Subword histories and Parikh matrices

Alexandru Mateescu

Faculty of Mathematics, University of Bucharest

Academiei, 14, Bucharest, Romania

alexmate@pcnet.ro

Arto Salomaa

Turku Centre for Computer Science

Lemminkäisenkatu 14, 20520 Turku, Finland

asalomaa@utu.fi

Sheng Yu

Department of Computer Science

University of Western Ontario

London, Ontario, Canada N6A 5B7

syu@csd.uwo.ca

Turku Centre for Computer Science
TUCS Technical Report No 442

ISBN 952-12-0951-8
ISSN 1239-1891

Abstract

Parikh matrices recently introduced give much more information about a word than just the number of occurrences of each letter. In this paper we introduce the general notion of a *subword history* and obtain a sequence of results, valid for arbitrary words. We also investigate overall methods for establishing the validity of such results. The paper can be viewed as a contribution toward *arithmetization* of questions concerning words.
Keywords word, string, subword, scattered subword, number of subwords

TUCS Research Group
Theory Group: Mathematical Structures in Computer Science

1 Introduction

This paper deals with quite general issues concerning *words* or *strings*, hereafter *words*, over a given alphabet. In particular, we are interested in the number of occurrences of a word u as a *subword* in a word w, in symbols, $|w|_u$. In this paper the term *subword* means that w, as a sequence of letters, contains u as a subsequence. This means that there exist words $x_1, \ldots, x_k$ and $y_0, \ldots, y_k$, some of them possibly empty, such that

$$u = x_1 \ldots x_k \text{ and } w = y_0 x_1 y_1 \ldots x_k y_k.$$

Subwords in this sense are often called *scattered subwords.* (For varying uses of the terminology, the reader is referred to [6].)

Assume that Σ is an alphabet containing the letters a and b (and possibly some other letters). A little reflection shows that, for any word w,

$$(|w|_a) \times (|w|_b) = |w|_{ab} + |w|_{ba}.$$

This simple equation illustrates two key issues investigated in this paper. Firstly, by moving all terms to the left side, we obtain the *polynomial* $a \times b - ab - ba$. (Here iuxtaposition refers to the catenation of words, whereas $\times$ refers to the product of two numbers.) According to the terminology introduced below, we speak of the *subword history* $a \times b - ab - ba$ in the word w, defined by the equation

$$SH(w, a \times b - ab - ba) = (|w|_a) \times (|w|_b) - |w|_{ab} - |w|_{ba}.$$

Thus, our simple equation tells us that, for any word w,

$$SH(w, a \times b - ab - ba) = 0.$$

Secondly, our equation can be written in the form

$$SH(w, a \times b) = SH(w, ab + ba).$$

In other words, independently of w, the subword history $a \times b$ assumes the same value as the subword history $ab + ba$ in w. We say that the two subword histories are *equivalent.* We will give a method of deciding the equivalence of two given subword histories, based on the elimination of the operation $\times$. We will show that this holds true in general: the operation $\times$ can be eliminated from all subword histories. This also gives rise to a (unique) *normal form* for subword histories.

A brief outline about the contents of this paper follows. Sections 2 and 3 contain the definitions, as well as some basic results, dealing with subword histories and Parikh matrices. In Section 4, we consider a general inequality about the occurrences of words of a certain type as subwords in an arbitrary word. The arguments do not use the theory of Parikh matrices, contrary to the proofs of the inequalities given in Section 5, whose proofs use a minor of the Parikh matrix. A method of eliminating the products $\times$ from subword histories is presented in Section 6. We already pointed out the consequences of this method. In Section 7, we show that all minors of a Parikh matrix are nonnegative, which again leads to inequalities between subword histories. Section 8 discusses the general problem of characterizing Parikh matrices. Finally, some open problems are mentioned in Section 9.

In this paper only basic facts concerning words are required on the part of the reader, [8] can be consulted if need arises. *Subwords* are understood in the sense described above. (We consider very little *factors*, that is, subwords where the letters appear consecutively.) The *length* of a word w is denoted by $|w|$, and the empty word by λ. Given an alphabet Σ, we usually use a, b, c, d, e (possibly with indices) to denote letters in Σ, and u, v, w, x, y, z to denote words in Σ^*.

2 Subword histories

We begin with more details about the notation $|w|_u$, the number of occurrences of the word u as a (scattered) subword of the word w. Clearly, $|w|_u = 0$ if $|w| < |u|$. We also make the *convention* that, for any w and the empty word λ,

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

In [7] the number $|w|_u$ is denoted as a "binomial coefficient"

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

Indeed, 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 $\binom{i}{0} = 1$.

