数理解析研究所講究録
1166 巻 2000 年 118-130

Some open problems in combinatorics of words and related areas *

J. Karhumäki
Department of Mathematics
and Turku Centre for Computer Science
University of Turku
FIN-20014 Turku, FINLAND
karhumak@cs.utu.fi

Abstract

We discuss several open problems in combinatorics of words and related areas. The problems fall into six different problem areas.

1 Introduction

Combinatorics of words is an enormous source of fascinating open problems. Problems might be of quite a different nature, some asking decision questions and some others existential questions of various types. Often problems are closely related to other areas of discrete mathematics, in particular to algebra or matrices. A striking feature of these problems is that they are very easy to formulate and understand, but in many cases very hard to solve. In other words, they are mathematically extremely challenging.

The goal of this paper is to discuss several such problems. More precisely, we introduce six problem areas and formulate on each of these 1-4 different problems. Problems are by no means "discovered" by the author, but each of those have fascinated the author over the past years.

*Supported under the grant 44087 of the Academy of Finland

The problems we introduce deal with the following questions on different problem areas: How large can an independent system of word equations on n variables be? When does a cumulative defect effect of words hold true? When an equality language of two morphisms is guaranteed to be "simple"? When is the equivalence of two finite substitutions on languages decidable? When do two finite languages commute? Are some simple problems on integer matrices decidable?

2 Preliminaries

As we hinted we need only very standard notions of mathematics, and in particular of words. Consequently, the following lines are mainly to fix our terminology. Whenever necessary the author is referred to [ChK] for undefined notions on words and [HU] on automata.

We denote by Σ a finite alphabet and by Σ^* (resp. Σ^+) the *free monoid* (resp. *free semigroup*) it generates. Elements of Σ^* are called *words* and subsets of Σ^* *languages*. The *empty word*, i.e. the neutral element of Σ^*, is denoted by 1, consequently $\Sigma^+ = \Sigma^* \setminus \{1\}$. Let X be a set of words. A word w admits an *X-factorization* if w can be written as $w = x_1 \dots x_n$ with each x_i in X. This notion extends directly to *one-way infinite* words, that is to set Σ^ω, as well as to *two-way infinite* words, that is to set ${}^\omega\Sigma^\omega$.

An *equation* over Σ^* (or Σ^+) with Ξ as a (finite) set of variables is any pair (u, v) of words in $(\Sigma \cup \Xi)^*$, usually written as $u = v$. A *solution* of an equation $u = v$ is a morphism $\varphi : (\Sigma \cup \Xi)^* \to \Sigma^*$ satisfying $\varphi(u) = \varphi(v)$ and $\varphi(a) = a$ for each $a \in \Sigma$. A system of equations is any set of equations. Two systems of equations are *equivalent* if they have exactly the same solutions. Finally a system is *independent* if it is not equivalent to any of its proper subsystems. In this note we consider only constant-free equations, i.e. equations where $u, v \in \Xi^*$.

For two morphisms $h, g : \Sigma^* \to \Delta^*$ we say that they *are equivalent on a word* $w \in \Sigma^*$ if $h(w) = g(w)$, and that they *are equivalent on the language* $L \subseteq \Sigma^*$ if they are equivalent on each of its words. This notion extends, in a natural way, to more general mappings like finite substitutions, i.e. to morphisms $\sigma : \Sigma^* \to 2^{\Delta^*}$, where 2^{Δ^*} denotes the monoid of finite languages. The maximal set of words on which two morphisms h, g are equivalent is referred to as their *equality language* and is denoted by $E(h, g)$. Hence,

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

Two languages $X, Y \subseteq \Sigma^*$ are said to *commute*, if $XY = YX$. It is straightforward to see that for a given X there exists the unique maximal set commuting with X. Such a set is called the *centralizer* of X and is denoted by $C(X)$. The multiplicative semigroup of $n \times n$ matrices with entries on a semiring S is denoted by $M_{n\times n}(S)$. The cardinality of a set X is denoted by $card(X)$.

3 Independent systems of equations

One of the fundamental properties of words is the Ehrenfeucht Compactness Property of word equations originally formulated as the Ehrenfeucht Conjecture in early 70's. In 1985 it was shown to hold independently in [AL] and [G]. It states that each system of equations over free semigroups having a finite number of variables is equivalent to some of its finite subsystems. In other words, each independent system is finite. The same compactness type of result holds for equations over commutative semigroups cf. e.g. [Re] or [KPII]. Concerning other finitely generated semigroups the property might hold or might not hold as was discussed in [HKP], but no characterization when it holds is known.

