RAIRO-Inf. Theor. Appl. **39** (2005) 661-675
DOI: 10.1051/ita:2005035

EQUALITY SETS FOR RECURSIVELY ENUMERABLE LANGUAGES

Vesa Halava[1], Tero Harju[1], Hendrik Jan Hoogeboom[2] and Michel Latteux[3]

Abstract. We consider shifted equality sets of the form $E_G(a, g_1, g_2) = \{w \mid g_1(w) = ag_2(w)\}$, where g_1 and g_2 are nonerasing morphisms and a is a letter. We are interested in the family consisting of the languages $h(E_G(J))$, where h is a coding and $E_G(J)$ is a shifted equality set. We prove several closure properties for this family. Moreover, we show that every recursively enumerable language $L \subseteq A^*$ is a projection of a shifted equality set, that is, $L = \pi_A(E_G(a, g_1, g_2))$ for some (nonerasing) morphisms g_1 and g_2 and a letter a, where π_A deletes the letters not in A. Then we deduce that recursively enumerable star languages coincide with the projections of equality sets.

Mathematics Subject Classification. 03D25, 68Q45.

1. Introduction

In formal language theory, languages are often determined by their generating grammars or accepting machines. It is also customary to say that languages generated by grammars of certain form or accepted by automata of specific type form a language family. Here we shall study a language family defined by simple generalized equality sets of the form $E_G(J)$, where $J = (a, g_1, g_2)$ is an instance of the *shifted Post Correspondence Problem* consisting of a letter a and two morphisms g_1 and g_2. Then the set $E_G(J)$ consists of the words w that satisfy $g_1(w) = ag_2(w)$.

Keywords and phrases. Morphism, equality set, shifted Post Correspondence Problem, closure properties, recursively enumerable sets.

[1] Department of Mathematics and TUCS – Turku Centre for Computer Science, University of Turku, 20014 Turku, Finland; vesa.halava@utu.fi; harju@utu.fi

[2] Department of Computer Science, Leiden University PO Box 9512, 2300 RA Leiden, The Netherlands; hoogeboom@liacs.nl

[3] Université des Sciences et Technologies de Lille, Bâtiment M3, 59655 Villeneuve d'Ascq Cedex, France; latteux@lifl.fr

Our motivation for these generalized equality sets comes partly from a result of [6], where it was proved that the family of regular valence languages is equal to the family of languages of the form $h(E_G(J))$, where h is a coding (*i.e.*, a letter-to-letter morphism), and, moreover, in the instance $J = (a, g_1, g_2)$ the morphism g_2 is periodic. Here we shall consider general case where we do not assume g_2 to be periodic. However, we do assume that both morphisms are nonerasing. We study the closure properties of this family $\mathcal{CE}$ of languages. In particular, we show that $\mathcal{CE}$ is closed under union, product, Kleene plus, intersection with regular sets. Also, more surprisingly, $\mathcal{CE}$ is closed under nonerasing inverse morphisms.

In the last section, we consider the connection of the sifted equality sets to recursively enumerable languages. In particular, we show that every every recursively enumerable language $L \subseteq A^*$ is a projection of a shifted equality set, that is, $L = \pi_A(E_G(a, g_1, g_2))$ for some (nonerasing) morphisms g_1 and g_2 and a letter a, where π_A deletes the letters not in A.

The results of Sections 2 and 3 have been proved in the authors' conference paper [7]. The characterization results of Section 4 concerning presentation of recursively enumerable sets by shifted equality sets are new. The problem of presenting recursively enumerable sets using (general) equality sets was initiated by Salomaa [14], Culik II [1], and Engelfriet and Rozenberg [2,3]; see also [4,11,15,16].

2. Preliminaries

Let A be an alphabet, and denote by A^* the monoid of all finite words under the operation of concatenation. Note that the *empty word*, denoted by ε, is in the monoid A^*. The semigroup $A^* \setminus \{\varepsilon\}$ generated by A is denoted by A^+. For a subset $L \subseteq A^*$, we denote by L^+ the set of all words of the form $w_1 w_2 \dots w_n$ for $w_i \in L$ with $n \geq 1$. Then $L^* = L^+ \cup \{\varepsilon\}$.

For two words $u, v \in A^*$, u is a *prefix* of v if there exists a word $z \in A^*$ such that $v = uz$. If $v = uz$, then we also write $u = vz^{-1}$ and $z = u^{-1}v$.

In the following, let A and B be alphabets and $g\colon A^* \to B^*$ a mapping. For a word $x \in B^*$, we denote by $g^{-1}(x) = \{w \in A^* \mid g(w) = x\}$ the inverse image of x under g. Then $g^{-1}(K) = \cup_{x \in K} g^{-1}(x)$ is the *inverse image* of $K \subseteq B^*$ under g, and $g(L) = \{g(w) \mid w \in L\}$ is the *image* of $L \subseteq A^*$ under g. Also, g is a *morphism* if $g(uv) = g(u)g(v)$ for all $u, v \in A^*$. A morphism g is a *coding*, if it maps letters to letters, that is, if $g(A) \subseteq B$. A morphism g is said to be *periodic*, if there exists a word $w \in B^*$ such that $g(A^*) \subseteq \{w\}^*$.

If A and B are alphabets such that $A \subseteq B$, then the morphism $\pi_A\colon B^* \to A^*$, defined by

$$\pi_A(a) = \begin{cases} a & \text{if } a \in A, \\ \varepsilon & \text{if } a \in B \setminus A, \end{cases}$$

is the *projection* of B^* onto A^*.

In the following section, for a given alphabet A, the alphabet $\bar{A} = \{\bar{a} \mid a \in A\}$ is a *copy* of A, if $A \cap \bar{A} = \emptyset$.

In the *Post Correspondence Problem*, PCP for short, we are given two morphisms $g_1, g_2 \colon A^* \to B^*$ and it is asked whether or not there exists a nonempty word $w \in A^+$ such that $g_1(w) = g_2(w)$. Here the pair (g_1, g_2) is an *instance* of the PCP, and the word w is called a *solution*. As a general reference to the problems and results concerning the Post Correspondence Problem, we give [8].

