JUHA HONKALA

A decision method for the recognizability of sets defined by number systems

Informatique théorique et applications, tome 20, nº 4 (1986), p. 395-403

<http://www.numdam.org/item?id=ITA_1986__20_4_395_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 20, n° 4, 1986, p. 395 à 403)

A DECISION METHOD FOR THE RECOGNIZABILITY OF SETS DEFINED BY NUMBER SYSTEMS (*)

by Juha HONKALA ([1])

Communicated by J. BERSTEL

Abstract. – *We show that it is decidable whether or not a k-recognizable set is recognizable. Consequently, it is decidable whether or not the set defined by a number system is recognizable.*

Résumé. – *Nous montrons qu'il est décidable si un ensemble k-reconnaissable est reconnaissable. En conséquence, il est décidable si l'ensemble défini par un système de numération est reconnaissable.*

1. INTRODUCTION

Recent work in the theory of codes and L codes has increased the importance of the study of arbitrary number systems (*see* [5]). Here "arbitrary" means that the digits may be larger than the base (in our considerations also negative) and that completeness is not required. Many basic facts about number systems were established by Culik and Salomaa, [1]. It was shown in [3] and [4] that the set of bases of the set represented by a number system strongly depends on whether or not the set is recognizable. If the set is not recognizable then the bases form a subfamily of an exponential family. This is not the case if the set is recognizable. It is often possible to determine the bases if it is known whether or not the set is recognizable. For the time being, however, no algorithm is known for determining the bases of the set given by a number system. Below we give an algorithm to decide whether or not the set defined by a number system is recognizable. The algorithm is, in

(*) Received in May 1985, revised in September 1985.
([1]) Mathematics Department, University of Turku, Turku, Finland.

0296-1598 86/04 395 09/$2.90/© Gauthier-Villars

fact, more general. It can be used to decide whether or not a k-recognizable set is recognizable.

The reader is assumed to know the basic facts concerning finite automata and k-recognizable sets (*see* [6] and [2]).

2. PRELIMINARIES

By a *number system* we mean a $(v+1)$-tuple $N=(n, m_1, \ldots, m_v)$ of integers such that $v \geqslant 1$, $n \geqslant 2$ and $m_1 < m_2 < \ldots < m_v$. The number n is referred to as the *base* and the numbers m_i as the *digits.*

A nonempty word

$$m_{i_k} m_{i_{k-1}} \ldots m_{i_1} m_{i_0}, \qquad 1 \leqslant i_j \leqslant v \tag{1}$$

over the alphabet $\{ m_1, \ldots, m_v \}$ is said to *represent* the integer

$$[m_{i_k} \ldots m_{i_0}] = m_{i_k} . n^k + m_{i_{k-1}} . n^{k-1} + \ldots + m_{i_1} . n + m_{i_0}. \tag{2}$$

The word (1) is said to be a *representation* of the integer (2). The set of all integers represented by N is denoted by $S(N)$. We denote by $\operatorname{Pos} S(N)$ the set

$$S(N) \cap \{ 0, 1, 2, \ldots \}$$

and by $\operatorname{Neg} S(N)$ the set

$$S(N) \cap \{ 0, -1, -2, \ldots \}.$$

A set K of integers is said to be *representable by a number system*, RNS for short, if there exists a number system N such that $K = S(N)$. An integer p is called a *base of an RNS set* K if there is a number system with base p representing K.

If $k \geqslant 2$ is an integer, define the mappings λ_k and ν_k from $\{ 0, 1, \ldots, k-1 \}^*$ to the set of nonnegative integers by

$$\lambda_k(w) = \sum_{i=0}^{m} w_i . k^{m-i}$$

and

$$\nu_k(w) = \sum_{i=0}^{m} w_i . k^i$$

where $w = w_0 w_1 \ldots w_m$ and $w_i \in \{0, 1, \ldots, k-1\}$. A subset A of the set of nonnegative integers is *k-recognizable* if there exists a regular language L over the alphabet $\{0, 1, \ldots, k-1\}$ such that $A = \lambda_k(L)$. The following theorem is a generalization of the translation lemma due to Culik and Salomaa, [1]. For a proof see [4].

