

ON BINARY EQUALITY SETS AND A
SOLUTION TO THE EHRENFEUCHT
CONJECTURE IN THE BINARY CASE

by

A. Ehrenfeucht
J. Karhumaki
G. Rozenberg

CU-CS-221-82

A. Ehrenfeucht
Dept. of Computer Science
Univ. of Colo. at Boulder
Boulder, Colorado 80309 USA

J. Karhumaki
Dept. of Mathematics
University of Turku
Turku
Finland

G. Rozenberg
Institute of Applied Mathematics and
Computer Science
University of Leiden
Leiden
The Netherlands

All correspondence to
G. Rozenberg.

ANY OPINIONS, FINDINGS, AND CONCLUSIONS
OR RECOMMENDATIONS EXPRESSED IN THIS PUB-
LICATION ARE THOSE OF THE AUTHOR AND DO NOT
NECESSARILY REFLECT THE VIEWS OF THE
NATIONAL SCIENCE FOUNDATION.

THIS MATERIAL IS BASED UPON WORK SUPPORTED
BY THE NATIONAL SCIENCE FOUNDATION UNDER
GRANT NO. MCS 79-03838.

ON BINARY EQUALITY SETS AND A SOLUTION TO THE
EHRENFUCHT CONJECTURE IN THE BINARY CASE

by

A. Ehrenfeucht
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado
U.S.A.

J. Karhumäki
Department of Mathematics
University of Turku
Turku
Finland

G. Rozenberg
Institute of Applied Mathematics and
Computer Science
University of Leiden
Leiden
The Netherlands

Abstract. In [4] it was conjectured that the equality set of two injective morphisms over a binary alphabet is of the form $\{u, v\}^*$ for some (possibly empty) words u and v . Here we show that such an equality set is always either of the above form or of the form $(uw^*v)^*$ for some words u, w and v . As an application we give a simple proof for the Ehrenfeucht conjecture in the binary case, cf. [6]. In fact, we show that a test set can always be chosen to contain no more than three words.

1. INTRODUCTION

In recent years a lot of research has been done to study the problem of whether two morphisms agree word by word on at least one or on all words of a given language. Such problems have turned out important for many areas of *mathematics*, for example for computability theory, for theory of equations in free monoids and for formal language theory in general. The Post Correspondence Problem [1], the Ehrenfeucht Conjecture [10] and the DOL equivalence problem [3] are typical examples.

The notion of an equality language, introduced in [3], is central when dealing with the above problems. Equality languages has been studied later e.g. in [2], [8] and [4]. In the last mentioned paper equality languages were studied in the case of the binary alphabet, and there it was conjectured that if at least one of the morphisms is injective, then the equality set is a free monoid generated by at most two words. Here we take a step in the direction to prove this conjecture. Namely, we show that such an equality set is either of the above form or generated by a regular language of the form uw^*v .

As an application we give a simple proof for the Ehrenfeucht Conjecture in the binary case. The conjecture is as follows:

EHRENFUECHT CONJECTURE: For each language L over a finite alphabet there exists a finite subset F of L such that if, for an arbitrary pair (h, g) of morphisms, $h(x) = g(x)$ holds true for all x in F , then also $h(x) = g(x)$ holds true for all x in L .

The algebraic importance of the Ehrenfeucht Conjecture was emphasized when it was pointed out in [5] that it is equivalent to the following statement: Each system of equations (with a finite number of variables) over a finitely generated free monoid has a finite equivalent subsystem.

The above subset F of L was called in [6] a test set for L . The existence of a test set for context-free languages was proved in [1] and for arbitrary languages over a binary alphabet in [6]. Here we give a new and shorter proof for this latter result. Moreover, we show that such a test set can always be chosen to contain no more than three words, thus sharpening the result of Culik and Salomaa.

2. PRELIMINARIES

In this paper only very basic notions of free monoids and formal languages are needed. As a general reference we mention [9]. To fix our notation we want to specify the following.

