Theoretical Computer Science 29 (1984) 285–308
North-Holland

THE EHRENFEUCHT CONJECTURE: A COMPACTNESS CLAIM FOR FINITELY GENERATED FREE MONOIDS

Juhani KARHUMÄKI

Department of Mathematics, University of Turku, SF-20500 Turku 50, Finland

Communicated by A. Salomaa
Received June 1983

Abstract. We survey recent results on the so-called Ehrenfeucht Conjecture which states: For each language L over a finite alphabet Σ there exists a finite subset F of L such that for each pair (g, h) of morphisms on Σ^* the equation $g(x) = h(x)$ holds for all x in L if and only if it holds for all x in F.

We point out that the conjecture is closely related to the theory of equations in free monoids. We also state a surprising consequence of the conjecture: If it holds (even noneffectively) for all D0L languages, then the HD0L sequence equivalence problem is decidable. Furthermore, we give examples of when the conjecture is known to hold. In particular, we establish it for all binary languages, as well as for all languages when attention is restricted to bounded delay morphisms of some fixed delay.

1. Introduction

In recent years there has been much research on morphisms of free monoids (cf. [6, 33]). Many of these problems are important not only in formal language theory but also in computability theory as well as in the theory of equations in free monoids.

Typical examples are morphic representation results for languages (cf. [33]), the Post Correspondence Problem (cf. [15]), the D0L sequence equivalence problem (cf. [7]) and the Ehrenfeucht Conjecture. Common to all these problems is that they are simple to formulate but difficult to solve, that is to say: they are mathematically challenging.

Our goal here is to survey recent results obtained on the Ehrenfeucht Conjecture. The conjecture was posed by Ehrenfeucht at the beginning of the 1970s and it is as follows.

Ehrenfeucht Conjecture. *For each language L over a finite alphabet Σ there exists a finite subset F of L such that for each pair (g, h) of morphisms on Σ^* the equation $g(x) = h(x)$ holds for all x in L if and only if it holds for all x in F.*

Basically, the conjecture is a statement on the topic "an infinite entity is equivalent to its finite subpart". Such, in a sense, compactness results are of course very important in all areas of mathematics, and not in the least in the areas which are connected with theoretical computer science.

The Ehrenfeucht Conjecture has a very natural interpretation in terms of equations in free monoids. We discuss this in Section 3 where we also state a result which emphasizes the importance of the conjecture from a purely algebraic point of view. Namely we show (cf. [9]) that the conjecture is equivalent to the following statement: Each system of equations over a finitely generated free monoid and with a finite number of variables has an equivalent finite subsystem.

Consequently, the Ehrenfeucht Conjecture is not only an important property of formal languages but also a fundamental property of free monoids.

As another example of connections of the Ehrenfeucht Conjecture to other problems we consider in Section 4 the so-called HD0L sequence equivalence problem (cf. [6]). The previously mentioned D0L sequence equivalence problem was for many years one of the most beautiful open problems within the theory of formal languages, until it was solved in [7] (cf. also [17]). The HD0L sequence equivalence problem is a natural generalization of the D0L sequence equivalence problem, and it has turned out that neither the method of [7] nor that of [17] applies to this generalized problem.

The decidability status of the HD0L sequence equivalence problem is still open. A surprising connection was, however, found in [9]: If the Ehrenfeucht Conjecture holds (even noneffectively) for all D0L languages, then the HD0L sequence equivalence problem is decidable. As regards some other generalizations of the D0L sequence equivalence problem, such as the DT0L sequence equivalence problem (cf. [6]) the situation is similar.

In Section 5 we turn to consider when the Ehrenfeucht Conjecture is known to hold. We give a simple proof for the conjecture in the case of regular (often called rational) languages. For context-free languages the proof is much more complicated (cf. [1]). Supports of formal power series form another generalization of regular languages for which the Ehrenfeucht Conjecture is known to hold (cf. [28]).

As we have already hinted, it is not known whether or not the conjecture holds for all D0L languages. However, in [10] a partial result in this direction was established when the conjecture was proved for all positive D0L languages, i.e., for D0L languages generated by D0L systems satisfying: For each pair (a, b) of letters, a occurs as a subword in $h(b)$, where h is the morphism of the system. As other examples we mention that in [2] the conjecture was proved for all commutatively closed languages, and in [10] a sufficient condition for a language L was given to guarantee that the conjecture holds for L.

Section 6 is devoted to establishing the Ehrenfeucht Conjecture for all languages over a binary alphabet. This result was originally proved in [13], and a shorter proof was given later in [14]. Moreover, the methods of the latter proof made it possible to show that for each binary language L a finite subset F can be chosen to contain

no more than three words, and that such an F can be found effectively if L satisfies certain relatively mild conditions.

In Section 7 we take a different approach to shed more light on the problem. Without restricting the language we now put restrictions on morphisms. We prove (cf. [14]) that the conjecture holds for all languages if attention is restricted to morphisms having a bounded delay equal to a fixed nonnegative integer p.

Finally, we point out, that, as a survey, this article contains neither any essentially new results nor detailed proofs of all results presented here. On the other hand, our purpose is to give a few detailed (and sometimes slightly modified) proofs to interest the reader not only in the results, but also in the problem as a whole.

2. Preliminaries

We assume the reader to be familiar with the basic facts concerning formal languages and free monoids (cf., e.g., [19, 24, 25]). Consequently, the following definitions are given mainly to fix our notation. Some terminology and notions are defined in the appropriate places when they are needed.

Let Σ be a finite alphabet. We denote by Σ^* the free monoid generated by Σ and by 1 its identity. We set $\Sigma^+ = \Sigma^* - \{1\}$. Elements of Σ^* are words, in particular, 1 is called the empty word. For a word x we denote by $|x|$ its length and by $|x|_a$ the number of times a occurs in x. If Σ is an ordered alphabet $\{a_1, a_2, \ldots, a_t\}$, then $\psi(x) = (|x|_{a_1}, \ldots, |x|_{a_t})$ is called the Parikh vector of x. The notation $\mathrm{pref}_k(x)$ for $k \geqslant 0$ is used to denote the longest prefix of x with length at most k.

For two words x and y the notation x pref y means that x is a prefix of y and the notation x Pref y means that either x pref y or y pref x holds. We say that x is a subword of y if there exist words z and z' such that $y = zxz'$. The left (resp. right) quotient of x by y is denoted by $y^{-1}x$ (resp. xy^{-1}). For the maximal common prefix of words x and y we use the notation $x \wedge y$. Clearly, the operation $\wedge$ is associative.

Subsets of Σ^* are called languages. For a language L the notation $\mathrm{pref}(L)$ denotes the set of all prefixes of words in L. Further, for two languages L and K their quotients are defined in a natural way via quotients of words.

The central notion of our studies is that of a morphism h from a free monoid Σ^* into some other free monoid Δ^*, in symbols $h: \Sigma^* \to \Delta^*$. In most of the problems considered here we may assume $\Sigma = \Delta$. We call a morphism *binary* if Σ is binary and *periodic* if there exists a word z such that $h(\Sigma) \subseteq z^*$. Further we say that a morphism $h: \Sigma^* \to \Delta^*$ has a *bounded delay* p (from left to right) if for all words u and v in Σ^* and all letters a and b in Σ we have: If $h(au)$ pref $h(bv)$ and $|u| \geqslant p$, then necessarily $a = b$. If h has bounded delay 0, then it is said to be a *prefix*, and if it has a bounded delay p, for some p, then it is said to have a *bounded delay*.

Let g and h be two morphisms on Σ^* and L a language over Σ^*. We say that g and h are *equivalent* on L or *agree* on L (in symbols, $g \equiv^L h$) if the equation $g(x) = h(x)$ holds for all x in L. Further we say that a word x is *morphically forced*

by a language L if whenever two morphisms agree on L they agree on x, too. Finally, a finite subset F of a language L is called a *test set* for L if for each pair (g, h) of morphisms the relation $g =^L h$ holds if and only if the relation $g =^F h$ holds. Let $\mathscr{L}$ be a family of languages. The *morphism equivalence problem* for $\mathscr{L}$ is to decide whether or not two given morphisms are equivalent on a given language L of $\mathscr{L}$.