THEOREM 1: *For every number system* $N = (n, m_1, \ldots, m_v)$ *the sets* Pos $S(N)$ *and* $-$Neg $S(N)$ *are n-recognizable.*

In Theorem 1 we denote $-\text{Neg}\, S(N) = \{x \mid -x \in \text{Neg}\, S(N)\}$.

If A is a subset of the set of nonnegative integers, define the ω-word $\omega(A) = a_0 a_1 a_2 \ldots$ by

$$a_i = \begin{cases} 0 & \text{if } i \notin A, \\ 1 & \text{if } i \in A. \end{cases}$$

If y is a word denote by y^ω the ω-word $yyy\ldots$ The set A is *recognizable* if there exist words y_1 and y_2 such that $\omega(A) = y_1 y_2^\omega$. The ω-word $y_1 y_2^\omega$ is called *the representation of* A if $\omega(A) = y_1 y_2^\omega$ and the following condition is satisfied: if $y_1 y_2^\omega = y_3 y_4^\omega$ for binary words y_3 and y_4 then either $|y_4| = |y_2|$ and $|y_3| \geq |y_1|$, or $|y_4| > |y_2|$. Here $|y|$ stands for the length of y. If $y_1 y_2^\omega$ is the representation of A, then $|y_1|$ is called *the index* of A and $|y_2|$ is called *the period* of A.

In what follows we assume that L is a fixed regular language and n is a fixed positive integer, $n \geq 2$, with the standard form $n = n_1^{v_1} \ldots n_s^{v_s}$ (i. e., each v_i is a positive integer and each n_i is a prime with $1 < n_1 < \ldots < n_s$).

We are going to show that if the index or the period of a recognizable set A is large then if $\mathscr{A}_1$ and $\mathscr{A}_2$ are finite deterministic automata recognizing $\lambda_n^{-1}(A)$ and $\nu_n^{-1}(A)$, respectively, then at least one of them has a great number of states. (In fact, both have. In the proofs below, however, it is more convenient to use first λ_n and then ν_n.) What remains in deciding whether $\lambda_n(L)$ is recognizable is to check a finite number of times whether $\lambda_n(L)$ equals a recognizable set. This can easily be done. By Theorem 1 we can then decide whether Pos $S(N)$ is recognizable for the number system N.

Example: Denote $N_k = (2, 1, k)$ and $S_m(N_k) = \{x \mid x$ has a representation of length m according to $N_k\}$. It is easy to see inductively that

$S_m(N_k)=\{x \mid 2^m-1 \leqq x \leqq k.2^m-k$ and $x \equiv 2^m-1 \pmod{k-1}\}$. Hence $S(N_k)=\bigcup_{m=1}^{\infty}\{x \mid 2^m-1 \leqq x \leqq k.2^m-k$ and $x \equiv 2^m-1 \pmod{k-1}\}$.

For $k=5$ we obtain $S(N_5)=\{1,5\} \cup \{x \mid x \equiv 3 \pmod 4\}$. Hence $S(N_5)$ is recognizable. Because $\omega(S(N_5))=010101(0100)^{\omega}$, the index of $S(N_5)$ is 6 and the period is 4.

For $k=6$ we obtain

$$
\begin{aligned}
S(N_6)=\bigcup_{m=0}^{\infty} [&\{x \mid x \equiv 1 \pmod 5 \text{ and } 2^{1+4m}-1 \leqq x \leqq 6.2^{1+4m}-6\} \\
&\cup \{x \mid x \equiv 3 \pmod 5 \text{ and } 2^{2+4m}-1 \leqq x \leqq 6.2^{2+4m}-6.\} \\
&\cup \{x \mid x \equiv 2 \pmod 5 \text{ and } 2^{3+4m}-1 \leqq x \leqq 6.2^{3+4m}-6\} \\
&\cup \{x \mid x \equiv 0 \pmod 5 \text{ and } 2^{4+4m}-1 \leqq x \leqq 6.2^{4+4m}-6\}].
\end{aligned}
$$

Suppose $S(N_6)$ were recognizable. Because $S(N_6)$ contains no element congruent to 4 modulo 5, the period of $S(N_6)$ should be a multiple of 5. This is, however, impossible because every residue class modulo 5 has arbitrarily long gaps. Thus $S(N_6)$ is not recognizable.

