A. N. Livshits

THE DECODING OF SEQUENCIES, GENERATED BY MORPHISMS, CONNECTED WITH PRIMITIVE SUBSTITUTIONS

There is some algorithmic construction, producing words, languages and sequences, which has many applications in different branches of mathematics and other sciences. It is so called morphic generation, which is intensively studied nowadays specifically in the theory of formal languages and in the symbolic dynamics. In both theories the structure of sequences and the questions of unicity of decoding (in some senses) are of importance . Here we shall discuss some notions and results in these topics (among which there are new ones).

If $\Sigma = \{\sigma_i\}$ is a finite alphabet, Σ^* is the free monoid, consisting of all the words over Σ (including λ), and $h : \Sigma^* \to \Sigma^*$ – morphism ($h(uv) = h(u)h(v)$, for any $u, v \in \Sigma^*$), then, beginning with any word $w \in \Sigma^+$, we can form the triple $G = (\Sigma, h, w)$, which is called DOL-system [1] (the pair (Σ, h) being sometimes called the DOL-scheme) and obtain:

1) the sequence of words $S(G) : w, h(w), h^2(w) = h(h(w)), \ldots$

2) the languages $L(G) = \{h^i(w) | i \geqslant 0\}$.

To the morphism h the matrix $\#\Sigma \times \#\Sigma$ corresponds. Its ij-th element in the number of occurrences of σ_j in the word $h(\sigma_i)$. If not specified we shall suppose this matrix to be primitive (some degree has only positive elements).

Let $A = \{a_j\}_{j=u}^{v}$ be arbitrary sequence of symbols (from Σ), $-\infty \leqslant u \leqslant v \leqslant \infty$. If $u \leqslant p \leqslant q \leqslant v$ then we shall denote by $A(p, q)$ the word (finite, one-sided) $a_p \ldots a_q$. If B is the word, then, in order to obtain $B(p, q)$ we have to convert B into a sequence with $u = 1$ for finite and right infinite words or to a sequence with $v = 0$ for left infinite words.

Now we present the well known space of two-sided infinite sequence X_h, connected with morphism. $X_h = \{x = \{x_j\}_{j=-\infty}^{\infty} |$ for every finite $p, q : p \leqslant q$, there exist k, l, r, s such that $x(p, q) = (h^k \sigma_l)(r, s)\}$.

It is clear that X_h coincides with the object , which is denoted in [2], as $X_{\mathcal{A}}$, where $\mathcal{A} = \{ha | a \in \Sigma\}$. X_h is a Cantor set and the shift T_h is its

homeomorphism. The elements of X_h are called admissible sequences and the sequence $A = \{a_j\}_{j=u}^{v}$ is called admissible if there exists $x \in X_h$ such that either $A = x$, or correctly defined and equal are the words $A(u, v)$ and $X(u, v)$. We call the word admissible if its conversion is admissible. Homeomorphism T_h is strictly ergodic and its topological and metric properties depend on properties of morphism. In some cases the combinatorial conditions are obtained for the spectrum of such system to be continuous, purely discrete and mixed ([3], [4], [5], [2] and bibliographies). For combinatorial investigation of morphically generated dynamical systems and of formal languages, generated by the DOL-systems, the coding properties of some collections of words are of great importance. In different circumstances different kinds of properties can be treated as coding ones.

Definition [1]. Nonvoid language C in alphabet Σ is called code if for any words $x_{i_1}, \ldots, x_{i_m}, x_{j_1}, \ldots, x_{j_n}$ such that $x_{i_1} \ldots x_{i_m} = x_{j_1} \ldots x_{j_n}$ the equality $x_{i_1} = x_{j_1}$ takes place.

We say that the code C has the delay p from the left if $''x_{i_1} \ldots x_{i_p}$ is the prefix of $x_{j_1} \ldots x''_{j_n}$ implies $x_{i_1} = x_{j_1}$.

