NUMBER THEORY AND FORMAL LANGUAGES

JEFFREY SHALLIT*

Abstract. I survey some of the connections between formal languages and number theory. Topics discussed include applications of representation in base k, representation by sums of Fibonacci numbers, automatic sequences, transcendence in finite characteristic, automatic real numbers, fixed points of homomorphisms, automaticity, and k-regular sequences.

Key words. finite automata, automatic sequences, transcendence, automaticity

AMS(MOS) subject classifications. Primary 11B85, Secondary 11A63 11A55 11J81

1. Introduction. In this paper, I survey some interesting connections between number theory and the theory of formal languages. This is a very large and rapidly growing area, and I focus on a few areas that interest me, rather than attempting to be comprehensive. (An earlier survey of this area, written in French, is [1].) I also give a number of open questions.

Number theory deals with the properties of integers, and formal language theory deals with the properties of strings. At the intersection lies

(a) the study of the properties of integers based on their *representation* in some manner — for example, representation in base k; and

(b) the study of the properties of strings of digits based on the integers they represent.

An example of a theorem of type (a) — perhaps the first significant one — is the famous theorem of Kummer [60, pp. 115–116], which states that the exponent of the highest power of a prime p which divides the binomial coefficient $\binom{n}{m}$ is equal to the number of "carries" when m is added to $n - m$ in base p.

For type (b) I do not know a theorem as fundamental as Kummer's. But here is a little problem that some may find amusing. Call a set of strings *sparse* if, as $n \to \infty$, it contains a vanishingly small fraction of all possible strings of length n. Then can one find a sparse set S of strings of 0's and 1's such that every string of 0's and 1's can be written as the concatenation of two strings from S? One solution is to let S be the set of all strings of 0's and 1's such that the number of 1's is a sum of two squares. Then by a famous theorem in number theory — Lagrange's theorem — every number n is the sum of *four* squares, so every string of 0's and 1's is a concatenation of two strings chosen from S. The sparseness of S follows

* Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. E-mail: `shallit@graceland.uwaterloo.ca` . Research supported in part by a grant from NSERC.

from an estimate in sieve theory [38]. Further examples of theorems of type (b) can be found in Section 8.1.

It may be objected that studying the formal language aspects of number theory is somewhat artificial, in the sense that it depends on choosing one particular representation — such as representation in base 2 — and that there is no reason to choose base 2 over any other base. For example, recall the famous objection of Hardy to certain kinds of digital problems[1]:

> These are odd facts, very suitable for puzzle columns and likely to amuse amateurs, but there is nothing in them which appeals much to a mathematician. The proofs are neither difficult nor interesting — merely a little tiresome. The theorems are not serious; and it is plain that one reason (though perhaps not the most important) is the extreme speciality of both the enunciations and the proofs, which are not capable of significant generalization. [46, p. 105]

I offer four answers to Hardy's objection. First, we attempt to make our theorems as general as possible. For example, we can try to prove theorems for all bases k rather than just a single base. Second, sometimes some bases *do* occur naturally in problems, and base 2 is one of them; see Section 4. Third, the area has proved to have many applications; perhaps the most dramatic examples are the recent simple proofs of transcendence in finite characteristic by Allouche and others; see Section 5. Finally, the area is "natural", and I submit as evidence the fact that many good mathematicians throughout history have worked in it, including Kummer, Lucas, and Carlitz.

2. Notation. I begin with some notation for formal languages, for which a good reference is the book of Hopcroft and Ullman [49].

Let Σ be a finite list of symbols, or *alphabet*, and let Σ^* denote the free monoid over Σ, that is, the set of all finite strings of symbols chosen from Σ, with concatenation as the monoid operation. Thus, if $\Sigma = \{0, 1\}$, then

$$\Sigma^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\},$$

where ϵ is the notation for the empty string. A *formal language*, or just *language*, is defined to be any subset of Σ^*.

Let L, L_1, L_2 be languages. We define the concatenation of languages as follows:

$$L_1 L_2 = \{x_1 x_2 \ : \ x_1 \in L_1, \ x_2 \in L_2\}.$$

[1] The two problems he cited as examples were (a) show that 8712 and 9801 are the only four-digit numbers which are nontrivial integral multiples of their reversals and (b) show that 153, 370, 371, and 407 are the only integers > 1 which are equal to the sum of the cubes of their decimal digits. Today, digital problems continue to attract attention and criticism; see, for example, [35].

Define $L^0 = \{\epsilon\}$, and $L^i = LL^{i-1}$ for $i \geq 1$. We define the *Kleene closure* of a language by

$$L^* = \bigcup_{i \geq 0} L^i.$$

A *regular expression* over an alphabet Σ is a way to denote certain languages — a finite expression using the symbols in Σ together with $+$ (to denote union), $*$ (to denote Kleene closure), ϵ (to denote the empty string), $\emptyset$ (to denote the empty set), and parentheses for grouping. For example, the regular expression $(\epsilon + 1)(0 + 01)^*$ denotes the set of all strings over $\{0, 1\}$ containing no two consecutive 1's. If a language can be represented by a regular expression, it is said to be *regular*.

3. Number representations. In order to talk about numbers in formal language theory terms, we need a way to represent numbers as strings of symbols over a finite alphabet. Let us begin with the integers. A classical way to do this is the canonical representation in base k:

THEOREM 3.1. *Let k be an integer ≥ 2. Then every positive integer n can be represented uniquely in the form $n = \sum_{0 \leq i \leq r} a_i k^i$, where the a_i are integers with $0 \leq a_i < k$, and $a_r \neq 0$.*

By associating n with the string $a_r a_{r-1} \cdots a_1 a_0$, this theorem gives a bijection between the positive integers and the set of strings given by the regular expression $(\Sigma_k - \{0\})\Sigma_k^*$, where $\Sigma_k = \{0, 1, 2, \ldots, k-1\}$. We define $(n)_k$ to be the string $a_r a_{r-1} \cdots a_1 a_0$ representing n in base k. We also define the inverse map $[w]_k$ to be the value of the string w when interpreted as a base-k number. We define $(0)_k = \epsilon$ and $[\epsilon]_k = 0$.

There are many relationships between base-k representation and elementary number theory. Here is just one example. Given an integer n, we may form $s_k(n)$, the sum of its base-k digits. For a prime p, let $\nu_p(n)$ denote the exponent of the highest power of p dividing n. Then we have the following classical theorem of Legendre [61, Vol. I, p. 10]:

THEOREM 3.2. *Let p be a prime number. Then for all $n \geq 0$ we have*

$$\nu_p(n!) = \frac{n - s_p(n)}{p - 1}.$$

One annoyance is that the canonical representation in base k suffers from the "leading zeros" problem — that is, the map $w \to [w]_k$ is not one-one if $w \in \Sigma_k^*$. For example, $[101]_2 = [0101]_2 = [00101]_2 = 5$. One way around this difficulty is the following simple "folk theorem", whose precise origins are unknown to me (but see [87, Note 9.1, pp. 90–91], [101, p. 24], and [40]):

THEOREM 3.3. *Let k be an integer ≥ 2. Then every non-negative integer can be represented uniquely in the form $n = \sum_{0 \leq i \leq r} a_i k^i$, where the a_i are integers with $1 \leq a_i \leq k$.*

For example, $13 = 2 \cdot 4 + 2 \cdot 2 + 1 \cdot 1$. This theorem gives a bijection between $\mathbb{N}$, the non-negative integers, and the regular language $(1 + 2 + \cdots + k)^*$.

There are many other ways to represent the non-negative integers. For example, let the Fibonacci numbers be defined by $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$. The following theorem gives the so-called *Zeckendorf* or *Fibonacci* representation [65,107]:

THEOREM 3.4. *Every non-negative integer can be represented uniquely in the form* $\sum_{2 \leq i \leq r} a_i F_i$, *where* $a_i \in \{0, 1\}$, *and* $a_i a_{i+1} \neq 1$.

This theorem gives a bijection between $\mathbb{N}$ and the regular language $\epsilon + 1(0 + 01)^*$. Notice that in all three cases we have examined, the set of "valid" representations is a regular language. This observation raises the question, for what numeration systems is the set of valid representations regular? See, for example, [91,48,67].