It is shown in [7] that the equation

$$|vb|_{ua} = |v|_{ua} + \delta_{a,b}|v|_u,\ a, b \in \Sigma;\ u, v \in \Sigma^*,$$

where

$$\delta_{a,b} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{if } a \neq b, \end{cases}$$

together with the already mentioned equations

$$|w|_\lambda = 1,\ |w|_u = 0 \text{ for } |w| < |u|$$

suffice to compute all values $|w|_u$. For instance, for $w = ababa$,

$$|w|_{ab} = 3 = |w|_{ba}.$$

If one knows the values $|w|_{u_i}$ for *several* subwords u_i, one has more information about the word w itself. For instance, the equations

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

imply, assuming $w \in \{a, b\}^*$, that

$$w = a^2bab^2.$$

In the sequel we assume that one knows a certain combination of the individual values $|w|_{u_i}$, obtained from the latter by additions, subtractions and multiplications. This means that one begins with a *polynomial* $p(u_1, \ldots, u_k)$ (with integer coefficients) in terms of the words $u_1, \ldots, u_k$. We denote the multiplication by $\times$, to distinguish the product $u_1 \times u_2$ from the catenation $u_1 u_2$. The combination of the individual words is referred to as a *subword history.*

Definition 1 *Consider an alphabet Σ and a word $w \in \Sigma^*$. A* subword history *in Σ and its* value *in w are defined recursively as follows.*

- *Every $u \in \Sigma^*$ is a subword history in Σ, referred to as* monomial, *and its value in w equals $|w|_u$.*
- *Assume that SH_1 and SH_2 are subword histories with values α_1 and α_2, respectively. Then*

  $$-(SH_1),\ (SH_1) + (SH_2),\ \text{and } (SH_1) \times (SH_2)$$

  are subword histories with values

  $$-\alpha_1, \alpha_1 + \alpha_2,\ \text{and } \alpha_1 \alpha_2,$$

  respectively.

Two subword histories are termed equivalent *if they assume the same value in any w. A subword history is* linear *if it is obtained without using the operation $\times$.*

In what follows we are interested only in the values of subword histories and will omit, therefore, unnecessary parentheses. We also use the notation $SH(w,p)$ for the value of the subword history p in w. Clearly, if associative, commutative or distributive laws are applied in polynomials p, the result will be an equivalent subword history. Equivalence will be denoted by the equality sign.

We have already pointed out that the subword histories $a \times b$ and $ab + ba$ are equivalent. This fact can also be expressed by saying that the subword history

$$a \times b - ab - ba$$

assumes the value 0 in any w. The next example illustrates the difference between $\times$ and catenation:

$$SH(a^2bab^2, ab \times b - a - b) = 18,\ SH(a^2bab^2, abb - a - b) = 1.$$

The equivalence between the subword histories $a \times b$ and $ab + ba$ can be extended to concern any number of letters. For instance,

$$a \times b \times c \text{ and } abc + acb + bac + bca + cab + cba$$

are equivalent. We will show in Section 6 that any subword history is equivalent to a *linear* subword history. We now present a preliminary result in this direction. We consider the *shuffle* $u \sqcup\!\sqcup\, v$ of two words u and v. By definition, $u \sqcup\!\sqcup\, v$ consists of all words

$$u_0 v_0 u_1 v_1 \ldots u_k v_k,\ \text{where } k \geq 0,\ u_i, v_i \in \Sigma^* \text{ for } 0 \leq i \leq k,\ \text{and}$$

$$u = u_0 \ldots u_k,\ v = v_0 \ldots v_k.$$

Theorem 1 *Assume that u and v are words over disjoint alphabets. Then the subword histories $u \times v$ and $\sum_{x \in u ⧢ v} x$ are equivalent.*

Proof. Consider an arbitrary word w. If the alphabet of w does not contain the alphabets of u and v, both of the subword histories of the theorem assume the value 0 in w. Thus, assume that the alphabet of w contains the alphabets of u and v. Consider a word $x \in u ⧢ v$ such that x has occurrences in w. Since the alphabets of u and v are disjoint, we can tell which letters in x come from u and which from v. Consequently, the set of occurrences of x in w determines a set of pairs of occurrences of u and v in w. This holds true for every $x \in u ⧢ v$. (If x has no occurrences in w, then the corresponding set of pairs of occurrences is empty.) Moreover, for two different words x and y in $u ⧢ v$, the corresponding sets of pairs are disjoint, and every pair of occurrences of u and v in w can be associated to a definite $x \in u ⧢ v$. This shows that

$$SH(w, u \times v) = SH(w, \sum_{x \in u ⧢ v} x),$$

which was the claim of the theorem.

