Cunsheng Ding | Arto Salomaa

On some problems of Mateescu concerning subword occurrences

Turku Centre *for* Computer Science

TUCS Technical Report
No 701, August 2005

On some problems of Mateescu concerning subword occurrences

Cunsheng Ding
Department of Computer Science
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong, China
cding@cs.ust.hk
Arto Salomaa
Turku Centre for Computer Science
Lemminkäisenkatu 14, 20520 Turku, Finland
asalomaa@it.utu.fi

TUCS Technical Report
No 701, August 2005

Abstract

The paper investigates inference based on quantities $|w|_u$, the number of occurrences of a word u as a scattered subword of w. Parikh matrices recently introduced are useful tools for such investigations. We introduce and study universal languages for Parikh matrices. We also obtain results concerning the inference from numbers $|w|_u$ to w, as well as from certain entries of a Parikh matrix to other entries.

Keywords: subword, scattered subword, number of subwords, Parikh matrix, inference from subwords, ambiguity

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

An aim encountered in many problems concerning words, languages and automata is to get rid of the mathematically awkward noncommutativity, at least to some extent. In *arithmetizing* the theory one reduces noncommutative properties to commutative numerical ones. This is an important feature in the theory of formal power series, [4, 11], and was emphasized in many of Alexandru Mateescu's last works, [6, 9].

The basic numerical quantity investigated in this paper is $|w|_u$, the number of occurrences of a word u as a scattered subword of a word w. Suppose we know the values $|w|_u$ for some specific words u. Can we infer the word w? Or can we infer the values $|w|_v$ for some other specific words v? These are typical problems arising in this context.

The notion of a *Parikh matrix* recently introduced, [7], is the basic technical tool for investigations concerning numbers $|w|_u$. Indeed, the numbers $|w|_u$ appear as entries of the matrix associated morphically to a word w, for some specific finite set U of words u. In the *generalized* version of a Parikh matrix, [18], the set U can include any prechosen words.

In Section 2 we will recall the basic definitions and introduce some new ones, notably the notion of an *M-universal language*. In the following section we develop some technical tools. Section 4 consists of an investigation about M-universal languages. Section 5 deals with both kinds of inference problems referred to above. Also several *open problems* will be mentioned.

We assume that the reader is familiar with the basics of formal languages. Whenever necessary, [11] may be consulted. As customary, we use small letters from the beginning of the English alphabet a, b, c, d, possibly with indices, to denote letters of our formal alphabet Σ. Words are usually denoted by small letters from the end of the English alphabet.

2 Subword occurrences. M-ambiguity and M-universality

The main notion studied in this paper is the number of occurrences of a word u as a *subword* in a word w, in symbols, $|w|_u$. For us the term *subword* means that w, as a sequence of letters, contains u as a subsequence. More formally, we begin with the following fundamental

Definition 1 *A word u is a* subword *of a word w if there exist words $x_1, \ldots, x_n$ and $y_0, \ldots, y_n$, some of them possibly empty, such that*

$$u = x_1 \ldots x_n \text{ and } w = y_0 x_1 y_1 \ldots x_n y_n.$$

The word u is a factor *of w if there are words x and y such that $w = xuy$. If the word x (resp. y) is empty, then u is also called a* prefix *(resp.* suffix*) of w.*

Throughout this paper, we understand subwords and factors in this way. In classical language theory, [11], our subwords are usually called "scattered subwords", whereas our factors are called "subwords". The notation used throughout the article is $|w|_u$, the number of occurrences of the word u as a subword of the word w. Two occurrences are considered different if they differ by at least one position of some letter. (Formally an occurrence can be viewed as a vector of length $|u|$ whose components indicate the positions of the different letters of u in w.)

Clearly, $|w|_u = 0$ if $|w| < |u|$. We also make the *convention* that, for any w and the empty word λ,

$$|w|_\lambda = 1.$$

(The convention is made also in [2, 12].) In [12] the number $|w|_u$ is denoted as a "binomial coefficient"

$$|w|_u = \binom{w}{u}.$$

If w and u are words over a one-letter alphabet,

$$w = a^i,\ u = a^j,$$

then $|w|_u$ equals the ordinary binomial coefficient: $|w|_u = \binom{i}{j}$. Our convention concerning the empty word reduces to the fact that $\binom{i}{0} = 1$.

A general problem arising in this context, and important in many applications, is: How can one construct a set of numbers $|w|_u$ such that the word w is uniquely, or "almost uniquely", determined? For instance, the reader should have no difficulties in proving that the word $w \in \{a, b, c\}^*$ is uniquely determined by the values

$$|w|_a = |w|_b = |w|_c = 3,\ |w|_{ab} = |w|_{bc} = 8.$$

Indeed, $w = a^2babcbc^2$. On the other hand, a word $w \in \{a, b\}^*$ of length 4 is not uniquely determined by the values $|w|_u,\ |u| \leq 2$. Either one of the words *abba* and *baab* can be chosen as w, and still the equations

$$|w|_a = |w|_b = |w|_{ab} = |w|_{ba} = 2,\ |w|_{aa} = |w|_{bb} = 1$$

are satisfied.