As above, if m and n are integers, then we can uniquely write $m = 2^{a_1} + \cdots + 2^{a_c}$ and $n = 2^{b_1} + \cdots + 2^{b_d}$, where $a_1 < \cdots < a_c$ and $b_1 < \cdots < b_d$. We clearly have

$$mn = \sum_{1 \leq i \leq c} \sum_{1 \leq j \leq d} 2^{a_i + b_j}.$$

Knuth [57] found a surprising generalization of this identity: if the Zeckendorf representation of m is $F_{a_1} + F_{a_2} + \cdots + F_{a_c}$, and the Zeckendorf representation of n is $F_{b_1} + F_{b_2} + \cdots + F_{b_d}$, define

$$m \circ n = \sum_{1 \leq i \leq c} \sum_{1 \leq j \leq d} F_{a_i + b_j}.$$

Then the $\circ$ multiplication is associative! Also see [7,43].

We now turn to the representation of rational numbers. Let $[a_0, \ldots, a_n]$ be an abbreviation for the *continued fraction*

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_n}}}. \tag{3.1}$$

THEOREM 3.5. *Every rational number in* $(0, 1)$ *can be expressed uniquely in the form*

$$[0, a_1, a_2, \ldots, a_n]$$

where the a_i *are positive integers and* $a_n \geq 2$.

As an application of this theorem, we prove the following theorem, inspired by [77]:

THEOREM 3.6. *There is a bijection* $r : \mathbb{N} \to \mathbb{Q}$ *such that both* r *and* r^{-1} *are computable in polynomial time.*

Proof. It suffices to give such a bijection between $\mathbb{N}$ and $\mathbb{Q} \cap (0,1)$.

Let $f_k : \mathbb{N} \rightarrow (1+2+\cdots+k)^*$ be the map that takes a non-negative integer to its representation in base k using digits $\{1,2\ldots,k\}$, as discussed in Theorem 3.3, and let f_k^{-1} be the inverse map. Let g be the map which takes a string over $(1+2+3)^*$ as an argument and returns a list of strings, where the 3's are treated as delimiters. For example, $g(121313322) = (121, 1, \epsilon, 22)$. Let h be the map such that

$$h(a_1, a_2, \ldots, a_k) = (0, a_1+1, \ldots, a_{k-1}+1, a_k+2).$$

Then we define the bijection r as follows:

$$r(n) = [h(f_2^{-1}(g(f_3(n))))],$$

where the function f_2^{-1} is extended in the obvious way to operate on lists of strings.

For example, consider the case $n = 12590$. Then its representation in base 3 using digits $\{1,2,3\}$ is 121313322. This is transformed by g into the list $(121, 1, \epsilon, 22)$, which is mapped by f_2^{-1} into $(9,1,0,6)$. Then h maps this to $(0,10,2,1,8)$. Hence $r(12590) = [0,10,2,1,8] = 26/269$.

It remains to see that r and r^{-1} can be computed in polynomial time. That f_3 and f_2^{-1} can be computed in polynomial time is easy, and is left to the reader. For the polynomial time computability of continued fractions, see, for example, [8, Chapter 4]. $\square$

There are many other formal language aspects of continued fractions. Some of these deal with the so-called "LR" or "Stern-Brocot" representation of rational numbers [44]. If

$$\theta = [a_0, a_1, a_2, \ldots],$$

then the LR-representation of θ is the string

$$R^{a_0} L^{a_1} R^{a_2} L^{a_3} \cdots.$$

Let a, b, c, d be integers with $ad - bc \neq 0$. Raney [83] gave a finite-state transducer to compute the LR-expansion of $\tau = (a\theta + b)/(c\theta + d)$ from that of θ. Using Raney's theorem, one can give a purely formal-language-theoretic proof of the fact that θ has bounded partial quotients iff τ does [90].

4. The Thue-Morse sequence. Recall from the previous section that $s_2(n)$ denotes the sum of the bits in the base-2 representation of n.

Now define an infinite word $\mathbf{t} = t_0 t_1 t_2 \cdots$ over $\{0,1\}$, as follows: $t_n = s_2(n) \bmod 2$. This infinite word is sometimes called the Thue-Morse sequence, because both Thue [99] and Morse [75] examined its properties near the beginning of this century. But Prouhet implicitly used the definition of $\mathbf{t}$ in an 1851 paper ([82]; also see [104]) that gave a solution to the multigrade problem.

The *multigrade problem* (or Tarry-Escott problem; see [62]) is to find disjoint sets U, V that $\sum_{u \in U} u^i = \sum_{v \in V} v^i$ for $i = 0, 1, \ldots, k-1$. Prouhet observed that one could take $U = \{0 \leq n < 2^k : t_n = 0\}$ and $V = \{0 \leq n < 2^k : t_n = 1\}$. For example, we have

$$0^i + 3^i + 5^i + 6^i = 1^i + 2^i + 4^i + 7^i$$

for $i = 0, 1, 2$.

Another result of number-theoretic interest related to the Thue-Morse sequence is the following. Woods [103] and Robbins [85] observed that

$$\prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_n}} = \frac{\sqrt{2}}{2}. \tag{4.1}$$

Here is a simple proof, due to Jean-Paul Allouche: Let $P = \prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_n}}$ and let $Q = \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_n}}$. Clearly

$$PQ = \frac{1}{2} \prod_{n \geq 1} \left(\frac{n}{n+1} \right)^{(-1)^{t_n}}.$$

Now break this infinite product into separate products over odd and even indices; we find

$$\begin{aligned} PQ &= \frac{1}{2} \prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_{2n+1}}} \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_n}} \\ &= \frac{1}{2} P^{-1} Q. \end{aligned}$$

It follows that $P^2 = \frac{1}{2}$. (Convergence and correctness of the rearrangements are left to the reader.)

But in fact, even more is true. Suppose one tries to express $\frac{\sqrt{2}}{2}$ as an infinite product of terms of the form $(\frac{2n+1}{2n+2})^{\pm 1}$, where the sign for $n = 0$ is chosen to be $+1$, and then iteratively chosen according to a greedy algorithm: if the product constructed so far is greater than $\frac{\sqrt{2}}{2}$, choose the sign $+1$, and if the product constructed so far is smaller than $\frac{\sqrt{2}}{2}$, choose the sign -1. Then the sequence of signs chosen is exactly $(-1)^{t_n}$. I conjectured this in 1983 [89], and it was proved by Allouche and Cohen in 1985 [5].

Notice that the technique used above does not let us conclude anything about the number Q. In analogy with (4.1), one may ask the following

OPEN QUESTION 1. *Is the number*

$$Q = \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_n}} \doteq 1.6281601297189$$

algebraic?
No simple formula for the number Q is known, although it appears in a somewhat disguised form in a paper of Flajolet and Martin [39, Theorem 3.A], where (using their notation) $\varphi = 2^{-1/2} e^{\gamma} Q^{-1}$.

5. Automatic sequences. The Thue-Morse sequence is a member of a much larger class of sequences called *k-automatic sequences*; more precisely, the Thue-Morse sequence is 2-automatic.

Let us recall the basics of finite automata. A *deterministic finite automaton*, or DFA, is a simple model of a computer. Formally it is a quintuple, $M = (Q, \Sigma, \delta, q_0, F)$, where

- Q is a finite set of *states*;
- Σ is a finite set of symbols, called the *input alphabet*;
- $q_0 \in Q$ is the *initial state*;
- $F \subseteq Q$ is the set of *final states*;
- $\delta : Q \times \Sigma \to Q$ is the *transition function.*

The transition function δ is extended in the obvious way to a map from $Q \times \Sigma^*$ into Q.

The *language accepted by M* is denoted by $L(M)$ and is given by $\{w \in \Sigma^* \mid \delta(q_0, w) \in F\}$. As an example, consider the automaton in Figure 5.1, which accepts exactly the strings over $\{0, 1\}$ that are the base-2 representations of the primes between 2 and 11.

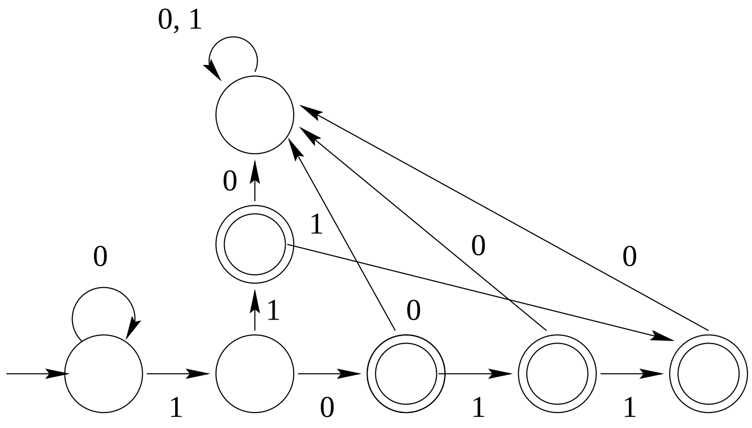

FIG. 5.1. *Automaton accepting the base-2 representations of the primes p where $2 \leq p \leq 11$*

Note that the start state is at the lower left, and is indicated, as is customary, by an unlabeled arrow with no source. Also, final states are denoted by double circles.