With the above terminology the Ehrenfeucht Conjecture can be restated as follows.

Ehrenfeucht Conjecture. *Each language over a finite alphabet possesses a test set.*

For a pair (g, h) of morphisms on Σ^* we define their *equality language* $E(g, h)$ to be

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

This notion was explicitly defined in [17] but used implicitly already in [7]. It has turned out to be very important. Indeed, all the problems mentioned at the very beginning of the previous section are connected to equality languages. For example, the famous Post Correspondence Problem is nothing but a question of asking for an algorithm to decide whether or not for two given morphisms g and h the relation $E(g, h) = \{1\}$ holds. As regards our current problem, the Ehrenfeucht Conjecture, it can be formulated by using equality languages as follows:

$$\forall L \subseteq \Sigma^*: \exists F \subseteq L, F \text{ is finite: } \forall (g, h) \text{ morphisms:}$$

$$L \subseteq E(g, h) \Leftrightarrow F \subseteq E(g, h).$$

Throughout this paper equality languages play an important role. Therefore we finish this section with a few examples and results concerning them.

Example 2.1. Let g and h be morphisms defined on $\{a, b\}^*$ by $g(a) = a = h(b)$ and $g(b) = aa = h(a)$. Then, clearly,

$$E(g, h) = \{x \in \{a, b\}^* \mid |x|_a : |x|_b = 1\} \cup \{1\}.$$

Example 2.2. Let g and h be morphisms defined as $g(a) = aab$, $g(b) = a$, $h(a) = a$ and $h(b) = baa$. In searching for elements of $E(g, h)$ we have to start with a, i.e., we have the following 'domino piece':

$g(a)$:	a	a	b
$h(a)$:	a		

We can continue only with a again yielding the following situation:

$g(aa)$:	a	a	b	a	a	b
$h(aa)$:	a	a				

Now, the only way to continue is to take b:

$g(aab)$:	a	a	b	a	a	b	a
$h(aab)$:	a	a	b	a	a		

Again we have to continue with b:

$g(aabb)$:	a	a	b	a	a	b	a	a
$h(aabb)$:	a	a	b	a	a	b	a	a

So we have found an element in $E(g, h)$ and, actually, our procedure shows that $E(g, h) = (aabb)^*$.

Example 2.3. Let g and h be morphisms defined as $g(a) = abab$, $g(b) = a$, $g(c) = ba$, $h(a) = ab$, $h(b) = a$ and $h(c) = baba$. Then $E(g, h) = \{a^n bc^n \mid n \geq 0\}^*$.

Example 2.4. Consider the morphisms g and h defined by the following table:

	a	b	c	d	e	f
g	$abcd$	$bcbc$	d	bd	cb	e
h	a	bc	d	db	$cbcb$	$dcbe$

Now, it is straightforward to see that

$$E(g, h) = (\{abcb^2c \ldots cb^{2^n}de^{2^n}c \ldots ce^2cef \mid n \geq 0\} \cup \{c\})^*.$$

Observe that both morphisms g and h are injective; in fact, g is the prefix and h the so-called suffix.

As our final example we state a result which shows the real generating power of equality languages. To be able to state the result we need one more notion. Let $e(g, h)$ denote the basis of an equality language $E(g, h)$, i.e.,

$$e(g, h) = (E(g, h) - \{1\}) - (E(g, h) - \{1\})^2.$$

Theorem 2.5. *For each recursively enumerable language L there exist morphisms g, h and π (defined on suitable free monoids) such that $L = \pi(e(g, h))$.*

Theorem 2.5 was proved in [5]; for similar results the reader is referred to [18] and [32].

Theorem 2.5 provides an explanation of why the Ehrenfeucht Conjecture, as well as the other previously mentioned problems involving morphisms, seems to be so difficult.

3. The Ehrenfeucht Conjecture and systems of equations

Let us consider the Ehrenfeucht Conjecture for the language $L=\{a^nb^n \mid n\geq 0\}$. Clearly, for each pair (g, h) of morphisms we have

$$g(ab)=h(ab)$$
$$\text{iff } g(a)g(b)=h(a)h(b)$$
$$\text{iff } (g(a), g(b), h(a), h(b)) \quad \text{is a solution of the equation } xy=uv.$$

Consequently, we are interested in all solutions of the following infinite system of equations:

$$S:\ x^ny^n=u^nv^n, \quad n\geq 1,$$

and the conjecture claims that the set of all such solutions is exactly the same as the set of all solutions of the system of equations defined by

$$S':\ x^ny^n=u^nv^n, \quad n\in\{n_1,\ldots n_t\} \text{ for some } t\geq 0.$$

Actually, it is not difficult to see that the above t can be chosen to be 2 and that n_1 and n_2 may be arbitrary unequal natural numbers. Therefore, e.g., $\{ab, aabb\}$ is a test set for L.

The above considerations show that the Ehrenfeucht Conjecture can be interpreted in a natural way in terms of equations of free monoids. To do this we call an equation $u=v$ *symmetric* if the sets X_u and X_v of variables of u and v, respectively, are disjoint and, moreover, there exist an isomorphism $\nu: X_u^*\to X_v^*$ such that $v=\nu(u)$. Using this notation the Ehrenfeucht Conjecture can be stated as: Each system of equations over a finitely generated free monoid containing only symmetric equations with the same isomorphism ν, and a finite number of variables is equivalent to its finite subsystem.

From this observation a very natural question arises. Is the special form of the equations important in the conjecture? Before giving an answer to this question we introduce, following [9], some terminology.

Let Σ be a finite alphabet and N a finite set such that $\Sigma\cap N=\emptyset$. A *system S of equations* over Σ with variables N is a binary relation $S\subseteq (N\cup\Sigma)^*\times(N\cup\Sigma)^*$. A pair (u, v) in S represents the equation $u=v$ and any word of Σ^+ occurring in u or v is called a *constant* of the equation. A *solution* of an equation $u=v$ is nothing but a morphism $h:(N\cup\Sigma)^*\to\Sigma^*$ such that $h(a)=a$ for all a in Σ and $h(u)=h(v)$. Of course, a morphism is a solution of a system of equations if it is a solution of all of its equations, and two systems are *equivalent* if they have exactly the same solutions. S' is called a *subsystem* of S if $S'\subseteq S$.

We say that a system S of equations is *rational* if the relation S is a rational subset of $(N\cup\Sigma)^*\times(N\cup\Sigma)^*$ (cf. [3]). By the well-known Nivat's theorem a subset S is rational if and only if there exist an alphabet Δ, a rational (or regular) language $L\subseteq\Delta^*$ and two morphisms g and h on Δ^* such that $S=\{(g(x), h(x)) \mid x\in L\}$. With this characterization as a motivation we say that a family $\mathcal{R}\subseteq(N\cup\Sigma)^*\times(N\cup\Sigma)^*$ of relations is *morphically characterized* by a family $\mathcal{L}$ of languages if the following

holds: $R \in \mathcal{R}$ if and only if there exist an alphabet Δ, a language $L \subseteq \Delta^*$ in $\mathcal{L}$ and two morphisms $g, h: \Delta^* \to (N \cup \Sigma)^*$ such that $R = \{(g(x), h(x)) \mid x \in L\}$, or briefly $R = [g, h]L$. By taking $\mathcal{L}$ equal to the family of algebraic languages we obtain the family of *algebraic relations*, which is known to coincide with the family of relation determined by pushdown transducers (cf. [3]).

We have the following obvious result.

Lemma 3.1. *The family of all binary relations $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ is morphically characterized by the family of all languages.*

Let $\mathcal{L}$ be a family of languages. We say that a system $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ of equations is of *type* $\mathcal{L}$ if it belongs to the family of relations morphically characterized by $\mathcal{L}$, i.e., $S = [g, h]L$ for some L in $\mathcal{L}$ and some morphisms g and h. This terminology is justified by Lemma 3.1 and the discussion before it.

