Combinatorics, Probability and Computing (2003) 12, 73–87. © 2003 Cambridge University Press
DOI: 10.1017/S0963548302005473 Printed in the United Kingdom

On the Distribution of the Number of Missing Words in Random Texts

SVEN RAHMANN[1] and ERIC RIVALS[2]†

[1] Department of Computational Molecular Biology, Max-Planck-Institut für Molekulare Genetik
Ihnestraße 63-73, D-14195 Berlin, Germany
Sven.Rahmann@molgen.mpg.de
[2] L.I.R.M.M., CNRS U.M.R. 5506,
161 rue Ada, F-34392 Montpellier Cedex 5, France
rivals@lirmm.fr

Determining the distribution of the number of empty urns after a number of balls have been thrown randomly into the urns is a classical and well understood problem. We study a generalization: Given a finite alphabet of size σ and a word length q, what is the distribution of the number X of words (of length q) that do not occur in a random text of length $n+q-1$ over the given alphabet? For $q = 1$, X is the number Y of empty urns with σ urns and n balls. For $q \geq 2$, X is related to the number Y of empty urns with σ^q urns and n balls, but the law of X is more complicated because successive words in the text overlap. We show that, perhaps surprisingly, the laws of X and Y are not as different as one might expect, but some problems remain currently open.

1. Introduction

Let $X^{(n,\sigma,q)}$ be the random number of missing words of length q (also called q-grams) in a random text of length $n+q-1$ over an alphabet Σ of size σ. The underlying probability space is $(\Sigma^{n+q-1}, \mathcal{P}(\Sigma^{n+q-1}), \pi^{n+q-1})$, where π is the uniform distribution on Σ. Let $Y^{(n,\sigma)} := X^{(n,\sigma,1)}$; this is classically interpreted as the number of empty urns after n balls have been thrown randomly and independently into σ urns.

It is interesting to compare the laws of $X^{(n,\sigma,q)}$ and $Y^{(n,\sigma^q)}$. In both cases, n q-grams are randomly drawn. The difference is that $Y^{(n,\sigma^q)}$ counts the number of missing q-grams when they are drawn independently, and $X^{(n,\sigma,q)}$ counts the number of missing q-grams when they are drawn from a single text of length $n+q-1$, such that two successive q-grams overlap by $q-1$ characters.

The law of $Y^{(n,\sigma^q)}$ is quite well understood (for example, see [7, 14]), but the law of $X^{(n,\sigma,q)}$ has received little attention so far. An algorithmic approach to compute the

† E.R. is supported by the Programme inter-EPST BioInformatique 2000, the Genopole from Montpellier, and by the University of Montpellier II.

expectation has been published by the authors in [10]. Here we consider an asymptotic framework, where the number of experiments n and the number of possible words σ^q both tend to infinity such that their ratio $\lambda := n/\sigma^q$ remains constant.

Section 2 presents our main results. A recurring theme is that the laws of $X^{(n,\sigma,q)}$ and $Y^{(n,\sigma^q)}$ do not differ as much as one might expect. In particular, we would like to draw the reader's attention to the surprising fact that $\mathbb{E}[Y^{(n,\sigma^q)}]$ is an excellent approximation of $\mathbb{E}[X^{(n,\sigma,q)}]$ with a relative error of only $O(\sigma^{-(q+1)})$; see Theorem 2.3.

Long proofs that would interrupt the flow of reading are deferred to Section 3. Some questions remain currently open, and we discuss them in Section 4.

We are aware of two immediate applications of this work. The first one concerns so-called "monkey tests" [8] for quality assessment of pseudo random number generators (PRNGs). The PRNG is used to generate many random texts of length $n+q-1$ over a given alphabet of size σ. One determines the number of missing q-grams in each text and compares the resulting empirical distribution function with the true law of $X^{(n,\sigma,q)}$. When there are significant deviations, the PRNG is rejected because it failed to model this aspect of randomness adequately. The name "monkey test" stems from the image of a monkey typing randomly on a keyboard. Compared to other tests, monkey tests have the advantage that they can detect dependencies in a q-instance context of the PRNG that may be overlooked by other tests. In the original paper [8], the variance of $X^{(n,\sigma,q)}$ was determined empirically, and the authors conjectured a wrong formula. Our Theorem 2.5 gives the asymptotically correct variance for $q = 2$ $\left(\sigma^2\, e^{-2\lambda}\,(e^\lambda - \lambda - 1) + O(\sigma)\right)$, and we conjecture that a similar formula $(\sigma^q\, e^{-2\lambda} \cdot (e^\lambda - \lambda - 1) + O(\sigma^{q-1}))$ is true for general q. These asymptotic formulas already give good estimates for small values of n and σ.

The second application concerns biological sequence analysis. DNA sequences can be seen as texts over a 4-letter alphabet. In [13], the *linguistic complexity* of a DNA sequence of length $n+q-1$ is defined for fixed q as the number of distinct q-grams in that sequence, i.e., as $4^q - X^{(n,4,q)}$. Linguistic complexity has been used successfully as a tool in DNA sequence analysis [1, 13]. High complexity regions of DNA should look essentially random, whereas low complexity regions should have significantly more missing q-grams than expected by chance. However, in this setting, a random model that allows different probabilities for the 4 DNA nucleotides would be more realistic. While no asymptotics are currently available for this case, the algorithmic approach described in [10] can be used to obtain exact results in this case.

We emphasize that the problem we study is structurally very different from the well-understood problem to find the distribution of the number of occurrences of a word or a set of words (see [11] for a review and further references): To obtain information about the number of missing words, we have to consider all words and their interactions simultaneously. This makes the asymptotic analysis much harder: The number of words under consideration is unbounded in our asymptotic framework, whereas a fixed set of words is usually assumed when one is interested in the number of occurrences.

To our knowledge, there is currently very little literature on this type of "global" word statistic, and we hope that the questions left open here will eventually be answered.

2. Results

In Section 2.1, we lay the foundations for our main results. Using the generating function approach of Guibas and Odlyzko [4, 5, 6], we establish asymptotic values of the absence probability of a given word, depending on its periodicity. In Section 2.2, we present new results on the expected number of missing words, and in Section 2.3, we prove a particular result about the variance for word length 2. Open problems are addressed in Section 4.

2.1. Absence Probabilities

Fix a word length $q \geq 1$ and an alphabet Σ of size $\sigma \geq 2$. A q-gram is a word of length q, i.e., an element of Σ^q.

Definition. Let Q be a q-gram. Let a_m^Q be the *absence probability* of Q in a text of length m, i.e., the probability that Q does not occur in a random text of length m.

It has been established previously (see Lemma 2.1) that the absence probability of a q-gram depends on its periodicity. Therefore the following definitions are useful.

Definition. Let $Q = Q[0] \ldots Q[q-1]$ be a q-gram over some alphabet.

