Estimating Statistics on Words Using Ambiguous Descriptions

Cyril Nicaud

Université Paris-Est, LIGM (UMR 8049), F77454 Marne-la-Vallée, France
cyril.nicaud@u-pem.fr

Abstract

In this article we propose an alternative way to prove some recent results on statistics on words, such as the expected number of runs or the expected sum of the run exponents. Our approach consists in designing a general framework, based on the symbolic method developed in analytic combinatorics. The descriptions obtained in this framework are built in such a way that the degree of ambiguity of an object O (i.e., the number of different descriptions corresponding to O) is exactly the value of the statistic under study for O. The asymptotic estimation of the expectation is then done using classical techniques from analytic combinatorics. To show the generality of our method, we not only apply it to obtain new proofs of known results, but also extend them from the uniform distribution to any memoryless distribution.

1998 ACM Subject Classification G.2.1 Combinatorics.

Keywords and phrases random words, runs, symbolic method, analytic combinatorics.

Digital Object Identifier 10.4230/LIPIcs.CPM.2016.9

1 Introduction

In this article we propose an alternative way to prove some recent results on statistics on words, such as the expected number of runs or the expected sum of the run exponents. Studying statistics on words is a classical topic in discrete probabilities, which has many fundamental applications in computer science, for instance in the fields of bioinformatics, information theory and average case analysis of algorithms.

We specially focus on statistics related to the runs in a random word (see Section 2.1 for the definition). Bounding the maximal number of runs in a word is a fundamental question in combinatorics of words, with consequences in text algorithms. Kolpakov and Kucherov proved that it is in $\mathcal{O}(n)$ in their seminal paper [12], and they conjectured that it is at most n. Banai and his coauthors proved this conjecture very recently [1]. Several other statistics, such as the total run length or the sum of exponents, have also been studied in the literature. Besides tightening lower and upper bounds in the worst case [4, 5, 8, 14, 16, 17, 18, 1], works have been done on the expected values of those statistics, for uniform distributions on words [15, 13, 11, 3]. It is the kind of questions we propose to study in this article.

Our main contribution is to provide a general framework, which proves quite useful to obtain asymptotic equivalents to the expectations of statistics related to runs. We follow and adapt the main ideas developed in the field of *analytic combinatorics* (see the textbook of Flajolet and Sedgewick [6]): First we explain how to build the formal power series $L_\chi(z)$ that corresponds to the statistic χ directly from a combinatorial specification on sets of words. Then, we use the techniques of complex analysis to estimate the expectation $\mathbb{E}_n[\chi]$ of χ for uniform random words of length n. The main difference with the classical framework is that the combinatorial specifications we use are *ambiguous*. Usually, unambiguity is mandatory for this combinatorial method to apply. However, if the degree of ambiguity of the specification

27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016).
Editors: Roberto Grossi and Moshe Lewenstein; Article No. 9; pp. 9:1–9:12

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

for a word w, i.e. the number of ways to produce w, is exactly $\chi(w)$, then we can directly get an expression of $L_\chi(z)$, or an equation it satisfies.

The net gain of this method is that once $L_\chi(z)$ is known, no tedious computations are needed to get the asymptotic equivalent of $\mathbb{E}_n[\chi]$. The tools from analytic combinatorics apply and directly yield the result. Moreover, this framework can be used to go beyond uniform distributions, since it can easily be extended to memoryless distributions, where each letter is chosen independently with some fixed probability on the alphabet.

The technique we propose is quite natural, and there are hints of its use, for instance, in [6, A.7.] and also in the study of hidden words [7]. However, it lacks a general framework, which is what we propose and illustrate in this article. This introduction is continued in Section 3, where we present the method on three basic examples, after the required notations given in Section 2. This is done in an informal way, but it should gives a fair picture of our method. The formalism of weighted sets is then introduced in Section 4. In Section 5, we propose alternative proofs to some results of the literature. Finally, we explain in Section 6 how to generalize them to memoryless distributions.

2 Preliminaries

For any two nonnegative integers i, j, let $[i, j]$ denote the integer interval $\{i, \dots, j\}$. By convention, $[i, j] = \emptyset$ if $j < i$. Let also $[i]$ denote the integer interval $[1, i]$.

The *mobius function* $\mu : \mathbb{Z}_{\geq 1} \to \{-1, 0, 1\}$ is defined as follows. If $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ is the decomposition of a positive n into prime numbers, then $\mu(n) = (-1)^k$ if all the α_i's are equal to 1, and $\mu(n) = 0$ otherwise. The main property of this function is that f and g are two functions from $\mathbb{Z}_{\geq 1}$ such that $f(n) = \sum_{d|n} g(d)$, then $g(n) = \sum_{d|n} \mu\left(\frac{n}{d}\right) f(d)$, where $d|n$ means that d ranges over the divisors of n.

2.1 Words and Probabilities on Words

In the sequel we consider words on a finite alphabet A, of cardinality $\ell \geq 2$. We assume the reader is familiar with the classical definitions on words, such as prefixes, suffixes, factors, subwords ... For $w \in A^*$ of length n and $i \in [n]$, let w_i (or $w[i]$) denote the i-th letter of w, with the convention that positions start at 1. The last letter of w is therefore $w_{|w|}$. Let also $w[i, j] = w_i \cdots w_j$ denote the factor of w that starts at position i and ends at position j.

Recall that a word w is *not primitive* when there exists a word v and an integer $k \geq 2$ such that $w = v^k$, and that it is *primitive* otherwise. Let $\mathcal{P}$ denote the set of all primitive words. A word w of length n is *periodic with period* $p \geq 1$ when $w[i] = w[i + p]$, for every $i \in [n - p]$. The *period* of a word is its smallest period. If w is periodic with period p, then its *exponent* is $\frac{|w|}{p}$. The exponent is not necessarily an integer: for instance the exponent of *ababa* is 5/2. A *run of period* p in a word w is a factor $w[i, j]$ of w with least period p, such that $p \geq 2$ and $w[i - 1, j]$ and $w[i, j + 1]$, when they exist, are not of period p (the factor $w[i, j]$ is "maximal" for the period p). We identify such a run by the triplet (i, j, p). Let $\textsc{runs}(v)$ denote the set of all runs in the word v.

