\mathrm{d}\rho\ \mathrm{d}\theta}{\mu(\Omega(\Gamma))} = \frac{2\ell(\Gamma)}{\delta(\Gamma)}.$$

Remark 2. The reader will have noted the relation between this theorem and the *Buffon needle* problem (see [7, p. 100–104]).

The theorem of Cauchy-Crofton leads us to the following definition.

Definition 3. Let Γ be a plane curve. Let $\ell(\Gamma)$ be its length and $\delta(\Gamma)$ the perimeter of its convex hull. The *inconstancy* of the curve Γ, denoted $\mathcal{I}(\Gamma)$, is defined by

$$\mathcal{I}(\Gamma) := \frac{2\ell(\Gamma)}{\delta(\Gamma)}.$$

Remark 4. The minimal value of $\mathcal{I}(\Gamma)$ is 1. It is obtained in particular when Γ is a segment.

3. First results

In [4] we compare inconstancy with residual variance for very simple sequences. Then we compute the inconstancy of classical infinite sequences. We give some of our results below.

Theorem 5. *Let* $(u_n)_{n\geq 0}$ *be an infinite sequence taking two values* 0 *and* $h > 0$*, with* $u_0 = 0$*. We make the assumption that the frequencies of occurrences of the blocks* 00*,* hh*,* $0h$*,* $h0$ *in the sequence exist and are respectively equal to* $\mathcal{F}_{00}, \mathcal{F}_{hh}, \mathcal{F}_{0h}, \mathcal{F}_{h0}$*. Then*

$$\mathcal{I}((u_n)_{n\geq 0}) = \mathcal{F}_{00} + \mathcal{F}_{hh} + (\sqrt{1+h^2})(\mathcal{F}_{0h} + \mathcal{F}_{h0}) = 1 + (\sqrt{1+h^2} - 1)(\mathcal{F}_{0h} + \mathcal{F}_{h0}).$$

Remark 6. A similar result holds with sequences taking any finite number of values.

Corollary 7. *We have in particular the following results for binary sequences (taking only values* 0 *and* 1*). Let* $((m_n)_{n\geq 0})$ *be the Thue-Morse sequence; let*

$((r_n)_{n\geq 0})$ be the Shapiro-Rudin sequence; let $((z_n)_{n\geq 0})$ be the regular paperfolding sequence.

$$\begin{array}{lcll}
\mathcal{I}((01)^\infty) & = & \sqrt{2} & = 1.414... \\
\mathcal{I}((0^21)^\infty) & = & \frac{1+2\sqrt{2}}{3} & = 1.276... \\
\mathcal{I}((m_n)_{n\geq 0}) & = & \frac{1+2\sqrt{2}}{3} & = 1.276... \\
\mathcal{I}((0^31)^\infty) & = & \frac{1+\sqrt{2}}{2} & = 1.207... \\
\mathcal{I}((r_n)_{n\geq 0}) & = & \frac{1+\sqrt{2}}{2} & = 1.207... \\
\mathcal{I}((z_n)_{n\geq 0}) & = & \frac{1+\sqrt{2}}{2} & = 1.207... \\
\mathcal{I}((u_n)_{n\geq 0}) & = & \frac{1+\sqrt{2}}{2} & = 1.207... \quad \text{(for almost all sequences } u\text{)} \\
\mathcal{I}(0^\infty)) & = & 1. &
\end{array}$$

4. Possible applications

We began checking whether inconstancy is a pertinent measure of fluctuation, or even a prediction tool in different domains: variations of BMI (*body mass index*) and metabolic syndrome (in relation with cardio-vascular diseases, see, e.g., [18]), smoothness of musical themes, and fluctuations of the stockmarket.

References

[1] J.-P. Allouche, Sur la complexité des suites infinies, *Bull. Belg. Math. Soc.* **1** (1994) 133–143.

[2] J.-P. Allouche, Automates et algébricités, *J. Théor. Nombres Bordeaux* **17** (2005) 1–11.

[3] J.-P. Allouche, M. Baake, J. Cassaigne, D. Damanik, Palindrome complexity, *Theoret. Comput. Sci.* **292** (2003) 9–31.

[4] J.-P. Allouche, L. Maillard-Teyssier, Inconstancy of finite and infinite sequences, Preprint 2009, `http://arxiv.org/abs/0910.1173`

[5] J.-P. Allouche, J. Shallit, *Automatic sequences. Theory, applications, generalizations*, Cambridge University Press, Cambridge, 2003.

[6] S. V. Avgustinovich, D. G. Fon-Der-Flaass, A. E. Frid, Arithmetical complexity of infinite words, in M. Ito and T. Imaoka, editors, *Words, Languages & Combinatorics III*, ICWLC 2000, Kyoto, Japan, March 14-18, 2000, World Scientific Publishing, Singapore, 2003, pp. 51–62.

[7] G. L. Leclerc Buffon, *Histoire naturelle générale et particulière*, Supplément, Tome quatrième, Imprimerie Royale, 1777.

[8] A. Cauchy, Notes sur divers théorèmes relatifs à la rectification des courbes, et à la quadrature des surfaces, *C. R. Acad. Sci. Paris* **13** (1841) 1060–1063. (Also in *Œuvres complètes* **6**, Gauthier-Villars, Paris, pp. 369–375, 1888.)

[9] A. Cauchy, Mémoire sur la rectification des courbes et la quadrature des surfaces courbes, *Mém. Acad. Sci. Paris* **22** (1850) 3–15. (Also in *Œuvres complètes* **2**, Gauthier-Villars, Paris, pp. 167–177, 1908.)

[10] M. W. Crofton, On the theory of local probability, applied to straight lines drawn at random in a plane; the methods used being also extended to the proof of certain new theorems in the Integral Calculus, *Philos. Trans. R. Soc. Lond.* **158** (1868) 181–199.

[11] S. Ferenczi, Z. Kása, Complexity for finite factors of infinite sequences, *Theoret. Comput. Sci.* **218** (1999) 177–195.

[12] T. Kamae, L. Q. Zamboni, Sequence entropy and the maximal pattern complexity of infinite words, *Ergodic Theory Dynam. Systems* **22** (2002) 1191–1199.

[13] T. Kamae, L. Q. Zamboni, Maximal pattern complexity of discrete dynamical systems, *Ergodic Theory Dynam. Systems* **22** (2002) 1201–1214.

[14] L. Ilie, Combinatorial Complexity Measures for Strings, in *Recent advances in formal languages and applications*, Studies in Computational Intelligence, 2006, Volume 25/2006, pp. 149–170.
[15] L. Ilie, S. Yu, K. Zhang, Word complexity and repetitions in words, Computing and Combinatorics Conference—COCOON'02, *Internat. J. Found. Comput. Sci.* **15** (2004) 41–55.
[16] M. Li, P. Vitanyi, *An introduction to Kolmogorov complexity and its applications*, Third edition, Springer Verlag, 2008.
[17] C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol, *Acta Arith.* **82** (1997) 365–377.
[18] A.-C. Vergnaud, S. Bertrais, J.-M. Oppert, L. Maillard-Teyssier, P. Galan, S. Hercberg, S. Czernichow, Weight fluctuations and risk for metabolic syndrome in an adult cohort, *Int. J. Obesity* **32** (2008) 315–321.

Word combinatorics, *S*-adic sequences and multidimensional continued fractions

Valérie Berthé

This survey lecture is about the numerous occurrences of Euclid's algorithm and continued fraction algorithms (in their usual form, or in generalized versions) in word combinatorics. One motivation is the well-known and particularly fruitful interaction between Sturmian sequences, rotations of $\mathbb{T}^1$, and regular continued fractions. For generalizations, see also the survey [4].

As another motivation let us recall Fine and Wilf's theorem. This theorem gives a condition on the length of the periods a finite word can have. More precisely, if w is a word having periods p and q with length greater than or equal to $p + q - \gcd(p, q)$, then w has period $\gcd(p, q)$. Asume now p and q coprime. The family of words with length $p + q - 2$ that are p and q periodic is particularly interesting. Such extremal words (with respect to Fine and Wilf's theorem) are known to be particular factors of Sturmian words, and their study involves once again Euclid's algorithm. For more details, see [9] and the references therein.

There exist two natural types of generalizations of Fine and Wilf's theorem, either by extending the size of the alphabet [7], or by considering multidimensional words [12]. Extremal words for these generalizations can also be described in terms of multidimensional continued fraction algorithms. In particular, in the former case, an algorithm in the flavour of the fully subtractive algorithm [11] allows the construction of extremal words [13]. See also the lecture by A. de Luca which evokes duality properties for Christoffel words and generalizations, obtained by reversing the corresponding generalized Euclid's algorithm.

The connection between word combinatorics and multidimensional continued fractions is particularly striking within the so-called S-adic framework. Let us recall that a substitution is a non-erasing morphism of the free monoid. A sequence is said to be S-adic if it is generated by an infinite composition of a finite number of substitutions. S-adic sequences generalize in a natural way substitutive sequences. The S-adic expansion of a Sturmian word can be described thanks to the continued fraction expansion of its slope [5]. More generally, infinite words having an at most linear number of factors of a given length (they are said to be

of linear complexity) are known to be S-adic, if they are furthermore assumed to be minimal [8]. For more details, see also the lecture by J. Cassaigne. This covers various families of infinite words with a rich dynamical behaviour. In order to understand the geometric and symbolic nature of the dynamical systems that are generated by such infinite words, we are mainly interested in the two following problems: first, finding geometric interpretations of various symbolic dynamical systems including those generated by substitutions or by S-adic generation, and secondly, developing multidimensional continued fraction algorithms reflecting the dynamics of the systems.

Several combinatorial questions can be formuated in an efficient way in this S-adic/continued fraction framework. Given an S-adic sequence, one can ask whether this sequence is substitutive, that is, whether it is a letter-to-letter projection of a fixed point of a substitution. Substitutive Sturmian sequences correspond to quadratic angles (for more details, see [5]). This result can be considered as a version of Galois' theorem for continued fraction expansions. See also [6] for a connected result: if all the parameters of an interval exchange belong to the same quadratic extension, the sequence of induced interval exchanges (by performing always the same induction process) is ultimately periodic.

More generally, convergence issues (and Diophantine approximation properties) for a multidimensional continued fractions algorithm underlying a family of infinite words correspond to the question of convergence toward frequencies of factors, which can themselves be expressed in measure-theoretic terms (in particular if one has unique ergodicity).

The study of S-adic words thus leads to numerous questions that are of a combinatorial, arithmetic or else dynamical nature. Among them, the so-called S-adic conjecture aims at finding a characterization of infinite words having linear complexity in S-adic terms. This conjecture is still open. Note that during the lecture, the following interesting decision problem has been raised by J. Shallit: Given a finite set of substitutions $\phi_1, \phi_2, \cdots, \phi_n$, and a word w, decide if w appears as a factor of some S-adic infinite word generated by the ϕ_i.

References

[1] P. Arnoux, S. Ito, *Pisot substitutions and Rauzy fractals*, Bull. Bel. Math. Soc. Simon Stevin **8** (2001), 181–207.

[2] P. Arnoux, S. Ito, Y. Sano, *Higher dimensional extensions of substitutions and their dual maps*, J. Anal. Math. **83** (2001), 183–206.

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

[4] V. Berthé, S. Ferenczi, L.Q. Zamboni Interactions between dynamics, arithmetics, and combinatorics: the good, the bad, and the ugly, Algebraic and Topological Dynamics, S. Kolyada, T. Manin, and T. Ward eds., Contemporary Mathematics (CONM), AMS, American Mathematical Society, 385 (2005), 333–364.

[5] V. Berthé, C. Holton, L. Q. Zamboni, *Initial powers of Sturmian sequences*, Acta Arith. **122** (2006), 315–347.

[6] M. D. Boshernitzan, C. R. Carroll, *An extension of Lagrange's theorem to interval exchange transformations over quadratic fields*, J. Anal. Math. **72** (1997), 21–44.

[7] M. G. Castelli, F. Mignosi, A. Restivo, *Fine and Wilf's theorem for three periods and a genereralization of Sturmian words*, Theoret. Comput. Sci. **218** (1999), 83–94.
[8] S. Ferenczi, *Rank and symbolic complexity*, Ergodic Theory Dynam. Systems **16** (1996) 663–682.
[9] M. Lothaire, *Algebraic combinatorics on words*, Cambridge University Press (2002).
[10] N. Pytheas Fogg, *Substitutions in dynamics, arithmetics and combinatorics*, Lecture Notes in Mathematics, **1794**, Edited by V. Berthé, S Ferenczi, C.Mauduit and A. Siegel, Springer-Verlag (2002).
[11] F. Schweiger, *Multi-dimensional continued fractions*, Oxford Science Publications, Oxford Univ. Press, Oxford (2000).
[12] R.J. Simpson, R.Tijdeman *Multi-dimensional versions of a theorem of Fine and Wilf and a formula of Sylvester*, Proc. Amer. Math. Soc. **131** (2003), 1661-1667.
[13] R. Tijdeman, L. Q. Zamboni, *Fine and Wilf words for any periods II.*, Theor. Comput. Sci. **410** (2009), 3027–3034.

Words of very low factor complexity

Julien Cassaigne

Let $u \in A^{\mathbb{N}}$ be an infinite word. The *factor complexity* of u is the function $p\colon \mathbb{N} \to \mathbb{N}$ defined by: $p(n)$ is the number of words of length n occurring in u (*factors* of u). Morse and Hedlund [5] proved that $p(n) \geq n+1$ for non-eventually-periodic words. Words for which $p(n) = n+1$ are called Sturmian words. Words for which $p(n) = n + c$ for some constant c can be deduced from them [3]. We are interested in words "just above" this, roughly $n + 1 \leq p(n) \leq 2n$. Let $\alpha_u = \liminf \frac{p(n)}{n}$ and $\beta_u = \limsup \frac{p(n)}{n}$, and $\Omega = \{(\alpha_u, \beta_u)\colon u \in A^{\mathbb{N}}\} \subseteq (\mathbb{R}^+ \cup \{+\infty\})^2$. Then the general problem, essentially open, is: what is the structure of Ω?

Heinis proved [4] that $\beta - \alpha \geq \frac{(2-\alpha)(\alpha-1)}{\alpha}$. In particular, $1 < \alpha = \beta < 2$ is impossible.[1] Aberkane [1] contructed a sequence of points of Ω converging to $(1, 1)$; on the other hand $(\frac{3}{2}, \frac{5}{3})$ seems to be an isolated point.

The main tool to study these words is the sequence of *Rauzy graphs*: Γ_n is the directed graph with vertices $L_n(u)$ (the factors of length n of u) and edges $L_{n+1}(u)$, with an edge from x to y labelled with z if and only if $z \in xA \cap Ay$. For Sturmian words, only two shapes of graphs are possible. For recurrent words with $p(n) \leq \frac{4}{3}n + 1$, two new shapes appear. Such a word is then defined (up to shift, etc.) by a path in the "graph of graphs", and some of its properties (α and β, frequencies, etc.) may be deduced from this path. The path also provides an *s-adic representation* (infinite composition of substitutions), which can be viewed as a generalized continued fraction expansion.

References

[1] Ali Aberkane, Words whose complexity satisfies $\lim \frac{p(n)}{n} = 1$, *Theoret. Comput. Sci* **307** (2003), 31–46.
[2] Julien Cassaigne and François Nicolas, Factor complexity, in *Combinatorics, automata and number theory*, éd. V. Berthé et M. Rigo, Cambridge University Press, 2010, 163–247.

[1] this is generalized in [2]: if $\lim \frac{p(n)}{n}$ exists, then it must be an integer.

