2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 1–18

On Correlation Polynomials and Subword Complexity

Irina Gheorghiciuc[1] and Mark Daniel Ward[2†]

[1] *Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA.* gheorghi@math.udel.edu
[2] *Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, USA.* ward2@math.upenn.edu

We consider words with letters from a q-ary alphabet $\mathcal{A}$. The kth subword complexity of a word $w \in \mathcal{A}^*$ is the number of distinct subwords of length k that appear as contiguous subwords of w. We analyze subword complexity from both combinatorial and probabilistic viewpoints. Our first main result is a precise analysis of the expected kth subword complexity of a randomly-chosen word $w \in \mathcal{A}^n$. Our other main result describes, for $w \in \mathcal{A}^*$, the degree to which one understands the set of all subwords of w, provided that one knows only the set of all subwords of some particular length k.

Our methods rely upon a precise characterization of overlaps between words of length k. We use three kinds of correlation polynomials of words of length k: unweighted correlation polynomials; correlation polynomials associated to a Bernoulli source; and generalized multivariate correlation polynomials. We survey previously-known results about such polynomials, and we also present some new results concerning correlation polynomials.

Keywords: Analytic methods, asymptotics, autocorrelation, average-case analysis, combinatorics on words, correlation polynomial, de Bruijn graph, depth, subword complexity, suffix trees.

1 Introduction.

For a fixed integer $q > 1$, we consider the q-ary alphabet $\mathcal{A} = \{a_1, \ldots, a_q\}$. The set of all words of length n on $\mathcal{A}$ is denoted by $\mathcal{A}^n$. The set of all finite words on $\mathcal{A}$ is denoted by $\mathcal{A}^*$. We use ε to denote the (empty) word of length zero. A subword of a finite or infinite word w over $\mathcal{A}$ is a finite block of consecutive letters of w. By $L_n(w)$ we denote the set of all the subwords of length n of w. The language of w is the set of all the subwords of w. Throughout this paper we denote the ith character of a word w by w_i. The concatenation of two words u and v is denoted by uv.

The subword complexity of w is the function $f_w : \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0}$ that assigns to each positive integer n the cardinality of $L_n(w)$. We say that $f_w(n)$ is the nth subword complexity of w. Clearly, $f_w(n) = 0$ if and only if w has length $|w| < n$. In the literature, subword complexity is sometimes referred to as symbolic or block complexity.

Intuitively, the subword complexity measures the degree of randomness of a word. For example, an infinite word w has a bounded subword complexity if and only if w is ultimately periodic. At the same time the expansion of a normal number has an exponential subword complexity function. We can also think of subword complexity as a characteristic of the "size" of the language of a word.

The subword complexity f_w of an infinite word w is a non-decreasing function with the property that, if $f_w(n) = f_w(n+1)$ for some $n \in \mathbb{Z}_{>0}$, then $f_w(i) = f_w(n)$ for all $i \geq n$. A list of other known properties of subword complexity of infinite words is given in (Fer99). Results on the subword complexity of finite words can be found in (dL99) and (JLS04). In particular, it was proved in (dL99) that the subword complexity of a finite word w is unimodal. Moreover, if $f_w(n) < f_w(n+1)$ for some $n \in \mathbb{Z}_{>0}$, then $f_w(i) = f_w(i+1) - 1$ for $n \leq i < |w|$.

In this paper we study the subword containment and subword complexity of finite words only. Whenever we randomly select a word, we assume that the letters are selected from $\mathcal{A}$ independently of each other and are each generated by a stationary Bernoulli source. In other words, there is a set of probabilities $\{p_1, \ldots, p_q\}$ such that letter $a_i \in \mathcal{A}$ has probability p_i of being selected. We write

$$\mathbf{P}(w) = \prod_{i=1}^{q} p_i^{b_i}$$

†MDW's research was supported by NSF Grant 0603821.

if w has b_i occurrences of letter a_i (for each i).

We concentrate on the kth subword complexity for a fixed k from several points of view. One of our goals is to characterize the random variable that represents the kth subword complexity of a randomly selected word $w \in \mathcal{A}^n$; in Theorem 2.1, we obtain the expected value of the kth complexity of a word of length n. A special case of this result, using uniform probabilities (i.e., $p_1 = p_2 = \cdots = p_q$), was obtained in (JLS04) using a different method.

Another interesting problem we consider is, for $w \in \mathcal{A}^*$, the extent to which we can draw conclusions about all the sets $L_m(w)$, provided that we only know the set $L_n(w)$ for some n. More precisely, let S be a set of "banned" subwords of length n, and let m be an integer, $1 \leq m \leq n$. We consider all the words $w \in A^*$ with the property that $L_n(w) \cap S = \emptyset$. In Theorem 2.4 we obtain the multivariate generating function that gives all the possible sets $L_m(w)$ and the frequency of each subword of length m in each such w. In the case $m = 1$ this problem was solved in (GO81b), (GJ79) and (NZ99).

To answer the questions posed above, we use combinatorics on words to precise characterize overlaps between words of length k. We use three kinds of correlation polynomials of words of length k: unweighted correlation polynomials (similar to the ones used by Guibas and Odlyzko in (GO78), (GO81a) and (GO81b)); correlation polynomials associated to a Bernoulli source (as defined by Régnier and Szpankowski in (RS98), (JS05) and (RD04)); and generalized multivariate correlation polynomials. A comparison of the methods used by Goulden-Jackson, Guibas-Odlyzko, Noonan-Zeilberger, and Régnier-Szpankowski can be found in (Kon05).

Nicodème et al. (PN02) consider automata and translation to generating functions by the Chomsky-Schützenberger algorithm. In particular, they analyze the statistics of the number of occurrences of a regular (contiguous) expression pattern in a regular text; an extension of the method is found in (Nic03). Both Park et al. (PHNS06) and Ward (War07) concern profiles of suffix-trees, which are intimately related to the number of repeated subwords. The studies (BK93), (Fay04), (JS94), (JS05), and (RR03) and are devoted to asymptotic analysis of related pattern matching problems.

In molecular biology the subword complexity of finite words is used to study DNA sequences, in particular the structure of certain genes. See, for example, (AC00) and (TAK$^+$02). Applications concerning subword complexity also include dynamical systems, ergodic theory, number theory and theoretical computer science. For surveys see, for instance, (All94) and (Fer99).

The definitions in Section 1.1 are necessary for the understanding of Theorem 2.1. The definitions in Sections 1.2 and 1.3 are necessary for understanding both Theorem 2.4 and the auxiliary results about correlation matrices in Section 5.

1.1 Univariate Correlation Polynomials

We define the correlation set of two words w and u, each of length k, as

$$\mathcal{S}_{w,u} = \{u_{i+1} \dots u_k \mid w_{k-i+1} \dots w_k = u_1 \dots u_i;\ \ 1 \leq i \leq k\}$$

(see (JS05) and (RD04)). The set of positions i used in defining the correlation set is defined as $\mathcal{P}(w,u)$, i.e.,

$$\mathcal{P}(w,u) = \{i \mid w_{k-i+1} \dots w_k = u_1 \dots u_i;\ \ 1 \leq i \leq k\}\,.$$

We note that v occurs as both a suffix of w (say, $w = xv$) and a prefix of u (say, $u = vy$) if and only if $y \in \mathcal{S}_{w,u}$ and $|v| \in \mathcal{P}(w,u)$. In other words, $y \in \mathcal{S}_{w,u}$ if and only if wy has u as a suffix, i.e., appending y to the end of w yields u as an overlapping suffix. The autocorrelation set of w, defined by

$$\mathcal{P}(w) = \{i \mid w_{k-i+1} \dots w_k = w_1 \dots w_i;\ \ 1 \leq i \leq k\}\,,$$

characterizes the overlaps of w with itself.

We define the unweighted correlation polynomial $C_{w,u}(z)$ as the generating function of $\mathcal{S}_{w,u}$ with unweighted coefficients, namely

$$C_{w,u}(z) = \sum_{i \in \mathcal{P}(w,u)} z^{k-i}\,.$$

In order to enumerate the overlaps of w with itself, we define the unweighted autocorrelation polynomial of w as $C_w(z) := C_{w,w}(z)$.

In the case of a stationary Bernoulli source, when the letters are selected from $\mathcal{A}$ independently of each other, we introduce the weighted correlation polynomial $S_{w,u}(z)$ as the generating function of $\mathcal{S}_{w,u}$ with weighted coefficients, namely

$$S_{w,u}(z) = \sum_{i \in \mathcal{P}(w,u)} \mathbf{P}(u_{i+1} \dots u_k) z^{k-i} \,.$$

As a special case of $S_{w,u}(z)$, we define the weighted autocorrelation polynomial of w as

$$S_w(z) := S_{w,w}(z) \,. \tag{1}$$

In the example below we demonstrate the definitions presented so far. To determine the set $\mathcal{S}_{w,u}$, for each i, $1 \le i \le k$, we place u under w such that the first character of u is under the ith character of w. If the characters of w and u in the overlapping segment are the same, then the suffix y of u that follows the overlap is in $\mathcal{S}_{w,u}$.

