Discrete Mathematics 43 (1983) 139–153
North-Holland Publishing Company

SYSTEMS OF EQUATIONS OVER A FREE MONOID AND EHRENFEUCHT'S CONJECTURE*

Karel CULIK II

Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

Juhani KARHUMÄKI

Department of Mathematics, University of Turku, Turku, Finland

Received 1 May 1981

Ehrenfeucht's conjecture states that every language L has a finite subset F such that, for any pair (g, h) of morphisms, g and h agree on every word of L if and only if they agree on every word of F. We show that it holds if and only if every infinite system of equations (with a finite number of unknowns) over a free monoid has an equivalent finite subsystem. It is shown that this holds true for rational (regular) systems of equations.

The equivalence and inclusion problems for finite and rational systems of equations are shown to be decidable and, consequently, the validity of Ehrenfeucht's conjecture implies the decidability of the HDOL and DTOL sequence equivalence problems. The simplicity degree of a language is introduced and used to argue in support of Ehrenfeucht's conjecture.

0. Introduction

The first step in showing the decidability of DOL sequence equivalence problem in [4] was to reformulate the problem in the following form for specific g, h and L. Given a language $L \subseteq \Sigma^*$ and two morphisms $g, h: \Sigma^* \to \Sigma^*$, test whether $g(x) = h(x)$ for each x in L, that is whether g and h are equivalent on L, in symbols $g \overset{L}{=} h$. The decidability of this problem has been studied for various types of languages (morphism equivalence for family $\mathcal{L}$), cf. survey [3].

Clearly related is the following conjecture by Ehrenfeucht (Problem 108 in [11]): For every language $L \subseteq \Sigma^*$ there exists a finite subset F of L such that for any pair of morphisms on Σ^*, $g \overset{L}{=} h$ iff $g \overset{F}{=} h$. Such a finite subset F has been called a test set for L in [7] where it has been shown that Ehrenfeucht's conjecture holds true for every language over a binary alphabet.

Ehrenfeucht's conjecture can be translated into the terminology of systems of equations over a free monoid. It says that every infinite system of equations of a certain very special form has an equivalent finite subsystem. In Section 2 we show a result which emphasizes the importance of Ehrenfeucht's conjecture from purely algebraic point of view, namely that it is equivalent to the following: Every

* This research was supported by the Natural Sciences and Engineering Council Canada, Grant A 7403.

0012-365X/83/0000–0000/$03.00

infinite system of equations (with finite number of unknowns) over a free monoid has an equivalent finite subsystem.

We also consider rational (regular) systems of equations. The effective existence of tests sets for rational (regular) languages [6] implies that for each rational system of equations a finite subsystem can be effectively constructed.

In Section 3 we show that the equivalence problem for finite systems of equations over Σ^* is decidable by reducing it to the solvability problem shown to be decidable by Makanin in [13]. This result is then extended to the equivalence and inclusion problems for rational systems of equations over Σ^*.

Among the applications of this result the most interesting is that for DOL and DTOL languages the existence of a test set implies its effective existence. This, in turn, means that the validity of Ehrenfeucht's conjecture implies the decidability of the well known HDOL sequence equivalence problem and the DTOL sequence equivalence problem, cf. [3].

In the last section we introduce the notion of simplicity degree of a language. It yields a classification of simplifiable languages introduced in [8]. We believe, it is of interest on its own but our main purpose is to show results which give some support for the validity of Ehrenfeucht's conjecture.

1. Preliminaries

For basic properties of free monoids and some elementary results of formal language theory we refer the reader to [9]. The following is mentioned mainly to establish notation.

A free monoid generated by a finite alphabet Σ is denoted by Σ^*. Its elements are called words or strings. The identity element of Σ^*, so-called empty word, is denoted by λ, and $\Sigma^+ = \Sigma^* - \{\lambda\}$. For a word x, $|x|$ denotes its length, and for a finite alphabet Σ, $|\Sigma|$ denotes its cardinality. For two words x and y, $x < y$ denotes that x is a prefix of y, i.e. there is a $z \in \Sigma^*$ so that $y = xz$. Prefix x is *proper* if $x \neq \lambda$ and $z \neq \lambda$. We call a set *prefix* if none of its elements is not a prefix of another.

Our basic notion is that of a morphism from a free monoid into another free monoid, in symbols $h : \Sigma^* \to \Delta^*$. We call h λ-*free* if $h(a) \neq \lambda$ for each a in Σ. By a *code* we mean an injective morphism, while a *coding* refers to a morphism h satisfying $|h(a)| = 1$ for each a in Σ. By a *prefix code* (resp. *suffix code*) we mean a morphism such that none of the images of generators is a prefix (resp. suffix) of another. By a *biprefix code* we mean a morphism which is both prefix and suffix code. All of these are special cases of so-called *bounded delay morphisms*, cf. e.g. [14]. We shall need also some other special types of morphisms. A morphism h is called *atomic* if it is of the form $h(a) = ab$, $h(x) = x$ otherwise, or of the form $h(a) = ba$, $h(x) = x$ otherwise, for some a and b in Σ. Any composition of atomic morphisms, including the zeroth power, i.e. the identity morphism I, is called *quasiatomic*. Finally, for two morphisms $g : \Sigma^* \to \Delta^*$ and $h : \Sigma_1^* \to \Delta^*$, with

$\Sigma \cap \Sigma_1 = \emptyset$, we define a morphism $g \cup h : (\Sigma \cup \Sigma_1)^* \to \Delta^*$ by setting $[g \cup h](a) = g(a)$, for each a in Σ, and $[g \cup h](a) = h(a)$, for each a in Σ_1.