[3] Ethan M. Coven, Sequences with minimal block growth II, *Math. Systems Theory* **8** (1975), 376–382.
[4] Alex Heinis, The $P(n)/n$ function for bi-infinite words, *Theoret. Comput. Sci* **273** (2002), 35–46.
[5] Marston Morse and Gustav A. Hedlund, Symbolic Dynamics II. Sturmian trajectories, *American J. Math.* **62** (1940), 1–42.

Power-free Sequences: Topology, Reachability and Curling Numbers

James D. Currie

A word is **repetitive** if it contains two identical blocks. A word containing k consecutive identical blocks is said to contain a k **power**. Dejean [5] also introduced the study of words containing **fractional** k **powers**.

Thue [11] showed that there are infinite words over the three letter alphabet $\{a, b, c\}$ which are non-repetitive. Infinite nonrepetitive words (sequences) have been used to build counter-examples in algebra [8], ordered sets [12], symbolic dynamics [7] and other areas. A non-empty set L of infinite words is **perfect** if for any $u \in S$ and any n there is a word $v \in L$, $v \neq u$ such that u and v have a common prefix of length at least n.

Open problem 1. *Let L be the set of infinite words over Σ which avoid pattern p. Is L perfect?*

Abusing notation, consider L to contain also the finite factors of its words. We consider the partial order where $u < v$ iff u is a prefix of v. The diagram of L under this order is a tree with root ϵ, the empty word. We will identify L with this tree. The **meet** of finite words u, v is their longest common prefix, denoted by $u \wedge v$.

Given words $u \leq v$ in L, the closed interval $[u, v]$ is the set $\{w \in L : u \leq w \leq v\}$. For notational convenience, we also define $[u, \infty] = \{w \in L : u \leq w\}$. Suppose that $u < v$ and

(1) $[\hat{v}, \infty]$ is infinite for at most one upper cover $\hat{v}$ of v.
(2) $[u, \infty] - [v, \infty]$ is finite.

In this case any path in L from u to ∞ must traverse the vertices of $[u, \hat{v}]$. We refer to the set $B_{\hat{v}}(u, v) = B(u, v) = [u, \infty] - [\hat{v}, \infty]$ as a **bottleneck**. The **length** of $B = B(u, v)$ is $|B| = |v| - |u| + 1$. The **index** of B is $\iota(B) = |u|$. Suppose that $B(u, v)$ is a bottleneck and $\hat{u} < \hat{v}$ are elements of $[u, v]$. It follows that at most one cover of $\hat{v}$ has an infinite extension and we can form a bottleneck $B(\hat{u}, \hat{v})$.

We consider the case where L is the language of square-free words over a three-letter alphabet. Long bottlenecks in L must occur far out, i.e. for a bottleneck B of L

$$\iota(B) \geq f(|B|), \text{ whenever } |B| > N_0. \tag{1}$$

Here N_0 is some constant, while f is eventually increasing and unbounded. Let g be a function such that for any non-negative integer M we have $f(x) > M$ whenever $x > g(M)$.

Theorem 8. *Language L is infinite if and only if it contains a word of length* $\max(N_0, g(0))$.

This is a special case ($u = \epsilon$) of the following:

Theorem 9. *Word $u \in L$ is a prefix of infinitely many words in L if and only if L contains a word $v > u$, with $|v| - |u| = \max(N_0, g(|u|))$.*

Inequality (1) can also be used to show that L is perfect.

Theorem 10. *If L is infinite, then L is perfect.*

Thus the infinite tree L is constantly branching. The proof is readily sharpened to put a bound on how far L can go without branching.

Theorem 11. *Suppose that u has an infinite extension in L. Then there is a word $v > u$, $|v| \leq |u| - 1 + g(|u|)$ such that v is the meet of two infinite extensions of u in L.*

Function f can be taken to have the form $f(x) = ax^{3/2}$, some positive constant a. Thus g can be taken to be $g(x) = \max(N + 1, (x/a)^{2/3})$. This implies the following:

Corollary 12. *The set of nonrepetitive words over $\{1, 2, 3\}$ of length n grows exponentially.*

A striking aspect is that all of this structural information about L is demonstrated non-constructively! [2, 3, 4] We will argue that these non-constructive methods should be used to attack the following three open problems:

Open problem 2. *(Restivo/Salemi) A reachability problem* [9]*: Suppose that u is a prefix of infinitely many words of L, and v is a suffix of infinitely many words of L. Does L contain a word of the form uwv?*

Open problem 3. *(Sloane et al.) Curling numbers* [1]*: Given a finite word w, the* **curling number** *of w is the largest integer n such that w can be written uv^n for some words u and v. Starting with any word w, we can form the* **curling number sequence** *of w: We let $w_0 = w$, and w_{i+1} is formed by appending to w_i its curling number. The conjecture is that if one starts with any finite word and begins to form the curling number sequence, one will eventually reach a 1.*

Open problem 4. *(Guay-Paquet and Shallit) Lexicographically least words* [6]*: It is conjectured that the lexicographically least infinite word over $\mathbb{N}$ avoiding $5/2$ powers uses only three letters.*

References

[1] Benjamin Chaffin & N. J. A. Sloane, The Curling Number Conjecture, arXiv:0912.2382.

[2] James. D. Currie, On the structure and extendibility of k-power free words, *Eur. J. Comb.* (1995) **16**, 111–124.

[3] James. D. Currie and Robert O. Shelton, Cantor sets and Dejean's conjecture, *J. Automata, Languages and Combin.* **1** (1996), 113–128.

[4] James. D. Currie and Robert O. Shelton, On the structure and extendibility of k-power free words II, submitted.
[5] Françoise Dejean, Sur un théorème de Thue, *J. Combin. Theory Ser. A* **13** (1972), 90–99.
[6] Mathieu Guay-Paquet & Jeffrey Shallit, Avoiding squares and overlaps over the natural numbers, *Discrete Mathematics* **309** (2009), 6245-6254.
[7] Marston Morse & Gustav A. Hedlund, Symbolic dynamics I, II, Amer. J. Math. **60** (1938), 815–866; **62** (1940) 142; MR **1**, 123d.
[8] P. S. Novikov & S. I. Adjan, Infinite periodic groups I, II, III, Izv. Akad. Nauk. SSSR Ser. Mat. **32** (1968), 212–244;251–524;709–731;MR **39** #1532a–c.
[9] Antonio Restivo & Sergio Salemi Words and Patterns *Developments in Language Theory 2001* 117–129.
[10] Robert O. Shelton, On the structure and extendibility of square-free words, *Combinatorics on Words: Progress and Perspectives* (1983) Academic Press, 101–188.
[11] Axel Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.* Christiana (1906), 1–22.
[12] William T. Trotter & Peter Winkler, Arithmetic Progressions in Partially Ordered Sets, *Order* **4** (1987) 37–42.

On a palindromization map on free monoids

Aldo de Luca

The right-palindromic closure of a word u of a free monoid A^* over the alphabet A, is the shortest palindrome $u^{(+)}$ of A^* having u as a prefix. In a seminal paper [6] of 1997, the author introduced in the case of a binary alphabet, an injective map, called *palindromization map*, associating to each word v a palindrome by an iterated application, 'directed' by the word v, of the right palindromic closure operator. More precisely, the palindromization map $\psi : A^* \to PAL$, where PAL is the set of palindromes of A^*, is inductively defined as: $\psi(\varepsilon) = \varepsilon$ (ε denotes the empty word) and for all $u \in A^*$ and $x \in A$,

$$\psi(ux) = (\psi(u)x)^{(+)} .$$

For all words v if u is a prefix of v, then $\psi(u)$ is a palindromic prefix (and suffix) of $\psi(v)$ and, conversely, every palindromic prefix of $\psi(v)$ is of the form $\psi(u)$ for some prefix u of v. For any $w \in \psi(A^*)$ the unique word u such that $\psi(u) = w$ is called the *directive word* of w. For instance, if $A = \{a, b, c\}$ and $v = aabc$, one has $\psi(a) = a$, $\psi(aa) = aa$, $\psi(aab) = (aab)^{(+)} = aabaa$, and $\psi(aabc) = (aabaac)^{(+)} = aabaacaabaa$.

It was proved in [6] that if the palindromization map is extended to infinite binary directive words such that each letter occurs infinitely many times in them, then one can construct all *standard Sturmian words*.

If one extends the action of palindromization map to infinite words over arbitrary finite alphabets, one can generate a wider class of words, called *standard episturmian*, introduced in 2001 by X. Droubay, J. Justin, and G. Pirillo in [12]. This class includes standard Sturmian words and Arnoux-Rauzy words [1]. In this extension of Sturmian words some properties are lost (for instance, the *balance* property) and other are preserved (for instance, the *richness* in palindromes of

their factors). In any case episturmian words satisfy very interesting combinatorial and structural properties and, in fact, many papers have been written on the subject (see, for instance, the overview papers [2, 13]).

In this theory a key role is played by the class of palindromic prefixes of all standard episturmian words over a given alphabet called *epicentral words*, and simply *central* in the case of a binary alphabet [15]. These words are precisely the images of a finitely generated free monoid by the palindromization map. Epicentral words satisfy interesting combinatorial properties since they can have several different representations. In fact, besides directive words which are related to the palindromization map, epicentral words can be represented by periods (period vector) and composition (Parikh vector). Further important representations can be done by using matrices, trees, and graphs [10, 11].

The previous representations are also useful for the problem of counting the epicentral words and the palindromes of any length in all episturmian words over a given alphabet [5]. For any k, the map P_k which counts for any n the epicentral words of length n on the k-letter alphabet, is a suitable extension to the case $k > 2$ of Euler's totient function φ. Indeed, for $k = 2$, $P_2(n) = \phi(n + 2)$[9]; for $k > 2$ a general arithmetic interpretation for $P_k(n)$ in terms of a multidimensional generalization of the Euclidean algorithm is in [16] (see also [10]). The behavior of the map P_k is quite irregular and oscillating. Some conjectures based on a table of numerical values of P_k ($3 \leq k \leq 6$ and $1 \leq n \leq 500$) are formulated. In [5] a formula for the map g_k counting for any n the palindromes of length n in all episturmian words over a k letter alphabet is given. This formula for g_k, extending a result found in [7] in the case $k = 2$, depends on the map P_k.

The palindromization map has been recently extended to the case of free-group F_2 by C. Kassel and C. Reutenauer [14]. Moreover, in [8] a (right) θ-palindromic closure operator, where ϑ is any involutory antimorphism of a free monoid, has been introduced. The fixed points of ϑ are called *ϑ-palindromes*. For any word u the ϑ-palindromic closure of u is the shortest ϑ-palindrome having u as a prefix. Similarly to the case of the reversal operator, a *ϑ-palindromization map* can be defined; it associates to each word v a ϑ-palindrome by an iterated application of ϑ-closure operator 'directed' by the word v. By acting with this operator on any infinite directive word one obtains a class of words larger than the class of standard episturmian, that we called *θ-standard words* (or simply *pseudostandard* if one does not refer to a particular ϑ); when θ is the reversal operator one obtains the class of standard episturmian words.

In [4, 3] two more general families of words have been introduced. The first is the family of the *ϑ-standard words with seeds*, that is the words obtained by iteration of the operator of θ-palindromic closure starting with a non-empty word called *seed*. A second family of words called *generalized pseudostandard words*, is formed by the pseudostandard words directed by 2-words: the directive word and a word describing the antimorphism to use at each iteration.

References

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

[2] J. Berstel, Sturmian and Episturmian words (A survey of some recent results), Lecture Notes in Computer Science, vol. 4728, Springer-Verlag Berlin 2007, pp. 23–47

[3] M. Bucci, A. de Luca, A. De Luca, L. Zamboni, On different generalizations of episturmian words, *Theoretical Computer Science* **393** (2008) 23–36]

[4] M. Bucci, A. de Luca, A. De Luca, L.Q. Zamboni, On some problems related to palindrome closure, *Theoretical Informatics and Applications* **42** (2008) 679–700

[5] M. Bucci, A. de Luca, A. De Luca, On the number of episturmian palindromes, *Theoretical Computer Science* **411** (2010) 3668–3684

[6] A. de Luca, Sturmian words: Structure, Combinatorics, and their Arithmetics, *Theoretical Computer Science* **183** (1997) 45–82

[7] A. de Luca, A. De Luca, Combinatorial properties of Sturmian palindromes, *International Journal of Foundations of Computer Science* **17** (2006) 557–573

[8] A. de Luca, A. De Luca, Pseudopalindrome closure operators in free monoids, *Theoretical Computer Science* **362** (2006) 282–300

[9] A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, *Theoretical Computer Science* **136** (1994) 361–385

[10] A. de Luca, L. Q. Zamboni, On graphs of central episturmian words, *Theoretical Computer Science* **411** (2010) 70–90

[11] A. de Luca, L. Q. Zamboni, Involutions of epicentral words, *European Journal of Combinatorics* **31** (2010) 867–886

[12] X. Droubay, J. Justin, and G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, *Theoretical Computer Science* **255** (2001) 539–553

[13] A. Glen, J. Justin, Episturmian words: A survey, *Theoretical Informatics and Applications* **43** (2009) 403–442

[14] C. Kassel, C. Reutenauer, A palindromization map for the free group, *Theoretical Computer Science* **409** (2008) 461–470

[15] M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press (Cambridge, 2002)

[16] F. Mignosi, L.Q. Zamboni, On the number of Arnoux-Rauzy words, *Acta Arithmetica* **101** (2002) 121–129

On (almost) rich words

Amy Glen

In recent years there has been growing interest in palindromes in the field of *combinatorics on words*, especially since the work of A. de Luca [8] and also X. Droubay and G. Pirillo [10], who showed that the well-known *Sturmian words* are characterised by their *palindromic complexity* [1, 3, 5]. A strong motivation for the study of palindromes, and in particular infinite words containing arbitrarily long palindromes, stems from applications to the modelling of *quasicrystals* in theoretical physics (see for instance [7, 17]) and to Diophantine approximation (e.g., see [14]).

In [9], X. Droubay, J. Justin, and G. Pirillo observed that any finite word w of *length* $|w|$ contains at most $|w|+1$ distinct palindromes (including the empty word). Even further, they proved that a word w contains exactly $|w|+1$ distinct palindromes if and only if the longest palindromic suffix of any prefix p of w occurs exactly once in p (i.e., every prefix of w has *Property Ju* [9]). Such words

are 'rich' in palindromes in the sense that they contain the maximum number of different palindromic factors. Accordingly, we say that a finite word w is *rich* if it contains exactly $|w| + 1$ distinct palindromes (or equivalently, if every prefix of w has *Property Ju*). Naturally, an infinite word is rich if all of its factors are rich. In independent work, P. Ambrož, C. Frougny, Z. Masáková, and E. Pelantová have considered the same class of words which they call *full words* in [2], following the earlier work of S. Brlek, S. Hamel, M. Nivat, and C. Reutenauer in [5].

In [9], X. Droubay *et al.* also showed that the family of *episturmian words* [9, 18], which includes the well-known *Sturmian words*, comprises a special class of rich infinite words. Specifically, they proved that if an infinite word $\boldsymbol{w}$ is episturmian, then any factor u of $\boldsymbol{w}$ contains exactly $|u| + 1$ distinct palindromic factors. (See [4, 15, 19] for recent surveys on the theory of Sturmian and episturmian words.) Another special class of rich words consists of S. Fischler's sequences with "abundant palindromic prefixes", which were introduced and studied in [13] in relation to Diophantine approximation (see also [14]). Other examples of rich words have appeared in many different contexts; they include the *complementation-symmetric Rote sequences* [1], certain words associated with β-expansions where β is a *simple Parry number* [2, 6], and symbolic codings of trajectories of symmetric interval exchange transformations [11, 12].