We may also provide our automaton with output. In this case we discard the set of final states from the definition of the DFA and add back Δ (the output alphabet) and $\tau : Q \to \Delta$ is the output mapping.

DEFINITION 5.1. *We say a sequence* $(s_i)_{i \geq 0}$ *over a finite alphabet* Δ *is* k*-automatic if there exists a deterministic finite automaton with output (DFAO)* $M = (Q, \Sigma_k, \Delta, \delta, \tau, q_0)$ *(where* τ *is a mapping taking* Q *to* Δ*) such that* $\tau(\delta(q_0, (n)_k)) = s_n$ *for all* $n \geq 0$*.*

These sequences are sometimes called *uniform tag sequences* [27] or k-recognizable sequences [37, p. 106] in the literature.

Another characterization of automatic sequences is the following. Suppose $(s(n))_{n \geq 0}$ is a sequence over a finite alphabet. Define $K_k(s)$, the k-kernel of s, to be the set of subsequences

$$K_k(s) = \{(s(k^i n + a))_{n \geq 0} \ : \ i \geq 0, \ 0 \leq a < k^i\}.$$

Then $(s(n))_{n \geq 0}$ is k-automatic iff the set $K_k(s)$ is finite.

Many sequences that occur in number theory turn out to be k-automatic for some small integer k. For example, let B be an integer ≥ 3, and consider the real number $f(B) = \sum_{k \geq 0} B^{-2^k}$. This is a transcendental number[2] ([53,15,71,68,56]; [76, Thm. 1.1.2]) whose continued fraction has bounded partial quotients [88,34]:

$$\begin{aligned} f(B) &= [a_0, a_1, a_2, \ldots] \\ &= [0, \ B-1, \ B+2, \ B, \ B, \ B-2, \ B, \ B+2, \ B, \ \ldots]. \end{aligned}$$

In fact, its continued fraction can be generated by the simple finite automaton with ten states in Figure 5.2.

For example, to compute a_{12}, we compute $(12)_2 = 1100$, and then feed the digits into the automaton, starting at the top. The output is the label of the last state reached, which is $B - 2$.

Probably the most interesting and useful number-theoretic aspect of automatic sequences is the following theorem of Christol [23,24]:

THEOREM 5.1. *Let* Δ *be a nonempty finite set,* $(a_i)_{i \geq 0}$ *be a sequence over* Δ*, and* p *be a prime number. Then* $(a_n)_{n \geq 0}$ *is* p*-automatic iff there exists an integer* $m \geq 1$ *and an injection* $\beta : \Delta \to GF(p^m)$ *such that the formal power series* $\sum_{n \geq 0} \beta(a_n) X^n$ *is algebraic over* $GF(p^m)(X)$*.*

As an example, consider the Thue-Morse sequence $(t_n)_{n \geq 0}$, which is 2-automatic. Let $T(X) = \sum_{n \geq 0} t_n X^n$.

$$T(X) = X + X^2 + X^4 + X^7 + X^8 + X^{11} + \cdots$$

Now

$$\begin{aligned} T(X) &= \sum_{n \geq 0} t_n X^n \\ &= \sum_{n \geq 0} t_{2n} X^{2n} + \sum_{n \geq 0} t_{2n+1} X^{2n+1} \end{aligned}$$

[2] Sometimes called the 'Fredholm number', although Fredholm apparently never worked on it.

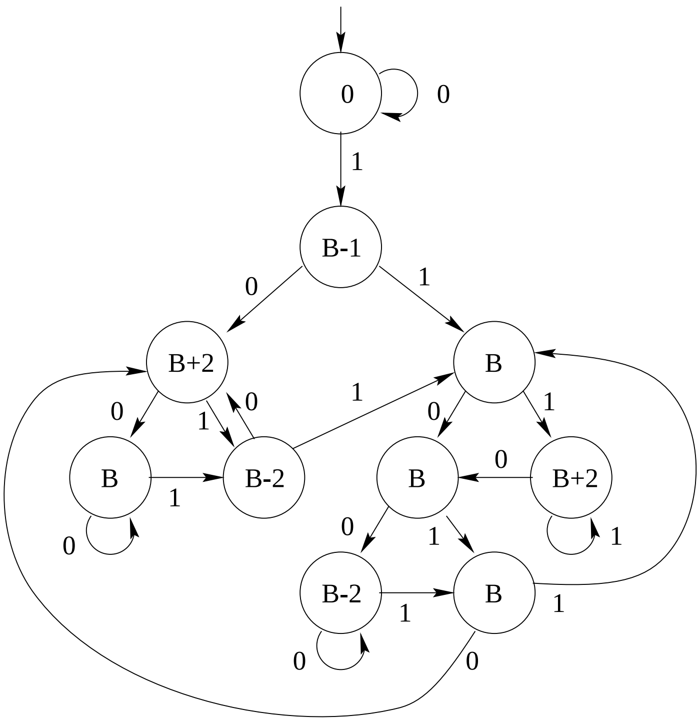

FIG. 5.2. *Automaton generating the continued fraction expansion of $f(B)$*

$$
\begin{aligned}
&= \sum_{n\geq 0} t_n X^{2n} + X \sum_{n \geq 0} (t_n + 1) X^{2n} \\
&= T(X^2) + XT(X^2) + X \frac{1}{1 - X^2}.
\end{aligned}
$$

Hence we have, over $GF(2)$,

$$(1+X)^3 T(X)^2 + (1+X)^2 T(X) + X = 0.$$

The theorem of Christol is remarkable because it relates a purely number-theoretic fact (algebraicity in finite characteristic) to a purely machine-theoretic fact (generation by a finite automaton). As a consequence, one may obtain transcendence results in finite characteristic by proving that no finite automaton can generate the sequence of coefficients of an appropriate formal power series. For example, Allouche [2] used this technique to give a new proof of the transcendence of π_q, the analogue of π in the field of formal Laurent series over $GF(q)$.

Other results along this line include those of Berthé [11,12], who proved that $\frac{\zeta_q(n)}{\pi_q^n}$ is transcendental for $1 \leq n \leq q-2$, a result previously proved by Yu [105] for every n such that $(q-1) \nmid n$. Here ζ_q is the Carlitz zeta-function, the formal power series analogue of the ordinary zeta-function. Recher [84] obtained transcendence results for periods of generalized Carlitz exponentials, i.e., of generalizations of π_q. Berthé [13] proved transcendence results for the Carlitz logarithm and gave results on linear expressions in $\frac{\zeta_q(n)}{\pi_q^n}$ for $1 \leq n \leq q-2$ [14]. Allouche [3] proved the transcendence of the values of the Carlitz-Goss gamma function for all p-adic rational arguments that are not natural numbers, and Mendès France and Yao [73] extended the result to *all* the values of the Carlitz-Goss gamma function at p-adic arguments that are not natural numbers. Thakur proved [98] that the period of the Tate elliptic curve is transcendental.

6. Automatic real numbers. Given a k-automatic sequence $(s_i)_{i \geq 0}$ over the alphabet $\Sigma = \{0, 1, 2, \ldots, b-1\}$, we may consider the sequence to represent the base-b representation of a real number. The number $\sum_{i \geq 0} b^{-2^i}$ is an example of such a number, discussed in the previous section.