Example 1.1 *Consider an alphabet of size $q = 3$, denoted as $\mathcal{A} = \{a_1, a_2, a_3\} = \{1, 2, 3\}$. Consider the words $w = 313212$ and $u = 212133$. The following matrix shows the contributions to $C_{w,u}(z)$ and $S_{w,u}(z)$ from various overlaps of w and u.*

$$\begin{array}{rccccccccccccc}
w: & 3 & 1 & 3 & 2 & 1 & 2 & & & & & & \textit{unweighted} & \textit{weighted} \\
u: & 2 & 1 & 2 & 1 & 3 & 2 & & & & & & 0 & 0 \\
 & & 2 & 1 & 2 & 1 & 3 & 2 & & & & & 0 & 0 \\
 & & & 2 & 1 & 2 & 1 & 3 & 2 & & & & 0 & 0 \\
 & & & & 2 & 1 & 2 & 1 & 3 & 2 & & & z^3 & \mathbf{P}(132)z^3 \\
 & & & & & 2 & 1 & 2 & 1 & 3 & 2 & & 0 & 0 \\
 & & & & & & 2 & 1 & 2 & 1 & 3 & 2 & z^5 & \mathbf{P}(12132)z^5
\end{array}$$

Thus $\mathcal{S}_{w,u} = \{132,\ 12132\}$. Also, $\mathcal{P}(w,u) = \{1,3\}$, since w has suffixes of lengths 1 and 3 in common with prefixes of u. We note that $C_{w,u}(z) = z^3 + z^5$ and $S_{w,u}(z) = \mathbf{P}(132)z^3 + \mathbf{P}(12132)z^5$.

Observation 1.2 *In the case where all words from $\mathcal{A}$ are equiprobable (i.e., $\mathbf{P}(a_i) = 1/q$ for each i), the weighted correlation polynomial $S_{w,u}(z) = \sum_{i \in \mathcal{P}(w,u)} \mathbf{P}(u_{i+1} \dots u_k) z^{k-i}$ and unweighted correlation polynomial $C_{w,u}(z) = \sum_{i \in \mathcal{P}(w,u)} z^{k-i}$ are related by*

$$C_{w,u}(z) = S_{w,u}(z/q) \,.$$

It is an interesting fact that the set of unweighted autocorrelation polynomials of words of length k over $\mathcal{A} = \{a_1, \dots, a_q\}$ does not depend on q (as long as $q \ge 2$) and is of order $k^{\log k}$ (see (GO81a) and (HHI00)). The unweighted correlation polynomial $C_w(z)$ has a probabilistic meaning. Consider the experiment which consists of repeated throws of a fair q-sided die with faces $a_1, a_2, \dots, a_q$. Then the expected waiting time until w appears is $q^{|w|} p_w(1/q)$. This result was first proved in (Sol66).

Unweighted correlation polynomials are also used for counting the number of words of any given length that do not contain subwords from a given set of "forbidden" subwords. Consider a "forbidden" set of i words $S = \{u_1, \dots, u_i\} \subseteq \mathcal{A}^k$. Without loss of generality, assume that the u_j are in increasing lexicographic order. We define $\mathrm{C}_S(z)$ as the matrix whose entries are the unweighted correlation polynomials associated with the words in the set S. In other words,

$$\mathrm{C}_S(z) = \begin{pmatrix} C_{u_1,u_1}(z) & C_{u_1,u_2}(z) & \cdots & C_{u_1,u_i}(z) \\ \vdots & \vdots & \ddots & \vdots \\ C_{u_i,u_1}(z) & C_{u_i,u_2}(z) & \cdots & C_{u_i,u_i}(z) \end{pmatrix} . \tag{2}$$

The following proposition is a direct consequence of Theorem 1 in (GO81b).

Proposition 1.3 *Let $S = \{u_1, \dots, u_i\} \subseteq \mathcal{A}^k$. By $c_S(n)$ we denote the number of words of length n over $\mathcal{A}$ that do not contain any subwords from the set S. Let $F_S(z)$ be the generating function of $c_S(n)$, namely*

$$F_S(z) = \sum_{n=0}^{\infty} c_S(n) z^n \,.$$

Then

$$F_S(z) = 1/[1 - zq + z^i \mathit{trace}(\mathrm{C}_S(z)^{-1}\mathrm{E})] \,,$$

where $\mathrm{C}_S(z)$ is the matrix defined above and E is the $i \times i$ matrix whose entries are all 1.

Proposition 1.3 is of peculiar interest to us. For fixed positive integers k and l, we can use the result of Proposition 1.3 in combination with the method of inclusion-exclusion to obtain a generating function for the number of words of length n whose k complexity equals l. However, this generating function will have 2^{q^k} terms and will be computationally ineffective.

1.2 Multivariate Correlation Polynomials

Throughout the following discussion, we consider an integer $m > 0$ that remains fixed.

We temporarily let $v_0, \ldots, v_{q^m-1}$ denote the q^m words of $\mathcal{A}^m$, listed in increasing lexicographic order. Then the type $\tau_m(w)$ of a word $w \in \mathcal{A}^*$ is the monomial $z_0^{k_0} z_1^{k_1} \ldots z_{q^m-1}^{k_{q^m-1}}$, where k_i is the number of occurrences of v_i as a subword in w. In other words, the type $\tau_m(w)$ gives the number of occurrences of each subword of length m in w.

Example 1.4 *For example, consider the case $q = 2$ and $\mathcal{A} = \{a_1, a_2\} = \{0, 1\}$. The binary word $w = 001111$ has type $\tau_3(001111) = z_1 z_3 z_7^2$, because w has the subwords $v_1 = 001$ occurring once, $v_3 = 011$ occurring once, and $v_7 = 111$ occurring twice.*

We note that, if $|w| < m$, then $\tau_m(w) = 1$, because w is too short to have any subwords of length m. The type $\tau_m(w)$ of a word w is a modification of the notion of type introduced by Goulden and Jackson in (GJ79).

Now we generalize the concept of the unweighted correlation polynomial $C_{w,u}(z)$, for $w, u \in \mathcal{A}^k$, where $k \geq 2m - 2$. Let $\mathbf{z} = (z_0, z_1, \ldots, z_{q^m-1})$. The mth multivariate unweighted correlation polynomial of $w, u \in \mathcal{A}^k$ is defined as

$$C_{w,u}^{(m)}(\mathbf{z}) = \sum_{y \in \mathcal{S}_{w,u}} \tau_m(y'),$$

where y' is defined as follows: We write $r = r_1 \ldots r_{2k-i} = xvy$, where $|v| = i$, and $w = xv$, and $u = vy$. Then $y' = r_{k-2m+3} \ldots r_{2k-i-(m-1)}$. In other words we remove the last $m - 1$ characters of r, and then y' is formed by taking a suffix of length $|y| + (m - 1)$ of what remains.

In order to enumerate the overlaps of w with itself, we define the mth multivariate unweighted autocorrelation polynomial of w as $C_w^{(m)}(\mathbf{z}) := C_{w,w}^{(m)}(\mathbf{z})$.

Example 1.5 *Consider the alphabet $\mathcal{A} = \{a_1, a_2\} = \{0, 1\}$ and the words $w = 100101$ and $u = 101011$. Then*

$$C_{v,w}^{(1)}(z_0, z_1) = \tau_1(011) + \tau_1(01011) = z_0 z_1^2 + z_0^2 z_1^3,$$
$$C_{v,w}^{(2)}(z_0, z_1, z_2, z_3) = \tau_2(0101) + \tau_2(010101) = z_1^2 z_2 + z_1^3 z_2^2,$$
$$C_{v,w}^{(3)}(z_0, z_1, \ldots, z_7) = \tau_3(01010) + \tau_3(0101010) = z_2^2 z_5 + z_2^3 z_5^2.$$

In the following table we compare the contributions to $S_{w,u}(z)$, $C_{w,u}^{(1)}(\mathbf{z})$, and $C_{w,u}^{(2)}(\mathbf{z})$, from various overlaps of w and u.

												$S_{w,u}(z)$	$C_{w,u}^{(1)}(z)$	$C_{w,u}^{(2)}(z)$
$w:$	1	0	0	1	0	1								
$u:$	1	0	1	0	1	1						0	0	0
		1	0	1	0	1	1					0	0	0
			1	0	1	0	1	1				0	0	0
				1	0	1	0	1	1			$\mathbf{P}(011)z^3$	$z_0 z_1^2$	$z_1^2 z_2$
					1	0	1	0	1	1		0	0	0
						1	0	1	0	1	1	$\mathbf{P}(01011)z^5$	$z_0^2 z_1^3$	$z_1^3 z_2^2$

The multivariate correlation polynomials generalize the correlation polynomials introduced earlier in this paper, as well as other types of correlation polynomials used in literature. If we substitute $\mathbf{z} = (z, z, \ldots, z)$ in the multivariate correlation polynomial $C_{v,w}^{(1)}(\mathbf{z})$, we get the unweighted correlation polynomial $C_{w,u}(z)$. If we substitute $\mathbf{z} = (zp_1, zp_2, \ldots, zp_q)$ in $C_{v,w}^{(1)}(\mathbf{z})$, where p_i is the probability of the letter $a_i \in \mathcal{A}$ in the case of a stationary Bernoulli source, we get the weighted correlation polynomial $C_{w,u}(z)$. Also, the autocorrelation polynomial for the Markovian model of order m in (RS98) is the specialization of $C_v^{(m+1)}(\mathbf{z})$ when $\mathbf{z} = (z\mathbf{P}(v_0), z\mathbf{P}(v_1), \ldots, z\mathbf{P}(v_{q^{m+1}-1}))$, where $\{v_0, \ldots, v_{q^{m+1}-1}\}$ are the words of length $m + 1$ over $\mathcal{A}$ arranged in increasing lexicographic order.

The reversal of a word $w = w_1 w_2 \ldots w_n$ is the word $w_n w_{n-1} \ldots w_1$, denoted by $\widetilde{w}$. The connector matrix for the cluster method, introduced by Goulden and Jackson in (GJ79) and used by Noonan and Zeilberger in (NZ99), has entries $e_{w,v}$, where $e_{w,w} = C_{w,w}^{(1)}(\mathbf{z}) - 1$ and $e_{w,v} = C_{\widetilde{v},\widetilde{w}}^{(1)}(\mathbf{z})$ when $w \neq u$.

