Counting occurrences for a finite set of words: an inclusion-exclusion approach

Pierre Nicodème

CNRS - LIX, École polytechnique

joint work with *Frédérique Bassino, Julien Clément* and *Julien Fayolle*

Problem setting

Compute separately the number of occurrences of a non-reduced set of words $\mathcal{U}$ in a random text under Bernoulli (non-uniform) model

Reduced set: no word is factor of another word

Reduced	Non-Reduced
$\mathcal{U} = \{aab, ba, bb\}$	$\mathcal{U} = \{aa, aab, bbaabb\}$

Methods

– Formal languages manipulations (Régnier-Szpankowski) (**it fails in the non-reduced case**)

– Aho-Corasick (automaton) + Chomsky-Schützenberger

– Inclusion-Exclusion (Goulden-Jackson, Noonan-Zeilberger)

Analytic Aim

$\mathcal{U} = \{u_1, \dots u_r\}$ non-reduced set of words

$\mathcal{O}_n^{(r)}$: random variable counting the number of occurrences of the word u_r in a random text of size n (Bernoulli model)

We want to compute

$$F(z, x_1, \dots, x_r) = \sum_{k_1 \geq 0, \dots, k_r \geq 0, n \geq 0} \Pr(\mathcal{O}_n^{(1)} = k_1, \dots, \mathcal{O}_n^{(r)} = k_r) x_1^{k_1} \dots x_r^{k_r} z^n$$

From there

$$\mathbf{E}\left(\mathcal{O}_n^{(1)} \times \cdots \times \mathcal{O}_n^{(r)}\right) = [z^n] \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_r} F(z, x_1, \dots, x_r)\Big|_{x_1 = \cdots = x_r = 1}$$

(Auto)-Correlation Set

auto-correlation

$$h = ababa \quad \rightsquigarrow \quad \begin{array}{l} ababa \\ ababa| \\ \quad aba\,ba \\ \qquad a\,baba \end{array} \quad \rightsquigarrow \quad \mathcal{C}_{ababa,ababa} = \{\epsilon, ba, baba\}$$

$$\mathcal{C}_{h,h} = \{\ w, \quad h.w = r.h \quad \text{and} \quad |w| < |h|\ \}$$

correlation

$$\mathcal{C}_{h_1,h_2} = \{\ w, \quad h_1.w = r.h_2 \quad \text{and} \quad |w| < |h_2|\ \}$$

$$h_1 = baba,\ h_2 = abaaba \quad \longrightarrow \quad \mathcal{C}_{baba,abaaba} = \{aba, baaba\}$$

Generating function of a language

language = set of words

alphabet $\mathcal{A} = \{a, b\}$

$\mathcal{A}^\star = \epsilon + \mathcal{A} + \mathcal{A}^2 + \cdots + \mathcal{A}^n + \ldots$ all the words

$\mathcal{L} \subset \mathcal{A}^\star \quad \rightsquigarrow \quad F_{\mathcal{L}}(a,b) = \sum_{w \in \mathcal{L}} \text{commute}(w)$

$(aabaa)^\star = \epsilon + aabaa + (aabaa)^2 + (aabaa)^3 + \cdots$

$$\mathcal{L} = (aabaa)^\star + bbb \quad \Longrightarrow \quad F_{\mathcal{L}}(a,b) = \frac{1}{1 - a^4 b} + b^3$$

if $\mathcal{X}.\mathcal{Y}$ non ambiguous, $F_{\mathcal{X}.\mathcal{Y}}(a,b) = F_{\mathcal{X}}(a,b) \times F_{\mathcal{Y}}(a,b)$

if $\mathcal{X}$ and $\mathcal{Y}$ disjoint, $F_{\mathcal{X}+\mathcal{Y}}(a,b) = F_{\mathcal{X}}(a,b) + F_{\mathcal{Y}}(a,b)$

if $\mathcal{X}^\star$ non ambiguous, $F_{\mathcal{X}^\star}(a,b) = \dfrac{1}{1 - F_{\mathcal{X}}(a,b)}$

