Counting occurrences for a finite set of words: combinatorial methods.

Frédérique Bassino, Julien Clément, Julien Fayolle, Pierre Nicodème

Counting occurrences for a finite set of words: combinatorial methods

FRÉDÉRIQUE BASSINO
LIPN UMR 7030, Université de Paris 13, CNRS, France
JULIEN CLÉMENT
GREYC UMR 6072, Université de Caen, Ensicaen, CNRS, France
JULIEN FAYOLLE
LIPN UMR 7030 , Université de Paris 13, CNRS, France
and
PIERRE NICODÈME
LIX, CNRS-UMR 7161, École polytechnique, Palaiseau, France

In this article, we give the multivariate generating function counting texts according to their length and to the number of occurrences of words from a finite set. The application of the inclusion-exclusion principle to word counting due to Goulden and Jackson (1979, 1983) is used to derive the result. Unlike some other techniques which suppose that the set of words is *reduced* (*i.e.*, where no two words are factor of one another), the finite set can be chosen arbitrarily. Noonan and Zeilberger (1999) already provided a MAPLE package treating the non-reduced case, without giving an expression of the generating function or a detailed proof. We provide a complete proof validating the use of the inclusion-exclusion principle. We also restate in modern terms the normal limit laws theorems of Bender and Kochman (1993), emphasising on the underlying analytic mean shifting method.

Categories and Subject Descriptors: F.2.2. [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems; G.2.1. [**Discrete Mathematics**]: Generating functions, Counting problems

General Terms: Algorithms

Additional Key Words and Phrases: Word Statistics, Inclusion-Exclusion, Generating Functions, Aho-Corasick Automaton

1. INTRODUCTION

Enumerating sequences with given combinatorial properties is rigorously formalized since the end of the seventies and the beginning of the eighties by Goulden and Jackson [Goulden and Jackson 1979; 1983] and by Guibas and Odlyzko [Guibas

Email: Frederique.Bassino@lipn.univ-paris13.fr, Julien.Clement@info.unicaen.fr, Julien.Fayolle@lipn.univ-paris13.fr, nicodeme@lix.polytechnique.fr

and Odlyzko 1981a; 1981b].

The former [Goulden and Jackson 1979; 1983] introduce a very powerful method of inclusion-exclusion to count occurrences of words from a *reduced* set of words (*i.e.*, a set where no word is factor of another word of the set) in texts; this method is characterized by counting texts where some occurrences are marked (other terms are pointed or anchored) and then removing multiple counts of the same text (text counted several times with different markings). We refer later to this by *inclusion-exclusion* method. Goulden-Jackson counting is typically multivariate, a formal parameter being associated to each word.

The latter [Guibas and Odlyzko 1981a; 1981b] introduce the notion of autocorrelation of a word that generalizes to correlation between words, this notion being implicit in Goulden and Jackson. Formal non-ambiguous manipulations over languages translate into generating functions; we refer to this later by *formal language* method. Unlike Goulden and Jackson, Guibas and Odlyzko consider univariate cases, like enumerating sequences avoiding a pattern, or sequences terminating with a first occurrence of a pattern in a text (see also [Sedgewick and Flajolet 1996]). Régnier and Szpankowski [Régnier and Szpankowski 1998] generalize the formal language approach by a bivariate analysis for counting the number of matches of a word in random texts (handling also a Markovian source on the symbol emission) and prove a normal limit law. Régnier [Régnier 2000] extends this further to multivariate analysis and simultaneous counting of several words. See also the books of Szpankowski [Szpankowski 2001] and Lothaire [Lothaire 2005]. Bourdon and Vallée [Bourdon and Vallée 2002; 2006] apply the previous analysis to dynamical sources. Prum *et al.* [Prum et al. 1995], Reinert and Schbath [Reinert and Schbath 1998], Reinert *et al.* [Reinert et al. 2000], and Roquain and Schbath [Roquain and Schbath 2007] follow a more probabilistic approach.

Noonan and Zeilberger [Noonan and Zeilberger 1999] extend the inclusion-exclusion method of Goulden and Jackson and solve the general non-reduced case (words may be factor of other words), implementing corresponding MAPLE programs, without however completely publishing the explicit result formulæ. Recently Kong [Kong 2005] applies the results of Noonan and Zeilberger for the reduced case to an asymmetrical Bernoulli (also called memoryless) model for the generation of symbols. He also compares the Goulden and Jackson method to the Régnier and Szpankowski method, emphasizing the conceptual simplicity of the inclusion-exclusion approach. It is however useful to note that the formal language approach provides access to information that the inclusion-exclusion method does not, such as the waiting time for a first match of a word or the time separating two matches of the same word or of two different words (in both case eventually forbidding matches with other words).

A third approach is possible by use of automata. Nicodème *et al.* [Nicodème et al. 2002] use classical algorithms to (1) build a marked deterministic automaton recognizing a regular expression and (2) translate into generating function (Chomsky-Schützenberger algorithm [Chomsky and Schützenberger 1963]); this provides the bivariate generating function counting the matches. A variation of the method extends the results to Markovian sources. This result applies immediately to a set of words considered as a regular expression. Nicodème [Nicodème 2003] extends

this to multivariate counting by taking the product of marked automata (with an automaton and a mark associated to a word) and to sets of words with possible errors [1]. Notice that, when handling finite languages, step (1) of this approach may be directly done by building the Aho-Corasick automaton, designed for pattern-matching.

Each of the three above-mentioned approaches did develop quite independently and partially unaware of each other.

Let $\mathcal{A}$ be the alphabet on which the words are written and $\mathcal{U} = \{u_1, u_2, \ldots, u_r\}$ be a finite set (or pattern) of distinct words on the alphabet $\mathcal{A}$. We note $\pi(w)$ the weight of the word w. The weight could be a formal weight over the commutative monoid $\mathcal{A}^*$ (*i.e.*, $\pi(ababab) = \alpha^3\beta^3$) or, the probability generating function in the Bernoulli (also called *memoryless*) setting, $\pi(w) = \Pr(w)$ (the probability of w in this model), or even $\pi(w) = 1$ for a uniformly weighted model over all words.

We set some more notations: given a r-row vector $\mathbf{x} = (x_1, \ldots, x_r)$ of formal variables and a r-row vector $\mathbf{j} = (j_1, \ldots, j_r)$ of integers, we will denote by $\mathbf{x}^{\mathbf{j}}$ the product $\prod_{i=1}^{r} x_i^{j_i}$.

In this article we describe two approaches to compute the multivariate generating function $F_{\mathcal{U}}$ counting texts according to their length and to their number of occurrences of words from the set $\mathcal{U}$:

$$F_{\mathcal{U}}(z, \mathbf{x}) = F(z, \mathbf{x}) := \sum_{w \in \mathcal{A}^*} \pi(w) z^{|w|} \mathbf{x}^{\boldsymbol{\tau}(w)}, \tag{1}$$

where $\boldsymbol{\tau}(w) = (|w|_1, \ldots, |w|_r)$, and $|w|_i$ is the total number of occurrences of u_i in w (with possible overlaps). Historically, research on counting occurrences for finite cases considered separately the so-called "reduced" case, which is easier and where no word of the pattern is factor of another word of the pattern; in the opposite or "non-reduced" case, there is no conditions on the pattern. We focus on methods which solve the problem in this latter case (as example the pattern $\mathcal{U}$ can contain $u_1 = abbababa$ and $u_2 = baba$ although u_2 is a factor of u_1). Note that in the non-reduced case, the count of matches that we consider here may exceed the count of positions of the texts at which an occurrence terminates; in contrary, in the reduced case, these two counts are identical. We aim at presenting for the general counting problem a novel approach and a full proof of results partially in Noonan and Zeilberger [Noonan and Zeilberger 1999].

In Section 2 we present an approach using the Aho-Corasick automaton that solves the general (non-reduced) problem; we also consider the complexity of this method. We present in Section 3 of the formal language approach of Régnier and Szpankowski. We describe and prove our results in Section 4 using the inclusion-exclusion principle. Algorithmic aspects are also considered in this section. We handle the asymptotic multivariate limit law for counts in Section 5 providing also an original application to multivariate normal limit laws in the Markovian contexts.

2. AUTOMATON APPROACH

We resort in this section to the well-known Aho-Corasick algorithm [Aho and Corasick 1975; Crochemore and Rytter 2002] which builds from a finite set of words $\mathcal{U}$

[1] Algorithms implemented in the package `regexpcount` of `algolib`, Algorithms Project, INRIA

a (not necessarily minimal) deterministic complete automaton recognizing the language $\mathcal{A}^*\mathcal{U}$. This automaton denoted by $\mathcal{A}_\mathcal{U}$ is the basis of many efficient algorithms on string matching problems and is often called the *string matching automaton*. This automaton is usually described by the trie of the set of words together with a failure function. Let $\mathcal{T}_\mathcal{U}$ be the ordinary trie representing the set $\mathcal{U}$, seen as a finite deterministic automaton $(Q, \delta, \varepsilon, T)$, where the set of states is $Q = \mathrm{Pref}(\mathcal{U})$ (prefixes of words in $\mathcal{U}$), the initial state is ε (denoting ε the empty word), the set of final states is $T = \mathcal{A}^*\mathcal{U} \cap \mathrm{Pref}(\mathcal{U})$ and the transition function δ is defined on $\mathrm{Pref}(\mathcal{U}) \times \mathcal{A}$ by

$$\delta(p, x) = \begin{cases} px & \text{if } px \in \mathrm{Pref}(\mathcal{U}), \\ \mathrm{Border}(px) & \text{otherwise,} \end{cases}$$

where the failure function Border() is defined by

$$\mathrm{Border}(v) = \begin{cases} \text{the longest proper suffix of } v \text{ which belongs to } \mathrm{Pref}(\mathcal{U}) \text{ if defined,} \\ \text{or } \varepsilon \text{ otherwise.} \end{cases}$$

In the following we identify a word $v \in \mathrm{Pref}(\mathcal{U})$ with the node reached by reading the letters of v and following the corresponding transitions on the tree seen as an automaton, so that Border() defines also a map from the set $\mathrm{Pref}(\mathcal{U})$ on the set of nodes of the tree. There are efficient $O(|\mathcal{U}|)$ algorithms [Aho and Corasick 1975; Crochemore and Rytter 2002] linear both in time and space to build such a tree structure and the auxiliary Border() function.

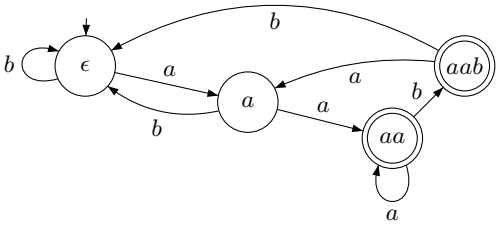
The matrix $\mathbb{T}(\mathbf{x})$ (with $\mathbf{x}$ a r-vector of formal variables) denotes the weighted transition matrix of the Aho-Corasick automaton where the variable x_i marks the states accepting the word u_i. The generating function is expressed as

$$F(z, \mathbf{x}) = \sum_{w \in \mathcal{A}^*} \pi(w) z^{|w|} \mathbf{x}^{\boldsymbol{\tau}(w)} = \begin{pmatrix} 1, & 0, & \cdots, & 0 \end{pmatrix} \left(\mathbb{I} - z\mathbb{T}(\mathbf{x})\right)^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad (2)$$

where $\pi(w)$ can be viewed as the weight of the word w.

Example 1. *Let $\mathcal{U} = \{aab, aa\}$. Ordering the states of the automaton following the lexicographical order, we have, with $\alpha = \pi(a)$ and $\beta = \pi(b)$,*

$$\mathbb{T}(x_1, x_2) = \begin{pmatrix} \beta & \alpha & 0 & 0 \\ \beta & 0 & \alpha x_1 & 0 \\ 0 & 0 & \alpha x_1 & \beta x_2 \\ \beta & \alpha & 0 & 0 \end{pmatrix},$$



and

$$F(z, x_1, x_2) = \frac{1 - \alpha(x_1 - 1)z}{1 - z(\alpha x_1 + \beta - \alpha\beta(x_1 - 1)z + \alpha^2\beta x_1(x_2 - 1)z^2)}.$$

For instance, the coefficient of $[z^n x_1^{n_1} x_2^{n_2}]F(z, x_1, x_2)$ is the probability in the Bernoulli model (if $\alpha + \beta = 1$) of having a text of size n with n_1 occurrences of aa and n_2 occurrences of aab.

Complexity. Let $L = \sum_{u \in \mathcal{U}} |u|$ be the sum of the lengths of the words of $\mathcal{U}$. We first have to compute the Aho-Corasick automaton and this can be done classically in time $O(L)$ for a finite alphabet. The automaton can have up to L states. Denoting by N the number of states of the Aho-Corasick automaton, the transitions matrix $\mathbb{T}$ is of size N^2, but in general this matrix is sparse: only $N \times \operatorname{Card} \mathcal{A}$ entries are non-zero (since the automaton is complete and deterministic with $\operatorname{Card} \mathcal{A}$ transitions from each state).

So the complexity to obtain the counting multivariate generating function by this approach is basically the one of inverting a relatively sparse matrix of the form $\mathbb{I} - z\mathbb{T}(\mathbf{x})$ all terms of which are linear polynomials in z with coefficients that are monomials of the form $\alpha \prod x_i^{\varepsilon_i}$ (with $\alpha = \pi(\ell)$ and $\ell \in \mathcal{A}$ and the ε_i's in $\{0, 1\}$); these coefficients correspond to the transition matrix of the automaton. The limit of this approach is the fact that the size of the transition matrix can grow rapidly if we consider many rather long words. In the two next sections, we adopt other approaches which lead also to solve a system of equations, but then the size of the system is $r \times r$ (where r is the number of words in $\mathcal{U}$).