For an instance $I = (g_1, g_2)$ of the PCP, let

$$E(I) = \{w \in A^* \mid g_1(w) = g_2(w)\}$$

be its *equality set*. It is easy to show that an equality set $E = E(g_1, g_2)$ is always a monoid, that is, $E = E^*$. In fact, it is a free monoid, and thus the algebraic structure of E is relatively simple, although the problem whether or not E is trivial is undecidable.

We shall now consider special instances of the *generalized* Post Correspondence Problem in order to have slightly more structured equality sets. In the *shifted Post Correspondence Problem*, or *shifted PCP* for short, we are given two morphisms $g_1, g_2 \colon A^* \to B^*$ and a letter $a \in B$, and it is asked whether there exists a word $w \in A^*$ such that

$$g_1(w) = ag_2(w). \tag{1}$$

The triple $J = (a, g_1, g_2)$ is called an *instance* of the shifted PCP and a word w satisfying equation (1) is called a *solution* of J. It is clear that a solution w is always nonempty. We let

$$E_G(J) = \{w \in A^+ \mid g_1(w) = ag_2(w)\}$$

be the *shifted equality set* of J.

We shall denote by $\mathcal{CE}$ the set of all languages $h(E_G(J))$, where h is a coding, and the morphisms in the instances J of the shifted PCP are both nonerasing.

In [6] $\mathcal{CE}_{\text{per}}$ is defined as the family of languages $h(E_G(J))$, where h is a coding, and one of the morphisms in the instance J of the shifted PCP is assumed to be periodic. It was proved in [6] that $\mathcal{CE}_{\text{per}}$ is equal to the family of languages defined by the regular valence grammars (see [12]). It is easy to see that the morphisms in the instances could have been assumed to be nonerasing in order to get the same result. Therefore, the family $\mathcal{CE}$ studied in this paper is a generalization of $\mathcal{CE}_{\text{per}}$ or, actually, $\mathcal{CE}_{\text{per}}$ is a subfamily of $\mathcal{CE}$.

3. Closure properties of $\mathcal{CE}$

The closure properties of the family $\mathcal{CE}_{\text{per}}$ follow from the known closure properties of regular valence languages. In this section, we study the closure properties of the more general family $\mathcal{CE}$ under various operations.

Before we start our journey through the closure results, we make first some assumptions of the instances of the shifted PCP defining the languages at hand.

An instance $J = (a, g_1, g_2)$ of the shifted PCP is said to be *frontal*, if the shift letter a appears only as the first letter in the images of g_1 and a does not occur at all in the images of g_2.

Lemma 3.1. *Let $L = h(E_G(J))$ for a instance $J = (a, g_1, g_2)$ of the shifted PCP and a coding h. There exists a frontal instance $J' = (\#, g_1', g_2')$ and a coding h' such that $L = h'(E_G(J'))$.*

Proof. Assume $g_1, g_2 \colon A^* \to B^*$ and $h \colon A^* \to C^*$. Let $\#$ be a letter not in B. We shall construct a new instance $J' = (\#, g_1', g_2')$, where $g_1', g_2' \colon (A \cup \bar{A})^* \to (B \cup \{\#\})^*$ and $\bar{A}$ is a copy of A, by setting for all $x \in A$ $g_2'(x) = g_2'(\bar{x}) = g_2(x)$, and $g_1'(x) = g_1(x)$ and

$$g_1'(\bar{x}) = \begin{cases} g_1(x), & \text{if } a \text{ is not a prefix of } g_1(x), \\ \#w, & \text{if } g_1(x) = aw. \end{cases}$$

Define a new coding $h' \colon (A \cup \bar{A})^* \to C^*$ by $h'(x) = h'(\bar{x}) = h(x)$ for all $x \in A$. It is now obvious that $L = h'(E_G(J'))$. □

The next lemma shows that we may also assume that the instance (g_1, g_2) does not have any nontrivial solutions, that is, $E(g_1, g_2) = \{\varepsilon\}$ for all instances $J = (a, g_1, g_2)$ defining the language $h(E_G(J))$. For this result we introduce two mappings which are used for desynchronizing a pair of morphisms. Let d be a new letter. For a word $u = a_1 a_2 \cdots a_n$, where each a_i is a letter, define

$$\ell_d(u) = da_1 da_2 d \cdots da_n \quad \text{and} \quad r_d(u) = a_1 da_2 d \cdots da_n d.$$

In other words, ℓ_d is a morphism that adds d in front of every letter and r_d is a morphism that adds d after every letter of a word. This is a standard technique in language theory, see *e.g.* [8].

Lemma 3.2. *For every instance J of the shifted PCP and coding h, there exists a frontal instance $J' = (a, g_1', g_2')$ and a coding h' such that $h(E_G(J)) = h'(E_G(J'))$ and $E(g_1', g_2') = \{\varepsilon\}$.*

Proof. By Lemma 3.1, we can assume that $J = (a, g_1, g_2)$ is a frontal instance of the shifted PCP. Let $g_1, g_2 \colon A^* \to B^*$, and let $h \colon A^* \to C^*$. We define new morphisms $g_1', g_2' \colon (A \cup \bar{A})^* \to (B \cup \{d\})^*$, where $d \notin B$ is a new letter and $\bar{A}$ is a copy of A, as follows. For all $x \in A$,

$$g_2'(x) = \ell_d(g_2(x)) \quad \text{and} \quad g_2'(\bar{x}) = \ell_d(g_2(x))d, \tag{2}$$

$$g_1'(x) = g_1'(\bar{x}) = \begin{cases} ad \cdot r_d(w), & \text{if } g_1(x) = aw, \\ r_d(g_1(x)), & \text{if } a \text{ is not a prefix of } g_1(x). \end{cases} \tag{3}$$

It is clear that J' is a frontal instance. Note also that, since the images $g'_2(\bar{x})$ start and end in d, the letters in $\bar{A}$ can be used only as the last letter of a solution of $J' = (a, g'_1, g'_2)$. Since every image by g'_2 begins with letter d and it is not a prefix of any image of g'_1, we obtain that $E(g'_1, g'_2) = \{\varepsilon\}$. On the other hand, (a, g'_1, g'_2) has a solution $w\bar{x}$ if and only if wx is a solution of (a, g_1, g_2). Therefore, we can define $h'\colon (A \cup \bar{A})^* \to C^*$ by $h'(x) = h'(\bar{x}) = h(x)$ for all $x \in A$. The claim of the lemma follows, since obviously $h(E_G(J)) = h'(E_G(J'))$. □

We call an instance (a, g_1, g_2) *reduced*, if it is frontal and $E(g_1, g_2) = \{\varepsilon\}$.

