Theoretical Informatics and Applications
Theoret. Informatics Appl. **35** (2001) 551–564

A SHARPENING OF THE PARIKH MAPPING *

Alexandru Mateescu[1], Arto Salomaa*[2], Kai Salomaa[3] and Sheng Yu[4]

Abstract. In this paper we introduce a sharpening of the Parikh mapping and investigate its basic properties. The new mapping is based on square matrices of a certain form. The classical Parikh vector appears in such a matrix as the second diagonal. However, the matrix product gives more information about a word than the Parikh vector. We characterize the matrix products and establish also an interesting interconnection between mirror images of words and inverses of matrices.

Mathematics Subject Classification. 68Q45, 68Q70.

Introduction

The Parikh mapping (vector) is an old and important tool in the theory of formal languages. This notion was introduced in [8]. One of the important results concerning this mapping is that the image by the Parikh mapping of a context-free language is always a *semilinear* set. (For details and ramifications, see [11].) The basic idea behind Parikh vectors is that properties of words are expressed as *numerical* properties of vectors. However, much information is lost in the transition from a word to a vector.

Keywords and phrases: Formal languages, Parikh mapping, scattered subwords.

* *Dedicated to Aldo de Luca on his 60th birthday. Aldo has been a great colleague and friend during the past decades, always willing to share his deep insights into combinatorial structures. We wish him all the best in the years to come.*

[1] Faculty of Mathematics, University of Bucharest, Academiei 14, Bucharest, Romania; e-mail: alexmate@pcnet.pcnet.ro

[2] Turku Centre for Computer Science, Lemminkäisenkatu 14, 20520 Turku, Finland; e-mail: asalomaa@utu.fi

[3] Department of Computing and Information Science, Queen's University, Kingston, Ontario K7L 3N6, Canada; e-mail: ksalomaa@cs.queensu.ca

[4] Department of Computer Science, University of Western Ontario, London, Ontario N6A 5B7, Canada; e-mail: syu@csd.uwo.ca

In this paper we introduce a sharpening of the Parikh mapping, where somewhat more information is preserved than in the original Parikh mapping. The new mapping is based on a certain type of matrices. The classical Parikh vector will appear in such a matrix as the second diagonal. All other entries above the main diagonal contain information about the *order* of letters in the original word. All matrices are triangular, with 1's on the main diagonal and 0's below it.

Two words with the same Parikh matrix always have the same Parikh vector, but two words with the same Parikh vector have in many cases different Parikh matrices. Thus, the Parikh matrix gives more information about a word than a Parikh vector. The exact meaning of the entries in a Parikh matrix is given below in Theorem 2.1. Our second main result, Theorem 3.2, shows an interesting interconnection between the inverse of a Parikh matrix and the Parikh matrix of the mirror image.

We start with some basic notations and definitions. The set of all nonnegative integers is denoted by N. Let Σ be an alphabet. The set of all words over Σ is Σ^* and the empty word is λ. If $w \in \Sigma^*$ then $|w|$ denotes the length of w.

In this paper we very often use "ordered" alphabets. An ordered alphabet is an alphabet $\Sigma = \{a_1, a_2, \dots a_k\}$ with a relation of order ("$<$") on it. If we have $a_1 < a_2 < \cdots < a_k$, then we use the notation

$$\Sigma = \{a_1 < a_2 < \cdots < a_k\}\,.$$

Let $a \in \Sigma$ be a letter. The number of occurrences of a in a word $w \in \Sigma^*$ is denoted by $|w|_a$. Let u, v be words over Σ. The word u is a *scattered subword* of v if there exists a word t such that $v \in u \sqcup\!\sqcup\, t$, where $\sqcup\!\sqcup$ denotes the shuffle operation. We now introduce a *notation* very important in our subsequent considerations.

If $u, v \in \Sigma^*$, then the number of occurrences of u in v as a scattered subword is denoted by $|v|_{\mathrm{scatt}-u}$. For instance,

$$|acbb|_{\mathrm{scatt}-ab} = 2,\ |acba|_{\mathrm{scatt}-ab} = 1 \text{ and } |aabbc|_{\mathrm{scatt}-abc} = 4.$$

Thus, partially overlapping occurrences of a word as a scattered subword are counted as distinct occurrences. The number $|v|_{\mathrm{scatt}-u}$ is denoted as a binomial coefficient in [10]. Indeed, we are back to ordinary binomial coefficients if we are dealing with a one-letter alphabet.

Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet. The *Parikh mapping* $\Psi : \Sigma^* \to N^k$, is defined by

$$\Psi(w) = (|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_k}).$$

The *Parikh vector* of w is $(|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_k})$. Note that the Parikh mapping Ψ is a morphism from the monoid $(\Sigma^*, \cdot, \lambda)$ to the monoid $(N^k, +, (0, 0, \dots, 0))$.

The *mirror image* of a word $w \in \Sigma^*$, denoted $\mathrm{mi}(w)$, is defined as: $\mathrm{mi}(\lambda) = \lambda$ and $\mathrm{mi}(b_1 b_2 \dots b_n) = b_n \dots b_2 b_1$, where $b_i \in \Sigma$, $1 \le i \le n$.

Our exposition is largely self-contained. The reader is referred to [9] as a comprehensive treatment on formal languages and diverse background material. The most fundamental applications and interconnections of Parikh vectors with language theory are presented in [11].

Semilinearity plays an important role in the study of language families suitable as models in linguistics, for instance, see [5]. We refer the reader also to [2] for recent results about the preservation of semilinearity under certain machine mappings and its significance to decision problems.

1. Parikh mapping extended to matrices

We consider a special type of matrices, called "triangle" matrices. A *triangle matrix* is a square matrix $M = (m_{i,j})_{1 \leq i,j \leq k}$, such that $m_{i,j} \in N$, for all $1 \leq i, j \leq k$, $m_{i,j} = 0$, for all $1 \leq j < i \leq k$, and, moreover, $m_{i,i} = 1$, for all $1 \leq i \leq k$.

The set of all triangle matrices is denoted by $\mathcal{M}$. The set of all triangle matrices of dimension $k \geq 1$ is denoted by $\mathcal{M}_k$. Clearly, $\mathcal{M}_k$ constitutes a monoid under matrix multiplication.

We are now ready to introduce the main notion of this paper.

Definition 1.1. Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet, where $k \geq 1$. The *Parikh matrix mapping*, denoted Ψ_{M_k}, is the morphism:

$$\Psi_{M_k} : \Sigma^* \to \mathcal{M}_{k+1},$$

defined by the condition: if $\Psi_{M_k}(a_q) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$, then for each $1 \leq i \leq (k+1)$, $m_{i,i} = 1$, $m_{q,q+1} = 1$, all other elements of the matrix $\Psi_{M_k}(a_q)$ being 0.

Consider the following examples. Let Σ be the ordered alphabet $\{a < b < c\}$ and assume that $w = bbaac$. Then $\Psi_{M_3}(w)$ is a 4×4 triangle matrix that can be computed as follows:

$$\begin{aligned}
\Psi_{M_3}(bbaac) &= \Psi_{M_3}(b)\Psi_{M_3}(b)\Psi_{M_3}(a)\Psi_{M_3}(a)\Psi_{M_3}(c) \\
&= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
\end{aligned}$$

Another example is given by $w' = aabbc$, for which we obtain

$$\Psi_{M_3}(w') = \Psi_{M_3}(aabbc) = \Psi_{M_3}(a)\Psi_{M_3}(a)\Psi_{M_3}(b)\Psi_{M_3}(b)\Psi_{M_3}(c) = \begin{pmatrix} 1 & 2 & 4 & 4 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

On the other hand,

$$\Psi_{M_3}(acb) = \Psi_{M_3}(cab) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, the Parikh matrix mapping is not injective. One of the major open problems is to characterize non-injectivity, that is, to provide some natural conditions for two words to possess the same Parikh matrix. This problem is closely linked with the fundamental problem about the information content of a Parikh matrix: how much does the Parikh matrix tell about a word?

It was brought to our attention by one of the referees that the term *Parikh matrix* was used in [7] for the growth matrix of a morphism (or a D0L system).

2. Significance of the entries of a Parikh matrix

In this section we characterize the entries of the Parikh matrix. We first introduce some notation that will be applied in our first theorem. Recall also the notation $|v|_{\mathrm{scatt}-u}$ defined in the Introduction.

Consider the ordered alphabet $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$, where $k \geq 1$. We denote by $a_{i,j}$ the word $a_i a_{i+1} \ldots a_j$, where $1 \leq i \leq j \leq k$.

We are now ready to prove the basic property of the Parikh matrix mapping.

Theorem 2.1. *Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet, where $k \geq 1$, and assume that $w \in \Sigma^*$. The matrix $\Psi_{M_k}(w) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$, has the following properties:*

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

Proof. Obviously the first two properties, *(i)* and *(ii)* are true. Now we prove the property *(iii)*. Assume that $|w| = n$. The proof is by induction on n. If $n \leq 1$, then clearly the assertion is true.