The class of codes among the languages is rather small but it should become greater if we introduce some restrictions to the classes of possible sequences $\{i_1, \ldots, i_m\}, \{j_1, \ldots, j_n\}$ and of possible "messages" $x_{i_1} x_{i_2} \ldots x_{i_m}$. We shall consider the concrete situations of this sort.

In the work [6] the next statement is proved.

Theorem 1. *Let h be a homeomorphism from Σ^* into Σ^* such that for at least one a in Σ $h(a) \neq \lambda$. Let $\#\Sigma = m$ and let w_1 and w_2 be the words over Σ. Then the existence of n such that $h^n(w_1) = h^n(w_2)$ is equivalent to $h^{m-1}(w_1) = y^{m-1}(w_2)$.*

It implies the following: if the language $\{y^{m-1}(\sigma_i)\}_{\sigma_i \in \Sigma}$ is code then every language $\{h^n(\sigma_i)\}_{\sigma_i \in \Sigma}$ for $n > m - 1$ is code . When investigating the dynamical systems of importance are the properties of these languages and of their union which is often far from being code (even with great restrictions, corresponding to the supply of admissible words and with taking into account the natural hierarchical relations which must not be treated as ambiguities-corresponding hierarchical codings of admissible words are connected with substitutional systems of numeration [7], but there are many reasons in investigation of actual ambiguities (in known cases discrete spectrum corresponds to great ambiguities).

In numerous works the notion of local catenativity of DOL systems is studied which provides the example of ambiguity.

DOL-system $G = \{\Sigma, h, w\}$ is called local catenative with cut p and depth n [7], [6] if there exists the vector $\langle i_1, \dots, i_k \rangle$, $n = \max\{i_1, \dots, i_k\}$ such that $h^q w = h^{q-i_1} w \dots h^{q-i_k} w$ for any $q \geqslant p$. Among the l.c. systems the class of standard l.c. systems is distinguished. Let Σ_n denote $\{0, \dots, n-1\}$. Then the DOL-scheme $S = (\Sigma_n, h)$ is called standard if $h(i) = i+1 (i = 1, \dots, n-2), h(n-1) = (n-i_1) \dots (n-i_k)$. The importance of this class is clear because of the following statement proved by Kobuchi.

Theorem 2. *A DOL-system $G = (\Sigma, h, w)$ is $\langle i_1, \dots, i_k \rangle$-locally- catenative if and only if there exists a standard DOL-scheme with $\langle i_1, \dots, i_k \rangle$ l.c. DOL-system $G' = (\Sigma_n, h'0)$ and a λ-free homomorphism $\gamma : \Sigma_n^+ \to \Sigma^+$, such that for any x in $S(G')h(\gamma(x)) = \gamma(h'(x))$.*

In [7] some types of standard l.c. systems are presented (parallel decomposable, cyclic, semi cyclic) for which the word $w \neq 0$ exists such that (Σ_n, h, w) is l.c.

Another property of DOL-system, which is not compatible with good coding properties, is periodicity. It was intensively studied too. The DOL-system is called periodic if there exist nonnegative i and natural p and e such that $h^{p+i}(w) = (h^i w)^e$ [9]. The coding properties of the DOL-languages for periodic DOL-systems, the properties (from the point of view of the theory of formal languages) of the set of all initial words of periodic DOL-systems [9] given DOL-scheme and other questions were studied. For dynamical systems the next situation is of interest : given morphism h there exists finite admissible word (remind: admissibility means the existence of such i, j, k, l that $w = (h^i \sigma_i)(k, l))$ such that system (Σ, h, w) is periodic. Such substitutions (in symbolic dynamic this term is often used and means the same as morphism) are called cyclic.

The concept of "decidability" (possibility to check algorithmically some property of morphism or system) is of significance for the dynamical systems too and some results (for example the sufficient condition [5] of pure discretness of spectrum) can be treated as the concrete algorithm.

Morphically generated sequences and their images (symbol by symbol) under homomorphisms of some kinds are studied as well in the theory of tag-systems [10], of semi groups, of avoidable patterns and so on. Note that in [10] the next generalization of morphism is discussed – the deterministic generalized sequential mapping, iterations of which can be used to simulate the computation of Turing machines.