Similar to the probabilistic meaning of $C_w(z)$ discussed earlier, the polynomial $C_w^{(1)}(\mathbf{z})$ also has a probabilistic meaning. Consider the experiment which consists of independent, repeated throws of a biased q-sided die with faces $a_1, a_2, \ldots, a_q$. Let p_i be the probability that letter a_i comes up on one throw. We assume that, for each a_i, the probability p_i is nonzero; otherwise, we can safely eliminate a_i from our alphabet. Let $\tau_1(w)[1/p_i]$ denote the specialization of the monomial $\tau_1(w)$ at $\mathbf{z} = (1/p_1, 1/p_2, \ldots, 1/p_q)$. It was proved in (Li80) that the expected waiting time until the word w appears is $\tau_1(w)[1/p_i]C_w^{(1)}(p_1, p_2, \ldots, p_q)$.

1.3 The de Bruijn Graph.

The de Bruijn graph $B_n(\mathcal{A})$ is the directed graph whose vertices are words from $\mathcal{A}^n$ and whose edges are words from $\mathcal{A}^{n+1}$, with the property that a directed edge of the form $w_1 \ldots w_{n+1}$ points from the vertex $w_1 \ldots w_n$ to the vertex $w_2 \ldots w_{n+1}$.

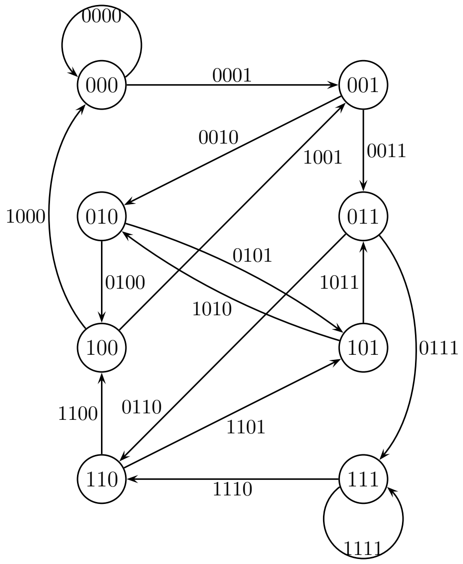

Fig. 1: The de Bruijn graph $B_3(\mathcal{A}_2)$

Observation 1.6 *Since the de Bruijn graph $B_n(\mathcal{A})$ is strongly connected and has the property $d^+(v) = d^-(v) = q$ for all vertices v of $B_n(\mathcal{A})$, then $B_n(\mathcal{A})$ is Eulerian. Also, for $n \geq 2$, the graph $B_n(\mathcal{A})$ is the line graph of $B_{n-1}(\mathcal{A})$, which implies that $B_n(\mathcal{A})$ is Hamiltonian as well. This proves the existence of a word w of length $q^n + n - 1$ with nth subword complexity $f_n(w) = q^n$.*

Let M_n denote the adjacency matrix of $B_n(\mathcal{A})$, whose rows and columns are indexed by the words of $\mathcal{A}^n$ arranged in increasing lexicographic order. The (i, j)th entry of M_n (for $0 \leq i, j \leq q^n - 1$) is "1" if

$$i = q^{n-1}a + b \qquad \text{and} \qquad j = qb + c$$

for $0 \leq a, c < q$ and $0 \leq b < q^{n-1}$, and the entry is 0 otherwise. To see this, note that the first q^{n-1} rows of M_n are just repeated over and over a total of q times, and each row has a form that is easily discernible. For instance, consider the following example.

Example 1.7 *Consider the case $q = 2$, so $\mathcal{A} = \{a_1, a_2\}$. Then the adjacency matrix of the de Bruijn graph $B_3(\mathcal{A})$ is*

$$\mathrm{M}_3 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

2 Main Results.

We use autocorrelation polynomials to compute the expected value of the kth subword complexity of a word $w \in \mathcal{A}^n$ when the letters of w are selected from $\mathcal{A}$ independently of each other and are each generated by a stationary Bernoulli source. By $\{p_1, \ldots, p_q\}$ we denote set of probabilities such that letter $a_i \in \mathcal{A}$ has probability p_i of being selected. A special case of the result below, using uniform probabilities (i.e., $p_1 = p_2 = \cdots = p_q$), was obtained in (JLS04) using a different method.

To formalize this notation, we let $Y = Y_1Y_2Y_3\ldots$ denote a sequence of symbols drawn from $\mathcal{A}$. We assume that the Y_i's are chosen independently, with $\mathbf{P}\{Y_i = a_j\} = p_j$. We let $X_{n,k}$ denote the number of distinct words from $\mathcal{A}^k$ which each appear as a subword of $Y_1Y_2\ldots Y_{n+k-1}$.

We also let $Y^{(l)} = Y_1^{(l)}Y_2^{(l)}Y_3^{(l)}\ldots$ denote, for each l, a sequence of symbols drawn from $\mathcal{A}$. Again, we assume that all of the $Y_i^{(l)}$'s are chosen independently, with $\mathbf{P}\{Y_i^{(l)} = a_j\} = p_j$. We let $\widehat{X}_{n,k}$ denote the number of distinct words in the collection $\{Y_1^{(l)}Y_2^{(l)}\ldots Y_k^{(l)} \mid 1 \le l \le n\}$.

Without loss of generality, we assume that $0 < p_1 \le p_2 \le \cdots \le p_q < 1$. We define $p := p_q$ for ease of notation. We also define $\delta = \sqrt{p}$. We choose $c > 0$ such that $p_1^{-c}\delta < 1$, and we choose ϵ with $0 < \epsilon < c$. Finally, we define $\mu = p_1^{-c}\delta$ for ease of notation.

The following two theorems are true regardless of the relationship between n and k. Even if k is a function of n or, on the other hand, k and n are treated independently of each other, the following two theorems hold. The ϵ and μ in this theorem and its corollary do *not* depend on n or k.

Theorem 2.1 *Recall that $0 < p_1 \le \cdots \le p_q < 1$, and also $p := p_q$ and $\delta = \sqrt{p}$. Consider $c > 0$ so that $\mu := p_1^{-c}\delta < 1$, and ϵ with $0 < \epsilon < c$. The difference of the average subword complexity $X_{n,k}$ compared to $\widehat{X}_{n,k}$ is asymptotically negligible. The difference satisfies*

$$\mathbb{E}[X_{n,k}] - \mathbb{E}[\widehat{X}_{n,k}] = O(n^{-\epsilon}\mu^k)\,. \tag{3}$$

The average subword complexity $X_{n,k}$ is

$$\mathbb{E}[X_{n,k}] = q^k - \sum_{w\in\mathcal{A}^k}(1-\mathbf{P}(w))^n + O(n^{-\epsilon}\mu^k)\,. \tag{4}$$

Corollary 2.2 *Consider the case where $p_1 = p_2 = \cdots = p_q = 1/q$. Recall $p := p_q = 1/q$ and $\delta = \sqrt{p} = 1/\sqrt{q}$. Consider $c > 0$ so that $\mu := p_1^{-c}\delta < 1$, and ϵ with $0 < \epsilon < c$. Then the average subword complexity $X_{n,k}$ is*

$$\mathbb{E}[X_{n,k}] = q^k - q^k(1-(1/q)^k)^n + O(n^{-\epsilon}\mu^k)\,,$$

where $\epsilon > 0$ and $\mu < 1$ are described above.

Note (see (12) below) that

$$\mathbb{E}[\widehat{X}_{n,k}] = \sum_{w\in\mathcal{A}^k}\left(1-(1-\mathbf{P}(w))^n\right)\,.$$

Equivalently,

$$\mathbb{E}[\widehat{X}_{n,k}] = q^k - \sum_{w\in\mathcal{A}^k}(1-\mathbf{P}(w))^n\,.$$

So (3) implies (4) immediately. Plugging in $p_1 = p_2 = \cdots = p_q = 1/q$ yields Corollary 2.2.

So we can simply focus our attention on proving (3); the proof of (3) begins in Section 4 below.

Next we would like to develop a tool to analyze the language of a given word w. We recall that $L_n(w)$ denotes the set of subwords of length n of w. We would like to be able to say as much as possible about the sets $L_m(w)$ provided that we know only the set $L_n(w)$ for some n. One approach is to obtain a generating function that, for given n and m, with $1 \le m \le n$, and for a set S of "banned" words of length n, gives the mth type $\tau_m(w)$ of any word w that contains no subwords from the set S. This result would give us all possible sets $L_m(w)$ and the frequencies of occurrence of subwords of length m in w when $L_n(w) \cap S = \emptyset$.

Fix integers m and n with $1 \le m \le n$. All the $q^n \times q^n$ matrices used here have their rows and columns indexed by the words from $\mathcal{A}^n$ arranged in increasing lexicographic order. Define $\mathrm{D}_n^{(m)}(\mathbf{z})$ to be the $q^n \times q^n$ diagonal matrix

$$\mathrm{D}_n^{(m)}(\mathbf{z}) = \mathrm{diag}(z_0, \ldots, z_0, z_1, \ldots, z_1, \ldots, z_{q^m-1}, \ldots, z_{q^m-1})\,; \tag{5}$$

each z_i occurs q^{n-m} times (consecutively) along the diagonal. Also define

$$\widetilde{\mathrm{P}} = (\mathrm{I} - \mathrm{D}_n^{(m)}(x\mathbf{z})\mathrm{M}_n)^{-1}\,, \tag{6}$$

where I is the identity matrix and M_n is the adjacency matrix of the de Bruijn graph $B_n(\mathcal{A})$.