If the alphabets of u and v are not disjoint, we obtain only the inequality

$$SH(w, \sum_{x \in u ⧢ v} x) \leq SH(w, u \times v),$$

where the equality is a rare exception. Consider our previous example $w = ababa$, and choose $u = ab$ and $v = ba$. Then $SH(w, u \times v) = 9$, whereas

$$u ⧢ v = \{abba,\ abab,\ baba,\ baab\}$$

and, consequently,

$$SH(w, \sum_{x \in u ⧢ v} x) = 3.$$

Two observations are needed to get a *linear* subword history equivalent to $u \times v$. Firstly, the words $abba$ and $baab$ can both be obtained in two different ways in $u ⧢ v$ and, consequently, should be counted twice. Secondly, also the "reduced" words aba and bab occurring in w give rise to pairs of occurrences of u and v in w. By these observations, it can be concluded that

$$\sigma = 2abba + abab + baba + 2baab + aba + bab$$

is a linear subword history equivalent to $ab \times ba$. (The term $2abba$ is, of course, an abbreviation of $abba + abba$.) Indeed, by the equations

$$|w|_{aba} = 4,\ |w|_{bab} = 1,$$

we conclude that $SH(w, \sigma) = 9$. The equation

$$SH(w, ab \times ba) = SH(w, \sigma)$$

holds for all w. For $w = abababa$, we have

$$SH(w, ab \times ba) = 36.$$

The individual six terms of σ (in the order listed above) occur in w separately $5, 5, 5, 1, 10, 4$ times, which gives the sum $SH(w, \sigma) = 36$, when one remembers to count the first and fourth numbers twice.

3 Parikh matrices

Parikh mappings (vectors) introduced in [5] express properties of words as *numerical* properties of vectors yielding some fundamental language-theoretic consequences, [8]. However, much information is lost in the transition from a word to a vector. A sharpening of the Parikh mapping, where more information is preserved than in the original Parikh mapping, was introduced in [3]. The new mapping uses upper triangular square matrices, with nonnegative integer entries, 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 general different Parikh matrices. Thus, the Parikh matrix gives more information about a word than a Parikh vector. The set of all triangular matrices described above 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 introduce the notion of a Parikh matrix mapping. (Before [3], the term *Parikh matrix* was used in [4] in an entirely different meaning, for the growth matrix of a morphism or a D0L system).

Definition 2 *Let $\Sigma = \{a_1, \ldots, a_k\}$ be an alphabet. 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.

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. Knowledge of the Parikh matrices for different orderings of the alphabet will increase our knowledge of the word in question.

If we consider letters without numerical indices, we assume the alphabetic ordering in the definition of Parikh matrices. The Parikh matrix mapping is not injective even for the alphabet $\{a, b\}$. For instance, 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 five words

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

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. This example becomes clearer in view of the following theorem, [3], where the entries of the Parikh matrix are characterized. For the alphabet $\Sigma = \{a_1, \ldots, a_k\}$, 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 2 *Consider* $\Sigma = \{a_1, \ldots, a_k\}$ *and* $w \in \Sigma^*$. *The matrix* $\Psi_{M_k}(w) = (m_{i,j})_{1 \le i,j \le (k+1)}$, *has the following properties:*

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

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. Theorem 2 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.

We still mention one result from [3], which establishes an interconnection between the inverse of a Parikh matrix and the Parikh matrix of $mi(w)$, the *mirror image* of w.

Definition 3 *Let* Σ *and* $w \in \Sigma^*$ *be as in Theorem 2. Assume that the Parikh matrix of* w *is* $\Psi_{M_k}(w) = (m_{i,j})_{1 \le i,j \le k+1}$. *The* alternate Parikh matrix of w, *denoted* $\overline{\Psi}_{M_k}(w)$, *is the matrix* $(m'_{i,j})_{1 \le i,j \le k+1}$, *where* $m'_{i,j} = (-1)^{i+j} m_{i,j}$, *for all* $1 \le i,j \le k+1$.

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 *Let* Σ *and* $w \in \Sigma^*$ *be as in Theorem 2. Then*

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

Theorem 3 has the following corollary. The corollary can be obtained also from the general considerations in Section 7. We consider minors of order k of a Parikh matrix, that is, determinants obtained from the matrix by omitting one row and one column, and leaving the order of the other rows and columns unchanged.

Corollary 1 *Let* A *be the Parikh matrix associated to a word* w *over an alphabet with* k *letters. Then every minor of order* k *of* A *assumes a nonnegative integer value.*

Proof. Consider the standard method of computing the inverse of A, using cofactors and determinants. The determinants involved, apart from the determinant of A which assumes the value 1, are the ones mentioned in our theorem. If one of them would assume a negative value, then the corresponding entry in the Parikh matrix of $mi(w)$ would be negative, by Theorem 3. However, this is impossible.