Assume now that the assertion *(iii)* is true for all words of length at most n and let w be of length $n+1$. Hence $w = w'a_i$, where $|w'| = n$ and $a_i \in \Sigma$ with $1 \leq i \leq k$.

It follows that:

$$\Psi_{M_k}(w) = \Psi_{M_k}(w'a_i) = \Psi_{M_k}(w')\Psi_{M_k}(a_i).$$

Assume that

$$\Psi_{M_k}(w') = \begin{pmatrix} 1 & m'_{1,2} & \dots & \dots & m'_{1,k+1} \\ 0 & 1 & \dots & \dots & m'_{2,k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & m'_{k,k+1} \\ 0 & 0 & \dots & \dots & 1 \end{pmatrix}.$$

By the inductive hypothesis the matrix $\Psi_{M_k}(w')$ has the property (iii).

From Definition 1.1, we deduce that

$$\Psi_{M_k}(a_i) = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & \dots & & 1 \end{pmatrix}.$$

All elements in the matrix $\Psi_{M_k}(a_i)$ are zero except that the elements on the main diagonal are 1 and also the element on the position $(i, i+1)$ is 1.

Therefore, the matrix $\Psi_{M_k}(w)$ equals

$$\begin{pmatrix} 1 & m'_{1,2} & \dots & \dots & m'_{1,k+1} \\ 0 & 1 & \dots & \dots & m'_{2,k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & m'_{k,k+1} \\ 0 & 0 & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & \dots & & 1 \end{pmatrix} = M.$$

The resulting matrix, $M = (m_{p,q})_{1 \le p,q \le k+1}$ has the property that $m_{j,i+1} = m'_{j,i} + m'_{j,i+1}$, for all j, $1 \le j \le i$ and, for all other indices, $m_{p,q} = m'_{p,q}$.

This completes the inductive step, because the number of occurrences of $a_{j,i} = a_j \dots a_i$ (as a scattered subword) in w equals the sum of the number of occurrences of $a_{j,i}$ in w' and the number of occurrences of $a_{j,i-1}$ in w'. Thus, Theorem 2.1 follows. □

Corollary 2.2. *The matrix $\Psi_{M_k}(w)$ has as the second diagonal (i.e., the vector $(m_{1,2}, m_{2,3}, \dots, m_{k,k+1})$) the Parikh vector of w, i.e., $(m_{1,2}, m_{2,3}, \dots, m_{k,k+1}) = (|w|_{a_1}, |w|_{a_2}, \dots, |w|_{a_k})$.*

As already pointed out, the Parikh matrix mapping gives more information about a word than the classical Parikh mapping, although the Parikh matrix

mapping is still not injective. Injectivity would of course mean that the information given by Parikh matrices is *complete*. This would be more than one can reasonably hope for: one cannot expect that words could be expressed as matrices in this fashion, which would give all information in a simple numerical form.

So far very little is known about *sets of Parikh matrices associated to languages belonging to a fixed family* such as the families in the Chomsky hierarchy. The following remark shows that the semilinearity result of context-free languages does not carry over to sets of matrices.

Remark 2.3. Consider the ordered alphabet $\{a < b\}$ and the context-free language $L = \{a^n b^n | n \geq 1\}$. Clearly,

$$\Psi_{M_2}(a^n b^n) = \begin{pmatrix} 1 & n & n^2 \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence $\Psi_{M_2}(L)$ cannot be a semilinear set (for any reasonable extension of the definition of semilinearity to matrices).

Clearly, every triangle matrix is not a Parikh matrix of some word. For instance, the matrix

$$\begin{pmatrix} 1 & 2 & 7 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$$

is not a Parikh matrix. This follows because ab occurs as a scattered subword at most 6 times in a word with the Parikh vector (2,3). In fact, we have the following immediate corollary of Theorem 2.1:

Corollary 2.4. *The entries $m_{i,j+1}$, $1 \leq i < j \leq k$ in a Parikh matrix $\Psi_{M_k}(w)$ satisfy the inequality*

$$m_{i,j+1} \leq m_{i,j} \cdot m_{i+1,j+1}.$$

Various strengthenings of Corollary 2.4 can be obtained. For instance, the product of the entries in the Parikh vector constitutes an upper bound for the entry $m_{1,k+1}$. Thus, the size of the entry $m_{1,3}$ in Remark 2.3 is maximal. On the other hand, upper bounds can be satisfied and yet the matrix is not a Parikh matrix. For instance, $x = rst$ is the only possible value of x for the matrix

$$\begin{pmatrix} 1 & r & rs & x \\ 0 & 1 & s & st \\ 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

to be a Parikh matrix, whereas in the matrix

$$\begin{pmatrix} 1 & r & x \\ 0 & 1 & s \\ 0 & 0 & 1 \end{pmatrix}$$

any x with $0 \leq x \leq rs$ is possible. Whether or not a given triangle matrix is a Parikh matrix is clearly a decidable question.