The *uniform distribution* on a finite set E is the probability p defined for all $e \in E$ by $p(e) = \frac{1}{|E|}$. By a slight abuse of notation, we will speak of the *uniform distribution on* A^* to denote the sequence $(p_n)_{n\geq 0}$ of uniform distributions on A^n. For instance, if $A = \{a, b, c\}$, then each element of A^n has probability 3^{-n} under this distribution.

Another very classical distribution on A^n is the *memoryless distribution of probability* p, where p is a probability on the alphabet A. Under this distribution, the probability of a word

$w = w_1 \cdots w_n \in A^n$ is $\mathbb{P}_p(w) = p(w_1) \cdots p(w_n)$. This distribution consists in generating each letter of the word independently, following p.

2.2 Elements of Analytic Combinatorics

We only present the parts of this well-established theory that will be needed in the sequel. For more information, the reader is referred to the book of Flajolet and Sedgewick [6].

A set $\mathcal{E}$ with a size function $s : \mathcal{E} \to \mathbb{N}$ is a *combinatorial set* if for every $n \in \mathbb{N}$, $\mathcal{E}_n := s^{-1}(n)$ is finite. The *generating series* $E(z)$ of $\mathcal{E}$ is defined by $E(z) := \sum_{e \in \mathcal{E}} z^{s(e)} = \sum_{n \geq 0} e_n z^n$, with $e_n = |\mathcal{E}_n|$. We will also use the notation $[z^n]E(z) := e_n$ to denote the n-th coefficient $E(z)$. If $\mathcal{E}$ and $\mathcal{F}$ are two combinatorial sets of size functions s and t, $\mathcal{E} \times \mathcal{F}$ is also a combinatorial set for the size function $r((e, f)) = s(e) + t(f)$, for every $e \in \mathcal{E}$ and $f \in \mathcal{F}$. This construction extends naturally to $\mathcal{E}_1 \times \cdots \times \mathcal{E}_k$ and to $\mathcal{E}^k$, for every $k \geq 2$.

The *symbolic method* consists in a dictionary to directly translate unambiguous combinatorial specifications into equations on generating series. In particular:

▶ **Theorem 1** ([6]). *For $\mathcal{E}$ and $\mathcal{F}$ two combinatorial sets of generating series $E(z)$ and $F(z)$:*

- *If $\mathcal{E}$ and $\mathcal{F}$ are two disjoint sets, then $\mathcal{G} = \mathcal{E} \dot\cup \mathcal{F}$ implies that $G(z) = E(z) + F(z)$.*
- *If $\mathcal{G} = \mathcal{E} \times \mathcal{F}$, then $G(z) = E(z)F(z)$.*
- *If $\mathcal{E}_0 = \emptyset$ and $\mathcal{G} = \mathcal{E}^* := \cup_{k \geq 0} \mathcal{E}^k$, then $G(z) = \frac{1}{1-E(z)}$.*

There are other basic constructions, but we will not need them in this article. However, there is a more advanced tool that is particularly useful for us: If $\mathcal{E}_0 = \emptyset$, a tuple of elements of $\mathcal{E}$ is *primitive* when, it is primitive as a word on the alphabet $\mathcal{E}$. From [6, A.4] we get that if $\mathcal{F}$ is the set of primitive tuples of elements of $\mathcal{E}$, then

$$F(z) = \sum_{k \geq 1} \frac{\mu(k)\, E(z^k)}{1 - E(z^k)}. \tag{1}$$

As an illustration, observe that the generating series of the alphabet is $A(z) = \ell z$, as there are ℓ letters, each of size 1. Since a word is a tuple of letters, the generating series of all words[1] is $\frac{1}{1-A(z)} = \frac{1}{1-\ell z}$. Moreover, by Equation 1, the generating series $P(z)$ of the set $\mathcal{P}$ of primitive words on A is

$$P(z) = \sum_{k \geq 1} \frac{\mu(k)\, \ell z^k}{1 - \ell z^k}. \tag{2}$$

The second part of the theory consists in considering generating series as analytic functions from $\mathbb{C}$ to $\mathbb{C}$, and then in using the powerful techniques of this field. We referred the reader to [6] for the classical definitions of the theory of analytic functions. In the sequel, we will only use the following theorem, which is a simplified version of the classical Transfer Theorem [6, p.393]. The full version is much more powerful, but it requires some analytic conditions that are too long to introduce for this extended abstract.

▶ **Theorem 2** (Simplified Transfer Theorem [6]). *Let r be a positive real number. Let f be a function from $\mathbb{C}$ to $\mathbb{C}$, which is analytic at 0, with radius of convergence greater than r. For any $k \in \mathbb{Z}_{\geq 1}$, we have the following asymptotic equivalent as n tends to infinity,*

$$[z^n] \frac{f(z)}{(1 - z/r)^k} \sim \frac{f(r)\, n^{k-1}}{(k-1)!\, r^n}.$$

[1] This elementary result can of course be obtained directly.

We will also use Theorem 2 the following way in the sequel: if $f_1, \ldots, f_k$ are analytic at 0 and of radius of convergence greater than r, then applying the theorem to each term yields

$$[z^n]\left(\frac{f_1(z)}{1-z/r} + \frac{f_2(z)}{(1-z/r)^2} + \ldots + \frac{f_k(z)}{(1-z/r)^k}\right) \sim \frac{f_k(r)\, n^{k-1}}{(k-1)!\, r^n}, \tag{3}$$

since the other terms are negligible when n tends to infinity.

Extracting the n-th coefficient of Equation (2) yields the well known fact that if P_n denote the number of primitive words, then $P_n = \sum_{d|n} \ell^{n/d} \sim \ell^n$. Hence, $P(z)$ is analytic at 0 and its radius of convergence is $1/\ell$. This simple fact will be quite useful in the sequel.

If χ is a statistic on a combinatorial set $\mathcal{E}$, i.e. a mapping from $\mathcal{E}$ to $\mathbb{R}$, the *cumulative generating series of* χ is the formal power series $L_\chi(z) = \sum_{e\in\mathcal{E}} \chi(e) z^{|e|}$. Observe that the expectation of χ for uniform random elements of $\mathcal{E}_n$ is given by $\mathbb{E}_n[\chi] = [z^n]L_\chi(z)/[z^n]E(z)$. Since we focus on statistics on words in this article, we will always have $[z^n]E(z) = \ell^n$, the number of words of length n, except in Section 6 where we directly work with probabilities.