4 An inequality

We consider in this section a general inequality concerning $|w|_u$. The inequality bears some resemblance to the classical *Cauchy inequality* in algebra and analysis. Notice that the different occurrences of u in w can be naturally *ordered* from left to right. For two occurrences u_1 and u_2, we consider the first letter that is positioned differently in w. According to its position, we say which one of u_1 and u_2 is to the left of the other. The occurrences can also be presented in terms of vectors. For $|u| = m$, each occurrence is an m-dimensional vector with increasing integer components, where each component indicates the position of the corresponding letter of u in w. For $w = aababb$, the occurrences of $u = abb$ are from left to right:

$$(1,3,5),\ (1,3,6),\ (1,5,6),\ (2,3,5),\ (2,3,6),\ (2,5,6),\ (4,5,6).$$

The occurrences can be ordered, similarly, from right to left. But the order is not necessarily (and also not in our example) the reverse of the order from left to right, although the first and last elements are the same.

We now consider a general inequality concerning the occurrences of subwords of a certain general type. We consider decompostions $u = xyz$, and the occurrences of u, y, xy, and yz in an arbitrary w.

Conjecture 1 *The inequality $|w|_{xyz}|w|_y \leq |w|_{xy}|w|_{yz}$ holds for arbitrary words w, x, y, z.*

Lemma 1 *Conjecture 1 holds if y is a letter.*

Proof. We use the notations

$$|w|_y = n,\ |w|_{xyz} = C,\ |w|_{xy} = A,\ |w|_{yz} = B.$$

(The reason for the notational difference for $|w|_y$ will become apparent below.) Thus, the claim is

$$nC \leq AB.$$

Order the n occurrences of y in w from left to right. Consider the first (leftmost) occurrence, and let α_1 be the number of occurrences of x (strictly) to the left of it, in w. For $i = 2, \ldots, n$, define the numbers α_i in the same way, with respect to the ith occurrence from the left of y in w. It follows that

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$$

and

$$A = \sum_{i=1}^{n} \alpha_i.$$

Similarly, let β_1 be the number of occurrences of z to the *right* of the first occurrence of y in w. Again, for $i = 2, \ldots, n$, define the numbers β_i in the same way, with respect to the ith occurrence of y in w. We have again

$$B = \sum_{i=1}^{n} \beta_i$$

but now the inequalities are in the reverse order:

$$\beta_1 \geq \beta_2 \geq \ldots \geq \beta_n.$$

(In fact, this is the only spot, where our argument fails for the general Conjecture 1; if y is a word, the reverse inequalities do not necessarily hold.) Observe now that

$$C = \sum_{i=1}^{n} \alpha_i \beta_i.$$

Denote further

$$D = \sum_{i,j=1,\ i<j}^{n} (\alpha_j - \alpha_i)(\beta_i - \beta_j).$$

Observe that, by our inequalities concerning α_i and β_i, we have $D \geq 0$. Our claim now follows by the algebraic identity

$$\sum_{i=1}^{n} \alpha_i \sum_{i=1}^{n} \beta_i = \sum_{i=1}^{n} \alpha_i \beta_i + \sum_{i,j=1,\ i<j}^{n} (\alpha_j - \alpha_i)(\beta_i - \beta_j).$$

The proof of the lemma shows that the validity of Conjecture 1 reduces to the question of the sum D being nonnegative. One can give examples, where D has negative terms. But in all examples known to us the positive terms in D cancel the negative ones. Examples of this kind are obtained by choosing

$$y = a^2ba^2,\ x = z = a^2,\ w = a^iba^j,\ \text{where } i, j > 4.$$

In these examples the inequality is an equality.

An easy calculation with binomial coefficients shows that Conjecture 1 is valid for one-letter alphabets. It holds also if x, y, z result from a 2×2 determinant in a Parikh matrix.

5 Canonical subword histories

We now prove that a certain subword history SH_Σ, referred to as *canonical*, having positive and negative terms in a symmetric fashion, assumes a nonnegative value in any w. The canonical subword history depends on the alphabet. As always in connection with Parikh matrices, we have a specific ordering of the alphabet in mind. However, we can fix the ordering arbitrarily. When dealing with letters from the beginning of the English alphabet, we tacitly assume the ordering to be the normal one.

We can also split the canonical subword history into two parts, referred to as *primary* and *secondary*, where the latter consists of the negative terms in the canonical history. Then the value assumed by the primary part is greater than or equal to the value assumed by the secondary part, in any w.

Definition 4 *Consider the alphabet $\Sigma = \{a_1, a_2, \ldots, a_k\}$ and the word $u_\Sigma = a_1a_2 \ldots a_k$. We now form from u_Σ (monomial) subword histories by inserting an arbitrary number i, $0 \leq i \leq k-1$, of product signs $\times$. If $i = (k-1)-j$, where j is even (resp. odd), the resulting monomial subword history is referred to as* primary *(resp.* secondary*). The* primary *(resp.* secondary*)* canonical subword history SH^1_Σ *(resp.* SH^2_Σ*) of* Σ *equals the sum of all monomial primary (resp. secondary) subword histories. Finally, the* canonical subword history of Σ *is defined by*

$$SH_\Sigma = SH^1_\Sigma - SH^2_\Sigma.$$

As an illustration, we list the canonical subword histories of certain small alphabets. The primary (resp. secondary) parts are obtained by considering only the positive (resp. negative) terms.

