What is the long range order in the Kolakoski sequence?

Report 95-100

F.M. Dekking

TU Delft

Faculteit der Technische Wiskunde en Informatica
Faculty of Technical Mathematics and Informatics

Technische Universiteit Delft
Delft University of Technology

ISSN 0922-5641

Copies of these reports may be obtained from the bureau of the Faculty of Technical Mathematics and Informatics, Julianalaan 132, 2628 BL Delft, phone +31152784568.
A selection of these reports is available in PostScript form at the Faculty's anonymous ftp-site. They are located in the directory /pub/publications/tech-reports at ftp.twi.tudelft.nl

WHAT IS THE LONG RANGE ORDER IN THE KOLAKOSKI SEQUENCE?

F.M. DEKKING
Delft University of Technology
Department of Mathematics and Informatics and
Thomas Stieltjes Institut for Mathematics
Mekelweg 4, 2628 CD Delft
The Netherlands

Abstract. This paper has a survey part in which we collect the known properties of the Kolakoski sequence $2211212212\cdots$ and compare these with the results known for the Prouhet-Thue-Morse sequence $122121122112\cdots$. In the second part we introduce a new object, the Kolakoski measure which presumably describes the frequency behaviour of all subwords of the Kolakoski sequence.

Keywords. Substitutions, block substitutions, Kolakoski sequence, Prouhet-Thue-Morse sequence, unique ergodicity.

1. Introduction

The Kolakoski sequence $x = x_1 x_2 \ldots$, introduced in (Kolakoski, 1965), is a sequence over the alphabet $\{1, 2\}$ defined by the property that the sequence of its runlengths is equal to x itself:

$$x = \underbrace{22}_{2}\ \underbrace{11}_{2}\ \underbrace{2}_{1}\ \underbrace{1}_{1}\ \underbrace{22}_{2}\ \underbrace{1}_{1}\ \underbrace{22}_{2}\ \underbrace{11}_{2}\ \underbrace{2}_{1}\ \begin{matrix}\cdots\\ \cdots\end{matrix}$$

Here a *run* is a maximal subsequence of consecutive identical letters. The only other sequence having this property is obviously given by $\hat{x} = 1x$, i.e., $\hat{x}_1 = 1, \hat{x}_k = x_{k-1}$ for $k \geq 2$.