For $v, w \in \mathcal{A}^n$, we let $\widetilde{p}_{v,w}$ denote the (v, w)th entry of $\widetilde{\mathrm{P}}$. By $v_0, \ldots, v_{q^n-1}$ we denote the q^n words of $\mathcal{A}^n$, listed in increasing lexicographic order.

Let K_w denote the sum of entries in the wth column of $\widetilde{\mathrm{P}}$, that is

$$K_w = \sum_{i=0}^{q^n-1} \widetilde{p}_{v_i,w}\,. \tag{7}$$

Let R_w^τ denote the weighted sum of entries in the wth row of $\widetilde{\mathrm{P}}$, as follows:

$$R_w^\tau = \sum_{i=0}^{q^n-1} \tau_m(v_i)\widetilde{p}_{w,v_i}\,. \tag{8}$$

Also let

$$G_n^{(m)}(x, \mathbf{z}) = \sum_{w\in\mathcal{A}^n} R_w^\tau. \tag{9}$$

Observation 2.3 *The multivariate functions R_w^τ, $K_w\tau_m(w)$ and $G_n^{(m)}$ have combinatorial meanings. The coefficient of $x^i\mathbf{z}^\mathbf{k}$ in the generating function R_w^τ is the number of words of length i with prefix w and m-type $\mathbf{z}^\mathbf{k}$. Similarly, the coefficient of $x^i\mathbf{z}^\mathbf{k}$ in $K_w\tau_m(w)$ is the number of words of length i with suffix w and m-type $\mathbf{z}^\mathbf{k}$. Also, the coefficient of $x^i\mathbf{z}^\mathbf{k}$ in $G_n^{(m)}$ is the number of words of length $i \geq n$ and m-type $\mathbf{z}^\mathbf{k}$.*

Theorem 2.4 *Let $S = \{u_1, u_2, \ldots, u_k\} \subseteq \mathcal{A}^n$, where the u_i are arranged in increasing lexicographic order. Let m be an integer, $1 \leq m \leq n$. We will denote the monomial $z_0^{k_0} z_1^{k_1} \ldots z_{q^m-1}^{k_{q^m-1}}$ by $\mathbf{z}^\mathbf{k}$, where $\mathbf{z} = (z_0, z_1, \ldots, z_{q^m-1})$ and $\mathbf{k} = (k_0, k_1, \ldots, k_{q^m-1})$.*

Define the multivariate generating function

$$F_S(x, \mathbf{z}) = \sum_{i\geq n,\, \mathbf{k}} f(i, \mathbf{k})x^i\mathbf{z}^\mathbf{k}\,,$$

where $f(i, \mathbf{k})$ is the number of words $w \in \mathcal{A}^i$ of type $\tau_m(w) = \mathbf{z}^\mathbf{k}$, with the restriction that w does not contain occurrences of any u_i (i.e., the u_i are forbidden from appearing as subwords of w). In other words, $F_S(x, \mathbf{z})$ gives the number of words of length $i \geq n$ that (1) do not contain any subwords from the set S, and (2) have a given list of subwords of length m and their frequencies.

We use $\widetilde{\mathrm{P}}_S$ to denote the $k \times k$ submatrix of $\widetilde{\mathrm{P}}$ by with the set of rows and columns S, where $\widetilde{\mathrm{P}}$ is defined by (6).

Let

$$\mathrm{M}_S = \begin{pmatrix} G_n^{(m)}(x, \mathbf{z}) & K_{u_1} & K_{u_2} & \ldots & K_{u_k} \\ R_{u_1}^\tau & \widetilde{p}_{u_1,u_1} & \widetilde{p}_{u_1,u_2} & \cdots & \widetilde{p}_{u_1,u_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{u_k}^\tau & \widetilde{p}_{u_k,u_1} & \widetilde{p}_{u_k,u_2} & \cdots & \widetilde{p}_{u_k,u_k} \end{pmatrix},$$

where K_w^τ, R_w^τ and $G_n^{(m)}(x, \mathbf{z})$ are defined by (7), (8) and (9) respectively.

Then

$$F_S(x, \mathbf{z}) = x^n \det \widetilde{\mathrm{P}}_S^{-1} \det \mathrm{M}_S\,. \tag{10}$$

3 Proof of Theorem 2.4

We recall that all the $q^n \times q^n$ matrices used here have their rows and columns indexed by the words from $\mathcal{A}^n = \{v_0, \dots, v_{q^n-1}\}$ arranged in increasing lexicographic order.

We define $\widetilde{\mathrm{I}}_S$ to be the $q^n \times q^n$ diagonal matrix with the (v,v)th entry equal to 0 if $v \in S$, or 1 otherwise. The $q^n \times q^n$ diagonal matrix T is defined to have the (v,v)th entry equal to $\tau_m(v)$.

The (u,v)th entry of the matrix $x^n(\mathrm{I} - x\mathrm{D}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^{-1}\mathrm{T}$ is equal to

$$\sum_w x^{|w|}\tau_m(w)\,,$$

where the sum is over all words w of length $\geq n$ with prefix u and suffix v.

Since we do not want to count words w that contain subwords from S, we need to delete the vertices $u_1, u_2, \dots, u_k$ from $B_n(\mathcal{A})$, so we use the adjacency matrix of the resulting matrix instead of M_n. Thus the (u,v)th entry in the matrix

$$x^n\left(\widetilde{\mathrm{I}}_S + \sum_{i\geq 1} x^i(\widetilde{\mathrm{I}}_S\mathrm{D}_n^{(m)}(\mathbf{z})\mathrm{M}_n\widetilde{\mathrm{I}}_S)^i\right)\mathrm{T} = x^n\widetilde{\mathrm{I}}_S(\mathrm{I} - x\mathrm{D}_n^{(m)}(\mathbf{z})\mathrm{M}_n\widetilde{\mathrm{I}}_S)^{-1}\mathrm{T}$$

is $\sum_w x^{|w|}\tau_m(w)$, where the sum is over all words w of length $\geq n$, with prefix u and suffix v, with the restriction that w does not contain any subwords from S. Hence

$$F_S(x,\mathbf{z}) = x^n \mathrm{trace}(\widetilde{\mathrm{I}}_S(\mathrm{I} - \mathrm{D}_n^{(m)}(x\mathbf{z})\mathrm{M}_n\widetilde{\mathrm{I}}_S)^{-1}\mathrm{T}\,\mathrm{E})\,,$$

where E is the matrix with all entries equal to 1.

Let $Q = \mathcal{A}^n - S = \{h_1, h_2, \dots, h_l\}$, where $l = q^n - k$ and h_i are ordered in increasing lexicographic order. We use A to denote the matrix obtained by deleting the S rows and columns from the matrix $\widetilde{\mathrm{P}}^{-1}$. By T_A we denote the matrix obtained by deleting the S rows and columns from the matrix T. Then $\widetilde{\mathrm{I}}_S(\mathrm{I}-\mathrm{D}_n^{(m)}(x\mathbf{z})\mathrm{M}_n\widetilde{\mathrm{I}}_S)^{-1}$ is the $q^n \times q^n$ matrix that has zeros in the S columns and rows and submatrix A^{-1} in the Q rows and columns. Thus

$$F_S(x,\mathbf{z}) = x^n \mathrm{trace}(\mathrm{A}^{-1}\mathrm{T}_\mathrm{A}\mathrm{E})\,,$$

where E is the $(q^n - k) \times (q^n - k)$ matrix with all entries equal to 1.

Let σ denote the sum of all entries of the matrix $\mathrm{A}^{-1}\mathrm{T}_\mathrm{A}$. Then $F_S(x,\mathbf{z}) = x^n\sigma$. The rows and columns of A and T_A are indexed by $Q = \{h_1, h_2, \dots, h_l\}$.

By Laplace's Extension Theorem

$$\sigma = \frac{1}{\det \mathrm{A}} \sum_{v\in Q} \tau_m(v) \sum_{w\in Q} \mathrm{K}_{v,w}\,,$$

where $\mathrm{K}_{v,w}$ is the cofactor of the (v,w)th entry of A. For a word $u \in \mathcal{A}^n$, let π_u denote the number of words in S that are less than u. Notice that, $(-1)^{\pi_v+\pi_w}\mathrm{K}_{v,w}$ is the cofactor of the $(k+1)\times(k+1)$ submatrix $\mathrm{M}_{v,w}$ of $\widetilde{\mathrm{P}}$ with row set $S \cup \{w\}$ and column set $S \cup \{v\}$ (the cofactor of a submatrix X of a matrix Y is $(-1)^\mu \det \mathrm{Z}$, where matrix Z is obtained by deleting the rows and columns of X from Y, and μ is the sum of the indices of rows and columns of X). Thus

$$\mathrm{K}_{v,w} = \frac{(-1)^{\pi_v+\pi_w}\det \mathrm{M}_{v,w}}{\det \widetilde{\mathrm{P}}}\,,$$

and

$$\sigma = \frac{1}{\det \mathrm{A} \det \widetilde{\mathrm{P}}} \sum_{v\in Q}(-1)^{\pi_v}\tau_m(v) \sum_{w\in Q}(-1)^{\pi_w}\det \mathrm{M}_{v,w}\,.$$