3. FORMAL LANGUAGES APPROACH

We briefly recall here the Régnier and Szpankowski [Régnier and Szpankowski 1998] approach that is a basis for the analysis by formal languages of reduced sets of words. Considering one word w, Régnier and Szpankowski use a natural parsing or decomposition of texts with at least one occurrence of w, separating unambiguously the texts as follows:

(1) the part of text from the beginning of the text to the first occurrence of the word belongs to the *Right* language,
(2) if there are any other occurrences of the word, each two consecutive occurrences are separated by a text from the *Minimal* language,
(3) the part of text from the last occurrence to the end of the text belongs to the *Ultimate* language.

Moreover, there is a language of texts without any occurrence of the considered word w. Régnier [Régnier 2000] further extended this decomposition approach to a reduced set of words.

We follow here the presentation of Lothaire [Lothaire 2005, Chapter 7].

Definition 3.1. Right, Minimal, Ultimate and Not languages. Let $\mathcal{V} = \{v_1, \ldots, v_r\}$ be a reduced set of words.

– The "Right" language $\mathcal{R}_i$ associated to the word v_i is the set of texts

$$\mathcal{R}_i = \{r \,|\, r = e \cdot v_i \text{ and there is no } v \in \mathcal{V} \text{ such that } r = xvy \text{ with } |y| > 0\}.$$

– The "Minimal" language $\mathcal{M}_{ij}$ leading from a word v_i to a word v_j is the set of texts

$$\mathcal{M}_{ij} = \{m \,|\, v_i{\cdot}m = e{\cdot}v_j \text{ and there is no } v \in \mathcal{V} \text{ such that } v_i{\cdot}m = xvy \text{ with } |x| > 0, |y| > 0\}.$$

– The "Ultimate" language completing a text after an occurrence of the word v_i is the set of texts

$$\mathcal{U}_i = \{u \mid \text{ there is no } v \in \mathcal{V} \text{ such that } v_i \cdot u = xvy \text{ with } |x| > 0\}.$$

– The "Not" language is the set of texts where no word from $\mathcal{V}$ occurs

$$\mathcal{N} = \{n \mid \text{there is no } v \in \mathcal{V} \text{ such that } n = xvy\}.$$

The notations $\mathcal{R}, \mathcal{M}, \mathcal{U}$ and $\mathcal{N}$ refer here to the Right, Minimal, Ultimate and Not languages of a single word.

We consider as example the word $w = ababa$; in the following texts, the underlined words belong to the set $\mathcal{M}$; the overlined text does not since the occurrence represented in bold faces is an intermediate occurrence.

$$ababa\underline{aaaaababa} \qquad ab\mathbf{aba}\overline{\mathbf{ba}bbbababa} \qquad ababa\underline{ba}.$$

Considering the matrix $\mathbb{M}$ such that $\mathbb{M}_{ij} = \mathcal{M}_{ij}$, we have, with $\delta_{ij} = 1$ if and only if $i = j$,

$$\bigcup_{k\geq 1} \left(\mathbb{M}^k\right)_{i,j} = \mathcal{A}^\star \cdot w_j + \mathcal{C}_{ij} - \delta_{ij}\varepsilon, \qquad \mathcal{U}_i \cdot \mathcal{A} = \bigcup_j \mathcal{M}_{ij} + \mathcal{U}_i - \varepsilon, \tag{3}$$

$$\mathcal{A} \cdot \mathcal{R}_j - (\mathcal{R}_j - w_j) = \bigcup_i w_i \mathcal{M}_{ij}, \qquad \mathcal{N} \cdot w_j = \mathcal{R}_j + \bigcup_i \mathcal{R}_i \left(\mathcal{C}_{ij} - \delta_{ij}\varepsilon\right). \tag{4}$$

If the size of the texts is counted by the variable z and the occurrences of the words $v_1, \ldots, v_r$ are counted respectively by $x_1, \ldots, x_r$, we get the matrix equation for the generating function of occurrences

$$F(z, x_1, \ldots, x_r) = \mathcal{N}(z) + (x_1\mathcal{R}_1(z), \ldots, x_r\mathcal{R}_r(z))\big(\mathbb{I} - \mathbb{M}(z, x_1, \ldots, x_r)\big)^{-1} \begin{pmatrix} \mathcal{U}_1(z) \\ \vdots \\ \mathcal{U}_r(z) \end{pmatrix}. \tag{5}$$

In this last equation, we have $\mathbb{M}_{ij}(z, x_1, \ldots, x_r) = x_j\mathcal{M}_{ij}(z)$ and the generating functions $\mathcal{R}_i(z)$, $\mathcal{M}_{ij}(z)$, $\mathcal{U}_j(z)$ and $\mathcal{N}(z)$ can be computed explicitly from the set of Equations (3) and (4).

In particular, when considering the Bernoulli weighted case where the probabilities of the letters sum up to 1 and a single word $w = w_1 w_2 \ldots w_{|w|}$ with $\pi(w) = \Pr(w) = \prod_{i=1}^{|w|} p_{w_i}$, we have the set of equations

$$\mathcal{R}(z) = \frac{\pi(w)z^{|w|}}{D(z)}, \quad \mathcal{M}(z) = 1 + \frac{z-1}{D(z)}, \quad \mathcal{U}(z) = \frac{1}{D(z)}, \quad \mathcal{N}(z) = \frac{\mathcal{C}(z)}{D(z)}, \tag{6}$$

where $D(z) = \pi(w)z^{|w|} + (1-z)\mathcal{C}(z)$. Finally, accordingly to general principles (see [Flajolet and Sedgewick 2009]), we use the combinatorial mapping from $\mathcal{A}^\star = \mathcal{N} + \mathcal{R}\mathcal{M}^\star\mathcal{U}$ to

$$F(z, x) = \frac{1}{1 - z + \pi(w)z^{|w|}\dfrac{1-x}{x + (1-x)\mathcal{C}(z)}} = \sum_{n,k} f_{n,k} x^k z^n. \tag{7}$$

In this last equation, $f_{n,k}$ is the probability that a text of size n has exactly k occurrences of w.

In the case of a single word, Régnier and Szpankowski prove a gaussian limit law when the number of occurrences is $\Theta(n)$, a Poisson-like law when the number of occurrences is $O(1)$ and they provide a large deviation result.

4. INCLUSION-EXCLUSION METHOD APPLIED TO WORD COUNTING

This section presents an approach that follows the same lines as [Goulden and Jackson 1983] but extended to the *non-reduced case*. See also [Noonan and Zeilberger 1999] that provides Maple scripts for this non-reduced case.

We aim to count texts according to their length and to their number of occurrences of words from a set $\mathcal{U}$. A text where some occurrences of words from $\mathcal{U}$ are marked is decomposed combinatorically as a sequence of letters from $\mathcal{A}$ and clusters (set of overlapping and marked occurrences of $\mathcal{U}$, noted $\mathcal{L}_{\mathcal{U}}$; see Definitions (4.2) and (4.3) in the next section). Each text is counted several times depending on which occurrences are marked (each text is counted as many times as the number of possible configurations of marking of occurrences). This multiple counting is eliminated by use of the inclusion-exclusion principle (see among others [Goulden and Jackson 1983], [Szpankowski 2001], and [Flajolet and Sedgewick 2009, III.6.4] for details). This gives an elegant solution to the problem.

4.1 Intuitive approach to counting by inclusion-exclusion

The core of the probabilistic or set theoretic point of view on the inclusion-exclusion principle relies on the equality (for instance for sets A and B)

$$\mathrm{Card}(A \cup B) = \mathrm{Card}(A) + \mathrm{Card}(B) - \mathrm{Card}(A \cap B),$$

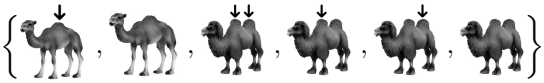
which can be further generalized into an alternative sum for a family $(A_i)_{1 \le i \le r}$ of subsets. However this formulation is sometimes difficult (although of course correct) when considering complex examples.

In this paper we use a symbolic alternative based on multivariate generating functions which is technically easier.

Camelus Genus. We consider an elementary example illustrating the symbolic exclusion-inclusion method: let $\mathcal{P}$ be the set of Camelus Genus (that is camel and dromedary). Each one is of size one (we count individuals), and we want to count the number of humps (using a formal variable u in the generating function). In a very standard manner (see [Flajolet and Sedgewick 2009]), we get

$\mathcal{P} =$  $, \qquad P(z,u) = z(u+u^2).$

Now consider the distinguished set $\mathcal{Q}$ defined as the set of *objects of $\mathcal{P}$ in which each elementary configuration (hump) is either distinguished or not* (we mark humps with variable v in the generating function)

$Q =$ 

$$Q(z,v) = z(v+1+v^2+v+v+1) = z(2+3v+v^2) = P(z,1+v).$$

Fig. 1. Two clusters of a text $aaaaa$ with respect to the word aaa, corresponding respectively to decompositions $aaa \cdot a \cdot a$ (left) and $aaa \cdot aa$ (right). Distinguished occurrences are represented with black rules.

So if $Q(z, v)$ is easy to obtain, then we have $P(z, u) = Q(z, u - 1)$ by a simple substitution $v \mapsto u - 1$. Hence the symbolic inclusion-exclusion principle can be summarized in the following way: counting objects that contain an exact number of pattern occurrences is reduced to counting objects that contains the pattern at distinguished places (the latter being usually a simpler problem).

Clusters: one word example. To make things more precise, let us consider the case of counting occurrences of one word w. One can look at all possibilities of distinguishing occurrences of a given text. Note that occurrences occur in "clusters" (a term which be made more precise later), that is occurrences of w which overlap. To build a cluster, we first distinguish an occurrence (the first one), and from this occurrence, add any non empty word from the autocorrelation set to get the next occurrence (overlapping the first one), then we can repeat this process again and again. Note that the same underlying word can give rise to several such clusters. For instance if $w = aaa$, the word $aaaaa$ can be decomposed as $aaa \cdot a \cdot a$, or $aaa \cdot aa$ which are both valid.

More formally, a text with distinguished occurrences can be described according to the combinatorial description (where Seq denotes a sequence of object of finite length greater or equal to zero)

$$\mathcal{Q} = \textsc{Seq}(\mathcal{A}) \cdot \textsc{Seq}\big(w \cdot \textsc{Seq}\,(\mathcal{C} \setminus \varepsilon) \cdot \textsc{Seq}(\mathcal{A})\big).$$

The expression $w \cdot \textsc{Seq}\,(\mathcal{C} \setminus \varepsilon)$ describes the structure of a cluster of overlapping occurrences. This specification is translated thanks to general principles ($\textsc{Seq}(\mathcal{L}) \mapsto (1 - L(z, v))^{-1}$ for instance) and yields the bivariate generating function

$$\begin{aligned} Q(z, v) &= \sum_{\text{Texts } t} \sum_{\substack{\text{all configurations of} \\ \text{distinguished occurrences of } w \text{ in } t}} \pi(t) z^{|t|} v^{\#\text{ distinguished occurrences}} \\ &= \frac{1}{1 - z\pi(\mathcal{A})} \frac{1}{1 - \pi(w) v z^{|w|} \dfrac{1}{1 - v(C(z) - 1)} \dfrac{1}{1 - z\pi(\mathcal{A})}} \\ &= \frac{1}{1 - z\pi(\mathcal{A}) - \dfrac{v\pi(w) z^{|w|}}{1 - v(C(z) - 1)}}. \end{aligned}$$

Then thanks to the symbolic exclusion-inclusion principle, denoting $|t|_w$ the number of occurrences of w in t, we get directly

$$P(z, u) = \sum_{t \in \mathcal{A}^*} \pi(t) z^{|t|} u^{|t|_w} = Q(z, u - 1).$$

Example 2. *Consider the word $w = aaa$ and the binary alphabet $\mathcal{A} = \{a, b\}$,*

and pose $\pi(a) = \pi(b) = 1$ *(so that we get the enumerative generating function, i.e., any word has weight* 1*). Then* $C(z) = 1 + z + z^2$*, and one gets*

$$Q(z, v) = \cfrac{1}{1 - 2z - \cfrac{vz^3}{1 - v(z + z^2)}}, \qquad P(z, u) = Q(z, u - 1).$$

Our goal is to generalize this process to any finite set of words.

4.2 Preliminary

We formally state the generating function in terms of occurrence positions.

Definition 4.1 *Occurrence positions set.* The occurrence positions set of a word u in a (longer) word w is the set of final positions of occurrences of u in w (indices start at 1 in w):

$$\mathrm{Occ}(u, w) = \big\{p \in \{|u|, |u| + 1, \ldots, |w|\} \quad \big| \quad w[(p-|u|+1) \ldots p] = u\big\}.$$

With this definition, we can rewrite the counting generating function of Equation (1) on p. 3

$$F(z, \mathbf{x}) = \sum_{w \in \mathcal{A}^*} \pi(w) z^{|w|} \prod_{i=1}^{r} x_i^{\mathrm{Card}(\mathrm{Occ}(u_i, w))}.$$

Note that the occurrences of two distinct words ending at the same position are both counted. We need here the following definitions.

Definition 4.2 *Clustering-word.* A clustering-word for the set $\mathcal{U} = \{u_1, \ldots, u_r\}$ is a word $w \in \mathcal{A}^*$ such that any two consecutive positions in w are covered by the same occurrence in w of a word $u \in \mathcal{U}$. The position i of the word w is covered by a word u if $u = w[(j - |u| + 1) \ldots j]$ for some $j \in \{|u|, \ldots, |w|\}$ and $j - |u| + 1 \leq i \leq j$. The language of all clustering-words for a given set $\mathcal{U}$ is noted $\mathcal{K}_{\mathcal{U}}$.