(i) Its *autocorrelation vector* $c^Q := (c_0^Q, \ldots, c_{q-1}^Q)$ is defined as follows: For $i = 0, \ldots, q-1$, set $c_i^Q := 1$ iff the pattern overlaps itself if slided i positions to the right, i.e., iff $Q[i+j] = Q[j]$ for all $j = 0, \ldots, (q-i-1)$. Otherwise, set $c_i^Q := 0$. Note that by this definition always $c_0^Q = 1$.

(ii) The corresponding *autocorrelation polynomial* $C^Q(z)$ is obtained by taking the c_i^Q as coefficients: $C^Q(z) := c_0^Q + c_1^Q z + c_2^Q z^2 + \cdots + c_{q-1}^Q z^{q-1}$.

(iii) The positions i for which $c_i^Q = 1$ are called *periods* of Q; every word has the trivial period 0.

(iv) If it exists, the smallest $p > 0$ for which $c_p^Q = 1$ is called the *basic period* of Q. If it does not exist, Q is said to be *period-free.*

As an example, consider the 11-gram Q=`ABRACADABRA`. Its periods are 0, 7, and 10. Therefore $c^Q = (10000001001)$ and $C^Q(z) = 1 + z^7 + z^{10}$.

The absence probability sequence a_m^Q for increasing text length m is now obtained via generating functions.

Lemma 2.1 (Guibas & Odlyzko [6]). *Let Q be a q-gram over some alphabet of size σ; let $C^Q(z)$ be its autocorrelation polynomial. Then the generating function $A^Q(z)$ of the absence probability sequence a_m^Q is given by*

$$A^Q(z) = \frac{C^Q\left(\frac{z}{\sigma}\right)}{\left(\frac{z}{\sigma}\right)^q + (1-z) \cdot C^Q\left(\frac{z}{\sigma}\right)}.$$

In other words, $a_m^Q = [z^m]\, A^Q(z)$, the Taylor coefficient of z^m in $A^Q(z)$.

For word length $q = 2$, a_{n+q-1}^Q can be obtained explicitly from Lemma 2.1.

Theorem 2.1 (Absence probabilities for two-letter words). *A 2-gram Q is either period-free, i.e., of the form* `ab` *with* $\mathtt{a} \neq \mathtt{b}$*, or periodic, i.e., of the form* `aa`*.*

(i) If Q is period-free and the alphabet size is $\sigma = 2$, then

$$a^Q_{n+1} = \frac{n+2}{2} \cdot 2^{-n}. \tag{1}$$

If Q is period-free and $\sigma \geq 3$, define

$$s(\sigma) := \sqrt{\sigma^2 - 4}, \tag{2}$$

$$\beta_0(\sigma) := \frac{\sigma}{2}\,(\sigma - s(\sigma))\,, \qquad \gamma_0(\sigma) := \frac{\sigma}{2}\,(\sigma + s(\sigma))\,. \tag{3}$$

Then

$$a^Q_{n+1} = \frac{\sigma^2 + \sigma\, s(\sigma) - 2}{2\,\sigma\, s(\sigma)} \cdot \beta_0(\sigma)^{-n} - \frac{\sigma^2 - \sigma\, s(\sigma) - 2}{2\,\sigma\, s(\sigma)} \cdot \gamma_0(\sigma)^{-n}. \tag{4}$$

(ii) If Q is periodic, define for all $\sigma \geq 2$

$$r(\sigma) := \sqrt{\frac{\sigma+3}{\sigma-1}}, \tag{5}$$

$$\beta_1(\sigma) := \frac{\sigma}{2}\,(r(\sigma) - 1)\,, \qquad \gamma_1(\sigma) := -\frac{\sigma}{2}\,(r(\sigma) + 1)\,. \tag{6}$$

Then

$$\begin{aligned} a^Q_{n+1} &= \frac{2}{\sigma(\sigma-1)} \cdot \frac{r(\sigma)+1}{r(\sigma)\,(r(\sigma)-1)^2} \cdot \beta_1(\sigma)^{-n} \\ &+ \frac{2}{\sigma(\sigma-1)} \cdot \frac{r(\sigma)-1}{r(\sigma)\,(r(\sigma)+1)^2} \cdot \gamma_1(\sigma)^{-n}. \end{aligned} \tag{7}$$

For word length $q \geq 3$, the following theorem provides an asymptotic value of a^Q_{n+q-1}, depending on the basic period of Q.

Theorem 2.2 (Asymptotic absence probabilities). *Let Q be a q-gram with $q \geq 3$. Let $n, \sigma \to \infty$ such that $\lambda := \frac{n}{\sigma^q}$ remains constant.*

(i) If Q is period-free, then

$$a^Q_{n+q-1} = e^{-\lambda} \cdot \bigl(1 - \lambda\,(q - 1/2)\,\sigma^{-q} + O(\sigma^{-2q})\bigr)\,. \tag{8}$$

(ii) If Q has a basic period p, then

$$a^Q_{n+q-1} = e^{-\lambda} \cdot \Bigl(1 + \lambda\,\sigma^{-p} + O(\sigma^{-(p+1)})\Bigr)\,. \tag{9}$$

One consequence of Theorem 2.2 is that for sufficiently large σ, the absence probability is slightly lower than $e^{-\lambda}$ for period-free words, and slightly higher for words with periods. Stated differently, self-overlapping words tend to appear later in a random text than period-free words. This is also a consequence of the stronger results proved in [2], where bounds on the difference of absence probabilities $a^Q_n - a^{Q'}_n$ of two words Q, Q' with different period sets are obtained.

2.2. Expected Number of Missing Words

This section contains our results on the expected number of missing words. We provide an exact formula for $q = 2$ and asymptotic formulas for $q \geq 3$, where we assume that q is fixed, and n and σ tend towards infinity such that the ratio $\lambda := \frac{n}{\sigma^q}$ remains constant.

From the following theorem (Theorem 2.3), we conclude the non-obvious fact that asymptotically $\mathbb{E}[X^{(n,\sigma,q)}] = \mathbb{E}[Y^{(n,\sigma^q)}] \approx \sigma^q\, e^{-\lambda}$. In other words, the expected *fraction* of missing q-grams is asymptotically $e^{-\lambda}$, whether one considers n independently generated q-grams (equivalently, n independent throws at σ^q urns), or n overlapping and hence dependent q-grams from a single text of length $n+q-1$. But more is true: In the asymptotic expansions of $\mathbb{E}[Y^{(n,\sigma^q)}]$ and $\mathbb{E}[X^{(n,\sigma,q)}]$, all of the first $q+1$ terms agree (compare Equation (10), replacing σ by σ^q, with Equation (13)). When $\mathbb{E}[X^{(n,\sigma,q)}]$ is approximated by (10), the relative error is only of order $O(\sigma^{-(q+1)})$.

Theorem 2.4 and Corollary 2.1 contain a more detailed comparison for two-letter words.

Theorem 2.3 (Expected number of missing words).