3. Inverses of matrices versus mirror images of words

This section investigates interrelations between the inverse of a Parikh matrix associated to a word w and the Parikh matrix of $\mathrm{mi}(w)$, the mirror image of w. Clearly, the set of all triangle matrices of order $k \geq 2$ with integer entries is a noncommutative group with respect to multiplication, the unit element being the unit matrix of order k. Consequently, for each Parikh matrix A, there exists the inverse matrix A^{-1}.

Definition 3.1. Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet and let $w \in \Sigma^*$ be a word. Assume that the Parikh matrix of w is $\Psi_{M_k}(w) = (m_{i,j})_{1 \leq i,j \leq k+1}$. The *alternate Parikh matrix of* w, denoted $\overline{\Psi}_{M_k}(w)$, is the matrix $(m'_{i,j})_{1 \leq i,j \leq k+1}$, where $m'_{i,j} = (-1)^{i+j} m_{i,j}$, for all $1 \leq i, j \leq k+1$.

Observe that the mapping $\overline{\Psi}_{M_k}(w)$ is a morphism of Σ^*. For the Parikh vector Ψ and for every word w, $\Psi(w) = \Psi(\mathrm{mi}(w))$. However, for the Parikh matrix mapping the situation is completely different. The next theorem reveals the interrelation between the inverse of the Parikh matrix of a word w and the alternate Parikh matrix of the mirror image of w.

Theorem 3.2. *Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet and let $w \in \Sigma^*$ be a word. Then:*

$$[\Psi_{M_k}(w)]^{-1} = \overline{\Psi}_{M_k}(\mathrm{mi}(w)).$$

Proof. The proof is by induction on $n = |w|$.

If $n = 1$, then $w = a_q$ for some $1 \leq q \leq k$. By Definition 1.1, $\Psi_{M_k}(a_q) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$, such that for each $1 \leq i \leq (k+1)$, $m_{i,i} = 1$, $m_{q,q+1} = 1$ and all other elements of the matrix $\Psi_{M_k}(a_q)$ are zero.

It is easy to verify that $[\Psi_{M_k}(a_q)]^{-1} = (m'_{i,j})_{1 \leq i,j \leq (k+1)}$, so that for each $1 \leq i \leq (k+1)$, $m'_{i,i} = 1$, $m'_{q,q+1} = -1$ and all other elements of the matrix $[\Psi_{M_k}(a_q)]^{-1}$ are zero.

Hence, $[\Psi_{M_k}(w)]^{-1} = \overline{\Psi}_{M_k}(\mathrm{mi}(w))$.

For the inductive step assume that the Theorem 3.2 is true for all words $u \in \Sigma^*$, with $|u| \leq n$ and let $w \in \Sigma^*$ be a word with $|w| = n+1$. Then $w = xa_p$ such that $|x| = n$ and $a_p \in \Sigma$ with $1 \leq p \leq k$.

We see that

$$[\Psi_{M_k}(w)]^{-1} = [\Psi_{M_k}(xa_p)]^{-1} = [\Psi_{M_k}(x)\Psi_{M_k}(a_p)]^{-1} = [\Psi_{M_k}(a_p)]^{-1}[\Psi_{M_k}(x)]^{-1}.$$

Assume that

$$\overline{\Psi}_{M_k}(\mathrm{mi}(x)) = \begin{pmatrix} 1 & m_{1,2} & \dots & \dots & m_{1,k+1} \\ 0 & 1 & \dots & \dots & m_{2,k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & m_{k,k+1} \\ 0 & 0 & \dots & \dots & 1 \end{pmatrix}.$$

By the inductive hypothesis $[\Psi_{M_k}(x)]^{-1} = \overline{\Psi}_{M_k}(\mathrm{mi}(x)) = (m_{i,j})_{1\le i,j\le k+1}$, where $m_{i,i} = 1$, $1 \le i \le k+1$, $m_{i,j} = 0$, $1 \le j < i \le k+1$ and $m_{i,j+1} = (-1)^{i+j+1}|\mathrm{mi}(x)|_{\mathrm{scatt}-a_{i,j}}$, for all $1 \le i \le j \le k$,

We know by the proof for $n = 1$ that

$$[\Psi_{M_k}(a_p)]^{-1} = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & \dots & & 1 \end{pmatrix}.$$

Thus, all elements in the matrix $[\Psi_{M_k}(a_p)]^{-1}$ are zero except that the elements on the main diagonal are 1, and the element on the position $(p, p+1)$ is -1.

