ON MORPHIC GENERATION OF REGULAR LANGUAGES

T. HARJU
Department of Mathematics, University of Turku, Turku, Finland

J. KARHUMÄKI
Department of Mathematics, University of Turku, Turku, Finland

H.C.M. KLEIJN
Institute of Applied Mathematics and Computer Science, University of Leiden, P.O. Box 9512, The Netherlands

Received 13 November 1984
Revised 24 January 1986

We prove that for any regular language \( L_0 \), \( \text{Reg} \neq \mathcal{H}^{-1}(L_0) \), where \( \text{Reg} \) denotes the family of all regular languages and \( \mathcal{H} \) (resp. \( \mathcal{H}^{-1} \)) denotes the family of all morphisms (resp. inverse morphisms).

1. Introduction

The search for representation results for families of languages is one of old classical themes of formal languages, cf., e.g., [8] or [12]. This topic has revived considerably during the last five or six years, and new, especially morphic, representation results for families of languages have been achieved, cf. [1], [5] and [3].

Particularly interesting is the case when a family of languages is generated from a single language (of this family) via some operations. A well-known theorem of Greibach, cf. [6] or [8], stating that each context-free language is obtainable as an inverse morphic image of a fixed context-free language is a typical example of such results. In [3] a similar result to that of Greibach was proved for the families of recursively enumerable and context-sensitive languages, and moreover, it was noted that no such characterization occurs for the family of regular languages. Indeed, the number of states of a finite automaton needed to recognize an inverse morphic image of a regular language \( L \) is not larger than that needed to recognize the language \( L \).

Actually, in [3] even a stronger negative result than mentioned above was established. In order to state it let \( \text{Reg} \) denote the family of all regular languages and \( \mathcal{H} \) (resp. \( \mathcal{H}^{-1} \)) the family of all morphisms (resp. inverse morphisms). Then for each regular language \( L_0 \), we have \( \text{Reg} \neq \mathcal{H}\mathcal{H}^{-1}(L_0) \), the proof being based on the infinite star height hierarchy of regular languages, cf., e.g., [11]. Concerning the
positive representation results it was proved in [2] that \( \text{Reg} = \mathcal{H}^{-1}(\mathcal{H}^{-1}(a^*b)) \). Later in [9] this was strengthened to the following form: \( \text{Reg} = \mathcal{H}^{-1}(\mathcal{H}(a^*b)) = \mathcal{H}^{-1}(\mathcal{H}^{-1}(b)) \), where, moreover, some of the morphisms may be assumed to be of special forms. Essentially the same result (and many others) were obtained in [13], too.

The question of whether or not the equality \( \text{Reg} = \mathcal{H}^{-1}(\mathcal{H}(L_0)) \) holds true for some regular language \( L_0 \) remained open, although some partial results were achieved in [13]. In this note we settle the problem by showing that the equality does not hold. In fact, we prove even a slightly stronger result, namely, that, for any regular \( L_0 \), the family \( \mathcal{H}^{-1}(\mathcal{H}(L_0)) \) does not contain even all finite languages.

2. Preliminaries

We assume that the reader is familiar with the basic facts of regular languages as well as free monoids, cf., e.g., [12] and [10]. Hence, the following lines are mainly to fix our terminology.

The free monoid (resp. free semigroup) generated by a finite alphabet \( \Sigma \) is denoted by \( \Sigma^* \) (resp. \( \Sigma^+ \)). Elements of \( \Sigma^* \) are called words; \( \lambda \) denotes the empty word. The length of a word \( u \) is denoted by \( |u| \). For two words \( u \) and \( v \) in \( \Sigma^* \) we say that \( u \) is a prefix (suffix, respectively) of \( v \) if \( v = uz \) (\( v = zu \), respectively) for some \( z \) in \( \Sigma^* \), and further that \( u \) is a factor of \( v \) if there exist words \( z \) and \( z' \) such that \( v = zuz' \); \( u \) and \( v \) are conjugate if there exist words \( z \) and \( z' \) such that \( u = zz' \) and \( v = z'z \). The period of a word \( u \), in symbols \( \pi(u) \), is the length of a shortest word \( v \) such that \( u \) is a factor in \( v^m \) for some \( m \geq 1 \). Further we say that \( u \) is primitive if it is not a proper power of any word, i.e., the relation \( u = z^m \) implies that \( z = u \) (and \( m = 1 \)). A total order on \( \Sigma \) can be extended in a natural way to a lexicographic order on \( \Sigma^* \). Having this order we say that a word is a Lyndon word if it is primitive and minimal in its conjugate class.