3.1. Rational operations

Theorem 3.3. *The family $\mathcal{CE}$ is closed under union and product of languages.*

Proof. Let $K, L \in \mathcal{CE}$ with $K = h_1(E_G(J_1))$ and $L = h_2(E_G(J_2))$, where $J_1 = (a_1, g_{11}, g_{12})$ and $J_2 = (a_2, g_{21}, g_{22})$ are reduced, and $g_{11}, g_{12}\colon \Sigma^* \to B_1^*$ and $g_{21}, g_{22}\colon \Omega^* \to B_2^*$. Without restriction we can suppose that $\Omega \cap \Sigma = \emptyset$. (Otherwise we take a copy of the alphabet Ω that is disjoint from Σ.) We can also assume that $B_1 \cap B_2 = \emptyset$. Let $B = B_1 \cup B_2$.

(1) For the closure under union, let $\#$ be a new letter. First replace every appearance of the shift letters a_1 and a_2 in J_1 and J_2 by $\#$. Define $g_1, g_2\colon (\Sigma \cup \Omega)^* \to B^*$ as follows: for all $x \in \Sigma \cup \Omega$,

$$g_1(x) = \begin{cases} g_{11}(x), & \text{if } x \in \Sigma \\ g_{21}(x), & \text{if } x \in \Omega \end{cases} \quad \text{and} \quad g_2(x) = \begin{cases} g_{12}(x), & \text{if } x \in \Sigma \\ g_{22}(x), & \text{if } x \in \Omega. \end{cases}$$

Define a coding $h\colon (\Sigma \cup \Omega)^* \to C^*$ similarly:

$$h(x) = \begin{cases} h_1(x), & \text{if } x \in \Sigma \\ h_2(x), & \text{if } x \in \Omega. \end{cases} \tag{4}$$

Since $\Sigma \cap \Omega = \emptyset$, and the instances J_1 and J_2 are reduced (*i.e.*, $E(g_{11}, g_{12}) = \{\varepsilon\} = E(g_{21}, g_{22})$), it follows that the solutions in $E_G(J_1)$ and $E_G(J_2)$ cannot be combined or mixed. Thus, it is easy to see that $h(E_G(\#, g_1, g_2)) = K \cup L$.

(2) For the closure under product, we assume that the length of the images of the morphisms are at least 2. (Actually, this is needed only for g_{11}). This can be assumed, for example, by the construction in the proof of Lemma 3.2.

We shall prove that $KL = \{uv \mid u \in K,\ v \in L\}$ is in $\mathcal{CE}$. For this, we define $g_1, g_2\colon (\Sigma \cup \Omega)^* \to B^*$ in the following way: for each $x \in \Sigma$,

$$g_1(x) = \begin{cases} \ell_{a_2}(g_{11}(x)), & \text{if } a_1 \text{ is not a prefix of } g_{11}(x), \\ a_1 y \ell_{a_2}(w), & \text{if } g_{11}(x) = a_1 y w \quad (y \in B_1), \end{cases}$$

and

$$g_2(x) = r_{a_2}(g_{12}(x)),$$

and for each $x \in \Omega$, $g_1(x) = g_{21}(x)$ and $g_2(x) = g_{22}(x)$. If we now define h by combining h_1 and h_2 as in (4), we obtain that $h(E_G(a_1, g_1, g_2)) = KL$. □

We shall now extend the above result by proving that $\mathcal{CE}$ is closed under Kleene plus, *i.e.*, if $K \in \mathcal{CE}$, then also $K^+ \in \mathcal{CE}$. Clearly $\mathcal{CE}$ is not closed under Kleene star, since the empty word does not belong to any language in $\mathcal{CE}$.

Theorem 3.4. *The family $\mathcal{CE}$ is closed under Kleene plus.*

Proof. Let $K = h(E_G(a, g_1, g_2))$, where $g_1, g_2 \colon A^* \to B^*$ are nonerasing morphisms, $h \colon A^* \to C^*$ is a coding and the instance (a, g_1, g_2) is frontal. Also, let $\bar{A}$ be a copy of A, and define $\bar{g}_1, \bar{g}_2 \colon (A \cup \bar{A})^* \to B^*$ in the following way: for each $x \in A$,

$$\begin{aligned}
&\bar{g}_1(x) = g_1(x) \quad \text{and} \quad \bar{g}_2(x) = g_2(x),\\
&\bar{g}_1(\bar{x}) = \begin{cases} \ell_a(g_1(x)), & \text{if } a \text{ is not a prefix of } g_1(x),\\ \ell_a(w), & \text{if } g_1(x) = aw, \end{cases}\\
&\bar{g}_2(\bar{x}) = r_a(g_2(x)).
\end{aligned}$$

Extend h also to $\bar{A}$ by setting $h(\bar{x}) = h(x)$ for all $x \in A$.

Now $h(E_G(a, \bar{g}_1, \bar{g}_2)) = K^+$, since $\bar{g}_1(w) = a\bar{g}_2(w)$ if and only if, $w = x_1 \cdots x_n x_{n+1}$, where $x_i \in \bar{A}^+$ for $1 \le i \le n$, $x_{n+1} \in A^+$, $\bar{g}_1(x_i)a = a\bar{g}_2(x_i)$ for $1 \le i \le n$ and $\bar{g}_1(x_{n+1}) = a\bar{g}_2(x_{n+1})$. After removing the bars form the letters x_i (by h), we obtain words in $E_G(a, g_1, g_2)$. □

3.2. Intersection with regular languages

We show now that $\mathcal{CE}$ is closed under intersections with regular languages.