For finite automata we use the notation of [6]. In particular, if the automaton $\mathscr{A}$ moves to q' when reading w in state q, we write $qw \Rightarrow {}^*q'$. We denote the number of states of $\mathscr{A}$ by $\#\mathscr{A}$. If $\mathscr{A}$ and $\mathscr{B}$ are finite automata we denote their product by $\mathscr{A} \times \mathscr{B}$ (*see* [2], p. 17).

If $w=a_0a_1a_2\ldots$ is an ω-word over the alphabet Σ and each a_i is a letter, we denote $w[i,j]=a_i a_{i+1} \ldots a_{i+j-1}$ for nonnegative integers i and j.

3. THE PERIOD CANNOT BE LARGE

We show first that if $\nu_n(L)=A$ for a recognizable set A, then the period of A cannot have a large factor prime to n.

If A is a set of nonnegative integers, define the equivalence relation $\sim_A$ by

$$m_1 \sim_A m_2, \qquad m_1, m_2 \in \mathbb{N}$$

if and only if

$$m_1 n^r + i \in A \quad \Leftrightarrow \quad m_2 n^r + i \in A$$

for all $r \in \mathbb{N}$ and $0 \leqq i < n^r$.

For a proof of the following lemma, see [2], p. 107.

LEMMA 2: *If the set A is recognizable then the number of equivalence classes of $\sim_A$ is finite and equals the number of states in a minimal finite deterministic automaton recognizing $\lambda_n^{-1}(A)$.*

The following lemma is obvious.

LEMMA 3: *Suppose $\omega(A)=a_0 a_1 a_2 \ldots$ If $\omega(A)[m_1 n^r, n^r] \neq \omega(A)[m_2 n^r, n^r]$ where m_1, m_2 and r are nonnegative integers, then m_1 and m_2 are not equivalent modulo $\sim_A$.*

LEMMA 4: *Let A be a recognizable set with the representation $y_1 y_2^\omega = a_0 a_1 a_2 \ldots$ Suppose $|y_2| = c . n_1^{u_1} \ldots n_s^{u_s}$ where $c, u_1, \ldots, u_s$ are nonnegative integers and c is prime to n. Choose k and m such that $n^k \geqq 2|y_2|$ and $m . n^k \leqq |y_1| < (m+1) . n^k$. Denote*

$$\alpha_{m+i} = (y_1 y_2^\omega)[(m+i) . n^k, n^k]$$

for $i \geqq 1$. Then $\alpha_{m+i} \neq \alpha_{m+j}$ if $1 \leqq i < j \leqq c$.

Proof: Assume on the contrary that $1 \leqq i < j \leqq c$ and $\alpha_{m+i} = \alpha_{m+j}$.

Denote $r = |y_2|$ and $y_2 = b_1 b_2 \ldots b_r$. Then there exist binary words β_1, β_2, β_3, β_4 and a positive integer t such that

$$\alpha_{m+i} = \beta_1 b_1 b_2 \ldots b_r \beta_2$$

and

$$\alpha_{m+j} = \beta_3 b_t b_{t+1} \ldots b_r b_1 \ldots b_{t-1} \beta_4$$

and $|\beta_1| = |\beta_3|$. Because $(m+i) n^k \not\equiv (m+j) n^k \pmod{c}$, we obtain $t \neq 1$. Hence the words $b_1 \ldots b_{t-1}$ and $b_t \ldots b_r$ are both nonempty. Because furthermore

$$(b_1 \ldots b_{t-1})(b_t \ldots b_r) = (b_t \ldots b_r)(b_1 \ldots b_{t-1})$$

there exists a nonempty word y such that

$$b_1 b_2 \ldots b_{t-1} = y^p \qquad \text{and} \qquad b_t \ldots b_r = y^q$$

for some positive integers p and q. Hence $y_1 y_2^\omega$ is not the representation of A. This contradiction shows that $\alpha_{m+i} \neq \alpha_{m+j}$. $\square$