Let \( A \) be a set of words and \( w \) a word. An \( A \)-interpretation of \( w \) is factorization \( w = x_0x_1\cdots x_rx_{r+1} \), where \( r \geq 0 \), \( x_0 \) is a suffix of a word in \( A \), \( x_{r+1} \) is a prefix of a word in \( A \) and, for \( i = 1, \ldots, r \), \( x_i \) is in \( A \). It is disjoint from another \( A \)-interpretation \( w = y_0y_1\cdots y_sy_{s+1} \) if \( x_0\cdots x_i \neq y_0\cdots y_j \) for all \( i \leq r \) and \( j \leq s \). The \( A \)-degree of \( w \) is the maximum number of pairwise disjoint \( A \)-interpretations of \( w \). The following important result was originally proved in [4], cf. also [10, Theorem 8.3.1].

**Theorem 0.** If the period of a word \( w \) is strictly greater than the periods of the words in \( A \), then the \( A \)-degree of \( w \) is at most \( \text{Card}(A) \). \( \square \)

Finally, let \( \text{Reg} \) denote the family of all regular languages, and \( \mathcal{H} \) (resp. \( \mathcal{H}^{-1} \)) the family of all morphisms (resp. inverse morphisms) between finitely generated free monoids. The family of compositions of morphisms and inverse morphisms — in this order — is denoted by \( \mathcal{H}^{-1}\mathcal{H} \).
3. Results

We start with a reduction result. This lemma was already announced in a slightly different form and without a proof in [13].

**Lemma 1.** For each regular language $L_0$ there exist an alphabet $\theta$ and a symbol $\# \in \theta$ such that for every $L = g^{-1} h(L_0)$ with $h, g \in \mathcal{K}$, $L = g_1^{-1} h_1(\theta^* \# \cup \{\lambda\} \cap h(L_0))$ for some $h_1, g_1 \in \mathcal{K}$, where $h_1$ is $\lambda$-free.

**Proof.** Let $L_0 \subseteq \Sigma^*$ be a regular language recognized by a finite automaton $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$. We define a transition system, cf. [8], $\mathcal{A}' = (Q', \Sigma, \delta', 0, \{N\})$ as follows. The state set of $\mathcal{A}'$ is $Q' = \{0, N\} \cup \overline{Q}$ where $\overline{Q}$ is assumed to be the set $\{1, \ldots, N-1\}$, and the transition set of $\mathcal{A}'$ is

$$\delta' = \delta \cup \{(q, \lambda, q) \mid q = 1, \ldots, N-1\} \cup \{(q, \lambda, N) \mid q \in F\}.$$ 

Clearly, $\mathcal{A}'$ accepts $L_0$.

Let $h : \Sigma^* \rightarrow \Gamma^*$ and $g : A^* \rightarrow \Gamma^*$ be in $\mathcal{K}$. Before constructing $\theta, h_1$ and $g_1$, we provide some intuition.