A *DOL-system* is a triple (Σ, h, x), where Σ is an alphabet, h is a morphism $\Sigma^* \to \Sigma^*$ and x is a nonempty word in Σ^*. A DOL system defines a *DOL sequence* when h is applied iteratively to $x : x, h(x), h^2(x), \ldots$. If such a sequence is mapped by another morphism, say $f : \Sigma^* \to \Delta^*$, an *HDOL sequence* is obtained. Thus, an *HDOL system* can be identified with a quadruple (Σ, h, x, f). For the basic properties of DOL systems as well as the definition of a DTOL system we refer to [14].

We say that two morphisms $g, h : \Sigma^* \to \Delta^*$ are *equivalent* on a language $L \subseteq \Sigma^*$, in symbols $g \overset{L}{=} h$, if $g(x) = h(x)$ for each x in L. We call a finite subset F of L a *test set for* L if, for any pair of morphisms (g, h), $g \overset{L}{=} h$ if and only if $g \overset{F}{=} h$. Our basic interest is in the following:

Ehrenfeucht's Conjecture. For every language over a finite alphabet there exists a test set.

Let L be a language and x a word over Σ^*. We say that x is *morphically forced* by L, if for all pairs of morphisms g and h, $g \overset{L}{=} h$ implies $g(x) = h(x)$. A language L is called *independent* if none of its words is morphically forced by the rest of the language. This means that for any x in L there exist morphisms g and h such that $g \overset{L'}{=} h$, where $L' = L - \{x\}$, and $g(x) \neq h(x)$.

Let $N = \{x_1, \ldots, x_n\}$ be a finite set of variables (unknowns) and Σ be a finite alphabet. An *equation* with n unknowns over Σ^* is of the form

$$u = v$$

where $u, v \in \{N, \Sigma\}^*$. A system of equations is a (finite or infinite) collection of equations. A *solution* of an equation (or a system of equations) with n unknowns over Σ^* is an n-tuple from $(\Sigma^*)^n$. Formally, a solution of a system S is nothing but a morphism $h : (N \cup \Sigma)^* \to \Sigma^*$ such that $h(a) = a$, for each a in Σ, and $h(u) = h(v)$ for each equation $u = v$ in S. Two systems S_1, S_2 of equations are called *equivalent* if they have exactly the same solutions. Finally, a system S of equations is called *independent* if it is not equivalent to any of its proper nonempty subsystems (sub-sets).

2. A generalization of Ehrenfeucht's conjecture

We now show that the existence of a test set for a language of certain type is equivalent to the existence of an equivalent finite subsystem for a system of equations over Σ^* of 'the same type'.

To have a convenient notation for systems of equations over Σ^* (both finite and infinite) we define it formally as follows.

A *system of equations over* Σ^* *with unknowns* N is a binary relation $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$. A pair $(u, v) \in S$ represents the equation $u = v$. We say that a system of equations is *rational* (*regular*) if N is finite and S is a rational (regular) relation [2]. Similarly, we have an *algebraic* (*push down*) system of equations.

A family of binary relations $\mathcal{R}$ is said to be *morphically characterized* by a family of languages $\mathcal{L}$ if the following holds: $R \in \mathcal{R}$ iff there exist two morphisms g, h and a language $L \in \mathcal{L}$ such that $R = \{(g(w), h(w)) \mid w \in L\}$ or briefly $R = [g, h]L$.

It is well known (Nivat's theorem) that the family of rational (regular) relations is morphically characterized by the family of rational (regular) sets, and the family of algebraic (push down) relations by the family of algebraic (context-free) languages.

Lemma 2.1. *The family of all binary relations is morphically characterized by the family of all languages.*

Proof. Obvious. □

Let $\mathcal{L}$ be a family of languages. We say that a system of equations S, $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$, is of *type* $\mathcal{L}$ (e.g. rational) if S belongs to the family of relations morphically characterized by $\mathcal{L}$, i.e. if $S = [f, g]L$ for $L \in \mathcal{L}$ and some morphisms f and g. The discussion above justifies this terminology in the case of rational, algebraic and arbitrary systems of equations.

We are ready to show the correspondence between the existence of test sets for languages of type $\mathcal{L}$ and the existence of equivalent finite subsystems of equations of type $\mathcal{L}$.