Since $\sum_{w\in Q}(-1)^{\pi_w}\det \mathrm{M}_{v,w} = \sum_{w\in\mathcal{A}^n}\det \mathrm{M}'_{v,w}$, where

$$\mathrm{M}'_{v,w} = \begin{pmatrix} \widetilde{p}_{w,v} & \widetilde{p}_{w,u_1} & \widetilde{p}_{w,u_2} & \cdots & \widetilde{p}_{w,u_k} \\ \widetilde{p}_{u_1,v} & \widetilde{p}_{u_1,u_1} & \widetilde{p}_{u_1,u_2} & \cdots & \widetilde{p}_{u_1,u_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widetilde{p}_{u_k,v} & \widetilde{p}_{u_k,u_1} & \widetilde{p}_{u_k,u_2} & \cdots & \widetilde{p}_{u_k,u_k} \end{pmatrix},$$

then $\sum_{w \in \mathcal{A}^n} \det \mathrm{M}'_{v,w} = \mathrm{M}_v$, where

$$M_v = \begin{pmatrix} K_v & K_{u_1} & K_{u_2} & \dots & K_{u_k} \\ \widetilde{p}_{u_1,v} & \widetilde{p}_{u_1,u_1} & \widetilde{p}_{u_1,u_2} & \cdots & \widetilde{p}_{u_1,u_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widetilde{p}_{u_k,v} & \widetilde{p}_{u_k,u_1} & \widetilde{p}_{u_k,u_2} & \cdots & \widetilde{p}_{u_k,u_k} \end{pmatrix},$$

and $\det \mathrm{A} \det \widetilde{\mathrm{P}} = \det \widetilde{\mathrm{P}}_S$. Thus

$$\sigma = \det \widetilde{\mathrm{P}}_S^{-1} \sum_{v \in Q} (-1)^{\pi_v} \mathbf{z}^{\tau_m(v)} \det \mathrm{M}_v = \det \widetilde{\mathrm{P}}_S^{-1} \det \mathrm{M}_S \,,$$

and (10) follows.

4 Proof of Theorem 2.1

We utilize some results from the literature of combinatorics on words. For a starting point to the theory of combinatorics on words, we refer the reader to (GO81a), (JS94), (JS05), (RD04), and (RS98); this is merely a sampling of the growing literature in this area. For a collection of recent results, see the three volumes edited by Lothaire, especially (Lot05).

Underlying much of the theory of combinatorics on words is a precise means of characterizing the extent to which a word overlaps with itself. For this purpose, for each word $w \in \mathcal{A}^m$, we recall from (1) that the autocorrelation polynomial of w is

$$S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \dots w_m) z^{m-i} \,,$$

where $\mathcal{P}(w)$ denotes the set of i's satisfying $w_1 \dots w_i = w_{m-i+1} \dots w_m$. In other words, for each $i \in \mathcal{P}(w)$, the prefix of w of length i is identical to the suffix of w of length i.

Now we define a useful language—and its associated generating function—frequently used in combinatorics on words. We write

$$\mathcal{R}_w = \{v \in \mathcal{A}^* \mid v \text{ contains exactly one occurrence of } w \text{, located at the right end}\} \,.$$

We write the generating function associated with this language as

$$R_w(z) = \sum_{v \in \mathcal{R}_w} \mathbf{P}(v) z^{|v|} \,.$$

It is well-known (see, for instance, (JS94), (JS05), (RD04), (RS98)) that this generating function can be expressed in terms of $S_w(z)$ as follows: If $w \in \mathcal{A}^m$, and if we define

$$D_w(z) = (1-z) S_w(z) + \mathbf{P}(w) z^m \,,$$

then we have

$$R_w(z) = \frac{\mathbf{P}(w) z^m}{D_w(z)} \,.$$

Next, we describe the generating functions associated with $\mathbb{E}[X_{n,k}]$ and $\mathbb{E}[\widehat{X}_{n,k}]$. Analogous, but more complicated, generating functions for the second moments of $X_{n,k}$ and $\widehat{X}_{n,k}$ can be established using a similar methodology, but more intricate word comparisons must be utilized.

The kth subword complexity of a word $w \in \mathcal{A}^n$ denotes the number of distinct subwords of length k (i.e., blocks of k contiguous letters) that appear in w. Our goal is to characterize the random variable $X_{n,k}$, defined as the kth subword complexity of a randomly selected word $w \in \mathcal{A}^n$.

We recall that $Y = Y_1 Y_2 Y_3 \dots$ denote a sequence of symbols drawn from $\mathcal{A}$. We assume that the Y_i's are chosen independently, with $\mathbf{P}\{Y_i = a_j\} = p_j$. We recall that $X_{n,k}$ denotes the number of distinct words from $\mathcal{A}^k$ which each appear as a subword of $Y_1 Y_2 \dots Y_{n+k-1}$.

We also recall that $Y^{(l)} = Y_1^{(l)} Y_2^{(l)} Y_3^{(l)} \ldots$ denote, for each l, a sequence of symbols drawn from $\mathcal{A}$. Again, we assume that all of the $Y_i^{(l)}$'s are chosen independently, with $\mathbf{P}\{Y_i^{(l)} = a_j\} = p_j$. We recall that $\widehat{X}_{n,k}$ denotes the number of distinct words in the collection $\{Y_1^{(l)} Y_2^{(l)} \ldots Y_k^{(l)} \mid 1 \le l \le n\}$.

Without loss of generality, we assumed that $0 < p_1 \le p_2 \le \cdots \le p_q < 1$. We recall that $p := p_q$ and $\delta = \sqrt{p}$. We choose $c > 0$ such that $p_1^{-c}\delta < 1$, and we have selected ϵ with $0 < \epsilon < c$. Finally, we defined $\mu = p_1^{-c}\delta$.

We define $G_k(z) = \sum_{n \ge 0} \mathbb{E}[X_{n,k}] z^n$ and $\widehat{G}_k(z) = \sum_{n \ge 0} \mathbb{E}[\widehat{X}_{n,k}] z^n$ as the ordinary generating functions for $\mathbb{E}[X_{n,k}]$ and $\mathbb{E}[\widehat{X}_{n,k}]$, respectively.

We observe that $w \in \mathcal{A}^k$ makes a contribution to $X_{n,k}$ if and only if Y begins with a word from $R_w\mathcal{A}^*$ of length $k+n-1$, which happens with probability

$$[z^{k+n-1}] \left(\frac{R_w(z)}{1-z} \right) .$$

It follows immediately that

$$\sum_{n \ge 0} \mathbb{E}[X_{n,k}] z^n = \sum_{w \in \mathcal{A}^k} \frac{R_w(z)}{(1-z)z^{k-1}} .$$

Since $R_w(z) = \mathbf{P}(w) z^k / D_w(z)$, it follows that

$$G_k(z) = \sum_{w \in \mathcal{A}^k} \frac{\mathbf{P}(w) z}{(1-z) D_w(z)} . \tag{11}$$

Next we observe that that $w \in \mathcal{A}^k$ makes a contribution to $\widehat{X}_{n,k}$ if and only if at least one $Y^{(l)}$ begins with w, which happens with probability

$$1 - (1 - \mathbf{P}(w))^n .$$

So

$$\mathbb{E}[\widehat{X}_{n,k}] = \sum_{w \in \mathcal{A}^k} \left(1 - (1 - \mathbf{P}(w))^n\right) . \tag{12}$$

Summing $\mathbb{E}[\widehat{X}_{n,k}] z^n$ over all $n \ge 0$, it follows that

$$\widehat{G}_k(z) = \sum_{w \in \mathcal{A}^k} \frac{\mathbf{P}(w) z}{(1-z)(1-(1-\mathbf{P}(w))z)} . \tag{13}$$

We have $\mathbf{P}\{Y_i = a_j\} = p_j$ and $\mathbf{P}\{Y_i^{(l)} = a_j\} = p_j$. Without loss of generality, we assumed that $0 < p_1 \le p_2 \le \cdots \le p_q < 1$. We observe that $p_q \le \sqrt{p_q} < 1$, so there exists $\rho > 1$ such that $\rho\sqrt{p_q} < 1$, and of course $\rho p_q < 1$ too. We recall that $\delta = \sqrt{p_q}$.

We recall from (1) the definition of the autocorrelation polynomial of a word w. The autocorrelation polynomial $S_w(z)$ records the extent to which w overlaps with itself. Of course, every word w has a trivial (complete) overlap with itself, which provides a contribution of "1" to $S_w(z)$. With high probability, we observe that the other overlaps of w with itself are very small, providing contributions to $S_w(z)$ of much smaller order. We formalize this notion with the following well-known lemma, which appears often throughout the literature of combinatorics on words (see, for instance, (JS05)). We use the Iverson notation $[\![A]\!] = 1$ if A holds, and $[\![A]\!] = 0$ otherwise.

Lemma 4.1 *Consider $\theta = (1 - p\rho)^{-1}$, $\delta = \sqrt{p}$, and $\rho > 1$ with $\rho\delta < 1$. When randomly selecting a binary word $w \in \mathcal{A}^k$, the autocorrelation polynomial $S_w(z)$ (at $z = \rho$) is approximately 1, with high probability. More specifically,*

$$\sum_{w \in \mathcal{A}^k} [\![\, |S_w(\rho) - 1| \le (\rho\delta)^k \theta \,]\!] \, \mathbf{P}(w) \ge 1 - \theta\delta^k .$$

Lemma 4.2 *Recall $\delta = \sqrt{p}$; also $\rho > 1$ is defined such that $\rho\delta < 1$. Consider the polynomial $D_w(z) = (1 - z)S_w(z) + \mathbf{P}(w) z^m$, where $S_w(z)$ denotes the autocorrelation polynomial of w (see (1)). There exists an integer K such that, for* every *word w with $|w| \ge K$, the polynomial $D_w(z)$ has exactly one root in the disk $|z| \le \rho$.*

Throughout the rest of the discussion below, we fix the "K" mentioned in the lemma above, and we consistently restrict attention to word lengths $k \geq K$.