Now we shall consider the special questions of coding for substitutional

dynamical systems. From the definition of admissible sequence it is easy to deduce that every language $R_l = \{h^l\sigma_i\}_{\sigma_i\in\Sigma}$ can be used to code it, i.e., if $x \in X_h$, then for every l there exists strictly monotonous infinite two-sided sequences of integers $n^l_{r,x}, -\infty < r < \infty$, such that $n^l_{-1,x} + 1 \leqslant 0 \leqslant n^l_{0,x}$ and for every r there exists $i = i^l(r)$, such that $x(n^l_{r,x}+1, n^l_{r+1,x}) = h^l(\sigma_i)$. One can say more: in spite of non-necessity for some of these codings to be unique: we can choose (may be non-unique for some cases) the ierarchical system of codings and functions $i_l(r)$, forming the so-called structurized admissible sequence. Before introducing this notion we shall introduce the auxiliary one-the notion of general structurized sequence, which is defined as a pair $\{\{S^l_x\}^\infty_{l=0}, \{\{n^l_{r,x}\}\}^\infty_{l=0}\} = x^s$, (where for every $l : S^l_x = \{\sigma_{i^l_{r,x}}\}^\infty_{r=-\infty}$)), such that:

1) If $l_1 < l_2$ then $\{n^{l_2}_{r,x}\} \subset \{n^{l_1}_{r,x}\}$; if $-\infty < r_1 < r_2 < \infty$ then $n^l_{r,x} < n^l_{r,x}$ for any $l \geqslant 0$.

2) $n^l_{-1,x} < 0 \leqslant n^l_{0,x}$ for any $l \geqslant 0$.

3)

$$\sigma_{i^0_{n^l_{r}+1,x}} \dots \sigma_{i^0_{n^l_{r+1},x}} = h^l(\sigma_{i^l_{r+1,x}}), \quad l \geqslant 0, \quad -\infty < r < \infty.$$

4) If for some $r\Delta, r'$ according to 1) $n^l_{r',x} = n^{l+1}_{r,x}$, $n^l_{r'+\Delta,x} = n^{l+1}_{r+1,x}$, then $h(\sigma_{i^{l+1}_{r+1,x}}) = \sigma_{i^l_{r'+1,x}} \dots \sigma_{i^l_{r'+\Delta,x}}$.

Of course 4) implies 3). Evidently $n^0_{r,x} = r$ for every r, x^s. We can define the shift $(T^s)^q$ (for any integer q) of x^s in the following way. Let for every $l \geqslant 0$ $H(l,q)$ stand for p, such that $n^l_{p-1,x} + 1 \leqslant q \leqslant n^l_{p,x}$ (for example $H(0,q) = q$). Denote by $\{n^{l,q}_{r,x}\}$ the sequence $\{n_{r+H(l,q),x} - q\}^\infty_{r=-\infty}$ and by $(T^s)^q x^s$ the pair $\{\{T^{H(l,q)}\{S^l_x\}\}^\infty_{l=0}, \{\{n^{l,q}_{r,x}\}^\infty_{l=0}\}\}$, where T is ordinary shift of sequence. It is easy to check 1)–4) for $(T^s)^q x^s$. Notice that by 1)–4) the knowledge of $S^{l_k}, n^{l_k}_{0,x}$ for some $l_k \to \infty$ uniquely determines x^s. Basing on this remark we shall present some class of concrete general structurized sequences. Quadruple (i_0, j_0, m_0, n'_0) is called generating quadruple [2], if the numbers p, q, u, v exist, such that σ_p, σ_q are the j_0-th and $j_0 + 1$-th symbols of $h\sigma_i, h^{m_0}\sigma_p = w\sigma_u, h^{m_0}\sigma_q = \sigma_v w', h^{n'_0}\sigma_u = w_1\sigma_u; h^{n'_0}\sigma_v = \sigma_v w'_1$. If it is the case we define $x_s(i_0, j_0, m_0, n'_0)$, as the general structurized sequence $\{\{S^l_x\}^\infty_{l=0}, \{\{n^l_{r,x}\}\}^\infty_{l=0}\}$, such that all $S^{kn'_0}_x$ are identical