Now, finally, we are ready to answer our question. The answer is affirmative, and, moreover, we are able to show the correspondence between the existence of a test set for languages of type $\mathcal{L}$ and the existence of an equivalent finite subsystem for systems of equations of type $\mathcal{L}$.

Theorem 3.2. *Let $\mathcal{L}$ be a family of languages. The following two conditions are equivalent:*

(i) *For each language L in $\mathcal{L}$ there exists a test set.*

(ii) *For each system of equations of type $\mathcal{L}$ there exists an equivalent finite subsystem.*

Proof. We prove the theorem in the case when $\mathcal{L}$ is the family of all languages and equations do not contain any constants. The main differences between this special case and the general case are notational difficulties (cf. [9]).

Following our considerations at the beginning of this section if every system S of equations without constants has an equivalent finite subsystem, then every language has a test set. So it is enough to prove the implication (i)$\Rightarrow$(ii).

Let $u = v$ be an equation containing the variables $\{x_1, \ldots, x_n\}$ and no constants. Let $X = \{x_1, \ldots, x_n\}$ and $\bar{X} = \{\bar{x}_1, \ldots, \bar{x}_n\}$ such that $X \cap \bar{X} = \emptyset$. We associate to the equation $u = v$ the word $u\bar{v}$, where $\bar{v}$ is obtained from v by replacing each variable x of v by $\bar{x}$. Now, a morphism $h: X^* \to \Sigma^*$ is a solution of the equation $u = v$ if and only if the morphisms $g, \bar{g}: (X \cup \bar{X})^* \to \Sigma^*$ defined by

$$g(x) = h(x) \quad \text{for } x \in X,$$

$$g(\bar{x}) = 1 \quad \text{for } \bar{x} \in \bar{X},$$

and

$$\bar{g}(x) = 1 \quad \text{for } x \in X,$$

$$\bar{g}(\bar{x}) = h(x) \quad \text{for } \bar{x} \in \bar{X},$$

agree on the word $u\bar{v}$. From this the implication (i)$\Rightarrow$(ii) follows. □

Theorem 3.2 remains true if the word 'exists' in (i) and (ii) is replaced by the words 'exists effectively'. As a consequence of Theorem 3.2 and Lemma 3.1 we can reformulate the Ehrenfeucht Conjecture once again.

Corollary 3.3. *The Ehrenfeucht Conjecture is equivalent with the statement: Each system of equations over a finitely generated free monoid and containing only a finite number of variables has an equivalent finite subsystem.*

Constants are allowed in the above equations. This, however, is not essential from the point of view of the Ehrenfeucht Conjecture since they can always be eliminated simply by replacing each occurrence of a letter a in the equations by a new variable X_a and by introducing new equations $X_a = a$. This leads us to the following important subproblem of the statement of Corollary 3.3.

Generalized Ehrenfeucht Conjecture with n variables (GEC(n) for short). *Every system of equations over a finitely generated free monoid containing n variables and no constants is equivalent to its finite subsystem.*

Theorem 3.4. GEC(2) *holds. Moreover, each system S of equations containing two variables and no constants is equivalent to its finite subsystem containing only two equations.*

Proof. Let $u = v$ be an equation with x and y as variables, i.e., $u, v \in \{x, y\}^+$. By the Defect Theorem (cf. [25]) we conclude that the general solution of the equation $u = v$ is of one of the following forms:

(1) $\{(x, y) | x, y \in \Sigma^*\}$,

(2) $\{(x, y) | \exists z \in \Sigma^*: x, y \in z^*\}$,

(3) $\{(x, y) | \exists z \in \Sigma^*: x, y \in z^* \text{ and } |x| : |y| = k\} \cup \{(1, 1)\}$ for some rational $k \geq 0$ or $k = \infty$,

(4) $\{(1, 1)\}$,

where we have the interpretation $n : 0 = \infty$ for $n > 0$.

Let K_1 and K_2 be sets of the above form. If both of them are not of form (3), then, clearly, $K_1 \cap K_2 \in \{K_1, K_2\}$. If, in turn, both of them are of form (3), then $K_1 \cap K_2 \in \{K_1, K_2, \{(1, 1)\}\}$. From these observations the theorem follows. □

In spite of the simple proof of Theorem 3.4 it is not known whether or not GEC(3) holds!

Theorem 3.4 can, however, be generalized to cover a subcase of the problem GEC(n). To do this we need some preliminary notions. Let N be the set of variables. For each equation $u = v$ we define first($u = v$) to be the set of rightmost variables of u and v. Further, for a system S of equations we define its graph G_S as follows. The set of nodes of G_S equals N, and for two nodes m and m' there exists an edge between them if and only if $\{m, m'\} = \text{first}(u = v)$ for some equation $u = v$ in S.

With the above terminology one can show, by using a result from [4] (cf. also [11]) and simple considerations from linear algebra, the following result.

Theorem 3.5. *Let S be a system of equations with n variables and no constants. If G_S is a complete n-element graph, then S is equivalent to its finite subsystem S'.*

If we are only interested in such solutions, where all the components are nonempty, then Theorem 3.5 holds under the assumption that G_S is a connected graph (cf. [11]).

We finish this section with the following related result of [20].

Theorem 3.6. *Each finite system S of equations over Σ^*, where Σ contains at least two letters, is equivalent to one equation alone.*

Proof. It is enough to show that a pair $u = v$ and $u' = v'$ is equivalent to a single equation. Let a and b be two different letters of Σ. Then we have

$$\left.\begin{matrix} u = v \\ u' = v' \end{matrix}\right\} \Leftrightarrow uau'ubu' = vav'vbv'$$

as is straightforward to see. □

Observe that Theorem 3.6 is not of the same nature as the Ehrenfeucht Conjecture since the new equation need not be among the original ones.

4. The Ehrenfeucht Conjecture and the HD0L sequence equivalence problem

In this section we establish a surprising consequence of the Ehrenfeucht Conjecture from [9], namely that it implies the decidability of the HD0L sequence equivalence problem (cf. [6]). This problem is as follows.

A *D0L sequence* is a sequence of words obtained by iterating a morphism starting at a given word, i.e., if h is the morphism and ω is the word, then they determine the D0L sequence $\omega, h(\omega), h^2(\omega), \ldots$, as well as the *D0L language* $\{h^n(\omega) \mid n \geq 0\}$. An *HD0L sequence* is a morphic image of a D0L sequence, i.e., if we map the above D0L sequence by a morphism f, we obtain an HD0L sequence $f(\omega), f(h(\omega)), f(h^2(\omega)), \ldots$. Consequently, both D0L and HD0L sequences are purely morphically defined.

The problem of deciding the equivalence of two given D0L sequences, usually referred to as the *D0L sequence equivalence problem*, was for many years a challenging open problem within the theory of formal languages. Finally, it was solved by Culik II and Fris [7]. Later a shorter proof was found in [17]. Both algorithms are very cumbersome, and only in the case that the sequences are over a binary alphabet, a simple algorithm is known (cf. [22]). It has also become clear that neither the algorithm of [7] nor that of [17] can be generalized to solve the equivalence of two HD0L sequences, i.e., the *HD0L sequence equivalence problem.*

The HD0L sequence equivalence problem is still an open problem. It was proved in [31] that this problem can be reduced to another famous open problem, namely the problem of determining whether or not a given Z-rational sequence contains 0 [cf. 34]. We give another similar reduction result.

We shall need a few auxiliary results, the first of which we believe is interesting in its own right.

Theorem 4.1. *The equivalence problem for finite systems of equations is decidable.*

The proof of Theorem 4.1 is based on the following two facts. First, as pointed out already in [8], inequality can be expressed by using a finite number of equalities, i.e., the relation $u \neq v$ is equivalent to the relation $u_1 = v_1 \vee u_2 = v_2 \vee \cdots \vee u_n = v_n$, where $\vee$ denotes disjunction and the words are chosen in a suitable way. Second, we use a deep result of Makanin [cf. 26], which states that it is decidable whether or not a given finite system of equations has a solution. Details can be found in [9].

As a corollary to Theorem 4.1 we obtain the following.