The first unified study of combinatorial and structural properties of rich words was carried out by myself and J. Justin, published in a joint paper with S. Widmer and L.Q. Zamboni who studied a wider class of words for which successive occurrences of any letter are separated by palindromes (called *weakly rich words*) – see [16].

In this talk, I will begin by giving a brief overview of some fundamental properties of rich words. In particular, I will show that rich words are characterised by the property that all *complete returns* to any palindromic factor are palindromes. I will also give a more explicit description of periodic rich infinite words. I will then discuss so-called *almost rich words*: they are infinite words for which only a finite number of prefixes do not satisfy *Property Ju*. Such words can also be defined in terms of the *defect* of a finite word w, which is the difference between $|w| + 1$ and the number of distinct palindromic factors of w (see the work of Brlek *et al.* in [5] where periodic infinite words with finite defect are characterised). With respect to this notion, rich words are those with defect 0 and almost rich words are infinite words with finite defect.

Lastly, I will consider the action of morphisms on (almost) rich words, with particular interest in morphisms that preserve (almost) richness. We have shown that such morphisms belong to the class of *P-morphisms* that was introduced by A. Hof, O. Knill, and B. Simon in [17] (see also the nice survey on palindromic complexity by J. Allouche *et al.* [1]), but it remains an **open problem** to characterise them. This is related to the following long-standing open question posed in [17]: are there (uniformly recurrent) infinite words containing arbitrarily long palindromes that arise from primitive morphisms, none of which belongs to class

P? The answer is believed to be no. Up until now, it has only been shown to hold in the periodic case (see [1]) and also in the 2-letter case (see [20]).

References

[1] J.-P. Allouche, M. Baake, J. Cassaigne, D. Damanik, Palindrome complexity, *Theoret. Comput. Sci.* 292 (2003) 9–31.

[2] P. Ambrož, C. Frougny, Z. Masáková, E. Pelantová, Palindromic complexity of infinite words associated with simple Parry numbers, *Ann. Inst. Fourier (Grenoble)* 56 (2006) 2131–2160.

[3] P. Baláži, Z. Masáková, E. Pelantová, Factor versus palindromic complexity of uniformly recurrent infinite words, *Theoret. Comput. Sci.* 380 (2007) 266–275.

[4] J. Berstel, Sturmian and episturmian words (A survey of some recent results), in: *Proceedings of CAI 2007*, Lecture Notes in Computer Science, vol. 4728, 2007, pp. 23–47.

[5] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, On the palindromic complexity of infinite words, *Internat. J. Found. Comput. Sci.* 15 (2004) 293–306.

[6] M. Bucci, A. De Luca, A. Glen, L.Q. Zamboni, A connection between palindromic and factor complexity using return words, *Adv. in Appl. Math.* 42 (2009) 60–74.

[7] D. Damanik, L.Q. Zamboni, Combinatorial properties of Arnoux-Rauzy subshifts and applications to Schrödinger operators, *Rev. Math. Phys.* 15 (2003) 745–763.

[8] A. de Luca, Sturmian words: structure, combinatorics and their arithmetics, *Theoret. Comput. Sci.* 183 (1997) 45–82.

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

[10] X. Droubay, G. Pirillo, Palindromes and Sturmian words, *Theoret. Comput. Sci.* 223 (1999) 73–85.

[11] S. Ferenczi, L.Q. Zamboni, Language of k-interval exchange transformations, *Bull. London Math. Soc.* 40 (2008) 705–714.

[12] S. Ferenczi, L.Q. Zamboni, Structure of k-interval exchange transformations: induction, trajectories, and distance theorems, Preprint (2008), `http://iml.univ-mrs.fr/~ferenczi/fz1.pdf`

[13] S. Fischler, Palindromic prefixes and episturmian words, *J. Combin. Theory Ser. A* 113 (2006) 1281–1304.

[14] S. Fischler, Palindromic prefixes and diophantine approximation, *Monatsh. Math.* 151 (2007) 11–37.

[15] A. Glen, J. Justin, Episturmian words: a survey, *Theoret. Inform. Appl.* 43 (2009) 402–433.

[16] A. Glen, J. Justin, S. Widmer, L.Q. Zamboni, Palindromic Richness, *European J. Combin.* 30 (2009) 510–531.

[17] A. Hof, O. Knill, B. Simon, Singular continuous spectrum for palindromic Schrödinger operators, *Comm. Math. Phys.* 174 (1995) 149–159.

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

[19] M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, 2002.

[20] B. Tan, Mirror substitutions and palindromic sequences, *Theoret. Comput. Sci.* 389 (2007) 118–124.

Weinbaum Factorizations

Tero Harju
(joint work with Volker Diekert, Dirk Nowotka)

C. M. Weinbaum [2] proved that for any primitive word w and a letter a occurring in w, there exists a conjugate of w that has a decomposition as $w' = uv$ such that (1) $u \in aA^* \cap A^*a$ but $v \notin aA^* \cup A^*a$, and (2) u and v have *unique positions* in w as cyclic factors, i.e., there is exactly one conjugate of w having u as a prefix and only one conjugate of w having v as a prefix.

Taken alone both conditions (1) and (2) are easily seen to be satisfied. For instance, given that $w = baababbabaa$, we find that the decompositions $w' = (aa)(babbabaab)$ and $w' = (aaba)(bbabaab)$ both satisfy (1) but not (2). On the other hand, the decomposition $w' = (abba)(baabaab)$ satisfies both conditions (1) and (2).

The following shows that the condition (2) is always satisfied.

Lemma 13. *Let $w = uv$ be a Lyndon word where v is the maximum suffix of w. Then u and v are uniquely positioned in w. Moreover, if v' is a cyclic factor of w such that $v \lhd v'$, then $v \leq_{\mathrm{p}} v'$.*

A simple proof of Weinbaum's result is given in Diekert, Harju and Nowotka [1], where also the following generalization of the result are proven.

Let w be a primitive word, and f, g its factors. A factorization $w' = uv$ of a conjugate of w is a *Weinbaum factorization of w for f and g*, if u and v are uniquely positioned in cyclic w and

$$u \in (fA^* \cap A^*f) \setminus (gA^* \cup A^*g),$$
$$v \in (gA^* \cap A^*g) \setminus (fA^* \cup A^*f).$$

Note that if $w' = uv$ is a Weinbaum factorization of w in the original setting, then it is a Weinbaum factorization of w for a and v.

Let now w be a primitive word, f a proper factor of w, and define

$$\begin{aligned} G(f) =& \{g \mid |fg| \leq |w|,\ fgf \text{ is a cyclic factor of } w^2,\ fgf \notin A^+fA^+\}, \\ R(f) =& \{g \in G(f) \mid g \text{ does not occur in any other element of } G(f), \text{ and} \\ & \quad f \text{ and } g \text{ do not intersect in } w\}. \end{aligned}$$

A word f is called a *Weinbaum factor* of w, if $R(f) \neq \varnothing$.

Theorem 14 ([1])**.** *Let w be a primitive word and let f be a Weinbaum factor of w with $g \in R(f)$. Then w has a Weinbaum factorization for f and g.*

Let then R^i denote the i-th iteration of the operation R. We can show that either $R^i(g) = R^{i+2}(g)$ or the set $R^{i+2}(g)$ contains a word having length at least twice the length of a word in $R^i(g)$. Hence, $R^i(g) = R^{i+2}(g)$ for some $i \leq 2\log_2(n)$. The bound can be improved so that we obtain the following result, where $\Phi = (1+\sqrt{5})/2$ is the golden ratio.

Theorem 15. *Let w be a primitive word with a Weinbaum factor f, and let $g \in R(f)$. If $2\ell \geq \log_\Phi(n)$, then $R^{2\ell-1}(g) \neq \varnothing$, for every $u \in R^{2\ell-1}(g)$, the set $R(u)$ is a singleton; and for $R(u) = \{v\}$ we obtain $R(v) = \{u\}$ and a Weinbaum factorization of $w = uv$ for f and g.*

References

[1] V. Diekert, T. Harju, and D. Nowotka, *Weinbaum factorizations for primitive words*, Izv. Vyssh. Uchebn. Zaved. Mat. (2010), no. 1, 21–33, In English: Russian Mathematics (Iz VUZ) 54 (1), 16-25. MR 2664483

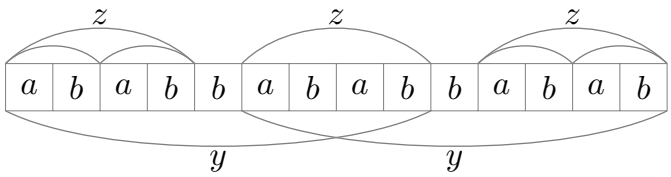
[2] C.M. Weinbaum, *Unique subwords in nonperiodic words*, Proc. Amer. Math. Soc. **109** (1990), no. 3, 615–619.

Periods and Unbordered Factors: The Ehrenfeucht-Silberger Problem

Štěpán Holub

(joint work with Dirk Nowotka)

The period of a word w, denoted by $\pi(w)$, is the length of the shortest word u such that w is a prefix of the infinite word $uuu\ldots$. Obviously, such a shortest word u is *primitive*, that is, it is not a power of a shorter word. An extremal situation is when $\pi(w) = |w|$, where $|w|$ denotes the length of w. In such a case, the word w is called *unbordered*, otherwise it is called *bordered*. The name is justified by the fact that a word w is bordered if and only if $w = uvu$ for some nonempty word u, called a *border* of w. More generally, a border of w is any nonempty word u that is both a prefix and a suffix of w, and $u \neq w$ allowing a border to be longer than the half of the length of w. However, it is easy to conclude (see the picture below) that any bordered word has a border shorter than the half of its length. It is also obvious that the shortest border of a word is itself unbordered.



If $w = uv$, then $\tilde{w} = vu = u^{-1}wu$ is called a *conjugate* of w. Any primitive word w has an unbordered conjugate w'; it is enough to consider a conjugate of w that is a *Lyndon word*, defined as the minimal conjugate w.r.t. to a chosen lexicographic order $\lhd$. The proof of the fact that any Lyndon word is unbordered illustrates how elegant and efficient proofs can be when using the properties of lexicographic orders. Suppose that $w = uvu$ is a Lyndon word w.r.t to $\lhd$. By definition, the word w is primitive, whence $uv \neq vu$ by the well known Periodicity lemma. If $uv \lhd vu$, then also $u(uv) \lhd u(vu)$; if, on the other hand, $vu \lhd uv$, then $(vu)u \lhd (uv)u$; a contradiction in both cases.

The above property of Lyndon words in particular shows that any word uu, with u primitive, contains an unbordered factor of length $|u|$. It is also obvious that any longer factor of uu is bordered. Therefore, we have $\tau(uu) = \pi(uu)$,

where $\tau(w)$ denotes the length of the longest unbordered factor of w. As already suggested, the inequality $\tau(w) \leq \pi(w)$ is obvious. Previous considerations also show that $\tau(w) = \pi(w)$ as soon as $|w| \geq 2\pi(w)$. Moreover, the multiplicative constant 2 is optimal: for the word $w = a^n b a^{n+1} b b a^{n+1} b a^n$ we have $\pi(w) = 2n+6$, $\tau(w) = 2n + 5$, and $|w| = 2\pi(w) - 4$.

It turns out that the question is much more difficult if we replace $\pi(w)$ by $\tau(w)$ obtaining the following problem:

The Ehrenfeucht-Silberger Problem: What is the smallest number c such that $|w| \geq c\,\tau(w)$ implies $\tau(w) = \pi(w)$?

History. The problem was raised by Ehrenfeucht and Silberger in 1979 [3]. They conjectured that $c = 2$, which was falsified shortly thereafter by Assous and Pouzet [1] by the following example:

$$w = a^n b a^{n+1} b a^n b a^{n+2} b a^n b a^{n+1} b a^n$$

where $n > 1$ and $\tau(w) = 3n + 6$ and $\pi(w) = 4n + 7$ and $|w| = 7n + 10$, that is, $\tau(w) < \pi(w)$ and $|w| = \frac{7}{3}\tau(w) - 4 > 2\tau(w)$. Assous and Pouzet in turn conjectured that $c = 3$. Duval [2] established in 1982 that $c \leq 4$ and made a conjecture for a special case: if w possesses an unbordered prefix of length $\tau(w)$, then $|w| \geq 2\tau(w)$ implies $\tau(w) = \pi(w)$. Duval's conjecture was only solved in 2004 [4] with a new proof given in [5]. The proof of Duval's conjecture lowered the bound for Ehrenfeucht and Silberger's problem to $c \leq 3$ reaching the value conjectured by Assous and Pouzet. However, there remained a gap of $\frac{2}{3}$ between that bound and the largest known example represented by the above Assous-Pouzet words. In 2009, Holub and Nowotka [6] proved that the bound $\frac{7}{3}$ is in fact optimal, solving the original Ehrenfeucht-Silberger problem.

The proof strategy. The proof has two main ideas. The first one is to factorize the studied word w as

$$v' u z u v,$$

where $|u|$ is the maximum length of an unbordered factor of w with at least two occurrences in w, and $|z|$ is the maximum distance between two such occurrences. In other words, uzu is the longest word with the longest possible shortest border.

The second idea of the proof is the concept of the *α-critical suffix*, which is best explained by the naive pattern matching algorithm. The algorithm scans a word w looking for an occurrence of α. At each position, the scanning continues until a mismatch with α is observed (or until α is found). When a mismatch occurs, the algorithm backtracks and starts to scan the next possible position. The α-critical suffix of w is defined as the position of the last mismatch.

The critical suffixes combined with the above mentioned factorization are an efficient tool for finding long unbordered factors of w. The machinery is likely to be useful even outside the Ehrenfeucht-Silberger problem. The factors found in the process induce seven constraints, which are shown to force either $|w| \leq \frac{7}{3}(\tau(w)-1)$ or $\tau(w) = \pi(w)$.

Open problems and challenges. Although the multiplicative constant $\frac{7}{3}$ is optimal, there remains space for improvement of the additive constant, bounded by -4 due to the Assous-Pouzet words.

Open problem 1: Does $|w| > \frac{7}{3}\tau(w) - 4$ imply $\tau(w) = \pi(w)$?

More generally:

Open problem 2: Find

$$\inf\{d \mid \text{there is a word with } |w| \geq \tfrac{7}{3}\tau(w) - d \text{ and } \tau(w) < \pi(w)\}.$$

The methods of the proof by Holub and Nowotka allow quite strong an insight into the structure of words satisfying $\tau(w) < \pi(w)$ and $|w| \doteq \frac{7}{3}\tau(w)$. The constraints suggest that Assous-Pouzet words can be the only words achieving the extremal bound. Whence the following problem.

Open problem 3: Is it true that any word satisfying $\tau(w) < \pi(w)$ and $|w| \geq \frac{7}{3}\tau(w) - 4$ is an Assout-Pouzet word?

Or, again more generally:

Open problem 4: For given δ, describe all words satisfying $\tau(w) < \pi(w)$ and $|w| \geq \frac{7}{3}\tau(w) - \delta$. What is the smallest δ such that there is more than one such a word for arbitrary fixed length?

References

[1] R. Assous and M. Pouzet, *Une caractérisation des mots périodiques*, Discrete Math. **25** (1979), no. 1, 1–5.

[2] J.-P. Duval, *Relationship between the period of a finite word and the length of its unbordered segments*, Discrete Math. **40** (1982), no. 1, 31–44.

[3] A. Ehrenfeucht and D. M. Silberger, *Periodicity and unbordered segments of words*, Discrete Math. **26** (1979), no. 2, 101–109.