Theorem 2.1. *Let* $\mathcal{L}$ *be a family of languages. The following two statements are equivalent:*

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

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

Proof. (ii) $\Rightarrow$ (i). Let $L \in \mathcal{L}$, $L \subseteq \Sigma^*$. Define $\bar{\Sigma} = \{\bar{a} \mid a \in \Sigma\}$, and morphism $\mu : \Sigma^* \to \bar{\Sigma}^*$ by $\mu(a) = \bar{a}$ for each $a \in \Sigma$.

Consider the system of equations $S = [I, \mu]L$, where I is the identity morphism. Clearly, for morphisms f, g, $f \overset{L}{=} g$ iff $f \cup (g \circ \mu^{-1})$ is a solution of S. System S is of type $\mathcal{L}$, therefore, by (ii), there exists an equivalent finite subsystem S' of S. Let F be a finite subset of L such that $S' \subseteq [I, \mu]F$. Now, if $f \overset{F}{=} g$, then $f \cup (g \circ \mu^{-1})$ is a solution of S', therefore it is also a solution of S and thus $f \overset{L}{=} g$. Hence F is a test set for L.

(i) $\Rightarrow$ (ii). Let $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ be a system of equations of type $\mathscr{L}$. That means that there is $L \in \mathscr{L}$, $L \subseteq \Delta^*$ for some alphabet Δ, and morphisms $f, g : \Delta^* \to (N \cup \Sigma)^*$ such that $S = [f, g]L$. By (i) there exists a test set F for L. Consider the system of equations S' defined by $S' = [f, g]F$. Let h be a solution of S', i.e. $h(u) = h(v)$ for each $(u, v) \in S'$. This means $h \circ f \overset{F}{\equiv} h \circ g$. Since F is a test set for L also $h \circ f \overset{L}{\equiv} h \circ g$ which, in turn, can be written as $h(u) = h(v)$ for each $(u, v) \in S$, i.e. h is also a solution of S. Thus every solution of subsystem S' is a solution of system S. Since the converse is obvious the system S and its finite subsystem S' are equivalent.

Finally, we observe that our constructions are effective, so if a test set for a language from $\mathscr{L}$ exists effectively, then also the finite subsystem S' exists effectively and vice versa. □

Now, we are ready to state the following conjecture and demonstrate its equivalence to the well known Ehrenfeucht's conjecture.

Conjecture A. For each arbitrary system S of equations (with a finite number of unknowns) over Σ^* there exists (noneffectively) an equivalent finite subsystem (subset of S). Or alternately: no infinite system of equations is independent.

Corollary 2.2. *Conjecture* A *holds iff Ehrenfeucht's conjecture holds.*

Proof. By Lemma 2.1 and Theorem 2.1. □

Since for every regular set there effectively exists a test set [6] we have the following.

Corollary 2.3. *For every rational system of equations there effectively exists an equivalent finite subsystem.*

The corresponding result for algebraic equations is shown in [1].

3. The equivalence problem for systems of equations over a free momoid, and applications

Now, we show that the equivalence of two infinite systems of equations (with a finite number of unknowns) over Σ^* is decidable. Clearly, this result extends to all types of systems of equations for which there effectively exists a finite subsystem, in particular, to rational systems of equations. We reduce these problems to the problem of the solvability of a system of equations shown decidable in [13]. Essential is the possibility to translate the nonequality into several systems of equations shown in [5].

Theorem 3.1. *The equivalence problem for finite systems of equations (with a finite number of unknowns) is decidable.*

Proof. Given two systems $S_1, S_2 \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ we construct a finite number of systems $Z_1, \ldots, Z_n$ such that S_1 and S_2 are equivalent iff for $i = 1, \ldots, n$ no Z_i has a solution.

Let $S_1 = \{\alpha_j = \beta_j \mid j = 1, \ldots, r\}$ and $S_2 = \{\gamma_j = \delta_j \mid j = 1, \ldots, s\}$ where each $\alpha_j, \beta_j, \gamma_j, \delta_j \in (N \cup \Sigma)^*$. Clearly, systems S_1 and S_2 are nonequivalent iff either for some values of the unknowns and some $k \in \{1, \ldots, s\}$

$$\alpha_j = \beta_j \quad \text{for } j = 1, \ldots, r \quad \text{and} \quad \gamma_k \neq \delta_k \tag{1}$$

or for some values of the unknowns and some $k \in \{1, \ldots, r\}$.

$$\alpha_k \neq \beta_k \quad \text{and} \quad \gamma_j = \delta_j \quad \text{for } j = 1, \ldots, s. \tag{2}$$

Similarly like in the proof of Theorem 3.1 in [5] we translate (1), for each value of k, into a finite number of systems of equations, that is to say, we construct a finite number of new systems of equations in such a way that at least one of these has a solution iff (1) is satisfied for some values of the unknowns N.