Definition 4.3 *Cluster.* A cluster of a clustering-word w in $\mathcal{K}_{\mathcal{U}}$ is a set of words with their occurrence positions $\{\, \mathcal{S}_u \subset \mathrm{Occ}(u, w) \mid u \in \mathcal{U} \,\}$ which covers exactly w; a cluster is therefore a subset of the occurrences in the text, with their positions. Moreover, every two consecutive positions i and $i + 1$ in w are covered by at least one same occurrence of some $u \in \mathcal{U}$. More formally

$$\forall i \in \{1, \ldots, |w| - 1\} \quad \exists\, u \in \mathcal{U}, \exists\, p \in \mathcal{S}_u \quad \text{such that} \quad p - |u| + 1 \leq i < i + 1 \leq p.$$

The set of clusters with respect to clustering-words built from some finite set of words $\mathcal{U}$ is noted $\mathcal{L}_{\mathcal{U}}$. We note $\mathcal{L}_{\mathcal{U}}(w)$ the subset of $\mathcal{L}_{\mathcal{U}}$ corresponding to the clustering-word $w \in \mathcal{K}_{\mathcal{U}}$. For a cluster $\mathfrak{C} = \{\mathcal{S}_u \mid u \in \mathcal{U}\}$, we also define $w(\mathfrak{C})$ the corresponding (unique) clustering-word and $|\mathfrak{C}|_u$ the number of marked occurrences of the word u in the cluster, *i.e.*,

$$|\mathfrak{C}|_u = \mathrm{Card}\, \mathcal{S}_u.$$

Example 3. *Let* $\mathcal{U} = \{baba, ab\}$ *and* $w = abababa$*, so that* $w \in \mathcal{K}_{\mathcal{U}}$*. We have*

$$\mathcal{L}_{\mathcal{U}}(w) = \Big\{ \big\{\mathcal{S}_{ab} = \{2, 4, 6\}, \mathcal{S}_{baba} = \{5, 7\}\big\}, \big\{\mathcal{S}_{ab} = \{2, 6\}, \mathcal{S}_{baba} = \{5, 7\}\big\}, \\ \big\{\mathcal{S}_{ab} = \{2, 4\}, \mathcal{S}_{baba} = \{5, 7\}\big\}, \big\{\mathcal{S}_{ab} = \{2\}, \mathcal{S}_{baba} = \{5, 7\}\big\} \Big\}.$$

Definition 4.4 *Factor relation for occurrences.* An occurrence q_u of a word u is factor of an occurrence $q_{u'}$ of a word u' in a cluster $\mathfrak{C}$ if

—the word u is a factor of the word u'

—and denoting $\mathrm{pos}(q_v)$ the position of the occurrence q_v of the word v in the clustering-word

$$0 \leq \mathrm{pos}(q_{u'}) - \mathrm{pos}(q_u) \leq |u'| - |u|$$

.

Definition 4.5 *Skeleton of a cluster or reduced cluster.* Given a cluster $\mathfrak{C}$, let $Q = \{q_1, \ldots, q_j\}$ be the set of occurrences of words of the cluster. The factor relation induces a partial order on the set Q:

$$q_x \prec q_y \text{ if } q_x \text{ is factor of } q_y \qquad (q_x, q_y \in Q).$$

Comparing by pairs all the elements of Q, we can remove any element that is smaller than another one; this gives a reduced set of occurrences $\underline{Q}$, that defines a *reduced cluster* of the original cluster. Then one has

LEMMA 4.6. *Given a cluster $\mathfrak{C}$ with clustering word w, let $\underline{\mathfrak{C}}$ be a reduced cluster as described in Definition (4.5). We have the following properties.*

(*1*) *$\underline{\mathfrak{C}}$ is uniquely defined.*

(*2*) *The clustering word of $\underline{\mathfrak{C}}$ is w.*

(*3*) *The occurrences $\{w_1, \ldots, w_k\}$ in $\underline{\mathfrak{C}}$ can be increasingly ordered with respect to their positions such that each occurrence overlap the following one, when it exists. This ordering is unique.*

PROOF. The removal process described in Definition (4.5) is independent of the order along which the pairs have been taken. This proves unicity. To prove the second assertion, let us start from the cluster $\mathfrak{C}$ and suppose that this cluster is not reduced. When removing the first factor occurrence u_{i_1}, we know that there exists a word u_x such that $u_{i_1} \prec u_x$ or equivalently that u_{i_1} is factor of u_x. This implies first that removing u_{i_1} cannot disconnect the cluster and second that if the last letter of u_x is the last letter of w or if the first letter of u_x is the first letter of w, this will remain unchanged after removing u_{i_1}. Therefore the clustering-word of the cluster obtained after removing u_{i_1} is w. Iterating this reasoning proves the second assertion since a cluster contains a finite number of words. The last assertion is proved by removing iteratively the leading occurrence of the cluster and numbering the occurrences correspondingly; the unicity in the choice of the leading term comes from the fact there are no factor occurrences. □

It will be important in the following that the reduced cluster is *unique*. We will call also this cluster *skeleton* of the original cluster.

EXAMPLE 4. *In Example 3, the cluster $\mathfrak{C} = \{\mathcal{S}_{ab} = \{2,4,6\}, \mathcal{S}_{baba} = \{5,7\}\}$ gives rise to the reduced cluster $\underline{\mathfrak{C}} = \{\mathcal{S}_{ab} = \{2\}, \mathcal{S}_{baba} = \{5,7\}\}$. Note that in this particular case $\underline{\mathfrak{C}}$ is the skeleton of all clusters of $\mathcal{L}_{\mathcal{U}}(w)$.*

In the non-reduced case, a word u_i may occur within some other word from $\mathcal{U}$. In order to properly generate the clusters we introduce the notion of *right extension* of a pair of words (h_1, h_2). This notion is a generalization of the correlation set of two words h_1 and h_2 but differs in that:

(i) overlapping is not allowed to occur at the beginning of h_1.

(ii) extension has to add some letters to the right of h_1.

More formally we have

Definition 4.7 *Right extension set.* The right extension set of a pair of words (h_1, h_2) is

$$\mathcal{E}_{h_1,h_2} = \{\, e \quad | \quad \text{there exists } e' \in \mathcal{A}^+ \quad \text{such that} \quad h_1 e = e' h_2 \quad \text{with } 0 < |e| < |h_2| \}.$$

Note that, when h_1 and h_2 have no factor relation, the right extension set $\mathcal{E}_{h_1,h_2}$ is the correlation set of h_1 to h_2. Moreover, when $h_1 = h_2$, the set $\mathcal{E}_{h_1,h_2}$ is the strict autocorrelation set of h_1 (the empty word does not belong to $\mathcal{E}_{h_1,h_2}$).

One can also define the right extension matrix of a vector of words $\mathbf{u} = (u_1, \ldots, u_r)$

$$\mathcal{E} = \left(\mathcal{E}_{u_i,u_j}\right)_{1 \le i,j \le r}.$$

Example 5. *We give some examples and their right extension matrices.*

(1) For $\mathbf{u} = (ab, aba)$*, we have* $\mathcal{E} = \begin{pmatrix} \emptyset & \emptyset \\ b & ba \end{pmatrix}$.

(2) For $\mathbf{u} = (a^3 = aaa, a^7 = aaaaaaa)$*, we have*

$$\mathcal{E} = \begin{pmatrix} a + aa & a^5 + a^6 \\ a + aa & a + a^2 + a^3 + a^4 + a^5 + a^6 \end{pmatrix}.$$

(3) For $\mathbf{u} = (aa, ab, ba, baaab)$*, we have* $\mathcal{E} = \begin{pmatrix} a & b & \emptyset & \emptyset \\ \emptyset & \emptyset & a & aaab \\ a & b & \emptyset & \emptyset \\ \emptyset & \emptyset & a & aaab \end{pmatrix}$.

4.3 Generating function of clusters

We define the generating function $\xi(z, \mathbf{t})$ of the set of clusters $\mathcal{L}_{\mathcal{U}}$ on $\mathcal{U}$ where the length of the corresponding clustering word is marked by the formal variable z and each marked occurrence of u_i in clusters is marked by the formal variable t_i. The set of all possible clusters is the disjoint union over all clustering-words w of the set of all the clusters built from w, hence

$$\xi(z, \mathbf{t}) = \sum_{w \in \mathcal{K}_{\mathcal{U}}} \sum_{\mathfrak{C} \in \mathcal{L}_{\mathcal{U}}(w)} z^{|w|} \pi(w) t_1^{|\mathfrak{C}|_{u_1}} \ldots t_r^{|\mathfrak{C}|_{u_r}}.$$

4.3.1 *Basic decomposition.* We establish a bijection between clusters and paths in a graph to derive an expression for the generating function $\xi(z, \mathbf{t})$ of clusters in $\mathcal{L}_{\mathcal{U}}$.

Definition 4.8 *Right extension graph.* Let $\mathcal{U} = \{u_1, \ldots, u_r\}$; the right extension graph $\mathcal{G}_{\mathcal{U}} = (V, E)$ of the set of words $\mathcal{U}$ is the directed labeled graph such that:

(a) the set of vertices is $V = \{\varepsilon\} \cup \mathcal{U}$;

(b) the set of edges is $E = \{\varepsilon \xrightarrow{u} u \mid u \in \mathcal{U}\} \cup \{u \xrightarrow{y} u' \mid u, u' \in \mathcal{U} \text{ and } y \in \mathcal{E}(u, u')\}$.

See an example on Figure 2 with $\mathcal{U} = \{aa, ab, ba, baaab\}$.

We now prove a key bijection between clusters and "decorated" paths in the extension graph.

THEOREM 4.9. *There exists a bijection between the set of clusters $\mathfrak{C}$ and the set of pairs $(c, \mathcal{F}_c)$ where c is a path in $\mathcal{G}$ (starting at ε) and $\mathcal{F}_c$ is a k-tuple (k is the length of the path c in terms of number of edges traversed) of sets of occurrence positions.*

PROOF. If the set $\mathcal{U}$ is reduced (*i.e.*, without factor relations) then a cluster is completely described by a path in this graph starting at ε. When the set is not reduced, this is no longer true. We need to associate along the path the possible occurrences of $\mathcal{U}$ within the last label read.

We partition the set of clusters with respect to the set of reduced clusters. For a given reduced cluster $\mathfrak{R}$, let

$$\mathfrak{G}_{\mathfrak{R}} = \{\mathfrak{C}; \quad \underline{\mathfrak{C}} = \mathfrak{R}\},$$

be the set of clusters having reduced cluster $\mathfrak{R}$.

For each reduced cluster $\mathfrak{R}$, we know by the last assertion of Lemma (4.6) that there is a unique ordering of occurrences $(w_1, \ldots, w_k) \in \mathcal{U}^k$. The same assertion states that for each w_i with $i \in \{1, \ldots, k-1\}$ there exists v and y_{i+1} with $|v| > 0$ and $|y_{i+1}| > 0$ such that $w_i \cdot y_{i+1} = v \cdot w_{i+1}$. This corresponds to a unique path $\varepsilon \xrightarrow{y_1} w_1 \xrightarrow{y_2} w_2 \xrightarrow{y_3} \ldots \xrightarrow{y_k} w_k$ in the right extension graph, which implies an injection from the set of reduced clusters into the set of paths of the right extension graph.

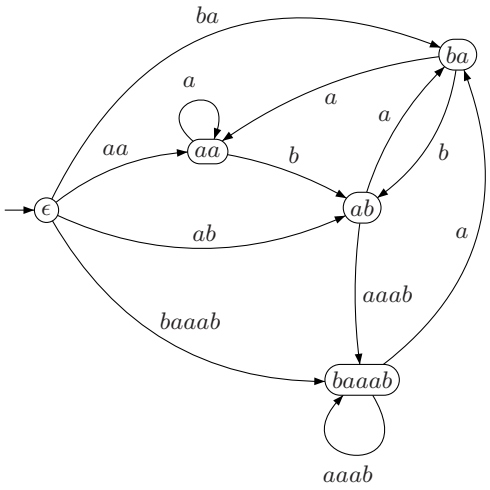
Reciprocally, considering any path $c = y_1 \cdot y_2 \ldots y_k$ in the graph $\mathcal{G}$

$$\varepsilon \xrightarrow{y_1} w_1 \xrightarrow{y_2} w_2 \xrightarrow{y_3} \ldots \xrightarrow{y_k} w_k,$$

where $w_i \in \mathcal{U}$, so that while reading transition y_j we get an occurrence of a word of $\mathcal{U}$ together with its position (which is just $\sum_{i<j} |y_j|$); this follows from the definition of the right extension sets, all transitions y_j, but y_1, being such extensions and y_1 being the occurrence u_{i_1}. Therefore any path c corresponds to a unique reduced cluster $\mathfrak{R}$.