Theorem 4.2. *Given two finite languages L_1 and L_2 such that $L_1 \subseteq L_2$. Then it is decidable whether or not L_1 is a test set for L_2.*

Now we are ready for the next theorem.

Theorem 4.3. *For every* D0L *language L, the existence of a test set implies that it can be effectively found.*

Proof. Let $L = \{h^i(\omega) \mid i \geq 0\}$ be a D0L language. Further let $L_j = \{h^i(\omega) \mid i \leq j\}$. Since L possesses a test set, there exists a minimal $p > 0$ such that L_{p-1} morphically forces $h^p(\omega)$, i.e., for an arbitrary pair (f, g) of morphisms if

$$f(h^i(\omega)) = g(h^i(\omega)) \quad \text{for } i = 0, \ldots, p-1,$$

then

$$f(h^p(\omega)) = g(h^p(\omega)).$$

The minimal p can effectively be found since, by Theorem 4.2, we can test whether L_{j-1} is a test set for L_j. We now show that, for each $n \geq p$, the set L_{n-1} morphically forces the word $h^n(\omega)$, which means that L_{p-1} is a test set for L.

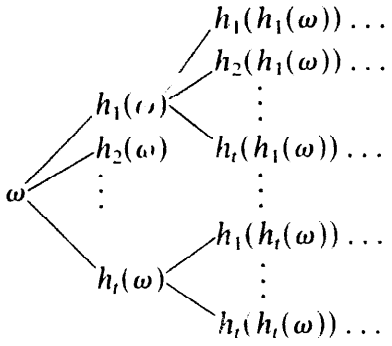
Assume that there exists an $N > p$ such that $h^N(\omega)$ is not morphically forced by L_{N-1}, that is to say, there exist morphisms α and β such that $\alpha(h^i(\omega)) = \beta(h^i(\omega))$ for $0 \leq i < N$ and $\alpha(h^N(\omega)) \neq \beta(h^N(\omega))$. Let $\gamma = \alpha h^{N-p}$ and $\delta = \beta h^{N-p}$. Then the morphisms γ and δ are equivalent on L_{p-1} but $\gamma(h^p(\omega)) = \alpha(h^N(\omega)) \neq \beta(h^N(\omega)) = \delta(h^p(\omega))$, a contradiction to the choice of p. □

The main theorem of this section now follows.

Theorem 4.4. *If the Ehrenfeucht Conjecture holds (even noneffectively) for* D0L *languages, then the* HD0L *sequence equivalence problem is decidable.*

Proof. As pointed out in [12] the HD0L sequence equivalence problem is equivalent to the morphism equivalence problem for D0L languages, the trick being a standard use of a 'barred alphabet'. Consequently, Theorem 4.4 follows from Theorem 4.3 since the effective existence of a test set for all languages in a family $\mathscr{L}$ surely implies the decidability of the morphism equivalence problem for $\mathscr{L}$. □

The above considerations can be generalized to cover so-called DT0L sequences as well. We recall (cf. [30]) that a *DT0L sequence* or *tree* is obtained by iterating a finite set of morphisms in all possible ways starting at a given word, i.e., if $h_1, \ldots, h_t$ are the morphisms and ω is the word, then they determine the DT0L sequence



The set of all words thus obtained is called a *DT0L language.* An *HDT0L sequence* is defined as a morphic image of a DT0L sequence. It has been proved in [29] that the problem of deciding whether two given DT0L languages coincide is undecidable. The same problem for DT0L sequences, i.e., the *DT0L sequence equivalence problem*, is open and, we believe, it should be decidable rather than undecidable. Moreover, it is known that the DT0L sequence equivalence problem and its generalization for HDT0L sequences are, from the point of view of their decidability status, equivalent (cf. [12]).

Corresponding to Theorem 4.4 we can prove (cf. [9]) the following theorem.

Theorem 4.5. *If the Ehrenfeucht Conjecture holds (even noneffectively) for all* DT0L *languages, then the* DT0L *sequence equivalence problem is decidable.*

In the next section we give a class of D0L languages for which the Ehrenfeucht Conjecture is known to hold.

5. The Ehrenfeucht Conjecture and families of languages

The Ehrenfeucht Conjecture in its full generality seems to be a very difficult problem; such a view is supported by the results of the previous section. One way

to obtain reasonable subproblems is to consider the conjecture only for some restricted classes of languages. This is our topic here.

First, we establish the conjecture for all regular (or rational) languages.

Lemma 5.1. *For all words $x, y, z, w, \bar{x}, \bar{y}, \bar{z}$ and $\bar{w}$ in Σ^* the following implication holds:*

$$\left.\begin{aligned} xy &= \bar{x}\bar{y} \\ xzy &= \bar{x}\bar{z}\bar{y} \\ xwy &= \bar{x}\bar{w}\bar{y} \end{aligned}\right\} \Rightarrow xzwy = \bar{x}\bar{z}\bar{w}\bar{y}.$$

Proof. Let $x = \bar{x}t$, and consequently, $\bar{y} = ty$. Then, by the second equation on the left side, we have $\bar{x}tzy = \bar{x}\bar{z}ty$, or equivalently, $tz = \bar{z}t$. Similarly, the third equation yields $tw = \bar{w}t$. Therefore, we obtain $xzwy = \bar{x}tzwy = \bar{x}\bar{z}twy = \bar{x}\bar{z}\bar{w}ty = \bar{x}\bar{z}\bar{w}\bar{y}$. The case $\bar{x} = xt$ is analogous, completing the proof. □

By Lemma 5.1 we easily obtain the following.

Theorem 5.2. *Let L be a regular language over Σ. The language $F = \{x \in L \mid |x| \leq 2\,\mathrm{card}(Q)\}$, where Q is the state set of a finite automaton $\mathcal{A}$ accepting L, is a test set for L.*

Proof. Lemma 5.1 can be interpreted as follows: Whenever two morphisms g and h agree on words uv, uqv and $uq'v$, they agree on the word $uqq'v$, too. Consequently, a set of words having a computation in $\mathcal{A}$ such that it does not pass any state more than twice is a test set for L, proving the theorem. □

Observe that automaton $\mathcal{A}$ in Theorem 5.2 need not be deterministic. Observe also that size of a test set is bounded by a relatively small constant depending on an automaton accepting the language.

The argument of the proof of Theorem 5.2 can be generalized for context-free languages as well. However, now the construction of a test set can be based on a context-free grammar generating the language rather than a pushdown automaton accepting it.

Theorem 5.3. *Let L be a context-free language over Σ generated by a context-free grammar G without erasing productions. Then the language $L = \{x \in L \mid |x| \leq m^{4n+1}\}$ is a test set for L, where n is the cardinality of the terminal alphabet of G and m is the length of the longest right-hand side of the productions of G.*

The proof of Theorem 5.3 is a rather complicated manipulation of equations and can be found in [1].

As we have already mentioned, the effective existence of a test set for all languages in a family $\mathscr{L}$ implies that the morphism equivalence problem is decidable for $\mathscr{L}$. So we obtain a result of [12].

Theorem 5.4. *The morphism equivalence problem is decidable for context-free languages.*

Theorem 5.4 means that for a given context-free language L and for two given morphisms g and h the question "Is $g(x)=h(x)$ for all x in L?" is decidable. It is instructive to notice that a related question "Is $g(L)=h(L)$?" is undecidable (cf. [12]).

Recalling our main theorem from Section 3, we obtain another corollary of Theorem 5.4.

Theorem 5.5. *Each algebraic system of equations possesses, effectively, an equivalent finite subsystem.*

To be able to state another generalization of Theorem 5.2 we need some terminology. Let Σ be a finite alphabet and k a field. A mappings $s: \Sigma^* \to k$ is called a *formal power series* over k. The *support* of s is the language $\{x \in \Sigma^* \mid s(x) \neq 0\}$. A formal power series s is called *rational* if it is obtained from polynomials, i.e., formal power series with finite supports, by applying rational operations union, product and quasi-inverse finitely many times (cf. [34]).

It is well known that the family of supports of rational formal power series contains all regular languages but is incomparable to the family of context-free languages (cf. [34]). Hence, the following result of [28] provides another way to generalize Theorem 5.2.