(i) *For word length $q = 1$, let $\lambda := n/\sigma$. Then*

$$\mathbb{E}[Y^{(n,\sigma)}] = \mathbb{E}[X^{(n,\sigma,1)}] \quad = \quad \sigma \cdot \left(1 - \frac{1}{\sigma}\right)^n \quad = \quad e^{-\lambda} \cdot \left(\sigma - \frac{\lambda}{2} + O(\sigma^{-1})\right). \tag{10}$$

(ii) *For word length $q = 2$ and alphabet size $\sigma = 2$,*

$$\begin{aligned}\mathbb{E}[X^{(n,2,2)}] \quad &= \quad \left(\frac{1}{2} + \frac{1}{\sqrt{5}}\right) \cdot \left(\sqrt{5} - 1\right)^{-n} \\ &+ \quad \left(\frac{1}{2} - \frac{1}{\sqrt{5}}\right) \cdot \left(-\sqrt{5} - 1\right)^{-n} \; + \; (n+1) \cdot 2^{-n}. \end{aligned} \tag{11}$$

For word length $q = 2$ and alphabet size $\sigma \geq 3$,

$$\begin{aligned}\mathbb{E}[X^{(n,\sigma,2)}] \quad &= \quad \frac{(\sigma-1)(\sigma^2 + \sigma\, s(\sigma) - 2)}{2\, s(\sigma)} \cdot \beta_0(\sigma)^{-n} \\ &+ \quad \frac{-(\sigma-1)(\sigma^2 - \sigma\, s(\sigma) - 2)}{2\, s(\sigma)} \cdot \gamma_0(\sigma)^{-n} \\ &+ \quad \frac{2}{\sigma-1} \cdot \frac{r(\sigma)+1}{r(\sigma)\,(r(\sigma)-1)^2} \cdot \beta_1(\sigma)^{-n} \\ &+ \quad \frac{2}{\sigma-1} \cdot \frac{r(\sigma)-1}{r(\sigma)\,(r(\sigma)+1)^2} \cdot \gamma_1(\sigma)^{-n}, \end{aligned} \tag{12}$$

where $s(\sigma)$, $\beta_0(\sigma)$, $\gamma_0(\sigma)$, $r(\sigma)$, $\beta_1(\sigma)$, and $\gamma_1(\sigma)$ are defined in Theorem 2.1, Equations (2), (3), (5), and (6), respectively. An asymptotic expansion of $\mathbb{E}[X^{(n,\sigma,2)}]$ is given in Theorem 2.4.

(iii) *For word length $q \geq 3$, let $\lambda := \frac{n}{\sigma^q}$.*

$$\mathbb{E}[X^{(n,\sigma,q)}] = e^{-\lambda} \cdot \left(\sigma^q - \frac{\lambda}{2} + O(\sigma^{-1})\right). \tag{13}$$

Theorem 2.4 (Comparison for two-letter words). *Let $X := X^{(n,\sigma,2)}$ be the number of missing words, and let $Y := Y^{(n,\sigma^2)}$ be the number of empty urns. Let $\lambda := \frac{n}{\sigma^2}$.*

Then

$$\begin{aligned}\mathbb{E}[X] &= e^{-\lambda}\cdot\left(\sigma^2-\frac{\lambda}{2}+\frac{\lambda\,(\lambda-2)}{2\,\sigma}+\frac{\lambda\,(4\,\lambda^2-33\,\lambda+40)}{24\,\sigma^2}+O(\sigma^{-3})\right),\\ \mathbb{E}[Y] &= e^{-\lambda}\cdot\left(\sigma^2-\frac{\lambda}{2}+\frac{3\,\lambda^2-8\,\lambda}{24\,\sigma^2}+O(\sigma^{-4})\right).\end{aligned}$$

Proof. The formula for $\mathbb{E}[Y]$ is an asymptotic expansion in σ^{-1} of (10) with alphabet size σ^2 in place of σ and $n=\lambda\sigma^2$. Likewise, the formula for $\mathbb{E}[X]$ is an asymptotic expansion in σ^{-1} of (12). □

Corollary 2.1.

$$\mathbb{E}[X]\left\{\begin{array}{c}<\\=\\>\end{array}\right\}\mathbb{E}[Y]\quad\textit{for}\quad\lambda\left\{\begin{array}{c}<\\=\\>\end{array}\right\}2+\epsilon(\sigma), \tag{14}$$

where for sufficiently large σ, $0\le\epsilon(\sigma)\to 0$ for $\sigma\to\infty$.

Proof. $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are equal up to the constant term and first differ in the σ^{-1}-term, which has the coefficient $c(\lambda):=\frac{\lambda(\lambda-2)}{2}$ in $\mathbb{E}[X]$, and is zero in $\mathbb{E}[Y]$. $c(\lambda)$ is negative for $0<\lambda<2$ and positive for $\lambda>2$. It vanishes for $\lambda=2$; in this case the coefficient for the σ^{-2}-term is $-\frac{5}{6}$ for $\mathbb{E}[X]$ and $-\frac{1}{6}$ for $\mathbb{E}[Y]$. Hence for sufficiently large σ, $\mathbb{E}[X]<\mathbb{E}[Y]$ for $\lambda=2$, and the point λ where $\mathbb{E}[X]=\mathbb{E}[Y]$ is at $2+\epsilon(\sigma)$, where $\epsilon(\sigma)$ has the stated properties. □

2.3. Variance of the Number of Missing Words

We restate the known variance for word length $q=1$, and prove an asymptotic formula for the variance when $q=2$. The difficulties that arise when one attempts to compute the variance for longer words are described in Section 4. However, we conjecture that in general $\mathrm{Var}[X^{(n,\sigma,q)}]=e^{-2\lambda}\,(e^{\lambda}-1-\lambda)\,\sigma^q+O(\sigma^{q-1})$.

Theorem 2.5 (Variance of the number of missing words).

(i) For word length $q=1$, let $\lambda:=n/\sigma$. Then

$$\begin{aligned}\mathrm{Var}[Y^{(n,\sigma)}] &= \sigma(\sigma-1)\left(1-\frac{2}{\sigma}\right)^n+\sigma\left(1-\frac{1}{\sigma}\right)^n-\sigma^2\left(1-\frac{1}{\sigma}\right)^{2n}\\ &= e^{-2\lambda}\cdot\left((e^{\lambda}-1-\lambda)\sigma+\left(\frac{3\lambda^2}{2}-\frac{\lambda e^{\lambda}}{2}\right)\right)+O(\sigma^{-1}).\end{aligned}$$

(ii) For $q=2$, let $\lambda:=n/\sigma^2$. Then

$$\mathrm{Var}[X^{(n,\sigma,2)}] = e^{-2\lambda}\cdot\left((e^{\lambda}-1-\lambda)\,\sigma^2+\lambda^2\,\sigma+\left(\frac{\lambda^3}{3}+\frac{\lambda^2}{2}-\frac{\lambda\,e^{\lambda}}{2}\right)\right)+O(\sigma^{-1}).$$

Corollary 2.2. *The variance of the number of missing 2-grams in a text of length $n+1$ is asymptotically equal to, but for sufficiently large alphabet sizes always larger than the* *variance of the number of empty urns after σ^2 balls have been randomly thrown into n urns.*

Proof. The first-order terms of $\mathrm{Var}[X^{(n,\sigma,2)}]$ and $\mathrm{Var}[Y^{(n,\sigma^2)}]$ are equal $(e^{-2\lambda} \cdot (e^{\lambda} - 1 - \lambda)\,\sigma^2)$. The coefficient of the σ-term of $\mathrm{Var}[X^{(n,\sigma,2)}]$ is $\lambda^2 > 0$, whereas it is zero for $\mathrm{Var}[Y^{(n,\sigma^2)}]$. □

3. Proofs

3.1. Absence Probabilities

Proof of Theorem 2.1.