Let $w = a_1 \cdots a_n \in L$, where $a_i \in \Sigma$ for $1 \leq i \leq n$, and assume that $h(w) \neq \lambda$. Then there exist $(i_1, a_1, i_2), (i_2, a_2, i_3), \ldots, (i_n, a_n, i_{n+1})$ in $\delta$ such that $i_1 = q_0$, $i_{n+1} \in F$, and $i_1, \ldots, i_{n+1} \in \{1, \ldots, N-1\}$. Let $j_1, \ldots, j_m \in \{1, \ldots, n\}$ be such that $m \geq 1$, $w = x_1 a_j x_2 a_j \cdots x_m a_j x_{m+1}$, $h(x_i) = \lambda$, for $1 \leq i \leq m+1$ and $h(a_j) \neq \lambda$, for $1 \leq i \leq m$. In $\mathcal{A}'$ there exist paths

from 0 to $i_{j_1}$ labeled by $x_1$,
from $i_{j_{l+1}}$ to $i_{j_{l+1}}$ labeled by $x_{l+1}$, for $1 \leq l \leq m-1$, and
from $i_{j_{m+1}}$ to $N$ labeled by $x_{m+1}$.

Note that $i_{j_{l+1}} = i_{j_{l+1}}$ if $x_{l+1} = \lambda$, $1 \leq l \leq m-1$. In $\theta$ we will have corresponding symbols $[0, i_{j_1}], [i_{j_{l+1}}, i_{j_{l+1}}], 1 \leq l \leq m-1$, and $[i_{j_{m+1}}, N]$.

Now, we define

$$\theta = \{(i, a, j) \mid i, j \in \{1, \ldots, N-1\}, a \in \Sigma, (i, a, j) \in \delta\}$$

$$\cup \{(i, j) \mid i, j \in \{0, \ldots, N\} \text{ such that } (i, j) \neq (0, N)\}.$$ 

Let $\#, p, \varepsilon$ and $\$ be new symbols. Next we define the morphism

$$h_1 : (\theta \cup \{\#\})^* \rightarrow (\Gamma \cup \{p, \varepsilon, \$\})^*$$

in such a way that $h_1$ simulates $h$ together with information coded in $p, \varepsilon, \$ on paths in $\mathcal{A}'$. (All 'superfluous' symbols in $\theta$ are mapped to $\varepsilon \varepsilon$.)

$$h_1([0, j]) = \$ p^{N-j}, \text{ for } j = 1, \ldots, N-1, \text{ if there exists a path in } \mathcal{A}' \text{ from } 0 \text{ to } j \text{ labeled by } w \text{ such that } h(w) = \lambda,$$
$h_1([i,N]) = p^i \$, for $i = 1, \ldots, N-1$, if there exists a path in $\mathcal{A}'$ from $i$ to $N$ labeled by $w$ such that $h(w) = \lambda$,

$h_1([i,j]) = p^i \$ \$ p^{N-j}$, for $i, j = 1, \ldots, N-1$, if there exists a path in $\mathcal{A}'$ from $i$ to $j$ labeled by $w$ such that $h(w) = \lambda$,

$h_1([i,a,j]) = p^i b_1 \$ \$ p^{N-j} \$ p^{N-j} \$ p^{N} \$ p^{N} \$ b_m p^{N-j}$, for $(i, a, j)$ in $\delta$ with $a \in \Sigma$, if $h(a) = b_1 \ldots b_m$ where $m \geq 1$ and $b_k \in \Gamma$ for $k = 1, \ldots, m$,

$h_1(\#) = \$, $h_1(x) = \$ \$, otherwise.

Finally we define the morphism $g_1 : A^* \rightarrow (\Gamma \cup \{p, \$, \$\})^*$.

$g_1(d) = \$ p^{N} b_1 \$ p^{N} \$ p^{N} \$ p^{N} \$ p^{N} \$ p^{N} \$ \$ p^{N}$, if $g(d) = b_1 \ldots b_m$ where $m \geq 1$ and $b_k \in \Gamma$ for $k = 1, \ldots, m$,

$g_1(d) = \lambda$, if $g(d) = \lambda$.

$g_1$ directly mimicks $g$ interspersed with $p^{N} \$ \$ p^{N}$. Every nonempty word in $g_1(A^*)$ starts with $\$ p^{N}$ and ends with $p^{N} \$ \$. In this way every word $u$ in $h_1(\theta^* \#) \cap g_1(A^*)$ corresponds to a path from $0$ to $N$ in $\mathcal{A}'$ labeled by $w$ and such that $h(w)$ equals $u$ after removing all symbols $p$, $\$ \$ and $\$.