Weighted and Counting Generating Function

Generating function of the language $\mathcal{L}$ $\quad M(a,b) = \sum_{\alpha\in\mathcal{L}} \text{commute}(\alpha)$

Weighted generating function $W(z) = M(\omega_a z, \omega_b z) = \sum_{\alpha\in\mathcal{L}} p_\alpha z^{|\alpha|} = \sum \pi_n z^n$

$\omega_a = \Pr(a)$, $\omega_b = \Pr(b)$, p_α proba. of word α, π_n proba. that a word of size n belongs to $\mathcal{L}$

Counting generating function $F(z) = M(z,z) = \sum_{\alpha\in\mathcal{L}} z^{|\alpha|} = \sum f_n z^n$

f_n number of words of the language of size n

Example

$\mathcal{L} = \{\epsilon, aa, ab, ba, aaab\}$ $\quad$ (ϵ empty word)

$$\Rightarrow \begin{cases} M(a,b) = 1 + a^2 + 2ab + a^3 b \\ F(z) = 1 + 3z^2 + z^3 \end{cases}$$

Formal Languages Analysis

(Régnier-Szpankowski - 1998)

"parse" the text with respect to the occurrences

Right $\mathcal{R}$ – set of texts obtained by reading up to the first occurrence

Minimal $\mathcal{M}$ – set of texts separating two occurrences

Ultimate $\mathcal{U}$ – set of texts following the last occurrence

Not $\mathcal{N}$ – set of texts with no occurrence

$$\mathcal{A}^\star = \mathcal{N} + \mathcal{R}.(\mathcal{M})^\star.\mathcal{U} \quad \Rightarrow \quad \mathcal{L}_x = \mathcal{N} + \mathcal{R}x.(\mathcal{M}x)^\star.\mathcal{U}$$

Equations over the langages

$\mathcal{C} = \mathcal{C}_{h,h} \qquad \pi_h = \Pr(h)$ (Bernoulli model)

(I) $\mathcal{A}^\star = \mathcal{U} + \mathcal{M}\mathcal{A}^\star$ (II) $\mathcal{A}^\star h = \mathcal{R}.\mathcal{C} + \mathcal{R}.\mathcal{A}^\star.h$

(III) $\mathcal{M}^+ = \mathcal{A}^\star.h + \mathcal{C} - \epsilon$ (IV) $\mathcal{N}.\mathcal{A} = \mathcal{R} + \mathcal{N} - \epsilon$

solving

$$R(z) = \frac{\pi_h z^{|h|}}{\pi_h z^{|h|} + (1-z)C(z)} \qquad U(z) = \frac{1}{\pi_h z^{|h|} + (1-z)C(z)}$$

$$N(z) = \frac{C(z)}{\pi_h z^{|h|} + (1-z)C(z)} \qquad M(z) = 1 + \frac{z-1}{\pi_h z^{|h|} + (1-z)C(z)}$$

$$L(z,x) = \frac{1}{1 - z + \pi_h z^{|h|} \dfrac{1-x}{x + (1-x)C(z)}}$$

Reduced sets (Régnier)

$\mathcal{R}_i, \mathcal{M}_{i,j}, \mathcal{U}_i \rightsquigarrow R_i(z), M_{i,j}(z), U_i(z)$

functions of $C_{h_1,h_1}(z), C_{h_2,h_2}(z), C_{h_1,h_2}(z), C_{h_2,h_1}(z)$

$$F(z, x_1, x_2) = N(z) + (x_1 R_1(z), x_2 R_2(z)) \begin{pmatrix} x_1 M_{1,1}(z) & x_2 M_{1,2}(z) \\ x_1 M_{2,1}(z) & x_2 M_{2,2}(z) \end{pmatrix}^\star \begin{pmatrix} U_1(z) \\ U_2(z) \end{pmatrix}$$

This collapses in case of non-reduced sets

Aho-Corasick

– **Input:** non-reduced set of words $\mathcal{U}$.

– **Output:** automaton $\mathcal{A}_{\mathcal{U}}$ recognizing $\mathcal{A}^*\mathcal{U}$.