[4] T. Harju and D. Nowotka, *Periodicity and unbordered words: A proof of the extended Duval conjecture*, J. ACM **54** (2007), no. 4.

[5] Š. Holub, *A proof of the extended Duval's conjecture*, Theoret. Comput. Sci. **339** (2005), no. 1, 61–67.

[6] Štěpán Holub and Dirk Nowotka, *On the relation between periodicity and unbordered factors of finite words*, Int. J. Found. Comput. Sci. **21** (2010), no. 4, 633–645.

Independent Systems of Word Equations and Related Topics

Juhani Karhumäki

Theory of word equations is a fundamental topic in Combinatorics on Words. In one hand it provides deep results – some jewels of discrete mathematics – and on the other hand amazing simply formulated problems. As a challenging example I urge the reader to conclude that the equation $x^2y^3x^2 = u^2v^3u^2$ has only periodic solutions, i.e. in any solution all unknowns are powers of a common word.

In this abstract we consider one fundamental property of word equations and, in particular, open problems related to that. The property is so-called *Ehrenfeucht compactness property*, formulated as:

> "Any system of word equations with a finite number of unknowns over a free monoid Σ^* is equivalent to some of its finite subsystems."

The property was formulated by A. Ehrenfeucht in 1970's in slightly different terms of formal languages. It was shown to hold by M. Albert and J. Lawrence [1] and simultaneously by G.S. Guba [6]. Actually, the result is a consequence of two well known facts: the existence of embeddings of free monoids into multiplicative monoids of integer matrices, and Hilbert's Basis Theorem for polynomial ideals, see e.g. [7].

The compactness property immediately proposes a question

> "What is the size of the above equivalent subsystem. Or more concretely can it be bounded by any function on the number of unknowns?"

This is the fundamental problem we are discussing here.

For clarity, we next recall the necessary terminology. Let Ξ be a finite set of variables and Σ a finite alphabet. We denote by Σ^* and Σ^+ the free monoid and free semigroup generated by Σ, respectively. Now, an *equation* is a pair $e = (u, v) \in \Xi^+ \times \Xi^+$ usually written as $u = v$. A *solution* of equation in Σ^* is a morphism $h : \Xi^* \to \Sigma^*$ satisfying $h(u) = h(v)$. These notions extend in a natural way to systems of equations.

Finally, we say that two systems of equations are *equivalent* if they possess exactly the same set of solutions, and a system is *independent* if it is not equivalent to any of its proper subsystems.

With these notions the compactness property can be reformulated:

> "Each independent system of equations with a finite number of unknowns over Σ^* is finite."

Accordingly, the second question, which is the main question in this presentation, asks:

> "Is the cardinality of the maximal independent system of equations on n unknowns bounded by a function on n?"

Actually, this question can be asked separately for each value of n. Really amazingly we do not know the answer even in the case $n = 3$! The case $n = 2$ is trivial.

Before we continue a few remarks are in order. Of course, the above questions can be asked for any semigroup or monoid, and not only for free ones. And the answer to the compactness question depends on the semigroup, see [8]. Particularly interesting is the case of commutative semigroups, which are in some sense complete opposites to free ones where no elements commute. For commutative monoids the compactness property does hold, but even in the case of only one unknown no bound for the size of the maximal independent system of equations exists, that is there are arbitrarily large, but finite, such systems, see [10]. This also implies indirectly that the known methods of using Hilbert's Basis Theorem to prove the compactness property cannot give a solution to our main question.

We start the other remark with an example.

Example 16. Let $\Xi = \{x, y, z\}$ and S the pair of equations $S : xyz = zyx, xyyz = zyyx$.

We leave it to the reader to figure out that this pair is independent. It also has a nonperiodic solution $x = z = a$ and $y = b$. A question is can we add into S a third equation with Ξ as the set of unknowns, such that it would still remain independent and possess a nonperiodic solution

In this case this is not possible, but we have the following amazing open problem from [2]:

> "Does there exist an independent system of three equations with three unknowns such that it possesses a nonperiodic solution?"

This problem has been studied quite intensively, see [4], [9], [5] or [3]. However, it is only conjectured that the answer is "no". If this would be true it would follow easily that independent systems of three unknown equations are of cardinality at most 3 – a solution to our main question in the case $n = 3$.

However, as an indication of our very poor knowledge on word equations, it is not known whether the maximal size of independent systems of equations exist – even in case $n = 3$!

Nontrivial lower bounds for the size of maximal independent systems of equations with n unknowns were given in [10]. One of the examples was as follows:

Example 17. Let $\Xi = \{x_i, y_i, u_i, w_i, v_i \mid i = 1, \ldots, n\}$ and S the following set of equations

$$S : x_i u_j w_k v_j y_i = y_i u_j w_k v_j x_i \quad \text{for } i, j, k = 1, \ldots, n.$$

Hence S contains $5n$ unknowns and is of cardinality n^3. We claim that S is independent over free semigroup Σ^+. Let us fix i, j, k and denote the corresponding equation in S by $S(i, j, k)$. Next we consider the morphism h defined by

$$h(x_t) = \begin{cases} b^2ab & \text{if } t = i \\ a & \text{otherwise} \end{cases} \quad h(u_p) = \begin{cases} ba & \text{if } p = j \\ bab & \text{otherwise} \end{cases} \quad h(w_q) = \begin{cases} bab^2 & \text{if } q = k \\ b & \text{otherwise} \end{cases}$$

$$h(y_t) = \begin{cases} b & \text{if } t = i \\ a & \text{otherwise} \end{cases} \quad h(v_p) = \begin{cases} ba & \text{if } p = j \\ a & \text{otherwise} \end{cases}$$

Straightforward calculations show that h is not a solution of the equation $S(i, j, k)$, but is that of any other equation of S. This means that S is independent.

The conclusion of the above example is:

> "The size of maximal independent system of equations with n unknowns is $\Omega(n^3)$ in the free semigroup Σ^+."

Similarly, it is shown in [10] that in free monoids we can do a bit better:

> "The size of maximal independent system of equations with n unknowns is $\Omega(n^4)$ in the free monoid Σ^*."

In [11] the hidden constants of the last result are slightly improved.

The above asymptotic lower bound does not give anything for small values of n, say $n = 3$. As we hinted in this case we know only that independent systems can contain three equations, and that all such systems are finite! A simple example of an independent three unknown system of equations is: $x^2 = y$, $y^2 = z$ and $z^2 = x$.

We conclude with a related problem. Let $\mathrm{Sol}(S)$ denote the set of all solutions of a system S of equations. For a sequence $s_1, \ldots, s_m$, we say that it is *descending chain* of equations if

$$\mathrm{Sol}\{s_i \mid i \leq j\} \not\subseteq \mathrm{Sol}\{s_i \mid i < j\} \quad \text{for all } j = 1, \ldots, m.$$

This means that whenever a new equation is introduced the set of solutions decreases. Therefore we are formalizing the question how many *constrains* for a set of n words we can introduce such that in each step the set of words satisfying these constrains becomes smaller.

An obvious connection between independent systems and descending chains of equations is that any nonrepetitive sequence of equations from an independent system is a descending chain. Apart from this no interesting connections seem to be known.

In fact, the results for descending chains are similar to independent systems. We state:

"Any descending chain of equations over Σ^* or Σ^+ is finite."

The known asymptotic lower bounds for the maximal lengths of such chains are those obtained for independent systems of equations. However, in the case of three unknowns we can do better. As shown in [11] descending chain of length seven can be constructed:

Example 18. The following sequence of seven equations provides a descending chain. Here on the right a solution which is not anymore a solution of the next one is shown.

$$\begin{aligned}
xyz &= zxy, & x &= a,\ y = b,\ z = abab\\
xyxzyz &= zxzyxy, & x &= a,\ y = b,\ z = ab\\
xz &= zx, & x &= a,\ y = b,\ z = 1\\
xy &= yx, & x &= a,\ y = a,\ z = a\\
x &= 1, & x &= 1,\ y = b,\ z = a\\
y &= 1, & x &= 1,\ y = 1,\ z = a\\
z &= 1, & x &= 1,\ y = 1,\ z = 1.
\end{aligned}$$

To conclude our knowledge on descending chains, we recall that all such chains are finite, but we do not know whether they can be longer than seven.

References

[1] M. H. Albert and J. Lawrence, *A proof of Ehrenfeucht's conjecture*, Theoret. Comput. Sci. **41** (1985), no. 1, 121–123.

[2] Karel Culik, II and Juhani Karhumäki, *Systems of equations over a free monoid and Ehrenfeucht's conjecture*, Discrete Math. **43** (1983), no. 2-3, 139–153.
[3] Elena Czeizler, Stepan Holub, Juhani Karhumäki, and Markku Laine, *Intricacies of simple word equations: an example*, Internat. J. Found. Comput. Sci. **18** (2007), no. 6, 1167–1175.
[4] Elena Czeizler and Juhani Karhumäki, *On non-periodic solutions of independent systems of word equations over three unknowns*, Internat. J. Found. Comput. Sci. **18** (2007), no. 4, 873–897.
[5] Elena Czeizler and Wojciech Plandowski, *On systems of word equations over three unknowns with at most six occurrences of one of the unknowns*, Theoret. Comput. Sci. **410** (2009), no. 30-32, 2889–2909.
[6] V. S. Guba, *Equivalence of infinite systems of equations in free groups and semigroups to finite subsystems*, Mat. Zametki **40** (1986), no. 3, 321–324.
[7] Tero Harju and Juhani Karhumäki, *Morphisms*, Handbook of Formal Languages (Grzegorz Rozenberg and Arto Salomaa, eds.), vol. 1, Springer-Verlag, 1997, pp. 439–510.
[8] Tero Harju, Juhani Karhumäki, and Wojciech Plandowski, *Independent systems of equations*, Algebraic Combinatorics on Words (M. Lothaire, ed.), Cambridge University Press, 2002, pp. 443–472.
[9] Tero Harju and Dirk Nowotka, *On the independence of equations in three variables*, Theoret. Comput. Sci. **307** (2003), no. 1, 139–172.
[10] Juhani Karhumäki and Wojciech Plandowski, *On the size of independent systems of equations in semigroups*, Theoret. Comput. Sci. **168** (1996), no. 1, 105–119.
[11] Juhani Karhumäki and Aleksi Saarela, *On maximal chains of systems of equations*, Manuscript.

Word periods under involution

Dirk Nowotka
(joint work with Bastian Bischoff)

This talk addresses questions about unbordered words, local and global periods generalized by considering involutions. Apart from general interest, this topic draws its motivation from combinatorial questions in computational biology and DNA computing. The complementation of a single stranded DNA, understood as a word over the alphabet $\{A, C, G, T\}$, constitutes what we call an antimorphic involution where A is mapped to T and C to G and the order of letters is reversed; see also the work of Lila Kari et.al. [2, 5, 6]. The presented material results from work in progress.

Let A denote an alphabet and A^* the free monoid of all finite words over A. Let $w \in A^*$ denote a word over A. Let $|w|$ denote the length of w. Let θ denote an involution on A^*, that is, $\theta(\theta(w)) = w$. Then θ is called morphic, if $\theta(uv) = \theta(u)\theta(v)$, and antimorphic, if $\theta(uv) = \theta(v)\theta(u)$. A word w is called primitive, if $w = u^i$ implies $i = 1$ for any $u \in A^*$. The primitive root of w is the shortest word u such that $w = u^i$ for some $i \geq 1$. A word w is called θ-primitive, if $w \in \{u, \theta(u)\}^i$ implies $i = 1$ for any $u \in A^*$. The θ-primitive root of w is the shortest word u such that $w \in \{u, \theta(u)\}^i$ for some $i \geq 1$. If w is a prefix of u^ω for some u then $|u|$ is called (global) period of w. Let $\Pi(w)$ denote the set of all periods of w, and let $\pi(w)$ denote the shortest period of w.

The notion of periodicity under involution can be defined in several ways. If w is a prefix of $\{u, \theta(u)\}^\omega$ for some u then $|u|$ is called θ-period of w. Let $\Pi_\theta(w)$ denote the set of all θ-periods of w, and let $\pi_\theta(w)$ denote the shortest θ-period of w. If w is a prefix of $(u\,\theta(u))^\omega$ for some u then $|u|$ is called alternating θ-period of w. Let $\Pi_\theta^{alt}(w)$ denote the set of all alternating θ-periods of w, and let $\pi_\theta^{alt}(w)$ denote the shortest alternating θ-period of w.

Another θ-generalization of the notion of periodicity is the following. We say that a natural p is called weak θ-period of w, if $w_{[i]} = w_{[i+p]}$ or $w_{[i]} = \theta(w_{[i+p]})$ for all $1 \leq i \leq |w| - p$ where $w_{[j]}$ denotes the jth letter of w. Let $\Pi_\theta^{weak}(w)$ denote the set of all weak θ-periods of w. However, the following result shows that weak θ-periods do not seem to imply anything different than ordinary periods.

Theorem 19. *Let θ be an involution on A^* and $w \in A^*$. Let ψ be a suitable substitution for θ. Then $\Pi_\theta^{weak}(w) = \Pi(\psi(w))$.*

A suitable substitution ψ for θ is a substitution with the following properties: Let A' be an alphabet such that (1) $a \in A'$ or $\theta(a) \in A'$ for all $a \in A$ and (2) $a \in A'$ and $a \neq \theta(a)$ implies $\theta(a) \notin A'$. Then ψ is a substitution where $\psi(a) = a$, if $a \in A'$, and $\psi(a) = \theta(a)$, if $a \notin A'$. We consider only (alternating) θ-periods in the following.

A very natural question regarding periods of words is the effects caused by overlaps. The classical result by Fine and Wilf [4] considers periods without involutions.

Theorem 20 ([4])**.** *Let $p, q \in \Pi(w)$ for some word w. If $|w| \geq p + q - \gcd\{p, q\}$ then $\gcd\{p, q\} \in \Pi(w)$.*

Czeizler et.al. consider in [2] the general antimorphic case and state:

Theorem 21 ([2])**.** *Let θ be an anti-morphic involution on A^* and $w \in A^*$ and $p, q \in \Pi_\theta(w)$ where $p > q$. Let u and v be prefixes of w of length p and q, respectively. If $|w| \geq 2p + q - \gcd\{p, q\}$ then u and v have the same θ-primitive root, that is they have a common period not longer than q.*

We add the following for the morphic case.

Theorem 22. *Let θ be a morphic involution on A^* and $w \in A^*$.*
If $|w| \geq p + q - \gcd\{p, q\}$ with $p, q \in \Pi_\theta(w)$ then $\gcd\{p, q\} \in \Pi_\theta(w)$.
If $|w| \geq p + q$ with $p, q \in \Pi_\theta^{alt}(w)$ then $\gcd\{p, q\} \in \Pi_\theta^{alt}(w)$.

All bounds given so far are tight. The only open case is the one for alternating θ-periods where θ is antimorphic.

Theorem 23. *Let θ be an antimorphic involution on A^* and $w \in A^*$.*
If $|w| \geq p + q$ with $p, q \in \Pi_\theta^{alt}(w)$ then $\gcd\{p, q\} \in \Pi_\theta^{alt}(w)$.

This bound however could not shown to be tight. We conjecture that the alternating antimorphic case has actually the bound $|w| \geq p + q - \gcd\{p, q\}$. This conjecture is still open.