3 Three Introductory Examples

In this section we study three basic examples, to illustrate how some statistics on random words can be estimated using ambiguous specifications. We will not be fully formal, the rigorous framework will be presented in the next section.

We start with the classical question of estimating the expected number occurrences of a fixed pattern v of length m in a uniform random word w of length n. Occurrences may overlap: aaa has two occurrences of aa in our settings. Let α_v be the random variable that counts the number of occurrences of v in w. The classical probabilistic analysis of the expectation $\mathbb{E}_n[\alpha_w]$ of α_w for the uniform distribution on A^n is the following: for any $i \in [n]$ let X_i be the random variable that values 1 if there is an occurrence of v in w starting at position i and that values 0 otherwise. Then we have $\alpha_v = \sum_{i=1}^n X_i$. The X_i's are not independent, but since the expectation is linear, we have $\mathbb{E}_n[\alpha_v] = \sum_{i=1}^n \mathbb{E}[X_i]$. As a consequence, $\mathbb{E}[Z_n] = (n-m+1)\ell^{-m} \sim n\ell^{-m}$, as v is fixed in our settings.

As we are working with the uniform distribution, the probabilistic proof can also be established in a purely combinatorial manner: We just count the number of words of length n having an occurrence of v at position i, and get exactly the same computations.

There is another, more advanced, way to obtain this result using combinatorics. The symbolic method described in Section 2.2 works when one starts with an unambiguous combinatorial specification. If the regular expression is ambiguous, then applying blindly the rules of transformation does not produce the correct generating series. Nonetheless, the resulting series $L(z)$ can still be useful: roughly speaking, if $\kappa(w)$ denote the number of different ways that the word w can be parsed in the expression (we call this quantity the *degree of ambiguity of* w), then $L(z) = \sum_w \kappa(w) z^{|w|}$. We can take advantage of this property, provided we can design an ambiguous expression such that for every word, *the value of the statistic is equal to its degree of ambiguity.* Back to our example, it is not difficult to see that for the ambiguous expression $\mathcal{L} = A^* v A^*$, each word w can be parsed in a number of ways equal to the number of occurrences of v in w. Hence, using the dictionary of the symbolic method, we get that $L_{\alpha_v}(z) = \frac{z^m}{(1-\ell z)^2}$. From this expression we obtain:

$$\sum_{|w|=n} \alpha_v(w) = [z^n]\frac{z^m}{(1-\ell z)^2} = [z^{n-m}]\frac{1}{(1-\ell z)^2} = (n-m+1)\ell^{n-m}.$$

We just have to divide by ℓ^n to get the expectation of α_v. Instead, we can use the Simplified Transfer Theorem directly on $\frac{z^m}{(1-\ell z)^2}$ to obtain that $\mathbb{E}_n[\alpha_v] \sim n\ell^{-m}$. It is probably too

complicated to use analytic combinatorics here, but in many situations, we will not want to find an exact expression for the n-th coefficient, if it can be avoided. Using the Transfer Theorem, we can find asymptotic equivalents without first computing the coefficients.

Let us consider another simple example. Assume that we are now interested in the number $\beta_v(w)$ of occurrences of v as a subword of w. The expectation of β_v for random words of length n can be established using probabilities and the linearity of the expectation as for α_v. However, we want to illustrate the use of analytic tools once more. It is not difficult to verify that the ambiguous expression $\mathcal{L} = A^*v_1A^*v_2A^* \cdots A^*v_mA^*$ corresponds to our needs. Its associated generating series is $L(z) = \frac{z^m}{(1-\ell z)^{m+1}}$, which satisfies the conditions of the Simplified Transfer Theorem. This yields that $[z^n]L(z) \sim \frac{\ell^{n-m}n^m}{m!}$. As there are ℓ^n words of length n, the expected number of occurrences of v as a subword of a random word of length n is asymptotically equivalent to $\frac{n^m}{m!\ell^m}$. See [7] for more information on statistics related to subwords.

We conclude this section with a last elementary example. Let $\pi(w)$ denote the length of the largest word v such that $w \in vA^*\overline{v}$, where $\overline{v}$ denote the reverse (or mirror) of v. The description $\mathcal{L} = \cup_{v\in A^+} vA^*\overline{v}$ is ambiguous, but a word w is in exactly $\pi(w)$ sets of this union, since the number of nonempty prefixes of a word is equal to its length. The specification $\mathcal{L}$ can be rewritten $\mathcal{E}\times A^*$, where $\mathcal{E}$ is the set of pairs $(v, \overline{v})$ for nonempty v. The generating series of $\mathcal{E}$ is $E(z) = \frac{\ell z^2}{1-\ell z^2}$, and the symbolic method yields that $L(z) = \frac{E(z)}{1-\ell z}$. As $E(z)$ is analytic at 0 with radius of convergence $\frac{1}{\sqrt{\ell}} > \frac{1}{\ell}$, the Simplified Transfer Theorem applies and yields that $[z^n]L(z) \sim E(\ell^{-1})\ell^n = \frac{1}{\ell-1}\ell^n$. Hence, the expected value of π tends to $\frac{1}{\ell-1}$.

In the sequel, we define a framework on *sets of weighted words* to formalize what we did for our three introductory examples. It is directly inspired from the simple remarks we just made, on how ambiguity can be used to estimate statistics. However, this is done in a more sophisticate way. We will be able, for instance, to handle non-integer degrees of ambiguity, which will prove useful in Section 5.

4 Combinatorics of Sets of Weighted Words

In this section we introduce the framework that will be used throughout this article. The idea is to formalize the notion of "number of time an ambiguous expression is parsed", and to do it in a way similar to the symbolic method. For this purpose, we have to introduce some formalism on sets of weighted words. The definitions we propose are natural extensions of the classical ones on sets.

Consider the two sets of words $\mathcal{E} = \{a, ab, aa\}$ and $\mathcal{F} = \{\varepsilon, a, b\}$. We interpret them as "each word of $\mathcal{E}$ has weight 1", and the same for $\mathcal{F}$. Since a is in both $\mathcal{E}$ and $\mathcal{F}$, we would like a to have weight two in $\mathcal{E} \cup \mathcal{F}$. Similarly, since $ab = a \cdot b = ab \cdot \varepsilon$, we would like ab to have weight two in $\mathcal{E} \cdot \mathcal{F}$. Finally, since $aaa = a \cdot a \cdot a = a \cdot aa = aa \cdot a$, we would like aaa to have weight three in $\mathcal{E}^*$. A relevant way to handle this is to use multisets, that is, sets where an element may appear more than once. We will need a bit more in the sequel, and thus allow the weights to take any real positive value in the definitions below.