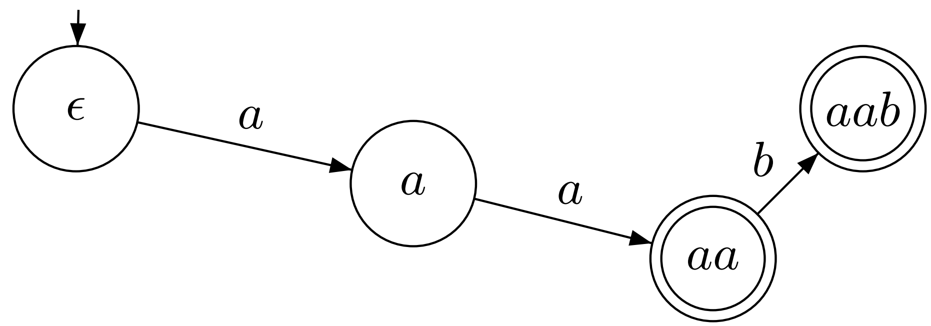
Algorithm:

1. build $\mathcal{T}_{\mathcal{U}}$, the ordinary trie representing the set $\mathcal{U}$
2. build $\mathcal{A}_{\mathcal{U}} = (\mathcal{A}, Q, \delta, \epsilon, T)$:
   – $Q = \mathrm{Pref}(\mathcal{U})$
   – $T = \mathcal{A}^*\mathcal{U} \cap \mathrm{Pref}(\mathcal{U})$
   – $\delta(q, x) = \begin{cases} qx & \text{if } qx \in \mathrm{Pref}(\mathcal{U}), \\ \mathrm{Border}(qx) & \text{otherwise,} \end{cases}$

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

Example

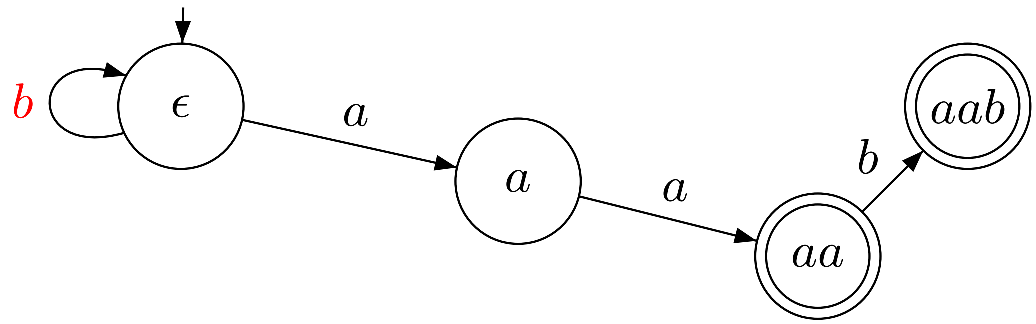
$\mathcal{U} = \{aab, aa\}$



Trie $\mathcal{T}_{\mathcal{U}}$ of $\mathcal{U}$

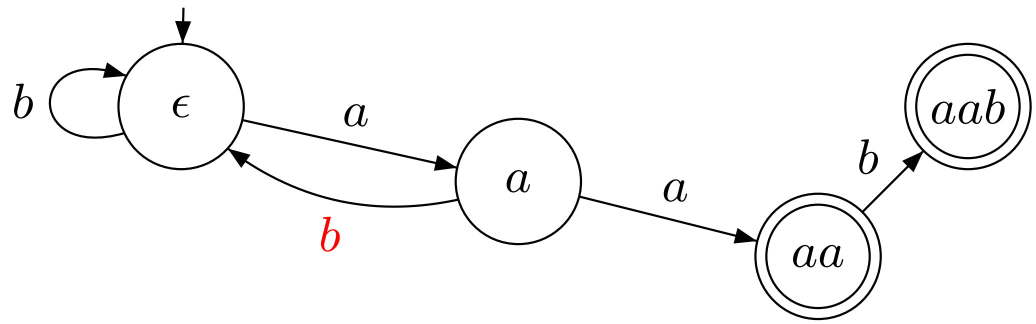
Example

$\mathcal{U} = \{aab, aa\}$ $\qquad$ $\delta(\epsilon, b) = \text{Border}(b) = \epsilon$



