Subword histories and associated matrices

Arto Salomaa *

Turku Centre for Computer Science, Joukahaisenkatu 3–5 B, 20520 Turku, Finland

A R T I C L E   I N F O

Article history:
Received 29 January 2008
Received in revised form 14 April 2008
Accepted 30 May 2008
Communicated by M. Ito

Keywords:
Subword history
Subword
Scattered subword
Parikh matrix

A B S T R A C T

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 $w$. Arithmetical combinations of such quantities yield a so-called subword history. We investigate the information content of subword histories. Reducing subword histories to linear ones, as well as the recently introduced Parikh matrices, will be important tools. Simple polynomial formulas for computing the value of a subword history for arbitrary powers of a word are obtained.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

This paper studies methods of computing $|w|_u$, the number of occurrences of a word $u$ as a (scattered) subword of a word $w$. The computations are extended to concern so-called subword histories, arithmetical combinations of numbers $|w|_u$. It is important in many problems concerning words, languages and automata to get rid of the mathematically awkward noncommutativity, at least to some extent. This is seen, for instance, in many cases in [7]. In arithmetizing the theory one reduces noncommutative properties to commutative numerical ones. This makes the constructions easier, as seen in many instances in the theory of formal power series, [7,12]. The study of subword histories belongs to this line of research: words can be characterized by subword history values. This offers also an alternative way of defining languages, [15,19,2,17].

A brief outline of the contents of the paper follows. We first introduce the basic concepts of a Parikh matrix and a subword history, and discuss some results and examples needed later on in the paper. In Section 3 we also present a technique of making a given subword history linear. Section 4 discusses, from various points of view, the interconnection between Parikh matrices and subword histories. Thereby a specific number $\mu(F)$, associated to a finite language $F$, plays a central role. By definition, $\mu(F)$ equals the length of the shortest word where each word in $F$ appears as a factor.

Section 5 presents a method for computing, for a word $w$ and subword history $SH$, the value of $SH$ for an arbitrary word $w^n$ in $w^*$. It turns out that the value is always a polynomial in $n$, with rational coefficients. The polynomial remains unchanged in the process of making the subword history linear.

We assume that the reader is familiar with the basics of formal languages. Whenever necessary, [12] 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 $\Sigma$. Words are usually denoted by small letters from the end of the English alphabet.

2. Subwords and Parikh matrices

Throughout this paper, we consider the number of occurrences of a word $u$ as a subword in a word $w$, in symbols, $|w|_u$. Here the term subword means that $w$, as a sequence of letters, contains $u$ as a subsequence. More formally, we have 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

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

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 $w$.

We note that, in classical language theory, [12], 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 $\lambda$,

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

We would like to point out that in [4] 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,

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

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

A general problem, [12], arising in this context, and important in many applications, is: How can one construct a set of numbers $|w|_u$, or some arithmetical combination of such numbers, such that the word $w$ is uniquely, or "almost uniquely", determined? For instance, the reader should have no difficulties in proving that any word $w \in \{a, b, c\}^*$ is, for each $n \geq 1$, $n \neq 2$, uniquely determined by the values

$$|w|_a = |w|_b = |w|_c = n, \quad |w|_{ab} = |w|_{ac} = n^2 - 1.$$  

Indeed, for $n = 1$, we have $w = cba$ and, for $n \geq 3$, $w = a^{n-1}baba^2cabca^{n-1}$. For $n = 2$, both of the words $abcab$ and $ababc$ satisfies the conditions. On the other hand, a word $w \in \{a, b, c\}^*$ of length 4 is not uniquely determined by the values $|w|_{ab}$, $|u| \leq 2$, the words $abba$ and $baab$ constituting a counterexample.

For handling such problems a specific tool, referred to as the Parikh matrix was introduced in [9], and investigated further in [1,3,6,10,11,13,14,16,21]. The formal definition given below uses the extended notion due originally to [20].

The Parikh matrix is a powerful generalization of a Parikh mapping (vector). 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 $\mathcal{M}_k$.

