NOTE

A HOMOMORPHIC CHARACTERIZATION
OF REGULAR LANGUAGES

Karel CULIK II
Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

Faith E. FICH
Computer Science Division, University of California, Berkeley, CA, USA

Arto SALOMAA
Mathematics Department, University of Turku, Turku, Finland

Received 14 May 1980
Revised 30 June 1981

Every regular language \( R \) (over any alphabet) can be represented in the form \( R = h_2 h_1^{-1} h_3 h_1 (1^* 0) \) where \( h_1, h_2, h_3 \), and \( h_4 \) are homomorphisms. Furthermore, if \( n \) is sufficiently large, then \( R = g_3 g_2^{-1} g_1 (\{1, \ldots, n\}^* 0) \) where \( g_1, g_2, \) and \( g_3 \) are homomorphisms.

1. Introduction

Recently there has been a vivid interest in homomorphic representations of language families. A purely homomorphic characterization of recursively enumerable sets is given in [1] and related questions are investigated in [2] and [3].

The general idea is to try to find a particular language \( L_0 \) such that every language in the family under consideration is obtained from \( L_0 \) by homomorphisms and inverse homomorphisms. Thus, \( L_0 \) can be viewed as a morphic generator or basis for the family in question.

For three of the four language families in the basic Chomsky hierarchy, the families of context-free, context-sensitive, and recursively enumerable languages, a particularly simple morphic characterization exists: there is a language \( L_0 \) in the family such that every language in the family is of the form \( h^{-1}(L_0) \), where \( h \) is a homomorphism. (This statement is true only modulo the empty word \( \lambda \), two generators \( L_0 \) and \( L_0 \cup \{\lambda\} \) have to be used in order to obtain languages with and without \( \lambda \).) The reader is referred to [2], [3], and [5] for further discussion and references.

However, the fourth family in the hierarchy, the family of regular languages, is exceptional: it is shown in [3] that for every regular language \( R \) there exists a regular
language \( R' \) such that
\[
R' \neq g(h^{-1}(R))
\]
for all homomorphisms \( g \) and \( h \). (Here we are dealing with an essential inequality, i.e., the inequality is not only modulo \( \lambda \).) Thus, an inverse morphism followed by a morphism is not sufficient for generation. The purpose of this note is to establish the following rather surprising result, which shows that a sequence of four morphisms is sufficient.

2. The result

**Theorem.** For every alphabet \( \Sigma \) and every regular language \( R \subseteq \Sigma^* \) there are homomorphisms \( h_1, h_2, h_3, \) and \( h_4 \) with the property that \( R = h_4h_3^{-1}h_2h_1^{-1}(1*0) \).

**Proof.** For all unexplained notions in language theory, the reader is referred to [4].

Let \( \Sigma \) be an arbitrary alphabet and let \( R \) be any regular language. The language \( R \setminus \{\lambda\} \) is accepted by a finite nondeterministic automation \( A \) with state set \( Q \), such that \( q_0 \in Q \) is the only initial state, \( q_1 \in Q \) is the only final state and, furthermore, there is no transition to \( q_0 \). Also we may assume, without loss of generality, that \( Q \cap \Sigma = \emptyset \).

We consider the alphabet \( \Sigma_A \) whose letters are triples \((q, a, q') \in Q \times \Sigma \times Q\) such that there is a transition in \( A \) from \( q \) to \( q' \) labelled by \( a \). Let \$ be a new symbol not in \( Q \cup \Sigma \).

Four homomorphisms
\[
\begin{align*}
h_1 : (\Sigma_A \cup \{$\})^* & \to \{0, 1\}^*; \\
h_2 : (\Sigma_A \cup \{$\})^* & \to (\Sigma \cup Q \cup \{$\})^*; \\
h_3 : (\Sigma \cup Q \cup \{$\})^* & \to (\Sigma \cup Q \cup \{$\})^*; \\
h_4 : (\Sigma \cup Q \cup \{$\})^* & \to \Sigma^*
\end{align*}
\]
are now defined by
\[
\begin{align*}
h_1((q, a, q')) & = 1 \quad \text{for all} \ (q, a, q') \in \Sigma_A, \quad h_1(\$) = 0; \\
h_2((q, a, q')) & = qa'q' \quad \text{for all} \ (q, a, q') \in \Sigma_A, \quad h_2(\$) = \$; \\
h_3(a) & = a \quad \text{for all} \ a \in \Sigma, \quad h_3(q_0) = q_0, \\
h_3(q) & = qq \quad \text{for all} \ q \in Q \setminus \{q_0\}, \quad h_3(\$) = q_1\$; \\
h_4(a) & = a \quad \text{for all} \ a \in \Sigma, \quad h_4(q) = \lambda \quad \text{for all} \ q \in Q, \quad h_4(\$) = \lambda.
\end{align*}
\]