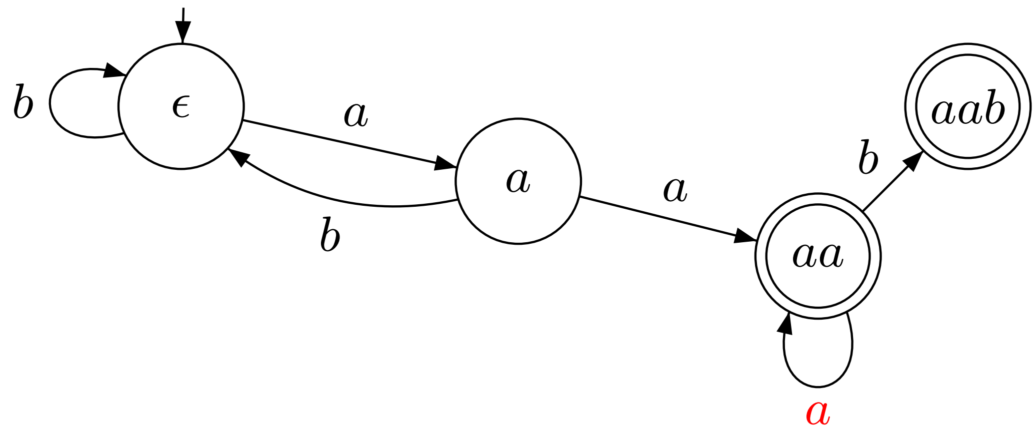
Example

$\mathcal{U} = \{aab, aa\}$ $\quad$ $\delta(a, b) = \text{Border}(a.b) = \epsilon$