For handling such problems a specific tool, referred to as the *Parikh matrix* was introduced in [7]. When dealing with the *extended* notion, [18], one has more leeway in the choice of the words u.

The Parikh matrix is a powerful generalization of a *Parikh mapping (vector)* introduced in [10]. While a Parikh vector only indicates the number of occurrences of each letter in a word, the Parikh matrix gives also information about the mutual positions of the occurrences. The Parikh matrix mapping uses upper triangular square matrices, with nonnegative integer entries, 1's on the main diagonal and 0's below it. The set of all such triangular matrices is denoted by $\mathcal{M}$, and the subset of all matrices of dimension $k \geq 1$ is denoted by $\mathcal{M}_k$.

We are now ready to give the formal definition of a Parikh matrix.

Definition 2 *Let $\Sigma_k = \{a_1, \ldots, a_k\}$ be an alphabet. The* Parikh matrix mapping*, denoted Ψ_k, is the morphism:*

$$\Psi_k : \Sigma_k^* \to \mathcal{M}_{k+1},$$

defined by the following condition. Let $1 \leq q \leq k$ and $\Psi_k(a_q) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$. Then for each $1 \leq i \leq (k+1)$, $m_{i,i} = 1$, $m_{q,q+1} = 1$, all other elements of the matrix $\Psi_k(a_q)$ being 0. Matrices of the form $\Psi_k(w), w \in \Sigma_k^*$, are referred to as Parikh matrices.

Observe that when defining the Parikh matrix mapping we have, similarly as when defining the Parikh vector, in mind a specific *ordering* of the alphabet. The ordering will be clear from the context. If we consider letters without

numerical indices, we assume the alphabetic ordering when numbering the rows and columns of the matrix.

For instance, the Parikh matrix associated to the word *abcbadbcdd* is

$$\Psi_4(abcbadbcdd) = \begin{pmatrix} 1 & 2 & 4 & 5 & 11 \\ 0 & 1 & 3 & 4 & 9 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The following theorem, [7], makes this example clearer. The theorem characterizes the entries of a Parikh matrix in terms of some subword occurrences $|w|_u$. For the alphabet $\Sigma_k = \{a_1, \ldots, a_k\}$, we denote by $a_{i,j}$ the word $a_i a_{i+1} \ldots a_j$, where $1 \leq i \leq j \leq k$.

Theorem 1 *Consider $\Sigma_k = \{a_1, \ldots, a_k\}$ and $w \in \Sigma^*$. The matrix $\Psi_k(w) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$, has the following properties:*

- $m_{i,j} = 0$, *for all* $1 \leq j < i \leq (k+1)$,
- $m_{i,i} = 1$, *for all* $1 \leq i \leq (k+1)$,
- $m_{i,j+1} = |w|_{a_{i,j}}$, *for all* $1 \leq i \leq j \leq k$.

By the *second diagonal* (and similarly the *third diagonal*, etc.) of a matrix in $\mathcal{M}_{k+1}$, we mean the diagonal of length k immediately above the main diagonal. (The diagonals from the third on are shorter than k and will be important in our subsequent discussions.) Theorem 1 tells that the second diagonal of the Parikh matrix of w gives the Parikh vector of w. The next diagonals give information about the order of letters in w by indicating the numbers $|w|_u$ for certain specific words u. Indeed, all factors of the word $a_1 a_2 \ldots a_k$ appear among the words u. The *generalized Parikh matrices*, [18], give information about the numbers $|w|_u$, where the words u can be chosen arbitrarily. The dimension of the matrices grows, depending on the choice of the words u. In this paper we are mainly concerned with the notion of a Parikh matrix given by Definition 2. Thus, for any word w over the alphabet $\{a, b, c\}$, Theorem 1 implies that

$$\Psi_3(w) = \begin{pmatrix} 1 & |w|_a & |w|_{ab} & |w|_{abc} \\ 0 & 1 & |w|_b & |w|_{bc} \\ 0 & 0 & 1 & |w|_c \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This paper deals with some general problems concerning the information content of a Parikh matrix, as well as concerning languages associated to Parikh matrices. In his last research, *Alexandru Mateescu* investigated problems about the *ambiguity* of a matrix. (See, for instance, [6, 8, 9].) To what extent does a matrix determine a word? Sometimes the word is uniquely determined, sometimes this is not the case. These questions have been widely investigated, [3, 6, 7, 9, 13, 14, 15]. There is always a language, sometimes empty but always finite, consisting of words having a given matrix as their associated Parikh matrix. (The language being empty means that the given matrix is not at all a Parikh matrix.) These considerations, as well as the ambiguity issues, lead to many natural problems, as pointed out in [6].