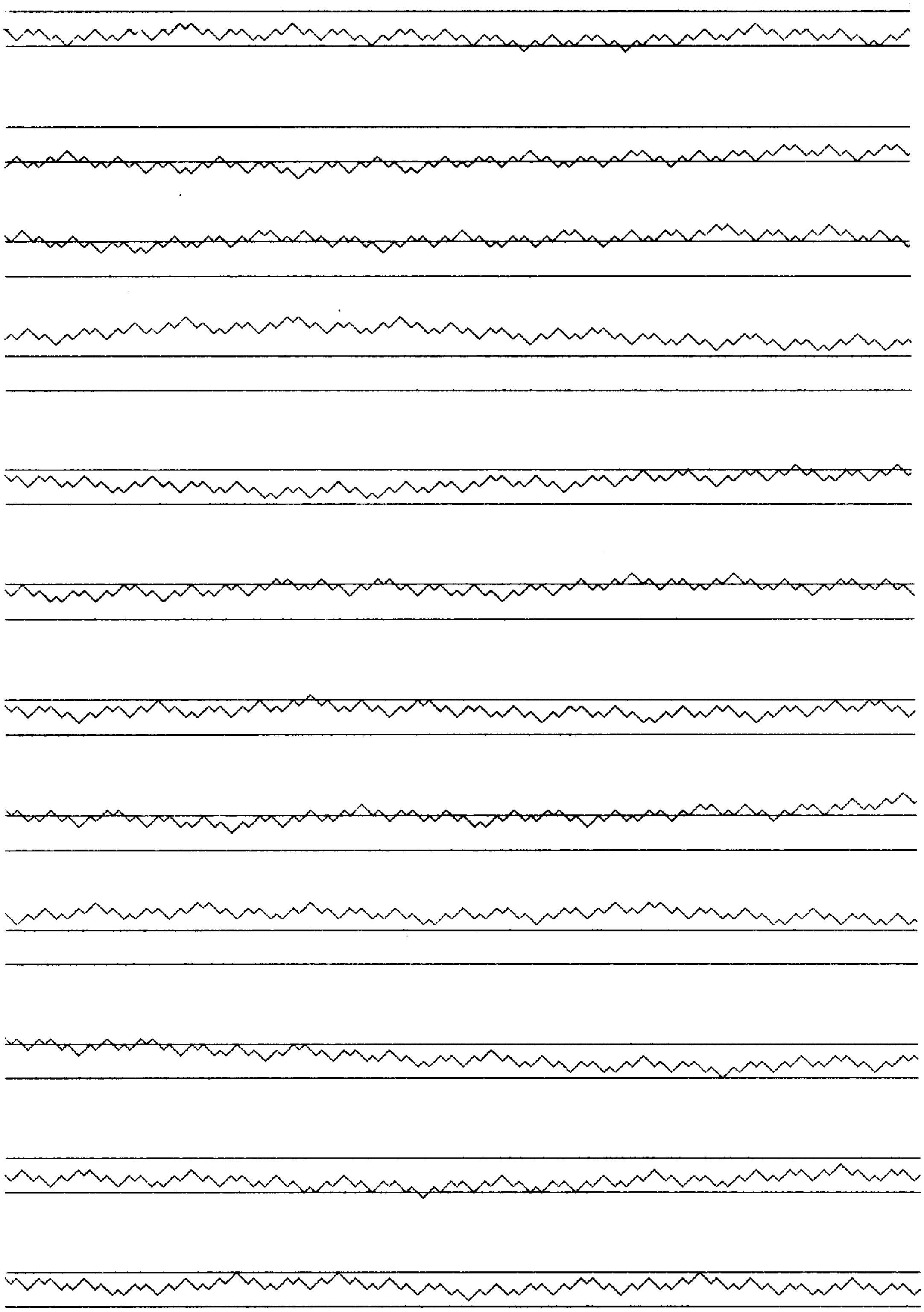

Figure 1. A walk generated by the Kolakoski sequence

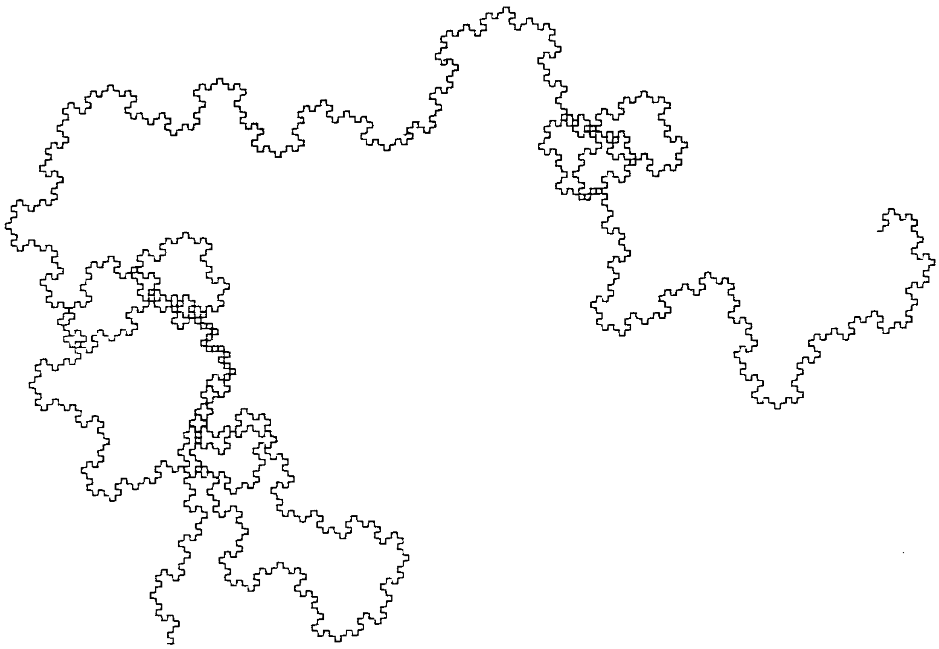

Figure 2. A walk in the plane generated by the Kolakoski sequence

Although this self similarity property might seem somewhat peculiar, the sequence x is in some sense a prototype of a sequence out of a class of sequences which are the next to be studied after the sequences generated by substitutions.