We consider now the factor occurrences. Each set in $\mathcal{F}_c$ is composed of positions of occurrences of words from $\mathcal{U}$ that end within the label of the corresponding edge of the path. If $\mathfrak{R}$ is a reduced cluster where there are h potential or *flip-flop* marked factor occurrences, there are $|\mathfrak{G}_{\mathfrak{R}}| = 2^h$ possible configurations, each one being unique. There are also h possible factor occurrences induced by right extensions when going through the path c corresponding to $\mathfrak{R}$; this gives 2^h different configurations. Going through the edges of the path c provides also the positions of the factor occurrences, and therefore to each pair $(c, \mathcal{F}_c)$ corresponds a cluster $\mathfrak{C} \in \mathfrak{G}_{\mathfrak{R}}$. This completes the proof. □

It remains to translate this combinatorial description of clusters to a formal description in terms of generating functions. We introduce hereafter two notations



$$\mathcal{E} = \begin{pmatrix} a & b & \emptyset & \emptyset \\ \emptyset & \emptyset & a & aaab \\ a & b & \emptyset & \emptyset \\ \emptyset & \emptyset & a & aaab \end{pmatrix}, \quad \mathfrak{E} = \begin{pmatrix} z\pi(a) & z\pi(b) & 0 & 0 \\ 0 & 0 & z\pi(a) & z^4\pi(aaab)(1+t_{ba})(1+t_{aa})^2(1+t_{ab}) \\ z\pi(a) & z\pi(b) & 0 & 0 \\ 0 & 0 & z\pi(a) & z^4\pi(aaab)(1+t_{ba})(1+t_{aa})^2(1+t_{ab}) \end{pmatrix}.$$

Fig. 2. Graph $\mathcal{G}$ for $\mathcal{U} = \{aa, ab, ba, baaab\}$ (top). On the bottom, the right extension matrix $\mathcal{E}$ (rows and columns are ordered as in $aa, ab, ba, baaab$). We remark that several paths may correspond to the same labelling. For instance the word $baaab$ corresponds to different paths

$$\varepsilon \xrightarrow{baaab} baaab, \quad \text{and} \quad \varepsilon \xrightarrow{ba} ba \xrightarrow{a} aa \xrightarrow{a} aa \xrightarrow{b} ab$$

with different skeletons.

$\mathfrak{u}^\circ$ and $\mathfrak{E}$ in order to express the formal equivalent of $\mathbf{u}$ and $\mathcal{E}$. Basically, when a word u_m is factor of a word u_i, it may be marked or remains unmarked during the marking process; this induces a term $(1 + t_m)$ in the associated multivariate generating function.

Definition 4.10 *Inclusion-exclusion formal marking.* Let $\mathbf{u} = (u_1, \ldots, u_r)$ be a vector of r words, the formal marking of $\mathbf{u}$ denoted by $\mathfrak{u}^\circ$ has its i-th coordinate element equal to

$$\mathfrak{u}_i^\circ = \pi(u_i) z^{|u_i|} \prod_{m \neq i} (1 + t_m)^{|u_i|_m},$$

where $|u_i|_m$ is the number of occurrences of the word u_m in the word u_i. The formal marking of the matrix of right extension sets $\mathcal{E}$ is the matrix $\mathfrak{E}$, where the element of indices (i, j) is

$$\mathfrak{E}_{i,j} = \sum_{e \in \mathcal{E}_{i,j}} \pi(e) z^{|e|} \prod_{m \neq j} (1 + t_m)^{\min(|u_i e|_m - |u_i|_m, |u_j|_m)}. \tag{8}$$

There is a need of two different definitions since when considering the right extension matrix we are only counting factor occurrences that begin and finish within the last occurrence considered (u_j in the definition) and also finish inside the extension; in contrary, when considering a single word u we consider *all* factor occurrences, and we note this particular case by the symbol $\circ$ in exponent.

EXAMPLE 6. *We develop further the Example 5 p.11 by taking* $\pi(a) = \pi(b) = 1$.

(1) For $\mathbf{u} = (ab, aba)$, *we have* $\mathfrak{E} = \begin{pmatrix} 0 & 0 \\ z & z^2(1+t_1) \end{pmatrix}$.

(2) For $\mathbf{u} = (a^3 = aaa, a^7 = aaaaaaa)$ *(see Figure 2), the matrix* $\mathfrak{E}$ *is equal to*

$$\begin{pmatrix} z+z^2 & (1+t_1)^5(z^5+z^6) \\ z+z^2 & (1+t_1)z + (1+t_1)^2z^2 + (1+t_1)^3z^3 + (1+t_1)^4z^4 + (1+t_1)^5(z^5+z^6) \end{pmatrix}.$$

Here the crucial point is to only consider which factor occurrences are within the last occurrence read (hence the min *operator in Equation (8)) and finish inside the extension.*

(3) For $\mathbf{u} = (aa, ab, ba, baaab)$, *we have*

$$\mathfrak{E} = \begin{pmatrix} z & z & 0 & 0 \\ 0 & 0 & z & z^4(1+t_1)^2(1+t_2)(1+t_3) \\ z & z & 0 & 0 \\ 0 & 0 & z & z^4(1+t_1)^2(1+t_2)(1+t_3) \end{pmatrix}.$$

With these notations, we get to the following proposition.

PROPOSITION 4.11. *The generating function* $\xi(z, \mathbf{t})$ *of clusters built from the set* $\mathcal{U} = \{u_1, \ldots, u_r\}$ *is given by*

$$\xi(z, \mathbf{t}) = \mathfrak{u}^{\circ}\Delta(\mathbf{t}) \cdot \Big(\mathbb{I} - \mathfrak{E}\Delta(\mathbf{t})\Big)^{-1} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{9}$$

where $\mathbf{u} = (u_1, \ldots, u_r)$, $\mathbf{t} = (t_1, \ldots, t_r)$, *and the matrix* $\Delta(\mathbf{t})$ *is the* $r \times r$ *diagonal matrix with entries* $t_1, \ldots, t_r$.

PROOF. The matrix $\mathfrak{E}$ is the formal expression of the transition matrix of the graph $\mathcal{G}$ where the vertex ε and its corresponding edges have been removed. Some occurrences of the word u_i (for each $i \in \{1, \ldots, n\}$) are marked with the formal variables t_i in the labels of $\mathcal{G}$. More precisely, a word occurrence u_i obtained when visiting a vertex u_i is marked by the formal variable t_i (and appears in the calculus through the diagonal matrix $\Delta(\mathbf{t})$ in (9)); in contrary, a factor occurrence can be marked or not (this does not change the path in the graph), hence providing a term of the form $\prod_{m \neq i}(t_m + 1)^k$ where k is the number of possible new occurrences. The first transition from ε to any $u \in \mathcal{U}$ is handled similarly. So the paths with $k+1$ transitions in $\mathcal{G}$ starting from ε have generating function

$$\mathfrak{u}^{\circ}\Delta(\mathbf{t}) \cdot \Big(\mathfrak{E}\Delta(\mathbf{t})\Big)^{k} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Finally we use the quasi-inverse notation $\sum_{j=0}^{\infty} \big(\mathfrak{E}\Delta(\mathbf{t})\big)^j = \big(\mathbb{I} - \mathfrak{E}\Delta(\mathbf{t})\big)^{-1}$ to get the result. □

4.3.2 *Applications.* We examine special cases where we apply Proposition 4.11.

Reduced set. When the set $\mathcal{U}$ is reduced, that is, no word of $\mathcal{U}$ is factor of another, the clusters are uniquely defined by a path in the previous graph $\mathcal{G}$. So $\mathfrak{u}^\circ$ and $\mathfrak{E}$ do not depend on any of the variables t_i's. Hence in Equation (9), the variables t_i's appear only inside $\Delta(\mathbf{t})$. This is another formulation of the result of Goulden and Jackson [Goulden and Jackson 1983].

One word. For $\mathcal{U} = \{u\}$, we get

$$\xi(z,t) = \frac{t\mathfrak{u}}{1-t\mathfrak{E}} = \frac{t\pi(u)z^{|u|}}{1-t\widehat{c}(z)} = \frac{t\pi(u)z^{|u|}}{1-t(c(z)-1)}, \tag{10}$$

where $c(z)$ and $\widehat{c}(z)$ respectively are the autocorrelation polynomial and the strict autocorrelation polynomial (empty word ε omitted) of u.

Two words. For a set of two words $\{u_1, u_2\}$, we can compute explicitly $\xi(z, t_1, t_2)$ by the Cramer's rule,

$$\xi(z,t_1,t_2) = \frac{t_1\mathfrak{u}_1 + t_2\mathfrak{u}_2 - t_1t_2\big(\mathfrak{u}_1\big[\mathfrak{E}_{2,2} - \mathfrak{E}_{1,2}\big] + \mathfrak{u}_2\big[\mathfrak{E}_{1,1} - \mathfrak{E}_{2,1}\big]\big)}{1 - t_2\mathfrak{E}_{2,2} - t_1\mathfrak{E}_{1,1} + t_1t_2\big(\mathfrak{E}_{1,1}\mathfrak{E}_{2,2} - \mathfrak{E}_{2,1}\mathfrak{E}_{1,2}\big)}, \tag{11}$$

and this expression is computable from the right extension matrix of $\{u_1, u_2\}$.

EXAMPLE 7. *Let* $\mathbf{u} = (a^3, a^7)$. *Recall the right extension matrix (see Example 6(2)) is:*

$$\mathcal{E} = \begin{pmatrix} a + a^2 & a^5 + a^6 \\ a + a^2 & a + a^2 + a^3 + a^4 + a^5 + a^6 \end{pmatrix}.$$

If we consider $\pi(w) = 1$ *for all words* w *(the unweighted "enumerative" model where each word has weight 1), we have* $\mathfrak{u}^\circ = \big(z^3, (1+t_1)^5 z^7\big)$*;*

$$\begin{aligned}
\mathfrak{E}_{1,1} =& z + z^2, \quad \mathfrak{E}_{1,2} = (1+t_1)^5(z^5+z^6), \quad \mathfrak{E}_{2,1} = z + z^2,\\
\mathfrak{E}_{2,2} =& (1+t_1)z + (1+t_1)^2z^2 + (1+t_1)^3z^3 + (1+t_1)^4z^4\\
&+ (1+t_1)^5(z^5+z^6),
\end{aligned}$$

By substituting these values in Equation (11), the generating function $\xi(z, t_1, t_2)$ *can be written as*

$$\frac{z^3\left(t_2\,(1+t_1)^4\,z^4 - t_2t_1\,(1+t_1)\,z^3 - t_2t_1\,(1+t_1)^2\,z^2 - t_2t_1\,(1+t_1)\,z + t_1\right)}{1 - z^3t_2(1+t_1)\left((1+t_1)^3\,z^3 + \left({t_1}^3 + 3\,{t_1}^2 + 2\,t_1 + 1\right)z^2 + \left({t_1}^3 + 2\,{t_1}^2 + t_1 + 1\right)z + 1 - {t_1}^2 - 2\,t_1\right) - (t_2t_1 + t_2 + t_1)\,(z^2+z).}.$$

4.4 Generating function of texts

A text is decomposed combinatorically as a sequence of letters from $\mathcal{A}$ (with generating function $A(z)$) and clusters (or more rigorously of clustering words) from $\mathcal{L}_\mathcal{U}$ (with generating function $\xi(z, \mathbf{t})$). The multivariate generating function $F(z, \mathbf{x})$ of Equation (1) p.3 is derived by substituting $t_i \mapsto x_i - 1$ for $i \in \{1, \ldots, r\}$ in each $(A(z) + \xi(z, \mathbf{t}))^k$, where k is the number of combinatorial objects in the decomposition.

To summarize, we have the following theorem.

THEOREM 4.12. *Let* $\mathbf{u} = (u_1, \ldots, u_r)$ *be a finite vector of words in* $\mathcal{A}^*$ *and* $\mathcal{E}$ *the associated right extension matrix. The multivariate generating function* $F(z, \mathbf{x})$

counting texts the length of which is counted by the variable z and where the occurrences of u_i are counted by the vector of formal variables $\mathbf{x} = (x_1, \ldots, x_r)$ is

$$F(z, \mathbf{x}) = \frac{1}{1 - A(z) - \xi(z, \mathbf{x} - \mathbf{1})}, \tag{12}$$

where $A(z) = \sum_{\sigma \in \mathcal{A}} \pi(\sigma) z$ is the generating function of the alphabet and $\xi(z, \mathbf{t})$ is defined in Equation (9).

Proof. The proof relies on two main points. On one hand, the generating function $\xi(z, \mathbf{t})$ counts all the clusters (see Proposition 4.11 in Section 4.3.1). On the other hand, the inclusion-exclusion principle yields the final result by the substitutions $t_i \mapsto x_i - 1$. □

The application of the standard techniques of analytic combinatorics (see [Flajolet and Sedgewick 2009]) to the multivariate generating function $F(z, \mathbf{x})$ gives access to many statistics (*e.g.* mean, variance, covariance...).

4.5 Algorithmic construction of the Right Extension Sets

We present here a general method in order to compute the generating function $\xi(z, t)$. We compute the $r \times r$ right extension matrix $\mathfrak{E}$ (where r is the number of words in $\mathcal{U}$) with the help of the Aho-Corasick automaton $\mathcal{A}_{\mathcal{U}}$. We remark that the coefficients of the matrix are polynomials whose degree (in any variable) is bounded by $\max_{u \in \mathcal{U}} |u| - 1$. Here the $r \times r$ matrix is smaller and more compact than the linear system obtained by applying the Chomsky-Schützenberger algorithm on the Aho-Corasick automaton of Section 2 which has size $O((\sum_{u \in \mathcal{U}} |u|)^2)$ since there are $O\left(\sum_{u \in \mathcal{U}} |u|\right)$ states in the automaton.

In the following, we provide an algorithm derived from the Aho-Corasick automaton (represented by a failure function) which computes the multivariate matrix $\mathfrak{E}$ and the vector $\mathfrak{u}^\circ$ in time $O(r^2 \times s + \sum_{u \in \mathcal{U}} |u|)$ where r is the cardinality of $\mathcal{U}$ and s is the size of the longest suffix chain[2] of a word $u \in \mathcal{U}$.