The Critical Factorization theorem (CFT) is fundamental in the investigation of local periods. Let $u \sim_p v$ denote the fact that either u is a prefix of v or vice versa. Similarly, we write $u \sim_s v$, if u is a suffix of v or vice versa. Consider a factorization $w = uv$, then x is called a repetition word for this factorization of w, if $u \sim_s x$ and $v \sim_p x$. The length of x is called local period for the factorization uv. The smallest local period is denoted by $\pi(u, v)$. It is straightforward to see that $\pi(u, v) \leq \pi(w)$. A factorization is called critical if $\pi(u, v) = \pi(w)$. The critical factorization theorem was developed in several papers, see [7, 1, 3], and can be stated as follows.

Theorem 24 (CFT)**.** *Among any $\pi(w)$ many consecutive factorizations uv of a word w exists at least one that is critical, that is, where $\pi(u, v) = \pi(w)$.*

It is a natural generalization to consider local periods under an involution as we did for the global periods. Given a factorization uv of w, we call x a θ-repetition word, if $u \sim_s x$ and $v \sim_p \theta(x)$. The length of x is called local θ-period for the factorization uv. The smallest local θ-period is denoted by $\pi_\theta(u, v)$. However, this notion does not yield a structural property like the CFT, neither when $\pi_\theta(u, v)$ is related to the ordinary nor the θ-period of w, and neither for the morphic nor antimorphic case, and neither for the alternating nor non-alternating case. Let the following propositions exemplify this for the case of alternating θ-periods where θ is morphic.

Proposition 25. *Let $|A| \geq 3$ and θ be a morphic involution (not the identity) on A^*. Then there exists for all $p \geq 1$ a word w such that $p = \pi_\theta^{alt}(w)$ and $p = \pi_\theta(u, v)$ for all factorizations uv of w.*

Proposition 26. *Let θ be a morphic involution (not the identity) on A^*. Then there exists for all $p \neq 3$ a word w such that $p = \pi_\theta^{alt}(w)$ and $\pi_\theta(u, v) \in \{1, 2\}$ for all factorizations uv of w.*

The inability to directly transfer a strong result on local periods to the case of local θ-periods suggests a deeper investigation of the local θ-periodic structure of words. Unbordered factors are an obvious subject of interest here. A word is bordered if there exists a proper prefix that is also a suffix. Otherwise, a word is called unbordered. Let $\tau(w)$ denote the maximal length of unbordered factors in w. The following are straightforward observations: $\tau(w) \leq \pi(w)$, the shortest border of a bordered word is unbordered, a shortest local repetition word is unbordered. Moreover, a short argument using Lyndon words establishes that, if $|w| \geq 2\pi(w)-1$ then $\tau(w) = \pi(w)$. A more involved proof shows that, if $|w| \geq 7/3\,\tau(w)$ then $\tau(w) = \pi(w)$ (see the abstract by Štěpán Holub on page 2213). How does that property of unbordered factors relate to the θ-bordered case? A word w is called θ-bordered, if w has a proper prefix u such that $\theta(u)$ is a suffix of w. Otherwise, w is called θ-unbordered. Consider the morphic case and alternating θ-periods. The following sequence of words has a ration of word length to the maximal length of unbordered factors that approaches 3 and in the limit yet $\tau_\theta(w_i) \neq \pi_\theta^{alt}(w)$ thereby

showing that the $7/3\,\tau_\theta(w)$ bound does not hold. Let

$$w_i = (ab)^i abb(ab)^i aab(ab)^i a$$

for any $i \geq 2$, and let $\theta(a) = b$ and $\theta(b) = a$. Then $|w_i| = 6(i+1)+1$ and $\tau_\theta(w_i) = 2i+4$ and $\pi_\theta^{alt}(w_i) = 4i+5$. We have $|w_i| \geq 7/3\,\tau_\theta(w_i)$ for all $i \geq 2$ and $\lim_{i\to\infty} |w_i|/\tau_\theta(w_i) = 3$. We conjecture that this example is the best possible.

Conjecture 27. Let θ be a morphic involution on A^* and $w \in A^*$.
If $|w| \geq 3\tau_\theta(w)$ then $\tau_\theta(w) = \pi_\theta^{alt}(w)$.

References

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

[2] E. Czeizler, L. Kari, and S. Seki, *On a special class of primitive words*, Theoret. Comput. Sci. **411** (2010), no. 3, 617–630.

[3] J.-P. Duval, *Périodes et répétitions des mots du monoïde libre*, Theoret. Comput. Sci. **9** (1979), no. 1, 17–26.

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

[5] L. Kari and K. Mahalingam, *Involutively bordered words*, Internat. J. Found. Comput. Sci. **18** (2007), no. 5, 1089–1106.

[6] L. Kari and K. Mahalingam, *Watson–Crick palindromes in DNA computing*, Natural Computing **9** (2010), 297–316.

[7] M.-P. Schützenberger, *A property of finitely generated submonoids of free monoids*, Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976) (Amsterdam), Colloq. Math. Soc. János Bolyai, vol. 20, North-Holland, 1979, pp. 545–576.

Repetition-free colorings of trees

Pascal Ochem

The notion of square-freeness of words naturally extends to colored graphs. A factor of a colored graph is given by the sequence of colors on a non-intersecting path. A coloring of a graph is non-repetitive if and only if none of these factors is a square. The non-repetitive chromatic number of a graph class is the smallest integer k such that every graph in the class has a non-repetitive coloring using at most k colors.

The famous result of Thue [7] that there exists an infinite square-free word over a 3-letter alphabet is equivalent to the statement that the non-repetitive chromatic number of paths is 3. A major open question in this area is whether the non-repetitive chromatic number of planar graphs is bounded [4]. It is known to be $O(\Delta^2)$ for graphs with maximum degree Δ [2] and at most 4^t for graphs with treewidth k [5]. So planar graphs form an interesting class to study in this respect, since it is a small and natural class such that both the maximum degree and the treewidth are unbounded. The non-repetitive chromatic number of planar graphs is at least 10, and this lower bound already holds for the weaker *star coloring* [1], such that the only forbidden squares are aa and $abab$ for any letters a and b.

In this talk, I focus on trees, which form an intermediate class between paths and planar graphs. First, a negative result about the list version of non-repetitive coloring. Given an integer ℓ and a list assignment on the vertices such that each list contains exactly ℓ colors, we wonder whether we can obtain a non-repetitive coloring of the graph such that the color of a vertex is chosen from its associated list.

Theorem 28 ([3])**.** *For every integer ℓ, there exists a tree T and a list assignment L with lists of size ℓ, such that every coloring of T obtained from L contains a square.*

Theorem 28 implies that some classical methods will fail to prove that the non-repetitive chromatic number of planar graphs is bounded: those methods that produce the same upper bound for both the "list" and the "non-list" version of the studied coloring.

In the case of words or paths, we do not only know that the non-repetitive chromatic number is 3. Now that Dejean's conjecture is proved, we know the repetition threshold of words over a k letter alphabet for every $k \geq 2$. For an integer k and graph class $\mathcal{C}$, we similarly define the repetition threshold $RT(k, \mathcal{C})$. The next result gives the repetition thresholds for the class of trees $\mathcal{T}$.

Theorem 29 ([6])**.**

(1) $RT(2, \mathcal{T}) = \frac{7}{2}$
(2) $RT(3, \mathcal{T}) = \frac{5}{2}$
(3) $RT(k, \mathcal{T}) = \frac{3}{2}$, *for* $k \geq 4$.

References

[1] M.O. Albertson, G.G. Chappell, H. A. Kierstead, A. Kndgen, and R. Ramamurthi. Coloring with no 2-Colored P_4's, *Electron. J. Comb.* **11** (2004), #R26

[2] N. Alon, J. Grytczuk, M. Hałuszczak, and O. Riordan. Non-repetitive colourings of graphs, *Random Struct. Alg.* 21 (2002), 336-346.

[3] F. Fiorenzi, P. Ochem, P. Ossona de Mendez, and X. Zhu. Thue choosability of trees *submitted*

[4] J. Grytczuk. Nonrepetitive colorings of graphs - a survey, *Int. J. Math. Math. Sci.* (2007), Art. ID 74639, 10 pp.

[5] A. Kündgen and M. J. Pelsmajer. Nonrepetitive colourings of graphs of bounded treewidth, *Discrete Math.* 308(19) (2008), Simonovits '06, pp. 4473–4478.

[6] P. Ochem and E. Vaslet. Repetition thresholds for subdivided graphs and trees, Journées montoises 2010.

[7] A. Thue, Über unendliche Zeichenreichen, Norske Vid. Selsk. Skr., I Mat. Nat. Kl., Christiania, 7 (1906), 1-22.

On the Minimal Uncompletable Word Problem

Elena V. Pribavkina

A finite set S of (finite) words over an alphabet Σ is said to be *complete* if $Fact(S^*)$, the set of factors of S^*, is equal to Σ^*, that is, if every word of Σ^* is a factor of, or can be completed by multiplication on the left and on the right as, a word of S^*. If S is not complete, $\Sigma^* \setminus Fact(S^*)$ is not empty and a word in this set of minimal length is called a *minimal uncompletable word* (with respect to the non-complete set S).

Here we state some open questions related to the notion of a non-complete set and give a brief overview of partial results obtained so far towards solving these questions. The first natural question related to complete sets is the following:

Question 30. *Is it decidable whether a given set S is complete?*

To answer this question we can associate with the set S a non-deterministic finite-state automaton $\widehat{\mathscr{F}}(S)$ recognizing S^* in such a way that testing the property of completeness of the set S is equivalent to testing the synchronizability property of the associated automaton. For more details on this question see the paper [2]. Once Question 30 is solved, another natural question is whether there is an efficient algorithm of testing $Fact(S^*) = \Sigma^*$. In other terms,

Question 31. *What is the computational complexity of testing $Fact(S^*) = \Sigma^*$?*

Some results related to Question 31 were obtained by Rampersad, Shallit and Xu in [3]. Given a language L they studied the computational complexity of problems $Pref(L) = \Sigma^*$, $Suff(L) = \Sigma^*$ and $Fact(L) = \Sigma^*$. They showed that if L is given by a DFA M, then testing $Pref(L(M)) = \Sigma^*$ can be performed in linear time, the decision problem $Suff(L(M)) = \Sigma^*$ is PSPACE-complete and the problem $Fact(L(M)) = \Sigma^*$ is solvable in polynomial time. In contrast, if the language L is given by an NFA, all three problems become PSPACE-complete. In [3] it is also proved, that in particular case $L = S^*$ one can test $Pref(S^*) = \Sigma^*$ and $Suff(S^*) = \Sigma^*$ in linear time, while the computational complexity of testing $Fact(S^*) = \Sigma^*$ is still unknown.

Another rather natural question that might give some hint on solving the Question 31 is about the possible length of words that cannot be completed in S. Formally we state it as follows:

Question 32. *Given a non-complete set S, what is the minimal length $uwl(S)$ of words in $\Sigma^* \setminus Fact(S^*)$?*

From the connection between the set S and synchronizability property of the associated automaton $\widehat{\mathscr{F}}(S)$ one can deduce an exponential upper bound on the value $uwl(S)$:

$$uwl(S) \leq 2^{\|S\|-m+1},$$

where m is the number of elements in S and $\|S\|$ is the *size* of S, i.e. the sum of lengths of all elements in S. However this bound is not likely to be precise.

The Question 32 was introduced by Restivo in 1981. In his paper [4] he conjectured that a non-complete set S always possesses an uncompletable word w of length at most $2k^2$, where k is the maximal length of words in S, and w is of the form $w = uv_1uv_2 \cdots uv_{k-1}u$, where $u \notin S$, $|u| = k$ and $|v_i| \leq k$ for all $i = 1, 2, \ldots, k-1$. This conjecture is appeared to be false by means of a counterexample found in [1]. Namely, let $k > 6$ and let

$$S_k = \Sigma^k \setminus \{a^{k-2}bb\} \cup \Sigma ba^{k-4}\Sigma \cup \Sigma ba \cup b^4 \cup J_k$$

where $J_k = \bigcup_{i=1}^{k-3}(ba^i\Sigma \cup a^ib)$. In [1] the authors computed for $7 \leq k \leq 12$ that the word

$$\begin{aligned} w \quad = \quad & (a^{k-2}bb)a^{k-1}(a^{k-2}bb)ba^{k-4}((a^{k-2}bb)ba(a^{k-2}bb)bba^{k-5})^{k-6} \\ & (a^{k-2}bb)ab(a^{k-2}bb)bba^{k-3}(a^{k-2}bb)ba^{k-3}(a^{k-2}bb) \end{aligned}$$

is a minimal uncompletable word for S_k, and its length is $|w| = 3k^2 - 9k + 1$. Nevertheless it is not proved that such a word is a minimal uncompletable (or even uncompletable) for each $k > 6$. Thus the possible directions of the future work towards solving the Question 32 are the following:

- show that the word w is a minimal uncompletable word for the set S_k for each $k > 6$;
- find another infinite series of non-complete sets having minimal uncompletable words of length greater that $2k^2$;
- perform a series of computational experiments to find non-complete sets among all such sets with a fixed parameter k having the maximal possible length of minimal uncompletable words;
- give some (polynomial, if possible) upper bound on the length of minimal uncompletable word in terms of k.

References

[1] G. Fici, E. Pribavkina, J. Sakarovitch. *On the Minimal Uncompletable Word Problem*, CoRR, http://arxiv.org/abs/1002.1928, 2010.

[2] E. V. Pribavkina. *Slowly synchronizing automata with zero and incomplete sets*, CoRR, http://arxiv.org/abs/0907.4576, 2009.

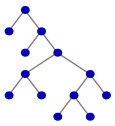
[3] N. Rampersad, J. Shallit, Z. Xu *The computational complexity of universality problems for prefixes, suffixes, factors, and subwords of regular languages*, CoRR, http://arxiv.org/abs/0907.0159, 2009.

[4] A. Restivo. *Some remarks on complete subsets of a free monoid*, Quaderni de "La ricerca scientifica", CNR Roma **109** (1981) 19–25.

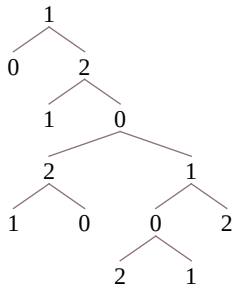
Ambiguity in a certain context-free grammar

Eric Rowland
(joint work with Bobbe Cooper, Doron Zeilberger)

Let G be the context-free grammar with start symbols $0, 1, 2$ and formation rules $0 \to 12$, $0 \to 21$, $1 \to 02$, $1 \to 20$, $2 \to 01$, $2 \to 10$. An n-leaf tree T *parses* a length-n word w on $\{0, 1, 2\}$ if T is a valid derivation tree for w under the grammar G; that is, there is a labeling of the vertices of T compatible with the formation rules such that the leaves of T, from left to right, are labeled with the letters of w. For example, the tree



parses the word 0110212:



The grammar G is ambiguous — there exist distinct trees that parse the same word; for example, the trees

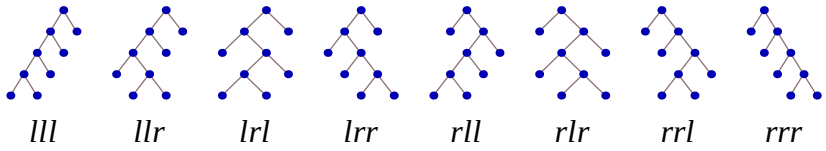
both parse 010. However, something much stronger can be said about this grammar.

Theorem 33. *Let $n \geq 1$, and let T_1 and T_2 be n-leaf binary trees. Then T_1 and T_2 parse a common word under G.*

Kauffman [4] proved this theorem by showing that it is equivalent to the four color theorem — the statement that every planar graph is four-colorable. The four color theorem was proved by Appel, Haken, and Koch [1, 2] and employed a large case analysis carried out by machine. The hope of the present authors [3] is that a direct proof will be shorter than the known proofs of the four color theorem, thereby providing a shorter proof of the four color theorem. In this direction, we enumerate the common parse words for some infinite families of tree pairs and discuss ways to reduce the problem of finding a parse word for a pair of trees to that for a smaller pair.