This is the class of sequences which are generated by 2-block substitutions (terminology from (Dekking, 1980), in computer science one speaks of D1L-systems). The 2-block substitution σ which generates x is given by

$$\begin{aligned}\sigma(11) &= 21\\ \sigma(12) &= 211\\ \sigma(21) &= 221\\ \sigma(22) &= 2211.\end{aligned}$$

Iterating σ starting from 22 we obtain

$$22 \xrightarrow{\sigma} 2211 \xrightarrow{\sigma} 221121 \xrightarrow{\sigma} 221121221 \xrightarrow{\sigma} 2211212212211 \xrightarrow{\sigma} \cdots.$$

There is a small problem here: the word $w = \sigma^3(22)$ has odd length, so formally $\sigma(w)$ is not defined. The rule will be that the last letter is ignored if w has odd length. Clearly the iterates $\sigma^n(22)$ converge to the Kolakoski sequence x, and x is the unique fixed point of σ.

A different way to obtain x, by iterating alternatingly the two substitutions $\sigma_1(1 \to 1, 2 \to 11)$ and $\sigma_2(1 \to 2, 2 \to 22)$ has been proposed in (Culik, et al, 1992).

In Figure 1 and 2 we show two walks generated by x in the spirit of (Dekking, 1995). In Figure 1 each 1 in x gives a step $(1,1)$, each 2 a step $(1,-1)$. In Figure 2 the walk is made out of steps of length one, and the symbols in x are now instructions to turn right over 90 degrees ($x_k = 1$) or left over 90 degrees ($x_k = 2$). It is apparent from these figures that x has some long range order, e.g. the number of ones and two's seems very well balanced, patterns occur regularly etc. Mathematical formulations of this long range structure will be given in the next section. It turns out that all these properties can be resolved for fixed points of substitutions (we shall consider our favorite fixed point, the Prouhet-Thue-Morse sequence), but hardly *anything* is known for the Kolakoski sequence.

We remark here that for some of the obvious generalizations of the Kolakoski sequence all these properties are known. Consider e.g. the Kolakoski sequence y on the alphabet $\{1,3\}$:

$$y = 33311133313133311133313331\cdots$$

(again the sequence of runlengths of y is equal to y). This sequence has a much regular structure than x, because ((Dekking, 1980)) it is a letter to letter projection of a *fixed* point of a substitution τ on the four letter alphabet $\{1,3,\overline{1},\overline{3}\}$. The substitution τ is defined by $\tau(1) = 3, \tau(3) = 3\overline{3}3, \tau(\overline{1}) = \overline{1}$, $\tau(\overline{3}) = \overline{1}1\overline{1}$, and the projection by $1, \overline{1} \to 1$ and $3, \overline{3} \to 3$.

In y the digits 1 and 3 do *not* have the same frequency. A simple computation involving the matrix of τ show that the frequency of 3's equals

$$\frac{1}{2}(\frac{43}{54} + \frac{1}{18}\sqrt{177})^{\frac{1}{3}} + \frac{2}{9}(\frac{43}{54} + \frac{1}{18}\sqrt{177})^{-\frac{1}{3}} - \frac{1}{6} = \cdot 60278\cdots.$$

2. Long range properties

Let $z = z_1 z_2 \ldots$ be an infinite sequence with $z_k \in A$ for some finite set A. The first question to ask is whether z is not *eventually periodic*, i.e., whether there exists M and p such that $z_{i+1} \ldots z_{i+p} = z_{i+p+1} \ldots z_{i+2p}$ for all $i \geq M$. This was the problem posed on the Kolakoski sequence at its first appearance in 1965 (Kolakoski, 1965). A stronger version of this asks whether cubes occur in z, i.e., do there exist M and p such that

$$z_{M+1} \cdots z_{M+p} = z_{M+p+1} \cdots z_{M+2p} = z_{M+2p+1} \cdots z_{M+3p}.$$

Recently, it has been shown that the Kolakoski sequence does not contain any cubes (Lepistö, 1993), (Carpi, 1993).