For w with $|w| = k \geq K$, since $D_w(z)$ has a unique root in the disk $|z| \leq \rho$, we denote this root as A_w, and we write $B_w = D'_w(A_w)$. Using bootstrapping, we have

$$\begin{aligned} A_w &= 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O(\mathbf{P}(w)^2)\,, \\ B_w &= -S_w(1) + \left(k - \frac{2S'_w(1)}{S_w(1)}\right)\mathbf{P}(w) + O(\mathbf{P}(w)^2)\,. \end{aligned} \tag{14}$$

Next we compare $\sum_{n\geq 0}\mathbb{E}[X_{n,k}]z^n$ to $\sum_{n\geq 0}\mathbb{E}[\widehat{X}_{n,k}]z^n$.

We define

$$Q_k(z) = G_k(z) - \widehat{G}_k(z) = \sum_{n\geq 0}\left(\mathbb{E}[X_{n,k}] - \mathbb{E}[\widehat{X}_{n,k}]\right)z^n$$

and the contribution to $Q_k(z)$ from w as

$$Q^{(w)}(z) = \frac{\mathbf{P}(w)z}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-(1-\mathbf{P}(w))z}\right).$$

By (11) and (13), we know that

$$Q_k(z) = \sum_{w\in\mathcal{A}^k} Q^{(w)}(z)\,.$$

We also define $Q_{n,k} = [z^n]Q_k(z)$ and $Q_n^{(w)} = [z^n]Q^{(w)}(z)$. So $Q_{n,k}$ is exactly $\mathbb{E}[X_{n,k}] - \mathbb{E}[\widehat{X}_{n,k}]$, and $Q_n^{(w)}$ is the contribution to $Q_{n,k}$ from w. Our ultimate goal is to prove that $Q_{n,k}$ is asymptotically negligible, i.e., $\mathbb{E}[X_{n,k}]$ and $\mathbb{E}[\widehat{X}_{n,k}]$ have the same asymptotic growth.

Using Cauchy's Integral Formula, we have

$$Q_n^{(w)} = \frac{1}{2\pi i}\oint \frac{\mathbf{P}(w)z}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-(1-\mathbf{P}(w))z}\right)\frac{dz}{z^{n+1}}\,,$$

where the path of integration is a circle about the origin with counterclockwise orientation. Using a counterclockwise, circular path of radius ρ about the origin, we also define

$$I_n^{(w)}(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \frac{\mathbf{P}(w)z}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-(1-\mathbf{P}(w))z}\right)\frac{dz}{z^{n+1}}\,, \tag{15}$$

and by Cauchy's theorem, it follows that

$$\begin{aligned} Q_n^{(w)} = I_n^{(w)}(\rho) &- \operatorname*{Res}_{z=A_w}\frac{\mathbf{P}(w)z}{(1-z)D_w(z)z^{n+1}} + \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))}\frac{\mathbf{P}(w)z}{(1-z)(1-(1-\mathbf{P}(w))z)z^{n+1}} \\ &- \operatorname*{Res}_{z=1}\frac{\mathbf{P}(w)z}{(1-z)D_w(z)z^{n+1}} + \operatorname*{Res}_{z=1}\frac{\mathbf{P}(w)z}{(1-z)(1-(1-\mathbf{P}(w))z)z^{n+1}}\,. \end{aligned}$$

We compute the four relevant residues, namely

$$\begin{aligned} \operatorname*{Res}_{z=A_w}\frac{\mathbf{P}(w)z}{(1-z)D_w(z)z^{n+1}} &= \frac{\mathbf{P}(w)}{(1-A_w)B_wA_w^n}\,, \\ \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))}\frac{\mathbf{P}(w)z}{(1-z)(1-(1-\mathbf{P}(w))z)z^{n+1}} &= (1-\mathbf{P}(w))^n\,, \\ \operatorname*{Res}_{z=1}\frac{\mathbf{P}(w)z}{(1-z)D_w(z)z^{n+1}} &= -1\,, \\ \operatorname*{Res}_{z=1}\frac{\mathbf{P}(w)z}{(1-z)(1-(1-\mathbf{P}(w))z)z^{n+1}} &= -1\,. \end{aligned}$$

We define

$$f_w(x) = -\frac{\mathbf{P}(w)}{(1-A_w)B_w A_w^x} + (1-\mathbf{P}(w))^x .$$

We want to prove that $\sum_{w\in\mathcal{A}^k} f_w(x)$ is asymptotically small. We first observe that $\sum_{w\in\mathcal{A}^k} f_w(x)$ is absolutely convergent for all x. Then we define $\bar{f}_w(x) = f_w(x) - f_w(0)e^{-x}$. Next we utilize the Mellin transform of $\bar{f}_w(x)$. (See (FGD95) and (Szp01) for details about the Mellin transform.) Since $\bar{f}_w(x)$ is exponentially decreasing as $x \to +\infty$, and is $O(x)$ when $x \to 0$, then the Mellin transform of $\bar{f}_w(x)$, namely

$$\bar{f}_w^*(s) = \int_0^\infty \bar{f}_w(x) x^{s-1}\, dx ,$$

is well-defined for $\Re(s) > 1$. We have

$$\bar{f}_w^*(s) = -\frac{\mathbf{P}(w)}{B_w(1-A_w)} \int_0^\infty \left(\frac{1}{A_w^x} - 1\right) x^{s-1}\, dx + \int_0^\infty ((1-\mathbf{P}(w))^x - 1)x^{s-1}\, dx .$$

Using the well-known properties of the Mellin transform (see (FGD95) and (Szp01)), it follows that

$$\bar{f}_w^*(s) = -\frac{\mathbf{P}(w)}{B_w(1-A_w)}\Gamma(s)\left((\log A_w)^{-s} - 1\right) + \left(\left(\log\frac{1}{1-\mathbf{P}(w)}\right)^{-s} - 1\right)\Gamma(s) .$$

From the bootstrapped equations for A_w and B_w given in (14), it follows that

$$\bar{f}_w^*(s) = -\left(1 + O(|w|\mathbf{P}(w)^2)\right)\Gamma(s)\left(\left(\frac{\mathbf{P}(w)}{S_w(1)}\right)^{-s}(1 + O(\mathbf{P}(w))) - 1\right) + \left(\mathbf{P}(w)^{-s}(1 + O(\mathbf{P}(w))) - 1\right)\Gamma(s) ,$$

which simplifies to

$$\bar{f}_w^*(s) = \Gamma(s)\mathbf{P}(w)^{-s}\left(1 - \frac{1}{S_w(1)^{-s}}\right)(1 + O(\mathbf{P}(w))) .$$

Now we define $g^*(s) = \sum_{w\in\mathcal{A}^k} \bar{f}_w^*(s)$. We compute

$$\begin{aligned} g^*(s) &= \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)^{-s}\Gamma(s)\left(1 - \frac{1}{S_w(1)^{-s}}\right) O(1) \\ &= \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)^{-s-1}\Gamma(s)\left(\frac{\mathbf{P}(w)(S_w(1)^{-s} - 1)}{S_w(1)^{-s}}\right) O(1) \\ &= (\sup\{p_q^{-\Re(s)-1}, 1\})^k \delta^k \Gamma(s) O(1) , \end{aligned}$$

where the last equality follows from Lemma 4.1, which precisely describes the fact that the autocorrelation polynomial is close to 1 with very high probability. We note that when $s = 0$, the pole at $\Gamma(s)$ is canceled.

We note that there exists $c > 0$ such that $p_1^{-c}\delta < 1$. So $g^*(s)$ is analytic in $\Re(s) \in (-1, c)$. We choose $\epsilon > 0$ with the property that $0 < \epsilon < c$. Then we have

$$Q_{n,k} - I_n^{(w)}(\rho) = \frac{1}{2\pi i}\int_{\epsilon - i\infty}^{\epsilon + i\infty} g^*(s) n^{-s}\, ds + \sum_{w\in\mathcal{A}^k} f_w(0)e^{-x} .$$

The first term is $O(n^{-\epsilon})O((p_1^{-c}\delta)^k)$ since $g^*(s)$ is analytic in the strip $\Re(s) \in (-1, c)$. The second term is $O(e^{-x})$. Finally, Lemma 4.3 (given below) concerning $I_n^{(w)}(\rho)$ allows us to complete the proof of Theorem 2.1. In the statement of Theorem 2.1, we use $\mu = p_1^{-c}\delta < 1$.

Lemma 4.3 *Consider $\delta = \sqrt{p}$, and $\rho > 1$ with $\rho\delta < 1$. Recall from (15) that*

$$I_n^{(w)}(\rho) = \frac{1}{2\pi i}\int_{|z|=\rho} \frac{\mathbf{P}(w)z}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-(1-\mathbf{P}(w))z}\right)\frac{dz}{z^{n+1}} ,$$

where $D_w(z) = (1-z)S_w(z) + \mathbf{P}(w)z^k$ for $w \in \mathcal{A}^k$. The sum of $I_n^{(w)}(\rho)$ over all words $w \in \mathcal{A}^k$ is asymptotically negligible, namely

$$\sum_{w \in \mathcal{A}^k} I_n^{(w)}(\rho) = O(\rho^{-n})O((\rho\delta)^k)\,.$$

Proof: There exist constants C_1, C_2, and K_2 such that, for all $k \geq K_2$ and all $|z| = \rho$, we have $\frac{1}{|D_w(z)|} \leq C_1$ and $\frac{1}{|1-(1-\mathbf{P}(w))z|} \leq C_2$ for all w with $|w| = k$. The proof of this useful fact is straightforward. Thus

$$|I_n^{(w)}(\rho)| = \frac{2\pi\rho}{2\pi} \frac{\mathbf{P}(w)z}{1-z} \frac{\mathbf{P}(w)z + 1 - z - D_w(z)}{D_w(z)(1-(1-\mathbf{P}(w))z)} \frac{1}{\rho^{n+1}}\,.$$

We note that $|D_w(z) - (1-z)| \leq |1-z||S_w(z) - 1| + |z|^k\mathbf{P}(w) \leq (1+\rho)(S_w(\rho) - 1) + (p_q\rho)^k$. Finally, using Lemma 4.1, which formalizes the notion that the autocorrelation polynomial is close to 1 with high probability, the result follows.