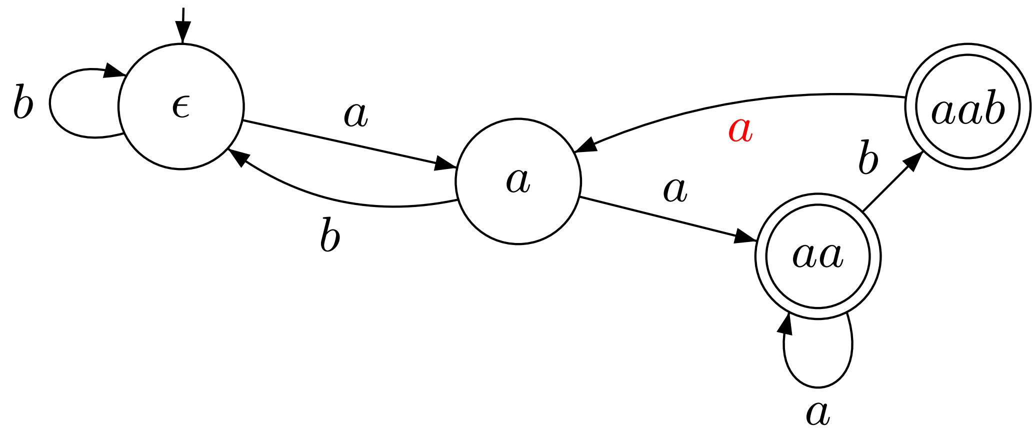
Example

$\mathcal{U} = \{aab, aa\}$ $\delta(aa, a) = \text{Border}(aa.a) = aa$



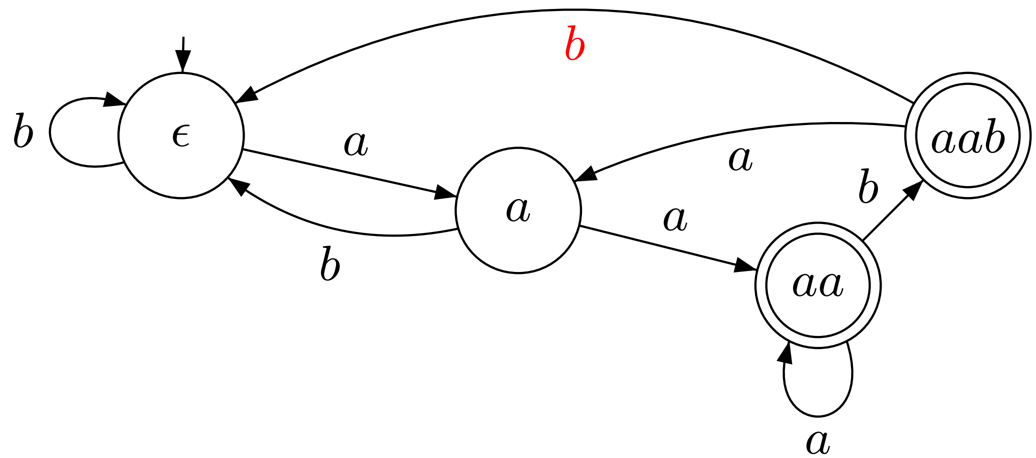
Example

$\mathcal{U} = \{aab, aa\}$ $\quad\quad$ $\delta(aab, a) = \text{Border}(aab.a) = a$



Example

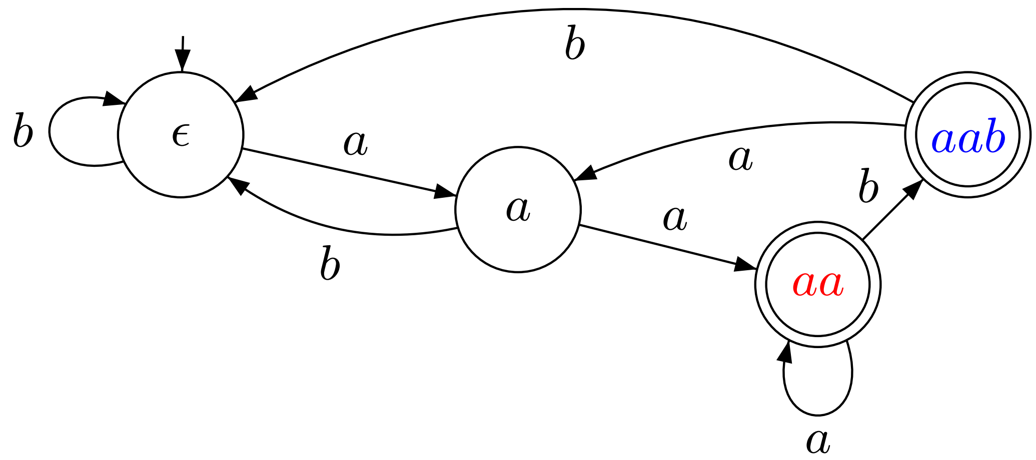
$\mathcal{U} = \{aab, aa\}$ $\quad\quad$ $\delta(aab, b) = \text{Border}(aab.b) = \epsilon$



Example

$\mathcal{U} = \{aab, aa\}$

$$\mathbb{T}(x_1, x_2) = \begin{pmatrix} b & a & 0 & 0 \\ b & 0 & ax_2 & 0 \\ 0 & 0 & ax_2 & bx_1 \\ b & a & 0 & 0 \end{pmatrix},$$