Assume first that \( R \) does not contain the empty word, i.e., \( R \setminus \{\lambda\} = R \). Then it is easy to verify the following four facts, which together imply the result.

(1) \( h_1^{-1}(1*0) = \Sigma_A^*\$.}
A homomorphic characterization of regular languages

(2) \( h_2 h_1^{-1}(1*0) \) consists of the string $ and all sequences of the form

\[ q_1 a_1 q'_1 q_2 a_2 q'_2 \cdots q_n a_n q'_n $ \]

where \( n \geq 1 \) and there is a transition in \( A \) from \( q_i \) to \( q'_i \) labelled by \( a_i \) for \( i = 1 \cdots n \).

(3) \( h_3^{-1} h_2 h_1^{-1}(1*0) \) consists of all sequences of the form

\[ q_1 a_1 q_2 a_2 q_3 \cdots q_n a_n $ \]

where \( n \geq 1 \), \( q_1 = q_0 \), and there is a transition in \( A \) from \( q_n \) to \( q_f \) labelled by \( a_n \), and there are transitions in \( A \) from \( q_i \) to \( q_{i+1} \) labelled by \( a_i \) for \( i = 1, \ldots, n - 1 \).

(4) \( h_4 h_3^{-1} h_2 h_1^{-1}(1*0) = R \).

Intuitively, \( h_1^{-1} \) produces an arbitrary sequence of transitions, each of which is in \( A \). After \( h_2 \) changes the representation of this sequence, \( h_3^{-1} \) ensures that the sequence of transitions represents an accepting computation of \( A \). Finally, \( h_4 \) erases everything except the letters causing that computation.

If \( \lambda \in R \), a new letter, \( c \), is added to the domain alphabet of \( h_3 \) and \( h_4 \). The definitions of \( h_3 \) and \( h_4 \) are extended by

\[ h_3(c) = $, \quad h_4(c) = \lambda. \]

The homomorphisms \( h_1 \) and \( h_2 \) remain unchanged. This adds the one additional word, \( c, \) to \( h_3^{-1} h_2 h_1^{-1}(1*0) \), which in turn causes \( \lambda \) to be added to \( h_4 h_3^{-1} h_2 h_1^{-1}(1*0) \). \( \Box \)

Corollary. For every alphabet \( \Sigma \) and every regular language \( R \subseteq \Sigma^* \) there are homomorphisms \( g_1, g_2, \) and \( g_3 \) such that \( R = g_3 g_2^{-1} g_1(\{1, \ldots, n\}*0) \), where \( n \) is sufficiently large.

Proof. Let \( m \) denote the cardinality of \( \Sigma_A \), let \( n \geq m \), and let \( f \) be any bijection from \( \{1, \ldots, n\} \) to \( \Sigma_A \). Defining \( g_1 : \{0, 1, \ldots, n\} \to (\Sigma \cup \emptyset \cup \{\$\})^* \) by

\[ g_1(0) = $, \quad g_1(x) = h_2 f \quad \text{for} \ 1 \leq x \leq m, \quad g_1(x) = \lambda \quad \text{for} \ m < x \leq n \]

if follows easily that \( g_1(\{1, \ldots, n\}*0) = h_2 h_1^{-1}(1*0) \). Hence, with \( g_2 = h_3 \) and \( g_3 = h_4 \), we have that \( R = g_3 g_2^{-1} g_1(\{1, \ldots, n\}*0) \). \( \Box \)

We leave as an open problem the question of whether the representation in our theorem is optimal or whether an analogous result can be obtained using only three morphisms.

References