Theorem 5.6. *Each support of rational formal power series possesses a test set.*

As we pointed out in the previous section it would be important to know whether the Ehrenfeucht Conjecture holds for all DOL languages. This question becomes even more interesting when one notices that both the conjecture and the notion of a DOL language are defined by only using morphisms. Although we do not know the answer to this question we have a partial result in this direction. We call a DOL language $L=\{h^i(\omega) \mid i \geq 0\}$ *positive* if for each pair (a, b) of letters, a occurs in $h(b)$.

Theorem 5.7. *Each positive* DOL *language possesses a test set.*

The proof of Theorem 5.7 can be found in [10]. It is very long and does not appear to be extendable to all DOL languages.

We conclude this section by mentioning two families of languages for which the Ehrenfeucht Conjecture is known to hold. Common to these results is that the

considered languages have to satisfy some quite strong structural properties. First, we call a language $L \subseteq \Sigma^*$ *commutatively closed* if whenever $x \in L$ and $\psi(x) = \psi(y)$, then also $y \in L$. For commutatively closed languages we have a positive result (cf. [2]).

Theorem 5.8. *Each commutatively closed language has a test set of the cardinality at most* $2^n(n!+n)+5n^2$, *where* n *is the cardinality of the alphabet.*

In order to be able to state our second result we again need some terminology. Let L be a language over an alphabet $\{a_1, \ldots, a_t\}$. We define a language $\mathrm{sp}(L)$ by setting

$$\mathrm{sp}(L) = \psi^{-1}(\langle \psi(L) \rangle \cap \mathbb{N}^t),$$

where ψ denotes the Parikh mapping, $\langle \psi(L) \rangle$ denotes the vector space (over rationals) generated by $\psi(L)$ and $\mathbb{N}$ denotes the set of nonnegative integers. Now, our central notion is that of the *deviation* of a word w with respect to a language L, in symbols $d_L(w)$. It is defined by the formula

$$d_L(w) = \mathrm{Min}\{z \in \mathbb{N}^t \mid \psi(w) \in \psi(\mathrm{sp}(L)) + z\},$$

where Min refers to minimality with respect to the usual $\leq$-relation in $\mathbb{N}^t$. Using this notion we define a structural property of languages. We say that L has a *bounded prefix deviation* if there exists a constant C such that for all words w in $\mathrm{pref}(L)$ we have

$$\min\{|z_1| + \cdots + |z_t| \mid (z_1, \ldots, z_t) \in d_L(w)\} \leq C,$$

where min refers to the minimal element of a set of numbers.

Intuitively, L has a bounded prefix deviation, if every prefix of words in L is 'close' to a word having the 'numerical properties' of some of the words in L.

We still need another structural property of languages. We say that L has a *fair distribution* of letters if there exists a constant q such that whenever u is a subword of a word in L with length at least q, then u contains all letters of the alphabet.

Now, we can state the following theorem.

Theorem 5.9. *Every language* L *with a bounded prefix deviation and a fair distribution of letters has a test set.*

The proof of this result, as well as more discussion on the required notions, can be found in [10] (cf. also [11]).

6. The Ehrenfeucht Conjecture in the binary case

This section is devoted to giving a solution to the Ehrenfeucht Conjecture in the case of a binary alphabet. This problem was first solved in [13], and later a simpler

proof was found in [16]. Our presentation here is a slight modification of the proof in [16]. The whole section is strongly connected and based on a result characterizing equality languages of binary morphisms.

We start with some terminology. Throughout this section let $\Sigma = \{a, b\}$. We define the *ratio* of a nonempty word w, in symbols $r(w)$, by $r(w) = |w|_a : |w|_b$ and say that w is *ratio-primitive* if none of its proper prefixes has the same ratio as the whole word w. Clearly, each ratio-primitive word is *primitive* in the ordinary sense, i.e., is not a proper power of any word.

Our starting point is the following simple result (cf. [8]), which is based on length considerations only. We again use the interpretation $n:0=\infty$ for any number n different from 0.

Theorem 6.1. *Let g and h be periodic binary morphisms. Then $E(g, h) = \{1\}$ or $E(g, h) = \{1\} \cup \{x \in \Sigma^+ \mid r(x) = k\}$ for some rational $k \geq 0$ or $k = \infty$.*

In the case of nonperiodic morphisms a crucial fact is the following simple result, which, we believe, is also interesting in its own right. This lemma was first formulated in [23]. By a *periodic* subset X of Σ^* we mean, of course, a subset X satisfying $X \subseteq z^*$ for some word z in Σ^*.

Lemma 6.2. *Let $X = \{x, y\}$ be an aperiodic subset of Σ^*. Then we have the following implication:*

$$\left.\begin{array}{l} u \in xX^*, \quad |u| \geq |xy| - 1 \\ v \in yX^*, \quad |v| \geq |xy| - 1 \end{array}\right\} \Rightarrow u \wedge v = xy \wedge yx.$$

The proof of Lemma 6.2 is straightforward and does not require the fact that Σ is binary. We leave it to the reader. The important message of the lemma is that $u \wedge v$ is independent of the pair (u, v), i.e., a characteristic constant of X.

Before stating our characterization result we need another notion. We call a triple (β, γ, δ) of words *reduced* if γ is primitive and it is neither a suffix of β nor a prefix of δ. Now, we are ready for the next theorem.

Theorem 6.3. *Let g and h be binary morphisms such that at least one of them is aperiodic. Then the equality language $E(g, h)$ is either of the form $\{\beta, \gamma\}^*$ for some words β and γ in Σ^* or of the form $(\beta\gamma^*\delta)^*$ for some reduced triple (β, γ, δ).*

Proof. If one of the morphisms g and h is periodic and the other is not, then $E(g, h) = \beta^*$ for some possibly empty word β. This can be seen, e.g., as a simple consequence of the Defect Theorem [25].

So we assume that both g and h are aperiodic. Let $\alpha_g = g(ab) \wedge g(ba)$ and $\alpha_h = h(ab) \wedge h(ba)$. By Lemma 6.2, $|\alpha_g| < |g(ab)|$ and $|\alpha_h| < |h(ab)|$.

We call the elements of $E(g, h)$ *solutions* and elements of $\mathrm{pref}(E(g, h))$ *presolutions*. Furthermore, for each presolution w we define *w-overflow*, in symbols $o(w)$, by

$$o(w) = \begin{cases} (h(w)^{-1}g(w), g) & \text{if } |g(w)| \geq |h(w)|, \\ (g(w)^{-1}h(w), h) & \text{if } |g(w)| < |h(w)|. \end{cases}$$

A pair in $\Sigma^* \times \{h, g\}$ is called an *overflow* if it is a w-overflow for some presolution w in $\mathrm{pref}(E(g, h))$. Finally, we say that a presolution w, or a w-overflow $o(w)$, *admits an a-continuation*, for a in Σ, if wa is a presolution, too. It is clear that for any two presolutions w and w' satisfying $o(w) = o(w')$ they admit a-continuations simultaneously and therefore we may talk about overflows admitting a-continuations. Of course, all the above notions are defined with respect to a given pair (g, h).

Let w be a presolution such that $o(w) = (z, h)$; the case when $o(w) = (z, g)$ is symmetric. In the spirit of Example 2.2 in Section 2 we want to continue w and we are especially interested in the case when w admits two different continuations, i.e., both a- and b-continuations. We assume that this is the case, which means that there exist words w_a and w_b such that waw_a and wbw_b are solutions. Since $E(g, h)$ is closed under the catenation we may assume that

$$(1) \qquad \begin{cases} \min\{|g(aw_a)|, |g(bw_b)|\} > |g(ab)|, \\ \min\{|h(aw_a)|, |h(bw_b)|\} > |h(ab)|. \end{cases}$$

We also have

$$(2) \qquad zh(aw_a) = g(aw_a) \quad \text{and} \quad zh(bw_b) = g(bw_b),$$

and moreover, by Lemma 6.2 and (1),

$$(3) \qquad g(aw_a) \wedge g(bw_b) = \alpha_g \quad \text{and} \quad h(aw_a) \wedge h(bw_b) = \alpha_h.$$

Therefore, under our assumption, it follows from (1), (2) and (3) that the word z must be equal to $\alpha_g \alpha_h^{-1}$ (cf. also Fig. 1).