For each $a \in \Sigma$ we define systems of equations with unknowns $N \cup \{x\}$ over Σ^* by

$$S_1 \cup \{\gamma_k = \delta_k a x\}. \tag{3}$$

and

$$S_1 \cup \{\gamma_k a x = \delta_k\}. \tag{4}$$

Furthermore, for each pair $(a, b) \in \Sigma \times \Sigma$, $a \neq b$, we construct systems of equations with unknowns $N \cup \{x, y, z\}$ over Σ^* as follows:

$$S_1 \cup \{\gamma_k = xay, \ \delta_k = xbz\}. \tag{5}$$

Note that, for each $k = 1, \ldots, s$, (3), (4) and (5) represent several systems each but altogether a finite number of systems. Symmetrically, we can translate (2) into a finite number of systems.

Using the result of Makanin [13] we can test for each of the above systems, say $Z_1, \ldots, Z_n$, whether it has a solution. If at least one of them has a solution, then systems S_1 and S_2 are not equivalent, otherwise they are equivalent. □

Corollary 3.1. *The equivalence problem for rational systems of equations is decidable.*

Proof. By Corollary 2.2 and Theorem 3.1. □

Corollary 3.2. *Given two infinite languages L_1 and L_2, $L_1 \subseteq L_2$, it is decidable whether L_1 is a test set for L_2.*

Proof. By Theorem 3.1 after translating the problem into equations as shown in Section 2. □

Now we show that the inclusion problem for solutions of systems of equations can be reduced to the equivalence problem.

The *inclusion problem* for systems of equations is: given two systems S_1 and S_2, test whether each solution of S_1 is also a solution of S_2.

Corollary 3.3. *The inclusion problem for finite or rational systems of equations is decidable.*

Proof. Given two finite systems of equations S_1 and S_2, every solution of S_1 is a solution of S_2 iff $S_1 \cup S_2$ and S_2 are equivalent, which is decidable by Theorem 3.1. For rational systems we first construct equivalent finite systems by Corollary 2.2. □

Now, we will use Corollary 3.2 to show a rather surprising result, namely that Ehrenfeucht's conjecture implies the effective existence of test sets for DOL and DTOL languages and thus, in turn, the decidability of well known open problems, for example the HDOL sequence equivalence problem, cf. [3].

Theorem 3.2. *For every DOL language L, the existence of a test set implies that a test set for L can be effectively found.*

Proof. Let $L = L(G)$ where $G = (\Sigma, h, w)$ is a DOL system. Since L possesses a test set, there exists the minimal $p > 0$ such that the set $\{w, \ldots, h^{p-1}(w)\}$ morphically forces $h^p(w)$. i.e. if for arbitrary morphisms f and g

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

then

$$f(h^p(w)) = g(h^p(w)).$$

This minimal p can be found effectively, since by Corollary 3.2 we can test whether $\{w, h(w), \ldots, h^{p-1}(w)\}$ is a test set for the language $\{w, h(w), \ldots, h^p(w)\}$. We now show that for each $n \geq p$, the set $\{w, h(w), \ldots, h^{n-1}(w)\}$ morphically forces the string $h^n(w)$, which means that $\{w, h(w), \ldots, h^{p-1}(w)\}$ is a test set for L.

Assume that there is $N > p$ such that $h^N(w)$ is not morphically forced by $\{w, h(w), \ldots, h^{N-1}(w)\}$, that is there exist morphisms α, β such that $\alpha(h^k(w)) = \beta(h^k(w))$ for $0 \leq k < N$ and $\alpha(h^N(w)) \neq \beta(h^N(w))$. Let $\gamma = \alpha \circ h^{N-p}$ and $\delta = \beta \circ h^{N-p}$. Morphisms γ and δ are equivalent on the set $\{w, h(w), \ldots, h^{p-1}(w)\}$ but $\gamma(h^p(w)) = \alpha(h^N(w)) \neq \beta(h^N(w)) = \delta(h^p(w))$, a contradiction with the choice of p. □

We may generalize Theorem 3.2 as follows:

Theorem 3.3. *For every DTOL language L, the existence of a test set implies that a test set for L can be effectively found.*

Proof. An easy modification of the proof of Theorem 3.2. Here Ehrenfeucht's conjecture implies the existence of the minimal p such that the set F of strings derived in less than p steps by a given DTOL system G morphically forces each string derived in exactly p steps in G. Almost the same arguments as in the proof of Theorem 3.2 then show that F is a test set for $L(G)$. □

Corollary 3.3. *If Ehrenfeucht's conjecture holds true, then for every DOL (DTOL) language there effectively exists a test set.*

Proof. By Theorem 3.2 and Theorem 3.3. □

Corollary 3.4. *If Ehrenfeucht's conjecture holds, then the following problems are decidable.*

(a) *Morphism equivalence on DOL languages, i.e. given a DOL system G and morphisms g, f, is* $g \overset{L(G)}{\equiv} f$?

(b) *HDOL sequence equivalence problem.*

(c) *Morphism equivalence on DTOL languages.*

(d) *DTOL sequence equivalence problem, see* [6] *or* [3].