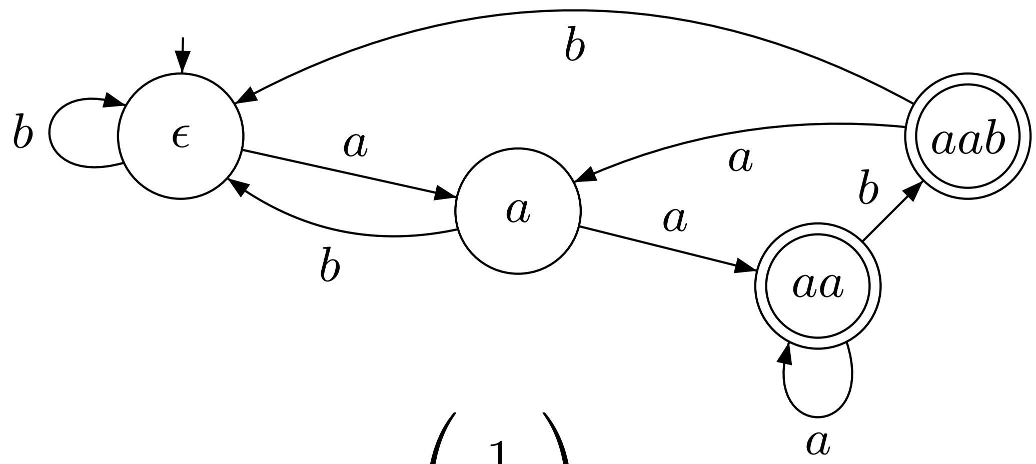


x_1, x_2 marks for aab, aa

Example

$\mathcal{U} = \{aab, aa\}$

$$\mathbb{T}(x_1, x_2) = \begin{pmatrix} b & a & 0 & 0 \\ b & 0 & ax_2 & 0 \\ 0 & 0 & ax_2 & bx_1 \\ b & a & 0 & 0 \end{pmatrix},$$

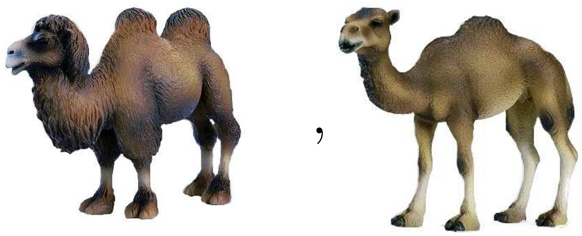


$$F(a, b, x_1, x_2) = (1, 0, 0, 0)(\mathbb{I} - \mathbb{T}(a, b, x_1, x_2))^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

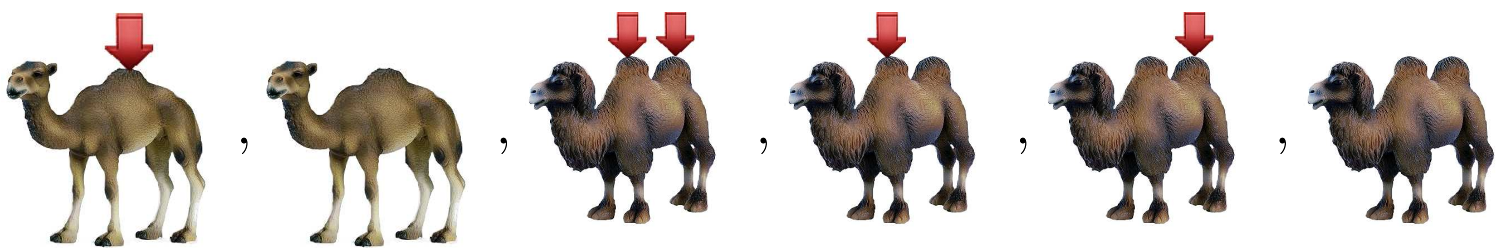
$$= \frac{1 - a(x_2 - 1)}{1 - ax_2 - b + ab(x_2 - 1) - a^2 b x_2 (x_1 - 1)^2}.$$

Inclusion-Exclusion Principle - Analytic Version

Set of *camelus genus* (camel and dromedary); the number of humps is counted by the formal variable x.

$\mathcal{F} = \{$  $\},\qquad F(x) = x^2 + x$

$\Phi = \{$"objects of $\mathcal{P}$ in which each elementary configuration (hump) is either distinguished or not"$\}$

$= \{$  $\}$

$$\Phi(t) = t + 1 + t^2 + t + t + 1 = 2 + 3t + t^2 = F(1 + t)$$

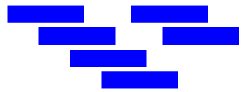
Inclusion-Exclusion principle

If $\Phi(t)$ is easy to get, then $F(x) = \Phi(x - 1)$.