The next problem is to say something about the *subword-complexity* of z, i.e., the cardinality $P_z(n)$ of the set of words of length n that occur in z. The conjecture is that for the Kolakoski sequence

$$c_1 n^\alpha \leq P_x(n) \leq c_2 n^\alpha$$

for some constansts $c_1, c_2 > 0$ and $\alpha = \log 3/\log(3/2)$ ((Dekking, 1981)). It has been proved (Dekking, 1981) that

$$P_x(n) \leq n^{7.2}.$$

This of course implies that the entropy $h(x) := \lim_{n\to\infty} \frac{1}{n} \log P_x(n) = 0$ for the Kolakoski sequence.

When a sequence z is not eventually periodic it might still be that z is *recurrent*, i.e., any word that occurs in z occurs infinitely often, or even *uniformly recurrent*, i.e., any word that occurs in z occurs with bounded gaps (this is sometimes called repetitiveness or Delaunay-recurrence). It is not known whether the Kolakoski sequence is recurrent.

In the case of a two symbol set $A = \{1,2\}$ there are two simple structural properties sequences might have. We call $z \in A^{\mathbb{N}}$ *mirror invariant* if

$$w \text{ occurs in } z \quad \Leftrightarrow \quad \tilde{w} \text{ occurs in } z$$

where $\tilde{w} = \tilde{w}_1 \ldots \tilde{w}_n$ if $w = w_1 \ldots w_n$ and $\tilde{1} = 2, \tilde{2} = 1$.
We call $z \in A^{\mathbb{N}}$ *reversal invariant* if

$$w \text{ occurs in } z \quad \Leftrightarrow \quad \overleftarrow{w} \text{ occurs in } z$$

where $\overleftarrow{w} = w_n \ldots w_2 w_1$ if $w = w_1 w_2 \ldots w_n$.
It is not known whether the Kolakoski sequence is mirror invariant nor whether it is reversal invariant.

Finally we arrive at the existence of frequencies of words, for the word 1 also known as Keane's problem for the Kolakoski sequence ((Keane, 1991)). Does the frequency of 1 in $x = 221121 \cdots$ exist, and is it equal to $\frac{1}{2}$? Despite tremendous efforts this is still an open problem. For some recent work around this problem see (Chvátal, 1994) and (Steacy, 1995). A stronger property is *strict transitivity* , i.e., for each word w in z the frequency of w in $z_k z_{k+1} \cdots$ exists uniformly in k (this implies unique ergodicity of the closed orbit of z).

We summarize the properties discussed in this section in the following table, where we compare the Kolakoski sequence with the Prouhet-Thue-Morse sequence $t = 1221211221121221\ldots$, which is one of the fixed points of the substitution σ given by

$$\sigma(1) = 12$$
$$\sigma(2) = 21.$$

	Prouhet-Thue-Morse	Kolakoski
non-periodicity	easy	Kolakoski, 1965
no cubes	Thue, 1906	Lepistö, Carpi, 1993
subword complexity	$2n \leq P_t(n) \leq \frac{10}{3}n$ Morse and Hedlund, 1938	CONJ: $P_x(n) \asymp n^{\log 2/\log(3/2)}$
recurrence	implied by uniform rec.	CONJECTURED
uniform recurrence	Gottschalk, 1963	CONJECTURED
mirror invariance	easy	CONJECTURED
reversal invariance	easy	CONJECTURED
frequency p_w of w	$p_w \in \{2^{-n}, \frac{1}{3}2^{-n}\}$ Dekking, 1992	CONJ: $p_w \in \{3^{-n}, \frac{1}{2}3^{-n}\}$
strict transitivity	Michel, 1964	CONJECTURED

The conjecture on the values of the frequencies of words in the Kolakoski sequence will be discussed in Section 5.