Throughout this paper I denotes a binary alphabet, say $I = \{0, 1\}$. A free monoid generated by I is denoted by I^* and its identity, so called empty word, by λ . As usual we set $I^+ = I^* - \{\lambda\}$. For a word x in I^* and a letter c in I , $\#_c(x)$ means the number of c 's in x , and $|x|$ the length of x . For two words x and y the notation yx^{-1} is used to denote the right quotient of y by x , and the notation $x \text{ pref } y$ is used to denote that x is a prefix (not necessarily proper) of y . The prefix of the length k of a word x is denoted by $\text{pref}_k(x)$. If $|x| < k$, then we set $\text{pref}_k(x) = x$. By the relation $x \text{ Pref } y$ we mean that either x is a prefix of y or y is a prefix of x . We call a nonempty word x primitive if it is not a proper power of any word, i.e. the relation $x = z^n$ implies that $x = z$ and $n = 1$. The ratio of a word x in $\{0, 1\}^+$ is defined to be $\#_0(x) : \#_1(x)$ and is denoted by $r(x)$. By a ratio-primitive word, or r-primitive word in short, we mean a word such that none of its proper prefixes has the same ratio as the whole word.

Our basic notion is that of a morphism from a free monoid I^* into another free monoid A^* . Because of the nature of the problems we are interested in we may assume that $Z = A$. So we shall deal with morphisms $h: \{0, 1\}^* \rightarrow \{0, 1\}^*$. A morphism h is λ -free if $h(a) \neq \lambda$ for all a in I . We call a morphism h periodic

if there exists a word p such that $h(I) \subseteq p^*$. By a marked morphism $h: \{0, 1\}^* \rightarrow \{0, 1\}^*$ we mean a λ -free morphism satisfying $\text{pref}_1(h(0)) \neq \text{pref}_1(h(1))$.

It is well known that nonperiodic morphisms over a binary alphabet can be characterized as follows.

LEMMA 1. A morphism $h: \{0, 1\}^* \rightarrow \{0, 1\}^*$ is nonperiodic if and only if it is injective if and only if $h(01) \neq h(10)$.

Following [13] we define the equality set (or equality language) of the pair (h, g) of morphisms on I^* , in symbols $E(h, g)$, by

$$E(h, g) = \{x \in I^* \mid h(x) = g(x)\}.$$

We shall also need a little bit generalized notion defined as follows. For a pair (h, g) of morphisms on I^* and a word a in I^* , the a -shifted equality set of (h, g) , in symbols $E_a(h, g)$, is defined by

$$E_a(h, g) = \{x \in I^* \mid ah(x) = g(x)a\}.$$

It is easy to see that for a given equality set, and hence also for a given a -shifted equality set, all of its words has the same ratio. In the case when at least one of the morphisms is periodic even more can be said about the structure of an equality set. Indeed, we have, see [5],

THEOREM 1. If h and g are periodic, then either $E(h, g) = \{\lambda\}$ or $E(h, g) = \{\lambda\} \cup \{x \in I^* \mid r(x) = k\}$ for some $k \geq 0$ or $k = \infty$. If h is periodic and g is not, then $E(h, g) = u^*$ for some (possibly empty) word u .

We finish this section with the following notions. Let (h, g) be a pair of morphisms on L^* . We say that h and g agree on a word x from L^* if $h(x) = g(x)$ and that they agree on a language L if they agree on each word of L . Using this terminology the Ehrenfeucht Conjecture, cf. [7], can be stated as: For each language L (over a finite alphabet) there exists a finite subset F of L such that any pair of morphisms agree on L if and only if they agree on F . Following [6] we refer ¹⁰ such a finite subset F of L ~~to~~ as a test set for L .

3. CHARACTERIZATION

Here we give a partial characterization for equality sets of injective morphisms in the binary case. Our result can be seen as a step in the direction to prove the conjecture presented in [4].