LEMMA 5: *Let A, y_1, y_2 and c be as in Lemma 4. Then every finite deterministic automaton recognizing the language $\lambda_n^{-1}(A)$ has at least c states.*

Proof: By Lemmas 3 and 4 $m+i \nsim_A m+j$ if $1 \leqq i < j \leqq c$. The claim follows by Lemma 2. □

Next we show that if $\nu_n(L)=A$ for a recognizable set A, then no high power of any factor of n can divide the period of A.

LEMMA 6: *Let A be a recognizable set. Assume that the period of A is $c.n_1^{u_1}\ldots n_s^{u_s}$ where c is prime to n. Denote*

$$B_i = A \cap \{x \mid x \equiv i \pmod{c}\}$$

for $0 \leqq i < c$. Then there exists an integer i such that the period of B_i is $c.n_1^{t_1}\ldots n_s^{t_s}$ where $\max t_r = \max u_r$.

Proof: If B_i is not empty then c divides the period of B_i.

To avoid notational complications we assume that $B_i \neq \varnothing$ for $0 \leqq i < c$ and that the index of A is 0.

Assume without loss of generality that $\max u_r = u_1$. Let the period of B_i be $c.n_1^{u_{i1}}\ldots n_s^{u_{is}}$. Denote $u = \max u_{i1}$. We show that $u = u_1$.

Assume on the contrary that $u < u_1$. Then for $0 \leqq i < c$ there exist words $w_i = b_{i0} b_{i1}\ldots b_{i,q-1}$ of length $q = n_1^u n_2^{u_2}\ldots n_s^{u_s}$ such that

$$\omega(B_i) = (0^i b_{i0} 0^{c-1} b_{i1} 0^{c-1} b_{i2}\ldots 0^{c-1} b_{i,q-1} 0^{c-1-i})^\omega.$$

Then

$$\omega(A) = (b_{00} b_{10} b_{20}\ldots b_{c-1,0} b_{01} b_{11}\ldots b_{c-1,1}\ldots b_{0,q-1} b_{1,q-1}\ldots b_{c-1,q-1})^\omega$$

which shows that the period of A divides $c.n_1^u n_2^{u_2}\ldots n_s^{u_s}$. This contradiction shows that the assertion is correct. □

LEMMA 7: *Let A, c, u_i and B_i be as in Lemma 6. Choose an integer i such that the period of B_i is $c.n_1^{t_1}\ldots n_s^{t_s}$ with* $\max t_r = \max u_r$. *Then every finite deterministic automaton recognizing the language $\nu_n^{-1}(B_i)$ has at least u states, where* $u = \max u_r$.

Proof: Assume on the contrary that there exists a finite deterministic automaton $\mathscr{B}$ such that $L(\mathscr{B}) = \nu_n^{-1}(B_i)$ and $\#\mathscr{B} \leqq u-1$.

Let w be a word over the alphabet $\{0, 1, \ldots, n-1\}$ such that

(1) $\nu_n(w) \equiv i \pmod{c}$,

(2) $|w| \geqq u-1$,

(3) $w' w_4 \in B_i$ for some word w_4, where $w = w'w''$ and $|w'| = u-1$.

Then there exist words w_1, w_2, w_3, states q_1, q_2, q_3 and a final state q_F such that $w'=w_1 w_2 w_3$, $w_2 \neq \lambda$, $q_0 w_1 \Rightarrow^* q_1$, $q_1 w_2 \Rightarrow^* q_1$, $q_1 w_3 \Rightarrow^* q_2$, $q_2 w_4 \Rightarrow^* q_F$, $q_2 w'' \Rightarrow^* q_3$, where q_0 is the initial state. Choose an integer k such that no prime factor of c divides $n^{|w_2|}-1$ more than k times. Then we obtain (φ stands for Euler's function):