(i) For period-free 2-grams Q, the generating function of the absence probabilities is

$$A^Q(z) = \frac{\sigma^2}{z^2 - \sigma^2\, z + \sigma^2}.$$

For $\sigma = 2$, this is $A^Q(z) = \frac{4}{(z-2)^2}$, and therefore by $(n{+}1)$-fold differentiation,

$$a^Q_{n+1} = [z^{n+1}]\, A^Q(z) = \frac{(A^Q)^{(n+1)}(0)}{(n+1)!} = \frac{n+2}{2} \cdot 2^{-n},$$

as claimed. For $\sigma \geq 3$, $A^Q(z)$ has two distinct poles, namely $\beta_0(\sigma)$ and $\gamma_0(\sigma)$, as defined in (3). Then

$$A^Q(z) = \frac{B_0(\sigma)}{z - \beta_0(\sigma)} + \frac{-B_0(\sigma)}{z - \gamma_0(\sigma)},$$

where $B_0(\sigma) = \frac{-\sigma}{s(\sigma)}$, and $s(\sigma)$ was defined in (2). Again by $(n{+}1)$-fold differentiation, one obtains

$$a^Q_{n+1} = \frac{(A^Q)^{(n+1)}(0)}{(n+1)!} = \frac{-B_0(\sigma)}{\beta_0(\sigma)^2} \cdot \beta_0(\sigma)^{-n} + \frac{B_0(\sigma)}{\gamma_0(\sigma)^2} \cdot \gamma_0(\sigma)^{-n},$$

which is exactly (4).

(ii) For a periodic 2-gram Q, one has

$$A^Q(z) = \frac{1 + \left(\frac{z}{\sigma}\right)}{\left(\frac{z}{\sigma}\right)^2 + (1-z) \cdot (1 + \left(\frac{z}{\sigma}\right))} = \frac{\frac{\sigma}{1-\sigma}\, z + \frac{\sigma^2}{1-\sigma}}{z^2 + \sigma z + \frac{\sigma^2}{1-\sigma}}.$$

For all $\sigma \geq 2$, $A^Q(z)$ has two distinct poles; these are $\beta_1(\sigma)$ and $\gamma_1(\sigma)$, as defined in (6). One obtains

$$A^Q(z) = \frac{B_1(\sigma)}{z - \beta_1(\sigma)} + \frac{C_1(\sigma)}{z - \gamma_1(\sigma)},$$

where

$$B_1(\sigma) = \frac{-\sigma}{2\,(\sigma-1)} \cdot \left(1 + \frac{1}{r(\sigma)}\right), \quad C_1(\sigma) = \frac{-\sigma}{2\,(\sigma-1)} \cdot \left(1 - \frac{1}{r(\sigma)}\right),$$

and $r(\sigma)$ was defined in (5). As above, one obtains

$$a^Q_{n+1} = \frac{-B_1(\sigma)}{\beta_1(\sigma)^2} \cdot \beta_1(\sigma)^{-n} \; + \; \frac{-C_1(\sigma)}{\gamma_1(\sigma)^2} \cdot \gamma_1(\sigma)^{-n},$$

which is (7). This completes the proof.

□

Proof of Theorem 2.2. The general idea is to obtain a partial fraction decomposition of $A^Q(z)$. We show that for $q \geq 3$, the denominator $D^Q(z) := (z/\sigma)^q + (1-z)C^Q(z)$ has a unique and simple root β of smallest absolute value. Furthermore β is slightly larger than 1. These statements are proved in Lemma 3.1 below.

Then it follows (see e.g. [3, Section 7.3]) that

$$a^Q_{n+q-1} = [z^{n+q-1}]\, A^Q(z) = \frac{C(\beta/\sigma)}{-(D^Q)'(\beta) \cdot \beta^q} \cdot \beta^{-n} + r_n,$$

where r_n is a sum of terms of the form $c \cdot \gamma^{-n}$ with γ larger than and bounded away from β, and c grows polynomially with σ. Therefore r_n can be neglected in the following asymptotic expansions.

(i) For period-free words Q, $C^Q(z) \equiv 1$, and $D^Q(z) = (z/\sigma)^q + 1 - z$. With $\lambda = n/\sigma^q$, it is sufficient to show that

$$\frac{1}{-(D^Q)'(\beta) \cdot \beta^q} \cdot \beta^{-n} = e^{-\lambda} \cdot \left(1 - \lambda(q-1/2)\sigma^{-q} + O(\sigma^{-2q})\right). \qquad (15)$$

This is proved below.

(ii) For words Q with basic period p, the autocorrelation polynomial has the form $C^Q(z) = 1 + z^p +$(possibly higher terms), and we have $D^Q(z) = (z/\sigma)^q + (1-z)C^Q(z)$. It is sufficient to show that

$$\frac{C^Q(\beta)}{-(D^Q)'(\beta) \cdot \beta^q} \cdot \beta^{-n} = e^{-\lambda} \cdot \Big(1 - \lambda\sigma^{-p} + O(\sigma^{-(p+1)})\Big), \qquad (16)$$

which is proved below.

□

Proof of Equation (15). The first step is to show that

$$\beta = 1 + \sigma^{-q} + q\sigma^{-2q} + O(\sigma^{-3q}).$$

We show constructively that β can be computed by fixed point iteration. By Lemma 3.1, $\beta \in [1,2]$. We rewrite $D^Q(z) = 0$ as the fixed point equation $z = 1 + \sigma^{-q}z^q =: f(z)$. Then $f'(z) = q\sigma^{-q}z^{q-1}$. In the interval $I := [1,2]$, $0 < f'(z) \leq q\,2^{q-1}\,\sigma^{-q} =: \rho_\sigma$. Thus $\rho_\sigma = O(\sigma^{-q}) < 1$ for sufficiently large σ. Furthermore f is strictly increasing on I, and $1 < 1 + \sigma^{-q} = f(1) < f(2) = 1 + (2/\sigma)^q < 2$. Therefore, f is a contraction on I with contraction factor $\rho_\sigma = O(\sigma^{-q})$. By Banach's Fixed Point Theorem (the contraction mapping principle), it follows that the solution of $g(z) = 0$ can be obtained by iterating f, starting with $z_0 = 1$. The error decreases in each step by a factor of ρ_σ, and is at most 1 in the beginning. For $z_3 = f(f(f(1)))$, the error is at most $O(\sigma^{-3q})$, and we find

$$\beta = z_3 + O(\sigma^{-3q}) = 1 + \sigma^{-q} + q\,\sigma^{-2q} + O(\sigma^{-3q}),$$

as claimed.