A Parikh matrix associated to a word $w$, as originally defined in [9], tells us the values $|w|_x$, where $x$ is an arbitrary factor of the ordered product $a_1 \ldots a_k$ of all letters of the alphabet. When considering generalized Parikh matrices, arbitrary values $|w|_x$ can be obtained as entries. The dimension of the matrix depends on the values $|w|_x$ wanted as entries.

In the formal definition, we use the "Kronecker delta". For letters $a$ and $b$,

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

Definition 2. Let $u = b_1 \ldots b_k$ be a word, where each $b_i$, $1 \leq i \leq k$, is a letter of the alphabet $\Sigma$. The Parikh matrix mapping with respect to $u$, denoted $\Psi_u$, is the morphism:

$$\Psi_u : \Sigma^* \rightarrow \mathcal{M}_{k+1},$$

defined, for $a \in \Sigma$, by the condition: if $\Psi_u(a) = (m_{i,j})_{1 \leq i,j \leq (k+1)}$, then for each $1 \leq i \leq (k+1)$, $m_{i,i} = 1$, and for each $1 \leq i \leq k$, $m_{i,i+1} = \delta_{a,b_i}$, all other elements of the matrix $\Psi_u(a)$ being 0. Matrices of the form $\Psi_u(w)$, $w \in \Sigma^*$, are referred to as generalized Parikh matrices.

Thus, the matrix $\Psi_u(a)$ associated to a letter $a$ has 1’s everywhere in the main diagonal and in those entries of the second diagonal that correspond to occurrences of $a$ in $u$, and 0’s elsewhere. The matrix $\Psi_u(w)$ associated to a word $w$ is obtained by multiplying the matrices $\Psi_u(a)$ associated to the letters $a$ of $w$, in the order in which the letters appear in $w$. The above definition implies that if a letter $a$ does not occur in $u$, then the matrix $\Psi_u(a)$ is the identity matrix.
Let Theorem 1\textsuperscript{[9,20]}. For all \(i \) and \(j \), \(1 \leq i \leq j \leq k \), we have \(m_{i+1,j} = |w|_{a_{ij}} \).

The following example of a generalized Parikh matrix might at this stage seem a bit complicated and artificial. However, it is needed in our considerations below. Consider the binary alphabet \( \{a, b\} \), as well as the words \( u = aaaaab \) and \( w = abhabaabbaababba \). (Observe that \( w \) is a prefix of the well-known Thue-Morse word.) By Theorem 1, the generalized Parikh matrix \( \Psi_u(w) \) satisfies, for an arbitrary word \( w \),

\[
\Psi_{aaaab}(w) = \begin{pmatrix}
1 & |w|_a & |w|_{aa} & |w|_{aaa} & |w|_{aaaa} & |w|_{aaaaa} \\
0 & 1 & |w|_a & |w|_{aa} & |w|_{aaa} & |w|_{aaaa} \\
0 & 0 & 1 & |w|_a & |w|_{aa} & |w|_{aaa} \\
0 & 0 & 0 & 1 & |w|_a & |w|_{aa} \\
0 & 0 & 0 & 0 & 1 & |w|_a \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

For our particular choice of \( w \) (the prefix of the Thue-Morse word), we obtain

\[
\Psi_{aaaab}(w) = \begin{pmatrix}
1 & 8 & 28 & 56 & 70 & 87 \\
0 & 1 & 8 & 28 & 56 & 98 \\
0 & 0 & 1 & 8 & 28 & 70 \\
0 & 0 & 0 & 1 & 8 & 32 \\
0 & 0 & 0 & 0 & 1 & 8 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

The significance of the boldface numbers will become clear below.

### 3. Subword histories and linearization

The definition given below for the notion of a subword history, \( SH \), follows essentially \[10\]. A subword history is a numerical quantity, associated to a variable word \( w \), polynomial in some numbers \( |w|_u \), where each \( u \) is a word over the basic alphabet \( \Sigma \). Thus, given a word \( w \), we do not consider only the number of occurrences of one word \( u \) as a subword of \( w \). There may be (finitely) many such words \( u \), and we may form sums, differences and products between the various quantities \( |w|_u \).