Or consider the Thue-Morse real number $\sum_{i \geq 1} t_{i-1} 2^{-i}$, whose base-2 representation is

$$\mathcal{T} = .0110100110010110 \cdots.$$

It follows from a general result of Mahler [71] that $\mathcal{T}$ is transcendental. Mahler's proof technique was later rediscovered by Cobham [26] and Dekking [30].[3]

It may be amusing to note that the number $\mathcal{T}$ appears "naturally" as a certain probability in formal language theory. Let $\mathcal{P}$ be the probability that a randomly-chosen language over $\{0, 1\}$ contains at least one word of every possible length. (Our model is to decide the membership of each word in L uniformly at random, with probability $\frac{1}{2}$.) Then

$$\mathcal{P} \quad = \quad \prod_{i \geq 0} (1 - 2^{-2^i}) = \sum_{j \geq 0} \frac{(-1)^{t_j}}{2^j} = \sum_{j \geq 0} \frac{1 - 2t_j}{2^j} = 2 - 4\mathcal{T}.$$

This result suggests the following

CONJECTURE 2. *Let k, b be integers ≥ 2. If $(s_i)_{i \geq 0}$ is a non-ultimately-periodic k-automatic sequence over the alphabet $\Sigma = \{0, 1, 2, \ldots, b-1\}$, then the number $\sum_{i \geq 0} s_i b^{-i}$ is transcendental.*

For some time it was believed that Loxton and van der Poorten had completely resolved this problem [69,70], but gaps in the proof have been pointed out by Paul-Georg Becker.

CONJECTURE 3. *No number of the form $\sum_{i \geq 0} s_i b^{-i}$, where $(s_i)_{i \geq 0}$ is a k-automatic sequence, and b is an integer ≥ 2, is a Liouville number.*

[3] Michel Dekking has kindly pointed out a minor, easily-repairable flaw in his proof.

Becker conjectures (personal communication, 1993) that in fact these numbers, when transcendental, are S-numbers in Mahler's classification ([72], [58, p. 63]).

Recently there have been some other interesting results on real numbers whose base-b expansions are k-automatic. Denoting the set of such numbers as $L(k, b)$, we have the following theorem of Lehr [63]:

THEOREM 6.1. *The set $L(k, b)$ forms a $\mathbb{Q}$-vector space.*

However, it can be shown that the set $L(k, b)$ is not closed under product; that is, $L(k, b)$ is not a ring [64]. The structure of $L(k, b)$ is still somewhat mysterious, although it is known that $L(k, b)$ is infinite dimensional over $\mathbb{Q}$. In fact, for each $B \geq 2$, we have $\mathbb{Q}[f(B)] \subset L(2, B)$, where f is the function defined in Section 5. Since $f(B)$ is transcendental over $\mathbb{Q}$, we have $\mathbb{Q}[f(B)]$ is infinite dimensional over $\mathbb{Q}$. See [64].

It would be nice to prove that some classical real numbers are not automatic numbers. For example, we have

CONJECTURE 4. *The numbers π, e, and $\ln 2$ are not in $L(k, b)$ for any $k, b \geq 2$.*

This conjecture would follow, for example, if it were proved that these numbers were normal.

7. Fixed points of homomorphisms. As Cobham observed [27], the k-automatic sequences discussed in the previous section can also be characterized as images (under a length-preserving homomorphism, or *coding*) of fixed points of *uniform* homomorphisms (i.e., homomorphisms φ with $|\varphi(a)| = k$ for all $a \in \Sigma$). For example, the Thue-Morse word is the unique fixed point, starting with 0, of the map which sends 0 to 01 and 1 to 10.

One can also study the fixed points of homomorphisms that are not necessarily uniform. The *depth* of a homomorphism $\varphi : \Sigma \to \Sigma^*$ is defined to be $|\Sigma|$, and the *width* is $\max_{a \in \Sigma} |\varphi(a)|$.

Suppose that $\varphi : \Sigma \to \Sigma^*$ is a homomorphism with the property that $\varphi(a) = ax$ for some letter $a \in \Sigma$. (We call such a homomorphism *extendable* on a.) Then

$$ax\varphi(x)\varphi^2(x)\varphi^3(x)\cdots$$

is a fixed point of φ, and if x contains at least one letter which is not ultimately sent to ϵ by repeated applications of φ, then this fixed point is infinite.

OPEN QUESTION 5. *Given a homomorphism φ extendable on a, of depth m and width n, can one compute the ith letter of the fixed point starting with a in time polynomial in m, n, and $\log i$?*

Note that this question is easily answerable in the affirmative when the homomorphism is uniform.

A particular fixed point that has been studied extensively is the so-called infinite *Fibonacci word*

$$\mathbf{f} = f_1 f_2 f_3 \cdots = 0100101001001\cdots,$$

which is the fixed point of the map $\varphi(0) = 01$ and $\varphi(1) = 0$ [9,10]. It can be shown that

$$f_n = 1 - \lfloor (n+1)\alpha \rfloor + \lfloor n\alpha \rfloor,$$

where $\alpha = (\sqrt{5} - 1)/2$.

One may generalize the concept of fixed points of homomorphisms by considering fixed points of finite-state transducers. The most famous example of this type is the Kolakoski word [59]

$$\mathbf{k} = 12211212212211211221211212 2\cdots$$

which is a fixed point of the transducer in Figure 7.1.

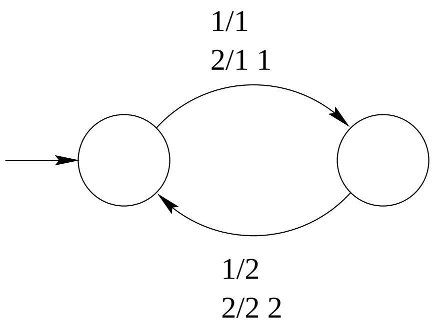

FIG. 7.1. *The Kolakoski transducer*

Despite much work on this sequence (e.g., [54,31,32,102,52,29,28] and [79,20,66,25,21,33,96]), the following conjecture is still open:

CONJECTURE 6. *The limiting frequencies of* 1 *and* 2 *in* $\mathbf{k}$ *exist, and are equal to* $\frac{1}{2}$.

8. Automaticity. In Section 5 we discussed languages that are accepted by finite automata and sequences that are generated by finite automata. However, "most" languages and sequences are not of this type. For the rest of these languages and sequences, can we somehow evaluate how "close" these objects are to being regular or automatic?

In this section, we introduce a measure of descriptional complexity called *automaticity.* Our complexity measure is a *function*, and is designed so that regular languages have $O(1)$ automaticity, and languages "close" to regular have "small" automaticity.

Let

$$\Sigma^{\leq n} = \epsilon + \Sigma + \Sigma^2 + \cdots + \Sigma^n,$$

the set of all strings in Σ^* of length $\le n$. We say a language $L \subseteq \Sigma^*$ is an *nth order approximation* to a language L' if $L \cap \Sigma^{\le n} = L' \cap \Sigma^{\le n}$. Let DFA be the class of all deterministic finite automata over a finite alphabet Σ. We can now informally define the automaticity of a language L to be the function which counts the number of states in the smallest DFA that accepts some nth order approximation to L. Formally, if $|M|$ is defined to be the number of states in the DFA M, we define the automaticity $A_L(n)$ of a language L as follows:

$$A_L(n) = \min\{|M| \ : \ M \in \text{DFA} \ \text{ and } L(M) \cap \Sigma^{\le n} = L \cap \Sigma^{\le n}\}.$$

The following basic properties of the function $A_L(n)$ are easy to prove:

1. $A_L(n) \le A_L(n+1)$.
2. L is regular iff $A_L(n) = O(1)$.
3. $A_L(n) = A_{\overline{L}}(n)$.
4. $A_L(n) \le 2 + \Sigma_{w \in L \cap \Sigma^{\le n}} |w|$.

We now make the following

DEFINITION 8.1. *Two strings w, w' are called n–*dissimilar for *L if there exists a string v with $|wv|, |w'v| \le n$ and either*

(i) $wv \in L$, $w'v \notin L$; or

(ii) $wv \notin L$, $w'v \in L$.

Then we have [36,50,94]:

THEOREM 8.1. *$A_L(n)$ = the maximum number of distinct pairwise n–dissimilar strings for L.*

As an example, consider the language

$$L = \{0^n 1^n \ : \ n \ge 0\}.$$

This language is clearly not regular. What is its automaticity?

It can be shown that the automaticity of L is $A_L(n) = 2\lfloor n/2 \rfloor + 1$ for $n \ge 2$. To see the upper bound, note that we can accept an nth order approximation to L (for $n = 9$) with DFA in Figure 8.1.

To get the lower bound for $n = 9$, note that we may take

$$\{\epsilon, \ 0, \ 00, \ 000, \ 0000, \ 1, \ 01, \ 001, \ 0001\}$$

as our set of n-dissimilar strings. This easily generalizes to larger n.