First we compute an auxiliary function which associates to any prefix w of the set $\mathcal{U}$ a vector $\mathbf{f}_w = (f_1(w), \ldots, f_r(w))$ defined by

$$f_i(w) = \sum_{\substack{v \\ w \cdot v = u_i}} \pi(v) z^{|v|} \prod_{m \neq i} (1 + t_m)^{|u_i|_m - |w|_m}.$$

Informally, given a particular left "context" w (a word in $\mathrm{Pref}(\mathcal{U})$), we examine the set of words v such that wv is a word u_i of $\mathcal{U}$, and we compute the number of new occurrences of u_m ($m \neq i$ and marked by $(1 + t_m)$) added by considering the extension v to w. We remark that $\mathfrak{u}^\circ = (f_1(\varepsilon), \ldots, f_r(\varepsilon))$ where ε is the empty word. Two examples are given on Figure 3. The Aho-Corasick construction is useful here because, intuitively, first it enables us, given a prefix w, to consider the set of suffixes v such that wv is a word of $\mathcal{U}$ (since they correspond to terminal states attainable from the current state), and secondly, the suffix links give access to the suffixes of any word (and so record possible correlations between words of $\mathcal{U}$).

[2]The suffix chain of $u \in \mathcal{U}$ is the sequence $(u_1 = u, u_2 = \mathrm{Border}(u_1), u_3 = \mathrm{Border}(u_2), \ldots, u_s = \mathrm{Border}(u_{s-1}) = \varepsilon)$.

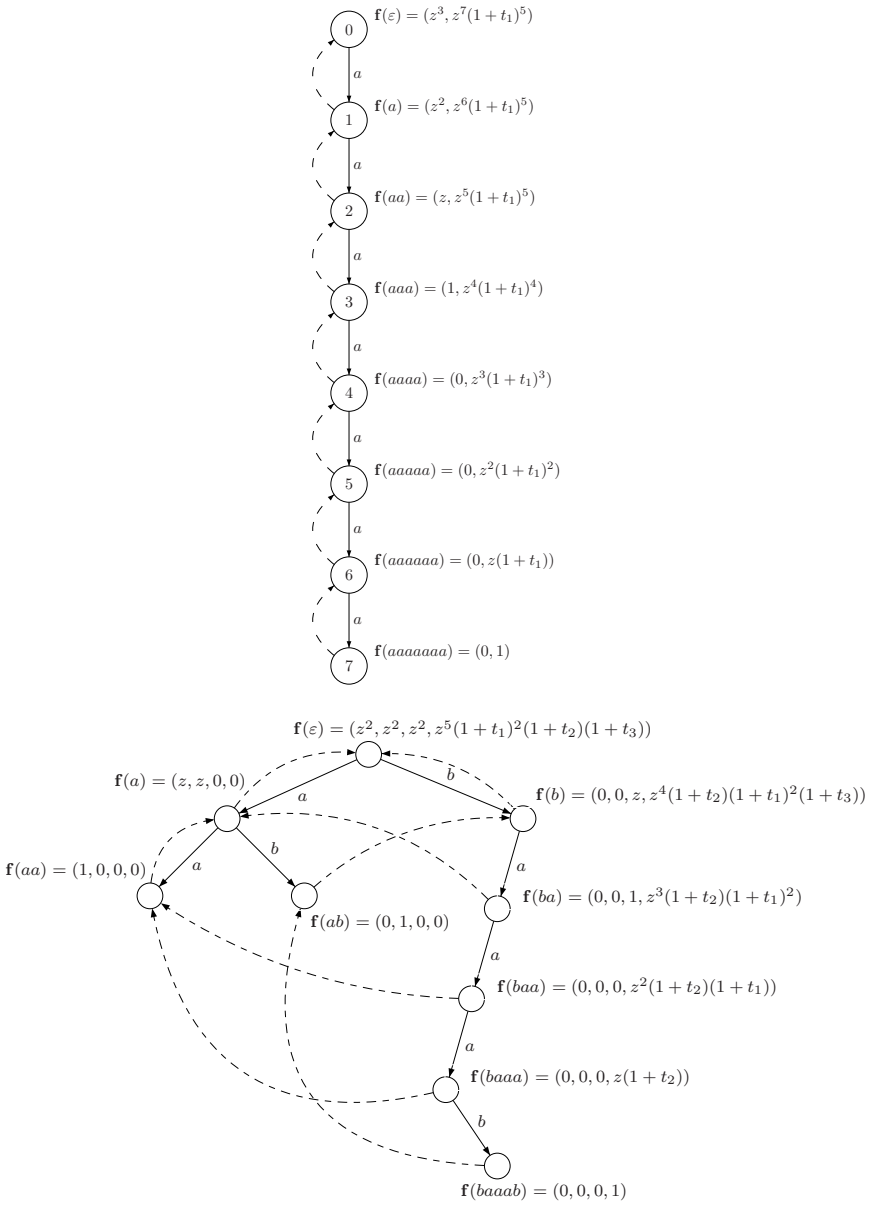


Fig. 3. Illustration for the computation of the functions $\mathbf{f} = (f_i)_{i=1}^r$ for $\mathcal{U} = \{a^3, a^7\}$ (top) and $\mathcal{U} = \{aa, ab, ba, baaab\}$ (bottom). The Aho-Corasick automata are represented together with their failure functions (dotted arrow) and the value of $\mathbf{f}(w)$ for $w \in \mathrm{Pref}(\mathcal{U})$ in both cases.

The "time complexity" (measured as the number of updates of the $f_i(w)$'s) of the following algorithm is $O(r \times \sum_{u\in\mathcal{U}} |u|)$.

INIT($\mathcal{A}_\mathcal{U}$)
for $i \leftarrow 1$ **to** r **do**
  $f_i(u_i) \leftarrow 1$
for $w \in \mathrm{Pref}(\mathcal{U})$ by a postorder traversal of the tree **do**
  for $i \leftarrow 1$ **to** r **do**
    for $\ell \in \mathcal{A}$ such that $w \cdot \ell \in \mathrm{Pref}(u_i)$ **do**
      $f_i(w) \leftarrow \pi(\ell) z f_i(w \cdot \ell) \prod_{j\neq i} (1 + t_j)^{[\![u_j \text{ suffix of } w\cdot\ell]\!]}$
return $(f_i)_{1\leq i\leq r}$

One then can show that

$$\mathfrak{E}_{i,j} = \sum_{\substack{v \text{ suffix of } u_i \\ v\notin\{\varepsilon\}\cup\mathcal{U}}} f_j(v).$$

So the matrix $\mathfrak{E}_\mathbf{u}$ can be computed thanks to the following algorithm.

BUILD-EXTENSION-MATRIX($\mathcal{A}_\mathcal{U}$)
▷ Initialize the matrix $(\mathfrak{E}_{i,j})_{1\leq i,j\leq r}$
for $i \leftarrow 1$ **to** r **do**
  for $j \leftarrow 1$ **to** r **do**
    $\mathfrak{E}_{i,j} \leftarrow 0$
▷ Compute the maps $(f_i(w))$ for i from 1 to r and $w \in \mathrm{Pref}(\mathcal{U})$
$(f_i)_{1\leq i\leq r} \leftarrow$ INIT($\mathcal{A}_\mathcal{U}$)
▷ Main loop
for $i \leftarrow 1$ **to** r **do**
  $v \leftarrow u_i$
  do $v \leftarrow \mathrm{Border}(v)$
    for $j \leftarrow 1$ **to** r **do**
      if $f_j(v) \neq 1$ **then**
        $\mathfrak{E}_{i,j} \leftarrow \mathfrak{E}_{i,j} + f_j(v)$
  while $v \neq \varepsilon$
return $\mathfrak{E}$

The main loop is iterated $O(s \times r^2)$ where r is the number of words and s is the length of the longest suffix chain. Hence the total time complexity (considering the number of operations on polynomials such as the $f_{i,j}$'s or the entries of $\mathfrak{E}$) is $O(r \times L + s \times r^2)$, where r is the number of words, L is the total length of words in $\mathcal{U}$ and s is the length of the longest suffix chain.

5. ASYMPTOTIC LIMIT LAWS

In this section, we review and reshape some general results from Bender and Kochman [Bender and Kochman 1993] concerning central limit theorems for occurrence statistics. Our contribution is twofold: first we propose to use an Aho-Corasick automaton (usually smaller than the de Bruijn graph proposed in [Bender and Kochman 1993]); second, we make use of nowadays standard notions like correlations and autocorrelations on words, and right extensions on words that we introduce in this paper.

Using the large power theorem of [Hwang 1998] that applies in dimension one, [Nicodème et al. 2002] prove a normal limit law for number of occurrences of regular expressions in texts under Bernoulli and Markovian model. This result has been extended to dynamical sources by [Bourdon and Vallée 2006]. Bender and Kochman [Bender and Kochman 1993] state central and normal limit theorems for occurrences of generalized words in any finite dimension, with possibly some forbidden generalized words in random texts under a Bernoulli model (where a generalized word of length ℓ is a set of words of length ℓ). The overall number of words considered is always finite. Recall that in the following a *pattern* is always a finite set of finite words.

The main asymptotic results of Bender and Kochman state limit laws for a set of patterns, when conditioning on observed counts of another set of patterns. Their proofs rely importantly on previous works of Bender *et al* in a series of articles [Bender 1973; Bender and Richmond 1983; Bender et al. 1983a].

We skip in this section some proofs that would closely follow the paths of the series of articles of Bender *et al.*. In particular, Bender and Kochman use a de Bruijn graph of high enough order, and large powers of the associated weighted adjacency matrix; we claim that it is equivalent to use the adjacency matrix associated to the recurrent component of any deterministic automaton accepting the language considered (and in particular the Aho-Corasick automaton).

A very nice feature of Bender *et al.* results is that the limiting normal law may be proved by enumerating a finite number of cases; in the context of words counting, this means that, by considering a finite number of (short) texts, it is possible to deduce a multivariate normal limit law. Disproving a normal limit law of given order by this method is however more difficult.

We will consider in Section 5.4.1 the *conditioned case* (where we can constrain a part of the occurrence counts to have a certain average value) that, as an interesting application, provides an asymptotic limit law under an asymptotic Markovian model.

We define now the lattice introduced by Bender *et al.* We consider next the unconditioned case for which nice closed formulas are available.

5.1 Lattices of differences of word counts

Let $\mathcal{U} = W_0 \cup W_1 \cup \ldots \cup W_d$ be a finite union of patterns. The Aho-Corasick automaton recognizing $\mathcal{U}$ can be decomposed into two subautomata: the initial subautomaton and the recurrent subautomaton. Let $\mathbb{C}(\mathbf{x})$ be the adjacency matrix of the subautomaton obtained by considering only the states in the recurrent part of the Aho-Corasick automaton. Note that the Aho-Corasick is not necessarily a minimal automaton.

5.1.1 *Primitivity.* By definition, the matrix $\mathbb{C}(\mathbf{x})$ is primitive if there exists an integer j such that all entries of $\mathbb{C}^j(\mathbf{x})$ are strictly positive. Primitivity implies irreducibility. We say that a set of patterns $\{W_1, \ldots, W_d\}$ is primitive if the associated matrix $\mathbb{C}(\mathbf{x})$ is primitive. This definition extends to the case where a pattern W_0 is forbidden (the random sequences considered are the subset of all sequences containing no occurrence of words of this pattern) by forbidding access to states recognizing words of W_0.

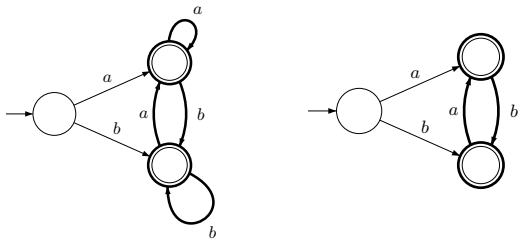


Fig. 4. (Left) Automaton for $\mathcal{U} = \{a, b\}$ over a binary alphabet. The connected automaton is drawn with bold lines. (Right) the same automaton with forbidden words $W_0 = \{aa, bb\}$; for this latter automaton, we do not have primitivity.

EXAMPLE 8. *We consider a binary alphabet and the pattern $W_1 = \{a, b\}$. Automata recognizing matches with W_1 respectively if $W_0 = \emptyset$ and $W_0 = \{aa, bb\}$ are depicted in Fig.4 (left) and Fig.4 (right). In the first case, the matrix associated to the recurrent part of the automaton is primitive. In the second case, this matrix is $\mathbb{J} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$; we have $\mathbb{J}^2 = \mathbb{I}_2$, the identity matrix and therefore no power of $\mathbb{J}$ has all entries different of zero, which excludes primitivity.*

We only consider in the following primitive sets of patterns $\{W_0, W_1, \ldots, W_s\}$.

5.1.2 *The lattice* Λ. We define now the lattice Λ associated to counts of a set of patterns $\{W_1, \ldots, W_s\}$.

Definition 5.1 *(Difference index set).* Let $\mathbb{C}(\mathbf{x})$ (with $\mathbf{x} = (x_1, \ldots, x_r)$) a matrix of polynomials in the variables $x_1, x_2, \ldots, x_r$. We consider all the power indices $\mathbf{i} = (i_1, \ldots, i_r)$ of the monomials $\mathbf{x}^{\mathbf{i}} = x_1^{i_1} \ldots x_r^{i_r}$ appearing in all elements $(\mathbb{C}^s(\mathbf{x}))_{\ell,m}$. We define now the difference index set $\mathcal{I}_s$ as the abelian group generated by all possible differences $\mathbf{i} - \mathbf{i}'$. We define the lattice of the system as

$$\Lambda = \bigcup_{s=\{1,2,3,\ldots\}} \mathcal{I}_s. \tag{13}$$

This definition generalizes immediately to a finite set of patterns by evaluating the variables associated to the words $w \in W_i$ to a variable y_i, and considering the vector of counts $\mathbf{y}$.

We have the following (degenerate) example.