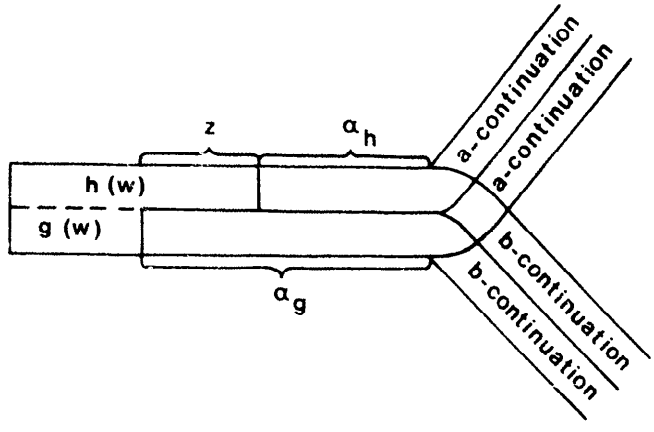


Fig. 1.

What we have proved is that there exists at most one overflow with two different continuations. We call such an overflow *critical*, and complete the proof of the theorem as follows.

First, if no critical overflow exists, then any presolution (including the empty word) can be continued in at most one way, and hence $E(g, h) = \beta^*$ for some possibly empty word β. If the unique critical overflow exists and is equal to $(1, g)$, then the same argument applies to all the presolutions which are not solutions showing that $E(g, h) = \{\beta, \gamma\}^*$ for some possibly empty words β and γ.

If, finally, the critical overflow is (α, g), or symmetrically (α, h) where $\alpha \neq 1$, we proceed as follows. Let w be a presolution such that $o(w) = \alpha$. We call a letter a *repetitive* (with respect to (α, g)) if there exists a word $\bar{w}_a$ such that

$$\alpha g(a\bar{w}_a) = h(a\bar{w}_a)\alpha \quad \text{and} \quad E(g, h) \cap \text{pref}(wa\bar{w}_a) = E(g, h) \cap \text{pref}(w).$$

Now, if neither of the letters a and b is repetitive, then $E(g, h) = \{\beta, \gamma\}^*$ for some nonempty words β and γ. This follows from the definition of the critical overflow. If exactly one of the letters a and b is repetitive, then $E(g, h) = (\beta\gamma^*\delta)^*$ for some reduced triple (β, γ, δ). Indeed, if a is the repetitive letter, then β equals the unique shortest word w such that $o(w) = (\alpha, g)$, $\gamma = a\bar{w}_a$ and δ equals the unique shortest word $b\hat{w}_b$ in $b\Sigma^*$ satisfying $\alpha g(b\hat{w}_b) = h(b\hat{w}_b)$. The definition of the critical overflow guarantees the existence of the above word $\hat{w}_b$. At the same time the argument above shows that the letters a and b cannot both be repetitive. □

It is an open question whether there really exist equality languages of binary morphisms of the form $(\beta\gamma^*\delta)^*$, and not expressable in the form $\{\beta, \gamma\}^*$. It was conjectured in [8], where the study of binary equality languages was initiated, that such equality languages do not exist. We still agree with this conjecture.

To prove the main result of this section we need the following lemma, the proof of which can be found in [16].

Lemma 6.4. *For two languages $L_i = \beta_i\gamma_i^*\delta_i$, 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 elements, then $L_1 = L_2$.*

Now, we are ready for the next theorem.

Theorem 6.5. *Each language L over a binary alphabet Σ possesses a test set of cardinality at most three.*

Proof. If L contains two words with different ratios, then these two words constitute a test set, since, as is straightforward to see, no equality language different from Σ^* can contain two words having different ratios (cf. also Theorem 6.1). Hence, we assume that all words of L have the same ratio.

By the definition of ratio-primitiveness, it is clear that each word x in $\{a, b\}^+$ possesses a unique decomposition $x = x_1 \ldots x_t$, where each x_i is ratio-primitive and

$r(x_i) = r(x)$ for $i = 1, \ldots, t$. We define L_r to be the language containing exactly those ratio-primitive words which occur in the above-mentioned decompositions when x ranges over L. Clearly, any two morphisms agree on L if and only if they agree on L_r. Moreover, any test set for L_r defines in a natural way a test set for L such that the cardinality of the test set does not increase. Therefore, it is enough to show that L_r has a test set containing no more than three words.

First, if L_r contains less than three words we are trivially done. Otherwise, we choose a three-element subset of L_r as follows. Let z_1 and z_2 be two different words of L_r. If they belong to a language of the form $\beta\gamma^*\delta$, where (β, γ, δ) is reduced, then, by Lemma 6.4, 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, including the possibility that $L(z_1, z_2)$ does not exist, z_3 is an arbitrary word from 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 morphisms are periodic. Now, by Theorem 6.1, any one-element set, and hence also $\{z_1, z_2, z_3\}$, tests whether such morphisms agree on L_r (remember that all words of L_r have the same ratio).

(II) The equality language of the pair (g, h) is of the form $\{\beta, \gamma\}^*$ for some words β and γ in Σ^*. Now, any three-element subset of L_r tests whether such morphisms agree on L_r. Indeed, since all words of L_r are ratio-primitive, the morphisms g and h actually disagree on any three-element subset of L_r.

(III) The equality language of the pair (g, h) is of the form $(\beta\gamma^*\delta)^*$ for some reduced triple (β, γ, δ). Now, if $\beta\gamma^*\delta = L(z_1, z_2)$, then, by the choice of z_3, the morphisms g and h do not agree on z_3, and hence the set $\{z_1, z_2, z_3\}$ tests whether they agree on L_r. If, in turn, $\beta\gamma^*\delta \neq L(z_1, z_2)$, including the case that $L(z_1, z_2)$ is not defined, then, by Lemma 6.4, both z_1 and z_2 cannot be in $\beta\gamma^*\delta$ and so also in this case $\{z_1, z_2, z_3\}$ tests whether g and h agree on L_r.

By Theorems 6.1 and 6.3, the classification (I)–(III) is exhaustive showing that the set $\{z_1, z_2, z_3\}$ is a test set for L_r. □

Of course, a test set for an arbitrary language over a binary alphabet cannot exist effectively (cf., e.g., [12]). However, it is not difficult to see from the proof of Theorem 6.5 that if a family $\mathcal{L}$ of languages satisfies the following three conditions, then a test set for each language L in $\mathcal{L}$ can be found effectively. Moreover, the cardinality of a test set is always at most three. The conditions are:

(i) Each language L in $\mathcal{L}$ is effectively recursively enumerable.

(ii) Given L in $\mathcal{L}$ and a regular language of the form $(\beta\gamma^*\delta)^*$, where β, γ and δ are words, it is decidable whether or not $(\beta\gamma^*\delta)^*$ includes L.

(iii) Given L in $\mathcal{L}$ it is decidable whether all words of L have the same ratio.

Consequently, we have the following.

Theorem 6.6. *Let $\mathcal{L}$ be a family of binary languages satisfying conditions (i)–(iii). Then each language L in $\mathcal{L}$ possesses effectively a test set of cardinality at most three.*

It is easy to see that conditions (i)–(iii) are shared by context-free languages over $\{a, b\}$ as well as by HDT0L languages over $\{a, b\}$ (cf. [30]). Consequently, Theorem 5.3 can be strengthened in the case of binary Σ as follows.

Theorem 6.7. *Each context-free language over a binary alphabet has effectively a test set of cardinality at most three.*

Similarly, we have a result which should be compared to results of Section 4.