Most of the fundamental notions needed below are introduced in the following definition. The letter "M" in the definition ("M-equivalent", "M-ambiguous", etc.) can be read as *matrix.* However, we would prefer reading it *Mateescu.*

Definition 3 *Let Σ_k and Ψ_k be as in Definition 2. Two words $w_1, w_2 \in \Sigma_k^*$ are termed* M-equivalent*, in symbols $w_1 \equiv_M w_2$, if $\Psi_k(w_1) = \Psi_k(w_2)$. A word $w \in \Sigma_k^*$ is termed* M-unambiguous *if there is no word $w' \neq w$ such that $w \equiv_M w'$. Otherwise, w is termed* M-ambiguous*. If $w \in \Sigma_k^*$ is M-unambiguous (resp. M-ambiguous), then also the Parikh matrix $\Psi_k(w)$ is called unambiguous (resp. ambiguous). The* M-extension *$ME(L)$ of a language L over the alphabet Σ_k consists of all words M-equivalent to some word in L:*

$$ME(L) = \{w' | w' \equiv_M w,\ w \in L\}.$$

Two languages K and L are M-equivalent *in the case their M-extensions coincide: $ME(K) = ME(L)$. A language L is* M-universal *if $ME(L) = \Sigma_k^*$.*

As shown in [9], each of the following three cases is possible for an infinite regular language L: (i) $ME(L)$ is regular, (ii) $ME(L)$ is context-free but not regular, (iii) $ME(L)$ is not context-free.

Clearly the language Σ_k^* is M-universal but there are much smaller M-universal languages. We will see below in Section 4 that both of the languages $a^*b^*a^*b^*a^*$ and $b^*a^*b^*a^*b^*$ are M-universal (with respect to the alphabet $\{a, b\}$), whereas neither one of the languages $a^*b^*a^*b^*$ and $b^*a^*b^*a^*$ is. We will also prove that there is a remarkable difference between two-letter and three-letter alphabets as regards M-universal languages.

The following *open problem* was originally posed by Mateescu. *Is the M-equivalence decidable for regular languages?* In other words, given two regular languages K and L, can we decide whether or not $ME(K) = ME(L)$?

The next result due to [9] (see also [3, 1]) characterizes M-unambiguous words (and matrices) in case of a two-letter alphabet.

Theorem 2 *A word in $\{a, b\}^*$ is M-ambiguous if and only if it contains disjoint occurrences of ab and ba. A word is M-unambiguous if and only if it belongs to the language denoted by the regular expression*

$$a^*b^* + b^*a^* + a^*ba^* + b^*ab^* + a^*bab^* + b^*aba^*.$$

We conclude this section with an example of an infinite sequence of words $\beta_i, i \geq 1$, over the alphabet $\{a, b, c\}$, significant also in considerations of M-universality. By definition,

$$\beta_i = ab^{2^i}cab^{2^{i-1}}cab^{2^{i-2}}c \ldots ab^{2^0}c,\ i \geq 1.$$

It is easy to see that

$$\Psi_3(\beta_i) = \begin{pmatrix} 1 & i+1 & 2^{i+2} - i - 3 & (i-1)2^{i+2} + i + 5 \\ 0 & 1 & 2^{i+1} - 1 & i2^{i+1} + 1 \\ 0 & 0 & 1 & i+1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

3 Auxiliary results

This section contains some notions and results needed below in our considerations about M-universality and interdependence of the elements in a Parikh matrix. Most of these results can be inferred rather directly from the ones presented in the literature. However, we have modified them to suit our specific needs.

All issues dealt with in this paper deal with the general *Inference Problem*: What can be inferred from some specific numbers $|w|_u$, both as regards the word w itself, and as regards some other numbers $|w|_v$? The following two problems are special cases. The first of them was already mentioned in the preceding section.

Word Inference Problem, WI. Considering an alphabet Σ, construct a language $U \subseteq \Sigma^*$ such that every word $w \in \Sigma^*$ is uniquely determined by some numbers $|w|_u,\ u \in U$.

Clearly, there are trivial solutions to Problem WI, for instance, $U = \Sigma^*$. The second of our problems deals with inferring some new numbers $|w|_v$ from given numbers $|w|_u$.

Number Inference Problem, NI. Given a language U and a word v, is the number $|w|_v$ uniquely determined by the numbers $|w|_u,\ u \in U$?

The answer is affirmative if $U = \{a, b, ab\}$ and $v = ba$, whereas $|w|_{abc}$ is not in general uniquely determined by the numbers $|w|_u,\ u \in \{a, b, c, ab, bc\}$. We will return to this question later.

¿From the point of view of Parikh matrices, Problem NI can be interpreted as follows: Which entries of a Parikh matrix are uniquely determined by other entries? (Having in mind the form of a Parikh matrix, we are talking here about entries *above the main diagonal*.) The problem is also closely connected with the problem of the characterization of Parikh matrices. A polynomial-time algorithm for deciding whether a given matrix is a Parikh matrix was given in [9], but a more compact characterization is missing. Let us now look at this problem more closely. How can we fill in the entries of a matrix, starting with the second diagonal, in such a way that the resulting matrix will be a Parikh matrix?