□

5 Auxiliary results.

In this section we show the connection between different types of correlation polynomial matrices, the adjacency matrix of the de Bruijn graph and subword complexity related generating functions.

Recall the definition of the matrix C_S given by Eq. 2. We are particularly interested in the case $S = \mathcal{A}^k$. The matrix in this case is $\mathrm{C}_{\mathcal{A}^k}(z)$, namely, the $q^k \times q^k$ matrix whose rows and columns are indexed by the words of length k over $\mathcal{A}$ arranged in increasing lexicographic order.

Example 5.1 *Consider the case $q = 2$ and $k = 3$. In this case, the unweighted correlation polynomial matrix of all words of length 3 over $\mathcal{A} = \{a_1, a_2\}$ is*

$$\mathrm{C}_{\mathcal{A}^3}(z) = \begin{pmatrix} z^2+z+1 & z^2+z & z^2 & z^2 & 0 & 0 & 0 & 0 \\ 0 & 1 & z & z & z^2 & z^2 & z^2 & z^2 \\ z^2 & z^2 & z^2+1 & z^2 & z & z & 0 & 0 \\ 0 & 0 & 0 & 1 & z^2 & z^2 & z^2+z & z^2+z \\ z^2+z & z^2+z & z^2 & z^2 & 1 & 0 & 0 & 0 \\ 0 & 0 & z & z & z^2 & z^2+1 & z^2 & z^2 \\ z^2 & z^2 & z^2 & z^2 & z & z & 1 & 0 \\ 0 & 0 & 0 & 0 & z^2 & z^2 & z^2+z & z^2+z+1 \end{pmatrix}.$$

The matrix $\mathrm{C}_{\mathcal{A}^k}(z)$ turns out to be very important because the matrices $\mathrm{C}_S(z)$ for $S \subseteq \mathcal{A}^k$, that are used in Proposition 1.3, are principal submatrices of $\mathrm{C}_{\mathcal{A}^k}(z)$. We will derive a simple formula for $\mathrm{C}_{\mathcal{A}^k}(z)$ in terms of the adjacency matrix of the de Bruijn graph.

Theorem 5.2 *Let I denote the $q^n \times q^n$ identity matrix, and let E denote the $q^n \times q^n$ matrix of all 1s. Let M_n denote the adjacency matrix of the de Bruijn graph $B_n(\mathcal{A})$. The unweighted correlation polynomial matrix $\mathrm{C}_{\mathcal{A}^n}(z)$ of all words of length n over $\mathcal{A}$ is*

$$\mathrm{C}_{\mathcal{A}^n}(z) = (\mathrm{I} - z^n\mathrm{E})(\mathrm{I} - z\mathrm{M}_n)^{-1}\,.$$

Proof: Consider two words $w, u \in \mathcal{A}^n$. The (w,u)th entry of $(\mathrm{M}_n)^i$ is the number of paths of length i in $B_n(\mathcal{A})$ that start in w and end in u. Equivalently, this is exactly the number of words in $\mathcal{A}^{n+i}$ with prefix w and suffix u. For $0 \leq i \leq n-1$ the (w,u)th entry of $(\mathrm{M}_n)^i$ is 1 if the suffix of w of length $n-i$ coincides with the prefix of u of length $n-i$; the (w,u)th entry is 0 otherwise. So, in either case, the (w,u)th entry of $(\mathrm{M}_n)^i$ is exactly the coefficient of z^i in $\mathrm{C}_{w,u}(z)$, for all $0 \leq i \leq n-1$.

For larger values of i, we note that the entries of $\mathrm{C}_{\mathcal{A}^n}(z)$ are each unweighted correlation polynomials. The degree of an unweighted correlation polynomial $\mathrm{C}_{w,u}(z)$ is at most $n-1$ for words $w, u \in \mathcal{A}^n$, since words of length n have overlaps of length at most $n-1$. So each entry of $\mathrm{C}_{\mathcal{A}^n}(z)$ has degree at most $n-1$.

Thus

$$\mathrm{C}_{\mathcal{A}^n}(z) = \mathrm{I} + z\mathrm{M}_n + \cdots + z^{n-1}(\mathrm{M}_n)^{n-1}\,,$$

which simplifies to

$$\mathrm{C}_{\mathcal{A}^n}(z) = (\mathrm{I} - z^n(\mathrm{M}_n)^n)(\mathrm{I} - z\mathrm{M}_n)^{-1}\,. \tag{16}$$

For every pair of words $w, u \in \mathcal{A}^n$, there is exactly one word of length $2n$ which starts with w and ends with u (namely, wu). Thus, each entry of $(\mathrm{M}_n)^n$ is simply "1". So $(\mathrm{M}_n)^n = \mathrm{E}$. From (16), we conclude $\mathrm{C}_{\mathcal{A}^n}(z) = (\mathrm{I} - z^n\mathrm{E})(\mathrm{I} - z\mathrm{M}_n)^{-1}$.

□

Corollary 5.3 *The unweighted correlation polynomial matrix $\mathrm{C}_{\mathcal{A}^n}(z)$ has eigenvalue $\lambda_1 = 1$ with multiplicity $q^n - 1$ and also eigenvalue $\lambda_2 = (1 - q^n z^n)/(1 - qz)$ with multiplicity 1.*

Proof: We recall that E denotes the $q^n \times q^n$ matrix of all 1s. So E has eigenvalue 0 with multiplicity $q^n - 1$ and also eigenvalue q^n with multiplicity 1. Since $(\mathrm{M}_n)^n = \mathrm{E}$, then M_n has eigenvalue 0 with multiplicity $q^n - 1$ and also eigenvalue q with multiplicity 1. It follows from Theorem 5.2 that $\mathrm{C}_{\mathcal{A}^n}(z)$ has eigenvalue $(1 - q^n z^n)/(1 - qz)$ with multiplicity 1 and also eigenvalue 1 with multiplicity $q^n - 1$.

□

Theorem 5.4 *Let $\mathbf{z} = (z_0, z_1, \ldots, z_{q^m-1})$ and $\mathbf{k} = (k_0, k_1, \ldots, k_{q^m-1})$. Let $\mathbf{z}^{\mathbf{k}}$ denote the monomial $z_0^{k_0} z_1^{k_1} \ldots z_{q^m-1}^{k_{q^m-1}}$. Consider the generating function*

$$G_n(x, \mathbf{z}) = \sum_{\mathbf{k} \in (\mathbb{Z}_{\geq 0})^{q^n}} \sum_{i \geq n} c(i, \mathbf{k}) x^i \mathbf{z}^{\mathbf{k}}\,,$$

where $c(i, \mathbf{k})$ is the number of words $w \in \mathcal{A}^i$ of type $\tau_n(w) = \mathbf{z}^{\mathbf{k}}$. Then

$$G_n(x, \mathbf{z}) = (1/q)x^{n-1}\mathrm{trace}((\mathrm{I} - x\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^{-1}\mathrm{E}) - (qx)^{n-1}\,,$$

where $\mathrm{D}_n^{(n)}$ is defined by Eq. 5, M_n is the adjacency matrix of the de Bruijn graph $B_n(\mathcal{A})$, and E is the $q^n \times q^n$ matrix of all 1s. .

Proof: For $u, v \in \mathcal{A}^n$, and for $i \geq 0$, the (u, v)th entry of $(\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^i$ is $\sum_w \tau_n(w_1 \ldots w_{n+i-1})$, where the sum is taken over all words $w \in \mathcal{A}^{n+i}$ with prefix u and suffix v. Therefore the number of words $w \in \mathcal{A}^{n+i-1}$ of type $\tau_n(w) = \mathbf{z}^{\mathbf{k}}$ is the coefficient of $\mathbf{z}^{\mathbf{k}}$ in the sum of all entries of $(1/q)(\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^i$, or equivalently, the coefficient of $x^{n+i-1}\mathbf{z}^{\mathbf{k}}$ in $(1/q)\mathrm{trace}(x^{n-1}(x\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^i\mathrm{E})$. Thus

$$\begin{aligned}
G_n(x, \mathbf{z}) &= (1/q)x^{n-1} \sum_{i=1}^{\infty} \mathrm{trace}((x\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^i\mathrm{E}) \\
&= (1/q)x^{n-1}\mathrm{trace}((\mathrm{I} - x\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^{-1}\mathrm{E}) - (1/q)x^{n-1}\mathrm{trace}(E) \\
&= (1/q)x^{n-1}\mathrm{trace}((\mathrm{I} - x\mathrm{D}_n^{(n)}(\mathbf{z})\mathrm{M}_n)^{-1}\mathrm{E}) - (qx)^{n-1}\,.
\end{aligned}$$