First we need some notions and lemmas. Following [7] we define a mapping $\text{cyc}_1: \{0,1\}^* \rightarrow \{0,1\}^*$ by

$$\begin{aligned} \text{cyc}_1(\lambda) &= \lambda, \\ \text{cyc}_1(cu) &= uc \quad \text{for } c \in \{0,1\} \text{ and } u \in \{0,1\}^*. \end{aligned}$$

Let $\text{cyc}_k = (\text{cyc}_1)^k$ for $k \geq 1$. It follows that for any mapping $f: \{0,1\}^* \rightarrow \{0,1\}^*$ and for any word x in $\{0,1\}^*$ the following holds true

$$\text{cyc}_k(f(x)) = (\text{pref}_{k_1}(f(x)))^{-1} f(x) \text{pref}_{k_1}(f(x)), \quad (*)$$

where $0 \leq k_1 < |f(x)|$ and $k_1 \equiv k \pmod{|f(x)|}$. We now assume that h is a nonperiodic morphism, i.e. $h(01) \neq h(10)$. Let z_h be the maximal common prefix of $h(01)$ and $h(10)$. Clearly, $|z_h| \leq |h(01)|$. We define a mapping $h^-: \{0,1\}^* \rightarrow \{0,1\}^*$ by setting

$$h^- = \text{cyc}_{|z_h|} \circ h.$$

The following result is not difficult to see.

LEMMA 2. The mapping h^- is a morphism and moreover marked.

Observe that, in general, for a morphism h the mappings of the form $\text{cyc}_k \circ h$ need not be morphisms.

Now, let (h, g) be a pair of nonperiodic morphisms and z_h and z_g the above defined words associated to h and g , respectively. We assume, because of symmetry, that $|z_h| \geq |z_g|$. Then we have

LEMMA 3. If z_g is not a prefix of z_h , then either $E(h, g) = \{\lambda\}$ or $E(h, g) = a^*$ for some $a \in \{0, 1\}$.

Proof. If $|h(a)| = |g(a)|$, for $a \in \{0, 1\}$, then lemma clearly holds true. So let $|h(0)| \neq |g(0)|$ and $|h(1)| \neq |g(1)|$.

Assume that $x \in E(h, g)$, $x \neq \lambda$. Clearly, x contains both 0 and 1. Consequently, by the definitions of z_h and z_g , z_h is a prefix of $h(x)$ and z_g is a prefix of $g(x)$. This implies that z_g is a prefix of z_h , a contradiction. Hence $E(h, g) = \{\lambda\}$.

Now, we assume that $z_g \text{ pref } z_h$. We define

$${}^a h, g = z_g^{-1} z_h$$

and derive

LEMMA 4. Let (h, g) be a pair of morphisms such that ${}^a h, g$ is defined. Then

$$E(h, g) = E({}^a h, g').$$

Proof. Immediate, by definitions and (*).

Before stating the main result of this section we still need one notion. Let (β, γ, δ) be a triple of words such that γ is primitive and it is neither a suffix of β nor a prefix of δ . We call such a triple reduced and define the language $L(\beta, \gamma, \delta)$

by setting

$$L(\beta, \gamma, \delta) = \beta \gamma^* \delta \quad (**)$$

THEOREM 2. Let (h, g) be a pair of injective morphisms over a binary alphabet. The equality set $E(h, g)$ is either of the form

- (i) $\{u, v\}^*$ for some (possibly empty) words u and v , or of the form
- (ii) $(L(u, w, v))^*$ for some reduced triple (u, w, v) .

Proof. By lemma 3, if ${}^a h, g$ is not defined we are done.

Consequently, we assume that ${}^a h, g$ is defined. Then, by lemma 4, it is enough to show that $E({}^a h, g')$, where (h', g') is an arbitrary pair of marked morphisms and a is an arbitrary word, is of the form (i) or of the form (ii).

We have two cases to be considered.

I $a = \lambda$. Since h' and g' are marked $E(h', g')$ may contain at most two (one starting with 0 and another with 1) r -primitive words. Hence, $E({}^a h', g')$ is of the form (i).

II $a \neq \lambda$. Let us refer nonempty words in $E({}^a h', g')$ as solutions, and let $i \in \{0, 1\}$ be such that $\text{pref}_i(g'(i)) = \text{pref}_i(a)$. Then the first letter of any solution x is i . This is because g' is marked and this first letter is determined by the condition $a \text{ Pref } g'(\text{pref}_i(x))$. Moreover, by the same reasoning, the prefixes of x are also uniquely determined up to the prefix x' where $ah'(x') = g'(x')$. If such an x' does not exist, then, clearly, $E({}^a h', g')$ contains at most one r -primitive word, i.e. $E({}^a h', g')$ satisfies (i).