Now the lemma is a straightforward consequence of the construction of $h_1$ and $g_1$.

Our second lemma, which we believe is interesting on its own, gives a solution to problem number 67 in Bulletin nr. 23 of EATCS. We repeat its proof here, cf. [7], for the sake of completeness.

**Lemma 2.** Let $F \subseteq \Sigma^+$ be a finite set and $w, w' \in A^+$ words such that $w'$ is a prefix of $w$. If $F^* \cap w^* w'$ is finite, then it has cardinality at most $2^{Card(F)}$.

**Proof.** Let $n = Card(F)$ and assume that $E = F^* \cap w^* w'$ is finite. Let $w = x^k$, for some primitive $x$ and $k \geq 1$. (We do not assume that $w$ is primitive, e.g., by replacing $w$ by $x$, because the finiteness of $E$ does not imply that $F^* \cap x^* w'$ is finite.) There exists a conjugate of $x$, say $v$, such that $|\pi(v)| = |x|$. This is achieved by taking $v$ equal to the Lyndon word of $x$, cf. [10]. Finally, without loss of generality we may suppose that each $u$ in $F$ does occur as a factor in some power of $w$, and therefore $|\pi(u)| \leq |x| = |v|$ for all $u$ in $F$.

Denote

$$P = \{ u \in F \mid |u| \geq |v| \},$$

$$p = Card(P) \quad \text{and} \quad q = n - p.$$  (1)
Let
\[ w'w' = u_1 \cdots u_r, \]
with \( u_i \) in \( F \),
be a factorization of a word in \( E \). Since \( E \) is finite, we have, for all \( i \) and \( j \geq 0 \),
\[ |u_i \cdots u_{i+j}| \neq 0 \text{ mod } |v|. \]
(3)

Otherwise \( u_i \cdots u_{i+j} (u_{i+j} \cdots u_r)^{-1} u_{i+j+1} \cdots u_r \in E \), for all \( l \geq 0 \).

Since \( v \) is a Lyndon word, every word \( u \) with \( |u| \geq |v| \) has at most one \( \{v\} \)-factorization. Hence, if \( w'w' = u_i \cdots u_{i+j} \cdots u_r \), with \( u_i = u_{i+j} = u \), for some \( u \in P \), \( i, i+j \in \{1, \ldots, r\} \) and \( j \geq 1 \), then \( |u_i \cdots u_{i+j-1}| \equiv 0 \text{ mod } |v| \), which contradicts (3). This implies that any word from \( P \) can occur at most once in a fixed factorization (2).

Assume that \( w'^m w' = u_1 u_2 u_3 u_4 u_5 \in E \) and \( w'^s w' = v_1 u_4 v_2 u_2 v_3 \in E \) for some \( m, s \geq 0 \), \( u_1, u_3, u_5, v_1, v_2, v_3 \in F^* \) and \( u_2, u_4 \in P \). Since \( u_2 \) and \( u_4 \) both have only one \( \{v\} \)-factorization, \( u_1 u_2 u_3 u_4 (u_2 u_3 u_4)^{-1} u_5 \in E \) with \( |v_2 u_2 u_3 u_4| \equiv 0 \text{ mod } v \). This contradicts (3). Hence, if \( u_i \) and \( u_{i+j} \) in a factorization (2) are from \( P \), then \( u_{i+j} \) appears in no factorization (2) before \( u_i \).

By the above observations we can divide the factorizations of the form (2) into different classes such that all factorizations in a fixed class contain exactly the same words from \( P \). Furthermore, the order in which these words appear in factorizations is the same for all classes. It is easy to see that the number of such classes is at most \( 2^p \). Now, we consider such a fixed class. Say all the factorizations (2) in this class contain exactly the words \( w_1, \ldots, w_s \) from \( P \) in this order. For such a factorization we can rewrite (2) as
\[ v_1 v^m v_2 = x_1 w_1 x_2 w_2 \cdots x_s w_s x_{s+1}, \]
(4)
where each \( x_i \) is in \( (F - P)^* \).