Theorem 6.8. *Each* HDT0L *language over a binary alphabet has effectively a test set of cardinality at most three.*

We want to close this section with the following remarks. Let us denote by EC(n) the Ehrenfeucht Conjecture restricted to an n-letter alphabet. Then in terms of Section 3 the problem studied in this section is a special case of GEC(4), or in general EC(n) is a special case of GEC($2n$). It also follows from the proof of Theorem 3.2 that EC(n) implies GEC(n). Therefore, we have the implications

$$\mathrm{GEC}(2n) \Rightarrow \mathrm{EC}(n) \Rightarrow \mathrm{GEC}(n).$$

As we have seen only problems EC(2) and GEC(2) have been solved. The simplest open problem seems to be that of Section 3, namely GEC(3).

7. The Ehrenfeucht Conjecture for bounded delay morphisms

In this section we turn to consider the Ehrenfeucht Conjecture from a different point of view. Without restricting the class of languages, as in the two previous sections, we now restrict the class of morphisms. Our purpose is to establish the conjecture for all languages if attention is restricted to morphisms having a bounded delay equal to a given integer $p \geq 0$. Such morphisms have been studied for quite a long time [27]. The results of this section are taken from [14].

Basically, our considerations here are similar to those of Section 6. We start by characterizing equality languages of bounded delay morphisms. A remarkable property of such morphisms is that their equality language is always regular [18, 33]. We sharpen this result in our next theorem. In order to be able to state it properly, we denote, for each nonnegative integer p, by $\mathscr{H}_p$ the class of all morphisms $h : \Sigma^* \to \Sigma^*$ having bounded delay p.

Theorem 7.1. *Let p be a nonnegative integer. Then there exists a regular language R over some alphabet V such that for any two morphisms $g, h : \Sigma^* \to \Sigma^*$ in $\mathscr{H}_p$ there exists a morphism $\tau : V^* \to \Sigma^*$ such that $E(g, h) = \tau(R)$.*

Outline of the proof. Without giving a detailed proof of the above theorem (cf. [14]), we just point out its main lines. We use the terminology of Theorem 6.3. In particular, the notion of a critical overflow is important here. It is a property of a bounded

delay which implies that there exist only finitely many different critical overflows, i.e., overflows which can be continued in two different ways leading to a solution. Moreover, the number of different critical overflows depends upon p only. Consequently, according to the considerations at the end of the proof of Theorem 6.3, one can conclude that there exists only a finite number of 'patterns' such that each equality language of morphisms in $\mathcal{H}_p$ is of one of these forms. Therefore, the result follows. □

In the case when Σ is binary, Theorem 7.1 can be written in the following stronger form. Remember that in this special case a morphism is aperiodic if and only if it is injective which, in turn, holds if and only if it has a bounded delay for some p. Now Theorem 6.3 shows that in the binary case the regular language R of Theorem 7.1 can be chosen independently of p.

Theorem 7.2. *For each aperiodic binary morphisms $g, h: \{a, b\}^* \to \{a, b\}^*$ there exists a morphism $\tau: \{0, 1, 2, 3\}^* \to \{a, b\}^*$ such that $E(g, h) = \tau((\{0\} \cup \{12^*3\})^*)$.*

Besides Theorem 7.1 we also need the following lemma for the main result of this section.

Lemma 7.3. *Let $R \subseteq V^*$ and $L \subseteq \Sigma^*$ be languages over alphabets V and Σ, respectively. There exists a finite subset F of L such that for each morphism $\tau: V^* \to \Sigma^*$ we have*

$$L \subseteq \tau(R) \Leftrightarrow F \subseteq \tau(R). \qquad (\bigstar)$$

Outline of the proof. The basic idea of the proof of the lemma is that the (unknown) morphism τ can be fixed, i.e., the values of $\tau(a)$ for a in V can be fixed, step by step, by taking words from L and requiring that $x \in \tau(R)$. Since V is finite, each sequence of words fixing the values of τ can be chosen to be finite, too. We give an example of this procedure.

Let $L = \{a^n b^n \mid n \geq 1\}$ and $R = (\{0\} \cup \{12^*3\})^*$. We first consider the word $x_0 = ab$. Now the requirement $x \in \tau(R)$ yields the following possibilities

τ_1: $\tau_1(0) = ab$, undefined otherwise,
τ_2: $\tau_2(1) = ab$, $\tau_2(3) = 1$, undefined otherwise,
τ_3: $\tau_3(2) = ab$, $\tau_3(1) = \tau_3(3) = 1$, undefined otherwise,
τ_4: $\tau_4(3) = ab$, $\tau_4(1) = 1$, undefined otherwise,
τ_5: $\tau_5(0) = a$, $\tau_5(1) = b$, $\tau_5(3) = 1$, undefined otherwise,
τ_6: $\tau_6(0) = a$, $\tau_6(2) = b$, $\tau_6(1) = \tau_6(1) = 1$,
τ_7: $\tau_7(0) = a$, $\tau_7(3) = 1$, $\tau_7(1) = 1$, undefined otherwise,
τ_8: $\tau_8(0) = b$, $\tau_8(1) = a$, $\tau_8(3) = 1$, undefined otherwise,
τ_9: $\tau_9(0) = b$, $\tau_9(2) = a$, $\tau_9(1) = \tau_9(3) = 1$,
τ_{10}: $\tau_{10}(0) = b$, $\tau_{10}(3) = a$, $\tau_{10}(1) = 1$, undefined otherwise,
τ_{ab}: $ab \notin \tau_{ab}(R)$.

The above classification divides the class of all morphisms $\tau: \{0, 1, 2, 3\}^* \to \{a, b\}^*$ into 11 classes. For morphisms τ in the class τ_{ab} we have $x \notin \tau(R)$ and so (★) holds for them. For morphisms τ in the classes τ_5–τ_{10}, clearly, $a^*b^* \subseteq \tau(R)$ independently of how the unspecified values are fixed. Therefore, (★) holds for these morphisms, too. So it remains to consider the morphisms in the classes τ_1, τ_2, τ_3 and τ_4.

Any morphism τ in the class τ_3 contains only one nonspecified value, namely the value of $\tau(0)$. This can be fixed by considering the word *aabb* in L. Indeed, if $aabb \in \tau(R)$, then necessarily $\tau(0) = aabb$. Consequently, for all morphisms τ in class τ_3 either $aabb \notin \tau(R)$ or τ is fixed as follows: $\tau(0) = aabb$, $\tau(1) = \tau(3) = 1$ and $\tau(2) = ab$. In the latter case, a^3b^3 does not belong to $\tau(R)$ showing that (★) holds for all morphisms in class τ_3 with $F = \{ab, aabb, aaabbb\}$.

Morphisms τ in class τ_2 contain two unspecified values. Now, the word *aabb* yields the following partition of this class:

$$\begin{aligned}
\tau_{2,1}&: \tau_{2,1}(1) = ab,\ \tau_{2,1}(3) = 1,\ \tau_{2,1}(0) = aabb, \quad \text{undefined otherwise,}\\
\tau_{2,2}&: \tau_{2,2}(1) = ab,\ \tau_{2,2}(3) = 1,\ \tau_{2,2}(0) = a,\ \tau_{2,2}(2) = b,\\
\tau_{2,aabb}&: aabb \notin \tau_{2,aabb}(R).
\end{aligned}$$

As above, sets $\{ab, aabb, aaabbb\}$ and $\{ab, aabb\}$ are suitable subsets for morphisms in classes $\tau_{2,2}$ and $\tau_{2,aabb}$, respectively. The morphisms τ in class $\tau_{2,1}$ are not totally specified; however, in whichever way the value $\tau(2)$ is chosen the word *aaabbb* does not belong to $\tau(R)$, proving that (★) holds also for these morphisms with $F = \{ab, aabb, aaabbb\}$.

Class τ_4 is symmetric to class τ_2. Hence, the above applies for τ_4 as well. Finally, for morphisms in class τ_1 some straightforward considerations are needed to see that the above three-element subset of L satisfies (★) for all morphisms in this class, too. □

What we have proved is, by Theorems 6.1, 6.3 and 7.2, that the language $\{ab, aabb, aaabbb\}$ is a test set for the language $\{a^n b^n \mid n \geq 1\}$. Similarly, Theorem 7.1 and Lemma 7.3 yield the following general result.