Now, we assume that all the solutions have a common prefix x'

such that $ah'(x') = g'(x')$. We have three subcases.

a) $E(h', g') = \{\lambda\}$. Now, by the fact that h' and g' are marked, there are at most two words satisfying the conditions $h'(z) = g'(z)a$ and $x'z$ is r -primitive. Consequently, $E_a(h', g')$ is of the form (i).

b) $E(h', g') = y^*$ for some nonempty word y . If for some nonempty prefix y' of y we have $h'(y') = g'(y')a$, then again $E_a(h', g')$ is of the form (i). If, on the other hand, such a prefix of y does not exist, then $E_a(h', g')$ is of the form (ii) or contains only λ depending on whether ~~or not~~ there exists or not a word z (with $\text{pref}_1(z) \neq \text{pref}_1(y)$) such that $h'(z) = g'(z)a$.

c) $E(h', g') = \{y_1, y_2\}^*$ for some nonempty words y_1 and y_2 with $\text{pref}_1(y_1) \neq \text{pref}_1(y_2)$. Now, if neither y_1 nor y_2 has a prefix z such that $h'(z) = g'(z)a$, then, clearly, $E_a(h', g') = \{\lambda\}$. If only one of the words y_1 and y_2 has the above mentioned prefix, then $E_a(h', g')$ is of the form (ii). Finally, if both y_1 and y_2 ^{have} such a prefix, then $E_a(h', g')$ is of the form (i).

Since h' and g' are marked the classification a) - c) in the case II is exhaustive, and so our proof for Theorem 2 is complete.

By careful analysis of the above proof we can say even more about the languages of the form (ii) in Theorem 2. Indeed, words u, w , and v satisfy: $\text{pref}_1(w) \neq \text{pref}_1(v)$, w contains both 0 and 1, and each of the words w, vu and $uw^i v$ for $i \geq 0$ is ratio-primitive.

We conclude this section by noting that we do not know whether there exists any equality set of the form (ii). As already con-

jectured in [4] we believe that there does not exist such sets. We also want to emphasize the following interesting property: Any finitely generated equality set in the binary case is generated by at most two words. As shown in [4], there really are equality sets (different from $\{*\}$) freely generated by two words.

4. APPLICATION TO THE EHRENFUCHT CONJECTURE

As was already mentioned the Ehrenfeucht conjecture was proved to hold in the case of the binary alphabet in [6]. As an application of Theorem 2 we give here a simple proof for the result. We also give a very small upper bound for the cardinality of such a test set: we show that it can always be chosen to contain no more than three words.

Recalling the definition of the languages of the form $L(\beta, \gamma, \delta)$ given in Section 3 we first prove

LEMMA 5. For two languages $L_1 = \beta_1 \gamma_1^* \delta_1$ and $L_2 = \beta_2 \gamma_2^* \delta_2$, where the triples $(\beta_i, \gamma_i, \delta_i)$ for $i = 1, 2$ are reduced, if $L_1 \cap L_2$ contains at least two words, then $L_1 = L_2$.

Proof. Assume that $L_1 \cap L_2$ contains two words, say

$$\beta_1 \gamma_1^t \delta_1 = \beta_2 \gamma_2^r \delta_2 \quad \text{and} \quad \beta_1 \gamma_1^q \delta_1 = \beta_2 \gamma_2^s \delta_2 \quad \text{with } t > q.$$

Let $|\beta_1 \gamma_1^q| \leq |\beta_2 \gamma_2^s|$ (the other case is symmetric). Then there exists a word u such that

$$\beta_1 \gamma_1^q u = \beta_2 \gamma_2^s \quad \text{and} \quad \delta_1 = u \delta_2 \quad (1)$$

and hence also

$$\gamma_1^{t-q} \delta_1 = u \gamma_2^{r-s} \delta_2 \quad \text{and} \quad \beta_1 \gamma_1^t u = \beta_2 \gamma_2^r. \quad (2)$$