Theorem 3.5. *The family $\mathcal{CE}$ is closed under intersections with regular languages.*

Proof. Let $J = (a, g_1, g_2)$ be an instance of the shifted PCP, $g_1, g_2 \colon \Sigma^* \to B^*$. Let $L = h(E_G(J))$, where $h \colon \Sigma^* \to C^*$ is a coding.

We shall prove that $h(E_G(J)) \cap R$ is in $\mathcal{CE}$ for all regular $R \subseteq B^*$. We note first that $h(E_G(J)) \cap R = h(E_G(J) \cap h^{-1}(R))$, and therefore it is sufficient to show that, for all regular languages $R \subseteq \Sigma^*$, $h(E_G(J) \cap R)$ is in $\mathcal{CE}$. Therefore, we shall give a construction for instances J' of the shifted PCP such that $E_G(J') = E_G(J) \cap R$.

Assume $R \subseteq \Sigma^*$ is regular, and let $G = (N, \Sigma, P, S)$ be a right linear grammar generating R (see [13]). Let $N = \{A_0, \ldots, A_{n-1}\}$, where $S = A_0$, and assume without restriction, that S does not appear on the right hand side of any production. We consider the set P of the productions as an alphabet.

Let # and d be new letters. We define $g_1', g_2' \colon P^* \to (B \cup \{d, \#\})^*$ as follows. First assume that

$$g_1(a) = a_1 a_2 \dots a_k \quad \text{and} \quad g_2(a) = b_1 b_2 \dots b_m$$

for the (generic) letter a. We define

$$g_1'(p) = \begin{cases} \#d^n a_1 d^n a_2 d^n \dots a_k d^j \ , & \text{if } p = (A_0 \to a A_j) \\ d^{n-i} a_1 d^n a_2 d^n \dots a_k d^j \ , & \text{if } p = (A_i \to a A_j), \\ \#d^n a_1 d^n a_2 d^n \dots a_k \ , & \text{if } p = (A_0 \to a), \\ d^{n-i} a_1 d^n a_2 d^n \dots a_k \ , & \text{if } p = (A_i \to a), \end{cases}$$

and

$$g_2'(p) = d^n b_1 d^n b_2 \dots d^n b_m, \quad \text{if } p = (A \to aX),$$

where $X \in N \cup \{\varepsilon\}$.

As in [9], $E_G(J') = E_G(J) \cap R$ for the new instance $J' = (\#, g_1', g_2')$. The claim follows from this. □

3.3. Morphisms

Next we shall present a construction for the closure under nonerasing morphisms. This construction is a bit more complicated than the previous ones.

Theorem 3.6. *The family $\mathcal{CE}$ is closed under taking images of nonerasing morphisms.*

Proof. Let $J = (a, g_1, g_2)$ be an instance of the shifted PCP, where $g_1, g_2 \colon A^* \to B^*$. Let $L = h(E_G(J))$, where $h \colon A^* \to C^*$ is a coding. Assume that $f \colon C^* \to \Sigma^*$ is a nonerasing morphism. We shall construct h', g_1' and g_2' such that $f(L) = h'(E_G(J'))$ for the new instance $J' = (a, g_1', g_2')$.

First we show that we can restrict ourselves to cases where

$$\min\{|g_1(x)|, |g_2(x)|\} \geq |f(x)| \ \text{ for all } \ x \in A. \tag{5}$$

Indeed, suppose the instance J does not satisfy (5). We construct a new instance $\bar{J} = (\#, \bar{g}_1, \bar{g}_2)$ and a coding $\bar{h}$ such that $\bar{h}(E_G(\bar{J}) = h(E_G(J))$ and $\bar{g}_1$ and $\bar{g}_2$ do fulfill (5). Let $c \notin B$ be a new letter. Let $k = \max_{x \in A}\{|f(x)|\}$. We define $\bar{g}_1(x) = \ell_c^k(g_1(x))$ and $\bar{g}_2(x) = \ell_c^k(g_2(x))$ for all $x \in A$. We also need a new copy x' of each letter x for which a is a prefix of $g_1(x)$. If $g_1(x) = aw$, where $w \in B^*$, then define $\bar{g}_1(x') = \#\ell_c^k(w)$. It now follows that if $u \in E_G(\bar{J})$, then $u = x'v$ for some word $v \in A^*$ and $xv \in E_G(J)$. Therefore, by defining $\bar{h}$ as follows

$$\bar{h}(y) = \begin{cases} h(y), & \text{if } y \in A, \\ h(x), & \text{if } y = x', \end{cases}$$

we have $\bar{h}(E_G(\bar{J}) = h(E_G(J))$ as required.

Now assume that (5) holds in $J = (a, g_1, g_2)$ and for f. Let us consider the nonerasing morphism $fh\colon A^* \to \Sigma^*$. Note that also the composition fh satisfies (5). In order to prove the claim, it is clearly sufficient to consider the case, where h is the identity mapping, that is, $f = fh$.

First we define for every image $f(x)$, where $x \in A$, a new alphabet $A_x = \{b_x \mid b \in \Sigma\}$. We consider the words

$$(b_1 b_2 \dots b_m)_x = (b_1)_x (b_2)_x \dots (b_m)_x,$$

for $f(x) = b_1 \dots b_m$.

Let c and d be new letters and let $n = \sum_{x \in A} |f(x)|$. Assume that $A = \{x_1, x_2, \dots, x_q\}$.

Partition the integers $1, 2, \dots, n$ into q sets such that for the letter x_i there corresponds a set, say $S_i = \{i_1, i_2, \dots, i_{|f(x_i)|}\}$, of $|f(x_i)|$ integers.