It remains to evaluate $\frac{1}{-(D^Q)'(\beta)\cdot\beta^q}\cdot\beta^{-n}$. From $(D^Q)'(z) = q\sigma^{-q}z^{q-1} - 1$, one can see that

$$\begin{aligned}
-(D^Q)'(\beta) &= 1 - q\,\sigma^{-q} - q(q-1)\,\sigma^{-2q} + O(\sigma^{-3q}),\\
\beta^q &= 1 + q\,\sigma^{-q} + \left(\frac{q(q-1)}{2} + q^2\right)\sigma^{-2q} + O(\sigma^{-3q}),\\
\beta^{-n} &= e^{-\lambda}\cdot\bigl(1 - \lambda\,(q-1/2)\,\sigma^{-q} + O(\sigma^{-2q})\bigr)\,;
\end{aligned}$$

the latter because $n = \lambda\sigma^q$. The correctness of (15) follows. □

Proof of Equation (16). The proof is similar to the previous one. The first step is to show that

$$\beta = 1 + \sigma^{-q} + \sigma^{-(q+p)} + O(\sigma^{-(q+p+1)});$$

this is done by rewriting $D^Q(z) = 0$ as the fixed point equation

$$z = 1 + \sigma^{-q}\frac{z^q}{C^Q(z/\sigma)} =: f(z).$$

We show that $f(z)$ is a contraction on the interval $I := [1,2]$ for sufficiently large σ with contraction factor $\rho_\sigma = O(\sigma^{-q})$ for large σ. Then $f(f(1))$ is an approximation to β with error $O(\sigma^{-2q})$. Note that f is increasing on I, since $C^Q(z/\sigma)/z^q$ is decreasing for $z \geq 1$. This also establishes $f'(z) > 0$ on I. It is furthermore clear that $f(1) > 1$. Also, $f(2) = 1 + \frac{2^q}{\sigma^q\, c(2/\sigma)} \leq 1 + \frac{2^q}{\sigma^q}$, as $C^Q(z/\sigma) \geq 1$ for $z \geq 0$. Hence $f(2) < 2$ for sufficiently large σ, and f maps I to a subinterval of I. To show that $f' \leq \rho_\sigma < 1$ on I, note that

$$\begin{aligned}
f'(z) &= \sigma^{-q}\cdot\frac{q\,z^{q-1}\,C^Q\bigl(\frac{z}{\sigma}\bigr) - \frac{1}{\sigma}\cdot(C^Q)'\bigl(\frac{z}{\sigma}\bigr)\,z^q}{C^Q\bigl(\frac{z}{\sigma}\bigr)^2}\\
&= \sigma^{-q}\cdot O(1) \;=\; O(\sigma^{-q}),
\end{aligned}$$

as $z = O(1)$, $C^Q\bigl(\frac{z}{\sigma}\bigr) = O(1)$, and $(C^Q)'\bigl(\frac{z}{\sigma}\bigr) = O(\sigma^{-(p-1)})$. Then one computes $f(f(1)) = 1 + \sigma^{-q} - \sigma^{-(q+p)} + O(\sigma^{-(q+p+1)})$, which is the claimed approximation for β.

It remains to prove that (16) holds with this value of β, which involves complex but straightforward algebra; the details are omitted at this point. □

Lemma 3.1 (Smallest Root of $D^Q(z)$). *Let $q \geq 3$ and $\sigma \geq 3$. Let Q be any q-gram over an alphabet of size σ. Then the denominator polynomial of $A^Q(z)$, i.e., $D^Q(z) := (z/\sigma)^q + (1-z)\,C^Q(z)$, has a unique simple real root β of smallest absolute value, and the universal bounds $1 < \beta < 1.0631$ hold for all $\sigma \geq 3$ and $q \geq 3$. All other roots γ fulfill $|\gamma| > 1.6993$.*

We remark that this is a stronger version of Lemma 3 from [4] (where a different notation is used); we provide explicit and universal bounds for β.

Proof. Here is a heuristic argument first: If we show that the roots of $C^Q(z)$ have large absolute values, then $(1-z)\,C^Q(z)$ has indeed a unique simple root of smallest absolute value, namely $z = 1$. At $z = 1$, $(z/\sigma)^q$ is a small quantity and should therefore have little effect when added to $(1-z)\,C^Q(z)$. Therefore $D^Q(z)$ should have a root β with the stated properties.

This argument can be made precise with a special case of Rouché's Theorem (proved in any textbook on complex analysis), applied to the functions $f_1(z) := (1-z)\,C^Q(z)$ and $f_2(z) := (z/\sigma)^q$ such that $f_1(z)+f_2(z) = D^Q(z)$:

Special case of Rouché's Theorem. *Let f_1 and f_2 be analytic functions on the complex plane. Let K be a disc around the origin, and let ∂K be its boundary. If $|f_2| < |f_1|$ on ∂K, then f_1 and f_1+f_2 have the same number of roots inside K.*

Let K_δ be the disc $|z| \le \sigma^\delta$ in the complex plane for $0 < \delta < 1$. Note that all K_δ contain the unit disc. We distinguish two cases, depending on the value of c_1 of the autocorrelation c^Q. In both cases, we prove that there is an interval $\Delta \subset (0,1)$ of values of δ where the following hold: (i) $|C^Q(z)|$ is bounded away from zero in K_δ. Hence, $f_1(z) := (1-z)\,C^Q(z)$ has a unique simple root inside K_δ, namely $z=1$. (ii) On ∂K_δ, $(|z|/\sigma)^q = |f_2(z)| < |f_1(z)|$, and Rouché's Theorem applies. Then we show that the intersection of the Δ sets from both cases is nonempty, and derive a bound for β.

Case $c_1 = 1$. Here Q consists of the repetition of a single letter. Hence $c_j = 1$ for all $j = 0, \ldots, q-1$ and $C^Q(z/\sigma) = \frac{1-(z/\sigma)^q}{1-z/\sigma}$. It follows that in K_δ

$$\left|C^Q\Big(\frac{z}{\sigma}\Big)\right| = \frac{\left|1-\left(\frac{z}{\sigma}\right)^q\right|}{|1-z/\sigma|} \ge \frac{1-\sigma^{-q}|z|^q}{1+\sigma|z|} \ge \frac{1-\sigma^{q(\delta-1)}}{1+\sigma^{\delta-1}} \ge \frac{1-3^{3(\delta-1)}}{1+3^{\delta-1}},$$

which proves (i). To prove (ii), note that on ∂K_δ,

$$|f_1(z)| = |1-z| \cdot \left|C^Q\Big(\frac{z}{\sigma}\Big)\right| \ge (3^\delta - 1)\cdot \frac{1-3^{3(\delta-1)}}{1+3^{\delta-1}},$$

and

$$|f_2(z)| = (|z|/\sigma)^q = \sigma^{q(\delta-1)} \le 3^{3(\delta-1)}.$$

One verifies numerically that for $\delta \in \Delta_1 := [0.05568, 0.73299]$ the inequality $|f_2(z)| < |f_1(z)|$ holds on ∂K_δ.