$$S^{kn'_0}_x(-\infty, 0) = h^{n'_0\infty}(\sigma_u); \quad S^{kn'_0}_x(1, \infty) = h^{n'_0\infty}(\sigma_v),), \quad 0 \leqslant k \leqslant \infty$$

and $n'_{0,x} = 0$ for all l, what determines $x^s(i_0, j_0, m_0, n'_0)$. Notice that $S_x^{kn'_0}$ is w_{ab} of [11], where $a = \sigma_u, b = \sigma_v$, and that $S^{kn'_0} \in X_h$, what is easy to check. We say that general structurized sequence is structurized admissible sequence if it satisfies the following additional condition.

5) Three possibilities take place

a) $n^l_{-1,x} \to \infty$, $n^l_{0,x} \to \infty$, when $l \to \infty$,

b) there exist k, L, such that $n^l_{-1,x} = k < 0$ for $l > L$ – in this case for some generating quadruple we have

$$x^s = (T^s)^{-k} x^s(i_0, j_0, m_0, n'_0);$$

c) there exist k, L, such that $n^l_{0,x} = k \geqslant 0$ for $l > L$ – in this case for some generating quadruple we have too

$$x^s = (T^s)^{-k} x^s(i_0, j_0, m_0, n'_0).$$

The space of all structurized admissible sequences with natural metric is a Cantor set. We denote it by X_h^s and the homeomorphism of shift – by T_h^s. It is easy to see that the projection $p : X_h^s \to X_h$, $p(x^s) = S_x^0$ turns to be surjection. For metric and topological theory it is interesting, for which h it is bijection. So the coding properties of some words, connected with h, are to be studied. We present three definitions of such properties, some of which are adapted here for more general situation (in our case only formally). These properties of h are: to be UAD – yield unique admissible decoding [2], to be recognizable [3] (in the recent work [12] the notion of bilateral recognizability is introduced and investigated) and to be determined ([13], [14]).

Let us notice that in spite of coding properties of h the transformation T_h^s is defined correctly (and metrically isomorphic to its analog , appearing as adic representation [2], which is still not always good from the topological point of view).

The word $w \in \Sigma^+$ is called determined if for any admissible words $\sigma_{j_0^1} B'_1 \sigma_{j_1^1} = B_1$; $\sigma_{j_0^2} B'_2 \sigma_{j_1^2} = B_2$ and natural numbers

$$m^l_j, 1 \leqslant l, j \leqslant 2, \quad \text{such that} \quad (hB_l)(m^l_1, m^l_2) = w;$$

$m^l_1 \leqslant |h\sigma_{j_0^l}| < |h(\sigma_{j_0^l} B'_l)| < m^l_2 \leqslant |hB_l|$, $1 \leqslant l \leqslant 2$ the equalities $B'_1 = B'_2$, $|\sigma_{j_0^1}| - m^1_1 = |h\sigma_{j_0^2}| - m^2_1$ take place.

It is possible that $m^1_1 = m^2_1 = 1$, $j^1_0 \neq j^2_0$.

The morphism h is called determined if there exists such N, that every $A \in \Sigma^+$ is determined whenever $|A| > N$.

In [13] there is a definition of determinism to order k, which is the same as the determinism relatively to h^k.

The morphism h is called recognizable if there exists such k, that if for some natural numbers r, s, t and admissible word A the following 3 condition are fulfilled:

1) $|hA| \geqslant \max(r+k, s+k)$,

2) $(hA)(r+1, r+k) = (hA)(s+1, s+k)$,

3) $|h(A(1,t))| = r$,

then there exists $t' \geqslant 0$ such that $|hA(1,t')| = s$. We can compare this definition with one of code with delay p.

Remind that we assume the primitively of morphism's matrix, that means, that such l exists, that for any $a \in \Sigma$ $h^l a$ contains all letters from Σ.