**Definition 3.** Let \( \Sigma \) be an alphabet and \( w \in \Sigma^* \). A subword history in \( \Sigma \) and its value for \( w \) are defined recursively as follows. For every \( x \in \Sigma^* \), \(|x|_x \) is a subword history in \( \Sigma \), referred to as monomial, and its value for \( w \) equals \(|w|_x \). Assume that \( SH_1 \) and \( SH_2 \) are subword histories in \( \Sigma \), with values \( \alpha_1 \) and \( \alpha_2 \) for \( w \), respectively. Then so are

\[-(SH_1), \quad (SH_1) + (SH_2), \quad \text{and} \quad (SH_1) \times (SH_2),\]

with values for \( w \)

\[-\alpha_1, \alpha_1 + \alpha_2, \quad \text{and} \quad \alpha_1\alpha_2,\]

respectively.

A subword history is linear if it is obtained without using the operation \( \times \). Two subword histories \( SH_1 \) and \( SH_2 \) are termed equivalent, written \( SH_1 = SH_2 \), if they assume the same value for any \( w \).
Subword histories have been used also as a tool in language theory, [2,15,17–19]. We will use here natural abbreviations. For instance, instead of $||ab + ||ab + ||ab + ||ab$ we write $4||ab$. The alphabet $\Sigma$ is understood as the minimal alphabet for the words $u$ appearing in the given $SH$. Thus,

$$SH = ||ab \times ||bc - ||abc - ||abc - 2||c + 3||bc$$

is a subword history over the alphabet $\{a, b, c\}$. For the word $w = ababc2$, it assumes the value $3 \cdot 5 - 7 - 2 - 2 \cdot 3 + 3 \cdot 4 = 12$. This will also be denoted by

$$SH(ababc^2, ab \times bc - abc - babc - 2c + 3bcc) = 12.$$ 

The following result is due to [10].

**Lemma 1.** For every subword history an equivalent linear subword history can be effectively constructed.

The construction of a linear subword history equivalent to a given subword history is important in the remainder of this paper. Therefore, we now outline the construction, somewhat simplifying the ideas from [10]. The idea is to replace the product $||u \times ||v$, with an equivalent sum.

We have to consider the shuffle $u \cup v$ of two words $u$ and $v$, consisting of all words

$$u_0v_0u_1v_1 \ldots u_kv_k,$$

where $k \geq 0$, $u_i, v_i \in \Sigma^*$ for $0 \leq i \leq k$, and

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

It is easy to see that, if $u$ and $v$ are words over disjoint alphabets, then the product of the subword histories determined by $u$ and $v$ is equivalent to the subword history determined by $\sum_{x \in \Sigma^*} x$. If the alphabets of $u$ and $v$ are not disjoint, we obtain only the inequality

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

where the equality is a rare exception.

We now simply force the alphabets to be disjoint. For $\Sigma$, we consider the primed version $\Sigma'$, $\Sigma' = \{a' \mid a \in \Sigma\}$. Let $g : \Sigma^* \rightarrow \Sigma'^*$ be the morphism defined by $g(a) = a'$. Also, let $h : (\Sigma' \cup \Sigma'^*)^* \rightarrow \Sigma^*$ be the morphism defined by $h(a) = h(a') = a$. Consider the set of rewriting rules $\{aa' \rightarrow a \mid a \in \Sigma\}$.

For two words $u, v$, we define $G(u, v) = u \cup v$.

Consider $x, y \in (\Sigma' \cup \Sigma'^*)^*$. The relation of $m$-reduction, denoted $\vdash_m$, for $m \geq 0$, holds exactly in case $y$ can be obtained from $x$ by applying in parallel $m$ rewriting rules. (If $m = 0$, then $x \equiv y$.)