Now, let's turn to another example. Consider the set

$$P = \{10, 11, 101, 111, 1011, 1101, 10001, 10011, \ldots\},$$

the set of primes expressed in base 2. A classical (1966) theorem due to Minsky and Papert [74] shows that P is not a regular language. However, this raises the question, how "far" from regular is P? We have the following theorem [92]:

THEOREM 8.2. *The automaticity of P^R is $\Omega(2^{n/43})$.*

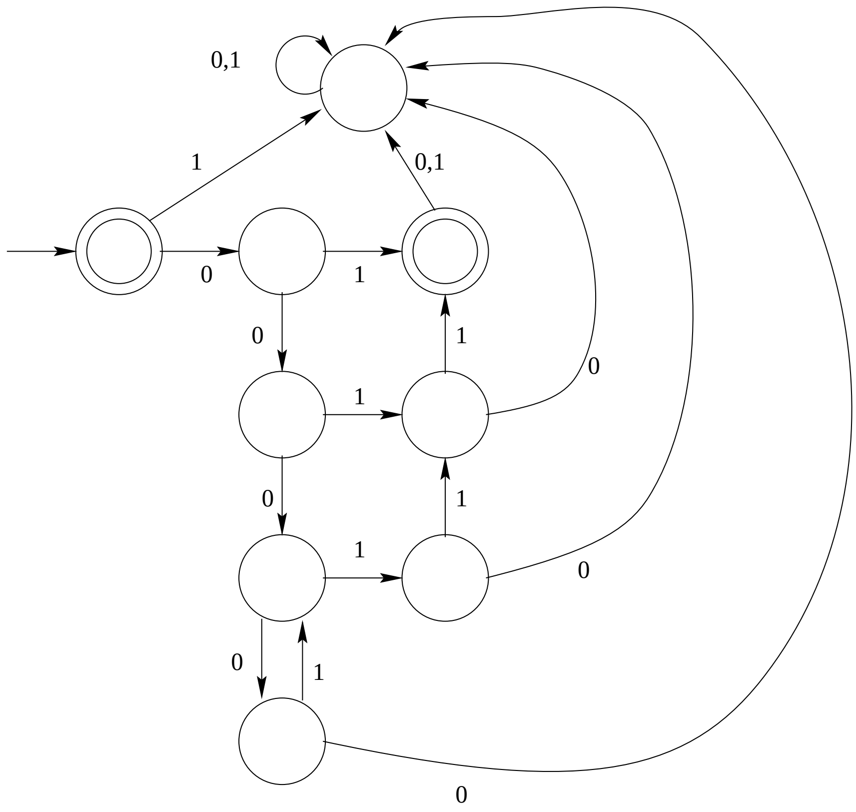

FIG. 8.1. *Automaton acccepting 9th order approximation to L*

(Here P^R denotes the reversal of the set P, i.e., the primes expressed with least significant digit first.)

The basic idea is to prove the following

LEMMA 8.1. *Given integers r, a, b with $r \geq 2$, $1 \leq a, b < r$ with $\gcd(r,a) = \gcd(r,b) = 1$, and $a \neq b$, there exists $m = O(r^{165/4})$ such that $rm + a$ is prime and $rm + b$ is composite.*

The proof of this lemma is an easy consequence of a deep theorem of Heath-Brown [47] on the distribution of primes in arithmetic progressions ("Linnik's Theorem").

Taking $r = 2^n$, the lemma implies that there are at least $2^{n/43}$ n-dissimilar strings for the language P^R.

Automaticity has been examined by Trakhtenbrot [100]; Grinberg & Korshunov [45]; Karp [51]; Breitbart [16,17,18]; Dwork and Stockmeyer [36]; Kaneps & Freivalds [50]; Shallit & Breitbart [93,94], Pomerance, Robson, & Shallit [80], Glaister & Shallit [42], and Shallit [92]. Koskas and de Mathan (work in progress, 1996) show how to apply automaticity to obtain irrationality measures in finite characteristic.

One of the nicest results in the area is Karp's theorem [51]:

THEOREM 8.3. *Let $L \subseteq \Sigma^*$ be a nonregular language. Then*

$$A_L(n) \geq (n+3)/2$$

for infinitely many n.

It can be shown that the constants 3 and 2 in Karp's theorem are best possible, in the sense that the theorem would be false if 2 were replaced with any smaller number, or if 3 were replaced with any larger number [94].

The case of unary alphabets has only recently begun to be studied. In this case, we have $A_L(n) \leq n+1$, for all L and for all n. The following theorems can be proved [80]:

THEOREM 8.4. *Let $L \subseteq 0^*$. Then*

$$A_L(n) \leq n + 1 - \lfloor \log_2 n \rfloor$$

for infinitely many n.

THEOREM 8.5. *Let $L \subseteq 0^*$. Then for "almost all" L we have*

$$A_L(n) > n - 2\log_2 n - 2\log_2 \log_2 n$$

for all sufficiently large n.

Recall that Karp proved that if L is not regular, then $A_L(n) \geq (n+3)/2$ infinitely often. This implies that

$$\limsup_{n\to\infty} \frac{A_L(n)}{n} \geq \frac{1}{2}$$

for all nonregular L. However, it seems that one can do better in the unary case. In 1994, I made the following conjecture [93,80]:

CONJECTURE 7. *There exists a real number $\gamma > 1/2$ such that if $L \subseteq 0^*$ is not regular, then*

$$\limsup_{n\to\infty} \frac{A_L(n)}{n} \geq \gamma.$$

In fact, I had conjectured that $\gamma = (\sqrt{5}-1)/2 \doteq .61803$. However, recently J. Cassaigne has shown that the proper constant is

$$\gamma = (60 - 2\sqrt{10})/89 \doteq .60309$$

and this constant is best possible [22]. (Partial results had previously been obtained by Allouche and Bousquet-Mélou [4].)

Finally, it is known that the maximum possible automaticity for a language $L \subseteq (0+1)^*$ is $O(2^n/n)$. An example of a context-free language (CFL) with automaticity $\Omega(2^n/n)$ is not known, although there are examples with automaticity $\Omega(2^{n(1-\epsilon)})$ for all $\epsilon > 0$ [42]. This suggests the following open problem:

OPEN PROBLEM 8. *Develop an efficient algorithm for computing the automaticity of a CFL, given its representation as a context-free grammar.*

8.1. Nondeterministic Automaticity. Let NFA be the class of all nondeterministic finite automata.

A *nondeterministic finite automaton (NFA)* is like a deterministic one, except now there can be $0, 1, 2$, or more arrows with the same label leaving any state. A string w is accepted by an NFA if there exists some path labeled w from the initial state to some final state.

The function $N_L(n)$ is the *nondeterministic automaticity* of the language L, where

$$N_L(n) = \min\{|M| \ : \ M \in \text{NFA} \ \text{ and } L(M) \ \cap \ \Sigma^{\leq n} = L \ \cap \ \Sigma^{\leq n}\}.$$

Then by the classical subset construction, we have

THEOREM 8.6. *Suppose $L \subseteq \Sigma^*$. If L is not regular, then $N_L(n) \geq \log_2((n+3)/2)$ for infinitely many n.*

This lower bound is best possible, up to a constant, since the Stearns-Hartmanis-Lewis language

$$\{2\ (1)_2^R\ 2\ (2)_2^R\ 2\ (3)_2^R\ 2\ (4)_2^R\ 2 \cdots\ 2\ (n)_2^R : n \geq 1\}$$

has nondeterministic automaticity $O(\log n)$. Here, as in Section 3, $(k)_2$ is the representation of k in base 2, and w^R denotes the reversal of the string w.

We can use some classical estimates from number theory to produce an example of a language with low nondeterministic automaticity [94]:

THEOREM 8.7. *Define*

$$L = \{w \in (0+1)^* \ : \ |w|_0 \neq |w|_1\}.$$

Then L is nonregular and

$$N_L(n) = O((\log n)^2/(\log\log n)).$$

Proof. We need the following fact from number theory:

LEMMA 8.2. *Let $n \geq 2$ and suppose $0 \leq i, j < n$. Then $i \neq j$ iff there exists a prime $p \leq 4.4 \log n$ such that $i \not\equiv j \pmod p$.*

Thus, to nondeterministically accept some nth order approximation to L, we can

- guess the correct prime $p \leq 4.4 \log n$;
- verify that $|w|_0 \not\equiv |w|_1 \pmod p$.

This construction uses at most

$$1 + \sum_{p \leq 4.4 \log n} p = O((\log n)^2/(\log\log n))$$

states. The construction is illustrated in Figure 8.2.