The morphism is called UAD if for any $x \in X_h$ there exists precisely one sequence $y = \{y_j\}_{j=-\infty}^{\infty} \in X_h$, such that there exists sequence $\{n_r\}_{r=-\infty}^{\infty}$, $-\infty < \ldots < n_{-l} < 0 \leqslant n_0 < \ldots < \infty$, for which $x(n_{r-1}+1, n_r) = hy_r$; $-\infty < r < \infty$. Note that if $x = p(x^s)$ for some $x^s \in X_h^s$, then the $y = S_x^1$ and $n_{r,x}^1$ are suitable. UAD is just equivalent to bijectivity of p.

It is easy to see, that determined morphism is UAD, that non cyclic recognizable is UAD, that cyclic morphism is neither UAD nor determined. Still it can be recognizable. The following statement is obvious.

Proposition. *Cyclic morphism is recognizable if there exist the admissible word C, containing all the letters of Σ and not equal to D^v for any D, $v > 1$, and the collection of natural numbers r, $n_1, \ldots, n_s$, $r \leqslant s$ such that if to denote $m_0 = 0$; $m_j = \sum_{l=1}^{j} n_l$, $1 \leqslant j \leqslant s$, then*

1) $m_s = |C|$;

2) $h(c(l,l)) = c(m_{(r+l-1) \mod s} + 1, m_{1+(r+l-1) \mod s})$.

For noncyclic case the question about the decoding was considered by several authors (for example, [3], [14]). There are some definitive results for different kinds of decoding in [12]. Particularly, the assertion of the following theorem 4 can be deduced from the result of [12] or be proved almost identically. Still it admits formulation and some proofs (together with theorems 5,6) in terms of structurized sequences and substitutional systems of numeration. Here are the formulation of the main results about decoding for substitutions.

Theorem 3. *Every noncyclic morphism with primitive matrix is UAD.*

Theorem 4. *Morphism h with p.m. is nonrecognizable if there exists collection of admissible words $M, M_1, M_2, X, Y, D_1, D_2, W_1, W_2$, symbols $e_1, e_2 \in \Sigma$ and natural number T, for which the following holds: the words W_1e_1X and W_2e_2Y are admissible, $h^T(W_1e_1X) = D_1W_1e_1XM$, $h^T(W_2e_2Y) = D_2W_2e_2YM$, $M \neq \lambda$, $M_1he_1 = M_2he_2$; $hX = hY$.*

Theorem 5. *Morphism h with p.m. is not determined if for some collection, consisting of admissible words $M, M_1, M_2, X, X', Y, Y', D_1, D_2, D, W_1, W_2$, symbols $e_1, e_1', e_2, e_2' \in \Sigma$, $e_1 \neq e_2$ and natural number T takes place either*

(i) words $W_1e_1'e_1X$ and $W_2e_2'e_2y$ are admissible,

$$h^T(W_1e_1'e_1X) = D_1W_1e_1'e_1XM; \quad h^T(W_2e_1'e_2Y) = D_2W_2e_2'e_2YM$$
$$M \neq \lambda, \quad h(X) = h(Y), \quad h(e_1'e_1) = X'D,$$
$$h(e_2'e_2) = Y'D, \quad |D| > \min(|he_1|, |he_2|)$$

or (ii) words $Xe_1e_1'W_1$ and $Ye_2e_2'W_2$ are admissible,

$$h^T(Xe_1e_1'W_1) = MXe_1e_1'W_1D_1, \quad h^T(Ye_2e_2'W_2) = MYe_2e_2'W_2D_2,$$
$$M \neq \lambda, \quad hX = hY,$$
$$h(e_1e_1') = DX', \quad h(e_2e_2') = DY', \quad |D| > \min(|he_1|, |he_2|).$$