$$SH_{\{a,b\}} = a \times b - ab,$$

$$SH_{\{a,b,c\}} = a \times b \times c + abc - a \times (bc) - (ab) \times c,$$

$$SH_{\{a,b,c,d\}} = a \times b \times c \times d + a \times (bcd) + (ab) \times (cd) + (abc) \times d$$

$$-a \times b \times (cd) - a \times (bc) \times d - (ab) \times c \times d - abcd.$$

Consider now the Parikh matrix $M(w)$ of a word w over the alphabet $\{a, b, c, d\}$. Omit the first column and last row from $M(w)$. The determinant of the resulting 4×4 matrix is

$$\begin{vmatrix} a & ab & abc & abcd \\ 1 & b & bc & bcd \\ 0 & 1 & c & cd \\ 0 & 0 & 1 & d \end{vmatrix}$$

Here in each entry we have written simply v, instead of $|w|_v$. Thus, the entry in position (2,4), denoted x_{24}, equals $|w|_{bcd}$. The value of the determinant equals the sum of products

$$(-1)^{t(p)} x_{1p(1)} x_{2p(2)} x_{3p(3)} x_{4p(4)},$$

where p is a permutation of the numbers $1, 2, 3, 4$, and $t(p)$ equals 2 or 1, according to p being even or odd. Note that, because of the 0's in the lower left-hand corner, the possibilities of obtaining a nonzero term are limited. If we fix the element x_{14}, the only possibility is to take the three 1's, the resulting permutation (1432) being odd. Thus, the term in the determinant is $-abcd$. If we fix the element x_{12}, we may choose either the first 1 and c and d, or else the first and the last 1's and cd. The former alternative leads to the odd permutation (12), the latter to the even permutation (12)(34). The terms thus obtained correspond to the part

$$-abcd - (ab) \times c \times d + (ab) \times (cd)$$

in the canonical subword history.

In this way it can be immediately verified that the value of the our determinant equals the value of the canonical subword history $SH_{\{a,b,c,d\}}$ in w. Moreover, instead of the four-letter alphabet, the argument generalizes immediately to arbitrary alphabets. Hence, by Corollary 1, we obtain the following result.

Theorem 4 *The value of the canonical subword history SH_Σ is nonnegative in every word w. The value of the primary canonical history in w is never less than the value of the secondary canonical history in w.*

Using the notion $|w|_u$, we still state Theorem 4 in case of the alphabet $\{a, b, c, d\}$. For any word w, we have

$$|w|_{abcd} + |w|_a|w|_b|w|_{cd} + |w|_a|w|_{bc}|w|_d + |w|_{ab}|w|_c|w|_d$$

$$\leq |w|_a|w|_{bcd} + |w|_{ab}|w|_{cd} + |w|_{abc}|w|_d + |w|_a|w|_b|w|_c|w|_d.$$

6 Reduction to linear subword histories

In this section we will give a method of eliminating products from subword histories. In other words, we show how to construct a *linear* subword history equivalent to a given subword history. This will give us a method of *deciding the equivalence* of two given subword histories, as well as *normal form* for subword histories.

The rudiments of the reduction argument were already given in Section 2, in connection with Theorem 1. We now go into formal details. For Σ, we consider the primed version Σ', $\Sigma' = \{a' \mid a \in \Sigma\}$. Let $g : \Sigma^* \longrightarrow \Sigma'^*$ be the morphism defined by $g(a) = a'$. Also, let $h : (\Sigma \cup \Sigma')^* \longrightarrow \Sigma^*$ be the morphism defined by $h(a) = h(a') = a$.

For two words u and v, we consider their *ground shuffle*:

$$G(u, v) = u \sqcup\!\sqcup\, g(v).$$

We also define a set of rewriting rules:

$$P = \{aa' \longrightarrow a \mid a \in \Sigma\}.$$

Using these rules we now get a reduction relation.

Definition 5 *Consider $x, y \in (\Sigma \cup \Sigma')^*$. The relation of m-reduction, denoted $\vdash_m$, for $m \geq 0$, is defined as follows: $x \vdash_m y$ holds iff y can be obtained from x by applying* in parallel *m rewriting rules from P. In other words, x has a decomposition $x = x_0 a_1 a_1' x_1 \ldots x_{m-1} a_m a_m' x_m$ such that $y = x_0 a_1 x_1 \ldots x_{m-1} a_m x_m$, where $a_i \in \Sigma$ for $1 \leq i \leq m$ and $x_i \in (\Sigma \cup \Sigma')^*$ for $0 \leq i \leq m$. (If $m = 0$, then $x = y$.)*