EXAMPLE 9. *We consider again a binary alphabet and $W_1 = \{0, 1\}$ while $W_0 = \emptyset$. The automaton and its decomposition are depicted on Fig. 4 (left). The transition matrix of the connected automaton (posing $\pi(\alpha) = 1$ for all symbols α since the weights, if non null, do not influence the definition of the lattice) is*

$$\mathbb{C}(x_1, x_2) = \begin{pmatrix} x_1 & x_2 \\ x_1 & x_2 \end{pmatrix}, \quad \text{and} \quad \mathbb{C}^s(x_1, x_2) = (x_1 + x_2)^{s-1} \begin{pmatrix} x_1 & x_2 \\ x_1 & x_2 \end{pmatrix}.$$

The lattice obtained is $\Lambda = (1, -1) \times \mathbb{Z}$, and we remark that Λ has dimension 1; in particular, we have $\Lambda \neq \mathbb{Z}^2$.

5.2 Number of occurrences for a non-reduced pattern

We consider here a Bernoulli model and extend the probability measure π defined for words to sets of words in the following way

$$\pi(\mathcal{U}) = \sum_{u \in \mathcal{U}} \pi(u).$$

We provide for the case of a pattern $\mathcal{U} = \{u_1, \ldots, u_k\}$ expressions for the expected value and variance of X_n counting the number of occurrences of $\mathcal{U}$ in a text of size n. Section 4.3 gives a mean to obtain the generating function for clusters $\xi(z, t_1, \ldots, t_k)$ where word $u_i \in \mathcal{U}$ is marked by variable t_i. The cluster generating function $\mathfrak{C}(z, t)$ related to occurrences of $\mathcal{U}$ is then defined by

$$\mathfrak{C}(z, t) = \xi(z, t, \ldots, t). \tag{14}$$

Finally the generating function of occurrences is by Equation (12)

$$F(z, x) = \frac{1}{1 - z - \mathfrak{C}(z, x - 1)},$$

and, since $F(z, 1) = 1/(1 - z)$, we have $\mathfrak{C}(z, 0) = 0$. Setting $\mathfrak{C}_t(z) = \frac{\partial}{\partial t}\mathfrak{C}(z, t)\big|_{t=0}$ and $\mathfrak{C}_{tt}(z) = \frac{\partial^2}{\partial t^2}\mathfrak{C}(z, t)\Big|_{t=0}$ and using basic algebra, we have

$$\begin{aligned}
\sum_{n \geq 0} \mathbf{E}[X_n] z^n &= \tfrac{\partial}{\partial x} F(z, x)\big|_{x=1} \\
&= \frac{\mathfrak{C}_t(z)}{(1 - z)^2} \\
\sum_{n \geq 0} \mathbf{E}[X_n^2] z^n &= \frac{\partial^2}{\partial x^2} F(z, x)\Big|_{x=1} + \tfrac{\partial}{\partial x} F(z, x)\big|_{x=1} \\
&= \frac{2\mathfrak{C}_t(z)^2}{(1 - z)^3} + \frac{\mathfrak{C}_{tt}(z) + \mathfrak{C}_t(z)}{(1 - z)^2}.
\end{aligned}$$

It is easy to see that $\mathfrak{C}_t(z) = \sum_{u \in \mathcal{U}} \pi(u) z^{|u|}$ (the clusters with one and only one marked occurrence). The expression for $\mathfrak{C}_{tt}(z)$ takes into account that some words of $\mathcal{U}$ are factor of other ones

$$\mathfrak{C}_{tt}(z) = \sum_{u, v \in \mathcal{U}} (2\pi(u)|u|_v z^{|u|} + \sum_{e \in \mathcal{E}_{u,v}} 2\pi(ue) z^{|ue|}),$$

with the convention that $|u|_u = 0$. After some algebra, we get the following result.

PROPOSITION 5.2. *Let $\mathcal{U} = \{u_1, \ldots, u_k\}$ be a pattern. The expected value and variance of the variable X_n counting the number of occurrences of $\mathcal{U}$ in a text of*

size n satisfies

$$\mathbf{E}[X_n] = \sum_{u\in\mathcal{U}} \pi(u)(n-|u|+1)$$
$$\frac{1}{n}\mathbf{Var}[X_n] = \sum_{u\in\mathcal{U}} \pi(u) - \sum_{u,v\in\mathcal{U}} \pi(u)\pi(v)(|u|+|v|-1) + \sum_{u,v\in\mathcal{U}} 2\pi(u)\pi(\mathcal{E}_{u,v}) + \sum_{u,v\in\mathcal{U}} 2\pi(u)|u|_v + o(1).$$

We point out that the last sum is a correcting factor and is non zero only if the set is non reduced.

If the set contains only one word u, we obtain (as we should!) the classical result for the variance

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{Var}(X_n) = \pi(u) + 2\pi(u)\pi(\mathcal{E}_{u,u}) - (2|u|-1)\pi(u)^2. \tag{15}$$

5.3 Covariance matrix for a non-conditioned set of patterns

Here we consider again a Bernoulli model and provide for the case of a (loosely denoted as a list) set of patterns $(W_1,\dots,W_d)$ an alternative form to Bender and Kochman theorem. We analyse covariance by considering these motifs by pairs and choosing any pair $(\mathcal{U},\mathcal{V})$ amongst the preceding set of patterns. As an application of Theorem 4.12, we find an alternative form of Bender and Kochman [Bender and Kochman 1993, Theorem 2], where we use the notion of right extensions sets introduced in this paper.

THEOREM 5.3. *The normalized asymptotic correlation coefficient of co-occurrences of two patterns W_i and W_j verifies*

$$\mathbb{B}_{ij} = \sum_{\substack{u\in W_i\\ v\in W_j}} \Big(\pi(u)\pi(\mathcal{E}_{u,v}) + \pi(v)\pi(\mathcal{E}_{v,u}) - (|u|+|v|-1)\pi(u)\pi(v)\Big) + \pi(W_i\cap W_j) + \sum_{\substack{u\in W_i\\ v\in W_j}} (|u|_v\pi(u) + |v|_u\pi(v)) + o(1), \tag{16}$$

with the convention $|u|_u = 1$. Moreover the non-singularity[3] of the matrix $(\mathbb{B}_{ij})$ is equivalent to the d-dimensionality of the lattice Λ.

PROOF. We consider here the weighted case where $A(z) = z$. Let $\mathcal{U}$ and $\mathcal{V}$ be two sets of words. We first decompose as a direct sum the set $\mathcal{U}\cup\mathcal{V}$:

$$\mathcal{U}\cup\mathcal{V} = (\mathcal{U}\setminus\mathcal{V}) \oplus (\mathcal{V}\setminus\mathcal{U}) \oplus (\mathcal{U}\cap\mathcal{V}).$$

In order to ease notation, we index variables in the generating function $\xi(z,\mathbf{t})$ by words, i.e., the variable t_u corresponds to word u. Then we consider the generating

[3]In dimension two, this condition is equivalent to the non-nullity of the Hessian of the system [Heuberger 2007]. Note also that the large powers $\kappa(n)$ considered by Hwang and Heuberger are any function of n tending to infinity. We have here the particular case $\kappa(n) = \Theta(n)$.

function of clusters for the three disjoint sets $\mathcal{U}' = \mathcal{U} \setminus \mathcal{V}$, $\mathcal{V}' = \mathcal{V} \setminus \mathcal{U}$ and $\mathcal{W} = \mathcal{U} \cap \mathcal{V}$, with $\mathbf{t} = (t_u)_{u \in \mathcal{U} \cup \mathcal{V}}$ with the respective variables t_1, t_2 and t_3

$$\mathfrak{C}(z, t_1, t_2, t_3) = \xi(z, \mathbf{u})|_{\substack{t_u = t_1 \text{ for } u \in \mathcal{U} \setminus \mathcal{V} \\ t_u = t_2 \text{ for } u \in \mathcal{V} \setminus \mathcal{U} \\ t_u = t_3 \text{ for } u \in \mathcal{U} \cap \mathcal{V}}}, \tag{17}$$

that is we simply substitute variables for words appearing in each of the three sets with t_1, t_2 and t_3.

Let $F(z, x, y)$ be the corresponding generating function counting occurrences. We have by Equation (12) and since occurrences in $\mathcal{U} \cap \mathcal{V}$ are marked two times

$$F(z, x, y) = \frac{1}{1 - z - \mathfrak{C}(z, x - 1, y - 1, xy - 1)}. \tag{18}$$

By construction, since $F(z, 1, 1) = \frac{1}{1-z}$, one has

$$\mathfrak{C}(z, 0, 0, 0) = 0.$$

To simplify notation, we set

$$\mathfrak{C}_i(z) = \frac{\partial}{\partial t_i} \mathfrak{C}(z, t_1, t_2, t_3)\Big|_{(t_1,t_2,t_3)=(0,0,0)} \quad \text{for } i = 1, 2, 3$$
$$\mathfrak{C}_{ij}(z) = \frac{\partial^2}{\partial t_i \partial t_j} \mathfrak{C}(z, t_1, t_2, t_3)\Big|_{(t_1,t_2,t_3)=(0,0,0)} \quad \text{for } i, j \in \{1, 2, 3\}.$$

By general mechanisms [Flajolet and Sedgewick 2009] we get from Equation (18)

$$\sum_{n \geq 0} \mathbf{E}(X_n) z^n = \left(\frac{\partial}{\partial x} F(z, x, y)\Big|_{x=y=1}\right. = \frac{1}{(1-z)^2} \left(\mathfrak{C}_1(z) + \mathfrak{C}_3(z)\right)$$
$$\sum_{n \geq 0} \mathbf{E}(Y_n) z^n = \left(\frac{\partial}{\partial y} F(z, x, y)\Big|_{x=y=1}\right. = \frac{1}{(1-z)^2} \left(\mathfrak{C}_2(z) + \mathfrak{C}_3(z)\right),$$

which gives

$$\mathbf{E}(X_n) = \sum_{u \in \mathcal{U}} (n - |u| + 1)\pi(u), \qquad \mathbf{E}(Y_n) = \sum_{u \in \mathcal{V}} (n - |u| + 1)\pi(u).$$

We have also easy access to the covariance since

$$\begin{aligned}
\sum_{n \geq 0} \mathbf{E}(X_n Y_n) z^n &= \frac{\partial^2}{\partial x \partial y} F(z, x, y)\Big|_{x=y=1} \\
&= 2\frac{(\mathfrak{C}_1(z) + \mathfrak{C}_3(z))(\mathfrak{C}_2(z) + \mathfrak{C}_3(z))}{(1-z)^3} + \\
&\quad \frac{\mathfrak{C}_{12}(z) + \mathfrak{C}_{13}(z) + \mathfrak{C}_{23}(z) + \mathfrak{C}_{33}(z) + \mathfrak{C}_3(z)}{(1-z)^2} \\
&= 2\frac{\mathfrak{C}_1(z)\mathfrak{C}_2(z)}{(1-z)^3} + \frac{\mathfrak{C}_{12}(z)}{(1-z)^2} + 2\frac{\mathfrak{C}_3(z)^2}{(1-z)^3} + \frac{\mathfrak{C}_{33}(z) + \mathfrak{C}_3(z)}{(1-z)^2} \\
&\quad + 2\frac{\mathfrak{C}_3(z)(\mathfrak{C}_1(z) + \mathfrak{C}_2(z))}{(1-z)^3} + \frac{\mathfrak{C}_{13}(z) + \mathfrak{C}_{23}(z)}{(1-z)^2}.
\end{aligned}$$

A Taylor expansion at $z=1$ gives for $i=1,2,3$ (with $f'(z)=\frac{\partial}{\partial z}f(z)$)

$$\mathfrak{C}_i(z)=\mathfrak{C}_i(1)-(1-z)\mathfrak{C}_i'(1)+o(1-z).$$

Hence we get

$$\begin{aligned}\mathbf{E}(X_nY_n)=(n+1)(n+2)\big(\mathfrak{C}_1(1)+\mathfrak{C}_3(1)\big)\big(\mathfrak{C}_2(1)+\mathfrak{C}_3(1)\big)\\ +(n+1)\Big(\big(\mathfrak{C}_1'(1)+\mathfrak{C}_3'(1)\big)\big(\mathfrak{C}_2(1)+\mathfrak{C}_3(1)\big)\\ +\big(\mathfrak{C}_1(1)+\mathfrak{C}_3(1)\big)\big(\mathfrak{C}_2'(1)+\mathfrak{C}_3'(1)\big)\\ +\mathfrak{C}_{12}(1)+\mathfrak{C}_{13}(1)+\mathfrak{C}_{23}(1)+\mathfrak{C}_{33}(1)+\mathfrak{C}_3(1)\Big)\\ +o(n).\end{aligned}$$

Now we can interpret each of the coefficients: we have for instance

$$\begin{aligned}\mathfrak{C}_1(1)+\mathfrak{C}_3(1)&=\sum_{u\in\mathcal{U}}\pi(u),\\ \mathfrak{C}_{13}(1)&=\sum_{u\in\mathcal{U}\setminus\mathcal{V}}\sum_{v\in\mathcal{V}\cap\mathcal{U}}\big(\pi(u)|u|_v+\pi(u)\pi(\mathcal{E}_{u,v}))+\pi(v)|v|_u+\pi(v)\pi(\mathcal{E}_{v,u}))\big).\end{aligned}$$

Summarizing and after some computations, we find that

$$\begin{aligned}\frac{1}{n}\mathbf{Cov}(X_n,Y_n)=\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}}\Big(\pi(u)\pi(\mathcal{E}_{u,v})+\pi(v)\pi(\mathcal{E}_{v,u})-(|u|+|v|-1)\pi(u)\pi(v)\Big)\\ +\pi(\mathcal{U}\cap\mathcal{V})+\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}}\big(|u|_v\pi(u)+|v|_u\pi(v)\big)+o(1),\end{aligned}$$

where, by convention, $|u|_u=0$. Substituting $\mathcal{U}$ and $\mathcal{V}$ respectively by W_i and W_j leads to Equation (16). □

The following theorem is proved in Bender and Kochman [Bender and Kochman 1993]; its proof suppose to unwind the proofs of the series of articles [Bender 1973; Bender and Richmond 1983; Bender et al. 1983a; Bender and Kochman 1993].