Formally, if $\mathcal{E}$ be a nonempty set, a *weighted set*[2] on $\mathcal{E}$ is a mapping $\mathcal{M}$ from $\mathcal{E}$ to $\mathbb{R}_{\geq 0}$. For $e \in \mathcal{E}$, we say that e is in $\mathcal{M}$ (written $e \in \mathcal{M}$) if $\mathcal{M}(e) \neq 0$, and we write $e \notin \mathcal{M}$ otherwise. A set $\mathcal{M}$ is viewed as a weighted set where every element of e has weight 1: for every $e \in \mathcal{E}$, $\mathcal{M}(e) = 1$ if $e \in \mathcal{M}$ and $\mathcal{M}(e) = 0$ otherwise.

[2] We use the terminology "weighted set on $\mathcal{E}$" for "set of weighted elements of $\mathcal{E}$", as a weighted graph is a graph of weighted vertices.

If $\mathcal{E}$ is a combinatorial set of size function s, we define *the generating series* $M(z)$ *of a weighted set* $\mathcal{M}$ *on* $\mathcal{E}$ by $M(z) = \sum_{e\in\mathcal{E}} \mathcal{M}(e)z^{s(e)}$. Observe that if $\mathcal{M}$ is a set, then the generating series of $\mathcal{M}$ viewed as a weighted set or as a set coincide.

From now on, we only work on weighted sets of words on A. To simplify the notations, we will sometimes write $\mathcal{M} = \{a \mapsto \frac{1}{2}, ba \mapsto 3, baba \mapsto 11\}$ for the weighted set defined by $\mathcal{M}(a) = \frac{1}{2}$, $\mathcal{M}(ba) = 3$, $\mathcal{M}(baba) = 11$, and $\mathcal{M}(x) = 0$ for every $x \notin \{a, ba, baba\}$.

If $\mathcal{M}$ and $\mathcal{M}'$ are two weighted sets of words, the *sum* $\mathcal{M} \oplus \mathcal{M}'$ is the weighted set $\mathcal{N}$ defined by $\mathcal{N}(w) = \mathcal{M}(w) + \mathcal{M}'(w)$, for every $w \in A^*$. The *concatenation* $\mathcal{M} \odot \mathcal{M}'$ of the weighted sets $\mathcal{M}$ and $\mathcal{M}'$ is defined by

$$\mathcal{M} \odot \mathcal{M}' = \bigoplus_{\substack{v\in\mathcal{M}\\ v'\in\mathcal{M}'}} \{vv' \mapsto \mathcal{M}(v)\mathcal{M}'(v')\}.$$

That is, every pair (v, v') contributes additively to $\mathcal{M}(v)\mathcal{M}(v')$ to the weight of the word vv'. For instance, if $\mathcal{M} = \{a \mapsto 1/2, ab \mapsto 3\}$ and $\mathcal{M}' = \{\varepsilon \mapsto 5, b \mapsto 7\}$, then their concatenation is $\mathcal{M} \odot \mathcal{M}' = \{a \mapsto 5/2, ab \mapsto 37/2, abb \mapsto 21\}$.

If $\varepsilon \notin \mathcal{M}$, the *star* $\mathcal{M}^\star$ is defined by $\mathcal{M}^\star = \oplus_{k\geq 0}\mathcal{M}^k$, where $\mathcal{M}^0 = \{\varepsilon \mapsto 1\}$ and $\mathcal{M}^{k+1} = \mathcal{M}^k \odot \mathcal{M}$ for every $k \geq 0$. Observe that if $\varepsilon \in \mathcal{M}$, then this operation is not well defined, as ε is in every $\mathcal{M}^k$ and therefore has infinite weight in $\mathcal{M}^*$.

The following proposition extends the symbolic method to weighted sets of words.

▶ **Proposition 3.** *If* $\mathcal{M}$ *and* $\mathcal{M}'$ *are two weighted sets of words, then*

$$\begin{aligned}
\mathcal{N} = \mathcal{M} \oplus \mathcal{M}' \quad &\Rightarrow \quad N(z) = M(z) + M'(z),\\
\mathcal{N} = \mathcal{M} \odot \mathcal{M}' \quad &\Rightarrow \quad N(z) = M(z)M'(z),\\
\mathcal{N} = \mathcal{M}^\star \quad &\Rightarrow \quad N(z) = \frac{1}{1 - M(z)}, \qquad (\text{if } \varepsilon \notin \mathcal{M}).
\end{aligned}$$

In the sequel, we will implicitly use the following lemma, which was already presented informally in Section 3.

▶ **Lemma 4.** *Let* $\alpha_v(w)$ *denote the number of occurrences of* v *as a factor of* w*. The generating series of the weighted set* $A^* \odot \{v \mapsto 1\} \odot A^*$ *(the weighted set version of* A^*vA^**) is equal to* $L_{\alpha_v}(z)$*, the cumulative generating series of the statistic* α_v*.*

Proof. As A^*v is a unambiguous expression, every element of $A^* \odot \{v \mapsto 1\}$ has weight 1, and the same holds for A^*. Thus, by definition, if $\mathcal{N} = (A^* \odot \{v \mapsto 1\}) \odot A^*$, then $\mathcal{N}(w) = |\{(w_1, w_2) \in A^* \times A^* : w = w_1 v \cdot w_2\}|$, which is exactly $\alpha_v(w)$, as announced. ◀

5 Application to Run Statistics

5.1 The Expected Number of Runs

For any given word v, let $\rho(v)$ denote its number of runs. In [15], Puglisi and Simpson established the following result.

▶ **Theorem 5** ([15])**.** *The expected number of runs in a word of length* n *on an alphabet of size* ℓ *satisfies asymptotically*

$$\mathbb{E}_n[\rho] \sim \left(\frac{\ell - 1}{\ell} \sum_{k\geq 1} \frac{\mu(k)}{\ell^{2k-1} - 1}\right) n.$$

To prove Theorem 5, they proceed as follows. For every given p, they compute the total number of runs of period p in the set of all words of length n. Then, they sum these values for all possible p. Finally, they obtain an asymptotic equivalent of this quantity using elementary, but technical, computations.