Therefore $[\Psi_{M_k}(w)]^{-1}$ equals

$$\begin{pmatrix} 1 & 0 & & \dots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & \dots & & 1 \end{pmatrix} \begin{pmatrix} 1 & m_{1,2} & \dots & \dots & m_{1,k+1} \\ 0 & 1 & \dots & \dots & m_{2,k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & m_{k,k+1} \\ 0 & 0 & \dots & \dots & 1 \end{pmatrix} = M'.$$

The resulting matrix, $M' = (m'_{i,j})_{1\le i,j\le k+1}$ has the property that

$$\begin{aligned} m'_{p,j} &= m_{p,j} - m_{p+1,j} \\ &= (-1)^{p+j}|\mathrm{mi}(x)|_{\mathrm{scatt}-a_{p,j-1}} - (-1)^{p+j+1}|\mathrm{mi}(x)|_{\mathrm{scatt}-a_{p+1,j-1}} \\ &= (-1)^{p+j}|a_p\mathrm{mi}(x)|_{\mathrm{scatt}-a_{p,j-1}} = (-1)^{p+j}|\mathrm{mi}(w)|_{\mathrm{scatt}-a_{p,j-1}}, \end{aligned}$$

for all $p < j \le k+1$. For all other indices, $m'_{i,q} = m_{i,q}$.

This completes the induction and proves Theorem 3.2. □

Observe that Theorem 3.2 provides a very simple method to compute the inverse of a Parikh matrix. One can also apply it directly to matrices: inverses of matrices of a certain type can be computed in this way.

As an example, consider the ordered alphabet $\Sigma = \{a < b < c\}$ and assume that $w = cbbaa$. Then

$$\Psi_{M_3}(cbbaa) = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since $\mathrm{mi}(cbbaa) = aabbc$, we have by Theorem 3.2:

$$[\Psi_{M_3}(cbbaa)]^{-1} = \overline{\Psi}_{M_3}(aabbc) = \begin{pmatrix} 1 & -2 & 4 & -4 \\ 0 & 1 & -2 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A special relation between $|w|_{\mathrm{scatt}-a_{i,j}}$ and $|\mathrm{mi}(w)|_{\mathrm{scatt}-a_{i,j}}$ is obtained in the next corollary. In the statement the last vertical bars stand for the absolute value.

Corollary 3.3. *Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet and let $w \in \Sigma^*$ be a word. Assume that the Parikh matrix of w is $\Psi_{M_k}(w) = (m_{i,j})_{1 \leq i,j \leq k+1}$, and that $[\Psi_{M_k}(w)]^{-1} = (m'_{i,j})_{1 \leq i,j \leq k+1}$. Then $|\mathrm{mi}(w)|_{\mathrm{scatt}-a_{i,j}} = |(m'_{i,j+1})|$ for all $1 \leq i, j \leq k$.*

4. Computing the inverse of a Parikh matrix

We consider now another method to compute the inverse of a Parikh matrix. We begin with some further definitions and notations.

Let $(A, <)$ be an ordered set. The *dual order* of the order $<$, denoted $<^\circ$, is defined as:

$$a <^\circ b \text{ iff } b < a.$$

Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet. The *dual ordered alphabet*, denoted $\Sigma_\circ$, is $\Sigma_\circ = \{a_k < a_{k-1} < \cdots < a_1\}$.

Consider the ordered alphabet $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ and let $w \in \Sigma^*$ be a word. The Parikh matrix associated to w with respect to the dual order on Σ is denoted by $\Psi_{M_k,\circ}(w)$.

Let $v = (v_1, v_2, \ldots, v_n)$ be a vector. The *reverse* of v, denoted $v^{(\mathrm{rev})}$, is the vector $v^{(\mathrm{rev})} = (v_n, v_{n-1}, \ldots, v_1)$.