Clearly, the second diagonal can be filled with arbitrary nonnegative integers: the Parikh vector can be quite arbitrary. It is also not difficult to characterize the third diagonal and, thus, the following result, [9], holds. For matrix entries we use the notation of Theorem 1.

Lemma 1 *Arbitrary nonnegative integers may appear on the second diagonal of a Parikh matrix. Arbitrary integers* $m_{i,i+2},\ 1 \leq i \leq k-1$, *satisfying the condition*

$$0 \leq m_{i,i+2} \leq m_{i,i+1}m_{i+1,i+2}$$

(but no others) may appear on the third diagonal of a $(k+1)$*-dimensional Parikh matrix.*

Nothing similar is known for diagonals beginning with the fourth. However, in special cases *all entries* of a Parikh matrix can be inferred from those on the second and third diagonals. In order to quote this result from [14], we first define the notion of a γ*-property*.

Consider the function γ defined by

$$\gamma(m,n) = \begin{cases} \{i | 0 \leq i \leq mn\} & \text{if } m \leq 1 \text{ or } n \leq 1, \\ \{0, 1, mn, mn-1\} & \text{if } m > 1 \text{ and } n > 1. \end{cases}$$

If we use the notation from above, we can write the entries in the third diagonal in the form $m_{i,i+2},\ 1 \leq i \leq k-1$. We say that the *corresponding* entries in the second diagonal are $m_{i,i+1}$ and $m_{i+1,i+2}$. Now the second and third diagonals of a matrix in $\mathcal{M}_{k+1}$, $k \geq 2$, are said to possess the *γ-property* if each entry in the third diagonal is in the set $\gamma(m,n)$, where m and n are the corresponding entries in the second diagonal.

Theorem 3 *Assume that the second, as well as third diagonals, in two matrices M_1 and M_2 in $\mathcal{M}_{k+1}$, $k \geq 2$, coincide and, moreover, possess the γ-property. Then either $M_1 = M_2$, or else the matrices are not both Parikh matrices.*

Theorem 3 says that if we have filled in the entries in the second and third diagonals in such a way that they possess the γ-property, then there is exactly one way (recall here also Lemma 1) to fill the remaining entries to make the matrix a Parikh matrix. Theorem 3 is also a contribution to Problem NI. In the special case of γ-property, the set of numbers $|w|_u$, where u ranges over all factors of length 1 and 2 of the word $a_1 \dots a_k$, determines uniquely the set of numbers $|w|_v$, where v ranges over all factors of the word $a_1 \dots a_k$. This holds for an arbitrary word w.

Consider again the alphabet $\Sigma_k = \{a_1, \dots, a_k\}$. Assume now that $k \geq 3$. Clearly, any word w' obtained from a word w by applying the rewriting rules

$$a_i a_{i+j} \to a_{i+j} a_i, \text{ and } a_{i+j} a_i \to a_i a_{i+j},\ 1 \leq i \leq k-2,\ 2 \leq j \leq k-i,$$

is M-equivalent to w. This follows because these rules do not alter any of the numbers $|w|_u$, where u is a factor of the word $a_1 \dots a_k$. We say that w' is *trivially M-equivalent* to w if it results from w by a sequence of applications of these rules. (In case of a two-letter alphabet, we have no rules of the form considered. Thus, trivial M-equivalence reduces to equality.)

Clearly, two words w and w' are trivially M-equivalent exactly in case the *projections* of w and w' into $\{a_i, a_{i+1}\}^*$ coincide, for each $i,\ 1 \leq i \leq k-1$.

We will only briefly describe the *generalized Parikh matrix* originally due to [18]. While the Parikh matrix mapping yields the numbers $|w|_u$, where u is a factor of the word $a_1 \dots a_k$, the generalized mapping is similarly based on an arbitrary word $v = b_1 ... b_t$, where each b is a letter. (Repetitions of letters are allowed.) The matrices belong now to $\mathcal{M}_{t+1}$, and the matrix corresponding to a letter b has in its second diagonal 1's in the positions corresponding to the occurrences of b in v. (For instance, [18, 13, 15] contain each the formal details.) Then, according to a result corresponding to Theorem 1, the entry $m_{i,1+j},\ 1 \leq i \leq j \leq t$, in the matrix corresponding to a word w equals the number $|w|_{b_i \dots b_j}$.

For instance, choosing $v = baaa$, we obtain for the word $w = abbabaab$

$$\Psi_{baaa}(abbabaab) = \begin{pmatrix} 1 & 4 & 8 & 7 & 2 \\ 0 & 1 & 4 & 6 & 4 \\ 0 & 0 & 1 & 4 & 6 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

This example will be used also in Section 5 when considering the interdependence of matrix elements.

The following theorem, [8, 13, 18], is a powerful tool in establishing interconnections and inequalities between different numbers $|w|_u$.