Theorem 7.4. *Let p be a nonnegative integer. For each language $L \subseteq \Sigma^*$ there exists a finite subset F of L such that for all morphisms $g, h : \Sigma^* \to \Sigma^*$ in $\mathcal{H}_p$ the equation $g(x) = h(x)$ holds for all x in L if and only if it holds for all x in F.*

As in the previous section the subset F of L cannot be found effectively in general. However, our next result shows that if a family $\mathcal{L}$ of languages satisfies the following conditions, then this finite 'bounded delay p test set' can be found effectively for each L in $\mathcal{L}$. The conditions are the following:

(i) Each L in $\mathcal{L}$ is effectively recursively enumerable.

(ii) For each L in $\mathcal{L}$ and each regular language R the language $L \cap R$ is effectively in $\mathcal{L}$.

(iii) It is decidable whether or not a given L in $\mathcal{L}$ is empty.

With these conditions we have the next theorem.

Theorem 7.5. *Let $\mathscr{L}$ be a family of languages satisfying conditions* (i)–(iii) *above. Then for each L in $\mathscr{L}$ a finite subset F of Theorem* 7.4 *can be found effectively.*

Theorem 7.5 has some consequences concerning the morphism equivalence problem for languages. First we state the following theorem.

Theorem 7.6. *Let $\mathscr{L}$ be a family of languages satisfying conditions* (i)–(iii) *above. The morphism equivalence problem restricted to morphisms having bounded delay is decidable for $\mathscr{L}$.*

It is worth noting that in Theorem 7.6 class $\bigcup_{p\geq 0} \mathscr{H}_p$ can be used instead of class $\mathscr{H}_p$ for some fixed p. An example of quite a large family of languages satisfying our conditions (i)–(iii) is the family of indexed languages (cf. [21]). Hence, Theorem 7.2 together with the fact that the minimal bounded delay of a bounded delay morphism can be effectively found yields the following result, which is a slight generalization of a result in [12]. Our proof, however, uses a different approach.

Theorem 7.7. *It is decidable whether or not two bounded delay morphisms agree on a given indexed language.*

It is not known whether this theorem can be generalized for all morphisms, i.e., whether the morphism equivalence problem for indexed languages is decidable.

8. Concluding remarks

We have discussed the Ehrenfeucht Conjecture quite extensively, having pointed out its connections to some other problems and also shown the main results known about the conjecture at the present time. We have included in our presentation several proofs as well as outlines of the proofs in order to give the reader enough motivation to become interested in the problem.

Our feeling is that the Ehrenfeucht Conjecture is a well motivated and fundamental problem not only from the point of view of formal languages but also from the point of view of free monoids as a whole.

To conclude this article we give a list of open problems which, in our estimation, contain the most important and attractive open subproblems of the Ehrenfeucht Conjecture.

(1) Is every system of equations with three variables and no constants equivalent to its finite subsystem? In other words: Does GEC(3) hold?

(2) Does every three-element language possess a test set? In other words: Does EC(3) hold?

(3) Does the Ehrenfeucht Conjecture hold for all D0L languages?

(4) When does the Ehrenfeucht Conjecture hold for all codes, i.e., for all injective morphisms?

(5) Does the Ehrenfeucht Conjecture hold for all bounded delay morphisms, i.e., can the class $\mathscr{H}_p$, for $p \geq 0$, in Theorem 7.4 be replaced by the class $\bigcup_{p \geq 0} \mathscr{H}_p$?

Acknowledgment

This work could not exist without the support of the Academy of Finland.

References

[1] J. Albert, K. Culik II and J. Karhumäki, Test sets for context-free languages and algebraic systems of equations in a free monoid, *Inform. Control* **52** (1982) 172–186.

[2] J. Albert and D. Wood, Checking sets, test sets, rich languages and commutatively closed languages, *J. Comput. System Sci.* **26** (1983) 82–91.

[3] J. Berstel, *Transductions and Context-Free Languages* (Teubner, Stuttgart, 1979).

[4] J. Berstel, D. Perrin, J.F. Perrot and A. Restivo, Sur le Théorème du defaut, *J. Algebra* **60** (1979) 169–180.

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

[6] K. Culik II, Homomorphisms: Decidability equality and test sets, in: R. Book, ed., *Formal Language Theory, Perspectives, and Open Problems* (Academic Press, New York, 1980).

[7] K. Culik II and I. Fris, The decidability of the equivalence problem for D0L systems, *Inform. Control* **35** (1977) 20–39.

[8] K. Culik II and J. Karhumäki, On the equality sets for homomorphisms on free monoids with two generators, *RAIRO Inform. Théor.* **14** (1980) 349–369.

[9] K. Culik II and J. Karhumäki, Systems of equations over a free monoid and Ehrenfeucht's Conjecture, *Discrete Math.* **43** (1983) 139–153.

[10] K. Culik II and J. Karhumäki, On the Ehrenfeucht Conjecture for D0L languages, *RAIRO Inform. Théor.* **17** (1983) 205–230.

[11] K. Culik II and J. Karhumäki, On test sets and the Ehrenfeucht Conjecture, in: *Lecture Notes in Computer Science* **140** (Springer, Berlin, 1982) pp. 128–140.

[12] K. Culik II and A. Salomaa, On the decidability of homomorphism equivalence for languages, *J. Comput. System Sci.* **17** (1978) 163–175.

[13] K. Culik II and A. Salomaa, Test sets and checking words for homomorphism equivalence, *J. Comput. System Sci.* **19** (1980) 379–395.

[14] C. Choffrut and J. Karhumäki, Test sets for bounded delay morphisms, in: *Lecture Notes in Computer Science* **154** (Springer, Berlin, 1983) 118–127.

[15] A. Ehrenfeucht, J. Karhumäki and G. Rozenberg, The (generalized) Post Correspondence Problem with lists consisting of two word is decidable, *Theoret. Comput. Sci.* **21** (1982) 119–144.

[16] A. Ehrenfeucht, J. Karhumäki and G. Rozenberg, On binary equality sets and a solution to the Test Set Conjecture in the binary case, *J. Algebra* **85** (1983) 76–85.

[17] A. Ehrenfeucht and G. Rozenberg, Elementary homomorphisms and a solution to the D0L sequence equivalence problem, *Theoret. Comput. Sci.* **7** (1978) 169–183.

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

[19] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).

[20] Y.I. Hmelevskii, Equations in free semigroups, *Proc. Steklov Inst. Mat., Am. Math. Soc. Transl.* **107** (1976).
[21] J. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, MA, 1979).
[22] J. Karhumäki, On the equivalence problem for binary D0L systems, *Inform. Control* **50** (1981) 276–284.
[23] J. Karhumäki, A note on intersections of free submonoids of a free monoid, *Semigroup Forum*, to appear.
[24] G. Lallement, *Semigroups and Combinatorial Applications* (Addison-Wesley, Reading, MA, 1979).
[25] M. Lothaire, *Combinatorics on Words* (Addison-Wesley, Reading, MA, 1983).
[26] G.S. Makanin, The problem of solvability of equations in a free semigroup, *Mat. Sb.* **103** (1977) 147–236 (English transl. in: *Math. USSR Sb.* **32** (1977) 129–198).
[27] M. Nivat, Elements de la théorie generale des codes, in: E.R. Caianello, ed., *Automata Theory* (Academic Press, New York, 1966).
[28] A. Restivo and C. Reutenauer, On cancellation properties of languages which are supports of rational formal power series, Unpublished manuscript, 1983.
[29] G. Rozenberg, The equivalence problem for deterministic T0L-systems is undecidable, *Inform. Process. Lett.* 1 (1972) 201–204.
[30] G. Rozenberg and A. Salomaa, *The Mathematical Theory of* L *Systems* (Academic Press, New York, 1980).
[31] K. Ruohonen, Zeros of Z-rational functions and D0L equivalence, *Theoret. Comput. Sci.* **3** (1976) 283–292.
[32] A. Salomaa, Equality sets for homomorphisms of free monoids, *Acta Cybernet.* **4** (1978) 127–139.
[33] A. Salomaa, *Jewels of Formal Language Theory* (Computer Science Press, Rockville, MD, 1981).
[34] A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series* (Springer, Berlin, 1978).