Proof. Obviously morphism equivalence is decidable for every family of languages $\mathscr{L}$ such that for each L in $\mathscr{L}$ a test set effectively exists. Hence we have parts (a) and (b) by Corollary 3.3. The equivalence of (a) and (c) and also of (b) and (d) is shown in [3]. □

4. Simplicity degree of a language

The aim of this section is to introduce a notion which, we believe, not only gives support for the validity of Ehrenfeucht's conjecture but is interesting on its own, too. The notion, simplicity degree of a language, yields a classification for simplifiable languages defined in [8].

Definition. Let $L \subseteq \Sigma^*$. The *simplicity degree of L*, in symbols $\mathrm{sd}(L)$, is defined by

$$\mathrm{sd}(L) = \min\{|F| \mid F \subseteq \Sigma^*,\ L \subseteq F^*\}.$$

For a morphism $h : \Sigma^* \to \Delta^*$ the *simplicity degree of h*, in symbols $\mathrm{sd}(h)$, is defined by $\mathrm{sd}(h) = \mathrm{sd}(h(\Sigma))$.

It is clear that for any language L

$$\mathrm{sd}(L) \leqslant \min\{|\Sigma|,\ |L|\}$$

and for any morphism h

$$\mathrm{sd}(h) \leqslant |\Sigma|.$$

Following [8] we call a language or a morphism *simplifiable* if its simplicity degree is strictly smaller than $|\Sigma|$. By a *periodic* language or a morphism we mean a language or a morphism having the simplicity degree equal to 1.

Our next result, a characterization result for λ-free morphisms, provides a tool to show some simplifiability properties of morphisms. The proof of this result goes along the lines of the proof of Theorem 1 in [10].

Theorem 4.1. *For each λ-free morphism $h: \Sigma^* \to \Delta^*$ there exist a biprefix $f: \Sigma_1^* \to \Delta^*$ for some $\Sigma_1 \subseteq \Sigma$, a coding $c: \Sigma^* \to \Sigma_1^*$, and a quasiatomic morphism $\pi: \Sigma^* \to \Sigma^*$ such that $h = f \circ c \circ \pi$.*

Proof. We first assume that the set $L = \{h(a) \mid a \in \Sigma\}$ is a biprefix. For each a in Σ let $[a] = \{b \in \Sigma \mid h(a) = h(b)\}$. This yields a partition of Σ. Let a_1 be a fixed element of the equivalence class $[a]$. Now the decomposition follows when we define $\Sigma_1 = \{a_1 \mid a \in \Sigma\}$, $f = h/\Sigma_1$, $\pi = I$ and c by the condition $c(a) = a_1$, for each a in Σ.

Secondly, assume that L is not a biprefix, say L is not a prefix (the other case is symmetric). Then there exist letters a and b such that $h(a) = h(b)u$ for some nonempty u. We define morphisms $h': \Sigma^* \to \Delta^*$ and $p: \Sigma^* \to \Sigma^*$ by

$$\begin{cases} h'(a) = u, \\ h'(x) = h(x) \quad \text{otherwise,} \end{cases}$$

and

$$\begin{cases} p(a) = ba, \\ p(x) = x \qquad \text{otherwise.} \end{cases}$$

Clearly, $h = h' \circ p$ and p is atomic. Moreover,

$$(*) \qquad \sum_{a \in \Sigma} |h(a)| > \sum_{a \in \Sigma} |h'(a)|.$$

If the set $\{h'(a) \mid a \in \Sigma\}$ is not a biprefix we apply the above procedure to h'. By $(*)$, we finally encounter the assumptions of the first part of this proof. Hence, the theorem follows. □

As an immediate corollary we obtain.

Corollary 4.1. *The semigroup of λ-free morphisms from Σ^* into itself is generated by biprefixes, codings and atomic morphisms.*

As an application to simplifiability results we state:

Corollary 4.2. *Let $\mathcal{H}$ be any set of λ-free morphisms of Σ^* containing biprefixes and atomic morphisms and closed under composition. If h is not in $\mathcal{H}$, then h is simplifiable.*

Proof. If h is not λ-free, then, clearly, h is simplifiable. So assume that h is λ-free and not in $\mathcal{H}$. By Theorem 4.1, h has a representation $h = f \circ c \circ \pi$, where f is a biprefix, c is a coding and π is a composition of atomic morphisms. If c would be a permutation, then, by the properties of $\mathcal{H}$, h would be in $\mathcal{H}$. So $|c(\Sigma)| < |\Sigma|$, i.e. h is simplifiable. □

Corollary 4.2 provides a uniform and simple proof for a result originally proved for codes in [8] and strengthened for bounded delay codes in [12].

Corollary 4.3. *If h is not a bounded delay code, then it is simplifiable.*

Proof. The properties of $\mathcal{H}$ in Corollary 4.2 are obviously satisfied by codes and easy to verify also for bounded delay codes (in the same direction). □

As a final corollary we give a simple proof of Theorem 1.3.4 from [9].

Corollary 4.4. *Let x and y be words in Σ^+. If they satisfy a non-trivial identity, i.e. $xu = yv$ for some u and v in $\{x, y\}^*$, then* $\mathrm{sd}(\{x, y\}) = 1$.