Case $c_1 = 0$. In this case $C^Q(z/\sigma) = 1 + \sum_{j=2}^{q-1} c_j^Q \cdot (z/\sigma)^j$ with $c_j^Q \in \{0,1\}$. Therefore in K_δ,

$$\left|C^Q\Big(\frac{z}{\sigma}\Big)\right| \ge 1 - \sum_{j=2}^{q-1}\left(\frac{|z|}{\sigma}\right)^j \ge 1 - \sum_{j=2}^{\infty} 3^{j(\delta-1)} = 2 + 3^{\delta-1} - \frac{1}{1-3^{\delta-1}},$$

proving (i). To prove (ii), note that on ∂K_δ,

$$|f_1(z)| = |1-z| \cdot \left|C^Q\Big(\frac{z}{\sigma}\Big)\right| \ge (3^\delta - 1)\cdot\left(2 + 3^{\delta-1} - \frac{1}{1-3^{\delta-1}}\right),$$

and as before $|f_2(z)| \le 3^{3(\delta-1)}$. In this case one verifies numerically that for $\delta \in \Delta_2 := [0.04741, 0.48265]$ indeed $|f_2(z)| < |f_1(z)|$ on ∂K_δ.

The intersection $\Delta := \Delta_1 \cap \Delta_2 = [0.05568, 0.48265]$ is non-empty. Taking $\delta = 0.05568$, we obtain that $|\beta| \le 1.0631$. Furthermore β is real, and it is easy to see that $\beta > 1$. Taking $\delta = 0.48265$, we see that all other roots γ must lie outside the circle $|z| = 3^\delta = 1.6993$. This completes the proof. □

3.2. Expectation

Proof of Theorem 2.3.

(i) For word length $q = 1$, the absence probability of each of the σ words is $(1 - 1/\sigma)^n$. Equation (10) and its asymptotic expansion in σ^{-1} are immediate.

(ii) For word length $q = 2$, there are σ periodic words Q and $\sigma^2 - \sigma$ period-free words Q'. For $\sigma = 2$, $\mathbb{E}[X^{(n,2,2)}]$ is therefore $2 \cdot a^{Q}_{n+1}$ (as given by (7) for σ=2) plus $2 \cdot a^{Q'}_{n+1}$ (as given by (1)). The sum is exactly (11). For $\sigma \geq 3$, $\mathbb{E}[X^{(n,\sigma,2)}]$ is $\sigma \cdot a^{Q}_{n+1}$ (as given by (7)) plus $(\sigma^2 - \sigma) \cdot a^{Q'}_{n+1}$ (as given by (4)). This sum is (12).

(iii) For word length $q \geq 3$, careful bookkeeping is required. Fix q, let $N_p(\sigma)$ be the number of q-grams with basic period p $(1 \leq p < q)$, and let $N_q(\sigma)$ be the number of period-free q-grams. Now $\sum_{p=1}^{q} N_p(\sigma) = \sigma^q$. We show that for each $p = 1, \ldots, q$, $N_p(\sigma) = \sigma^p - O(\sigma^{p-1})$: There are exactly σ words with period 1, namely the words that consist of the repetition of a single letter. For $p > 1$, a word has (not necessarily basic) period p iff the word consists of repetitions of its first p letters. Therefore, there are σ^p possible words with period p. If in addition p is to be the basic period, the first p letters are further constrained by the requirement that no smaller number $p' < p$ must be a period. Therefore, the number of words with basic period $p' < p$ and additional period p must be subtracted from σ^p. By induction, this number is at most $O(\sigma^{p-1})$.

Set $a_{n,p}(\sigma) := a^{Q}_{n+q-1}$ (see (9)), where Q is a q-gram over an alphabet of size σ with basic period p, and set $a_{n,q}(\sigma) := a^{Q}_{n+q-1}$ (see (8)) for a period-free q-gram. Then

$$\begin{aligned}
\mathbb{E}[X^{(n,\sigma,q)}] &= \sum_{p=1}^{q} N_p(\sigma) \cdot a_{n,p}(\sigma) \\
&= \sum_{p=1}^{q-1} N_p(\sigma) \cdot e^{-\lambda} \cdot \Big(1 + \lambda\,\sigma^{-p} + O(\sigma^{-(p+1)})\Big) \\
&\quad + N_q(\sigma) \cdot e^{-\lambda} \cdot \big(1 - \lambda\,(q - 1/2)\,\sigma^{-q} + O(\sigma^{-2q})\big) \\
&= e^{-\lambda} \cdot \left[\left(\sum_{p=1}^{q} N_p(\sigma)\right) + \sum_{p=1}^{q-1} N_p(\sigma)\,(\lambda\,\sigma^{-p} + O(\sigma^{-(p+1)}))\right] \\
&\quad + e^{-\lambda} \cdot N_q(\sigma) \cdot (\lambda\,(q-1/2)\,\sigma^{-q} + O(\sigma^{-2q})) \\
&= e^{-\lambda} \cdot \big(\sigma^q + (q-1)\,\lambda + O(\sigma^{-1}) - (q-1/2)\,\lambda + O(\sigma^{-q})\big) \\
&= e^{-\lambda} \cdot (\sigma^q - \lambda/2 + O(1/\sigma))\,.
\end{aligned}$$

This completes the proof of Theorem 2.3. □

3.3. Variance

In order to prove Theorem 2.5, we need a generalization of the absence probabilities of one word to both of two words.

Definition. Let $Q \neq R$ be q-grams. Let a_m^{QR} be the *absence probability* of both Q and R in a text of length m, i.e., the probability that both Q and R do not occur in a random text of length m.

Definition. Let Q and R be q-grams. Their *correlation vector* $c^{QR} = (c_0^{QR}, \dots, c_{q-1}^{QR}) \in \{0,1\}^q$ is defined by

$$c_i^{QR} := 1 \quad \text{iff} \quad Q[i+j] = R[j] \quad (j = 1, \dots, (q-i)).$$

The *correlation polynomial* $C^{QR}(z)$ is defined by $C^{QR}(z) := \sum_{i=0}^{q-1} c_i^{QR}\, z^i$.

In general $c^{QR} \neq c^{RQ}$. By definition, the autocorrelation of Q is c^{QQ}. For two words, there are four correlation polynomials to consider. These are conveniently summarized in a 2×2 *correlation matrix* $\begin{pmatrix} C^{QQ}(z) & C^{RQ}(z) \\ C^{QR}(z) & C^{RR}(z) \end{pmatrix}$.