Theorem 4 *Every minor in a Parikh matrix and in a generalized Parikh matrix assumes a nonnegative integer value.*

Assume that $w \in \{a, b, c, d\}^*$, and denote

$$\Psi_4(w) = \begin{pmatrix} 1 & A & E & H & x \\ 0 & 1 & B & F & I \\ 0 & 0 & 1 & C & G \\ 0 & 0 & 0 & 1 & D \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We consider the element x in the upper right-hand corner as an unknown, and try to express it in terms of the other elements. Considering the two- and three-dimensional minors in the upper right-hand corner, we obtain by Theorem 4 the following result. (The result can of course be expressed also in terms of the numbers $|w|_u$, where each u is a factor of the word *abcd*.)

Lemma 2 *Using the notation introduced above, the following inequalities hold for every word $w \in \{a, b, c, d\}^*$, assuming that $0 < F < BC$:*

$$\frac{CEI + BGH - HI - EFG}{BC - F} \leq x \leq \frac{HI}{F}.$$

In many instances the inequalities suffice to determine x uniquely. Since x is always an integer, the inequalities are actually strict if the fractions are not integers. The assumption concerning F excludes the trivial cases, where every b precedes every c, or every c precedes every b, in w.

For a word $w \in \{a, b, c\}^*$, using the notation

$$\Psi_3(w) = \begin{pmatrix} 1 & A & E & x \\ 0 & 1 & B & F \\ 0 & 0 & 1 & C \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

we obtain similarly:

Lemma 3 *The following inequalities hold for every word $w \in \{a, b, c\}^*$, assuming that $B > 0$:*

$$AF + CE - ABC \leq x \leq \frac{EF}{B}.$$

4 M-universality

In this section we will investigate M-universal languages (recall Definition 3) over the alphabet $\Sigma_k = \{a_1, \ldots, a_k\}$. It turns out that there is an essential difference between the values $k = 2$ and $k = 3$.

We say that a M-universal language is *minimal* if no proper subset of it is M-universal.

We say that a language L is *t-letter bounded*, $t \geq 1$, if

$$L \subseteq b_1^* \dots b_t^*, \ b_i \in \Sigma, \ \text{for } 1 \leq i \leq t.$$

A language is *letter bounded* if it is t-letter bounded, for some t.

Clearly, every letter bounded language is bounded in the usual sense, [11], but bounded languages are not in general letter bounded. If L is t-letter bounded and $w \in L$, then we meet at most $t-1$ changes of letters when reading through w (in either direction).

We begin with the following simple result.

Theorem 5 *No 4-letter bounded language is M-universal with respect to the two-letter alphabet* $\{a, b\}$.

Proof. In this case every 4-letter bounded language is a subset of $a^*b^*a^*b^*$ or of $b^*a^*b^*a^*$. However, neither one of these languages contains the language

$$a^*b^* + b^*a^* + a^*ba^* + b^*ab^* + a^*bab^* + b^*aba^*,$$

consisting of M-unambiguous words, by Theorem 2. Since clearly every M-universal language must contain all M-unambiguous words, our theorem follows.

In contrast to this result, we will show that the following two languages are minimal M-universal:

$$K_a = a^*b^+ab^*a^* \cup a^*b^*, \ K_b = b^*a^*ba^+b^* \cup a^*b^*.$$

Hence, the 5-letter bounded language $a^*b^*a^*b^*a^*$ (as well as $b^*a^*b^*a^*b^*$) is M-universal.

Lemma 4 *Assume that* $L \subseteq \{a, b\}^*$. *Then there is a unique language* $\mu_a(L) \subseteq K_a$ *M-equivalent to* L.

Proof. Consider a word $w \in L$. We determine a unique word $\mu_a(w) \in K_a$ M-equivalent to w. Finally, we define

$$\mu_a(L) = \{\mu_a(w) | w \in L\}.$$

If $|w|_a = 0$, we choose $\mu_a(w) \in b^*$. Thus, assume that $|w|_a \geq 1$. We denote

$$|w|_a = m \geq 1, \ |w|_b = n \geq 0, \ |w|_{ab} = p \geq 0.$$

By Lemma 1, we have $0 \leq p \leq mn$. If $p = mn$, we choose $\mu_a(w)$ in a unique way from the second term of the union K_a. (Every word in the first term of the union contains the factor ba.) Hence, we may assume that $p < mn$. We may write

$$p = \alpha n + \beta, \ 0 \leq \alpha \leq m-1, \ 0 \leq \beta \leq n-1.$$

(Observe that if $n = 0$, then necessarily $p = 0$, and we may again use the second term of the union K_a.) We now choose

$$\mu_a(w) = a^\alpha b^{n-\beta} a b^\beta a^{m-1-\alpha} \in K_a.$$

Clearly,

$$|\mu_a(w)|_a = m, \ |\mu_a(w)|_b = n, \ |\mu_a(w)|_{ab} = p,$$

as required. We still show that the chosen $\mu_a(w)$ is unique, that is, no other word

$$x = a^i b^j a b^r a^s \in K_a,\ i, r, s \geq 0,\ j \geq 1,$$

satisfies these conditions. Thus, we obtain

$$i + s + 1 = m,\ j + r = n,\ i(j + r) + r = p,$$

whence

$$r \leq n - 1,\ p = in + r,\ i \leq m - 1.$$

Consequently,

$$i = \alpha,\ r = \beta,\ j = n - \beta,\ s = m - 1 - \alpha,$$

and we are back in the chosen word $\mu_a(w)$. This completes the proof of Lemma 4.