Proof. By the identity $xu = yv$, $\{x, y\}$ is not a code. Consequently, by Corollary 3, $\{x, y\}$ is simplifiable, i.e. periodic. □

We shall generalize Corollary 4 to the three word case later.

Next we turn to consider Ehrenfeucht's conjecture in the light of simplifiability. We first derive two simple reduction results.

Theorem 4.2. *Any system S of equations over Σ^* is equivalent to a system S', where constants occur only in equations of the form $x_a = a$ with $a \in \Sigma$ and x_a is an unknown.*

Proof. For each a in Σ occurring in equations of S as a constant we introduce a new variable x_a. Now S' is obtained from S by replacing all occurrences of letters in equations by the corresponding variables and adding to the system equations $x_a = a$. □

To be able to state our second reduction result we need the following notion. Let S be a system of equations without constants. We say that a solution of S is *nonsingular* iff all its components are nonempty. We also say that two systems of equations are *non-singularly equivalent* iff they have exactly the same nonsingular solutions.

Theorem 4.3. *If any system of equations (without constants) is nonsingularly equivalent to its finite subsystem, then Conjecture* A *holds.*

Proof. Let S be a system of equations with unknowns N. For each subset A of N we define a morphism $\phi_A : N^* \to N^*$ by setting

$$\phi_A(x) = x \quad \text{if } x \notin A,$$
$$\phi_A(x) = \lambda \quad \text{if } x \in A.$$

Moreover, let S_A be the system of equations obtained from S by applying ϕ_A to each of its equations.

Now we assume that each system of equations has nonsingularly equivalent subsystem. Let F_A be such a system for each S_A. Let further F be a finite subsystem of S satisfying: if $u = v$ is an equation in some F_A, then there exist in F an equation $u' = v'$ such that $\phi_A(u') = u$ and $\phi_A(v') = v$. By the construction, it is immediate that F is equivalent to S. Hence, the theorem follows. □

Example 4.1. Let $L = \{aab, baa, x\}$, with $x \notin \{aab, baa\}^*$, and (h, g) a pair of morphisms, $h \neq g$, agreeing on L. We consider $\mathrm{sd}(L(h, g))$, where $L(h, g) = \{h(\Sigma), g(\Sigma)\}$. Clearly, $\mathrm{sd}(L(h, g)) \leqslant 4$. Moreover, since $h(aab) = g(aab)$, $h(a)$, $h(b)$, $g(a)$ and $g(b)$ satisfy a nontrivial identity and thus, by Corollary 4.3 $\mathrm{sd}(L(h, g)) \leqslant 3$. Indeed, as shown in [5], h and g, assuming that $|h(a)| > |g(a)|$, are of the form

$$(**) \qquad h: \begin{matrix} a \to \alpha(\beta\alpha)^t\beta\alpha, \\ b \to \gamma, \end{matrix} \qquad g: \begin{matrix} a \to \alpha(\beta\alpha)^t, \\ b \to \beta\alpha\gamma\alpha\beta \end{matrix}$$

for some words α, β and γ with $\alpha\beta \neq \lambda$ and $t \geqslant 0$. Now, substituting $(**)$ into $h(baa) = g(baa)$ we obtain a nontrivial identity for α, β and γ. Consequently, the set $\{\alpha, \beta, \gamma\}$ is simplifiable, and therefore $\mathrm{sd}(L(h, g)) \leqslant 2$. In fact, in [5] it is shown that h and g are obtained from several formulas involving two variables and some parameters. Again we substitute these formulas to $h(x) = g(x)$ and obtain, in all the cases, a nontrivial identity for these two variables, as shown in [5]. Therefore, $\mathrm{sd}(L(h, g)) = 1$, i.e. h and g are periodic.

We believe that the above is true in general, too. That is to say: if two distinct morphisms h and g over $\{a, b\}^*$ agree on one word, then $\mathrm{sd}(L(h, g)) \leqslant 3$, if they agree on two distinct r-primitive words, then $\mathrm{sd}(L(h, g)) \leqslant 2$, and, finally, if they agree on three distinct r-primitive words, then $\mathrm{sd}(L(h, g)) = 1$, i.e. h and g are periodic. Here, the r-primitiveness means that none of the prefixes of a word has the same ratio of the occurrences of letters as the whole word, cf. [5]. If our belief is true, then it would immediately imply the validity of a conjecture made in [5]: any regular equality set over $\{a, b\}^*$, i.e. the set of the form $\{x \in \{a, b\}^* \mid h(x) = g(x)\}$, is generated by at most two words.

Even in the general case, i.e. when the number of variables is arbitrary and we are not restricted to equations of the special form $h(x) = g(x)$, we do not know any example violating the statement: If S is the system of $n < |\Sigma|$ independent equations without constants, then each of its nonsingular solutions is of the

simplicity degree of at most $|\Sigma|-n$. In particular, if $n \geq |\Sigma|-1$, then all nonsingular solutions of S are periodic.