Consequently,

$$\beta_1 \gamma_1^{t+(t-q)} \delta_1 = \beta_1 \gamma_1^t u \gamma_2^{r-s} \delta_2 = \beta_2 \gamma_2^{r+(r-s)} \delta_2$$

which implies that $\beta_1 \gamma_1^{2t-q} \delta_1 \in \beta_2 \gamma_2^* \delta_2$, and so we conclude inductively that $L_1 \cap L_2$ is infinite. From this and from the

primitiveness of γ_1 and γ_2 it follows that γ_1 and γ_2 are conjugates, i.e. there exist words σ and ρ such that

$$\gamma_1 = \sigma \rho \quad \text{and} \quad \gamma_2 = \rho \sigma. \quad (3)$$

Now, we show that

$$u = \sigma \quad \text{and} \quad \beta_1 u = \beta_2. \quad (4)$$

Since $L_1 \cap L_2$ is infinite we may assume in (2) that t and r are arbitrarily large. So, by the form of γ_1 and γ_2 , the equality $\beta_1 \gamma_1^t u = \beta_2 \gamma_2^r$ implies that $u \in (\sigma \rho)^* \sigma$. Moreover, since $\delta_1 = u \delta_2$ and δ_2 does not contain the word $\sigma \rho = \gamma_1$ as a prefix, we conclude that $u = \sigma$. Now, the equality $\beta_1 u = \beta_2$ follows from the first equality of (1) since the triples $(\beta_i, \gamma_i, \delta_i)$ are reduced.

This completes the proof of Lemma 5. Indeed, the equality

$$\beta_1 \gamma_1^* \delta_1 = \beta_2 \gamma_2^* \delta_2 \quad \text{is a trivial consequence of the second equation of (1), (3) and (4).}$$

Now, we are ready for

THEOREM 3. Each language L over a binary alphabet has a test set of the cardinality at most three.

Proof. Let $L \subseteq \{0, 1\}^*$. If L contains two words with different ratios, then these two words constitute a test set, since no equality set different from L^* can contain two words of different ratios. So we assume that all the words of L ^{have} the same ratio.

By the definition of r -primitiveness, it is clear that each word x in $\{0, 1\}^*$ possesses a unique decomposition in the form

$x = x_1 \dots x_q$ where each x_i is r -primitive and $r(x) = r(x_i)$ for $i = 1, \dots, q$. We define L_r to be the language which contains exactly those r -primitive words which occur in the above mentioned decompositions when x ranges over L . Clearly, any pair of morphisms agrees on L if and only if it agrees on L_r , i.e. any test set for L_r is a test set for L and vice versa. Therefore it is enough to show that L_r has a test set containing no more than three words.

First we observe that if L_r contains less than three words we are trivially done. So assume that the cardinality of L_r is at least three. We choose a three-element subset of L_r as follows. Let z_1 and z_2 be arbitrary two words from L_r . If they belong to a language of the form $(**)$ (see Section 3), then, by lemma 5, they determine this language uniquely. Let L_{z_1, z_2} be this language (assuming that it exists). Now if $L_r \not\subseteq (L_{z_1, z_2})^*$, then we choose z_3 such that $z_3 \in L_r - (L_{z_1, z_2})^*$. Otherwise z_3 is an arbitrary word of L_r different from z_1 and z_2 . We claim that $\{z_1, z_2, z_3\}$ is a test set for L_r . We consider **different kinds of pairs** of morphisms separately.

I Both of the morphisms are periodic. Now, by Theorem 1, any one-element set, and hence also $\{z_1, z_2, z_3\}$, tests whether such morphisms agree on L_r (remember that all word of L_r have the same ratio).

II One of the morphisms is periodic and the other is not. In this case Theorem 1 guarantees that any two-element subset of L_r tests whether such morphisms agree on L_r .

III Both of the morphisms are injective. We have two sub-cases.

(i) The equality set of the morphisms is generated by at most two words. Now, the conclusion of case II is valid when instead of two-element sets three-element sets are considered.

(ii) The equality set of the morphisms is of the form

$(uw^*v)^*$ for some reduced triple (u, w, v) . If $uw^*v = L_{z_1, z_2}$ then, by the choice of z_3 , the set $\{z_1, z_2, z_3\}$ tests whether two morphisms of the considered kind agree on L_r . If, on the other hand, $uw^*v \neq L_{z_1, z_2}$ then, by lemma 5, z_1 and z_2 both can not be in uw^*v , and so also in this case $\{z_1, z_2, z_3\}$ tests whether the morphisms considered now agree on L_r .