Application: counts for one word

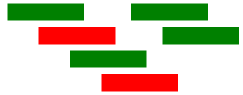
word $\color{red}{aaa}$ $\qquad$ $f(x)$: unknown p.g.f of counts of aaa

$bbbbbaaaaaaaabbbbb$

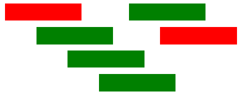


each occurrence is distinguished or not (flip-flop) $\Rightarrow$ 2^k configurations for a text with k occurrences

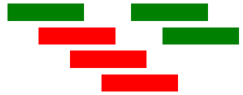
$bbbbbaaaaaaaabbbbb$



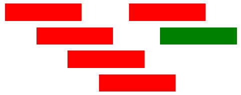
$bbbbbaaaaaaaabbbbb$

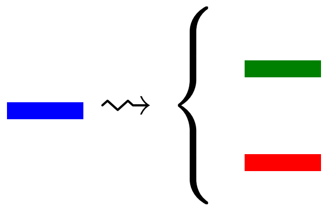


$bbbbbaaaaaaaabbbbb$



$bbbbbaaaaaaaabbbbb$





$$x \rightsquigarrow \begin{cases} 1 \\ +x \end{cases} \qquad f(x) \rightsquigarrow f(1+x) = \phi(x) \rightsquigarrow f(x) = \phi(x-1)$$

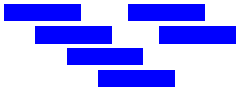
computing easier $\phi(t)$ **and substituting** $t \rightsquigarrow x-1$ **give harder** $f(x)$

(Inclusion-Exclusion paradigm)

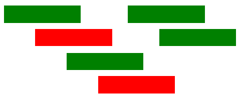
One word - Clusters

word *aaa* $\qquad C_{aaa,aaa} = \{\epsilon, a, aa\}$

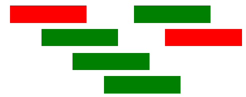
$bbbbbaaaaaaabbbbb$



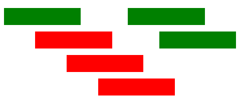
$bbbbbaaaaaaabbbbb$



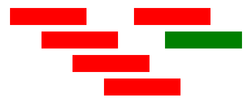
$bbbbbaaaaaaabbbbb$



$bbbbbaaaaaaabbbbb$



$bbbbbaaaaaaabbbbb$



clusters $\mathfrak{C}$

$$\begin{aligned} \mathfrak{C}_{aaa} &= aaa\bullet(\epsilon + a\bullet + aa\bullet + a\bullet a\bullet + a\bullet a\bullet a\bullet + a\bullet aa\bullet + aa\bullet a\bullet + \dots) \\ &= aaa\bullet \left(\epsilon + \left(\left(C_{aaa,aaa} - \epsilon\right)\bullet\right)^{+}\right) \end{aligned}$$

double counting (further removed by the inclusion-exclusion principle):

$$\begin{aligned} \left(C_{aaa,aaa} - \epsilon\right)^{+}(z) = \frac{z+z^2}{1-(z+z^2)} &= z + 2z^2 + 3z^3 + 5z^4 + 8z^5 + 13z^6 + \dots \\ &\neq z + z^2 + z^3 + z^4 + z^5 + z^6 + \dots \end{aligned}$$

Word aaa - Clusters - Generating function

$$C_{aaa,aaa} = \{\epsilon, a, aa\} \qquad C_{aaa,aaa}(z) = 1 + z + z^2$$

$$\begin{aligned}\mathfrak{C}_{aaa} &= aaa\bullet(\epsilon + a\bullet + aa\bullet + a\bullet a\bullet + a\bullet a\bullet a\bullet + a\bullet aa\bullet + aa\bullet a\bullet + \dots) \\ &= aaa\bullet\left(\epsilon + ((C_{aaa,aaa} - \epsilon)\bullet)^{+}\right)\end{aligned}$$

$$\begin{aligned}\mathfrak{C}_{aaa}(z, x) &= zzzx(1 + zx + zzx + zxzx + zxzxzx + zxzzx + zzxzx + \dots) \\ &= z^3 x\left(\epsilon + (C_{aaa,aaa}(z) \times x)^{+}\right) \\ &= xz^3\left(1 + \frac{xz + xz^2}{1 - (xz + xz^2)}\right) = \frac{xz^3}{1 - (xz + xz^2)}\end{aligned}$$