In this section, we propose an alternative proof of Theorem 5 using our framework. Recall that $\mathcal{P}$ is the set of all primitive words and that $P(z)$ is its associated generating series. Let $\mathcal{C} = \{ww \mapsto 1 : w \in \mathcal{P}\}$ and let $\mathcal{D} = \{aww \mapsto 1 : w \in \mathcal{P} \text{ and the last letter of } w \neq a\}$.

▶ **Lemma 6.** *The generating series of the weighted set $(\mathcal{C} \odot A^*) \oplus (A^* \odot \mathcal{D} \odot A^*)$ is the cumulative generating series of the statistic ρ.*

Proof. For the weighted set $\mathcal{M} = \mathcal{C} \odot A^* = \oplus_{w\in\mathcal{P}}\{ww \mapsto 1\} \odot A^*$, $\mathcal{M}(w)$ is the number of prefixes of the form ww for $w \in \mathcal{P}$, that is, $\mathcal{M}$ counts the number of runs at the beginning of the word. Similarly, for $\mathcal{N} = A^* \odot \mathcal{D} \odot A^* = \oplus_{w\in\mathcal{D},a\neq w[|w|]} A^* \odot \{aww \mapsto 1\} \odot A^*$, $\mathcal{N}(w)$ is the number of runs of w that does not start at the first position, since each run is identified by the factor aww. Hence, $\mathcal{M} \oplus \mathcal{N}$ counts the number of runs, concluding the proof. ◀

The generating series of $\mathcal{C}$ and $\mathcal{D}$ are $C(z) = P(z^2)$ and $D(z) = (\ell - 1)zP(z^2)$, respectively. Hence, the cumulative generating series $L_\rho(z)$ of the number of runs can be obtained using Proposition 3:

$$L_\rho(z) = \frac{P(z^2)}{1-\ell z} + \frac{(\ell-1)zP(z^2)}{(1-\ell z)^2}.$$

Since the radius of convergence of $P(z^2)$ is $\frac{1}{\sqrt{\ell}} > \frac{1}{\ell}$, we are in the settings of Equation (3) and the Simplified Transfer Theorem yields that $[z^n]L_\rho(z) \sim n\frac{\ell-1}{\ell}P(\ell^{-2})\ell^n$. Dividing by ℓ^n gives another expression for the result of Theorem 5:

$$\mathbb{E}_n[\rho] \sim \frac{\ell-1}{\ell}\, P\left(\frac{1}{\ell^2}\right) n. \tag{4}$$

In particular, the infinite sum of Theorem 5 is just $P(\ell^{-2})$. Indeed, by Equation (2) we have

$$P\left(\frac{1}{\ell^2}\right) = \sum_{k\geq 1} \frac{\mu(k)\,\ell\cdot\ell^{-2k}}{1-\ell\cdot\ell^{-2k}}.$$

Multiplying the numerator and denominator by ℓ^{2k-1} yields the formula of Theorem 5.

5.2 The Expected Total Run Length

The *total run length* of a word is the sum of the lengths of its runs. We denote by $\tau(w)$ the total run length of w. In [11], Glen and Simpson proved the following result.

▶ **Theorem 7** ([11])**.** *The expected total run length of a uniform random word of length n asymptotically satisfies*

$$\mathbb{E}_n[\tau] \sim \left(\sum_{k\geq 1} P_k \frac{2k(\ell-1)+1}{\ell^{2k+1}}\right) n,$$

where P_k is the number of primitive words of length k.

Their techniques follows the steps of the proof of Theorem 5 given in Section 5.1.

In order to prove Theorem 7 with our framework, we first focus on another statistic. For any word w, let $\delta(w)$ denote the sum of the periods of the runs of w. We are interested in the expected value of δ for uniform random words of length n. Consider the weighted set $\overline{\mathcal{C}} = \{ww \mapsto |w| : \ w \in \mathcal{P}\}$, where the weight of each ww is the length of w. A direct computation yields that the generating series of $\overline{\mathcal{C}}$ is $\overline{C}(z) = z^2P'(z^2)$. Similarly the generating series of the weighted set $\overline{\mathcal{D}} = \{aww \mapsto |w| : \ w \in \mathcal{P}$ and the last letter of $w \neq a\}$ is $\overline{D}(z) = (\ell-1)z^3P'(z^2)$.

We can now reuse the ambiguous specification of Lemma 6, with $\overline{\mathcal{C}}$ and $\overline{\mathcal{D}}$ instead of $\mathcal{C}$ and $\mathcal{D}$, and get that the cumulative generating series of δ is

$$L_\delta(z) = \frac{z^2P'(z^2)}{1-\ell z} + \frac{(\ell-1)z^3P'(z^2)}{(1-\ell z)^2}, \text{ with } P'(z) = \frac{d}{dz}P(z).$$

By Equation 3, from this expression of $L_\delta(z)$ we directly get the following proposition.

▶ **Proposition 8.** *The expected sum of the periods of the runs in a uniform random word of length n asymptotically satisfies $\mathbb{E}_n[\delta] \sim \frac{\ell-1}{\ell^3}P'(\ell^{-2})\, n$.*

We can now proceed with our proof of Theorem 7. Consider the ambiguous specification $\mathcal{L} = \cup_{w\in\mathcal{P}} A^*wwA^*$. Observe that a run $r = (i, j, p)$ in a word v matches the expression of $\mathcal{L}$ exactly once for every $w = v[k, k+p-1]$, with $k \in \{i, \dots, j-2p+1\}$. That is, the pair (v, r) matches the specification exactly $|r| - 2p + 1$ times. In other words, the generating series of the weighted set $A^* \odot \mathcal{C} \odot A^*$ is the cumulative generating series of the statistic $\tau - 2\delta + \rho$ (recall that τ is the total run length, δ is the sum of periods and ρ is the number of runs). Thus, Proposition 3 directly yields:

$$\frac{P(z^2)}{(1-\ell z)^2} = L_\tau(z) - 2L_\delta(z) + L_\rho(z) \Rightarrow L_\tau(z) = \frac{P(z^2)}{(1-\ell z)^2} + 2L_\delta(z) - L_\rho(z).$$