Since the classification I-III is exhaustive, $\{z_1, z_2, z_3\}$ is a test set for L_r , and therefore our proof for Theorem 3 is complete.

We want to finish this section with the following remarks.

Of course, a test set for an arbitrary language can not exist effectively, in general. However, our proof for Theorem 3 shows that if a family \mathcal{L} of languages satisfies the following three conditions, then a test set for each L in \mathcal{L} can be effectively found. Moreover, the cardinality of a test set is always at most three. The conditions are:

- (i) Each L in \mathcal{L} is recursively enumerable.
- (ii) Given L in \mathcal{L} and a regular language of the form $(uw^*v)^*$ for some words u, w and v , it is decidable whether $(uw^*v)^* \subseteq L$.
- (iii) Given L in \mathcal{L} , it is decidable whether all words of L has the same ratio.

We give two examples of the families satisfying the above conditions.

As shown in [1] each context free language, cf. [9], has effectively a test set. However, according to that proof a test set is quite large. In the case of binary context free languages, i.e. when languages are over a binary alphabet, we have a sharper result.

COROLLARY 1. Each binary context free language has effectively a test set of the cardinality at most three.

Proof. Clearly, conditions (i) - (iii) are satisfied for binary context free languages, (iii) being based on the fact that the Parikh image of a context free language is effectively semilinear, cf. [9].

As another example we consider so-called HDTOL languages, cf. [2], which are defined as follows. Let h_1, \dots, h_k and h be morphisms of a finitely generated free monoid Σ^* and x an element of Σ^+ . The languages of the form $\{h_1^{i_1} \dots h_k^{i_k} (x) \mid s \geq 0, i_j \in \{1, \dots, k\}\}$ are called HDTOL languages. Such a language is called binary if h is into a binary alphabet. We have the result.

COROLLARY 2. Each binary HDTOL language has effectively a test set of the cardinality at most three.

Proof. Now, condition (i) is trivial, condition (ii) is a known fact, cf. [2], and condition (iii) is a simple exercise on rational formal power series, cf. [7].

REFERENCES

- [1] J. Albert, K. Culik II and J. Karhumäki, Test sets for context free languages and algebraic systems of equations, Research Report CS-81-16, Department of Computer Science, University of Waterloo, Canada, 1981.
- [2] K. Culik II, A purely homomorphic characterization of recursively enumerable sets, JACM 26 (1979), 345-350.
- [3] K. Culik II and I. Fris, The decidability of the equivalence problem for DOL-systems, Information and Control 35 (1977), 20-35.
- [4] K. Culik II and J. Karhumäki, On the equality sets for homomorphisms on free monoids with two generators, R.A.I.R.O. Theoretical Informatics 14 (1980), 349-369.
- [5] K. Culik II and J. Karhumäki, Systems of equations over a free monoid and Ehrenfeucht conjecture, Research Report CS-81-15, Department of Computer Science, University of Waterloo, Canada, 1981.
- [6] K. Culik II and A. Salomaa, Test sets and checking words for homomorphism equivalence, JCSS 20 (1980), 379-396.
- [7] A. Ehrenfeucht, J. Karhumäki and G. Rozenberg, The (generalized) Post correspondence problem with lists consisting of two words is decidable, Theoret. Comput. Sci., to appear.
- [8] J. Engelfriet and G. Rozenberg, Fixed point languages, equality languages and representation of recursively enumerable languages, JACM 27 (1980) 493-518.
- [9] M. Harrison, "Introduction to Formal Language Theory," Addison-Wesley, Reading, Massachusetts, 1978.
- [10] M. Karpinski, ed., New Scottish Book of Problems, in preparation.
- [11] E. Post, A variant to a recursively unsolvable problem, Bull. of the Am. Math. Soc. 52 (1946), 264-268.
- [12] G. Rozenberg and A. Salomaa, "The Mathematical Theory of L Systems," Academic Press, New York, 1980.
- [13] A. Salomaa, Equality set for homomorphisms of free monoids, Acta Cybernetica 4 (1978), 127-139.
- [14] A. Salomaa and M. Soittola, "Automata-Theoretic Aspects of Formal Power Series," Springer-Verlag, Berlin, 1978.