*A word $r \in \Sigma^*$ is called an m-*reduction *of u, v iff there are a word $x \in G(u, v)$ and a word $y \in (\Sigma \cup \Sigma')^*$ such that $x \vdash_m y$ and, moreover, $r = h(y)$. The* multiplicity *of r, denoted $t(r)$, is defined as:*

$$t(r) = card\{x \in G(u, v) \mid x \vdash_m y \text{ and } h(y) = r, \text{ where } y \in (\Sigma \cup \Sigma')^*\}.$$

Finally, the reduced set *of u, v, denoted $R(u, v)$, is:*

$$R(u, v) = \{r \mid r \text{ is an } m - \text{reduction of } u, v \text{ for some } m \geq 0\}.$$

As an illustration, assume that $u = ab$ and $v = ab$. Then

$$G(u, v) = \{aba'b', aa'bb', aa'b'b, a'abb', a'ab'b, a'b'ab\}.$$

The 0-reductions are the words $abab$ with the multiplicity 2 and $aabb$ with the multiplicity 4. The $\vdash_1$ reduction relation is:

$$aa'bb' \vdash_1 aa'b,\ aa'bb' \vdash_1 abb',\ aa'b'b \vdash_1 ab'b,\ a'abb' \vdash_1 a'ab$$

.

Thus, we obtain the 1-reductions aab with the multiplicity 2 and abb with the multiplicity 2. The $\vdash_2$ reduction relation is $aa'bb' \vdash_1 ab$. Thus, the only 2-reduction is ab with the multiplicity 1. For $m > 2$ there are no m-reductions.

In conclusion, the reduced set of u, v is

$$R(u, v) = \{abab, aabb, aab, abb, ab\}.$$

The elimination of products is based on the following theorem.

Theorem 5 *For arbitrary words u and v, the subword histories satisfy*

$$u \times v = \sum_{r \in R(u,v)} t(r)r.$$

Proof. The proof runs along the lines of the proof of Theorem 1. Consider an arbitrary word w. We have to compare the values of the subword histories $u \times v$ and $\sum_{r \in R(u,v)} t(r)r$ in w. We replace the sum with multiplicities with a sum, where in connection with each r the element of $G(u, v)$ giving rise to r is marked. The occurrences in w of the elements x belonging to this new marked sum determine a set of pairs of occurrences of u and v in w. The sets of pairs determined in this way by two different elements of the marked sum are disjoint. Moreover, every pair of occurrences of u and v in w is determined by some x in the marked sum. This 1-to-1 correspondence shows that the claimed equation holds.

Consider our previous example. We obtain the equation

$$(ab) \times (ab) = 2abab + 4aabb + 2aab + 2abb + ab$$

between subword histories. Assume that $w = a^2b^2$. We obtain $|w|_{ab}|w|_{ab} = 16$. Using Theorem 5, we obtain

$$|w|_{ab}|w|_{ab} = 2|w|_{abab} + 4|w|_{aabb} + 2|w|_{aab} + 2|w|_{abb} + |w|_{ab} = 16.$$

Assume, secondly, that $w = a^2bab^2$. We have now $|w|_{ab}|w|_{ab} = 64$. On the other hand,

$$|w|_{ab}|w|_{ab} = 2|w|_{abab}+4|w|_{aabb}+2|w|_{aab}+2|w|_{abb}+|w|_{ab} = 8+20+14+14+8 = 64.$$

By successive applications of Theorem 5 (and by the distributive law), we now obtain the following result.

Corollary 2 *Every subword history is equivalent to a linear subword history. Moreover, given a polynomial $p(u_1, \ldots, u_n))$, a linear polynomial representing an equivalent subword history can be effectively constructed.*

We have also the following straightforward result.

Lemma 2 *Two linear subword histories are equivalent iff they are identical, apart from the order of terms.*

Proof. Assume that SH_1 and SH_2 are equivalent linear subword histories that are not identical. If the same term is added or removed from both of them, again non-identical equivalent linear subword histories result. Therefore, we may assume that a word u is in SH_1 but not in SH_2 and, moreover, SH_1 does not contain a word shorter than u. If SH_2 contains a word v shorter than u, the two subword histories assume a different value in $w = v$. Otherwise, the two subword histories assume a different value in $w = u$. This is a contradiction.

Corollary 3 *The equivalence of two given subword histories is effectively decidable.*

In view of these results, we can view linear subword histories as *normal forms* of arbitrary subword histories. By agreeing on the order of terms, we obtain a *unique* normal form.

Apart from equivalence (that is, equality between subword histories), one may investigate *inequalities* between subword histories using Theorem 5. It is a consequence of Lemma 1 that

$$b \times (abc) \leq (ab) \times (bc).$$

We now investigate this inequality more closely. We obtain first subword histories equivalent to the left and right sides:

$$b \times (abc) = babc + 2abbc + abcb + abc,$$

$$(ab) \times (bc) = babc + 2abbc + abcb + bacb + bcab + abc.$$

(We have written first the 0-reductions, then the 1-reductions.) Comparing the two, we obtain the following result.