Assume that $f(x_i) = b_1 \dots b_m$, $g_1(x_i) = a_1 a_2 \dots a_\ell$, and $g_2(x_i) = a'_1 a'_2 \dots a'_k$. We define new morphisms g'_1 and g'_2 as follows:

$$\begin{aligned}
g'_1((b_1)_{x_i}) &= c^n d^n a_1 c^{i_1}, \\
g'_1((b_j)_{x_i}) &= c^{n-i_{j-1}} d^n a_j c^{i_j} \quad \text{for } j = 2, \dots, m-1, \\
g'_1((b_m)_{x_i}) &= c^{n-i_{m-1}} d^n a_m c^n d^n \dots c^n d^n a_\ell,
\end{aligned}$$

and

$$\begin{aligned}
g'_2((b_1)_{x_i}) &= c^n d^n a_1 c^n d^{i_1}, \\
g'_2((b_j)_{x_i}) &= d^{n-i_{j-1}} a'_j c^n d^{i_j} \quad \text{for } j = 2, \dots, m-1, \\
g'_2((b_m)_{x_i}) &= c^n d^{n-i_{m-1}} a'_m c^n d^n \dots c^n d^n a'_k.
\end{aligned}$$

Then

$$\begin{aligned}
g'_1((b_1 \dots b_m)_{x_i}) &= c^n d^n a_1 c^n d^n a_2 \dots c^n d^n a_\ell, \\
g'_2((b_1 \dots b_m)_{x_i}) &= c^n d^n a'_1 c^n d^n a'_2 \dots c^n d^n a'_k.
\end{aligned}$$

The beginning has to be still fixed. For the cases, where $a_1 = a$, we need new letters $(b_1)'_{x_i}$, for which we define

$$g'_1((b_1)'_{x_i}) = a c^{i_1} \text{ and } g'_2((b_1)'_{x_i}) = c^n d^n a_j c^n d^{i_1}.$$

Now our constructions for the morphisms g'_1 and g'_2 are completed.

Next we define h', by setting $h'((b_i)_x) = b_i$ and $h'((b_1)'_x) = b_1$ for all i and x. We obtain that $h'(E_G(J')) = f(h(E_G(J))$, which proves the claim. □

Next we shall prove that the family $\mathcal{CE}$ is closed under inverse of nonerasing morphisms.

Theorem 3.7. *The family $\mathcal{CE}$ is closed under nonerasing inverse morphisms.*

Proof. Consider an instance $h(E_G(J))$, where $J = (a, g_1, g_2)$ with $g_i\colon A^* \to B^*$ and $h\colon A^* \to C^*$ is a coding. We may assume that $h(A) = C$.

Moreover, let $g\colon \Sigma^* \to C^*$ be a nonerasing morphism.

For each $x \in \Sigma$, let $h^{-1}g(x) = \{v_{x,1}, v_{x,2}, \ldots, v_{x,k_x}\}$ and let

$$\Sigma_x = \left\{x^{(1)}, \ldots, x^{(k_x)}\right\}$$

be a set of new letters for x. Denote $\Theta = \cup_{x\in\Sigma}\Sigma_x$, and define the morphisms $g_1', g_2'\colon \Theta^* \to B^*$ and the coding $t\colon \Theta^* \to \Sigma^*$ by

$$g_j'\left(x^{(i)}\right) = g_j(v_{x,i}) \quad \text{for } j = 1, 2, \quad \text{and} \quad t\left(x^{(i)}\right) = x$$

for each $x^{(i)} \in \Theta$.

Consider the instance $J' = (a, g_1', g_2')$.

Now, assume that $u = a_1a_2 \ldots a_n \in g^{-1}h(E_G(J))$ (with $a_i \in \Sigma$). Then there exists a word $w = w_1w_2 \ldots w_n$ such that $g_1(w) = ag_2(w)$ and $a_i \in g^{-1}h(w_i)$, that is, $w_i = v_{a_i,r_i} \in h^{-1}g(a_i)$ for some r_i, and so $g_1'(w') = ag_2'(w')$ for the word $w' = a_1^{(r_1)}a_2^{(r_2)} \ldots a_n^{(r_n)}$, for which $t(w') = u$. Therefore $u \in t(E_G(J'))$.

The converse inclusion $t(E_G(J')) \subseteq g^{-1}h(E_G(J))$ is clear by the above constructions. □

Let A and B be two alphabets. A mapping $\tau\colon A^* \to 2^{B^*}$, where 2^{B^*} denotes the set of all subsets of B^*, is a *substitution* if for all $u, v \in A^*$

$$\tau(uv) = \tau(u)\tau(v).$$

Note that τ is actually a morphism from A^* to 2^{B^*}.

A substitution τ is called *finite* if $\tau(a)$ is a finite set for all $a \in A$, and *nonerasing* if $\varepsilon \notin \tau(a)$ for all $a \in A$.

Corollary 3.8. *The family $\mathcal{CE}$ is closed under nonerasing finite substitutions.*

Proof. Since $\mathcal{CE}$ is closed under nonerasing morphisms, inverses of nonerasing morphisms, it is closed under nonerasing finite substitutions. Indeed, as is immediate, every finite substitution is a composition of an inverse of a coding and a nonerasing morphism. □

Note that $\mathcal{CE}$ is almost a *trio*, see [5], but it seems that it is not closed under *all* inverse morphisms. It is also almost a *bifaithful rational cone*, see [10], but since the languages do not contain the empty word, $\mathcal{CE}$ is not closed under the bifaithful finite transductions.

4. Equality sets and recursively enumerable languages

The following result of Engelfriet and Rozenberg [3] gives a classical morphic representation of recursively enumerable languages; see also Salomaa [15] (see Th. 6.9, p. 111). Recall that π_A denotes the projection onto A^*.

Theorem 4.1. *For every recursively enumerable language $L \subseteq A^*$, there are two morphisms h_1, h_2 and a regular language R such that $L = \pi_A(E(h_1, h_2) \cap R)$.*

A slight modification of its proof permits to strengthen this theorem:

Lemma 4.2. *For every recursively enumerable language $L \subseteq A^*$, there are two nonerasing morphisms h_1, h_2 and a regular language R such that $L = \pi_A(E(h_1, h_2) \cap R)$. Moreover, one can take $R = KA^*K'$ where K and K' are regular languages defined on an alphabet B disjoint from A.*

Proof. Assume first that $\varepsilon \notin L$. Let $G = (N, A, P, S)$ be a type 0 grammar generating L, where we can assume that the productions have no terminal letters on the right hand side, *i.e.*, $P \subseteq N^+ \times (N \cup A)^+$. Let $\bar{A}$ be a copy of A that is disjoint from the other alphabets. Also, let $V = N \cup \bar{A}$ and $R = KA^*K'$ with

$$K = S_0 \triangleright (V^* P V^* \triangleright)^* \text{ and } K' = F\#^*,$$

where S_0, $\triangleright$, F and $\#$ are new symbols.