Let $\mathrm{ParseWords}(T_1, T_2)$ be the set of equivalence classes, under permutations of the alphabet $\{0, 1, 2\}$, of words parsed by both trees T_1 and T_2. We will abuse notation slightly by writing a representative of each equivalence class.

A *path tree* is a binary tree with at most two vertices in each level. The 5-leaf path trees are as follows.



The illustrated bijection between n-leaf path trees and length-$(n-2)$ words on $\{l, r\}$ can be used to define families of trees. For example, let LeftCombTree(n) be the n-leaf path tree corresponding to the word l^{n-2}, and let RightCombTree(n) be the n-leaf path tree corresponding to r^{n-2}. There is only one equivalence class of words parsed by both a left comb tree and a right comb tree.

Theorem 34. ParseWords(LeftCombTree(n), RightCombTree(n)) =

$$\begin{cases} \{01^{n-2}2\} & \text{if } n \geq 2 \text{ is even} \\ \{01^{n-2}0\} & \text{if } n \geq 3 \text{ is odd.} \end{cases}$$

Similar results for other pairs of one-parameter families of trees can be established.

A simple two-parameter family is the $(m+n)$-leaf path tree corresponding to $l^m r^{n-2}$, which we call LeftTurnTree(m, n). Let RightTurnTree(m, n) be the left–right reflection of LeftTurnTree(m, n) — the tree corresponding to $r^m l^{n-2}$. The next three theorems collectively determine the number of parse words of LeftTurnTree(m, n) and RightTurnTree$(k, m+n-k)$.

Theorem 35. *For $m \geq 1$, $k \geq 1$, and $\max(2, k-m+2) \leq n \leq k$,*

$$|\text{ParseWords}(\text{LeftTurnTree}(m, n), \text{RightTurnTree}(k, m+n-k))| = 1.$$

Let

$$a(m, k) = |\text{ParseWords}(\text{LeftTurnTree}(m, k+1), \text{RightTurnTree}(k, m+1))|.$$

By considering the left–right reflections of these two trees, we see that $a(m, k) = a(k, m)$.

Theorem 36. *For $m \geq 1$ and $k \geq 1$,*

$$a(m+3, k) - 2a(m+2, k) - a(m+1, k) + 2a(m, k) = 0.$$

This recurrence can be written

$$(M-2)(M-1)(M+1)\, a(m, k) = 0,$$

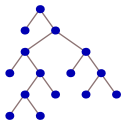
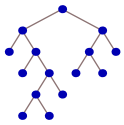
where M is the forward shift operator in m, so for fixed k the solution $a(m, k)$ is a linear combination of $2^m, 1, (-1)^m$. Unfortunately, we do not know a simple combinatorial proof of the recurrence.

Theorem 37. *For $m \geq 1$, $k \geq 1$, and $n \geq k+2$,*

$$|\text{ParseWords}(\text{LeftTurnTree}(m, n), \text{RightTurnTree}(k, m+n-k))| = 2a(m, k).$$

In addition to enumerating parse words, we are interested in pursuing existence results. There are several ways to reduce the problem of finding a parse word for a pair of trees to finding parse words for smaller pairs. Here we describe two — one easily provable and one conjectural.

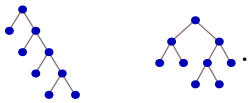
If T_1 and T_2 are n-leaf trees that have a common branch system in the same position, then we can decompose the pair into two smaller pairs. For example, the 8-leaf trees

$T_1 =$  $T_2 =$ 

share the branch system

$S =$ 

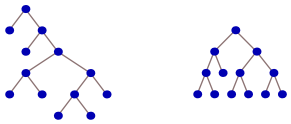
in the second through fifth leaves, which we may remove to obtain the 5-leaf trees

 .

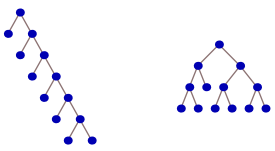
Given a common parse word $w_1w_2w_3w_4w_5$ of this pair of 5-leaf trees, we can find a common parse word of the original pair of 8-leaf trees by taking any valid labeling of S and permuting the alphabet so that the root receives the label w_2.

More generally, to decompose a pair of trees we only require a vertex in T_1 with dangling subtree S_1 and a vertex in T_2 with dangling subtree S_2 such that same set of leaves appears in S_1 and S_2. A parse word for T_1 and T_2 can then be constructed from a parse word for S_1 and S_2 and a parse word for the chopped trees T_1' and T_2'.

A pair of n-leaf trees T_1 and T_2 is *mutually crooked* if there is no pair of consecutive leaves that have an uncle–nephew relationship in both trees. For example,



are mutually crooked, while the following are not.



We conjecture that a pair of trees that is not mutually crooked has a common parse word in which the uncle and nephew leaves receive the same label. There is reason to believe that finding this parse word is reducible to finding a parse word for the pair obtained by deleting the ∴ structure holding one of the two leaves.

References

[1] Kenneth Appel and Wolfgang Haken, Every planar map is four colorable I: discharging, *Illinois Journal of Mathematics* **21** (1977) 429–490.

[2] Kenneth Appel, Wolfgang Haken, and John Koch, Every planar map is four colorable II: reducibility, *Illinois Journal of Mathematics* **21** (1977) 491–567.

[3] Bobbe Cooper, Eric Rowland, and Doron Zeilberger, Toward a language theoretic proof of the four color theorem, available at `http://arxiv.org/abs/1006.1324`.

[4] Louis Kauffman, Map coloring and the vector cross product, *Journal of Combinatorial Theory, Series B* **48** (1990) 145–154.

The enumeration of squares and runs in the Fibonacci words revisited

Kalle Saari

The problem of enumerating repetitions of different types in finite words is a recurrent research topic in combinatorics on words. Here are three seminal results in this area: Crochemore [1] showed that the number of primitively rooted squares, counting multiplicities, in a word of length n is in $O(n \log n)$; Fraenkel and Simpson [2] showed that a word of length n contains no more than $2n$ distinct squares; Kolpakov and Kucherov [4] showed that the number of runs, counting multiplicities, in a word is linearly bounded. For more information about repetitions in words and resent results, consult the surveys [5, Ch. 8] and [6, Ch. 8] and the website [7].

The exact number of repetitions in interesting classes of words is naturally also of interest. Let us recall that Fibonacci words f_n for $n \geq 0$ are defined by

$$f_0 = 0, \qquad f_1 = 01, \qquad f_n = f_{n-1}f_{n-2} \quad (n \geq 2);$$

their connection to Fibonacci numbers $F_0 = 0, F_1 = 1, \ldots$ is that the length of f_n equals F_{n+2}.

The number of squares and runs in Fibonacci words were worked out by Fraenkel and Simpson [3] and Kolpakov and Kucherov [4], respectively. The number of distinct squares in f_n is given by the formula $2(F_n - 1)$ for all $n \geq 4$; the number of runs in f_n, counting multiplicities, is precisely $2F_n - 3$, and this holds for all $n \geq 3$.

The purpose of this abstract is to outline a novel way for enumerating squares and runs, both distinct and with multiplicities, in finite Fibonacci words. Our technique uses properties of central and singular factors of the infinite Fibonacci word $\lim_{n\to\infty} f_n$. Central factors, denoted by p_n, are obtained from f_n by erasing the last two letters; singular factors, denoted by c_n, are obtained from f_n by removing the last letter and appending its complement to the beginning. Thus, for all $n \geq 1$, we have $f_n = p_n ab$ and $c_n = a p_n a$, where $ab \in \{01, 10\}$. It turns out that runs occurring within a central factor p_n coincide with occurrences of factors of the form $c_{k+1}c_k c_{k+1}$ and $c_k c_{k+1} c_k$. Also, the runs that occur as prefixes or suffixes of p_n are the central factors $p_4, p_5, \ldots, p_n$. Furthermore, each occurrence of c_{k+1} in p_n extends to an occurrence of $c_k c_{k+1} c_k$. Finally, it can be shown that

the number of occurrences of a singular factor c_k in a central factor p_n equals $F_{n-k} - 1$.

Using these observations, counting the number of squares and runs in a central factor p_n is straightforward, and extending these formulas for f_n is easy. In particular, it can be verified that the number of squares with multiplicities in f_n is given by

$$R(n) := \frac{2}{5}(n-5)(F_{n+2} + F_n) + \frac{4}{5}F_n + n + 2 \qquad (n \geq 3).$$

A formula for $R(n)$ was first derived by Fraenkel and Simpson in [3], but their formula has a misprint: it differs from ours by $3F_n$.

References

[1] Maxime Crochemore, *An optimal algorithm for computing the repetitions in a word*, Inf. Process. Lett. **12** (1981), no. 5, 244–250.

[2] Aviezri S. Fraenkel and Jamie Simpson, *How many squares can a string contain?*, J. Comb. Theory, Ser. A **82** (1998), no. 1, 112–120.

[3] Aviezri S. Fraenkel and Jamie Simpson, *The exact number of squares in Fibonacci words*, Theor. Comput. Sci. **218** (1999), no. 1, 95–106.

[4] Roman M. Kolpakov and Gregory Kucherov, *On maximal repetitions in words*, FCT (Gabriel Ciobanu and Gheorghe Paun, eds.), Lecture Notes in Computer Science, vol. 1684, Springer, 1999, pp. 374–385.

[5] M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, Cambridge, 2002.

[6] M. Lothaire, *Applied combinatorics on words*, Encyclopedia of Mathematics and its Applications, vol. 105, Cambridge University Press, Cambridge, 2005.

[7] L. Tinta M. Crochemore, L. Ilie, *The "runs" conjecture*, `http://www.csd.uwo.ca/faculty/ilie/runs.html`, September 2010.

On Patterns and Pattern Avoidance

Jeffrey Shallit

This talk centered around four themes, all concerned with patterns and pattern avoidance: (1) the complexity of matching words specified by abstract machines to patterns; (2) some aspects of the Thue-Morse word; (3) avoiding powers over $\mathbb{N}$, the natural numbers; and (4) the Pirillo-Varicchio-Halbeisen-Hungerbühler problem.

1. Complexity of pattern matching

Let Σ be an alphabet, i.e., a nonempty, finite set of symbols. By Σ^* we denote the set of all finite words over Σ, and by ϵ, the empty word. If $w = xyz$, then y is said to be a *factor* of w.

A *pattern* is a non-empty word p over a *pattern alphabet* Δ. The letters of Δ are called *variables*. A *morphism* is a map $h : \Sigma^* \to \Delta^*$ such that $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma$; a morphism is *non-erasing* if $h(a) \neq \epsilon$ for all $a \in \Sigma$. A word $w \in \Sigma^*$ *matches* a pattern p if there exists a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p) = w$. For example, the word $w =$ `oberwolfach` matches the pattern $xyxz$.

Some patterns play special roles in combinatorics on words. For example, words matching the pattern xx are called *squares*. An example in German is `nennen`, and in French is `chercher`. Words matching xxx are called *cubes*; an example is the English sort-of-word `shshsh`.

An *overlap* is a word of the form $axaxa$ where a is a single symbol and x is a (possibly empty) word. No single pattern captures the notion of overlap, but it is easy to see that a word w has a factor that is an overlap iff some factor of w matches either of the two patterns xxx and $xyxyx$.

Much study has been devoted to *fractional powers*. We say a word x is an exact $\frac{p}{q}$-power, for integers $p > q \geq 1$, if x can be written in the form $y^e y'$ for some $e \geq 1$, and y' is a prefix of y, and $|x| = p$, $|y| = q$. For example, the German word `schematische` is a $\frac{12}{8}$-power. If x has a factor that is an exact $\frac{p}{q} - power$, for some $\frac{p}{q} \geq \alpha$, we say that x has an occurrence of an α-power.

This suggests the following:

Problem 38. *Let $\alpha > 1$ be a real number that is not an integer. Is there any set of patterns P, finite or infinite, such that x has an occurrence of an α-power iff some factor of x matches some pattern $p \in P$?*

Remark. After my talk both James Currie and Julien Cassaigne suggested ways to show this is not possible, at least for some restricted ranges of α. But some further work is still needed to cover all $\alpha > 1$.

The complexity of pattern matching is an old problem, going back to fundamental results of Ehrenfeucht and Rozenberg [9] and (independently) Angluin [4]. They showed that the problem of determining if an arbitrary word w matches an arbitrary pattern p is NP-complete.

Recently I (and co-authors) have explored the computational complexity of other kinds of pattern matching, where we replace a word by a set of words specified by some abstract machine M (e.g., DFA, NFA, or PDA). For example, Anderson et al. [3] showed that the problem of deciding if some word in $L(M)$ matches a given pattern p is PSPACE-complete, if M is an NFA. Furthermore, the problem remains PSPACE-complete even the NFA is deterministic and the pattern p is restricted to belong the class of patterns $\{xx, xxx, xxxx, \ldots\}$. The idea behind the proof is very simple. Given NFA's $M_1, M_2, \ldots M_n$, we construct an NFA accepting $L = L(M_1)\# \cdots \# L(M_n)\#$. Then a word of L matches x^n if and only if $x \in L(M_1) \cap \cdots \cap L(M_n)$. This shows PSPACE-hardness. It is slightly harder to verify that the problem is in PSPACE. Here the idea is to show that if a match occurs, it cannot be too larger. One way to show membership in PSPACE is to guess, for each variable x_i in the pattern, a boolean matrix B_i that specifies the transition table in the automaton induced by x_i. Then we simply need to verify that indeed each B_i has a corresponding word, and multiply the matrices together.

On the other hand, if M is a PDA (or CFG), then the problem becomes undecidable, even if p is the single fixed pattern xx. The idea here is to reduce from the Post correspondence problem.

It seems reasonable to guess that at least some of this problem's complexity comes from the fact that $L(M)$ can be infinite. So we also explored finite analogues of the problem. Here we are given a DFA or NFA M accepting a *finite* language, and we ask if some word of $L(M)$ matches a given pattern p. In this case we recently showed [13] that the problem is NP-complete.

In the case that M is a PDA (or CFG), the finite problem is PSPACE-complete, even in the case where $p = ww$. Again, the hard part is showing membership in PSPACE. The idea is to guess the configuration C after reading the first copy of w. This configuration consists of a state and the contents of the stack, so we need to make sure that the stack contents cannot be arbitrarily large.

2. Around the Thue-Morse sequence

The Thue-Morse sequence $\mathbf{t} = t_0 t_1 t_2 \cdots = \mathtt{0110100110010110}\cdots$ is a familiar object in combinatorics on words. For one thing, it avoids overlaps. Recently I showed (with co-authors) that $\mathbf{t}$ is *fragile* with respect to this property. More precisely, $\mathbf{t}$ is overlap-free, but if the bits in any finite nonempty set of positions are flipped (0 becoming 1 and vice versa), then the resulting sequence $\mathbf{t}'$ is no longer overlap-free [7].

This suggests considering the same problem for squares. For example,

Problem 39. *Are there infinite words over* $\{0, 1, 2\}$ *that are fragile with respect to the property of squarefreeness?*

My guess is that you can get such a word by the usual trick of counting the number of 1's between consecutive 0's in the Thue-Morse word, and then dropping the first symbol, but I haven't proved it yet.

One can also consider words at the opposite spectrum. Call an infinite word $\mathbf{w}$ *robust* with respect to property P if

(1) $\mathbf{w}$ has property P; and
(2) there are infinitely many disjoint finite sets S of positions, such that changing every symbol at the positions in S to some other symbol results in a word that still has property P.