Clearly, by Theorems 4.2 and 4.3, the above statement, if valid, implies a positive answer to Conjecture A. Although we are not able to prove the statement even in the case of three variables (for a special case, cf. Theorem 4.4) we feel that our considerations, especially the notion of the simplicity degree and Corollary 4.3, gives a support for the validity of Ehrenfeucht's conjecture. In the case of two variables the above statement is nothing else but our Corollary 4.4. Consequently, Conjecture A is valid in this case.

The difficulties in trying to prove the above statement in general arise as follows. Given a finite system S of independent equations and a single equation $u = v$ such that also $S \cup \{u = v\}$ is independent, i.e. all of the solutions of S are not solutions of $u = v$. Assume that the simplicity degree of any solution of S is at most m, i.e. any solution is obtained from an expression involving at most m variables. Now, is it possible that for *some* expressions when substituted in $u = v$ no restriction for the m variables is obtained? If this is not possible, then the inductive application of Corollary 4.3 would imply the above statement.

The nonsingularity in the statement is needed to avoid some trivialities. Indeed, if we apply the morphism ϕ_A of the proof of Theorem 4.3 to nonidentical equations there is no guarantee that the equations remain nonidentical, i.e. when considered singular solutions the step-by-step simplification, as shown in Example 4.1, need not work. Indeed, we have a counter example.

Example 4.2. Let us consider the following system of equations

$$\begin{cases} xyz = yzx, \\ xzy = zyx. \end{cases}$$

The general solution is

$$\begin{cases} x, y, z \in \alpha^* & \text{for some } \alpha, \\ x = \lambda,\ y = \beta,\ z = \gamma & \text{for some } \beta \text{ and } \gamma. \end{cases}$$

This is seen, for example, as an application of Theorem 4.4. Clearly, our system is independent. Its all nonsingular solutions are of the simplicity degree 1, as the statement demands, but it has singular solutions of the simplicity degree 2.

We finish this section with a partial solution for Conjecture A in the case of three variables. Our result is, in a sense, a generalization of Corollary 4.4, and thus we believe it is of interest on its own.

Theorem 4.4. *Every nonsingular solution of the system*

$$S(0): \begin{cases} x\alpha = y\beta, \\ x\gamma = z\delta, \end{cases}$$

where α, β, γ and δ are words over $\{x, y, z\}$, is of the simplicity degree 1.

Proof. Let $u = v$ be an arbitrary equation over three variables and without constants such that neither u is a prefix of v nor v is a prefix of u. Further let the maximal common prefix of u and v be of length k and let u_{k+1} and v_{k+1} denote the $(k+1)$st variables of u and v, respectively. We define $\text{first}(u = v) = \{u_{k+1}, v_{k+1}\}$. By definition, whenever $\text{first}(u = v)$ is defined, its cardinality is two. Clearly, the system $S(0)$ has the property:

$$\text{first}(x\alpha = y\beta) \neq \text{first}(x\gamma = z\delta). \tag{1}$$

Now, we claim that this property is preserved in the following operation: Let $x = \tau x'$, where $\tau \prec y\beta$ and x' is a new variable, or $y = \tau' y'$, where $\tau' \prec x\alpha$ and y' is a new variable. A new system $S(1)$ is obtained from $S(0)$ by replacing each occurrence of x by $\tau' x'$ or, in the second case, by replacing each y by $\tau' y'$.

Let us denote the new equations, now over $\{x', y, z\}$ or $\{x, y', z\}$, by $s_1(1)$ and $s_2(1)$. Assume first that $S(1)$ is obtained by the replacement of x by $\tau x'$. Then, for the first equation $\text{first}(s_1(1)) = \{x', w\}$, where $w \in \{y, z\}$, and for the second equation $\text{first}(s_2(1)) = \{y, z\}$. Consequently, $\text{first}(s_1(1)) \neq \text{first}(s_2(1))$. If, on the other hand, $S(1)$ is obtained by the replacement of y by $\tau' y'$, then $\text{first}(s_1(1)) = \{w, y'\}$, where $w \in \{x, z\}$, and $\text{first}(s_2(1)) = \{x, z\}$, and so again $\text{first}(s_1(1)) \neq \text{first}(s_2(1))$.

By induction, we now conclude that we may iterate the above operation and still the nonequality of (1) remains valid for new systems. Let us denote the $(i+1)$st system by $S(i)$ and its equations by $s_1(i)$ and $s_2(i)$.

Our second claim is that if we make the following operation for $S(0)$ (or the corresponding operation for any $S(i)$), then the second equation of $S(0)$ (or $S(i)$) does not change into the identity. The operation is: any occurrence of x in $S(0)$ is replaced by a non-empty prefix of $y\beta$, or any occurrence of y is replaced by a nonempty prefix of $x\alpha$, in other words, we make our first operation by assuming that x' and y' are empty. The validity of this claim is easy to verify. Indeed, in both the cases the first variable of $s_2(0)$ on the right hand side remains unchanged, i.e. z, while the first variable of $s_2(1)$ on the left hand side will be y or x.