Lemma 3 *For any word w, we have*

$$|w|_b |w|_{abc} \leq |w|_{ab} |w|_{bc}.$$

Moreover, the equality holds exactly in case w contains neither of the subwords $bacb$ and $bcab$.

Finally, we investigate the inequality

$$A = (ab) \times (aabb) \leq (aab) \times (abb) = B.$$

It falls within Conjecture 1 but does not follow by Lemma 1. The equivalent linear subword histories are now more complicated. We obtain

$$A = aba^2b^2 + 4a^2bab^2 + 9a^3b^3 + a^2b^2ab + abab^2 + a^2bab + 6a^2b^3 + 6a^3b^2 + 4a^2b^2.$$

Moreover, the linear subword history equivalent to B is obtained by adding

$$a^2bab^2 + a^2b^2ab + aba^2b^2 + ababab + ab^2a^2b + 2abab^2 + a^2bab + ab^2ab + abab$$

to A. This gives the following conclusion.

Lemma 4 *For any word w, we have*

$$|w|_{ab} |w|_{aabb} \leq |w|_{aab} |w|_{abb}.$$

The equality holds exactly in case w does not contain the subword $abab$.

7 Minors of a Parikh matrix

In this section we generalize Corollary 1 to concern all minors of a Parikh matrix. For each minor, the result can be interpreted as an inequality for subword histories. We will mention only a couple of examples of this, rather than enter a systematic study of the inequalities.

Theorem 6 *The value of each minor of an arbitrary Parikh matrix is a nonnegative integer.*

Proof. Consider the Parikh matrix associated to a word w. The proof is by induction on the length of w. If $w = \lambda$, the assertion holds because each minor of the identity matrix assumes the value 0 or the value 1. Assume that the assertion holds for the Parikh matrix A associated to a word w, and consider the Parikh matrix A' associated to the word wa_i, $a_i \in \Sigma$. Then the matrix A' coincides with A, except that the $(i+1)$st column of A' is obtained by adding the ith column of A to the $(i+1)$st column of A.

Let D be a minor of A'. If the $(i+1)$st column of A' is not involved in D, then D is a minor of A, and we can use the inductive hypothesis. If the $(i+1)$st column but not the ith column of A' is involved in D, then we can write D in the form $D = D_1 + D_2$, where both D_1 and D_2 are minors of A. Again the inductive hypothesis is applicable. It is also applicable in the last case, where both the ith and the $(i+1)$st column of A' are involved in D. Now D can be written in the form $D = D_1 + D_2$, where D_1 is a minor of A and D_2 has two identical columns. Thus, $D_2 = 0$, which completes the proof.

The next two corollaries are examples of applications of Theorem 6 to subword histories.

Corollary 4 *Consider* $\Sigma = \{a_1, \ldots, a_k\}$, *and assume that* xyz *is a factor of the word* $u_\Sigma = a_1 \ldots a_k$. *Then, for any word* w,

$$|w|_y |w|_{xyz} \leq |w|_{xy} |w|_{yz}.$$

Proof. Consider the appropriate minor of order 2 of the Parikh matrix.

Corollary 5 *The subword history*

$$SH = (abcde)\times(c\times d - cd) + (abcd)\times(cde - c\times(de)) + (abc)\times((cd)\times(de) - d\times(cde))$$

in the alphabet $\{a, b, c, d, e\}$ *assumes always a nonnegative value.*

Proof. Consider the minor of order 3 of the Parikh matrix for a word w, obtained by omitting the first 3 columns and 2nd, 5th and 6th rows. The value of SH in w equals the value of the determinant

$$\begin{vmatrix} abc & abcd & abcde \\ c & cd & cde \\ 1 & d & de \end{vmatrix}.$$

Observe that the middle term in SH may assume a negative value, since we may have $|w|_{cde} < |w|_c |w|_{de}$.

8 On the characterization of a Parikh matrix

Two quite general questions about Parikh matrices are:

- Which matrices are Parikh matrices?
- How much ambiguity does a given Parikh matrix leave about a word? When does a given Parikh matrix determine a unique word?

We present in this section some observations about the first question. Clearly, arbitrary nonnegative integers can appear in the second diagonal of a Parikh matrix. Above the second diagonal the choices for the entries are restricted by the already existing entries. In particular, they have to satisfy Theorem 6. As regards the entries in the *third* diagonal, there are no further restrictions.