Now we introduce the notion of a reverse of a triangle matrix. Let $M = (m_{i,j})_{1 \leq i,j \leq n}$ be a triangle matrix. The *reverse* of M, denoted $M^{(\mathrm{rev})}$, is the matrix $M^{(\mathrm{rev})} = (m'_{i,j})_{1 \leq i,j \leq n}$, where $m'_{i,j} = m_{n+1-j,n+1-i}$, for all $1 \leq i < j \leq n$. (The entries on and below the main diagonal are the same in M and $M^{(\mathrm{rev})}$.)

Note that $M^{(\mathrm{rev})}$ is also a triangle matrix. An easy way to obtain $M^{(\mathrm{rev})}$ is to reverse in M all diagonals that are parallel to the main diagonal. For instance,

$$\text{If } M = \begin{pmatrix} 1 & 2 & 3 & 7 \\ 0 & 1 & 4 & 5 \\ 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ then } M^{(\mathrm{rev})} = \begin{pmatrix} 1 & 6 & 5 & 7 \\ 0 & 1 & 4 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A further method to compute $M^{(\mathrm{rev})}$ is to consider the transpose of $M = (m_{i,j})_{1\le i,j\le n}$, *i.e.*, $M^t = (m'_{i,j})_{1\le i,j\le n}$, where $m'_{i,j} = m_{j,i}$, for all $1 \le i, j \le n$.

The matrix M^t defines in a natural way a square. Let c be the geometrical center of this square. In order to compute $M^{(\mathrm{rev})}$, one has to replace each element $m_{i,j}$ by the element symmetric with respect to c. (Note that the geometrical center is an entry of M exactly in the case when the dimension of M is odd.)

The reader can easily verify the following proposition. (Observe that Def. 3.1 can be immediately extended to concern arbitrary matrices A.)

Proposition 4.1. *Let A, B be two triangle matrices of the same dimension. Then*

(i) $[A^{(\mathrm{rev})}]^{(\mathrm{rev})} = A$;
(ii) $(AB)^{(\mathrm{rev})} = B^{(\mathrm{rev})} A^{(\mathrm{rev})}$;
(iii) $\overline{\overline{A}} = A$;
(iv) $\overline{AB} = \overline{A}\,\overline{B}$.

The next theorem gives another method of computing the inverse of a Parikh matrix:

Theorem 4.2. *Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet and let $w \in \Sigma^*$ be a word. Then*

$$[\Psi_{M_k}(w)]^{-1} = [\overline{\Psi}_{M_k,\circ}(w)]^{(\mathrm{rev})}.$$

Proof. The proof is by induction on $n = |w|$. If $n = 1$, then $w = a_q$ for some $1 \le q \le k$. It follows that $\Psi_{M_k}(a_q)$ is the triangle matrix having 1 on the main diagonal and on the position $(q, q+1)$, all other entries being 0.

In the dual order the letter a_q appears on the position $k+1-q$. Hence, the matrix $\Psi_{M_k,\circ}(a_q)$ is the triangle matrix having the entry 1 on the main diagonal and on the position $(k+1-q, k+2-q)$, all other entries being 0.

The alternate Parikh matrix $\overline{\Psi}_{M_k,\circ}(a_q)$ is the triangle matrix having the entry 1 on the main diagonal, -1 on the position $(k+1-q, k+2-q)$, all other entries being 0.

We deduce that $[\overline{\Psi}_{M_k,\circ}(a_q)]^{(\mathrm{rev})}$ is the triangle matrix, where 1 is on the main diagonal, the value on the position $(k+2-(k+2-q), k+2-(k+1-q)) = (q, q+1)$ is -1 and all other elements are 0.

One can easily verify that this is exactly the inverse of the matrix $\Psi_{M_k}(a_q)$.

For the inductive step, assume that $w = w'a_i$, where $|w'| = n$ and $a_i \in \Sigma$.

Now using the inductive hypothesis and Proposition 4.1(ii) we obtain (recall also that $\overline{\Psi}_{M_k,\circ}(w)$ is a morphism):

$$\begin{aligned}[\Psi(w)]^{-1} &= [\Psi(w'a_i)]^{-1} = [\Psi(a_i)]^{-1}[\Psi(w')]^{-1} \\ &= [\overline{\Psi}_{M_k,\circ}(a_i)]^{(\mathrm{rev})}[\overline{\Psi}_{M_k,\circ}(w')]^{(\mathrm{rev})} = [\overline{\Psi}_{M_k,\circ}(w')\overline{\Psi}_{M_k,\circ}(a_i)]^{(\mathrm{rev})} \\ &= [\overline{\Psi}_{M_k,\circ}(w'a_i)]^{(\mathrm{rev})} = [\overline{\Psi}_{M_k,\circ}(w)]^{(\mathrm{rev})}.\end{aligned}$$