We now turn to the question of lower bounds for nondeterministic automaticity in the unary case [80]:

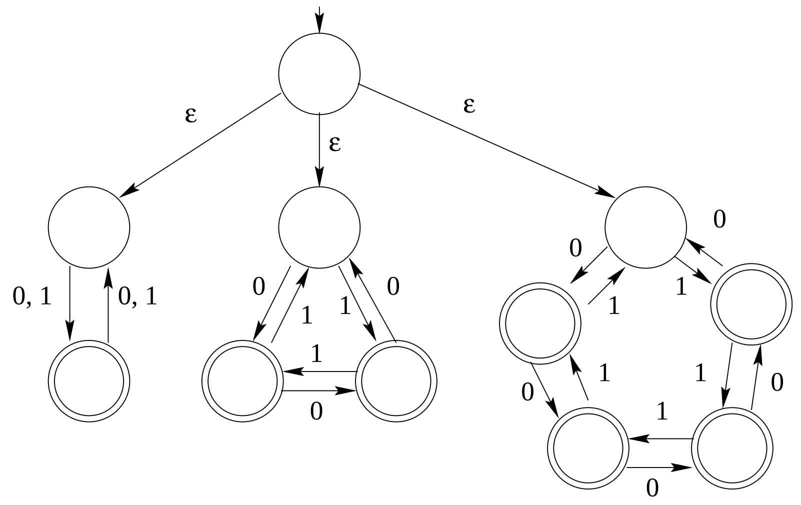

FIG. 8.2. *30th order approximation to L*

THEOREM 8.8. *There exists a constant c (which does not depend on L) such that if $L \subseteq 0^*$ is not regular, then*

$$N_L(n) \geq c(\log n)^2/(\log\log n)$$

infinitely often.

Pomerance has shown [80] that for all monotonically increasing functions f, there exists a language $L = L(f)$ such that

$$N_L(n) = O(f(n)(\log n)^2/(\log\log n)),$$

thus showing the lower bound is essentially tight. To give the flavor of his construction, we prove the following weaker result:

THEOREM 8.9. *Define $L = \{0^n \ : \ n \geq 1$ and the least positive integer not dividing n is not a power of $2\}$. Then L is nonregular and*

$$N_L(n) = O((\log n)^3/(\log\log n)).$$

Proof. The construction depends on the following two facts:

LEMMA 8.3. *If $0^n \in L$, then there exists a prime power p^k, $p \geq 3$, $k \geq 1$, $p^k \leq 5\log n$, such that $n \not\equiv 0 \pmod{p^k}$, and $n \equiv 0 \pmod{2^s}$, with $2^s < p^k < 2^{s+1}$. Further, if such a prime power p^k exists, then $0^n \in L$.*

An NFA accepting an n-th order approximation to L can now be constructed as follows:

- guess the correct odd prime power $p^k \leq 5\log n$;
- verify that, on input 0^r, we have
    - $r \not\equiv 0 \pmod{p^k}$;

$* \ r \equiv 0 \pmod{2^s}$, with $2^s < p^k < 2^{s+1}$.

This construction uses at most $O((\log n)^3/(\log\log n))$ states.

OPEN QUESTION 9. *What is a good lower bound on the nondeterministic automaticity of the set P^R, the (reversed) representations of primes in base 2?*

9. k-regular sequences. The last topic I wish to consider in this survey is k-regular sequences. These are generalizations of the automatic sequences mentioned above in Section 5.

While there are many examples of automatic sequences in number theory, their expressive power is somewhat limited because of the requirement that they take only a finite number of values. How can this be generalized? As we have seen above in Section 5, a sequence is k-automatic iff its k-kernel is finite. This suggests studying the class of sequences where the $\mathbb{Z}$-module generated by the k-kernel is *finitely generated.* We call such a sequence k-regular. The properties of such sequences and many examples were given in [6].

Here are some examples of k-regular sequences in number theory.

Example 1. The 3-adic valuation of a sum of binomial coefficients. Let $r(n) := \sum_{0\le i<n} \binom{2i}{i}$. Then $\nu_3(r(n))$ is 3-regular, as it can be shown that

$$\nu_3(r(n)) = \nu_3\left(n^2\binom{2n}{n}\right); \tag{9.1}$$

see [97]. In fact, Eq. (9.1) was first conjectured by applying a program which attempts to deduce the k-regularity of a given sequence. Zagier [106] found a beautiful proof based on 3-adic analysis.

Example 2. Propp's sequence. Jim Propp [81] introduced the sequence $(s(n))_{n\ge 0}$, defined to be the unique monotone sequence such that $s(s(n)) = 3n$. The table below gives the first few terms:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$s(n)$	0	2	3	6	7	8	9	12	15	18	19	20	21	22

It is sequence M0747 in the book of Sloane and Plouffe [95]. Patruno [78] showed that

$$s(n) = \begin{cases} n + 3^k, & \text{if } 3^k \le n < 2\cdot 3^k; \\ 3(n-3^k), & \text{if } 2\cdot 3^k \le n < 3^{k+1}. \end{cases}$$

This sequence is 3-regular, and satisfies the recurrence

$$\begin{aligned} s(3n) &= 3s(n); \\ s(9n+1) &= 6s(n) + s(3n+1); \end{aligned}$$

$$\begin{aligned} s(9n+2) &= 6s(n) + s(3n+2); \\ s(9n+4) &= 2s(3n+1) + s(3n+2); \\ s(9n+5) &= s(3n+1) + s(3n+2); \\ s(9n+7) &= -6s(n) + 3s(3n+1) + 2s(3n+2); \\ s(9n+8) &= -12s(n) + 6s(3n+1) + s(3n+2). \end{aligned}$$

Example 3. A greedy partition of the natural numbers into sets avoiding arithmetic progressions. Suppose we consider the integers $0, 1, 2, \ldots$ in turn, and place each new integer i into the set of lowest index S_k $(k \geq 0)$ so that S_k never contains three integers in arithmetic progression. For example, we put 0 and 1 in S_0, but placing 2 in S_0 would create an arithmetic progression of size 3 (namely, $\{0, 1, 2\}$), so we put 2 in S_1, etc.

Now define the sequence $(a_k)_{k \geq 0}$ as follows: $a_k = n$ if k is placed into set S_n. Here are the first few terms of this sequence:

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
a_k	0	0	1	0	0	1	1	2	2	0	0	1	0	0

This is Sloane and Plouffe's sequence M0185.

Gerver, Propp, and Simpson [41] showed that $a_{3k+r} = \lfloor (3a_k + r)/2 \rfloor$ for $k \geq 0$, $0 \leq r < 3$. It follows that $(a_k)_{k \geq 0}$ is 3-regular.

We now give some open problems on k-regular sequences.

CONJECTURE 10. *Suppose $(A(n))_{n \geq 0}$ and $(B(n))_{n \geq 0}$ are k-regular sequences with $B(n) \neq 0$ for all n. If $A(n)/B(n)$ is always an integer, then $(A(n)/B(n))_{n \geq 0}$ is also k-regular.*

OPEN QUESTION 11. *Show that $(\lfloor \frac{1}{2} + \log_2 n \rfloor)_{n \geq 0}$ is not a 2-regular sequence.*

We may also consider an extension of k-regular sequences to other types of representation; e.g., Fibonacci representation. Let us consider, for example, the problem of determining the number of partitions k_n of a number n as a sum of distinct Fibonacci numbers [55,19,86]. In other words, we are interested in the coefficient k_n of X^n in the infinite product

$$(1+X)(1+X^2)(1+X^3)(1+X^5)(1+X^8)(1+X^{13}) \cdots .$$

Here are the first few terms of this sequence:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
k_n	1	1	1	2	1	2	2	1	3	2	2	3	1	3

Then it is not hard to see that

$$k_n = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \cdot M_{w^R} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{9.2}$$

where w is the Fibonacci expansion of n, and

$$M_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 1 \end{bmatrix}; \quad M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}. \tag{9.3}$$

In particular, this allows computation of k_n in time polynomial in $\log n$, and gives a simple proof of Theorem 1 of [86].

10. Conclusions. Both number theory and formal language theory have a large body of research associated with them. At their intersection, however, is a new and growing area which promises to enrich them both.

11. Acknowledgments. Jean-Paul Allouche read a draft of this survey and made many helpful comments. I also express my gratitude to the referee, who read this survey with care and corrected several errors.

I would like to express my appreciation to the Institute for Mathematics and its Applications at the University of Minneapolis, where he spent a very pleasant and productive week in the summer of 1996 as an invited speaker at the conference entitled Emerging Applications of Number Theory.