Lemma 5 *Assume that $L \subseteq \{a, b\}^*$. Then there is a unique language $\mu_b(L) \subseteq K_b$ M-equivalent to L.*

Proof. The proof is similar to that of Lemma 4. After excluding the special cases, we obtain now

$$p = \gamma m + \delta,\ 0 \leq \gamma \leq n - 1,\ 0 \leq \delta \leq m - 1,$$

and choose

$$\mu_b(w) = b^{n-1-\gamma} a^\delta b a^{m-\delta} b^\gamma \in K_b.$$

Uniqueness is shown as before.

As an example, consider the word

$$w = a^2 b^3 a^5 b^7 a^{11} b^{13} a^{17} b^{19} a^{23} b^{29},$$

(indicating the first primes). We obtain

$$\mu_a(w) = a^{37} b^{62} a b^9 a^{20} \text{ and } \mu_b(w) = b^{25} a^{26} b a^{32} b^{45}.$$

Lemmas 4 and 5 yield immediately the following theorem. We assume that the languages K_a and K_b, as well as the mappings μ_a and μ_b, are defined as above.

Theorem 6 *The languages K_a and K_b are minimal M-universal languages. Moreover, two languages over a binary alphabet, $L_1, L_2 \subseteq \{a, b\}^*$, are M-equivalent if and only if $\mu_a(L_1) = \mu_a(L_2)$ (resp. $\mu_b(L_1) = \mu_b(L_2)$).*

Theorem 6 does not yield directly any decision method for the M-equivalence of regular languages over a binary alphabet. The language $\mu_a(L)$ is not necessarily context-free for a regular language L. For instance,

$$\mu_a((abab)^+) = \{a^n b^n a b^n a^{n-1} |\ n \geq 1\}.$$

M-universal languages are much more complicated if the alphabet consists of at least three letters. In fact, we do not know any minimal M-universal language in this case. The following result indicates the complexity in comparison with a binary alphabet.

Theorem 7 *No M-universal language over an alphabet with at least three letters is letter bounded.*

Theorem 7 is established by considering an infinite sequence of words over the alphabet $\{a, b, c\}$. One of the letters is viewed as a boundary marker, and it is shown that an unbounded number of occurrences of the marker, separated by other letters, is required to generate M-equivalent words to all words in the sequence. Such a sequence is defined, for $i \geq 2$, by

$$\delta_i = \prod_{j=1}^{i} (abc^{r^{i-j}}),\ r = i^2 + 1.$$

(Thus, $\delta_3 = abc^{100}abc^{10}abc$.) Then all of the i occurrences of the letter b in δ_i are separated in any word M-equivalent to δ_i. This is seen by an induction on i, viewing the entries $|w|_u$ as integers to the base r.

We still consider two other sequences with the "separation property", interesting also on their own right. Some words in the sequence β_i considered at the end of Section 2 possess M-equivalent words that are not trivially M-equivalent, for instance,

$$\beta_2 = ab^4cab^2cabc \equiv_M b^2a^2b^2cb^2cabc,$$

but the requirement concerning the separation is satisfied. A more sophisticated example is the sequence

$$\alpha_i = \prod_{j=0}^{i} (a^{2^{i-j}}bc^{2^j}),\ i \geq 1.$$

Thus, for $i = 4$, we obtain

$$\alpha_4 = a^{16}bca^8bc^2a^4bc^4a^2bc^8abc^{16},$$

and the matrix

$$\Psi_3(\alpha_4) = \begin{pmatrix} 1 & 31 & 129 & 3216 \\ 0 & 1 & 5 & 129 \\ 0 & 0 & 1 & 31 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In the general case, we indicate only the entries above the main diagonal:

$$\Psi_3(\alpha_i) = \begin{pmatrix} 2^{i+1}-1 & i \cdot 2^{i+1}+1 & \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor}((j+1)i+1-j^2)(2^{2i-j}+2^{i+j}) \\ \dots & i+1 & i \cdot 2^{i+1}+1 \\ \dots & \dots & 2^{i+1}-1 \end{pmatrix}.$$

Here it is understood that if $2i - j = i + j$, that is, if i is even and $j = \frac{i}{2}$, then the sum $2^{2i-j} + 2^{i+j}$ is reduced to 2^{i+j}. The rewriting rules

$$ac \to ca \text{ and } ca \to ca$$

can be applied to α_i to yield trivially M-equivalent words. There are also non-trivial equivalences, for instance,

$$\alpha_3 = a^8bca^4bc^2a^2bc^4abc^8 \equiv_M a^8ba^5bc^3ba^2c^5bc^7,$$

but the requirement of separation is satisfied. In the following table we indicate, for $w = \alpha_i,\ i \leq 5$, the values $|w|_a = |w|_c$, $|w|_b$, $|w|_{ab} = |w|_{bc}$, $|w|_{abc}$, as well as the bounds for $|w|_{abc}$, resulting by Lemma 3, based on the other values of the matrix.