Theorem 2 applies and we obtain that

$$\mathbb{E}_n[\tau] = \frac{1}{\ell^n}[z^n]L_\tau(z) \sim \left(\frac{2(\ell-1)}{\ell^3}P'\left(\frac{1}{\ell^2}\right) + \frac{1}{\ell}P\left(\frac{1}{\ell^2}\right)\right) n, \tag{5}$$

which is another formulation of Theorem 7. Indeed, since $P(z) = \sum_{k\geq 1} P_k z^k$, we have

$$\frac{1}{\ell}P\left(\frac{1}{\ell^2}\right) = \frac{1}{\ell}\sum_{k\geq 1}\frac{P_k}{\ell^{2k}} = \sum_{k\geq 1}\frac{P_k}{\ell^{2k+1}}.$$

Moreover, $P'(z) = \sum_{k\geq 1} kP_k z^{k-1}$, and thus

$$\frac{2(\ell-1)}{\ell^3}P'\left(\frac{1}{\ell^2}\right) = \frac{2(\ell-1)}{\ell^3}\sum_{k\geq 1}\frac{kP_k}{\ell^{2k-2}} = \sum_{k\geq 1}P_k\frac{2k(\ell-1)}{\ell^{2k+1}}.$$

Summing the two terms yields the formula of Theorem 7.

5.3 The Expected Sum of Exponents

For any word $v \in A^*$, let $\gamma(v)$ denote the sum of the exponents of the runs of v. In [13], Kusano, Matsubara, Ishino and Shinohara proved the following result.

▶ **Theorem 9** ([13]). *The expected sum of the exponents of runs for uniform random words of length n satisfies asymptotically:*

$$\mathbb{E}_n[\gamma] \sim \left(\sum_{k\geq 1} \mu(k) \left(\frac{2(\ell-1)}{\ell^{2k}-\ell} + \frac{1}{k\ell} \log\left(\frac{\ell^{2k}}{\ell^{2k}-\ell} \right) \right) \right) n.$$

We follow the analysis of the previous section: A run $r=(i,j,p)$ in a word v matches the expression $\mathcal{L} = \cup_{w\in\mathcal{P}} A^* w w A^*$ exactly $|r|-2p+1$ times. Since we want to compute the statistic γ, we have to divide the contribution of each run (i,j,p) by p.

Let $\tilde{\mathcal{C}} = \{ww \mapsto \frac{1}{|w|} : w \in \mathcal{P}\}$ and let $\tilde{\mathcal{D}} = \{aww \mapsto \frac{1}{|w|} : w \in \mathcal{P} \text{ and } w_{|w|} \neq a\}$. Let $\tilde{C}(z)$ and $\tilde{D}(z)$ denote their generating series. By Proposition 3, the generating series of $\tilde{\mathcal{L}} = A^* \odot \tilde{\mathcal{C}} \odot A^*$ is $\tilde{L}(z) = \frac{\tilde{C}(z)}{(1-\ell z)^2}$. Moreover, it satisfies:

$$\tilde{L}(z) = \sum_{v\in A^*} \sum_{\substack{r\in \text{RUNS}(v) \\ r=(i,j,p)}} \frac{|r|-2p+1}{p} z^{|v|} = L_\gamma(z) - 2L_\rho(z) + \sum_{v\in A^*} \sum_{\substack{r\in \text{RUNS}(v) \\ r=(i,j,p)}} \frac{z^{|v|}}{p} \tag{6}$$

Let $\xi(v)$ be the sum of $\frac{1}{p}$ for every $(i,j,p) \in \text{RUNS}(v)$. Using exactly the same idea as in Section 5.1, its cumulative series is $L_\xi(z) = \frac{\tilde{C}(z)}{1-\ell z} + \frac{\tilde{D}(z)}{(1-\ell z)^2}$. Hence, Equation (6) rewrites

$$L_\gamma(z) = 2L_\rho(z) + \frac{\tilde{C}(z)-\tilde{D}(z)}{(1-\ell z)^2} - \frac{\tilde{C}(z)}{1-\ell z}.$$

Since the radius of convergence of both $\tilde{C}(z)$ and $\tilde{D}(z)$ is $1/\sqrt{\ell}$, the Simplified Transfer Theorem applies. We obtain that the expected value of γ asymptotically satisfies

$$\mathbb{E}_n[\gamma] \sim \left(\frac{2(\ell-1)}{\ell} P\left(\frac{1}{\ell^2}\right) + \frac{1}{\ell} Q\left(\frac{1}{\ell^2}\right) \right) n, \tag{7}$$

where the function $Q(z) = \int_0^z P(t)t^{-1}dt$ naturally appears when simplifying $\tilde{C}(\ell^{-1}) - \tilde{D}(\ell^{-1})$.

One can check that Equation (7) is just another formulation of Theorem 9. Indeed, we have

$$2\frac{\ell-1}{\ell} P\left(\frac{1}{\ell^2}\right) = \sum_{k\geq 1} \mu(k) \frac{2(\ell-1)}{\ell(\ell^{2k-1}-1)} = \sum_{k\geq 1} \mu(k) \frac{2(\ell-1)}{\ell^{2k}-\ell}.$$

And since everything is normally convergent,

$$\left(\frac{1}{\ell^2}\right) = \int_0^{1/\ell^2} P(t)t^{-1}dt = \int_0^{1/\ell^2} \sum_{k\geq 1} \frac{\mu(k)}{t} \frac{\ell\, t^k}{1-\ell\, t^k} dt = \sum_{k\geq 1} \mu(k) \int_0^{1/\ell^2} \frac{\ell\, t^{k-1}}{1-\ell\, t^k} dt.$$

Observe that the derivative of $t \mapsto -\log(1-\ell\, t^k)$ is $t \mapsto \frac{k\ell\, t^{k-1}}{1-\ell\, t^k}$. Thus

$$Q\left(\frac{1}{\ell^2}\right) = \sum_{k\geq 1} \frac{\mu(k)}{k} \log \frac{1}{1-\ell^{1-2k}} = \sum_{k\geq 1} \frac{\mu(k)}{k} \log \frac{\ell^{2k}}{\ell^{2k}-\ell}.$$

This gives the announced result.

6 Generalization to Memoryless Sources

In this section, we show how our formalism can be used to generalize the results to memoryless sources (see Section 2.1 for the definition). From now on, the alphabet is $A = \{a_1, \dots, a_\ell\}$ and we have a probability function p on A that charges at least two letters:[3] $p(a_i) < 1$ for every $i \in [\ell]$. Let $\vec{p}$ be the vector $\vec{p} = (p(a_1), \dots, p(a_\ell))$.