A word $r \in \Sigma^*$ is called an $m$-reduction of the pair $(u, v)$, if and only if 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:

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

Finally, we denote

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

The products can now be eliminated, using the formula

$$SH(w, u \times v) = SH\left(w, \sum_{r \in R(u, v)} t(r) r\right),$$

valid for all words $w, u, v$. Some examples will be considered in the next section. □

**Remark.** The multiplicity $t(r)$ was defined in our original manuscript, and also in [10], by

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

We thank the referee for the following observation. In some cases (one could explicitly characterize them) it is not sufficient to count the x's but the number of pairs $(x, y)$ should be counted, as done in the definition of $t(r)$ in the proof of Lemma 1. For instance, consider the subword history $||a^2 \times ||a^2$. We obtain

$$G(aa, aa) = [aa'a', aa'aa', aaaa', aa'a', a'aa', aa'a']$$

Now the multiplicity of the 1-reduction $aaa$ of the pair $(aa, aa)$ is 5 according to the old formula, whereas it is 6 according to the new formula. The difference is due to the fact that the value $x = aa'a'$ should be counted twice because it gives rise to two different values of $y$. And only the value 6 leads to the correct formula

$$||a^2 \times ||a^2 = 6||a^4 + 6||a^3 + ||a^2.$$ □

The equivalence of two given subword histories is decidable, [10]. On the other hand, the decidability of the inequality problem, [19], is open: given $SH_1$ and $SH_2$, is the value of $SH_1$ at most that of $SH_2$ for all words $w$? Significant contributions towards the solution of this problem, as well as a general conjecture, are contained in [5].

Many specific inequalities between subword histories can be established using Parikh matrices, [10,11,14]. Of special interest is the Cauchy inequality, [10],

$$||y \times ||z \leq ||x \times ||yz,$$

valid for all words $x, y, z$. The inequality contains essential information because it reduces to an equality in numerous cases.
4. Matrices associated to subword histories

We now develop Definition 3 further. For a finite language $F$, we consider words $x_F$ such that every word in $F$ is a factor of $x_F$, as well as the shortest possible length $\mu(F)$ among such words. For $y$ being a factor of $x$ we use the notation $y|x$.

**Definition 4.** For a nonempty language $F$, define

$$\text{factor}(F) = \{x_F| \text{for all } y \in F, \ y|x_F\}.$$ 

Furthermore, let $\mu(F)$ be the smallest length of the words in factor($F$).

Clearly, the catenation of all words in $F$, taken in any order, belongs to factor($F$). This gives an upper bound for $\mu(F)$: the sum of the lengths of all words in $F$. In most cases the actual value of $\mu(F)$ is much smaller. Very little is known in the general case. If $F$ consists of two words of the same length, then $\mu(F)$ is smaller than twice the length exactly in case some nonempty suffix of one word is a prefix of the other. For languages $F$ of cardinality at least 3, the determining of $\mu(F)$ is more involved and leads to several cases.

If $F$ consists of all words up to a specific length $m$, then words in factor($F$) are customarily referred to as de Bruijn words, and

$$\mu(F) = k^m + m - 1$$

if the alphabet contains $k$ letters. For a proof of this equation see, for instance, [8], p. 20.

Of particular interest for us is the case where $F$ consists of all words appearing in a given subword history $SH$. Then, for any $u \in \text{factor}(F)$, the value of $SH$ for a word $w$ can be computed (by additions, subtractions and multiplications) from entries of the Parikh matrix $\Psi_u(w)$.

Thus, let $F_{SH}$ be the set of all words appearing in a subword history $SH$. This notation should be clear, for instance,

$$F_{SH} = \{ab, bc, abc, bab, c, bcc\}$$

if $SH$ is the subword history considered before Lemma 1. We define now

$$\mu(SH) = \mu(F_{SH}).$$

The next result is an immediate consequence of Theorem 1 and Definition 4.

**Theorem 2.** Consider a subword history $SH$ and a word $w$. Let $u$ be a word in factor($F_{SH}$) and $||_u$ be an arbitrary monomial component in $SH$. Then $|w|_u$ appears as an entry in the generalized Parikh matrix $\Psi_u(w)$. Consequently, the value of $SH$ for $w$ is obtained from the entries of the matrix $\Psi_u(w)$ by the arithmetical operations present in $SH$. The matrix can be chosen to be of dimension at most $\mu(SH) + 1$.

By Lemma 1, we can in Theorem 2 restrict the arithmetical operations to additions and subtractions. However, then we may have to operate with matrices of a higher dimension because $F_{SH}$ may change in the linearization process of Lemma 1.

The trade-off between the dimension of matrices and absence of multiplication must be considered in each particular situation.