Lemma 3.2. *Let $Q \neq R$ be q-grams. Define*

$$\begin{aligned} K^{QR}(z) &:= C^{QQ}(z)\, C^{RR}(z) - C^{QR}(z)\, C^{RQ}(z), &(17)\\ \kappa^{QR}(z) &:= (C^{QQ}(z) + C^{RR}(z)) - (C^{QR}(z) + C^{RQ}(z)). &(18)\end{aligned}$$

Then the generating function $A^{QR}(z)$, such that $a_m^{QR} = [z^m]\, A^{QR}(z)$, is given by

$$A^{QR}(z) = \frac{K^{QR}\big(\frac{z}{\sigma}\big)}{(1-z)\, K^{QR}\big(\frac{z}{\sigma}\big) + \sigma^{-q}\, z^q\, \kappa^{QR}\big(\frac{z}{\sigma}\big)}. \tag{19}$$

As expected, $A^{QR}(z)$ remains unchanged when Q and R are exchanged, i.e., $A^{QR}(z) = A^{RQ}(z)$. The same is true for $K^{QR}(z)$ and $\kappa^{QR}(z)$. A proof of Lemma 3.2 can be found, for example, in [6]. Simple proofs from first principles that include more complicated random text models for the absence probabilities of one or several words are furthermore given in [9].

Proof of Theorem 2.5. The variance of $X^{(n,\sigma,2)}$ can be expressed as

$$\mathrm{Var}[X^{(n,\sigma,2)}] = \mathbb{E}[X^{(n,\sigma,2)}\,(X^{(n,\sigma,2)} - 1)] + \mathbb{E}[X^{(n,\sigma,2)}] - (\mathbb{E}[X^{(n,\sigma,2)}])^2. \tag{20}$$

An asymptotic expansion of $\mathbb{E}[X^{(n,\sigma,2)}]$ up to $O(\sigma^{-3})$ for $\lambda = n/\sigma^2$ is given in Theorem 2.4. From this, one obtains

$$(\mathbb{E}[X^{(n,\sigma,2)}])^2 = e^{-2\lambda} \cdot \left(\sigma^4 - \lambda\,\sigma^2 + \lambda(\lambda-2)\,\sigma + \lambda\left(\frac{\lambda^2}{3} - \frac{5\lambda}{2} + \frac{10}{3}\right)\right) + O(\sigma^{-1}). \tag{21}$$

It remains to find an asymptotic expression for $\mathbb{E}[X^{(n,\sigma,2)}\,(X^{(n,\sigma,2)}-1)] = \sum_{Q \neq R} a_{n+1}^{Q,R}$, where the sum extends over all $\sigma^2\,(\sigma^2-1)$ pairs of unequal 2-grams.

Without loss of generality, we may assume that $\sigma \geq 4$. There are six different types of correlation matrices for 2-grams. (We treat correlation matrices that arise from each other by flipping rows and columns as identical, because $A^{QR}(z)$ is unaffected by this operation).

Table 1 The six types of correlation matrices for pairs of 2-grams. $O(\cdot)$ is a shortcut for $O(\sigma^{-1})$. See the text of the proof of Theorem 2.5 for further information.

Type 1:	$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{array}{rcl} K_1(z) & = & 1, \\ \kappa_1(z) & = & 2, \end{array}$ $\beta_1 = 1 + 2\,\sigma^{-2} + 8\,\sigma^{-4} + 40\,\sigma^{-6} + O(\sigma^{-8})$. $N_1 := \sigma(\sigma-1)^2(\sigma-2)$ pairs share M_1, being $\sigma(\sigma-1)(\sigma-2)(\sigma-3)$ pairs of type (AB, CD), and $\sigma(\sigma-1)(\sigma-2)$ pairs each of types (AB, AC) and (AB, CB). $V_1 = e^{-2\lambda} \cdot \big(\sigma^4 - 4\sigma^3 + (5-6\lambda)\sigma^2 + (-2+24\,\lambda)\sigma + \big(4 - \frac{170}{3}\lambda + 18\lambda^2\big)\big) + O(\cdot)$.
Type 2:	$M_2 = \begin{pmatrix} 1+z & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{array}{rcl} K_2(z) & = & 1+z, \\ \kappa_2(z) & = & 2+z, \end{array}$ $\beta_2 = 1 + 2\,\sigma^{-2} - \sigma^{-3} + 9\,\sigma^{-4} - 11\,\sigma^{-5} + O(\sigma^{-6})$. $N_2 := 2\sigma(\sigma-1)(\sigma-2)$ pairs share M_2, being $\sigma(\sigma-1)(\sigma-2)$ pairs each of types (AA,BC) and (AB,CC). $V_2 = e^{-2\,\lambda} \cdot \big(2\sigma^3 + (2\lambda-6)\sigma^2 + (\lambda^2-20\lambda+4)\sigma + \big(\frac{1}{3}\lambda^3-17\lambda^2+64\lambda-2\big)\big) + O(\cdot)$.
Type 3:	$M_3 = \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}$, $\begin{array}{rcl} K_3(z) & = & 1, \\ \kappa_3(z) & = & 2-z, \end{array}$ $\beta_3 = 1 + 2\,\sigma^{-2} - \sigma^{-3} + 8\,\sigma^{-4} - 10\,\sigma^{-5} + O(\sigma^{-6})$. $N_3 := 2\sigma(\sigma-1)(\sigma-2)$ pairs share M_3, being $\sigma(\sigma-1)(\sigma-2)$ pairs each of types (AB,CA) and (AB,BC). $V_3 = e^{-2\,\lambda} \cdot \big(2\sigma^3 + (2\lambda-6)\sigma^2 + (\lambda^2-18\lambda+4)\sigma + \big(\frac{1}{3}\lambda^3-15\lambda^2+56\lambda-2\big)\big) + O(\cdot)$.
Type 4:	$M_4 = \begin{pmatrix} 1 & 0 \\ z & 1+z \end{pmatrix}$, $\begin{array}{rcl} K_4(z) & = & 1+z, \\ \kappa_4(z) & = & 2, \end{array}$ $\beta_4 = 1 + 2\,\sigma^{-2} - 2\,\sigma^{-3} + 10\,\sigma^{-4} - 22\,\sigma^{-5} + O(\sigma^{-6})$. $N_4 := 4\sigma(\sigma-1)$ pairs share M_4, being $\sigma(\sigma-1)$ pairs each of types (AB,BB), (AA,AB), (AA,BA), and (AB,AA). $V_4 = e^{-2\,\lambda} \cdot \big(4\sigma^2 + (8\lambda-4)\sigma + \big(8\lambda^2-40\lambda\big)\big) + O(\cdot)$.
Type 5:	$M_5 = \begin{pmatrix} 1+z & 0 \\ 0 & 1+z \end{pmatrix}$, $\begin{array}{rcl} K_5(z) & = & (1+z)^2, \\ \kappa_5(z) & = & 2+2z, \end{array}$ $\beta_5 = 1 + 2\,\sigma^{-2} - 2\,\sigma^{-3} + 10\,\sigma^{-4} - 22\,\sigma^{-5} + O(\sigma^{-6})$. $N_5 := \sigma(\sigma-1)$ pairs share M_5, being those of type (AA,BB). $V_5 = e^{-2\,\lambda} \cdot \big(\sigma^2 + (2\lambda-1)\sigma + \big(2\lambda^2-10\lambda\big)\big) + O(\cdot)$.
Type 6:	$M_6 = \begin{pmatrix} 1 & z \\ z & 1 \end{pmatrix}$, $\begin{array}{rcl} K_6(z) & = & 1-z^2, \\ \kappa_6(z) & = & 2-2z, \end{array}$ $\beta_6 = 1 + 2\,\sigma^{-2} - 2\,\sigma^{-3} + 10\,\sigma^{-4} - 22\,\sigma^{-5} + O(\sigma^{-6})$. $N_6 := \sigma(\sigma-1)$ pairs share M_6, being those of type (AB,BA). $V_6 = e^{-2\,\lambda} \cdot \big(\sigma^2 + (2\lambda-1)\sigma + \big(2\lambda^2-10\lambda\big)\big) + O(\cdot)$.