$$\nu_n(w_2^{l\varphi(c^{k+1})}) = \nu_n(w_2)(1+n^{|w_2|}+\ldots+n^{(l\varphi(c^{k+1})-1)|w_2|})$$
$$= \nu_n(w_2) \cdot \frac{n^{l\varphi(c^{k+1})|w_2|}-1}{n^{|w_2|}-1} \equiv 0 \pmod{c}, \qquad l \in \mathbb{N},$$

because by Euler's theorem $n^{\varphi(c^{k+1})} \equiv 1 \pmod{c^{k+1}}$. Let $\bar{w}$ be a word over the alphabet $\{0, 1, \ldots, n-1\}$. Then

$$\nu_n(w_1 w_2^{l\varphi(c^{k+1})+1} \bar{w}) = \nu_n(w_1) + n^{|w_1|} \nu_n(w_2^{l\varphi(c^{k+1})})$$
$$+ n^{|w_1|+l\varphi(c^{k+1})|w_2|} \nu_n(w_2 \bar{w}) \equiv \nu_n(w_1) + n^{|w_1|} \nu_n(w_2 \bar{w})$$
$$= \nu_n(w_1 w_2 \bar{w}) \pmod{c}.$$

Choose l such that $|w_1 w_2^{l\varphi(c^{k+1})+1}|$ exceeds u and the index of B_i.

Because $q_0 w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w_4 \Rightarrow^* q_F$ we have

$$\nu_n(w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w_4) \in B_i.$$

The word $w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w''$ has the same first u letters as the word

$$w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w_4.$$

Furthermore, $\nu_n(w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w'') \equiv \nu_n(w_1 w_2 w_3 w'') = \nu_n(w) \equiv i \pmod{c}$. Hence $\nu_n(w_1 w_2^{l\varphi(c^{k+1})+1} w_3 w'') \in B_i$, which implies that q_3 is a final state. Because $q_0 w_1 w_2 w_3 w'' \Rightarrow^* q_3$, the word $w=w_1 w_2 w_3 w''$ belongs to $L(\mathscr{B})$. This shows that the period of B_i is smaller than $c \cdot n_1^{t_1} \ldots n_s^{t_s}$. This contradiction proves the lemma. □

LEMMA 8: *Let A, c and u_i be as in Lemma 6. Then every finite deterministic automaton recognizing the language $\nu_n^{-1}(A)$ has at least* max u_r/c *states.*

Proof: Let $\nu_n^{-1}(A) = L(\mathscr{A})$ where $\mathscr{A}$ is a finite deterministic automaton. Let $\mathscr{C}_i$ be a finite deterministic automaton, which has c states and which recognizes the language $\nu_n^{-1}(\{x \mid x \equiv i \pmod{c}\})$. Then $\mathscr{A} \times \mathscr{C}_i$ recognizes the language $\nu_n^{-1}(B_i)$. Furthermore, $\mathscr{A} \times \mathscr{C}_i$ has $\# \mathscr{A} \cdot c$ states. By Lemma 7 $\# \mathscr{A} \cdot c \geq \max u_r$. □

4. THE INDEX CANNOT BE LARGE

We still have to prove that if $\nu_n(L)=A$ then the index of A cannot be arbitrarily large.

LEMMA 9: *Let A be a recognizable set. Suppose that the period of A divides $c.n^u$ where c is prime to n, and that m is the index of A. If $m \geqq n^{u+v-2}+1$ for a positive integer v, then any finite deterministic automaton recognizing $\nu_n^{-1}(A)$ has at least v states.*

Proof: Assume on the contrary that there is a finite deterministic automaton $\mathscr{A}$ such that $\nu_n^{-1}(A)=L(\mathscr{A})$ and $\# \mathscr{A} \leqq v-1$.