Consider the subword history $||_a \times ||_{ab}$. By Theorems 1 and 2, the value of a word $w$ for this subword history equals the square of the entry in the upper right-hand corner of the matrix $\Psi_{ab}(w)$. For the (already considered) prefix of the Thue-Morse word, we obtain

$$\Psi_{ab}(ababaabbaababba) = \begin{pmatrix} 1 & 8 & 32 \\ 0 & 1 & 8 \\ 0 & 0 & 1 \end{pmatrix}.$$ 

Thus, the value we are looking for is $32^2 = 1024$.

We can also use the construction of Lemma 1, and obtain an equivalent linear subword history:

$$||_a \times ||_{ab} = 2||_{aab} + 4||_{aabb} + 2||_{aab} + 2||_{abb} + ||_{ab}.$$

Now we have to consider words in

$$\text{factor}(F_{SH}) = \text{factor}(aab, aabb, ab, abb, abab).$$

(The words have been permuted to get a correspondence with the matrix positions when the latter are in the natural order.) It turns out that $\mu(SH) = 8$, and $u = aabbabab \in \text{factor}(F_{SH})$. Consequently, the matrices $\Psi_u(w)$ will be 9-dimensional but for our subword history it suffices to know the values in the positions

$$(1, 4), (1, 5), (2, 4), (2, 5), (5, 9).$$

For our particular $w$, they are

$$(70, 98, 32, 70, 160),$$

respectively. Substituted in the linear subword history above they yield the correct sum 1024.
Consider next the subword history
\[ SH_1 = (||a||^4 + ||ab||. \]

By Theorem 2, the values of \(SH_1\) can be computed from the entries \((1, 2)\) and \((1, 3)\) of the matrices \(\Psi_{ab}(w)\). Using the same word \(w\) as above, as well as the associated matrix, we obtain the value \(8^4 + 32 = 4128\). However, by the construction of Lemma 1, we obtain an equivalent linear subword history
\[ SH_2 = 24||a^4| + 36||a^3 + 14||a^2| + ||a| + ||ab|. \]

Now we have \(\mu(SH_2) = 5\) and \(a^4b \in \text{factor}(F_{SH_1})\). The matrix \(\Psi_{ab}(w)\) was already computed above, at the end of Section 2. The entries we need are marked by boldface, and give the required result
\[ 24 \cdot 70 + 36 \cdot 56 + 14 \cdot 28 + 8 + 32 = 4128. \]

These constructions might give the false impression that linearization leads into more complicated calculations. However, the opposite is the case. Linear mappings based on matrices are simpler and also easier to handle theoretically.

5. Powers of words and subword histories

A general problem is to use the values of a subword history \(SH\) for some word(s) \(w\) to compute the values of \(SH\) for some other words. Not much is known about this problem, and we hope to return to it in another context. The matrix representation of Theorem 2 gives definite possibilities in this direction.

In this section we consider powers \(w^n\) of a given word \(w\). We begin with the following result concerning monomial subword histories.

**Theorem 3.** For all words \(w\) and \(u\) with \(|u| = k \geq 1\), there is a polynomial \(P_k(n)\) of degree \(k\) such that the equation
\[ |w^n|_u = P_k(n) \]
holds for every \(n \geq 0\). Given \(w\) and \(u\), the polynomial \(P_k(n)\) can be effectively constructed. It has rational coefficients and the constant term 0.

**Proof.** We denote by \(M\) the matrix \(\Psi_u(w)\). (Observe that \(M\) depends on \(u\) and \(w\) and is of dimension \(k + 1\).) Since the mapping \(\Psi_u\) is a morphism, we have \(\Psi_u(w^n) = M^n\) and thus we obtain by Theorem 1,
\[ |w^n|_u = p(M^n). \]

where \(p\) is the projection taking the upper right-hand corner entry from the matrix. The existence of the polynomial \(P_k\) now follows by the Cayley–Hamilton Theorem. The coefficients are determined by considering the first few powers of \(w^n\). This leads to a system of linear equations with integer coefficients, so the coefficients of the polynomial are rational. The claim about the constant term is obvious because \(u\) does not occur in the empty word. \(\square\)

Observe that, in the case of a one-letter alphabet, Theorem 3 is a direct consequence of the definition of a binomial coefficient.