Both of the above compactness results are based on Hilbert's Basis Theorem. In the abelian case it is not difficult to see that an independent system can be arbitrarily large, that is it is not bounded by any function on the number of variables. For the word case the problem is much more intriguing leading to the following general question:

> How large can an independent system of equations with n variables over a free semigroup be?

Here it is natural to assume that the equations are constant-free. The above question leads to the following more concrete problems:

Problem 1 *Does there exist a function $f : N \to N$ such that any independent system S of word equations in n variables satisfies $card(S) < f(n)$?*

As related modified problems we state:

Problem 2 *Can f (if exists) in Problem 1 be polynomially bounded or even the function $f(n) = 2^n$?*

Very little is known about these questions. The best known lower bound for the function $f(n)$ is cubic in n that is in the class $\Omega(n^3)$, cf. [KPlI] or [ChK].

The above problems were formulated for equations over free semigroups. If instead free monoids are used then the situation changes slightly, but both of the problems remain. In this case a lower bound for the function $f(n)$ is known to be in $\Omega(n^4)$, cf. again [KPlI] or [ChK].

Another extremely simply formulated problem on independent systems of equations is as follows:

Problem 3 *Does there exist an independent system over a free semigroup consisting of* three *equations with three variables and having a nonperiodic solution?*

Two observations related to this amazing problem are pointed out in the next examples.

Example 1 *Any system of equations with three unknowns and of the form*

$$S: \begin{array}{ll} eq_1: & x\ldots = y\ldots \\ eq_2: & x\ldots = z\ldots \\ eq_3: & anything \end{array}$$

has only periodic solutions in free semigroups. This, indeed follows from the Graph lemma of the next section.

Example 2 *The system*

$$\begin{cases} xyz &= zyx \\ xyyz &= zyyx \end{cases}$$

is an example of two independent equations having a nonperiodic solution. Indeed, it has a solution $x = z = a$ and $y = b$, and the system is independent as shown by the triples $(x, y, z) = (a, b, aba)$ and $(x, y, z) = (a, b, abba)$.

4 Cumulative defect effect

Another fundamental property of words is revealed in so-called *defect theorem*, which states that if a set $X \subseteq \Sigma^+$ of n nonempty words satisfies a nontrivial relation then there exists an $F \subseteq \Sigma^+$ such that

$$X \subseteq F^* \tag{1}$$

and

$$card(F) \leq card(X) - 1. \tag{2}$$

In other words, defining the *combinatorial rank* of X as the minimal cardinality of F satisfying (1), the defect theorem says that any nontrivial relation on X implies that the combinatorial rank of X is at most $n-1$, i.e. X possesses a *defect effect*. We denote by $r(X)$ the combinatorial rank of X.

As discussed in [ChK], there are many different formulations of the defect theorem based on different notions of the rank of a finite set. For our purposes the above combinatorial rank is most suitable.

Now, a natural question arises: If the n words of X satisfies 2, or in general $k \leq n-1$, "different" relations can 1 in (2) be replaced by 2, or in general by k. That is, would these assumptions imply a *cumulative defect effect*. This motivates to formulate a general question:

When does a cumulative defect effect of words hold true?

Two natural directions to study this question are to restrict

(i) type of relations, or

(ii) type of sets X.

Actually very little is known about these questions. Example 2 shows that in general a cumulative defect effect does not hold, for more see [KPlII] or [ChK].

As an example of a cumulative defect effect we recall the following so-called Graph Lemma, cf. [HK] of [ChK]. Let $X \subseteq \Sigma^+$ be a finite set of nonempty words. Define the graph $G_X = (V, E)$ by setting

$$V = X$$

and

u-v$\in E$ if and only if $uX^+ \cap vX^+ \neq \emptyset$. Then we have

Graph Lemma. *For each finite $X \subseteq \Sigma^+$ the combinatorial rank of X is at most the number of connected components of G_X.*

Note that the Graph Lemma was used in Example 1. As another application of it we formulate a variant of the defect theorem for bi-infinite

words. We can interpret a double X-factorization of a word w as a relation on X. Now, we consider bi-infinite relations, i.e. double factorizations of bi-infinite words. We call such a set of X-factorizations of w *disjoint* if no two factorizations match at any point inside w. Then we have, [KMI]

Defect theorem for bi-infinite words. *Let $X \subseteq \Sigma^+$ be finite. If a nonperiodic bi-infinite word possesses two disjoint X-factorizations then $r(X) \leq card(X) - 1$.*

The above result is the first defect theorem for bi-infinite words, and more interestingly it holds only with the combinatorial rank, but not with the other types of ranks considered e.g. in [ChK]. It also allows to formulate a nice problem on cumulative defect effect.