Here we prove the sufficiency for theorem 4. Let such collection exist. Consider the words $h^{Tl}(W_1e_1x) = h^{T(l-1)}D_1 \dots h^TD_1D_1W_1e_1XM \dots h^{T(l-1)}M$ and $h^{Tl}(W_2e_2Y) = h^{T(l-1)}D_2 \dots D_2W_2e_2YM \dots h^{T(l-1)}M$. Let Ω be the arbitrary admissible word of sort $\dots W_1e_1x \dots W_2e_2Y \dots$ (its existence is evident and connected with recurrence properties). Supposing the condition of recognizability to be fulfilled, choose l such that $|M \dots h^{T(l-1)}| > k$ and take the word $h^{Tl}\Omega$ as A. Consideration of subwords $h(e_1XM \dots h^{T(l-1)}M)$ and $h(e_2YM \dots h^{T(l-1)}M)$ leads to contradiction.

It is easy to check the fulfilling of conditions of the theorems.

Examples.

1) Morphism $h : h(0) = 101110$, $h(1) = 110$ is not recognizable.

2) Morphism $h : h(1) = 332$, $h(2) = 32$, $h(3) = 122$ is recognizable.

3) Morphism $h : h(1) = 413$, $h(2) = 132$, $h(3) = 24$, $h(4) = 4132$ is cyclic nonrecognizable, $h' : h'(1) = 213$, $h'(2) = 4$, $h'(3) = 4, h'(4) = 213$ – cyclic recognizable.

4) Let us consider the morphism h : $h(1) = 123312$, $h(2) = 123123$, $h(3) = 312123$. The listed words form the code from the point of view of definition, but do not from point of view of natural two-sided infinite "extension". If we claim "allowed" the sequences $\{i_k\}_{-\infty}^{\infty}$; $i_k i_{k+1} \neq 13$; $-\infty < k < \infty$, then any concatenation $\ldots h(i_k)h(i_{k+1})\ldots$ with allowed $\ldots i_k i_{k+1}\ldots$ has second representation $\ldots h(i'_k)h(i'_{k+1})\ldots$ with allowed $\ldots i'_k i'_{k+1}\ldots$. Still by theorem 3 our morphism is UAD.

Литература

1. A. Salomaa, *The jewels of formal language theory.* Turku, 1981.
2. A. N. Livshits, *A sufficient condition for weak mixing of substitutions and stationary adic transformations.* — Mat. Notes 44:6. (1988), 920–925.
3. B. Host, *Valeurs propres des systemes dynamiques definies par de longeur variable.* — Ergodic Theory Dynamical Systems, 6 (1986), 529–540.
4. M. Queffelec, *Substitutional dynamical systems, spectral analysis.* — Lecture Notes in. Math. 1294 (1986).
5. A. N. Livshits, *Some examples of adic transformations and automorphisms of substitutions.* — Selecta Mathematica Sovietica, 11, no. 1 (1992), 83–104.
6. A. Ehrenfeucht, G. Rosenberg, *Simplification of Homomorphisms.* — Information and Control, v. 38, no 3, (1978), 298–310.
7. J.-M. Dumont, A. Thomas, *Systemes de numeration et functions fractales relatifs aux substitutions.* — Theoretical Computer Science, 65, no.2 (1989), 153–169.
8. S. Seki, Y. Kobuchi, *On standard locally catenative L-schemes.* — Theoretical Computer Science, 83 (1991),237–248.
9. T. Head, B. Lando, *Regularity of sets of periodic DOL-systems.* — Theoretical Computer Science , 48. (1986), 101–108.
10. J. J. Pansiot, *On various classes of infinite words obtained by iterated morphisms.* — Lecture notes in computer sci. 192 (1985) 188–198.
11. W. H. Gottshalk, *Substitutions minimal sets.* — Trans. Amer. Math. Soc. 109, (1963), 467–491.
12. B. Mossé, *Puissances des mots et reconnaissabilité des points fixes d' une substitution.* — Theoretical Computer Science. 99, vol. 2 (1992), 227–234.
13. J. C. Martin, *Minimal flows arising from substitutions of nonconstant length.* — Math. Syst. Theory , 7, no 1 (1973) 73–82.
14. J. C. Martin, *Substitutional minimal flows.* — Amer. Journ. of Math. 93, no 2 (1971), 503–526.