As an illustration of the construction of Theorem 3, we consider the monomial subword history \(||b^3a\) and \(w = ab\). Thus, we determine a polynomial \(P_4(n)\) such that
\[ |(ab)^n|_{b^3a} = P_4(n). \]

Denote \(P_4(n) = e_0n^4 + e_1n^3 + e_2n^2 + e_3n\). From the values
\[ |(ab)^n|_{b^3a}, \quad 1 \leq n \leq 4, \]

we obtain the system of equations
\[ e_0 + e_1 + e_2 + e_3 = 0, \]
\[ 16e_0 + 8e_1 + 4e_2 + 2e_3 = 0, \]
\[ 81e_0 + 27e_1 + 9e_2 + 3e_3 = 0, \]
\[ 256e_0 + 64e_1 + 16e_2 + 4e_3 = 1. \]

This yields the solution (observe the connection to Vandermonde determinants!)
\[ e_0 = 1/24, \quad e_1 = -1/4, \quad e_2 = 11/24, \quad e_3 = -1/4. \]

Hence, we obtain the final result
\[ |(ab)^n|_{b^3a} = n(n^3 - 6n^2 + 11n - 6)/24, \quad n \geq 0. \]

We obtain also the following Corollary of Theorem 3.
Corollary 1. Let $w, w', u$, $|u| = k \geq 1$, be arbitrary words. If $|w^n|_u = |(w')^n|_u$ holds for every $n$, $0 \leq n \leq k$, it holds for all $n \geq 0$.

Theorem 3 deals with monomial subword histories $||_w$. It can be extended to concern arbitrary subword histories $SH$. We consider the set of words $F_{SH}$ present in $SH$ and choose a word $u \in \text{factor}(F_{SH})$. Given a word $w$, we compute the generalized Parikh matrix $\Psi_{w}(w)$. We now proceed similarly as in Theorem 3 and obtain, for every $u \in F_{SH}$, a polynomial $P_u(n)$ such that

$$|w^n|_u = P_u(n)$$

holds for all $n \geq 0$. These polynomials yield, by additions, subtractions and multiplications based on $SH$, a polynomial $P_{SH}$ such that the value of $SH$ for $w^n$, $n \geq 0$, equals $P_{SH}(n)$. Thus, we obtain the following result.

Theorem 4. Given a subword history $SH$ and a word $w$, one can effectively construct a polynomial $P$ in the variable $n$ such that, for all $n \geq 0$, the value of $SH$ for $w^n$ equals $P(n)$.

As a simple illustration, consider the subword history $SH = ||_{ab} \times ||_{ab}$ and the word $w = baa$. Now $F_{SH}$ consists of the word $ab$ alone, and we can take the mapping $\Psi_{ab}$. Thus, the matrices will be 3-dimensional. It suffices to consider the values $|baa|_{ab}$ and $|baaba|_{ab}$ to obtain the polynomial $P_{ab}(n) = n(n - 1)$. Consequently,

$$P_{SH}(n) = n^2(n - 1)^2.$$

The degree of the polynomial $P(n)$ constructed in Theorem 4 can be greater than $\mu(SH)$. This is due to the multiplications possibly present in $SH$, as is the case in the example. However, we can construct an equivalent subword history without multiplications, by Lemma 1. In this way we obtain an explicit upper bound for the degree of the polynomial, as stated in the following theorem.

Theorem 5. Given a linear subword history $SHL$ and a word $w$, one can effectively construct a polynomial $P$ in the variable $n$, of degree at most $\mu(SHL)$, such that the value of $SHL$ for $w^n$ equals $P(n)$, for all $n \geq 0$.

In most cases the degree of the polynomial will be less than $\mu(SHL)$, since in most cases the longest word in $F_{SHL}$ is shorter than $\mu(SHL)$.

Assume that $SHL$ is the linear subword history equivalent to a given subword history $SH$. Two linear subword histories are equivalent only if they are identical, up to the order of terms, [10]. Therefore, we may define

$$\mu_1(SH) = \mu(SHL).$$

Thus, $\mu_1(SH)$ equals the $\mu$-value of the linear subword history equivalent to $SH$. It follows that $\mu_1(SH)$ constitutes an upper bound for the degree of the polynomial $P$ in Theorem 4. We obtain also the following result, corresponding to Corollary 1.