Problem 4 *Let $X \subseteq \Sigma^+$ be finite. Is it true that whenever there exists a nonperiodic bi-infinite word having k disjoint factorizations, for $k \leq n$, then necessarily $r(X) \leq card(X) - k + 1$?*

The case $k = 2$ is answered affirmatively by the above formulation of the defect theorem, and the case $k = n$ is taken care by the famous Critical Factorization Theorem cf. [Lo] or [ChK]. About the other cases only the following is known: Problem 4 has an affirmative answer if $k = 3$ and X is a prefix set, cf. [KMII].

5 Equality languages

We recall that the *equality language* of two morphisms $h, g : \Sigma^* \to \Delta^*$ is the language

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

Consequently, the famous Post Correspondence Problem, cf. [HU] asks to decide whether a given equality language is empty (modulo 1). Since this is undecidable it is not surprising that there are challenging open problems connected to equality languages.

Problem 5 *It the equality language of two binary nonperiodic morphisms always of the form $\{\alpha, \beta\}^*$ for some words $\alpha, \beta \in \Sigma^*$?*

Here by a *binary* morphism we mean a morphism defined on a binary alphabet, and by a *periodic* morphism the one satisfying that all images are powers of a single word. Note also that the words α and β are allowed to be empty. By examples in [CuK] there exist equality languages of this form with α and β having different primitive roots. More interestingly, as was shown in [EKR], all equality languages of binary nonperiodic morphisms are either of the above form or of the form $(\alpha\beta^*\gamma)^*$ for some words α, β, $\gamma \in \Sigma^+$. So the problem is to rule out the second case, which, however, does not seem to be easy.

As another problem we state

Problem 6 *Find the smallest k such that the equality language of two* injective *morphisms h, $g : \Sigma^* \to \Delta^*$ with $card(\Sigma) = k$, can be nonregular.*

By the above discussion $k > 2$, and by an example in [K] $k \leq 5$. Hence, the exact value of k is 3, 4 or 5. This problem, although very special looking, might help to solve the Post Correspondence Problem in the case where $card(\Sigma) = 3$.

6 Equivalence of finite substitutions

In this section we slightly change our emphasis, namely from words to finite languages. We pose the following problem:

Problem 7 *Is it decidable whether two finite substitutions τ, $\sigma : \{a, b, c\}^* \to \Delta^*$ are equivalent on the language $L = ab^*c$.*

The problem might look very special and uninteresting. At the first glance it is clearly decidable. However, it has turned out very intriguing, and after all it is not ruled out that it were undecidable. We give some support for these views.

Firstly, as shown in [La], any fixed finite subset of L is not enough to test whether, for all τ and σ, they are equivalent on L. Secondly, the problem becomes undecidable if L is replaced by $a\{b, c\}^*d$, cf. [Li]. And finally, if the inclusion instead of the equivalence is asked also then the problem becomes undecidable, cf. [KL].

7 Commutation of finite languages

We continue with finite languages and study when do they commute, i.e. we consider $X, Y \subseteq \Sigma^*$ satisfying

$$XY = YX. \tag{3}$$

We recall that for a given X there exists the unique maximal Y satisfying (3). Such a Y is the centralizer of X, in symbols $C(X)$.

Problem 8 *For a given finite $X \subseteq \Sigma^*$ is its centralizer rational?*

This is a problem posed by Conway in [Co] (in a slightly general form assuming that X is rational). In some special cases the answer is known to be affirmative: this is the case if X is a prefix set, cf. [Ra], or $card(X) \leq 3$ cf. [CKO] and [KPe]. On the other hand, in the general case it is only known (and not very difficult to see) that $C(X)$ is in Co-RE. Consequently, also the following seems to be a nontrivial problem.

Problem 9 *For a given finite $X \subseteq \Sigma^*$ is its centralizer recursive?*

Another set of problems associated to the commutation is obtained when all Y's commuting with a given X are looked for. Clearly, any X, $Y \subseteq \Sigma^*$ of the forms

$$X=\cup_{i\in I} V^i \text{ and } Y = \cup_{j\in J} V^j \text{ with } I, J \subseteq N \text{ and } V \subseteq \Sigma^+ \tag{4}$$

commute. An interesting question is whether this, in certain cases, is also a necessary condition for the commutation.