3. The calculus of Kolakoski words

Let w be a word of $1's$ and $2's$. We call w *differentiable* if no 111 or 222 occurs in w. In that case the derivative w' of w is the word whose j^{th} symbol equals the length of the j^{th} run, *discarding* the first and/or last run if these have length 1.
Examples:
121112 is not differentiable
$(221122)' = 222$
$(221121)' = 221$
$(121)' = 1$
$(12)' = (21)' = 1' = 2' = \epsilon.$

In the last example ϵ denotes the empty word. It is differentiable by definition, with derivative $\epsilon' = \epsilon$.

A word v such that $v' = w$ is called a *primitive* of w. A word can have at most 8 primitives. The primitives of 22 are

1122, 2211, 21122, 12211, 11221, 22112, 211221, 122112.

By definition any word in the Kolakoski sequence x is differentiable, and any $x_1 x_2 \ldots x_k$ has a primitive $x_1 x_2 \ldots x_\ell$. From this we can deduce easily the following somewhat surprising fact about the Kolakoski sequence.

PROPOSITION Mirror invariance implies recurrence.

PROOF: It suffices to show that $x_1 \ldots x_k$ occurs *twice* in x. Choose ℓ such that $(x_1 \ldots x_\ell)' = x_1 \ldots x_k$. By assumption , $\tilde{x}_1 \ldots \tilde{x}_\ell$ also occurs in x. But then $(\tilde{x_1}\tilde{x_2} \ldots \tilde{x_\ell})' = x_1 \ldots x_k$ occurs in x, *not* at the beginning of x. Hence $x_1 \ldots x_k$ occurs at least twice. □.

We call a word a *C^∞-word* if it is arbitrarily often differentiable. Obviously all words occurring in the Kolakoski sequence are C^∞-words, but it is not clear whether the converse in true. In fact, if this were true than (by the previous proposition), the Kolakoski sequence would be recurrent, since we have the following.

PROPOSITION Mirror invariance holds if and only if each C^∞-word occurs in the Kolakoski sequence.

PROOF: Excercise (see also (Dekking, 1981)). □

We remark that the notion of derivative is also useful for obtaining upperbounds on the subword complexity $P_x(n)$ of the Kolakoski sequence. One can show that $|w'| \leq \frac{3}{4}|w|$, where $|w|$ denotes the length of a word. Since a word has no more than 8 primitives one deduces from this that $P_x(n) \leq n^{\log 8/\log(\frac{4}{3})}$ (see (Dekking, 1981) for more details, if necessary).

4. The Kolakoski shift

A natural question is "how many" infinite C^∞-words there are. Let this set be K, and let $D : K \rightarrow K$ be differentiation *without constants*, i.e., the j^{th} symbol of $D(x)$ equals the length of the j^{th} run of x, including the first run of x. So, e.g., $D(12211 \cdots) = 122 \cdots$

PROPOSITION (K, D) is conjugate to the full shift.

PROOF: Define $\varphi : K \rightarrow \{1,2\}^{\mathbb{N}}$ by $(\varphi x)_k = (D^k x)_1$. This is a continuous bijection which by construction is equivariant. □

A consequence of this proposition is that there are "generalized" Kolakoski sequences of any period. For example a period 2 Kolakoski sequence is given by

y	1	1	2	1	1	2	2	1	2	2	1	2	1	1	2	2	$\cdots$
$D(y)$	2		1	2		2		1	2		1	1	2		2		$\cdots$
$D^2(y)$	1		1	2				1	1		2		2				$\cdots$

Does *this* sequence have frequency of 1 equals frequency of 2, equals $\frac{1}{2}$? This question is closely related to the unique ergodicity conjecture which we make in the next section.