Corollary 2. Let $SH$ be a subword history and let $w, w'$, be arbitrary words. If, for $0 \leq n \leq \mu_1(SH)$, the subword history $SH$ assumes the same value for both $w^n$ and $(w')^n$, then $SH$ assumes the same value for the two words whenever $n \geq 0$.

We conclude the paper with some further illustrations of the constructions. Consider the generalized Parikh matrix mapping $\Psi_{aaba}$. Consequently, for every word $w$, we have by Theorem 1

$$\Psi_{aaba}(w) = \begin{pmatrix}
1 & |w|_a & |w|_{aa} & |w|_{aab} & |w|_{aaba} \\
0 & 1 & |w|_a & |w|_{ab} & |w|_{aba} \\
0 & 0 & 1 & |w|_b & |w|_{ba} \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$ 

Now it turns out that both the subword history

$$SH = ||_{ab} \times ||_{a} + ||_{aaba}$$

and the equivalent linear subword history

$$SHL = 2||_{aab} + ||_{aba} + ||_{ab} + ||_{aaba}$$

can be fully characterized in terms of the matrices $\Psi_{aaba}(w)$.

Consider the word $w = ababaaab$. The matrices needed for computations are, for $u = aaba$,

$$\Psi_u(w) = \begin{pmatrix}
1 & 4 & 6 & 7 & 2 \\
0 & 1 & 4 & 8 & 10 \\
0 & 0 & 1 & 4 & 8 \\
0 & 0 & 0 & 1 & 4 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}, \quad \Psi_u(w^2) = \begin{pmatrix}
1 & 8 & 28 & 70 & 120 \\
0 & 1 & 8 & 32 & 84 \\
0 & 0 & 1 & 8 & 32 \\
0 & 0 & 0 & 1 & 8 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix},$$

$$\Psi_u(w^3) = \begin{pmatrix}
1 & 12 & 66 & 253 & 706 \\
0 & 1 & 12 & 72 & 286 \\
0 & 0 & 1 & 12 & 72 \\
0 & 0 & 0 & 1 & 12 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.$$
\[
\Psi_u(w^4) = \begin{pmatrix}
1 & 16 & 120 & 620 & 2368 \\
0 & 1 & 16 & 128 & 680 \\
0 & 0 & 1 & 16 & 128 \\
0 & 0 & 0 & 1 & 16 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

The coefficients of the relevant linear systems of equations are seen from the matrices, and we obtain the polynomials

\[
|w^n_a| = 4n, \quad |w^n|_{ab} = 8n^2, \quad |w^n|_{aba} = n(32n^2 - 2)/3, \\
|w^n|_{aab} = n(32n^2 - 12n + 1)/3, \quad |w^n|_{aaba} = n(32n^3 - 16n^2 - 2n - 8)/3.
\]

Consequently, the value of the original subword history \( ||ab \times ||a + ||aaba \) for the word \((abbabaab)^n\) equals \( n(32n^3 + 80n^2 - 2n - 8)/3 \).

This polynomial is obtained both from the polynomials present in \( SH \) or from the ones present in the equivalent \( SHL \). It can also be computed directly from the first four values \(34, \ 376, \ 1570, \ 4416\) of the subword history. For instance,

\[
SH((abbabaab)^{100}, ||ab \times ||a + ||aaba) = 1.093.326.400.
\]

6. Conclusion

We have seen that there is a simple connection between subword histories and generalized Parikh matrices. Each subword history \( SH \) is completely characterized by a suitably chosen matrix mapping \( \Psi_u \). For a word \( w \), values of \( SH \) for words in \( w^* \) can be expressed as polynomial functions. We hope to return in another paper to other applications of the matrix connection. Further facts about \( \text{factor}(F) \) and \( \mu(F) \) might be useful in studies concerning finite languages.

Acknowledgements

We want to thank the two referees for many useful suggestions. The important contribution of one of the referees was already acknowledged in the remark in Section 3.

References