Now, by the definition of \( P \), the periods of the words in \( F - P \) are strictly smaller than that of \( v \). Hence, it follows from Theorem 0 that the \((F - P)\)-degree of \( v \) is at most \( q \). Since \( \pi(u) < \pi(v) \), for all \( u \in F - P \), \( v \) is not a factor of any \( u \in F - P \), meaning that every such occurrence of \( v \) in a fixed factorization (4) which does not touch a \( w_i \) corresponds to an \((F - P)\)-interpretation of \( v \).

Let \( u = v_0 v_1 \cdots v_l v_{l+1} \) and \( v = y_0 y_1 \cdots y_j y_{j+1} \), be two \((F - P)\)-interpretations of \( v \), that are not disjoint. Hence \( v_0 \cdots v_l = y_0 \cdots y_j \), for some \( l \leq l \) and \( j \leq j \). Assume that both these \((F - P)\)-interpretations of \( v \) occur as in a factorization (4): \( v_1 v^m v_2 = u_1 \tilde{v}_1 \cdots u_l \tilde{v}_l \tilde{w}_2 \tilde{y}_1 \cdots y_j \tilde{y}_l \tilde{u}_3 \), where \( u_1, u_2, u_3 \in F^* \), \( v_0 \) and \( y_0 \) are suffixes of \( \tilde{v} \) and \( \tilde{y} \), respectively, \( \tilde{v}, \tilde{y}, \tilde{v}, \tilde{y} \in F - P \). Then \( |v_{l+1} \cdots v_l \tilde{v}_2 \tilde{y}_1 \cdots y_j| \equiv 0 \text{ mod } |v| \), which contradicts (3). Hence, all \((F - P)\)-interpretations of \( v \) are disjoint in a fixed interpretation (4). This means that in (4) there occur at most \( q \) words \( v \) which do not touch any \( w_i \) in this factorization, that is, the number of those occurrences of \( v \) which occur as factors in some \( x_i \) is at most \( q \). Each \( w_i \) in (4) has only one \( \{v\} \)-interpretation. In each fixed class of factorizations (4), \( 0, 1, \ldots, q \) \( v \)'s can be contributed by words \( x_i \) from \((F - P)\) apart from the contribution by \( w_1, \ldots, w_s \). Hence there exist at most \( q + 1 \)
different words in $E$ with factorizations in the same class. Consequently, the cardinality of $E$ is at most $2^P(q+1)\leq 2^n$.  

Now, we are ready for

**Theorem 1.** For each regular language $L_0$ we have $\text{Reg} \neq \mathcal{H}^{-1}(\mathcal{H}(L_0))$.

**Proof.** Assume to the contrary that $\text{Reg} = \mathcal{H}^{-1}(\mathcal{H}(L_0))$ for some regular $L_0$. Let $\theta$ and $\#$ be as in the statement of Lemma 1. Let $n = 2^{\text{Card}(\theta)}$. Consider the language $L = \{a, a^2, \ldots, a^{n+1}\}$. By our assumption there exist morphisms $h$ and $g$ such that $L = g^{-1}h(L_0)$. Since $\lambda \in L$ there exist morphisms $h_1$ and $g_1$ such that $L = g_1^{-1}h_1(\theta \#)$. This implies that $(g_1(a))^i \in (h_1(\theta))^* h_1(\#)$ if and only if $i \leq n+1$. This, however, contradicts Lemma 2 with $F = h_1(\theta)$, $w = g_1(a)$ and $w' = Z$, where $g_1(a) = Z h_1(\#)$.  

Actually, we have proved even a stronger result than Theorem 1, namely, that for any regular language $L_0$, the family $\mathcal{H}^{-1}(\mathcal{H}(L_0))$ does not contain all finite languages.

**Acknowledgement**

The authors are indebted to H.J. Hoogeboom for useful discussions on the topic of this paper and to an anonymous referee for a very careful reading of the manuscript.

**References**