6.1 Multivariate Generating Series and Memoryless Sources

For $v \in A^*$ and $i \in [\ell]$, let $|v|_i$ denote the number of occurrences of the letter a_i in v. In our settings, multivariate generating series are formal power series on the formal variables z, u_1, $\dots$, u_ℓ. When needed, we will use the vector $\vec{u} = (u_1, \dots, u_\ell)$ to simplify the notations. For any positive integer k, let $\vec{u}^k$ denote the vector $(u_1^k, \dots, u_\ell^k)$, and let $N_k(\vec{u}) = u_1^k + \ldots + u_2^k$.

The *multivariate generating series* $L(z, \vec{u})$ of a language $\mathcal{L}$ is defined by

$$L(z, \vec{u}) := \sum_{v \in A^*} z^{|v|} \prod_{i=1}^{\ell} u_i^{|v|_i} = \sum_{n, k_1, \dots, k_\ell \geq 0} L(n, k_1, \dots, k_\ell) z^n u_1^{k_1} \cdots u_\ell^{k_\ell},$$

where $L(n, k_1, \dots, k_\ell)$ is the number of words of length n of $\mathcal{L}$ with exactly k_i occurrences of a_i, for every $i \in [\ell]$.

Multivariate generating series are widely use in combinatorics and analytic combinatorics. In particular, when the parameters controlled by the u_i's are additive, the symbolic method can be extended, giving efficient techniques to build the series. We refer the interested reader to [6, Ch. III] for more information on this topic. Interestingly, we can also extend our framework to multivariate generating series, when the u_i's are associated with the number of occurrences of the letters. First, the definition is extended to a weighted set $\mathcal{M}$ by weighting each word: $M(z, \vec{u}) := \sum_{v \in A^*} \mathcal{M}(v)\, z^{|v|} \prod_{i=1}^{\ell} u_i^{|v|_i}$. Proposition 3 is then directly generalized: if $\mathcal{M}$ and $\mathcal{N}$ are two weighted sets then the multivariate series of $\mathcal{M} \oplus \mathcal{N}$ is $M(z, \vec{u}) + N(z, \vec{u})$, the one of $\mathcal{M} \odot \mathcal{N}$ is $M(z, \vec{u})N(z, \vec{u})$, and the one of $\mathcal{M}^\star$ is $\frac{1}{1 - M(z, \vec{u})}$.

The main reason to consider multivariate series is the following: if $L(z, \vec{u})$ is the series of a language $\mathcal{L}$, then if we instantiate every formal variable u_i with the value $p(a_i)$, which we simply write $L(z, \vec{p})$, then we obtain a univariate series such that $[z^n]L(z, \vec{p})$ is exactly the probability that a word of length n belongs to $\mathcal{L}$, for the memoryless model of probability p. Similarly, if the generating series $M(z)$ of the weighted set $\mathcal{M}$ is the cumulative generating series of a statistic χ (for the uniform distribution), then $\mathbb{E}_n[\chi] = [z^n]M(z, \vec{p})$ for the memoryless distribution of probability p. The proofs of these facts are completely straightforward. However, together with the symbolic method, this provides a useful framework to deal with statistics on random words for memoryless distributions.

As an example, let us consider our first introductory statistic, the number of occurrences of the pattern v in a word. We use the weighted set description $A^* \odot \{v \mapsto 1\} \odot A^*$. The multivariate series of $A^\star$ is $\frac{1}{1 - z\, N_1(\vec{u})}$, since it is the weighted star of A, whose multivariate series is $A(z, \vec{u}) = u_1 z + \ldots + u_\ell z = N_1(\vec{u})\, z$. The multivariate series of $\{v \mapsto 1\}$ is $V(z, \vec{u}) = z^{|v|} u_1^{|v|_1} \cdots u_\ell^{|v|_\ell}$. Hence, the multivariate series of the number of occurrences of v is $\frac{V(z, \vec{u})}{(1 - N_1(\vec{u})z)^2}$. For $\vec{u} = \vec{p}$, we have $N_1(\vec{p}) = 1$, since p is a probability, and $V(z, \vec{p}) = \mathbb{P}_p(v) z^{|v|}$, by definition of a memoryless model. Hence, the multivariate series for $\vec{u} = \vec{p}$ is equal to $\frac{\mathbb{P}_p(v)}{(1-z)^2}$. The Simplified Transfer Theorem yields that the expected number of occurrences of v in a word of length n is asymptotically $\mathbb{P}_p(v)n$, for this memoryless distribution.

[3] Everything is trivial if $p(a_i) = 1$ for some i, as the only word of A^n with positive probability is a_i^n.

6.2 Expected Number of Runs for Memoryless Sources

We start as in Section 5.1, and use the weighted set $(\mathcal{C} \odot A^*) \oplus (A^* \odot \mathcal{D} \odot A^*)$ to count the number of runs, with $\mathcal{C} = \{ww \mapsto 1 : w \in \mathcal{P}\}$ and $\mathcal{D} = \{aww \mapsto 1 : w \in \mathcal{P} \text{ and } w_{|w|} \neq a\}$. The associated multivariate series is therefore $L(z, \vec{u}) = \frac{C(z,\vec{u})}{1-N_1(\vec{u})z} + \frac{D(z,\vec{u})}{(1-N_1(\vec{u})z)^2}$, where $C(z, \vec{u})$ and $D(z, \vec{u})$ are the multivariate series of $\mathcal{C}$ and $\mathcal{D}$.

At this point we have to compute the multivariate generalization $P(z, \vec{u})$ of $P(z)$, the series of primitive words. We will also need to compute $P_i(z, \vec{u})$, the multivariate series of the primitive words that ends by a_i. This is done using Equation (1), which readily extends to multivariate series in our case, yielding

$$P(z, \vec{u}) = \sum_{k\geq 1} \frac{\mu(k) z^k N_k(\vec{u})}{1 - z^k N_k(\vec{u})} \text{ and } P_i(z, \vec{u}) = \sum_{k\geq 1} \frac{\mu(k) z^k u_i^k}{1 - z^k N_k(\vec{u})}.$$

Moreover, $C(z, \vec{u}) = P(z^2, \vec{u}^2)$ and it is easy to compute from $P_i(z, \vec{u})$ that

$$D(z, \vec{u}) = \sum_{i\in[\ell]} z v_i P_i(z^2, \vec{u}^2) = \sum_{k\geq 1} \mu(k) z^{2k+1} \frac{\sum_{i=1}^{\ell} v_i u_i^{2k}}{1 - N_{2k}(\vec{u}) z^{2k}}, \text{ with } v_i = \sum_{\substack{j\in[\ell] \\ j\neq i}} u_i.$$

This formula looks complicated, but it simplifies when evaluating it at $z = 1$, the dominant singularity, and at $\vec{u} = \vec{p}$. In particular, if $\vec{u} = \vec{p}$, then $v_i = 1 - p(a_i)$ and $\sum_{i=1}^{\ell} v_i u_i^k = N_k(\vec{p}) - N_{k+1}(\vec{p})$. Hence, applying the Simplified Transfer Theorem to the expression of $L(z, \vec{p})$ yields the following result.