5. The Kolakoski measure

We say a C^∞-word w has *degree* n if

$$w^{(n)} \neq \epsilon, \quad w^{(n+1)} = \epsilon.$$

Here $w^{(n)}$ denotes the n^{th} derivative of w. We have for example

$$deg(22) = 1, \quad deg(2) = 0, \quad deg(122112) = 2.$$

In the space $X = \{1,2\}^{\mathbb{N}}$ we consider the *cylinder sets* $[w]$ defined for words $w = w_1 \dots w_m$ by

$$[w] = \{x \in X : x_1 = w_1, \dots, x_m = w_m\}.$$

We define a function μ from the cylinder sets to [0,1] by $\mu([\epsilon]) = 1$, and by $\mu([w]) = 0$ if w is not a C^∞-word, and

$$\mu([w]) = \frac{1}{(1+|w^{(n)}|)3^n} \quad \text{if } w \text{ is } C^\infty\text{-word of degree } n.$$

Note that necessarily $w^{(n)} \in \{1, 2, 12, 21\}$, hence $|w^{(n)}| = 1$ or $|w^{(n)}| = 2$. Examples:

$$\begin{aligned} &\mu([1]) = \mu([2]) = \tfrac{1}{2} \quad (n = 0), \\ &\mu([12]) = \mu([21]) = \tfrac{1}{3} \quad (n = 0), \\ &\mu([11]) = \mu([22]) = \tfrac{1}{6} \quad (n = 1). \end{aligned}$$

For typographical clarity we write $\mu[w]$ instead of $\mu([w])$ in the sequel. The following property follows immediately from the definition of μ.

PROPERTY. If w is a C^∞-word with $|w| \geq 2$ then

$$(\star) \qquad \mu[w'] = 3\mu[w]$$

THEOREM. The function μ extends to a Borel-measure (also denoted μ) on $\{1,2\}^{\mathbb{N}}$. This measure is mirror invariant, reversal invariant and shift invariant.

PROOF: By Kolmogorov's theorem it suffices to check the consistency of μ, i.e., we have to see that for all words w

$$\mu[w1] + \mu[w2] = \mu[w].$$

This is done by induction. The formula holds for $|w| = 1$ - see the examples above. Suppose it holds for all words of length smaller or equal than $n-1$. Let $w = w_1 \dots w_n$ be a word of length n . If w is not a C^∞-word, then both sides equal 0. Otherwise we consider two cases.

Case 1. $w = v11$. Then

$$\mu[w1] + \mu[w2] = \mu[v111] + \mu[v112] = \mu[v11] = \mu[w],$$

since $v111$ is not differentiable, and since $v112$ has the same derivative as v11.

Case 2. $w = v12$. Then

$$(w1)' = (v121)' = w'1; \quad (w2)' = (v122)' = w'2.$$

So

$$\begin{aligned} \mu[w1] + \mu[w2] &\overset{(\star)}{=} \frac{1}{3}\mu[(w1)'] + \frac{1}{3}\mu[(w2)'] \\ &= \frac{1}{3}\mu[w'1] + \frac{1}{3}\mu[w'2] \\ &= \frac{1}{3}\mu[w'] \\ &\overset{(\star)}{=} \mu[w], \end{aligned}$$

where the last but one equality sign holds by the induction hypothesis. Since $\tilde{w}' = w'$ the two remaining cases $w = v21$ and $w = v22$ follow suit. The fact that a word and its mirror image have the same derivative also implies mirror invariance of μ. Reversal invariance follows since the reversal operation commutes with taking derivatives (hence $deg(\overleftarrow{w}) = deg(w)$ and $|w^{(n)}| = |\overleftarrow{w}^{(n)}|$). Since the cylinders are an intersection stable class of sets generating the Borel σ-algebra, for shift invariance it suffices to show that for all words w

$$\mu[w] = \mu(T^{-1}[w]) = \mu[1w] + \mu[2w]$$

where $T : X \to X$ is the shift defined by $(Tx)_k = x_{k+1}$ for $k = 1, 2, \ldots$. But this folllows directly from reversal invariance and consistency. □

Clearly the support of μ is the closed, shift invariant set of all C^∞-words. We conjecture that μ is the unique shift invariant measure on this set, and hence that μ is ergodic.