REFERENCES

[1] J.-P. Allouche. Automates finis en théorie des nombres. *Exposition. Math.* **5** (1987), 239–266.

[2] J.-P. Allouche. Sur la transcendance de la série formelle π. *Séminaire de Théorie des Nombres de Bordeaux* **2** (1990), 103–117.

[3] J.-P. Allouche. Transcendence of the Carlitz-Goss Gamma function at rational arguments. *J. Number Theory* **60** (1996), 318–328.

[4] J.-P. Allouche and M. Bousquet-Mélou. On the conjectures of Rauzy and Shallit for infinite words. *Comment. Math. Univ. Carolinae* **36** (1995), 705–711.

[5] J.-P. Allouche and H. Cohen. Dirichlet series and curious infinite products. *Bull. Lond. Math. Soc.* **17** (1985), 531–538.

[6] J.-P. Allouche and J. O. Shallit. The ring of k–regular sequences. *Theoret. Comput. Sci.* **98** (1992), 163–187.

[7] P. Arnoux. Some remarks about Fibonacci multiplication. *Appl. Math. Letters* **2** (1989), 319–320.

[8] E. Bach and J. Shallit. *Algorithmic Number Theory.* MIT Press, 1996.

[9] J. Berstel. Mots de Fibonacci. In *Séminaire d'Informatique Théorique*, pages 57–78. Laboratoire Informatique Théorique, Institut Henri Poincaré, 1980/81.

[10] J. Berstel. Fibonacci words—a survey. In G. Rozenberg and A. Salomaa, editors, *The Book of L*, pages 13–27. Springer-Verlag, 1986.

[11] V. Berthé. De nouvelles preuves "automatiques" de transcendance pour la fonction zêta de Carlitz. In D. F. Coray and Y.-F. S. Pétermann, editors, *Journées Arithmétiques de Genève*, Vol. 209 of *Astérisque*, pages 159–168, 1992.

[12] V. Berthé. Fonction ζ de Carlitz et automates. *J. Théorie Nombres Bordeaux* **5** (1993), 53–77.

[13] V. Berthé. Automates et valeurs de transcendance du logarithme de Carlitz. *Acta Arith.* **66** (1994), 369–390.

[14] V. Berthé. Combinaisons linéaires de $\zeta(s)/\pi^s$ sur $\mathbb{F}_q(x)$, pour $1 \le s \le q-2$. *J. Number Theory* **53** (1995), 272–299.

[15] H. Blumberg. Note on a theorem of Kempner concerning transcendental numbers. *Bull. Amer. Math. Soc.* **32** (1926), 351–356.

[16] Y. Breitbart. Realization of boolean functions by finite automata. *NTL (Novosti Technicheskoi Literature), Seria Automatica, Telemechanika i Priborostroenie* No. 4 (1970).

[17] Y. Breitbart. On automaton and "zone" complexity of the predicate "to be a kth power of an integer". *Dokl. Akad. Nauk SSSR* **196** (1971), 16–19. In Russian. English translation in *Soviet Math. Dokl.* **12** (1971), 10–14.

[18] Y. Breitbart. *Complexity of the calculation of predicates by finite automata.* PhD thesis, Technion, Haifa, Israel, June 1973.

[19] L. Carlitz. Fibonacci representations. *Fibonacci Quart.* **6** (1968), 193–220.

[20] A. Carpi. Repetitions in the Kolakovski [sic] sequence. *Bull. European Assoc. Theor. Comput. Sci.* (50) (1993), 194–196.

[21] A. Carpi. On repeated factors in C^∞-words. *Inform. Process. Lett.* **52** (1994), 289–294.

[22] J. Cassaigne. On a conjecture of J. Shallit. In *Proc. 24th Int'l. Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 1256 of *Lecture Notes in Computer Science*, pages 693–704. Springer-Verlag, 1997.

[23] G. Christol. Ensembles presque périodiques k-reconnaissables. *Theoret. Comput. Sci.* **9** (1979), 141–145.

[24] G. Christol, T. Kamae, M. Mendès France, and G. Rauzy. Suites algébriques, automates et substitutions. *Bull. Soc. Math. France* **108** (1980), 401–419.

[25] V. Chvátal. Notes on the Kolakoski sequence. Technical Report 93-84, DIMACS, March 1994. Revised.

[26] A. Cobham. A proof of transcendence based on functional equations. Technical Report RC-2041, IBM Yorktown Heights, March 25 1968.

[27] A. Cobham. Uniform tag sequences. *Math. Systems Theory* **6** (1972), 164–192.

[28] K. Culik II and J. Karhumäki. Iterative devices generating infinite words. In *STACS 92, Proc. 9th Symp. Theoretical Aspects of Comp. Sci.*, Vol. 577 of *Lecture Notes in Computer Science*, pages 531–543. Springer-Verlag, 1992.

[29] K. Culik II, J. Karhumäki, and A. Lepistö. Alternating iteration of morphisms and the Kolakovski [sic] sequence. In G. Rozenberg and A. Salomaa, editors, *Lindenmayer Systems*, pages 93–103. Springer-Verlag, 1992.

[30] F. M. Dekking. Transcendance du nombre de Thue-Morse. *C. R. Acad. Sci. Paris* **285** (1977), 157–160.

[31] F. M. Dekking. Regularity and irregularity of sequences generated by automata. In *Séminaire de Théorie des Nombres de Bordeaux*, pages 9.01–9.10, 1979–1980.

[32] F. M. Dekking. On the structure of self-generating sequences. In *Séminaire de Théorie des Nombres de Bordeaux*, pages 31.01–31.06, 1980–1981.

[33] F. M. Dekking. What is the long range order in the Kolakoski sequence? Technical report, Faculty of Technical Mathematics and Informatics, Delft University of Technology, 1995.

[34] F. M. Dekking, M. Mendès France, and A. J. van der Poorten. Folds! *Math. Intelligencer* **4** (1982), 130–138, 173–181, 190–195.

[35] U. Dudley. Smith numbers. *Math. Mag.* **67** (1994), 62–65.

[36] C. Dwork and L. Stockmeyer. On the power of 2-way probabilistic finite state automata. In *Proc. 30th Ann. Symp. Found. Comput. Sci.*, pages 480–485. IEEE Press, 1989.

[37] S. Eilenberg. *Automata, Languages, and Machines*, Vol. A. Academic Press, 1974.

[38] P. Enflo, A. Granville, J. Shallit, and S. Yu. On sparse languages L such that $LL = \Sigma^*$. *Disc. Appl. Math.* **52** (1994), 275–285.

[39] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *J. Comput. System Sci.* **31** (1985), 182–209.

[40] R. R. Forslund. A logical alternative to the existing positional number system. *Southwest J. Pure Appl. Math.* **1** (1995), 27–29.

[41] J. Gerver, J. Propp, and J. Simpson. Greedily partitioning the natural numbers

into sets free of arithmetic progressions. *Proc. Amer. Math. Soc.* **102** (1988), 765–772.

[42] I. Glaister and J. Shallit. Automaticity III: Polynomial automaticity, context-free languages, and fixed points of morphisms. To appear, *Computational Complexity*, 1996.

[43] P. J. Grabner, A. Pethö, R. F. Tichy, and G. J. Woeginger. Associativity of reccurrence multiplication. *Appl. Math. Letters* **7**(4) (1994), 85–90.

[44] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, 1989.

[45] V. S. Grinberg and A. D. Korshunov. Asymptotic behavior of the maximum of the weight of a finite tree. *Problemy Peredachi Informatsii* **2** (1966), 96–99. In Russian. English translation in *Problems of Information Transmission* **2** (1966), 75–78.

[46] G. H. Hardy. *A Mathematician's Apology.* Cambridge University Press, 1967.

[47] D. R. Heath-Brown. Zero-free regions for Dirichlet L-functions and the least prime in an arithmetic progression. *Proc. Lond. Math. Soc.* **64** (1992), 265–338.

[48] M. Hollander. Greedy numeration systems and recognizability. Unpublished manuscript, 1995.

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

[50] J. Kaneps and R. Freivalds. Minimal nontrivial space complexity of probabilistic one-way Turing machines. In B. Rovan, editor, *MFCS '90 (Mathematical Foundations of Computer Science)*, Vol. 452 of *Lecture Notes in Computer Science*, pages 355–361. Springer-Verlag, 1990.