Let w be the shortest word such that $\nu_n(w)=m-1$. Then $|w| \geqq u+v-1$. Hence there are words w_1, w_2, w_3, w_4 and states q_1, q_2, q_3 such that $w=w_1 w_2 w_3 w_4$, $w_3 \neq \lambda$, $|w_1|=u$, $q_0 w_1 \Rightarrow {}^* q_1$, $q_1 w_2 \Rightarrow {}^* q_2$, $q_2 w_3 \Rightarrow {}^* q_2$, $q_2 w_4 \Rightarrow {}^* q_3$, where q_0 is the initial state. In the same way as in the proof of Lemma 7 we see that

$$\nu_n(w_2 w_3^{\varphi(c^{k+1})+1} w_4) \equiv \nu_n(w_2 w_3 w_4) \pmod{c}.$$

Hence

$$\nu_n(w_1 w_2 w_3^{\varphi(c^{k+1})+1} w_4) \equiv \nu_n(w_1 w_2 w_3 w_4) \pmod{c.n^u}.$$

Because the index of A is m, one of the following two conditions holds:

(1) $m-1 \in A$ and if $x>m-1$ and $x \equiv m-1 \pmod{c.n^u}$ then $x \notin A$.

(2) $m-1 \notin A$ and if $x>m-1$ and $x \equiv m-1 \pmod{c.n^u}$ then $x \in A$.

If (1) holds then q_3 is a final state, which is impossible because

$$q_0 w_1 w_2 w_3^{\varphi(c^{k+1})+1} w_4 \Rightarrow {}^* q_3$$

and $\nu_n(w_1 w_2 w_3^{\varphi(c^{k+1})+1} w_4) \notin A$. If (2) holds then q_3 is not a final state, which is impossible because $q_0 w_1 w_2 w_3^{\varphi(c^{k+1})+1} w_4 \Rightarrow {}^* q_3$ and $\nu_n(w_1 w_2 w_3^{\varphi(c^{k+1})+1} w_4) \in A$. $\square$

5. DECIDABILITY

THEOREM 10: *Let k be a positive integer, $k \geqq 2$. It is decidable whether or not a k-recognizable set is recognizable.*

Proof: Let B be a k-recognizable set. By the definition there exist regular languages L_1 and L_2 such that $B=\lambda_k(L_1)=\nu_k(L_2)$. Thus we can calculate how many states the finite deterministic automata recognizing $\lambda_k^{-1}(B)$ and $\nu_k^{-1}(B)$ have. Consequently, by Lemmas 5, 8 and 9 it suffices to check whether $\lambda_k(L_1)=A$ when the period and the index of the recognizable set A are small. To check whether $\lambda_k(L_1)=A$ for a fixed recognizable set A form a regular language L' over the alphabet $\{0, 1, \ldots, k-1\}$ such that $A=\lambda_k(L')$. This can be done effectively (*see* [2], p. 108). Clearly $\lambda_k(L_1)=\lambda_k(L')$ if and only if

$$0^*((0^*)^{-1}L_1)=0^*((0^*)^{-1}L'),$$

where $(0^*)^{-1}L_1$ stands for $\{w \mid 0^* w \cap L_1 \neq \varnothing\}$. $\square$

Our main theorem now follows by Theorems 1 and 10.

THEOREM 11: *Given a number system N, it is decidable whether or not* $\operatorname{Pos} S(N)$ *is recognizable.*

In Theorem 11, $\operatorname{Pos} S(N)$ can be replaced by $-\operatorname{Neg} S(N)$.

REFERENCES

1. K. CULIK II and A. SALOMAA, *Ambiguity and Decision Problems Concerning Number Systems*, Information and Control, Vol. 56, 1983, pp. 139-153.
2. S. EILENBERG, *Automata, Languages and Machines*, Vol. A, Academic Press, New York, 1974.
3. J. HONKALA, *Bases and Ambiguity of Number Systems*, Theoret. Comput. Sci., Vol. 31, 1984, pp. 61-71.
4. J. HONKALA, *On Number Systems with Negative Digits*, 1984, submitted for publication.
5. H. MAURER, A. SALOMAA and D. WOOD, *L Codes and Number Systems*, Theoret. Comput. Sci., Vol. 22, 1983, pp. 331-346.
6. A. SALOMAA, *Formal Languages*, Academic Press, New York, 1973.