▶ **Theorem 10.** *For the memoryless distribution of probability p, the expected number of runs in a random word of length n satisfies asymptotically*

$$\mathbb{E}_n[\rho] \sim D(1, \vec{p})n = \left(\sum_{k\geq 1} \mu(k) \frac{N_{2k}(\vec{p}) - N_{2k+1}(\vec{p})}{1 - N_{2k}(\vec{p})} \right) n.$$

7 Conclusions

As illustrated throughout this article, the framework we propose is quite useful to study some statistics on random words. We choose to focus on presenting the technique itself in this extended abstract, to try to convince the reader that it is a precious tool to estimate the expectation of various parameters on words.

Due to the lack of space, we only generalized the result on the expected number of runs to memoryless distributions, but the other theorems of Section 5 can also be extended in a similar way. Some other kinds of generalizations can also be obtained. For instance, the expected number of cubic-runs (runs of exponent at least 3) is asymptotically equivalent to $\frac{\ell-1}{\ell} P(\ell^{-3})n$, which can be obtained as in Section 5.1. More generally, all results can readily be generalized to k-runs. Other known statistics can be studied using this method: as a last example, the expected number of squares χ in a word, i.e. the number of factors of the form vv for nonempty v was studied in [3]. In our framework, this corresponds to the weighted set $\oplus_{v\in A+} A^* \odot \{vv \mapsto 1\} \odot A^*$, thus $L_\chi(z) = \frac{\ell z^2}{(1-\ell z)^2(1-\ell z^2)}$ and $\mathbb{E}_n[\chi] \sim \frac{n}{\ell-1}$.

A natural extension of this work would be to provide similar tools to deal with higher moments, in particular with the variance. However, what we did in this article is related to the linearity of the expectation, and the variance is not linear. To compute higher moments, we have to handle dependencies between runs in a word, which is much more complicated. It would also be interesting to revisit some other probabilistic studies of the literature, such as [9, 2, 10], to see if they can be included in the framework of sets of weighted words.

References

1 Hideo Bannai, Tomohiro I, Shunsuke Inenaga, Yuto Nakashima, Masayuki Takeda, and Kazuya Tsuruta. The "Runs" Theorem. *CoRR*, abs/1406.0263, 2014.

2 Manolis Christodoulakis, Michalis Christou, Maxime Crochemore, and Costas S. Iliopoulos. Abelian borders in binary words. *Discrete Applied Mathematics*, 171:141–146, 2014.

3 Manolis Christodoulakis, Michalis Christou, Maxime Crochemore, and Costas S. Iliopoulos. On the average number of regularities in a word. *Theoretical Computer Science*, 525:3–9, 2014.

4 Maxime Crochemore and Lucian Ilie. Maximal repetitions in strings. *Journal of Computer and Systems Sciences*, 74(5):796–807, 2008.

5 Maxime Crochemore, Lucian Ilie, and Liviu Tinta. The "runs" conjecture. *Theoretical Computer Science*, 412(27):2931–2941, 2011.

6 Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2008.

7 Philippe Flajolet, Wojciech Szpankowski, and Brigitte Vallée. Hidden word statistics. *Journal of the ACM*, 53(1):147–183, 2006.

8 Frantisek Franek and Qian Yang. An asymptotic lower bound for the maximal number of runs in a string. *Intern. Journal of Foundations Computer Science*, 19(1):195–203, 2008.

9 Kimmo Fredriksson and Szymon Grabowski. Average-optimal string matching. *Journal of Discrete Algorithms*, 7(4):579–594, 2009.

10 Pawel Gawrychowski, Gregory Kucherov, Benjamin Sach, and Tatiana A. Starikovskaya. Computing the longest unbordered substring. In Costas S. Iliopoulos, Simon J. Puglisi, and Emine Yilmaz, editors, *String Processing and Information Retrieval – 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*, volume 9309 of *Lecture Notes in Computer Science*, pages 246–257. Springer, 2015.

11 Amy Glen and Jamie Simpson. The total run length of a word. *Theoretical Computer Science*, 501:41–48, 2013.

12 Roman Kolpakov and Gregory Kucherov. Finding maximal repetitions in a word in linear time. In *Proceedings of the 1999 Symposium on Foundations of Computer Science (FOCS'99), New York (USA)*, pages 596–604, New-York, October 17-19 1999. IEEE Computer Society.

13 Kazuhiko Kusano, Wataru Matsubara, Akira Ishino, and Ayumi Shinohara. Average value of sum of exponents of runs in a string. *Intern. Journal of Foundations of Computer Science*, 20(06):1135–1146, 2009.

14 Wataru Matsubara, Kazuhiko Kusano, Akira Ishino, Hideo Bannai, and Ayumi Shinohara. New lower bounds for the maximum number of runs in a string. In Jan Holub and Jan Zdárek, editors, *Proceedings of the Prague Stringology Conference 2008, Prague, Czech Republic, September 1-3, 2008*, pages 140–145, 2008.

15 Simon J. Puglisi and Jamie Simpson. The expected number of runs in a word. *Australasian Journal of Combinatorics*, 42:45–54, 2008.

16 Simon J. Puglisi, Jamie Simpson, and William F. Smyth. How many runs can a string contain? *Theoretical Computer Science*, 401(1-3):165–171, 2008.

17 Wojciech Rytter. The number of runs in a string. *Information and Computation*, 205(9):1459–1469, 2007.

18 Jamie Simpson. Modified Padovan words and the maximum number of runs in a word. *Australasian Journal of Combinatorics*, 46:129–145, 2010.