In Table 1, we give an overview over all six types. For each type $j = 1, \ldots, 6$, we show the correlation matrix M_j, the functions $K_j(z)$ (see (17)) and $\kappa_j(z)$ (see (18)), an asymptotic expansion of the unique and simple pole β_j of smallest absolute value of $A_j(z)$ (defined by (19) in terms of $K_j(z)$ and $\kappa_j(z)$). The values of β_j are accurate up to $O(\sigma^{-6})$. We list the different types of 2-gram pairs for each matrix, and the total number N_j of such 2-gram pairs. Finally, we show the total contribution V_j of a word pair of type j to $\mathbb{E}[X^{(n,\sigma,2)}\,(X^{(n,\sigma,2)}-1)]$, accurate up to $O(\sigma^{-1})$. If we define $D_j(z) := (1-z)\,K_j\big(\frac{z}{\sigma}\big)+\sigma^{-q}\,z^q\,\kappa_j\big(\frac{z}{\sigma}\big)$ (the denominator polynomial of $A_j(z)$), then asymptotically

$$V_j = N_j \cdot \frac{K_j(\beta_j/\sigma)}{-D_j'(\beta_j)\,\beta_j^2} \cdot \beta_j^{-\lambda\sigma^2}.$$

Summing over all six cases, we obtain

$$\begin{aligned} \mathbb{E}[X^{(n,\sigma,2)}\,(X^{(n,\sigma,2)}-1)] \;&=\; V_1+V_2+V_3+V_4+V_5+V_6 \\ =\;\; e^{-2\,\lambda}\cdot\left(\sigma^4-(1+2\lambda)\,\sigma^2+2\lambda(\lambda-1)\,\sigma+\lambda\left(\frac{2\lambda^2}{3}-2\lambda+\frac{10}{3}\right)\right)&+O(\sigma^{-1}). \quad (22) \end{aligned}$$

Using (20) and the values from (21), (22), and Theorem 2.4, we obtain the stated asymptotic expansion of $\mathrm{Var}[X^{(n,\sigma,2)}]$. □

4. Open Problems

We have proved a fairly general result about the expected number of missing words (Theorem 2.3) and a special result about the variance for word length $q = 2$ (Theorem 2.5). For general $q \geq 3$, we conjecture that

$$\mathrm{Var}[X^{(n,\sigma,q)}] = \sigma^q\,e^{-2\lambda}\,(e^{\lambda}-\lambda-1)+O(\sigma^{q-1}).$$

For a proof using (20) and (13), the missing piece is to show that

$$\mathbb{E}[X^{(n,\sigma,q)}\,(X^{(n,\sigma,q)}-1)] = e^{-2\lambda}\cdot\left(\sigma^{2q}-(2\lambda+1)\,\sigma^q+O(\sigma^{q-1})\right).$$

The problem is that the structure of the 2×2, and more generally $k \times k$, correlation matrices is not yet fully understood. This is in contrast to single word autocorrelations whose structure has been characterized in [5]. Further results on the combinatorics and on the enumeration of autocorrelations have appeared in [12]. No such results are available for two or more words. Hence, we would like to pose the following problem.

Problem. Characterize and efficiently enumerate 2×2, and more generally, $k \times k$ matrices of correlation vectors between k pairwise different q-grams, and find the number of such matrices. Compute the number of k-tuples of words that share a given correlation matrix.

The law of the number of empty urns $Y^{(n,\sigma)}$ can be given explicitly in terms of the Stirling numbers of the second kind. Furthermore, a central limit theorem for $Y^{(n,\sigma)}$ was proved in [14]. We conjecture that $X^{(n,\sigma,q)}$, properly normalized, also tends to a normal law in the limit.

Conjecture. Let $q \geq 1$ be fixed, and let n and σ tend towards infinity such that the ratio $\lambda := \frac{n}{\sigma^q}$ remains constant. Then

$$\frac{X^{(n,\sigma,q)} - \sigma^q \, e^{-\lambda}}{\sqrt{e^{-2\lambda} \left(e^{\lambda} - (1+\lambda) \right) \sigma^q}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

This could be proved, for example, by showing that all moments of $X^{(n,\sigma,q)}$ and $Y^{(n,\sigma^q)}$ are asymptotically equal. However, this requires first a solution of the above mentioned problem.

References

[1] A. Bolshoy, K. Shapiro, E. N. Trifonov, and I. Ioshikhes. Enhancement of the nucleosomal pattern in sequences of lower complexity. *Nucleic Acids Research*, 25:3248–3254, 1997.

[2] I. Cakir, O. Chryssaohinou, and M. Månsson. On a conjecture by Eriksson concerning overlap in strings. *Combinatorics, Probability and Computing*, 8:429–440, 1999.

[3] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, MA, second edition, 1994.

[4] L. J. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM Journal of Applied Mathematics*, 35(2):401–418, 1978.

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

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

[7] N. L. Johnson and S. Kotz. *Urn Models and Their Applications*. Wiley, New York, 1977.

[8] G. Marsaglia and A. Zaman. Monkey tests for random number generators. *Computers and Mathematics with Applications*, 26(9):1–10, 1993.

[9] S. Rahmann. Word statistics in random texts and applications to computational molecular biology. Diplomarbeit (Master's Thesis), Univerisät Heidelberg, July 2000.

[10] S. Rahmann and E. Rivals. Exact and efficient computation of the expected number of missing and common words in random texts. In D. Sankoff and R. Giancarlo, editors, *Proceedings of the 11th Symposium on Combinatorial Pattern Matching (CPM 2000)*, number 1848 in Lecture Notes in Computer Science, pages 375–387, Berlin, 2000. Springer-Verlag.

[11] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1/2):1–46, 2000.

[12] E. Rivals and S. Rahmann. Combinatorics of periods in strings. In P. Orejas, P. G. Spirakis, and J. van Leuween, editors, *Proceedings of the 28th International Colloquium on Automata, Languages, and Programming (ICALP 2001)*, number 2076 in Lecture Notes in Computer Science, pages 615–626, Berlin, 2001. Springer-Verlag.

[13] E. N. Trifonov. Making sense of the human genome. In R. H. Sarma and M. H. Sarma, editors, *Human Genome Initiative and DNA Recombination*, volume 1 of *Structure and Methods*, pages 69–77. Adenine Press, New York, 1990.

[14] I. Weiss. Limiting distribution in some occupancy problems. *Annals of Mathematical Statistics*, 28:878–884, 1958.