THEOREM 5.4. *Given a set of finite patterns $\{W_1,\dots,W_r\}$, the non-singularity of the covariance matrix $(\mathbb{B}_{ij})$ where $\mathbb{B}_{ij}$ is defined in Equation (16) is equivalent to the d-dimensionality of the lattice Λ.*

EXAMPLE 10. *We apply the results of this section to $W_1=a^3$ and $W_2=a^7$ in a Bernoulli model with $p=\pi(a)$. We get by Equation (16),*

$$\left\{\begin{array}{l}\pi(a^3)\pi(\mathcal{E}_{a^3,a^7})=p^3(p^5+p^6),\\ \pi(a^7)\pi(\mathcal{E}_{a^7,a^3})=p^7(p+p^2),\\ |a^7|_{a^3}\pi(a^7)=5p^7,\\ |a^3|_{a^7}\pi(a^3)=0,\\ (|a^7|+|a^3|-1)\pi(a^7)\pi(a^3)=9p^{10}\end{array}\right|\qquad\Longrightarrow\qquad \mathbb{B}_{12}=5p^7+2p^8+2p^9-9p^{10}.$$

Computing the full matrix $\mathbb{B}^{(a^3,a^7)}=(\mathbb{B}_{ij})$ for $i,j=\{1,2\}$, and the corresponding

determinant Δ*, we get*

$$\mathbb{B}^{(a^3,a^7)} = \begin{pmatrix} p^3+2p^3(p+p^2)-5p^6 & p^7(5+2p+2p^2-9p^3) \\ p^7(5+2p+2p^2-9p^3) & p^7+2p^7(p+p^2+p^3+p^4+p^5+p^6)-13p^{14} \end{pmatrix},$$

$$\Delta = |\mathbb{B}^{(a^3,a^7)}| = p^{10} + 4p^{11} + 8p^{12} + 5p^{13} - 25p^{14} - 20p^{15} - 24p^{16} + 67p^{17} - 16p^{20}.$$

Let us remark that Δ *is zero if* $p = 0$ *or* $p = 1$ *and nowhere else. This corresponds to a degeneracy of the system. In these cases, we have* $\Lambda = (0, 0)$*, and the dimension of* Λ *is zero, which corresponds to the lack of variability of the system. If, on the opposite, we consider* $0 < p < 1$*, and texts of size* 15*, we see that among the possibility of occurrences, we have*

$$(0,1), (0,2), (0,3), (0,4), (1,5), (1,6), (1,7), \ldots, (9,13);$$

applying the differences $\mathbf{i} - \mathbf{i}'$ *provides a lattice of dimension 2, which corresponds to the non-singularity of the matrix* $\mathbb{B}^{(a^3,a^7)}$*.*

5.4 General case

Considering two sets of patterns S_1 and S_2, Bender and Kochman use an analytic heuristic method to obtain the average values of the counts of S_2 when the counts for S_1 are fixed at any value possibly happening for a text. This method relies on the exponential shift $x^i \rightsquigarrow r^i x^i$ for the variables counting the patterns in S_1 and an *a posteriori* normalization to recover a probability distribution. Since this shift modifies the mean of the distribution, it is often called *mean-shifting* method. Note that the resulting distribution is distorted; however, it is still possible to get a matricial equation for the covariances of patterns in S_2.

We provide here an introduction to the mean-shifting method, restate Theorem 1 of [Bender and Kochman 1993], and apply the results to an asymptotic Markovian model for the sequences.

5.4.1 *Mean shifting.*

Principles. Bender and Kochman [Bender and Kochman 1993] use without explicitly mentioning it the analytical mean shifting method, that substitutes the formal variable $\mathbf{x}$ by $\boldsymbol{\rho}\mathbf{x}$ and normalize to recover a probability generating function. Bender-Kochman [Bender and Kochman 1993] use extensively this method to prove that, even when conditioning on a given ("analytic") value $\boldsymbol{\rho}$, there is a limiting normal distribution for the counts.

We consider as introductory example the bivariate case. Let

$$F(z,x) = \sum_{n\geq 0, 0\leq k\leq n} f_{n,k} x^k z^n = \sum_{n,k} \Pr(X_n = k) x^k z^n \tag{19}$$

be a multivariate generating function where X_n counts the number of some objects in a system of size n. We have

$$\mu_n^{(1)} = \mathbf{E}(X_n) = [z^n] \left. \frac{\partial F(z,x)}{\partial x} \right|_{x=1}.$$

We write $\phi(x) = [z^n]F(z,x)$, and use ρ as a shift variable. We consider

$$\psi^{(\rho)}(x) = \frac{\phi(\rho x)}{\phi(\rho)}; \tag{20}$$

since $\psi^{(\rho)}(1) = 1$, we did define the probability generating function of a random variable $X_n^{(\rho)}$. We have

$$\mu_n^{(\rho)} = \mathbf{E}\left(X_n^{(\rho)}\right) = \frac{\partial}{\partial x}\frac{\phi(\rho x)}{\phi(\rho)}\bigg|_{x=1} = \rho\frac{\phi'(\rho)}{\phi(\rho)}.$$

When ρ tends to infinity, if $\phi_d x^d$ is the monomial of highest degree of $\phi(x)$, we have

$$\lim_{\rho\to\infty}\mu_n^{(\rho)} = \lim_{\rho\to\infty}\frac{\rho\times d\phi_d\rho^{d-1}}{\phi_d\rho^d} = d,$$

and d is the largest possible count; similarly, if i is the smallest possible count, we have $\mu_n^{(\rho)} \to i$ as ρ tends to 0.

An important result is that, when ρ varies from 0 to ∞ the variable $\mu_n^{(\rho)}$ increases continuously and goes through all the possible (discrete) values that the variable X_n can take. Moreover, $\mu_n^{(\rho)}$ is a convex function of r. Therefore, for any "possible" value α of $\mu_n^{(r)}$, it is possible to find a value ρ^* for ρ such that $\mu_n^{(\rho^*)} = \alpha$. We have thus a way to condition the random variable to have a certain average value.

This result extends to several dimensions.

Perron-Frobenius dominant eigenvalue. We relate here the mean shifting method and the logarithmic derivatives used in Bender *et al.*.

Since we consider positive matrices and the induced rational generating functions, we have, asymptotically, by a Cauchy integration along a suitable contour,

$$\phi(x) = [z^n]F(z,x) = c(x)\lambda(x)^n \times \left(1 + O\left(\frac{1}{R^n}\right)\right), \qquad (R>1), \tag{21}$$

where $\lambda(x)$ is the dominant eigenvalue of the positive matrix of our system and $c(x)$ is analytic. We get, as n tends to infinity,

$$\begin{aligned}\mu_n^{(\rho)} = \frac{\partial}{\partial x}\frac{\phi(\rho x)}{\phi(\rho)}\bigg|_{x=1} &= \left(\frac{c'(\rho)}{c(\rho)} + \frac{\partial}{\partial x}\frac{\lambda(\rho x)^n}{\lambda(\rho)^n}\bigg|_{x=1}\right)\times\left(1+O\left(\frac{1}{R^n}\right)\right)\\ &\sim n\times\rho\frac{\lambda'(\rho)}{\lambda(\rho)} = n\times\frac{\partial\log(\lambda(\rho))}{\partial z}\end{aligned} \tag{22}$$

where $\rho = \log z$. This is precisely the definition given by Bender *et al.* who only consider the dominant asymptotic term. This approach extends obviously to second derivatives; see similar conditions in [Heuberger 2007].

Considering for sake of simplicity counting of matches of two patterns, we have

$$\phi(z,x_1,x_2) = [z^n]F(z,x_1,x_2) = c(x_1,x_2)\lambda(x_1,x_2)^n\times\left(1+O\left(\frac{1}{R^n}\right)\right), \qquad (R>1).$$

This gives

$$\mu_{n,1}^{(\rho_1,\rho_2)} = \frac{\partial}{\partial x_1}\frac{\phi(\rho_1 x_1, \rho_2 x_2)}{\phi(\rho_1,\rho_2)}\bigg|_{\substack{x_1=1\\x_2=1}} \sim n\rho_1 \times \frac{\lambda_1'(\rho_1,\rho_2)}{\lambda(\rho_1,\rho_2)},$$

$$\left(\mu_{n,1}^{(\rho_1,\rho_2)}, \mu_{n,2}^{(\rho_1,\rho_2)}\right) \sim n \times \left(\rho_1\frac{\lambda_1'(\rho_1,\rho_2)}{\lambda(\rho_1,\rho_2)}, \rho_2\frac{\lambda_2'(\rho_1,\rho_2)}{\lambda(\rho_1,\rho_2)}\right), \quad (23)$$

where $\mu_{n,i}$ is the expectation of the number of occurrences of the pattern i in texts of size n and $\lambda_i'(s,t)$ is the derivative of $\lambda(s,t)$ with respect to the parameter at position i.

Then, provided that the values $(\lfloor n\xi_1\rfloor, \lfloor n\xi_2\rfloor)$ are possible for the counts, we can compute the corresponding values ρ_1 and ρ_2 by *simultaneously* solving the equations

$$\left\{\frac{\lambda_1'(\rho_1,\rho_2)}{\lambda(\rho_1,\rho_2)} = \xi_1, \quad \frac{\lambda_2'(\rho_1,\rho_2)}{\lambda(\rho_1,\rho_2)} = \xi_2\right\}. \quad (24)$$

All this generalizes to multivariate generating functions with a higher number of parameters.

Bender and Kochman prove that the map from $(\rho_1,\ldots,\rho_r)$ to the possible limit frequencies of the patterns is bijective, and that the set of limit frequencies is convex.

5.4.2 *From average conditioning to exact conditioning.* In [Bender and Kochman 1993, Theorem 1], the authors give a limit law for counts $X_j^{(n)}$ of patterns W_j with $j \in \{c+1,\ldots,d\}$, assuming that the counts $X_i^{(n)}$ of patterns W_i with $i \in \{1,\ldots,c\}$ are $\lfloor n \times k_i\rfloor + o(n)$ for given $k_1,\ldots,k_c$. The result is also conditioned by $X_0^{(n)} = 0$.

We have seen previously that for any "possible" $(\xi_1,\ldots,\xi_d)$ we can find a corresponding $(\rho_1,\ldots,\rho_d)$. The corresponding values for $(\mu_{1,n},\ldots,\mu_{i,n}) = \lim \frac{1}{n}(\mu_{1,n}, \ldots,\mu_{d,n})$ do not depend on n.

We consider the probability multivariate generating function

$$F(z,\mathbf{x}) = F(z,x_0,x_1,\ldots,x_c,x_{c+1},\ldots,x_d)$$

where z counts the size of the texts and x_i counts the number of occurrences of the pattern W_i. We condition now on the forbidden set of words W_0 and on the expected counts for W_j with $j \in \{1,\ldots,c\}$. This gives

$$\Phi_n(x_{c+1},\ldots,x_d) = \frac{[z^n]F(z,0,\rho_1,\ldots,\rho_c,x_{c+1},\ldots,x_d)}{[z^n]F(z,0,\rho_1,\ldots,\rho_c,1,\ldots,1)}. \quad (25)$$

This last equation provides the conditioned variables $X_i^{(n)}$ with $i \in \{c+1,\ldots,d\}$. The dominant asymptotic term of the means of these variables is $n \times m_i$ where

$$m_i = \lim_{n\to\infty}\frac{1}{n}\left.\frac{\partial\Phi_n(1,\ldots,1,x_i,1,\ldots,1)}{\partial x_i}\right|_{x_i=1}.$$

We need for what follows a lemma on conditional distributions analysis (see [Gelman et al. 1995, p. 79]).

LEMMA 5.5. *We consider two random vectors* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *with respective size* c *and* $d-c$ *and the vector* $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2)$ *of size* d. *Let* $\mathbf{m}$ *be the vector of expected*

values of $\mathbf{X}$ *and* $\mathbb{B}$ *be its covariance matrix. If the expectation vector* $\mathbf{m}$ *and the covariance matrix* $\mathbb{B}$ *are partitioned as follows*

$$\mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} \qquad \text{with sizes} \quad \begin{pmatrix} c \times 1 \\ (d-c) \times 1 \end{pmatrix},$$

$$\mathbb{B} = \begin{pmatrix} \mathbb{B}_{11} & \mathbb{B}_{12} \\ \mathbb{B}_{21} & \mathbb{B}_{22} \end{pmatrix} \qquad \text{with sizes} \quad \begin{pmatrix} c \times c & c \times (d-c) \\ (d-c) \times c & (d-c) \times (d-c) \end{pmatrix},$$

and such that $\mathbb{B}_{11}$ *is invertible, then the distribution of* $\mathbf{X}_2$ *conditional on* $\mathbf{X}_1 = \mathbf{o}$ *is multivariate normal* $(\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{o}) \sim N(\overline{\mathbf{m}}_2, \overline{\mathbb{B}}_{22})$ *where*

$$\overline{\mathbf{m}}_2 = \mathbf{m}_2 + \mathbb{B}_{21}\mathbb{B}_{11}^{-1}(\mathbf{o} - \mathbf{m}_1), \tag{26}$$

$$\overline{\mathbb{B}}_{22} = \mathbb{B}_{22} - \mathbb{B}_{21}\mathbb{B}_{11}^{-1}\mathbb{B}_{12}. \tag{27}$$

The matrix $\mathbb{B}_{21}\mathbb{B}_{11}^{-1}$ *is known as the matrix of regression coefficients.*

Note that if the random vector $\mathbf{X}_1$ is gaussian, then $\mathbb{B}_{11}$ is invertible. Moreover, the matrix $\overline{\mathbb{B}}_{22}$ does *not* depend on the value of $\mathbf{o}$.