□

Corollary 5.5 *Let $\mathbf{z} = (z_0, z_1, \ldots, z_{q-1})$. Consider*

$$G_1(x, \mathbf{z}) = \sum_{\mathbf{k} \in (\mathbb{Z}_{\geq 0})^q} \sum_{i \geq 1} c(i, \mathbf{k}) x^i \mathbf{z}^{\mathbf{k}}\,,$$

where $c(i, \mathbf{k})$ is the number of words $w \in \mathcal{A}^i$ of type $\tau_1(w) = \mathbf{z}^{\mathbf{k}}$. In other words, $c(i, \mathbf{k})$ denotes the number of words $w \in \mathcal{A}^i$ such that, for each i, the ith letter a_i of $\mathcal{A}$ occurs exactly k_i times in w. Then using $n = 1$ in Theorem 5.4 yields

$$G_1(x, \mathbf{z}) = \frac{1}{1 - x\sum_i z_i} - 1\,.$$

Remark 5.6 *Let $\nu(\mathbf{k})$ denote the number of non-zero k_i's in $\mathbf{k} = (k_0, k_1, ..., k_{q^n-1})$. Then the number of words in $\mathcal{A}^i$ with nth subword complexity $f_w(n) = j$ is exactly the coefficient of $\sum_{\nu(\mathbf{k})=j} x^i\mathbf{z}^{\mathbf{k}}$ in $G_n(x, \mathbf{z})$.*

Let $m \geq 1$ and $n \geq 2m-2$. The generalized correlation polynomial matrix (on q^m variables $\mathbf{z} = (z_0, \ldots, z_{q^m-1})$) of all words of $\mathcal{A}^n$ is the $q^n \times q^n$ matrix $\mathrm{C}_n^{(m)}(\mathbf{z})$, whose rows and columns are indexed by the words of $\mathcal{A}^n$ arranged in increasing lexicographic order, with the (u, v)th entry defined to be $C_{v,w}^{(m)}(\mathbf{z})$.

Considering the relative simplicity of the adjacency matrix of the de Bruijn graph, we give a method of computing all the generalized correlation polynomials of words of length n.

Lemma 5.7 *Let $m \geq 1$ and $n \geq 2m-2$. Let $\mathbf{z} = (z_0, z_1, \ldots, z_{q^m-1})$ and $\mathbf{k} = (k_0, k_1, \ldots, k_{q^m-1})$. Consider the $q^n \times q^n$ matrices $\mathrm{H}_n^{(m)}(\mathbf{z})$ and $\mathrm{T}_n^{(m)}(\mathbf{z})$ with rows and columns indexed by the words of $\mathcal{A}^n$ in increasing lexicographic order. Define the (u, v)th entry of $\mathrm{T}_n^{(m)}(\mathbf{z})$ as $\tau_m(w)$, where w is the suffix of length $n + 2m - 2$ of uv. The diagonal matrix $\mathrm{H}_n^{(m)}$ is defined by*

$$\mathrm{H}_n^{(m)}(\mathbf{z}) = \mathrm{diag}(z_0, \ldots, z_0, \ldots, z_{q^m-1}, \ldots, z_{q^m-1}, \ldots, z_0, \ldots, z_0, \ldots, z_{q^m-1}, \ldots, z_{q^m-1})\,;$$

and consists of q^{n-2m+1} blocks $z_0, \ldots, z_0, \ldots, z_{q^m-1}, \ldots, z_{q^m-1}$ that appear repeatedly along the diagonal, and each z_i occurs in such a block q^{m-1} times consecutively. Recall that M_n denotes the adjacency matrix of the de Bruijn graph $B_n(\mathcal{A})$.

The generalized correlation polynomial matrix (on q^m variables $\mathbf{z} = (z_0, \ldots, z_{q^m-1})$) of all words of $\mathcal{A}^n$ is

$$\mathrm{C}_n^{(m)}(\mathbf{z}) = (\mathrm{I} - \mathrm{T}_n^{(m)}(\mathbf{z}))(\mathrm{I} - \mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^{-1}\,.$$

Proof: The proof of this lemma is similar to the proof of Lemma 5.2. The (u, v)th entry of $(\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^i$ is $\sum_w \tau_m(w')$, where the sum is taken over all words $w = w_1 w_2 \ldots w_{n+i}$ with prefix u and suffix v; here, $w' = w_{n-2m+3} \ldots w_{n-m+i+1}$. For $0 \leq i \leq n-1$, if the suffix of length $n-i$ of u coincides with the prefix of length $n-i$ of v, then the (u, v)th entry of $(\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^i$ is $\tau_m(w_{n-2m+3} \ldots w_{n-m+i+1})$, where $w = w_1 w_2 \ldots w_{n+i}$ has prefix u and suffix v; otherwise, the (u, v)th entry of $(\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^i$ is 0. Therefore, the sum of the (u, v)th entries of $(\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^i$ for $0 \leq i \leq n-1$ is $C_{u,v}^{(m)}(\mathbf{z})$. So we obtain

$$\mathrm{C}_n^{(m)}(\mathbf{z}) = \mathrm{I} + \mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n + \cdots + (\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^{n-1} = (\mathrm{I} - (\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^n)(\mathrm{I} - \mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^{-1}\,.$$

We observe that $\mathrm{T}_n^{(m)}(\mathbf{z})$ is $(\mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^n$. Thus

$$\mathrm{C}_n^{(m)}(\mathbf{z}) = (\mathrm{I} - \mathrm{T}_n^{(m)}(\mathbf{z}))(\mathrm{I} - \mathrm{H}_n^{(m)}(\mathbf{z})\mathrm{M}_n)^{-1}\,.$$

□

Acknowledgements

The authors are thankful for insightful comments from the referees.

A part of this work was done during I. Gheorghiciuc's stay at Institut des Hautes Études Scientifiques. She would like to thank this institution for excellent working conditions.

For readers who are interested in contacting the authors, please notice that both authors are headed to new institutions, effective fall 2007. I. Gheorghiciuc will be in the Department of Mathematics at Carnegie Mellon University. M. D. Ward will be in the Department of Statistics at Purdue University.

References

[AC00] A. de Luca A. Colosimo. Special factors in biological strings. *J. Theoretical Biology*, 204, no. 1:29–46(18), May 2000.

[All94] J.-P. Allouche. Sur la complexité des suites infinies. *Bull. Belg. Math. Soc.*, 1:133–143, 1994.

[BK93] E. Bender and F. Kochman. The distribution of subword counts is usually normal. *European Journal of Combinatorics*, 14:265–275, 1993.

[dL99] A. de Luca. On the combinatorics of finite words. *Theoretical Computer Science*, 218:13–39, 1999.

[Fay04] J. Fayolle. An average-case analysis of basic parameters of the suffix tree. In M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, editors, *Mathematics and Computer Science*, pages 217–227, Vienna, Austria, 2004. Birkhäuser.

[Fer99] S. Ferenczi. Complexity of sequences and dynamical systems. *Discr. Math.*, 206:145–154, 1999.

[FGD95] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[GJ79] I. P. Goulden and D. M. Jackson. An inversion theorem for cluster decomposition of sequences with distinguished subsequences. *J. London Math. Soc.*, 20:567–576, 1979.

[GO78] L. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM J. Appl. Math.*, 35:401–418, 1978.

[GO81a] L. Guibas and A. M. Odlyzko. Periods in strings. *J. Combinatorial Theory*, 30A:19–42, 1981.

[GO81b] L. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Combinatorial Theory*, 30A:183–208, 1981.

[HHI00] V. Halava, T. Harju, and L. Ilie. Periods and binary words. *J. Combin. Theory*, 89A:298–303, 2000.

[JLS04] S. Janson, S. Lonardi, and W. Szpankowski. On average sequence complexity. *Theoret. Comput. Sci.*, 326:213–227, 2004.

[JS94] P. Jacquet and W. Szpankowski. Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory*, A66:237–269, 1994.

[JS05] P. Jacquet and W. Szpankowski. *Applied Combinatorics on Words*, chapter 7, Analytic Approach to Pattern Matching. Cambridge, 2005. See (Lot05).

[Kon05] Y. Kong. Extension of Goulden-Jackson cluster method on pattern occurrences in random sequences and comparison with Régnier-Szpankowski method. *Journal of Difference Equations and Applications*, 11:1265–1271, 2005.

[Li80] S.-Y. R. Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Probab.*, 8:1171–1176, 1980.

[Lot05] M. Lothaire. *Applied Combinatorics on Words*. Cambridge, 2005.

[Nic03] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fundamenta Informaticae*, 56:71–88, 2003.

[NZ99] J. Noonan and D. Zeilberger. The Goulden-Jackson cluster method: extensions, applications, and implementations. *J. Difference Eq. Appl.*, 5:355–377, 1999.

[PHNS06] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. Submitted for publication, 2006.

[PN02] P. Flajolet P. Nicodème, B. Salvy. Motif statistics. *Theoretical Computer Science*, 287:593–617, 2002.

[RD04] M. Régnier and A. Denise. Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science*, 6:191–214, 2004.

[RR03] S. Rahmann and E. Rivals. On the distribution of the number of missing words in random texts. *Combinatorics, Probability and Computing*, 12:73–87, 2003.

[RS98] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22:631–649, 1998.

[Sol66] A. D. Solov'ev. A combinatorial identity and its application to the problem concerning the first occurrence of a rare event. *Theory Prob. Appl.*, 11:276–282, 1966.

[Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

[TAK$^+$02] O. G. Troyanskaya, O. Arbell, Y. Koren, G. M. Landau, and A. Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18:679–688, 2002.

[War07] M. D. Ward. The average profile of suffix trees. In *The Fourth Workshop on Analytic Algorithmics and Combinatorics*, pages 183–193, New Orleans, 2007.