For the family of all finite languages this is not the case, a counterexample being the following four-element set $X = \{a, ab, ba, bb\}$ which commutes with $Y = X \cup X^2 \cup \{bab, bbb\}$, cf. [CKO].

In order to formulate our further problems, let us say that a finite $X \subseteq \Sigma^+$ satisfies *BTC-condition* if, for any Y commuting with X, X and Y can be written in the form (4). Hence, as we saw four-element sets do not satisfy BTC-condition. On the other hand, all prefix sets X satisfy this condition, see [Ra], and so do all binary sets, see [CKO]. So there remains two interesting problems:

Problem 10 *Does every three-element set X satisfy the BTC-condition?*

Problem 11 *Does every finite code $X \subseteq \Sigma^+$ satisfy the BTC-condition?*

As a final comment we recall that all polynomials with noncommuting variables and coefficients in Q, i.e. all finite multisets, satisfy the condition very similar to that in (4), namely the one where unions of powers of V are replaced by one variable polynomials on V. This is a deep result of Bergman, see [Be], as well as the explanation for the abbreviation BTC: *Bergman's Type of Characterization.*

8 Matrix problems

In this final section we turn to problems which are still a bit farther away from words. A connecting point here is the embedding

$$\Sigma^* \hookrightarrow M_{2\times 2}(N)$$

known already to Nielsen in 20's. That is the fact that the word semigroups are subsemigroups of the multiplicative semigroup of matrices over nonnegative integers.

This embedding is extremely useful in both directions: In one hand, it motivates to extend the problems of words into matrices – undecidability results are neat examples of that. And conversely, results on matrices can be used to deduce properties of words – the Ehrenfeucht Compactness Property is a splendid example of that. Our emphasis here is in the first direction.

Let $\mathcal{M} = \{M_1, \ldots, M_t\}$, for $t \geq 1$, be a set of $n \times n$ matrices over a semiring S, i.e. $M_i \in M_{n\times n}(S)$ for all i. Actually, S will be mostly either the set of integers Z or the set of nonnegative integers N. We denote by $\mathcal{M}^*$ the multiplicative semigroup $\mathcal{M}$ generates.

We ask the following natural decision questions:

(i) Is $\mathcal{M}^*$ free?

(ii) Does $\mathcal{M}^*$ contain the zero matrix $\mathbf{0}$?

(iii) Does $\mathcal{M}^*$ contain the identity matrix I?

Problem (i) was shown to be undecidable in [KBS] for 3×3 matrices over N, and this was extended for 3×3 upper triangular matrices in [CHK]. What remains is the following interesting problem:

Problem 12 *Let* $\mathcal{M} = \{M_1, \ldots, M_t\} \subseteq M_{2\times 2}(N)$. *Is it decidable whether* $\mathcal{M}^*$ *is free?*

Surprisingly, even the case $t = 2$ of this problem is unanswered, see again [CHK].

Problem (ii) asks – essentially – whether the existence of the zero element is decidable for certain finitely generated semigroups. Amazingly, this problem is undecidable already in the case when only the semigroup of two integer matrices is considered, see [CaK]. Moreover, the dimension of the matrices can be assumed to be only 45. Further this result can be translated into a general interesting undecidability result of semigroups, namely to the result that the existence of the zero element in 2-generator semigroups is undecidable. The undecidability of Problem (ii) holds also for 3×3-integer matrices, as was shown in [P] already in 1970. The case of 2×2 matrices seems to be open.

A problem related to (ii) is the so-called *Skolem's Problem* asking, whether, for a given $n \times n$ integer matrix M, some of its powers contains zero in the right upper corner. Similarly as Problem (ii) it becomes undecidable if instead of one two matrices are considered, see again [CaK].

From algebraic point of view Problem (iii) asks a very similar question as Problem (ii) does, namely the existence of the unit element in a finitely generated semigroup. However, instead of results we have only problems:

Problem 13 *Let* $\mathcal{M} = \{M_1, \ldots, M_t\} \subseteq M_{n\times n}(Z)$ *(or* $\subseteq M_{n\times n}(Q)$*). Is it decidable whether* $\mathcal{M}^*$ *contains the identity matrix?*

As should be clear there are much more similar open problems on matrices. We have only tried to pick up a few most interesting ones.