5.4.3 *Normal limiting distributions.* We note in the following, for a set of words $\mathcal{U} = \{u_1, \ldots, u_r\}$

$$|w|_{\mathcal{U}} = \sum_{u \in \mathcal{U}} |w|_u.$$

We have the following theorem.

THEOREM 5.6 [BENDER AND KOCHMAN 1993]. *Suppose that* $\mathcal{W} = (W_0, \ldots, W_d)$ *is primitive and that* $\Lambda(\mathcal{W})$ *is d-dimensional.*

Then the set $\mathcal{F}$ *of accumulation points of*

$$\bigcup_n \bigcup_{w \in \mathcal{A}^n} \left\{ \frac{1}{n} \left(|w|_{W_1}, \ldots, |w|_{W_c} \right) \quad \middle| \quad |w|_{W_0} = 0 \right\}$$

is c-dimensional.

Considering in random texts of size n unconditioned counts $X_0^{(n)}, X_1^{(n)}, \ldots, X_c^{(n)}$ *of the patterns* $W_0, \ldots, W_c$*, unconditioned counts* $X_{c+1}^{(n)}, \ldots, X_d^{(n)}$ *of the patterns* $W_{c+1}, \ldots, W_d$*, and the counts* $\overline{X}_{c+1}^{(n)}, \ldots, \overline{X}_d^{(n)}$ *of* $W_{c+1}, \ldots, W_d$ *conditioned by* $X_0^{(n)} = 0$ *and* $X_i^{(n)} = \lfloor n j_i \rfloor$ *with* $i \in [1, c]$*, where* $(j_1, \ldots, j_i) \in \mathcal{F}$*, let* $(\rho_1, \ldots, \rho_c)$ *be the vector of mean shift parameters such that*

$$\lim_{n \to \infty} \frac{1}{n} \mathbf{E}(X_1^{(n)}, \ldots, X_c^{(n)}) = (j_1, \ldots, j_c);$$

considering now the conditioned variables $\overline{X}_i^{(n)}$*, the generating function of which is given by Equation (25), and the vector of variables* $\mathbf{V}_n = (Y_{c+1}^{(n)}, \ldots, Y_d^{(n)})$*, where*

$$Y_i^{(n)} = \frac{X_i^{(n)} - n \times m_i}{\sqrt{n}}, \text{ and } m_i = \frac{\lambda_i'(0, \rho_1, \ldots, \rho_c, 1, \ldots, 1)}{\lambda(0, \rho_1, \ldots, \rho_c, 1, \ldots, 1)}, \qquad (i = c+1, \ldots, d);$$

there exists a non-singular computable matrix $\overline{\mathbb{B}}$ *such that:*

(1) *the vector* $\mathbf{V}$ *verifies a central limit theorem with mean* $(0,\dots,0)$ *and covariance matrix* $\overline{\mathbb{B}}$*;*

(2) *if* $\Lambda(\mathcal{W}) = \mathbb{Z}^d$*, we have a local limit theorem,*

$$\lim_{n\to\infty} \sup_{k_{c+1},\dots,k_d} \left| n^{(d-c)/2} \Pr\left(X_{c+1}^{(n)} = k_{c+1},\dots,X_d^{(n)} = k_d\right) - \frac{\exp\left(-\frac{1}{2}\mathbf{V}_n \overline{\mathbb{B}}^{-1} \mathbf{V}_n'\right)}{\sqrt{(2\pi)^{d-c}\det\overline{\mathbb{B}}}} \right| = 0, \tag{28}$$

where $\mathbf{V}_n'$ *is the transpose of* $\mathbf{V}_n$*.*

Proof. (Sketch) The conditioning by $X_{n,0} = 0$ may be done directly on the Aho-Corasick automaton by removing all transitions pointing to a state recognizing a word of W_0. We remark that the transient part of the automaton provides only a finite number of occurrences in the counts (read at the beginning of the sequence), which has no influence when the size of the texts tends to infinity; on contrary the values of the limiting average counts come from the recurrent part of the automaton. We recall here that we supposed that this recurrent part is primitive after having forbidden access to states recognizing words of W_0; this corresponds to the hypothesis that $\mathcal{W} = (W_0,\dots,W_d)$ is primitive and this implies that the resulting automaton *cannot* be disconnected.

The proof follows from Theorem 1 of [Bender et al. 1983b] that considers large powers of matrices with entries that are multivariate polynomials with positive coefficients; this theorem applies to the case of the matrix of the Aho-Corasick automaton under consideration. At this stage, there is no conditioning.

Bender and Kochman then proceed in two steps.

(1) Apply the mean shift $x_i \rightsquigarrow \rho_i x_i$ ($i \in [1..c]$) such that the *expected* counts for $W_1,\dots,W_c$ are $j_1 n,\dots,j_c n$.

(2) Use the Lemma 5.5 to get the desired "section" of the distribution, which provides the value of the covariance matrix $\overline{\mathbb{B}}$, as a function of the components of the covariance matrix $\mathbb{B}$. The matrix $\mathbb{B}$, in turn, is given by its entries $\mathbb{B}_{ij}$ of Equation (16); its computation therefore follows from using asymptotics and singularity analysis, what was done in Section 5.3.

□

5.5 Application: asymptotic Markovian model

There is a very simple and direct application of Theorem 5.6 to the case of a model related to a Markovian source. Let us consider for instance a Markov model of order 1 and a binary alphabet. The model is totally described by the (trivial) system

$$\begin{cases} 1 &= \pi(0) + \pi(1) \\ \pi(0) &= \pi(00) + \pi(10) \\ \pi(1) &= \pi(01) + \pi(11) \end{cases}$$

so that the probabilities $\pi(00), \pi(01), \pi(10)$ (for instance) are sufficient to deduce the other ones. Asymptotically, by the strong law of large numbers [Grimmet and Stirzaker 1992], it is equivalent to consider a model where the probability of the

sequences xy is $\pi(xy)$ and a model where the probability that an y follows an x is $\pi(xy)$ for x and y equal to 0 or 1. Therefore, choosing $c = 3$ and $W_1 = \{00\}, W_2 = \{01\}, W_3 = \{10\}$ in Theorem 5.6, we get an asymptotic central theorem for any finite set of patterns $\{W_4, \ldots, W_d\}$, provided that $\Lambda(00, 01, 10, W_4, \ldots, W_d)$ is d-dimensional, and we get a local limit theorem if $\Lambda = \mathbb{Z}^d$, which is mostly often the case. When doing so, we only need Step 1 of the proof of Theorem 5.6, that ensures that the limiting conditioning frequencies fit to the frequencies of the Markov model. Note that we are not in an exact Markov model since we are only insured that the average value of frequencies tends to the one of the Markov model (hence the name of asymptotic Markov model).

This generalizes to an (asymptotic) Markov model of highest order. We should however notice that, in the case of a Markov model of order k, we should not over-constrain the problem in choosing the "Markov fitting words" $W_1, \ldots, W_c$. For instance, for $k = 1$, the frequencies are totally described by giving $\pi(00), \pi(01), \pi(10)$ since then we are able to compute $\pi(11) = 1 - \pi(00) - \pi(01) - \pi(10)$, $\pi(1) = \pi(01) + \pi(11)$, $\pi(0) = \pi(00) + \pi(10)$. In fact for a general order k we need to fix exactly $c = 2^{k+1} - 1$ frequencies to describe the model.

PROPOSITION 5.7. *Provided a description of the Markov model with the vector of words $(W_1, \ldots, W_{2^{k+1}-1})$, and considering a vector of patterns $(W_{2^{k+1}-1}, \ldots, W_d)$, in an asymptotic Markov model of order k, the vector of counts $(X_{2^{k+1}-1}, \ldots, X_d)$ verifies asymptotically*

(1) a central limit theorem if the lattice Λ of the system is d-dimensional

(2) a local limit theorem if we have $\Lambda = \mathbb{Z}^d$.

The vector of expected counts and the covariance matrix are given in Theorem 5.6.

PROOF. The proof follows from Theorem 5.6 and Lemma 5.5. □

Conclusion and perspectives

We obtained a detailed proof and an explicit expression of the multivariate generating function counting texts according to their length and to their number of occurrences of words from a finite set. This result facilitates access to various moments and may lead to limiting distributions. From Bender and Kochman [Bender and Kochman 1993], we expect to find mostly a multivariate normal law for word counts. Our approach can possibly provide simpler criteria to decide if such a limiting law holds or not. Another nice aspect of the inclusion-exclusion approach is that it provides explicit formulae like Equation (11), whereas the Aho-Corasick construction does not preserve the structure: even for a single pattern the autocorrelation polynomial does not come out easily from the (Morris-Pratt) automaton.

Ongoing work is more concerned with the complexity of the diverse approaches presented in this paper. Also we plan to extend the analysis to more complex sources, such as Markovian or dynamical sources (see Vallée [Vallée 2001]).

Acknowledgements. The authors thank Jérémie Bourdon and Philippe Dumas for fruitful discussions and for providing important feedback to this paper.

REFERENCES

Aho, A. and Corasick, M. 1975. Efficient String Matching: An Aid to Bibliographic Search. *Communications of the ACM 18*, 333–340.

Bender, E. 1973. Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory 15*, 91–111.

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

Bender, E. and Richmond, B. 1983. Central and local limit theorems applied to asymptotic enumeration II: Multivariate Generating Functions. *Journal of Combinatorial Theory Series A,* 34, 255–265.

Bender, E., Richmond, B., and Williamson, G. 1983a. Central and local limit theorems applied to asymptotic enumeration. III. Matrix recursions. *Journal of Combinatorial Theory 35,* 3, 264–278.

Bender, E. A., Richmond, L. B., and Williamson, S. 1983b. Central and local limits theorems applied to asymptotic enumeration. iii: Matrix Recursions. *Journal of Combinatorial Theory Series A,* 35, 263–278.

Bourdon, J. and Vallée, B. 2002. Generalized pattern matching statistics. In *Proc. Colloquium on Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities.* Birkhauser, Trends in Mathematics. 249–265.

Bourdon, J. and Vallée, B. 2006. Pattern matching statistics on correlated sources. In *Proc. of LATIN'06.* LNCS, vol. 3887. 224–237.

Chomsky, N. and Schützenberger, M. 1963. The algebraic theory of context-free languages. *Computer Programming and Formal Languages,*, 118–161. P. Braffort and D. Hirschberg, eds, North Holland.

Crochemore, M. and Rytter, W. 2002. *Jewels of Stringology.* World Scientific Publishing, Hong-Kong. 310 pages.

Flajolet, P. and Sedgewick, R. 2009. *Analytic Combinatorics.* Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 1995. *Bayesian Data Analysis.* Chapman & Hall. 526 pages.

Goulden, I. and Jackson, D. 1979. An inversion theorem for clusters decompositions of sequences with distinguished subsequences. *J. London Math. Soc. 2,* 20, 567–576.

Goulden, I. and Jackson, D. 1983. *Combinatorial Enumeration.* John Wiley. New-York.

Grimmet, G. and Stirzaker, D. 1992. *Probability and Random Processes.* Oxford Science Publications. Second Edition.

Guibas, L. and Odlyzko, A. 1981a. Periods in strings. *J. Combin. Theory A,* 30, 19–42.

Guibas, L. and Odlyzko, A. 1981b. Strings overlaps, pattern matching, and non-transitive games. *J. Combin. Theory A,* 30, 108–203.

Heuberger, C. 2007. Hwang's quasi-power-theorem in dimension two. *Quaest. Math. 30*, 507–512.

Hwang, H.-K. 1998. On convergence rates in the central limit theorems for combinatorial structures. *European J. Combinatorics 19*, 329–343.

Kong, Y. 2005. Extension of Goulden-Jackson cluster method on pattern occurrences in random sequences and comparison with Régnier Szpankowski method. *J. of Difference Equations and Applications 11,* 15, 1265–1271.

Lothaire, M. 2005. *Applied Combinatorics on Words.* Encyclopedia of Mathematics. Cambridge University Press.

Nicodème, P. 2003. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fundamenta Informaticae 56,* 1-2, 71–88.

Nicodème, P., Salvy, B., and Flajolet, P. 2002. Motif statistics. *Theoretical Computer Science 287,* 2, 593–618.

Noonan, J. and Zeilberger, D. 1999. The Goulden-Jackson Method: Extensions, Applications and Implementations. *J. of Difference Equations and Applications 5,* 4-5, 355–377.

Prum, B., Rodolphe, F., and de Turckheim, E. 1995. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. Statist. Soc. B 57,* 1, 205–220.

Régnier, M. 2000. A unified approach to word occurrences probabilities. *Discrete Applied Mathematics 104,* 1, 259–280. Special issue on Computational Biology.

Régnier, M. and Szpankowski, W. 1998. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica 22,* 4, 631–649.

Reinert, G. and Schbath. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol. 5*, 223–253.

Reinert, G., Schbath, S., and Waterman, M. 2000. Probabilistic and statistical properties of words: an overview. *J. Comp. Biol. 7*, 1–46.

Roquain, E. and Schbath, S. 2007. Improved compound poisson approximation for the number of occurrences of multiple words in a stationary markov chain. *Adv. Appl. Prob.* 39, 1–13.

Sedgewick, R. and Flajolet, P. 1996. *An Introduction to the Analysis of Algorithms.* Addison-Wesley Publishing Company.

Szpankowski, W. 2001. *Average Case Analysis of Algorithms on Sequences.* Series in Discrete Mathematics and Optimization. John Wiley & Sons.

Vallée, B. 2001. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica 29,* 1, 262–306.