w	$\lvert w\rvert_a$	$\lvert w\rvert_b$	$\lvert w\rvert_{ab}$	$\lvert w\rvert_{abc}$	bounds
α_1	3	2	5	12	12,12
α_2	7	3	17	92	91,96
α_3	15	4	49	576	570,600
α_4	31	5	129	3216	3193,3328
α_5	63	6	321	16704	16632, 17173

5 Inference problems and matrix elements

We will first discuss the *word inference problem, WI*, described in Section 3. It is clear that no such *finite* language U can exist: every finite t-tuple of numbers $|w|_u$ fails to characterize all words of a sufficient length. This fact is well known in the literature, [5, 2]. We present it in the form of the following lemma.

Lemma 6 *Assume that $U \subseteq \Sigma_k^*$, $k \geq 2$, is a set of words of a finite cardinality i, the longest word in U being of length j. Then there is a bound t_0 such that, whenever $t \geq t_0$, there are different words $w, w' \in \Sigma_k^*$ with $|w| = |w'| = t$ such that $|w|_u = |w'|_u$, for every $u \in U$.*

Proof. Choose t_0 large enough such that the inequality

$$k^{t_0} > (t_0)^{ij}$$

is satisfied. Let $t \geq t_0$. There are k^t words over Σ_k of length t. Clearly, a word of length j appears at most $\binom{t}{j} < t^j$ times as a subword of a word of length t. Hence, there are at most t^j possible values for each of the numbers $|w|_u$, $u \in U$, and, consequently, at most $(t^j)^i$ possibilities for the i-tuple of values

$$\{|w|_u | u \in U\}.$$

The choice of t_0 now guarantees that two different words w and w' of length t are assigned the same i-tuple. This proves the lemma.

A set U often considered in the literature is the *t-spectrum* (also called *t-deck*), [2, 5, 13], consisting of *all* words of length $\leq t$. What is the smallest number $f(t)$ such that two different words of length $f(t)$ have the same t-spectrum? No good estimates for $f(t)$ are known, only upper bounds such as the ones resulting from Lemma 6.

It is questionable whether a t-spectrum is at all a good choice for the set U. A much smaller set of words of a specific form may yield the inference of words of the same length as a considerably bigger t-spectrum. As an instance we quote the following result from [15].

Theorem 8 *Assume that w and w' are words over the alphabet $\{a, b\}$ with the same Parikh vector (r, s) and that*

$$|w|_{ab^i} = |w'|_{ab^i}, \ 1 \leq i \leq \min(r, s).$$

Then $w = w'$. For any integer l, a word w of length $\leq l$ over the alphabet $\{a, b\}$ can be uniquely inferred from at most $\lfloor l/2 \rfloor + 2$ specific values $|w|_u$.

For instance, the values

$$|w|_a, \ |w|_b, \ |w|_{ab}, \ |w|_{ab^2}, \ |w|_{ab^3}$$

determine uniquely a word w of length ≤ 7.

By Lemma 6, no finite set U suffices for the inference of all words. However, by Theorem 8, the set

$$U = \{a, b\} \cup \{ab^i | i \geq 1\}$$

suffices for this purpose, as regards words over the binary alphabet. This result can be extended to concern arbitrary alphabets by observing that two words $w, w' \in \Sigma_k^*$ must be equal if their projections to each submonoid

$$\{a_i, a_j\}^*, \ 1 \leq i < j \leq k,$$

coincide. Hence, we obtain the following conclusion.

Theorem 9 *Consider the alphabet* $\Sigma_k = \{a_1, \ldots, a_k\}, \ k \geq 2$, *and define*

$$U = \{a_k\} \cup \{a_i a_j^\nu | 1 \leq i < j \leq k, \ \nu \geq 0\}.$$

If two words $w, w' \in \Sigma_k^*$ *satisfy*

$$|w|_u = |w'|_u, \text{ for all } u \in U, \ |u| \leq \lfloor |w|/2 \rfloor + 1,$$

then $w = w'$.

We will discuss, secondly, the *number inference problem, NI*, described in Section 3. More specifically, we are concerned with the problem: to what extent do some entries in a Parikh matrix determine the other entries in the matrix? Here Theorem 4, as well as Lemmas 2 and 3 are central tools. For instance, the words

$$abcddcbadcba, \ abdccbddacba, \ adcbbacddcba, \ dcbaabcddcba,$$

$$adcdbbaccbda, \ dcdbaabccbad, \ dcdbcaabbacd, \ dcdbcabaabcd$$

are all M-equivalent (no two of them being trivially so). The words have been constructed from words over $\{a, b, c\}$ and $\{b, c, d\}$, each with the Parikh matrix

$$\begin{pmatrix} 1 & 3 & 4 & 4 \\ 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus, the resulting words have the matrix

$$\begin{pmatrix} 1 & 3 & 4 & 4 & x \\ 0 & 1 & 3 & 4 & 4 \\ 0 & 0 & 1 & 3 & 4 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Now it is a consequence of Lemma 2 that $x = 4$ and, thus, the words are M-equivalent.