Lemma 5 *Consider* $\Sigma = \{a_1, \ldots, a_k\}$. *For* $1 \leq i \leq k-1$, *any integer* x *satisfying*

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

can appear as the entry $m_{i,i+2}$ *in a Parikh matrix.*

Proof. The entry $m_{i,i+2}$ depends only on the mutual position of the occurrences of a_i and a_{i+1}. If every a_i precedes every a_{i+1} (resp. every a_{i+1} precedes every a_i), we get the upper bound $m_{i,i+1}m_{i+1,i+2}$ (resp. the lower bound 0). It is easy to see that all intermediate numbers (some of them corresponding to several words, as was seen in the examples in Section 3) can be obtained by interchanging some of the occurrences. Since i was arbitrary, the lemma follows.

The preceding lemma gives an exhaustive characterization of Parikh matrices of words over a two-letter alphabet.

Proceed inductively and assume that you have a characterization in case of k-letter alphabets, that is, matrices of dimension $k+1$. Given a matrix A of dimension $k+2$, we want to decide whether or not A is the Parikh matrix for a word over the alphabet $\{a_1, a_2, \ldots, a_{k+1}\}$. We first delete from A the last row and column, and check whether the resulting matrix gives a word over $\{a_1, \ldots, a_k\}$. If the answer is negative, then A is not a Parikh matrix. Otherwise, we delete from A the first row and column and check whether the resulting matrix gives a word over $\{a_2, \ldots, a_{k+1}\}$. If it doesn't, A is not a Parikh matrix. Otherwise, we have to check whether or not the entry in the *upper right-hand corner* of A is compatible with the other entries. For this we can offer no systematic method. The following case study illustrates the difficulties arising in particular examples.

Consider the matrix

$$A_x = \begin{pmatrix} 1 & 3 & 7 & x \\ 0 & 1 & 5 & 6 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

What are the possibilities for the entry x? Considering the minor of order 2 in the upper right-hand corner, we obtain the condition $42 - 5x \geq 0$ and, consequently, $0 \le x \le 8$. Indeed, all of these values are possible and are obtained, in an increasing order, by the following words:

$$bbcccabaabb,\ bcbcabcaabb,\ cbbabccaabb,\ babcccbabab,\ cbbcabaabcb,$$

$$cbabcbabcab,\ cbabcabbcba,\ cbacbabbbac,\ acbcbbabbac.$$

Next, make a slight change in the entry $(2,4)$ of A_x, giving rise to the matrix

$$A'_x = \begin{pmatrix} 1 & 3 & 7 & x \\ 0 & 1 & 5 & 7 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Now the condition for x is $0 \le x \le 9$. But the values 0 and 3 are not possible for x, there are no words having the corresponding matrix! This is true although all minors of the matrix with $x = 0$ or $x = 3$ are nonnegative. The remaining values of x are obtained, in an increasing order, by the following words:

$$bbccabcaabb,\ bcbabccaabb,\ bcabbccabab,\ cbbaabcbcab,$$

$$cbbabcaabcb,\ cbbcaabbabc,\ cbabcabbbca,\ cabbcbabbac.$$

Thus, the consideration of minors is not sufficient for a characterization of Parikh matrices.

9 Conclusion. Open problems

We believe that Parikh matrices constitute a powerful tool for translating problems concerning words into numerical problems. Other such tools have been presented recently, for instance in [2]. Although a Parikh matrix does not characterize a word exhaustively, the characterization given by it is sufficient for many purposes.

As we have notices in Section 4 and especially in Section 8, some important problems remain still open. We conclude by mentioning another open problem. It concerns the decidability of *inequalities between subword histories*. By Theorem 5, it suffices to consider the problem between *linear* subword histories. Contrary to Lemma 2, the situation is rather complicated for inequalities. Even some undecidability results (unexpected such as the results in [1]) are possible. The following simple result holds.

Lemma 6 *If the inequality $u \le v$ holds between two monomial subword histories, then v is a subword of u.*

However, the converse of the lemma does not hold. For instance, it is not the case that $|w|_{aa} \le |w|_a$. The inequality can depend on the appearance of several terms on the right side. For instance,

$$|w|_{baab} \le |w|_{bab} + |w|_{baaab}$$

holds for any w, but this is not any more true if one of the terms on the right side is omitted.

References

[1] Harju, T., Ibarra, O., Karhumäki, J. and Salomaa, A., Some decision problems concerning semilinearity and commutation. To appear in *J. Comput. System Sci.* (2002).

[2] Honkala, J., On Parikh slender languages and power series. *J. Comput. System Sci.* 52 (1996), 185–190.

[3] Mateescu, A., Salomaa, A., Salomaa, K. and Yu, S., A sharpening of the Parikh mapping. To appear in Theoretical Informatics and Applications.

[4] Pansiot, J.J., A decidable property of iterated morphisms. *Springer Lecture Notes in Computer Science* 104 (1981), 152–158.

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

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

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

[8] Salomaa, A., *Formal Languages.* Academic Press, New York (1973).

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi

University of Turku
- **Department of Information Technology**
- **Department of Mathematical Sciences**

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

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**