References

[AL] Albert, M. H. and Lawrence, J., *A proof of Ehrenfeucht's Conjecture*, Theoret. Comput. Sci **41**, 1985, 121–123.

[Be] Bergman, G. *Centralizers in free associative algebras*, Transactions of the AMS, **137**, 1969, 327–344.

[CHK] Cassaigne, J., Harju, T. and Karhumäki, J., *On the undecidability of freeness of matrix semigroups*, Int. J. Algebra Comput. **9**, 1999, 295–306.

[CaK] Cassaigne, J. and Karhumäki, J. *Examples if undecidable problems for 2-generator semigroups*, Theoret, Comput. Sci **204**, 1998, 29–34.

[ChK] Choffrut, C. and Karhumäki, J. *Combinatorics of words*, in: G. Rozenberg and A. Salomaa (eds), Handbook of Formal Languages, vol 1, 1997, 329-438, Springer-Verlag.

[CKO] Choffrut, C., Karhumäki, J. and Ollinger, N., *The commutation of a finite set: a challenging problem*, Theoret. Comput. Sci (to appear).

[Co] Conway, J. H., *Regular Algebra and Finite machines*, 1971, Chapman & Hall.

[CuK] Culik II, K. and Karhumäki, J, *On equality sets for homomorphisms on free monoids with two generators*, RAIRO, Theor. Informat. **14**, 1980, 349–369.

[EKR] Ehrenfeucht, A., Karhumäki, J. and Rozenberg, G., *On binary equality sets and a solution to the test conjecture in the binary case*, J. Alg. **85** ,1983, 76–85.

[G] Guba, V. S., *The equivalence of infinite systems of equations in free groups and semigroups with finite systems* (in Russian), Mat. Zametki, **40**, 1986, 321–324.

[HK] Harju, T. and Karhumäki, J., *On the defect theorem and simplifiability*, Semigroup Forum **33**, 1986, 199–217.

[HKP] Harju, T., Karhumäki, J. and Plandowski, W., *Compactness of systems of equations in semigroups*, Intern. J. Algebra Comput. **7**, 1997, 457–470.

[HU] Hopcroft, J. and Ullman, J. D., *Introduction to automata theory, languages and computation*, 1979, Addison-Wesley, Reading, MA.

[K] Karhumäki, J. *On the regularity of equality languages*, Ann. Univ. Turkuensis Ser AI, 186, 1984.

[KL] Karhumäki, J. and Lisovik, L. P., *A simple undecidable problem: The inclusion problem for finite substitutions on ab^*c*, manuscript.

[KMI] Karhumäki, J., and Manuch, J., *A defect theorem for bi-infinite words*, Theoret. Comput. Sci., invited.

[KMII] Karhumäki, J., and Manuch, J. *Multiple factorizations and defect effect*, Theoret. Comput. Sci., submitted.

[KPe] Karhumäki, J. and Petre, I., *On the centralizer of a finite set*, Springer LNCS **1853**, 2000, 536–546.

[KPlI] Karhumäki, J. and Plandowski, W., *On the size of independent systems of equations in semigroups*, Theoret. Comput. Sci **168**, 1996, 105–119.

[KPlII] Karhumäki, J. and Plandowski, W., *On the defect effect of many identities in semigroups*, in: Gh. Paun (ed), Mathematical Aspects of Natural and Formal Languages, 1994, 225–233, World-Scientific, Singapore.

[KBS] Klarner, D. A., Birget J.-C. and Satterfield, W., *On the Undecidability of the Freeness of Integer Matrix Semigroups*, Int J. Algebra Comput. **1**, 1991, 223–226.

[La] Lawrence, J., *The nonexistence of finite test set for set-equivalence of finite substitutions*, Bull. EATCS **28**, 1986, 34–37.

[Li] Lisovik, L. P., *The equivalence problem for finite substitutions on regular languages*, Doklady of Academy of Sciences of Russia **357**, 1997, 299–301.

[Lo] Lothaire, M., *Combinatorics on Words*, 1983, Adison-Wesley, Reading, MA.

[P] Paterson, M. S., *Unsolvability in 3×3 matrices*, Studies in Appl. Math. **49**, 1970, 105–107.

[Ra] Ratoandramanana, B., *Codes et motifs*, RAIRO, Theor. Informat., **23**, 1989, 425–444

[Re] Redei, L., *The Theory of Finitely Generated Commutative Semigroups*, Pergamon Press, Oxford, 1965.