This suggests

Problem 40. *Are there infinite words over* $\{0, 1, 2\}$ *that are robust with respect to the property of squarefreeness?*

Remarks. After my talk Pascal Ochem showed how to find such words. For example, Brinkhuis [6] constructs a substitution (in the genuine sense of theoretical computer science!) on three letters as follows:

$$\begin{aligned} h(0) &= \{0120210201021201020120210, 0120212010210120102120210\} \\ h(1) &= \{1201021012102012101201021, 1201020121021201210201021\} \\ h(2) &= \{2012102120210120212012102, 2012101202102012021012102\} \end{aligned}$$

Then $h^\omega(0)$ is an infinite set of infinite squarefree words, and furthermore by replacing any aligned block of size 25 with its corresponding block retains the property of squarefreeness.

James Currie also found a solution to Problem 40.

T. Cusick recently asked me a very interesting problem related to Thue-Morse. Let $s_2(n)$ denote the number of 1's in the binary expansion of n (not reduced modulo 2). For integers $k, t \geq 1$ define

$$\gamma_k(t) = \frac{1}{2^k} |\{n \ : \ 0 \leq n < 2^k \text{ and } s_2(n+t) \geq s_2(n)\}|.$$

It is not hard to see that as $k \to \infty$, the values of $\gamma_k(t)$ stabilize to some fixed value $\gamma_\infty(t)$. (In fact, it suffices to pick $k \geq 2\lceil \log_2 t \rceil$.) So $\gamma_\infty(t)$ measures the fraction of binary numbers n for which adding t gives more 1's in $n+t$ than in n. It is now not so difficult to prove that $\gamma_\infty(1) = 3/4$, and indeed $\gamma_\infty(2t) = \gamma_\infty(t)$. Cusick asked,

Problem 41. *Is it true that* $\gamma_\infty(t) > 1/2$ *for all* $t \geq 1$*? What is* $\liminf_{t\to\infty} \gamma_\infty(t)$*?*

We might also ask:

Problem 42. *Find a simple way to compute* $\gamma_\infty(t)$*.*

Finally, I close this section with another problem related to Thue-Morse — more precisely, the overlap-free property. Let $\mathbf{a} = (a_i)_{i\geq 0}$ be an infinite sequence of integers. The *Hankel determinant of order* n *beginning at position* k is the determinant

$$\begin{vmatrix} a_k & a_{k+1} & \cdots & a_{k+n-1} \\ a_{k+1} & a_{k+2} & \cdots & a_{k+n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k+n-1} & a_{k+n} & \cdots & a_{k+2n-2} \end{vmatrix}.$$

Evidently if $\mathbf{a}$ has an overlap $cxcxc$, then there is a 0 Hankel determinant — namely, the one corresponding to a matrix whose first and last rows are cxc. So we can ask,

Problem 43. *Is there an infinite sequence* $\mathbf{a}$ *over a finite subset of* $\mathbb{Z}$ *with the property that the Hankel determinants of all orders, beginning at all positions, are nonzero?*

I've done some computations on this problem. A simple backtracking algorithm shows that there is no such sequence over a subset of size 2, so we need at least three letters. I have constructed what appears to be the lexicographically least sequence with hundreds of letters over $\{1, 2, 3\}$, namely:

$$1121122121121231121122121122322112231121122121123112112212112232211 2 \cdots$$

Interestingly, no backtracking was required to generate this!

I have also constructed a morphism that appears to generate such a sequence, namely:

$$\begin{aligned} 1 &\to 12 \\ 2 &\to 23 \\ 3 &\to 14 \\ 4 &\to 32 \end{aligned}$$

I have tested this for hundreds of terms, and all Hankel determinants were nonzero.

Problem 44. *Can you show that this morphism generates a sequence with all Hankel determinants nonzero?*

A natural approach to this kind of problem would be to consider the Hankel determinants modulo m for some integer $m \geq 2$. However, this approach does not seem to work. In fact, I conjecture that the following is true:

Conjecture 45. For all integer moduli $m \geq 2$, there is no infinite sequence over a finite subset of $\mathbb{Z}$ such that all Hankel determinants are nonzero (modulo m).

3. Avoiding patterns over the natural numbers

Suppose we try to avoid squares, but don't allow backtracking. Having generated w, at each new position we write down the least natural number i such that wi is squarefree. It seems to be part of the folklore, and not too hard to prove, that the resulting word

$$\mathbf{w}_2 = (c_i)_{i\geq 1} = 01020103010201014\cdots$$

is the so-called "ruler sequence", defined by $c_i = \nu_2(i)$, the exponent of the largest power of 2 dividing i. This is the lexicographically least sequence over $\mathbb{N}$ avoiding squares (and also, by the way, the lexicographically least sequence avoiding abelian squares). It can be generated by the morphism that maps i to $0, (i+1)$.

Recently Guay-Paquet and I [10] considered the same question for overlap-free sequences over $\mathbb{N}$. Here the lexicographically least sequence

$$\mathbf{w}_{2^+} = 0010011001002\cdots$$

has a much more complicated structure. For example, we find that

0 occurs for the first time at position 1
1 occurs for the first time at position 3
2 occurs for the first time at position 13
3 occurs for the first time at position 79
4 occurs for the first time at position 633

and the sequence $(d_i)_{i\geq 1} = 1, 3, 13, 79, 633, \ldots$ satisfies the recurrence $d_{i+1} = 2id_i + 1$. These numbers are also the coefficients of the exponential generating function $e^x/(1-2x)$. Furthermore, they are given by the formula $\lfloor 2^n \cdot n! \cdot (\sqrt{e} - 1) \rfloor$. The word $\mathbf{w}_{2^+}$ is a natural example of a word with transcendental letter frequencies.

The word $\mathbf{w}_{2^+}$ can be written as $\varphi^\omega(0)$ where

$$\begin{aligned}\varphi(0) &= 001\\ \varphi(1) &= 1001002\\ \varphi(2) &= 2001001100100200100110010 03\\ &\vdots\end{aligned}$$

where $|\varphi(i)| = 2d_{i+1} + 1$. Here

$$\varphi(h) = (S(\varphi^h(00)) \cdot (h+1)$$

where $S(xc)$ is the shift operator defined by $S(xc) = cx$ for $x \in \Sigma^*$, $c \in \Sigma$.

We proved many other results about the word $\mathbf{w}_{2^+}$ and the morphism φ. Similar results can be proved for words avoiding k'th powers and k^+-powers, for any integer $k \geq 2$.

However, similar results for fractional powers have so far eluded us. For example, we computed the first $1,000,000$ terms of the lexicographically least sequence avoiding $\frac{5}{2}$-powers and did not find any term greater than 3. Nor were we able to characterize this sequence.

This leads to the following problem:

Problem 46. *Characterize the lexicographically least infinite α-power-free sequence over $\mathbb{N}$ for some fractional α.*

Remarks. After my talk Eric Rowland observed that the lexicographically least $\frac{3}{2}$-power-free sequence was "pseudoperiodic" with period 10, and we were able to show [14] that this sequence is in fact 6-regular in the sense of Allouche and Shallit [1, 2]. Rowland observed similar results for $\frac{4}{3}$ but this has yet to be proved.

4. The Pirillo-Varricchio-Halbeisen-Hungerbühler problem

Pirillo & Varricchio [12], and, independently, Halbeisen & Hungerbühler [11] posed the following problem:

Problem 47. *Is there an infinite seqence over a finite subset of $\mathbb{Z}$ with the property that it has no factors of the form xx' with $|x| = |x'|$ and $\sum x = \sum x'$?*

We might call such a factor xx' a PVHH-square. Recently there has been some attention paid to this problem [8], but it is still open.

Recently we found some results related to this problem [5]. For example, although with Van der Waerden's theorem one can show that for any m it is impossible to avoid xx' with $|x| = |x'|$ and $\sum x = \sum x' \pmod m$, it becomes possible if we allow the congruence to hold only if both sides are $0 \pmod m$.

We also considered the analogous problem for cubes instead of squares. Thomas Stoll and I performed some extensive calculations that suggest that a word avoiding PVHH cubes exists. We conjecture

Conjecture 48. The fixed point generated by the morphism

$$\begin{aligned} 0 &\to 03 \\ 1 &\to 43 \\ 3 &\to 1 \\ 4 &\to 01 \end{aligned}$$

avoids $xx'x''$ with $|x| = |x'| = |x''|$ and $\sum x = \sum x' = \sum x''$.

However, we have not yet been able to prove this.

Remarks. During the workshop Julien Cassaigne suggested another strategy which may suffice to prove the conjecture.

References

[1] J.-P. Allouche and J. Shallit. The ring of k-regular sequences. *Theoret. Comput. Sci.* **98** (1992), 163–197.

[2] J.-P. Allouche and J. Shallit. The ring of k-regular sequences, II. *Theoret. Comput. Sci.* **307** (2003), 3–29.

[3] T. Anderson, J. Loftus, N. Rampersad, N. Santean, and J. Shallit, Detecting palindromes, patterns and borders in regular languages, *Inform. and Comput.* **207** (2009), 1096–1118.

[4] D. Angluin, Finding patterns common to a set of strings, *J. Comput. System Sci.* **21** (1980) 46–62.

[5] Y.-H. Au, A. Robertson, and J. Shallit, Van der Waerden's theorem and avoidability in words, preprint, `http://arxiv.org/abs/0812.2466` .

[6] J. Brinkhuis, Non-repetitive sequences on three symbols, *Quart. J. Math. Oxford* **34** (1983), 145–149.

[7] S. Brown, N. Rampersad, J. Shallit, and T. Vasiga, Squares and overlaps in the Thue-Morse sequence and some variants, *RAIRO-Info. Theor. Appl.* **40** (2006), 473–484.

[8] J. Cassaigne, G. Richomme, K. Saari, and L. Q. Zamboni, Avoiding Abelian powers in binary words with bounded Abelian complexity, preprint, `http://arxiv.org/abs/1005.2514` .

[9] A. Ehrenfeucht and G. Rozenberg, Finding a homomorphism between two words is NP-complete, *Inform. Process. Lett.* **9** (1979) 86–88.

[10] M. Guay-Paquet and J. Shallit, Avoiding squares and overlaps over the natural numbers, *Discrete Math.* **309** (2009), 6245–6254.

[11] L. Halbeisen and N. Hungerbühler, An application of Van der Waerden's theorem in additive number theory, *INTEGERS: Elect. Journ. Comb. Number Theory* **0** (2000), #A7 (electronic), `http://www.integers-ejcnt.org/vol0.html` .

[12] G. Pirillo and S. Varricchio, On uniformly repetitive semigroups, *Semigroup Forum* **49** (1994), 125–129.

[13] N. Rampersad and J. Shallit, Detecting patterns in finite regular and context-free languages, *Inf. Process. Lett.* **110** (2010), 108–112.

[14] E. Rowland J. Shallit, Avoiding $\frac{3}{2}$-powers over the natural numbers, preprint, September 2 2010.

On the sum of digits of n and n^h

Thomas Stoll

1. Introduction

Let $q \geq 2$ and denote by $s_q(n)$ the sum of digits in the q-ary representation of an integer n. In recent years, much effort has been made to get a better understanding of the distribution properties of s_q regarding certain subsequences of the positive integers. We mention the classical work by Gelfond [5], and the recents papers by C. Mauduit and J. Rivat on the distribution in arithmetic progressions of s_q of primes [10] and of squares [11]. These questions are intimately connected to find the frequency of letters -1 and $+1$ in the classical Thue-Morse sequence

$$(t_n)_{n\in\mathbb{N}} = ((-1)^{s_2(n)})_{n\in\mathbb{N}} = +1, -1, -1, +1, -1, +1, +1, -1, \ldots$$

with respect to special subsequences of indices (e.g., $n \equiv a \bmod k$, $n = p$ prime, or $n = m^2$). In the case of indices of the form $P(n)$, where $P(n)$ denotes a fixed integer-valued polynomial of degree $h \geq 3$, very little is known. For the current state of knowledge, we refer to the work of C. Dartyge and G. Tenenbaum [2], who provided some lower density estimates for the evaluation of $s_q(P(n))$ in arithmetic progressions.

2. Oberwolfach talk

A related problem, however, of a more elementary nature, is to study extremal properties of $s_q(P(n))$. In the binary case when $q = 2$, B. Lindström [9] showed that

$$\limsup_{n\to\infty} \frac{s_2(P(n))}{\log_2 n} = h. \tag{1}$$

The special case $P(n) = n^2$ of (1) has been reproved by M. Drmota and J. Rivat [4] with constructions due to J. Cassaigne and G. Baron. Moreover, it is well-known [3, 13] that the average order of magnitude of $s_q(n)$ and $s_q(n^h)$ is

$$\sum_{n<N} s_q(n) \sim \frac{1}{h} \sum_{n<N} s_q(n^h) \sim \frac{q-1}{2\log q}\, N \log N, \tag{2}$$

so that the average value of $s_q(n^h)$ is h times larger than the average value of $s_q(n)$. It is an interesting question how often the ratio $s_q(n^h)/s_q(n)$ can be *very large* or *very small* and to quantify these notions. By naive methods, it can be quite hard to find even a single value n such that $s_2(n^h) < s_2(n)$ for some h. For instance, an extremely brute force calculation shows that the minimal n such that $s_2(n^3) < s_2(n)$ is $n = 407182835067 \approx 2^{39}$ where one finds $s_2(n^3) = 28$ and $s_2(n) = 29$.

In my Oberwolfach talk, I reported on recent work by K. G. Hare, S. Laishram and myself [7, 8] that settles an old conjecture of Stolarsky [14] from 1978. We find (up to a multiplicative absolute constant) the exact extremal orders of magnitude of the ratio $s_q(n^h)/s_q(n)$.

Theorem ([7]). *There exist c_1 and c_2, depending at most on q and h, such that for all $n \geq 2$,*

$$\frac{c_2}{\log n} \leq \frac{s_q(n^h)}{s_q(n)} \leq c_1 (\log n)^{1-1/h}.$$

This is best possible in that there exist c_1' and c_2', depending at most on q and h, such that

$$\frac{s_q(n^h)}{s_q(n)} > c_1' (\log n)^{1-1/h},$$

respectively,

$$\frac{s_q(n^h)}{s_q(n)} < \frac{c_2'}{\log n}$$

infinitely often.

The constants c_1, c_2, c_1' and c_2' are explicitly computable. The proofs involve the Bose-Chowla theorem on sets with the distinct sum property [1] and a combinatorial argument using a certain polynomial of degree four. For any $\varepsilon > 0$ we get a bound on the minimal n such that the ratio $s_q(n^h)/s_q(n) < \varepsilon$. For the example given above, the approach allows to show that

$$\min\{n : s_2(n^3) < s_2(n)\} < 2^{178}.$$

In the talk I also outlined the answer to the analogous question in the case of the Zeckendorf representation of integers [15].

3. Oberwolfach research

During the stay in Oberwolfach I worked with Jeffrey Shallit on the following related problem for the Thue-Morse word: Let $k \geq 1$, and denote $\mathcal{N}_k = \{n : t_{kn} = 1\}$ and $f(k) = \min\{n : n \in \mathcal{N}_k\}$. In other words, this is the first multiple of k such that we see a "+1" in the subword $(t_{kn})_{n\geq 1}$ of the Thue-Morse word. The first few values of $(f(k))_{k\geq 1}$ are given by

$$1,\ 1,\ 7,\ 1,\ 5,\ 7,\ 1,\ 1,\ 9,\ 5,\ 1,\ 7,\ 1,\ 1,\ 19,\ 1,\ 17,\ 9,\ 1,\ 5\ldots$$

From a well-known theorem of Gelfond [5] we get that $f(k) < \infty$ for all k, but can we give a sharp upper bound for $f(k)$ in terms of k? J.-P. Allouche and J. Shallit had already discussed the problem for some time, but the conjecture $f(k) \leq k + 4$ was still open. In Oberwolfach we tried to make some progress but we were finally left with some cases for k where we could not find such an n. After the Oberwolfach week we made some fresh effort with Johannes Morgenbesser (Marseille) to handle these tedious cases, too. So, we finally came up with a complete solution to the problem (joint work in progress).