Let us define the morphisms h_1 and h_2 by

	S_0	$\triangleright$	$p = (u, v)$	$X \in N$	$\bar{a} \in \bar{A}$	$a \in A$	F	$\#$
h_1	$S_0 \triangleright S$	$\triangleright$	v	X	a	$\#$	$\#$	$\#$
h_2	S_0	$\triangleright$	u	X	a	a	$\triangleright$	$\#\#$

Let us take $u \in \pi_A(E(h_1, h_2) \cap R)$. Then there exists a word $z \in E(h_1, h_2) \cap R$ such that $h_1(z) = h_2(z)$, and $u = \pi_A(z)$. Here

$$z = S_0 \triangleright z_1 \triangleright \ldots \triangleright z_n F \#^i,$$

where $z_1, \ldots, z_{n-1} \in V^* P V^*$, $u = z_n$, and $i \geq 0$. Hence, for $1 \leq j \leq n-1$,

$$h_2(z_j) \Longrightarrow_G h_1(z_j),$$
$$h_1(S_0 \triangleright z_1 \triangleright \ldots \triangleright z_j) = h_2(S_0 \triangleright z_1 \triangleright \ldots \triangleright z_{j+1}).$$

Therefore, $h_2(z_1) = S$ and $h_1(z_j) = h_2(z_{j+1})$ for $1 \leq j \leq n-1$. So we obtain that

$$\begin{aligned} S = h_2(z_1) &\Longrightarrow_G h_1(z_1) = h_2(z_2) \Longrightarrow_G \ldots \Longrightarrow_G h_1(z_{n-2}) \\ &= h_2(z_{n-1}) \Longrightarrow_G h_1(z_{n-1}) = h_2(z_n) = u \end{aligned}$$

and therefore $u \in L$.

Conversely, if $u \in L$, then we have a derivation

$$S = w_1 \Longrightarrow_G w_2 \Longrightarrow_G \ldots \Longrightarrow_G w_n = u$$

according to the grammar G. For each $1 \leq j \leq n-1$, one can find $z_j \in V^*PV^*$ such that $h_1(z_j) = w_{j+1}$ and $h_2(z_j) = w_j$. Set then

$$z = S_0 \triangleright z_1 \triangleright \ldots \triangleright z_{n-1} \triangleright uF\#^{i+1},$$

where i is the length of u. Then $z \in R$ and one can easily check that $h_1(z) = h_2(z)$. Hence, $u = \pi_A(z) \in \pi_A(E(h_1, h_2) \cap R)$.

Finally, if $\varepsilon \in L$, set $D = A \cup \{d\}$, where d is a new symbol. Then

$$Ld = \pi_D(E(h_1, h_2) \cap KD^*K') = \pi_D(E(h_1, h_2) \cap KA^*dK')$$

and hence $L = \pi_A(E(h_1, h_2) \cap KA^*dK')$. This completes the proof of the lemma. □

Note that the form of the regular language R and the fact that the two morphisms are nonerasing are crucial for the proofs of the following lemmata. The proof of the following lemma uses the methods from [9].

Lemma 4.3. *Let A and B be two disjoint alphabets and $h_1, h_2 \colon (A \cup B)^* \to C^*$ be two nonerasing morphisms. If K and K' are two regular languages included in B^+, then $\pi_A(E(h_1, h_2) \cap KA^*K') = \pi_A(E_G(\#, g_1, g_2))$ for some nonerasing morphisms g_1 and g_2.*

Proof. Let us take two nondeterministic finite automata $M = (Q, B, \Delta, q_0, F)$, $M' = (Q', B, \Delta', q_0', F')$ such that $L(M) = K$ and $L(M') = K'$. The transitions are triples of the form (q, b, p), that is, $\Delta \subseteq Q \times B \times Q$ and $\Delta' \subseteq Q' \times B \times Q'$. Clearly, one can assume that $Q = \{q_0, \ldots, q_n\}$ and $Q' = \{q_0', \ldots, q_n'\}$ with $Q \cap Q' = \emptyset$, and that the automata have unique final states $F = \{q_n\}$ and $F' = \{q_n'\}$. Also, we can assume that there are no transitions (q, b, q_0) and (q', b, q_0') that enter the initial states q_0 and q_0', and that there are no transitions (q_f, b, q) and (q_f', b, q') leaving from the final states q_f and q_f'.

First, we define three morphisms Θ, ℓ and r as follows.

Let $\theta \colon (A \cup \Delta \cup \Delta')^* \to (A \cup B)^*$ be the morphism defined by

$$\theta(a) = a \text{ for } a \in A, \text{ and } \theta((p, b, q)) = b \text{ for } (p, b, q) \in \Delta \cup \Delta'.$$