□

As an illustration, observe first that by Theorem 2.1 we have

$$\Psi_{M_3}(cbbaa) = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Consider the dual ordered alphabet $\Sigma_\circ = \{c < b < a\}$. Thus,

$$\Psi_{M_3,\circ}(cbbaa) = \begin{pmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The alternate Parikh matrix of the matrix $\Psi_{M_3,\circ}(cbbaa)$ is

$$\overline{\Psi}_{M_3,\circ}(cbbaa) = \begin{pmatrix} 1 & -1 & 2 & -4 \\ 0 & 1 & -2 & 4 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Finally, the reverse matrix of the above matrix is:

$$[\overline{\Psi}_{M_3,\circ}(cbbaa)]^{(\mathrm{rev})} = \begin{pmatrix} 1 & -2 & 4 & -4 \\ 0 & 1 & -2 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This last matrix is indeed the inverse of the matrix $\Psi_{M_3}(cbbaa)$.

The above Theorem 4.2 provides a simpler method to compute the inverse of a Parikh matrix. Here we have to reverse a matrix that is of a fixed size (card(Σ)+1), whereas in the case of Theorem 3.2 we have to reverse the word w that can be arbitrarily long.

From Theorems 3.2 and 4.2 we deduce:

Corollary 4.3. *Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet and let $w \in \Sigma^*$ be a word. Then:*

$$\Psi_{M_k}(\mathrm{mi}(w)) = \Psi_{M_k,\circ}(w)^{(\mathrm{rev})}.$$

The subsequent final observation concerning the functions introduced is rather obvious. Consider the following four functions from $\mathcal{M}_k$ to $\mathcal{M}_k$: the identity I, the mapping $-$ of A to $\overline{A}$, the mapping (rev) of A to $A^{(\mathrm{rev})}$ and the mapping $\overline{(\mathrm{rev})}$ of A to $\overline{A}^{(\mathrm{rev})}$. Then these four functions together with the operation of composition constitute a group and, moreover, this is the well-known Four-Group of Klein.

5. Some further problems

Our next proposition gives a simple method of deciding whether or not a nonzero number appears in the upper right-hand corner of some power of a given Parikh matrix.

Proposition 5.1. *Let $k \geq 1$ be an integer and let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet. Assume that $w \in \Sigma^*$ is a word with $\Psi_{M_k}(w) = M = (m_{i,j})_{1\leq i,j\leq k+1}$. The following assertions are equivalent:*

- *(i) there is an integer $p \geq 1$ such that for the matrix $M^p = (m_{i,j}^{(p)})_{1\leq i,j\leq k+1}$, $m_{1,k+1}^{(p)} \neq 0$;*
- *(ii) there is an integer $p' \geq 1$ such that in the matrix $M^{p'} = (m_{i,j}^{(p')})_{1\leq i,j\leq k+1}$, for all $1 \leq i \leq j \leq k+1$, $m_{i,j}^{(p')} \neq 0$;*
- *(iii) for all i, $1 \leq i \leq k$, we have $m_{i,i+1} \neq 0$, i.e., the Parikh vector of w has all components nonzero.*

Proof. $(i) \Rightarrow (ii)$ From Theorem 2.1 it follows that the word $v = w^p$ has as a scattered subword the word $a_1a_2\ldots a_k$. Therefore v has as scattered subwords all words of the form $a_{i,j}$, where $1 \leq i \leq j \leq k$. Thus, again by Theorem 2.1, we conclude that for all $1 \leq i \leq j \leq k+1$, $m_{i,j}^{(p)} \neq 0$.

$(ii) \Rightarrow (iii)$ Obviously, since $w^{p'}$ has the Parikh vector with all components nonzero, it follows that w has the Parikh vector with all components nonzero.

$(iii) \Rightarrow (i)$ Consider the word $u = w^k$. One can easily deduce that u has as a scattered subword the word $a_1a_2\ldots a_k$. Again by Theorem 2.1 we conclude the assertion (i). □

Observe that from the above proof of $(iii) \Rightarrow (i)$ we conclude an upper bound for the power p, namely, $p \leq k$. In the following remark we list some simple facts concerning injectivity.