By Theorem 3, the second and third diagonals of a Parikh matrix determine the matrix uniquely, provided the γ-property is satisfied. If it is not satisfied, then the fourth diagonals can be entirely different in two Parikh matrices with the same second and third diagonals. However, we present the following conjecture.

Conjecture 1 *For some $t \geq 4$, every Parikh matrix (no matter how high the dimension) is uniquely determined by its entries in the diagonals up to the tth.*

It is possible that this conjecture holds even for $t = 4$. Lemmas 2 and 3 can be used to establish it for $t = 4$ in some special cases, for instance, the case where all entries in the 2nd, 3rd and 4th diagonals are equal. Our conjecture does not concern generalized Parikh matrices. A counter example for $t = 4$ is obtained by continuing the example presented in Section 3. We have also

$$\Psi_{baaa}(baababba) = \begin{pmatrix} 1 & 4 & 8 & 7 & 4 \\ 0 & 1 & 4 & 6 & 4 \\ 0 & 0 & 1 & 4 & 6 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and, thus, we have two different generalized Parikh matrices, where the 2nd, 3rd and 4th diagonals coincide.

6 Conclusion

Matrix constructions have turned out to be very useful for inference problems WI and NI. Some basic questions remain open in this area, such as problems concerning the characterization of Parikh matrices and M-universal languages for arbitrary alphabets. An interesting problem area, [17, 16], lying outside the scope of this paper deals with the definition of languages in terms of numbers $|w|_u$. For instance, the language

$$b^*(a^2ba^2 + ab^2a)b^* + b^* \text{ (resp. } \{a^nb^nc^nd^n | n \geq 1\})$$

is defined by the condition $|w|_a = |w|_{aba}$ (resp.

$$|w|_a = |w|_b = |w|_c = |w|_d \& |w|_{abcd} = 24|w|_{a^4} + 36|w|_{a^3} + 14|w|_{a^2} + |w|_a.)$$

References

[1] Atanasiu, A., Martín-Vide, C. and Mateescu, A., On the injectivity of the Parikh matrix mapping. *Fund. Informaticae* 49 (2002) 289–299.

[2] Dudik, M. and Schulman, L.J., Reconstruction from subsequences. *J. Combin. Th. A* 103 (2002) 337–348.

[3] Fossé, S. and Richomme, G., Some characterizations of Parikh matrix equivalent binary words. *Inform. Proc. Lett.* 92 (2004) 77–82.

[4] Kuich, W. and Salomaa, A. *Semirings, Automata, Languages*. Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[5] Manvel, B., Meyerowitz, A., Schwenk, A., Smith, K., Stockmeyer, P., Reconstruction of sequences. *Discrete Math.* 94 (1991) 209–219.

[6] Mateescu, A., Algebraic aspects of Parikh matrices. In J. Karhumäki, H. Maurer, G. Păun and G. Rozenberg (eds.) *Theory is Forever, Springer Lecture Notes in Computer Science 3113* (2004) 170–180.

[7] Mateescu, A., Salomaa, A., Salomaa, K. and Yu, S., A sharpening of the Parikh mapping. *Theoret. Informatics Appl.* 35 (2001) 551–564.

[8] Mateescu, A., Salomaa, A. and Yu, S., Subword histories and Parikh matrices. *J. Comput. Syst. Sci.* 68 (2004) 1–21.

[9] Mateescu, A. and Salomaa, A., Matrix indicators for subword occurrences and ambiguity. *Int. J. Found. Comput. Sci* 15 (2004) 277–292.

[10] Parikh, R.J., On context-free languages. *J. Assoc. Comput. Mach.* 13 (1966) 570–581.

[11] Rozenberg, G. and Salomaa, A. (eds.), *Handbook of Formal Languages 1–3.* Springer-Verlag, Berlin, Heidelberg, New York (1997).

[12] Sakarovitch, J. and Simon, I., Subwords. In M. Lothaire: *Combinatorics on Words*, Addison-Wesley, Reading, Mass. (1983) 105–142.

[13] Salomaa, A., Counting (scattered) subwords. *EATCS Bulletin* 81 (2003) 165–179.

[14] Salomaa, A., On the injectivity of Parikh matrix mappings. *Fundamenta Informaticae* 64 (2005) 391–404.

[15] Salomaa, A., Connections between subwords and certain matrix mappings. *Theoretical Computer Science.* 340 (2005) 188–203.

[16] Salomaa, A., On languages defined by numerical parameters. TUCS Technical Report 663 (2005), submitted for publication.

[17] Salomaa, A. and Yu, S., Subword conditions and subword histories. TUCS Technical Report 633 (2004), submitted for publication.

[18] Şerbănuţă, T.-F., Extending Parikh matrices. *Theoretical Computer Science* 310 (2004) 233–246.

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Sciences

ISBN 952-12-1585-2
ISSN 1239-1891