[51] R. M. Karp. Some bounds on the storage requirements of sequential machines and Turing machines. *J. Assoc. Comput. Mach.* **14** (1967), 478–489.

[52] M. S. Keane. Ergodic theory and subshifts of finite type. In T. Bedford, M. Keane, and C. Series, editors, *Ergodic Theory, Symbolic Dynamics, and Hyperbolic Spaces*, pages 35–70. Oxford University Press, 1991.

[53] A. J. Kempner. On transcendental numbers. *Trans. Amer. Math. Soc.* **17** (1916), 476–482.

[54] C. Kimberling. Advanced problem 6281. *Amer. Math. Monthly* **86** (1979), 793.

[55] D. Klarner. Partitions of N into distinct Fibonacci numbers. *Fibonacci Quart.* **6** (1968), 235–243.

[56] M. J. Knight. An "ocean of zeros" proof that a certain non-Liouville number is transcendental. *Amer. Math. Monthly* **98** (1991), 947–949.

[57] D. E. Knuth. Fibonacci multiplication. *Appl. Math. Letters* **1** (1988), 57–60.

[58] J. F. Koksma. *Diophantische Approximationen.* Springer, 1936.

[59] W. Kolakoski. Elementary problem 5304. *Amer. Math. Monthly* **72** (1965), 674. Solution in **73** (1966), 681–682.

[60] E. E. Kummer. Über die Ergänzungssätze zu den allgemeinen Reciprocitätsgesetzen. *J. Reine Angew. Math.* **44** (1852), 93–146.

[61] A.-M. Legendre. *Théorie des Nombres.* Firmin Didot Frères, Paris, 1830.

[62] D. H. Lehmer. The Tarry-Escott problem. *Scripta Math.* **13** (1947), 37–41.

[63] S. Lehr. Sums and rational multiples of q–automatic sequences are q–automatic. *Theoret. Comput. Sci.* **108** (1993), 385–391.

[64] S. Lehr, J. Shallit, and J. Tromp. On the vector space of the automatic reals. *Theoret. Comput. Sci.* **163** (1996), 193–210.

[65] C. G. Lekkerkerker. Voorstelling van natuurlijke getallen door een som van getallen van Fibonacci. *Simon Stevin* **29** (1952), 190–195.

[66] A. Lepistö. Repetitions in Kolakoski sequence. In G. Rozenberg and A. Salomaa, editors, *Developments in Language Theory*, pages 130–143. World Scientific, 1994.

[67] N. Loraud. β-shift, systèmes de numération et automates. *J. Théorie Nombres Bordeaux* **7** (1995), 473–498.

[68] J. H. Loxton and A. J. van der Poorten. Algebraic independence properties of

the Fredholm series. *J. Austral. Math. Soc. A* **26** (1978), 31–45.

[69] J. H. Loxton and A. J. van der Poorten. Arithmetic properties of the solutions of a class of functional equations. *J. Reine Angew. Math.* **330** (1982), 159–172.

[70] J. H. Loxton and A. J. van der Poorten. Arithmetic properties of automata: regular sequences. *J. Reine Angew. Math.* **392** (1988), 57–69.

[71] K. Mahler. Arithmetische Eigenschaften der Lösungen einer Klasse von Funktionalgleichungen. *Math. Annalen* **101** (1929), 342–366. Corrigendum, *Math. Annalen* **103** (1930), 532.

[72] K. Mahler. Zur Approximation der Exponentialfunktion und des Logarithmus. I. *J. Reine Angew. Math.* **166** (1931/32), 118–136.

[73] M. Mendès France and J.-Y. Yao. Transcendence and the Carlitz-Goss gamma function. *J. Number Theory* **63** (1997), 396–402.

[74] M. Minsky and S. Papert. Unrecognizable sets of numbers. *J. Assoc. Comput. Mach.* **13** (1966), 281–286.

[75] M. Morse. Recurrent geodesics on a surface of negative curvature. *Trans. Amer. Math. Soc.* **22** (1921), 84–100.

[76] K. Nishioka. *Mahler Functions and Transcendence.* Lecture Notes in Mathematics, Vol. 1631, Springer-Verlag, 1996.

[77] J. Paradís, L. Bibiloni, and P. Viader. On actually computable bijections between $\mathbb{N}$ and $\mathbb{Q}^+$. *Order* **13** (1996), 369–377.

[78] G. Patruno. Solution to problem proposal 474. *Crux Math.* **6** (1980), 198.

[79] G. Păun. How much Thue is Kolakovski? [sic]. *Bull. European Assoc. Theor. Comput. Sci.* (49) (February 1993), 183–185.

[80] C. Pomerance, J. M. Robson, and J. Shallit. Automaticity II: Descriptional complexity in the unary case. *Theoret. Comput. Sci.* **180** (1997), 181–201.

[81] J. Propp. Problem proposal 474. *Crux Math.* **5** (1979), 229.

[82] E. Prouhet. Mémoire sur quelques relations entre les puissances des nombres. *C. R. Acad. Sci. Paris* **33** (1851), 225.

[83] G. N. Raney. On continued fractions and finite automata. *Math. Annalen* **206** (1973), 265–283.

[84] F. Recher. Propriétés de transcendance de séries formelles provenant de l'exponentielle de Carlitz. *C. R. Acad. Sci. Paris* **315** (1992), 245–250.

[85] D. Robbins. Solution to problem E 2692. *Amer. Math. Monthly* **86** (1979), 394–395.

[86] N. Robbins. Fibonacci partitions. *Fibonacci Quart.* **34** (1996), 306–313.

[87] A. Salomaa. *Formal Languages.* Academic Press, 1973.

[88] J. O. Shallit. Simple continued fractions for some irrational numbers. *J. Number Theory* **11** (1979), 209–217.

[89] J. O. Shallit. On infinite products associated with sums of digits. *J. Number Theory* **21** (1985), 128–134.

[90] J. O. Shallit. Some facts about continued fractions that should be better known. Technical Report CS-91-30, Department of Computer Science, University of Waterloo, July 1991.

[91] J. O. Shallit. Numeration systems, linear recurrences, and regular sets. *Inform. Comput.* **113** (1994), 331–347.

[92] J. O. Shallit. Automaticity IV: Sets, sequences, and diversity. *J. Théorie Nombres Bordeaux* **8** (1996), 347–367.

[93] J. Shallit and Y. Breitbart. Automaticity: properties of a measure of descriptional complexity. In P. Enjalbert et al., editor, *STACS '94: 11th Annual Symposium on Theoretical Aspects of Computer Science*, Vol. 775 of *Lecture Notes in Computer Science*, pages 619–630. Springer-Verlag, 1994.

[94] J. Shallit and Y. Breitbart. Automaticity I: Properties of a measure of descriptional complexity. *J. Comput. System Sci.* **53** (1996), 10–25.

[95] N. J. A. Sloane and S. Plouffe. *The Encyclopedia of Integer Sequences.* Academic Press, 1995.

[96] R. Steacy. Structure in the Kolakoski sequence. *Bull. European Assoc. Theor.*

Comput. Sci. (59) (1996), 173–182.

[97] N. Strauss and J. Shallit. Advanced problem 6625. *Amer. Math. Monthly* **97** (1990), 252.

[98] D. S. Thakur. Automata-style proof of Voloch's result on transcendence. *J. Number Theory* **58** (1996), 60–63.

[99] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetaforlaget, Oslo, 1977, pp. 139–158.

[100] B. A. Trakhtenbrot. On an estimate for the weight of a finite tree. *Sibirskii Matematicheskii Zhurnal* **5** (1964), 186–191. In Russian.

[101] K. Wagner and G. Wechsung. *Computational Complexity.* D. Reidel, 1986.

[102] W. D. Weakley. On the number of C^∞-words of each length. *J. Combin. Theory. Ser. A* **51** (1989), 55–62.

[103] D. R. Woods. Elementary problem proposal E 2692. *Amer. Math. Monthly* **85** (1978), 48.

[104] E. M. Wright. Prouhet's 1851 solution of the Tarry-Escott problem of 1910. *Amer. Math. Monthly* **66** (1959), 199–201.

[105] J. Yu. Transcendence and special zeta values in characteristic p. *Ann. Math.* **134** (1991), 1–23.

[106] D. Zagier. Solution to advanced problem 6625. *Amer. Math. Monthly* **99** (1992), 66–69.

[107] E. Zeckendorf. Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas. *Bull. Soc. Royale des Sciences de Liège* **41**(3–4) (1972), 179–182.