Also, let $\ell = \ell_{d^{2n}}$ and $r = r_{d^{2n}}$, that is, $\ell, r \colon C^* \to (C \cup \{d\})^*$, where d is a new symbol, and

$$\ell(c) = d^{2n}c, \text{ and } r(c) = cd^{2n} \text{ for } c \in C.$$

The morphism $g_2\colon (A \cup \Delta \cup \Delta')^* \to (C \cup \{\#, d\})^*$ becomes defined by

$$g_2 = rh_2\theta.$$

It is immediate that $g_2(z) \in (Cd^{2n})^*$ for all $z \in (A \cup \Delta \cup \Delta')^*$. The shift letter $\#$ does not belong to any image of g_2. The notation wd^{-m} means $w(d^m)^{-1}$, that is, d^m is removed as a suffix of the word w, and similarly $d^{-m}w$ is an abbreviation for $(d^m)^{-1}w$. The morphism $g_1\colon (A \cup \Delta \cup \Delta')^* \to (C \cup \{\#, d\})^*$ is defined by

$$\begin{aligned}
&g_1((q_0, b, q_j)) = \#rh_1(b)d^{-2j},\\
&g_1((q_i, b, q_j)) = d^{2i}rh_1(b)d^{-2j} \text{ for } i \neq 0,\\
&g_1(a) = \ell h_1(a) \text{ for } a \in A,\\
&g_1((q'_0, b, q'_n)) = \ell h_1(b)d^{2n},\\
&g_1((q'_0, b, q'_j)) = \ell h_1(b)d^{2j+1} \text{ for } j \neq n,\\
&g_1((q'_i, b, q'_n)) = d^{-(2i+1)}\ell h_1(b)d^{2n} \text{ for } i \neq 0,\\
&g_1((q'_i, b, q'_j)) = d^{-(2i+1)}\ell h_1(b)d^{2j+1} \text{ for } i \neq 0 \text{ and } j \neq n.
\end{aligned}$$

The morphism g_1 decodes the behaviour of the combined automata that accepts the language KA^*K' in the sense that $g_1(z) \in \#(Cd^{2n})^*$ if and only if $z = uvu'$ for some words $u \in \Delta^*$, $v \in A^*$, and $u' \in \Delta'^*$ such that $\theta(u) \in K$ and $\theta(u') \in K'$. Therefore, we have

$$g_1(z) \in \#(Cd^{2n})^* \iff \theta(z) \in KA^*K'. \tag{6}$$

Finally, let $\pi = \pi_C$ be the projection $\pi\colon (C \cup \{\#, d\})^* \to C^*$ that deletes the letters d and $\#$. Then we have

$$\pi g_1 = h_1\theta \quad \text{and} \quad \pi g_2 = h_2\theta. \tag{7}$$

Let v be a word in $\pi_A(E_G(\#, g_1, g_2))$ and let z be such that $v = \pi_A(z)$ and $g_1(z) = \#g_2(z)$. Since $g_2(z) \in (Cd^{2n})^*$, also $g_1(z) \in (Cd^{2n})^*$, and it follows by (6) that $\theta(z) \in KA^*K'$. Consequently, by (7), we have

$$h_1\theta(z) = \pi g_1(z) = \pi(\#g_2(z)) = \pi g_2(z) = h_2\theta(z).$$

Hence, $\theta(z) \in E(h_1, h_2) \cap KA^*K'$ and also $v = \pi_A\theta(z) \in \pi_A(E(h_1, h_2) \cap KA^*K')$ as required.

Conversely, let $v \in \pi_A(E(h_1, h_2) \cap KA^*K')$, say $v = \pi_A(kvk')$ with $k \in K$, $k' \in K'$ and $h_1(kvk') = h_2(kvk')$. Then there exists a word $z = uvu'$ with $u \in \Delta^+$, $u' \in \Delta'^+$, $\theta(u) = k$, $\theta(u') = k'$, $\theta(z) = kvk'$ and $g_1(z) = \#rh_1\theta(z) = \#rh_2\theta(z) = \#g_2(z)$. Therefore, $v = \pi_A(z) \in \pi_A(E_G(\#, g_1, g_2))$, which completes the proof. □

From the two above lemmata, we obtain immediately the following result.

Theorem 4.4. *Every recursively enumerable language $L \subseteq A^*$ is a projection of a shifted equality set, that is, $L = \pi_A(E_G(a, g_1, g_2))$ for a letter a and some nonerasing morphisms g_1 and g_2.*

We remark that from this result it is very easy to find again Theorem 4.1. Indeed, if $L = \pi_A(E_G(a, g_1, g_2))$ for some morphisms g_1 and g_2 defined on an alphabet X, one gets $L = \pi_A(E(h_1, h_2) \cap dX^*)$ where d is a new letter, $h_1(d) = d$, $h_2(d) = da$ and $h_i(x) = g_i(x)$ for $x \in X$. Note also that the regular language dX^* is quite simple!

Two morphisms, $g_1, g_2 \colon A^* \to B^*$ are said to be *prefix-incomparable*, if for each letter $a \in A$, $g_1(a)$ is not a prefix of $g_2(a)$ and $g_2(a)$ is not a prefix of $g_1(a)$.

Lemma 4.5. *Let $L = E_G(\#, h_1, h_2)$ where h_1 and h_2 are nonerasing morphisms defined on the alphabet A. Then $L = \pi_A(E_G(\#, g_1, g_2))$ for some prefix-incomparable nonerasing morphisms g_1 and g_2.*

Proof. Let $h_1, h_2 \colon A^* \to X^*$, and let c and d be new letters. Set $B = A \cup \{d\}$ and $Y = X \cup \{c, d\}$. Recall that $\ell_c, r_c \colon X^* \to Y^*$ are defied by $\ell_c(x) = cx$ and $r_c(x) = xc$. Let $g_1 \colon B^* \to Y^*$ and $g_2 \colon B^* \to Y^*$ be defined by

$$g_1(d) = d \quad \text{and} \quad g_1(a) = r_c h_1(a) \quad \text{for } a \in A,$$
$$g_2(d) = cd \quad \text{and} \quad g_2(a) = \ell_c h_2(a) \quad \text{for } a \in A.$$

Clearly, $g_1(b)$ and $g_2(b)$ are prefix-incomparable morphisms. We have, for each $u \in A^*$, that

$$\#g_2(ud) = \#g_2(u)cd = \#\ell h_2(u)cd = r(\#h_2(u))d. \tag{8}$$

Now, if $u \in L$, that is, $h_1(u) = \#h_2(u)$, then it follows from (8) that $\#g_2(ud) = r(\#h_2(u))d = rh_1(u) = g_1(ud)$. Hence $u = \pi_A(ud) \in \pi_A(E_G(\#, g_1, g_2))$.