Let us now refer our first operation by a and our second operation by b. Then, clearly, the word $a^n b$, for $n \geq 0$, makes $S(0)$ to the system, whose variables $x(n)$, $y(n)$, $z(n)$ and the original variables x, y, z are related as follows:

$$x(0) = x, \qquad y(0) = y, \qquad z(0) = z, \tag{2}$$

$$\begin{aligned} x(i) &= f_i(y(i), z(i))\, x(i+1) \quad \text{for } i < n, \\ y(i) &= g_i(x(i), z(i))\, y(i+1) \quad \text{for } i < n, \\ z(i) &= h_i(x(i), y(i))\, z(i+1) \quad \text{for } i < n, \end{aligned} \tag{3}$$

and

$$\begin{aligned} x(n) &= f_n(y(n), z(n)), \\ y(n) &= g_n(x(n), z(n)), \\ z(n) &= h_n(x(n), y(n)), \end{aligned} \tag{4}$$

where the functions f_i, g_i and h_i, for $i = 0, \ldots, n$, are used to describe what is the actual ith operation. So, in all cases, functions f_i, g_i and h_i are some fixed catenations of their arguments and moreover, for each i, exactly one of them yields a nonempty word.

Now we claim that any $(x(0), y(0), z(0))$ obtained in the above way using nonempty values of variables is of the simplicity degree 1. Indeed, by its recursive definition it is expressable by using only two words. Moreover, these two words satisfy a nontrivial identity, namely the identity which is obtained from $s_2(n-1)$ after the substitution of (4). Observe, however, that to get this nontrivial identity for these words we must use nonempty values of variables.

Finally, we are ready to finish this proof. It is enough to show that every nonsingular solution of $S(0)$ is obtained from formulas (2)–(4) for some choices of functions f_i, g_i and h_i. This is, actually, quite immediate. Indeed, if $(\bar{x}, \bar{y}, \bar{z})$ is a given triple of nonempty words satisfying $S(0)$, then it must satisfy $s_1(0)$ and hence, depending on the relative lengths of $\bar{x}$, $\bar{y}$ and $\bar{z}$, it defines the sequence of operations a followed by one operation b. That is to say, $(\bar{x}, \bar{y}, \bar{z})$ satisfies the formulas (2)–(4) for this fixation and for some nonempty values of variables. □

Without the assumption first $(s_1(0)) \neq$ first$(s_2(0))$ Theorem 4.4 does not hold, as it is seen by considering the system S' defined by

$$xy = yz, \qquad xxy = yzz$$

Now, first$(xy = yx) = \{x, y\} =$ first$(xxy = yzz)$ and, however. S_1 has a nonperiodic nonsingular solution $x = ab$, $y = a$, $z = ba$. Observe, however, that S' (or even the infinite system $\{x^i y = yz^i \mid i \geq 1\}$) is equivalent to a single equation $xy = yz$ only. Consequently, S_1 is not independent. So an open question arises: Is Theorem 4.4 valid for independent pairs of equations?

Note added in proof

C. Choffrut pointed out to the authors that the answer to the open question is negative (for more discussion see Springer Lecture Notes in Computer Science, Vol. 140, pp. 128–140).

Acknowledgement

The authors are grateful to S. Burris and J. Pachl for discussions of the reported research.

References

[1] J. Albert, K. Culik II and J. Karhumaki, Test sets for context free languages and systems of equations over a free monoid, Inform. and Control, to appear.

[2] J. Berstel, Transductions and Context Free Languages (B. G. Teubner, Stuttgart, 1979).
[3] 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).
[4] K. Culik II and I. Fris, The decidability of the equivalence problem for DOL systems, Inform. and Control 35 (1977) 20–39.
[5] K. Culik II and J. Karhumaki, On the equality sets for homomorphisms on free monoids with two generators, R.A.I.R.O. Theoretical Informatics 14 (1980) 349–369.
[6] K. Culik II and A. Salomaa, On the decidability of homomorphism equivalence for languages, JCSS 17 (1978) 163–175.
[7] K. Culik II and A. Salomaa, Test sets and checking words for homorphism equivalence, JCSS (1980), 379–395.
[8] A. Ehrenfeucht and G. Rozenberg, Simplication of homomorphisms, Inform. and Control 38 (1978) 289–309.
[9] M.A. Harrison, Introduction to Formal Language Theory, (Addison–Wesley, Reading, MA, 1978).
[10] J. Karhumaki and I. Simon, A note on elementary homomorphisms and the regularity of equality sets, EATCS Bulletin 9 (1979) 16–24.
[11] M. Karpinski, ed., New Scottish Book of Problems, in preparation.
[12] M. Linna, The decidability of the DOL prefix problem, Internat. J. of Computer Mathematics 6 (1977), 127–142.
[13] G.S. Makanin, The problem of solvability of equations in a free semigroup (in Russian), Matematiceskij Sbornik 103 (145) (1977) 148–236.
[14] G. Rozenberg and A. Salomaa, The Mathematical Theory of L Systems, (Academic Press, New York, 1980).