We now connect the measure μ to the (conjectured) existence of the frequencies p_w of words w in the Kolakoski sequence x.

PROPOSITION Suppose the frequencies p_w exist and that $p_w = p_{\tilde{w}}$ for all words w occurring in x. Then for all words w

$$p_w = \mu[w].$$

PROOF. We consider the $N's$ such that $x_{N-1}x_N = 11$ or 22. For any such N let $M = M(N)$ be such that $(x_1x_2 \ldots x_N)' = x_1x_2 \ldots x_M$. Let us write $A_N(w)$ for the number of times that the word w occurs in $x_1 \ldots x_N$. We then have

$$A_M(1) + 2A_M(2) = N.$$

We also have

$$A_M(w') = A_N(w) + A_N(\tilde{w}).$$

Hence

$$\frac{A_M(w')}{A_M(1) + 2A_M(2)} = \frac{A_N(w)}{N} + \frac{A_N(\tilde{w})}{N}.$$

Letting $N \to \infty$, we obtain

$$\frac{p'_w}{p_1 + 2p_2} = p_w + p_{\tilde{w}}.$$

Since, by assumption, $p_1 = p_2 = \frac{1}{2}$ this implies $p_{w'} = 3p_w$. Since $\mu[w'] = 3\mu[w]$, the propostion will follow by induction on the length of w. □

REMARK. A result similar to this proposition has been obtained by Y. Peres and M. Bozhernistan (Peres, 1994).

References

Carpi, A. (1993) Repetitions in the Kolakovski sequence, *Bull. of the Eur. Ass. for Theor. Comp. Sc.* **50**, pp. 194-196.

Chvátal, V. (1994) Notes on the Kolakoski sequence, *DIMACS Technical Report* **93-84** (revised).

Culik, K. and Karhumäki, J. (1992) Iterative devices generating infinite words, *Lec. Notes in Comp. Sc.* **577**, Springer, Berlin, pp. 531-544.

Dekking, F. M. (1980) Regularity and irregularity of sequences generated by automata, *Sém. Th. Nombres Bordeaux* '79 - '80, pp. 901-910.

Dekking, F. M. (1981), On the structure of selfgenerating sequences, *Sém. Th. Nombres Bordeaux* '80-'81, pp. 3101-3106.

Dekking, F. M. (1992) On the Thue-Morse measure, *Acta Univ. Carolin.- Math.Phys.* **33**, pp. 35-40.

Dekking, F. M. (1995) Random and automatic walks. In: *Beyond Quasicrystals*, F. Axel and D. Gratias (Eds.), les Editions de Physique and Springer, Berlin.

Gottschalk, W. (1963) Substitution minimal sets, *Trans.Amer.Math.Soc.* **109**, pp. 467-491.

Keane, M. S. (1991) Ergodic Theory and Subshifts of finite type, in: *Ergodic Theory, Symbolic Dynamics and Hyperbolic spaces.* T. Bedford, M. Keane, C. Series (Eds.), Oxford University Press, Oxford.

Kolakoski, W. (1965) Self generating runs, Problem 5304, *Amer. Math. Monthly* **72**, p. 674. Solution: *Amer.Math. Monthly* **73**, pp. 681-682.

Lepistö, A. (1993) Repetitions in the Kolakoski sequence, preprint.

Michel, P. (1974) Stricte ergodicité d'ensembles minimaux de substitutions, *C.R. Acad. Paris Sér. A* **278**, pp. 811-813.

Morse, M. and Hedlund, G. A. (1938) Symbolic Dynamics, *Amer. J. Math* **60**, pp. 815-866.

Peres, Y. (1994) Personal communication.

Steacy, R. (1995) Structure in the Kolakoski sequence, preprint.

Thue, A. (1906) Uber unendlichen Zeichenreihen, *Norske Vid.Selsk.Skr. I. Math. -Nat. KL (Kristiania)* **7**, pp. 1-22.