Theorem 49. *For all $k \geq 1$ we have $f(k) \leq k + 4$. Moreover, we have*

1. *$f(k) = k + 4$ if and only if $k = 2^{2r} - 1$ for some $r \geq 1$,*
2. *There are no k's with $f(k) = k + 3$ or $f(k) = k + 2$.*
3. *$f(k) = k + 1$ if and only if $k = 6$.*
4. *$f(k) = k$ if and only if $k = 1$ or $k = 2^r + 1$ for some $r \geq 2$.*

In fact, by construction, for all k there is an $n \in \mathcal{N}_k$ with $n \leq k+4$ and $s_2(n) \leq 3$. We currently try to generalize our approach to a more general setting for $s_q(n)$.

4. Open questions

We end this report with some challenging open questions in this field.

- Determine the frequency of $+1$ and -1 in the Thue-Morse word for the subsequence of indices $P(n) = n^3$. More generally, find the distribution in arithmetic progressions of $s_q(n^h)$, $n \geq 1$, for fixed $h \geq 3$ (see [5]).
- Find $\frac{1}{N}\sum_{n<N} s_2(n^h)/s_2(n)$ (see [14]).
- Find $\#\{n < N,\ n \text{ odd}: \quad s_2(n^2) = s_2(n)\}$ (see [12]).
- Fix $k \in \{9, 10, 11, 14, 15\}$. Decide whether the set

$$\{n < N,\ n \text{ odd}: \quad s_2(n^2) = s_2(n) = k\}$$

  is finite or infinite (see [8]).

References

[1] R. C. Bose, S. Chowla, Theorems in the additive theory of numbers, *Comm. Math. Helv.* **37** (1962/63), 141–147.

[2] C. Dartyge, G. Tenenbaum, Congruences de sommes de chiffres de valeurs polynomiales, *Bull. London Math. Soc.* **38** (2006), no. 1, 61–69.

[3] H. Delange, Sur la fonction sommatoire de la fonction "somme des chiffres", *Enseign. Math.* **21** (1975), 31–47.

[4] M. Drmota, J. Rivat, The sum-of-digits function of squares, *J. London Math. Soc. (2)* **72** (2005), no. 2, 273–292.

[5] A. O. Gel'fond, Sur les nombres qui ont des propriétés additives et multiplicatives données, *Acta Arith.* **13** (1967/1968), 259–265.

[6] H. Halberstam, K. F. Roth, *Sequences*, Second edition. Springer-Verlag, New York-Berlin, 1983.

[7] K. G. Hare, S. Laishram, T. Stoll, Stolarsky's conjecture and the sum of digits of polynomial values, *Proc. Amer. Math. Soc.*, to appear (2010), `arXiv:1001.4169`.

[8] K. G. Hare, S. Laishram, T. Stoll, The sum of digits of n and n^2, submitted (2010), `arXiv:1001.4170`.

[9] B. Lindström, On the binary digits of a power, *J. Number Theory* **65** (1997), 321–324.

[10] C. Mauduit, J. Rivat, Sur un problème de Gelfond: la somme des chiffres des nombres premiers, *Ann. Math.* **171** (2010), no.3, 1591–1646.

[11] C. Mauduit, J. Rivat, La somme des chiffres des carrés, *Acta Math.* **203** (2009), 107–148.

[12] G. Melfi, On simultaneous binary expansions of n and n^2, *J. Number Theory* **111** (2005), no. 2, 248–256.

[13] M. Peter, The summatory function of the sum-of-digits function on polynomial sequences, *Acta Arith.* **104** (2002), no. 1, 85–96.

[14] K. B. Stolarsky, The binary digits of a power, *Proc. Amer. Math. Soc.* **71** (1978), 1–5.

[15] T. Stoll, Extremal orders of the Zeckendorf sum of digits of powers, submitted (see author's webpage).

A Note on Coloring Factors of Words

Luca Q. Zamboni

1. A Theorem of Ramsey

For each set A, we denote by A_2 the set of all subsets of A consisting of 2 elements. We recall a well known theorem of Ramsey:

Theorem 50. *Let $\mathcal{N}$ be an infinite subset of the natural numbers $\mathbf{N}$, $\mathcal{C}$ a finite non-empty set (the set of colors), and $c : \mathcal{N}_2 \to \mathcal{C}$. Then there exists an element $B \in \mathcal{C}$ and an infinite subset $\mathcal{N}^*$ of $\mathcal{N}$, such that $c(X) = B$ for all $X \in \mathcal{N}_2^*$.*

As a consequence of Theorem 50, we deduce the following:

Corollary 51. *Let $\mathcal{A}$ be a non-empty set (not necessarily finite), and let $W = W_1W_2W_3\cdots \in \mathcal{A}^{\mathbf{N}}$ be an infinite word on the alphabet $\mathcal{A}$. Let $\mathcal{C}$ be a finite non-empty set, and $c : \mathcal{F}^+(W) \to \mathcal{C}$ a coloring of the set of all non-empty factors of W. Then there exists a factorization of W of the form $W = VU_0U_1U_2\cdots$ such that $c(U_i) = c(U_j)$ for all i and j.*

We now consider two variants of Corollary 51: Let W be a non-periodic word on a finite alphabet.

Question 52. *Does there exist a finite coloring*

$$c : \mathcal{F}^+(W) \to \mathcal{C}$$

with the property that for any factoring $W = U_0U_1U_2\cdots$, there exists $i \neq j$ for which $c(U_i) \neq C(U_j)$?

Question 53. *Let $c : \mathcal{F}^+(W) \to \mathcal{C}$ be any (finite) coloring of the set of all non-empty factors of W, and k a positive integer. Then does W contain a factor of the form $U_1U_2\ldots U_k$ with $c(U_i) = c(U_j)$ and $|U_i| = |U_j|$ for all $1 \leq i, j \leq k$?*

We note that both questions are easily answered in case W is periodic: the answer is negative in the case of Question 52 and positive in the case of Question 53. Indeed if $W = UUUU\cdots$ then given any finite coloring $c : \mathcal{F}^+(W) \to \mathcal{C}$, relative to the factorization $W = U_0U_1U_2\cdots$, where each $U_i = U$, we have that $c(U_i) = c(U_j)$ for all $i, j \geq 0$.

2. On Question 52

Let $\mathcal{A}$ be a non-empty finite set, and $W = W_1W_2W_3\cdots \in \mathcal{A}^{\mathbf{N}}$. It is easy to see that there exist non recurrent infinite words W, and colorings of $\mathcal{F}^+(W)$ such that for any factoring $W = U_0U_1U_2\cdots$, there exists $i \neq j$ for which $c(U_i) \neq C(U_j)$. For instance, consider the infinite word $W = 10^\infty$. Given a factor U of W, set $c(U) = 1$ if U contains 1, and $c(U) = 0$ otherwise. Then for any factoring $W = U_0U_1U_2\cdots$, we have $c(U_0) = 1$ while $c(U_i) = 0$ for all $i > 0$.

We next show the existence of recurrent infinite words W and a finite coloring of $\mathcal{F}^+(W)$ such that for every factoring of $W = U_0U_1U_2\cdots$, there exists $i \neq j$ such that $c(U_i) \neq c(U_j)$.

Proposition 54. *Let W be any square free infinite word. Define $c : \mathcal{F}^+(W) \to \{$blue, green$\}$ by*

- *$c(U) =$ green if U is a prefix of W,*
- *$c(U) =$ blue otherwise.*

Then for any factoring $W = U_0U_1U_2\cdots$ we have that $c(U_i) \neq c(U_j)$ for some $i \neq j$.

Proof. Consider any factoring $W = U_0U_1U_2\cdots$. Suppose to the contrary that each U_i has the same color. Then as U_0 is a prefix of W, it follows that each U_i is colored green, i.e., each U_i is a prefix of W. If $|U_i| \leq |U_{i+1}|$ for some $i \geq 0$, then U_i would be a prefix of U_{i+1} and W would contain the factor U_iU_i, contradicting that W is square free. Thus, we must have that $|U_{i+1}| < |U_i|$, a contradiction. □

Let $W \in \{0,1\}^{\mathbf{N}}$ be the Thue-Morse infinite word beginning in 0, that is the fixed point beginning in 0 of the morphism

$$0 \mapsto 01 \quad \text{and} \quad 1 \mapsto 10.$$

It is well known that W does not contain *overlaps,* i.e., factors of the form UUu with u a non-empty prefix of U. Unlike in the previous example, W may be written as a concatenation of prefixes of W (e.g., W may be expressed as a concatenation of the prefixes $0, 01$ and 011); however, we will show that:

Lemma 55. *Let W be the Thue-Morse infinite word beginning in 0, and $W = U_0U_1U_2\cdots$ any factoring of W with each U_i a non-empty prefix of W. Then there exist indices $i \neq j$ for which U_i and U_j terminate in a different letter.*

Proof. Suppose to the contrary that each U_i ends in the same letter $a \in \{0,1\}$. It follows that $|U_n| > |U_{n+1}|$ for every $n \geq 1$. In fact, if $|U_{n+1}| \geq |U_n|$, then W would contain the factor aU_nU_n as U_n is a prefix of U_{n+1}. Since U_n ends in the letter a, we have that W contains an overlap, a contradiction.

□

Proposition 56. *Let W be the Thue Morse infinite word beginning in 0, and define*

$$c : \mathcal{F}^+(W) \to \{\text{red, blue, green}\}$$

by

- *$c(U) =$ red if U is a prefix of W ending in 0,*
- *$c(U) =$ blue if U is a prefix of W ending in 1,*
- *$c(U) =$ green otherwise.*

Then, for any factoring of $W = U_0U_1U_2\cdots$ we have that $c(U_i) \neq c(U_j)$ for some $i \neq j$.

Proof. Let $W = U_0U_1U_2\cdots$ be a factoring of W and suppose to the contrary that $c(U_i) = c(U_j)$ for all i, j. Then, since U_0 is a prefix, it follows that each U_i is a prefix ending in the same letter as U_0. This contradicts the result of the previous lemma. □

Remark 57. A variation of the previous example applies to any binary overlap-free word W.

We now show that the same coloring scheme as in Proposition 56 may be used to show that a characteristic Sturmian word cannot be factored as a concatenation of words all having the same color. The following lemma is an analogue of Lemma 55 for characteristic Sturmian words:

Lemma 58. *Let W be a characteristic Sturmian word and $W = U_0U_1U_2\cdots$ any factoring of W with each U_i a non-empty prefix of W. Then there exist indices $i \neq j$ for which U_i and U_j terminate in a different letter.*

Proposition 59. *Let $W \in \{0,1\}^\infty$ be a characteristic Sturmian word and let*

$$c : \mathcal{F}^+(W) \to \{\textit{red, blue, green}\}$$

by

- $c(U) =$ *red if U is a prefix of W ending in* 0,
- $c(U) =$ *blue if U is a prefix of W ending in* 1,
- $c(U) =$ *green otherwise.*

Then for every factorization $W = U_0U_1U_2\cdots$, we have that $c(U_i) \neq c(U_j)$ for some $i \neq j$.

Proof. Suppose to the contrary that there exists a characteristic Sturmian word W admitting a factorization $W = U_0U_1U_2\cdots$ for which $c(U_i) = c(U_j)$ for all $i, j \geq 0$. Then since U_0 is a prefix of W, it follows that each U_i is a prefix of W ending in the same letter contradicting the result of the previous lemma. □

The above naturally suggest the following questions:

Question 60. *Given a Sturmian word W, does there exist a finite coloring*

$$c : \mathcal{F}^+(W) \to \mathcal{C}$$

with the property that for any factoring $W = U_0U_1U_2\cdots$, there exists $i \neq j$ for which $c(U_i) \neq C(U_j)$? If so, what is the smallest cardinality of $\mathcal{C}$?

Question 61. *Let W be any aperiodic infinite word. Does there exist a 2-coloring*

$$c : \mathcal{F}^+(W) \to \{\textit{red, blue}\}$$

such that for any factoring of $W = U_0U_1U_2\cdots$ there exists $i \neq j$ with $c(U_i) \neq c(U_j)$.

References

[1] J. Berstel, Sturmian and Episturmian words (A survey of some recent results), Lecture Notes in Computer Science, vol. 4728, Springer-Verlag Berlin 2007, pp. 23-47

[2] R. Graham, B. Rothshild, J. Spencer, Ramsey Theory, Wiley 1980.,

[3] M.-P. Schutzenberger, Quelques problemes combinatoires de la théorie des automates, Cours professé à l'Institut de Programmation en 1966/1967.

Reporter: Dirk Nowotka

Participants

Prof. Dr. Jean-Paul Allouche
Laboratoire de Rech. Informatique
CNRS, Universite Paris Sud (Paris XI)
Centre d'Orsay, Bat. 490
F-91405 Orsay Cedex

Prof. Dr. Valerie Berthe
LIRMM
CNRS, Universite Montpellier II
161, rue Ada
F-34392 Montpellier Cedex 5

Prof. Dr. Julien Cassaigne
Institut de Mathematiques de Luminy
CNRS
Case 907 - Luminy
F-13288 Marseille Cedex 9

Prof. Dr. James Currie
Department of Mathematics & Statistics
University of Winnipeg
Winnipeg MB R3B 2E9
CANADA

Dr. Amy Glen
School of Chemical & Mathematical Scien.
Murdoch University
South Street
Murdoch , WA 6150
AUSTRALIA

Prof. Dr. Tero Harju
Department of Mathematics
University of Turku
FIN-20014 Turku

Prof. Dr. Stepan Holub
Faculty of Mathematics and Physics
Charles University
Sokolovska 83
186 75 Praha 8
CZECH REPUBLIC

Prof. Dr. Juhani Karhumäki
Department of Mathematics
University of Turku
FIN-20014 Turku

Prof. Dr. Aldo de Luca
Dip. di Matematica e Applicazioni
Universita di Napoli, Federico II
Complesso Monte S. Angelo
Via Cintia
I-80126 Napoli

Dr. habil. Dirk Nowotka
Institute for Formal Methods in Computer Science (FMI)
Universität Stuttgart
Universitätsstr. 38
70569 Stuttgart

Prof. Dr. Pascal Ochem
Laboratoire de Rech. Informatique
CNRS, Universite Paris Sud (Paris XI)
Centre d'Orsay, Bat. 490
F-91405 Orsay Cedex

Dr. Elena Pribavkina
Department of Algebra & Discrete Math.
Faculty of Mathematics & Mechanics
Ural State University
Lenina 51
620083 Ekaterinburg
RUSSIA

Dr. Eric Rowland
Department of Mathematics
Tulane University
New Orleans , LA 70118
USA

Dr. Kalle Saari
Department of Mathematics
University of Turku
FIN-20014 Turku

Prof. Dr. Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo ONT N2L 3G1
CANADA

Dr. Thomas Stoll
Institut de Mathematiques de Luminy
CNRS
Case 907 - Luminy
F-13288 Marseille Cedex 9

Prof. Dr. Luca Zamboni
Batiment Braconnier 34
Boulevard du 11 Novembre 1918
F-69622 Villeurbanne Cedex