Remark 5.2. Let $\Sigma = \{a_1 < a_2 < \cdots < a_k\}$ be an ordered alphabet. The following assertions are true:

(i) if L is a strictly bounded language over Σ, *i.e.*, $L \subseteq a_1^+ a_2^+ \dots a_k^+$, then the restriction of Ψ_{M_k} to L is an injective mapping;
(ii) if L is a bounded language over Σ, such that $L \subseteq w^*$, where $w \in \Sigma^*$, then the restriction of Ψ_{M_k} to L is an injective mapping.

Observe that, in the cases listed in Remark 5.2, also the ordinary Parikh mapping (from words to vectors) is injective. Thus, these cases do not contribute towards the main goal of characterizing the additional information provided by Parikh matrices, as opposed to Parikh vectors.

The examples given in Section 1 are over a three-letter ordered alphabet. One could perhaps hope for that the Parikh matrix mapping is injective over two letters, say $\{a < b\}$. Also this is futile, as shown by the simple example:

$$\Psi_{M_2}(a^2b^2a^2b^2) = \Psi_{M_2}(ba^4b^3) = \begin{pmatrix} 1 & 4 & 12 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}.$$

Indeed, reference [1] investigates the injectivity of Parikh matrix mappings in more detail, especially in case of two-letter alphabets. We mention a couple of examples from [1]. When restricted to the language $(a^* \cup b^*)(a^* \cup b^*)$, or to the language a^*bab^*, Parikh matrix mappings are injective. Consider the matrices

$$\begin{pmatrix} 1 & 4 & 6 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 5 & 8 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then the four words

$$baabaab,\ baaabba,\ abbaaab,\ abababa$$

are exactly the ones having the first matrix as the Parikh matrix. Similarly, the six words

$$aababbaa,\ aabbaaba,\ abaababa,\ baaaabba,\ ababaaab,\ baaabaab$$

are exactly the ones having the second matrix as the Parikh matrix.

6. Conclusion

Parikh matrices constitute a new promising way of encoding words numerically and thus transferring problems about words and languages to problems about vectors and matrices.

We hope to return to this topic in the future. Of related problem areas, we mention here problems concerning *slenderness*, a notion studied rather intensively

in recent years, see [3,4,6]. A language L is *slender* if there is an integer t such that L contains no more than t words of equal length. Similarly, L is *Parikh-slender* (resp. *Parikh-matrix-slender*) if there is an integer t such that L contains no more than t words with the same Parikh vector (resp. Parikh matrix). How can we decide, say, of a given context-free language whether or not it is Parikh-matrix-slender?

A problem area we have not discussed at all in this paper concerns *sets of Parikh matrices* and *families* of such sets, analogous to the family of *semilinear* sets of Parikh vectors. Basic questions in this area deal with interconnections between language families and matrix families. For instance, in view of Remark 2.3, we might look for a "natural" family of sets of matrices such that the set of Parikh matrices resulting from a context-free language is in this family.

References

[1] A. Atanasiu, C. Martín–Vide and A. Mateescu, *On the injectivity of the Parikh matrix mapping* (submitted).
[2] T. Harju, O. Ibarra, J. Karhumäki and A. Salomaa, Some decision problems concerning semilinearity and commutation. *J. Comput. System Sci.* (to appear).
[3] J. Honkala, On slender languages. *EATCS Bull.* **64** (1998) 145-152.
[4] J. Honkala, On Parikh slender languages and power series. *J. Comput. System Sci.* **52** (1996) 185-190.
[5] L. Ilie, An attempt to define mildly context-sensitive languages. *Publ. Math. Debrecen* **54** (1999) 865-876.
[6] L. Ilie, G. Rozenberg and A. Salomaa, A characterization of poly-slender context-free languages. *RAIRO: Theoret. Informatics Appl.* **34** (2000) 77-86.
[7] J.J. Pansiot, A decidable property of iterated morphisms. Springer, *Lecture Notes in Comput. Sci.* **104** (1981) 152-158.
[8] R.J. Parikh, On context-free languages. *J. Assoc. Comput. Mach.* **13** (1966) 570-581.
[9] G. Rozenberg and A. Salomaa, *Handbook of Formal Languages 1–3.* Springer-Verlag, Berlin, Heidelberg, New York (1997).
[10] J. Sakarovitch and I. Simon, Subwords, edited by M. Lothaire, *Combinatorics on Words.* Addison-Wesley, Reading, Mass. (1983) 105-142.
[11] A. Salomaa, *Formal Languages.* Academic Press, New York (1973).

Received February 6, 2001. Revised December 20, 2001.