Conversely, assume that $u \in \pi_A(E_G(\#, g_1, g_2))$. Then there exists a word v such that $u = \pi_A(v)$ and $g_1(v) = \#g_2(v)$. By the definitions of the morphisms g_1 and g_2, we must have $v = ud$. From (8), we obtain $\#g_2(v) = r(\#h_2(u))d = g_1(v) = g_1(ud) = rh_1(u)d$. Thus $r(\#h_2(u))d = r(h_1(u))$, which implies $\#h_2(u) = h_1(u)$ and $u \in E_G(\#, h_1, h_2)$ as required. □

A language $L \subseteq A^*$ is a *star language*, if $L = L^*$, that is, if it is closed under concatenation.

As seen in the preliminaries, equality sets are star languages. So it is clear that projections of equality sets are recursively enumerable star languages. As a matter of fact, the following result shows that these two families coincide.

Theorem 4.6. *Every recursively enumerable star language is a projection of an equality set, that is, for every recursively enumerable $L \subseteq A^*$, there are nonerasing morphisms g_1 and g_2 such that $L^* = \pi_A(E(g_1, g_2))$.*

Proof. From Theorem 4.4, we have that $L^* = \pi_A(E_G(\#, h_1, h_2))$ for some nonerasing morphisms h_1 and h_2 defined on an alphabet X. When we apply Lemma 4.5 to the shifted equality set $E_G(\#, h_1, h_2)$, we can, without loss of generality, assume that the morphisms h_1 and h_2 prefix-incomparable. Let d be a new letter and set $Y = X \cup \{d\}$. Let us define the morphisms g_1 and g_2 by:

$$\begin{aligned} g_1(d) &= d \quad \text{and} \quad g_1(x) = h_1(x) \quad \text{for } x \in X, \\ g_2(d) &= d\# \quad \text{and} \quad g_2(x) = h_2(x) \quad \text{for } x \in X. \end{aligned}$$

Now, if $u \in L^*$, we have $g_1(du) = dh_1(u) = d\#h_2(u) = g_2(du)$ and from (8) $\#g_2(ud) = r(\#h_2(u))d = rh_1(u) = g_1(ud)$. Hence $u \in \pi_A(E(g_1, g_2))$.

Conversely, let $u \in \pi_A(E(g_1, g_2))$. Then $u = \pi_A(v)$ with $g_1(v) = g_2(v)$. Since for each $x \in X$, $h_1(x)$ and $h_2(x)$ are prefix-incomparable, we have $v = dv_1 \dots dv_n$ where each v_i is in X^*. Now,

$$g_1(v) = dh_1(v_1) \dots dh_1(v_n) = g_2(v) = d\#h_2(v_1) \dots d\#h_2(v_n).$$

Therefore $h_1(v_i) = \#h_2(v_i)$ for each i, and $\pi_A(v_i) \in L^*$. From these we obtain $u = \pi_A(v) = \pi_A(dv_1 \dots dv_n) = \pi_A(v_1 \dots v_n) \in L^*$, which proves the claim. □

We conclude with a remark that Theorem 4.1 is a direct consequence of this result. Indeed, let $L \subseteq A^*$ be a recursively enumerable language, d a new letter and set $D = A \cup \{d\}$. From Theorem 4.6, we obtain $(Ld)^* = \pi_D(E(g_1, g_2))$, and hence

$$Ld = (Ld)^* \cap A^*d = \pi_D(E(g_1, g_2) \cap \pi_D^{-1}(A^*d))$$

and so $L = \pi_A(E(g_1, g_2) \cap \pi_D^{-1}(A^*d))$ as required.

References

[1] K. Culik II, A purely homomorphic characterization of recursively enumerable sets. *J. Assoc. Comput. Mach.* **26** (1979) 345–350.

[2] J. Engelfriet and G. Rozenberg, Equality languages and fixed point languages. *Inform. Control* **43** (1979) 20–49.

[3] J. Engelfriet and G. Rozenberg, Fixed point languages, equality languages, and representation of recursively enumerable languages. *J. Assoc. Comput. Mach.* **27** (1980) 499–518.

[4] V. Geffert, A representation of recursively enumerable languages by two homomorphisms and a quotient. *Theoret. Comput. Sci.* **62** (1988) 235–249.

[5] S. Ginsburg, *Algebraic and Automata-theoretic Properties of Formal Languages*. North-Holland (1975).

[6] V. Halava, T. Harju, H.J. Hoogeboom and M. Latteux, Valence Languages Generated by Generalized Equality Sets. *J. Autom. Lang. Comb.*, to appear.

[7] V. Halava, T. Harju, H.J. Hoogeboom and M. Latteux, Languages defined by generalized equality sets, in *14th Internat. Symp. on Fundamentals of Computation Theory, FCT'03*, Malmö, Sweden, edited by A. Lingas and B.J. Nilsson. *Lect. Notes Comput. Sci.* **2751** (2003) 355–363.

[8] T. Harju and J. Karhumäki, *Morphisms*, Handbook of Formal Languages, edited by G. Rozenberg and A. Salomaa. Springer-Verlag **1** (1997).

[9] M. Latteux and J. Leguy, On the composition of morphisms and inverse morphisms. *Lect. Notes Comput. Sci.* **154** (1983) 420–432.
[10] M. Latteux and J. Leguy, On usefulness of bifaithful rational cones. *Math. Syst. Theor.* **18** (1985) 19–32.
[11] M. Latteux and P. Turakainen, On characterization of recursively enumerable languages. *Acta Informatica* **28** (1990) 179–186.
[12] Gh. Păun, A new generative device: valence grammars. *Revue Roumaine de Math. Pures et Appliquées* **6** (1980) 911–924.
[13] A. Salomaa, *Formal Languages*. Academic Press, New York (1973).
[14] A. Salomaa, Equality sets for homomorphisms of free monoids. *Acta Cybernetica* **4** (1978) 127–139.
[15] A. Salomaa, *Jewels of Formal Language Theory*. Computer Science Press (1981).
[16] P. Turakainen, A unified approach to characterizations of recursively enumerable languages. *Bulletin of the EATCS* **45** (1991) 223–228.

Communicated by J Karhumäki.
Received February 25, 2004. Accepted November 26, 2004.