Parsing of a text with respect to clusters

word h, $\qquad \mathcal{C} = \mathcal{C}_{h,h}$, $\qquad$ clusters $\mathfrak{C}$

$$\mathfrak{C} = h + h.\mathcal{C} + h\mathcal{C}\mathcal{C} + h\mathcal{C}\mathcal{C}\mathcal{C} + \dots \quad \Longrightarrow \quad \mathfrak{C}(z,x) = \frac{xh(z)}{1 - x(\mathcal{C}(z) - 1)}$$

When reading a random text T, at each position, either we read a letter of the alphabet A, either we begin a cluster $\mathfrak{C}$,

$$T = \epsilon + A + \mathfrak{C} + AA + A\mathfrak{C} + \mathfrak{C}A + \mathfrak{C}\mathfrak{C} + AAA + AA\mathfrak{C} + A\mathfrak{C}A + \mathfrak{C}AA + A\mathfrak{C}\mathfrak{C} + \dots$$
$$= \mathrm{Seq}(A + \mathfrak{C})$$

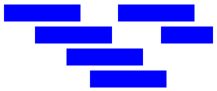
Therefore, counting with x the number of occurrences of the word h, we have, removing double counting by inclusion-exclusion,

$$F(z,x) = \frac{1}{1 - \big(A(z) + \mathfrak{C}(z, x-1)\big)} = \frac{1}{1 - A(z) - \dfrac{(x-1)h(z)}{1 - (x-1)(\mathcal{C}(z) - 1)}}$$

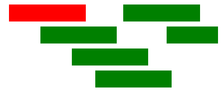
Reduced set - (Goulden-Jackson - 1979, 1983)

$\mathcal{U} = \{aba, bab, aa\}$

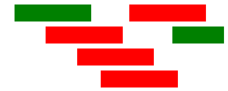
$bbbbbababaabbbbb$



$bbbbbababaabbbbb$



$bbbbbababaabbbbb$



clusters $\mathfrak{C}_{i,j}$ begin with w_i and finish with w_j

$$\mathfrak{C}_{i,j} = w_i \mathcal{C}_{w_i,w_j} + \sum_{1 \le k \le 3} \mathfrak{C}_{i,k}.(\mathcal{C}_{w_k,w_j} - \delta_{kj}\epsilon)$$

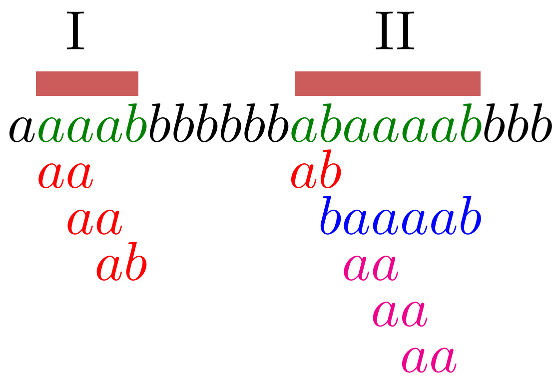
$$\mathfrak{C} = (w_1 \bullet, w_2 \bullet, w_3 \bullet) \left(\mathbf{I} - \begin{pmatrix} \mathcal{C}_{w_1,w_1}\bullet - \epsilon & \mathcal{C}_{w_1,w_2}\bullet & \mathcal{C}_{w_1,w_3}\bullet \\ \mathcal{C}_{w_2,w_1}\bullet & \mathcal{C}_{w_2,w_2}\bullet - \epsilon & \mathcal{C}_{w_2,w_3}\bullet \\ \mathcal{C}_{w_3,w_1}\bullet & \mathcal{C}_{w_3,w_2}\bullet & \mathcal{C}_{w_3,w_3}\bullet - \epsilon \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathcal{T} = \mathrm{Seq}(\mathcal{A} + \mathfrak{C}) \Longrightarrow \Phi(z, x_1, x_2, x_3) = \frac{1}{1 - A(z) - \mathfrak{C}(z, x_1, x_2, x_3)}$$

$$F(z, x_1, x_2, x_3) = \Phi(z, x_1 - 1, x_2 - 1, x_3 - 1) = \frac{1}{1 - A(z) - \mathfrak{C}(z, x_1 - 1, x_2 - 1, x_3 - 1)}$$

General Case: Non Reduced Set of Words

$\mathcal{U} = \{aa, ab, baaaab\}$



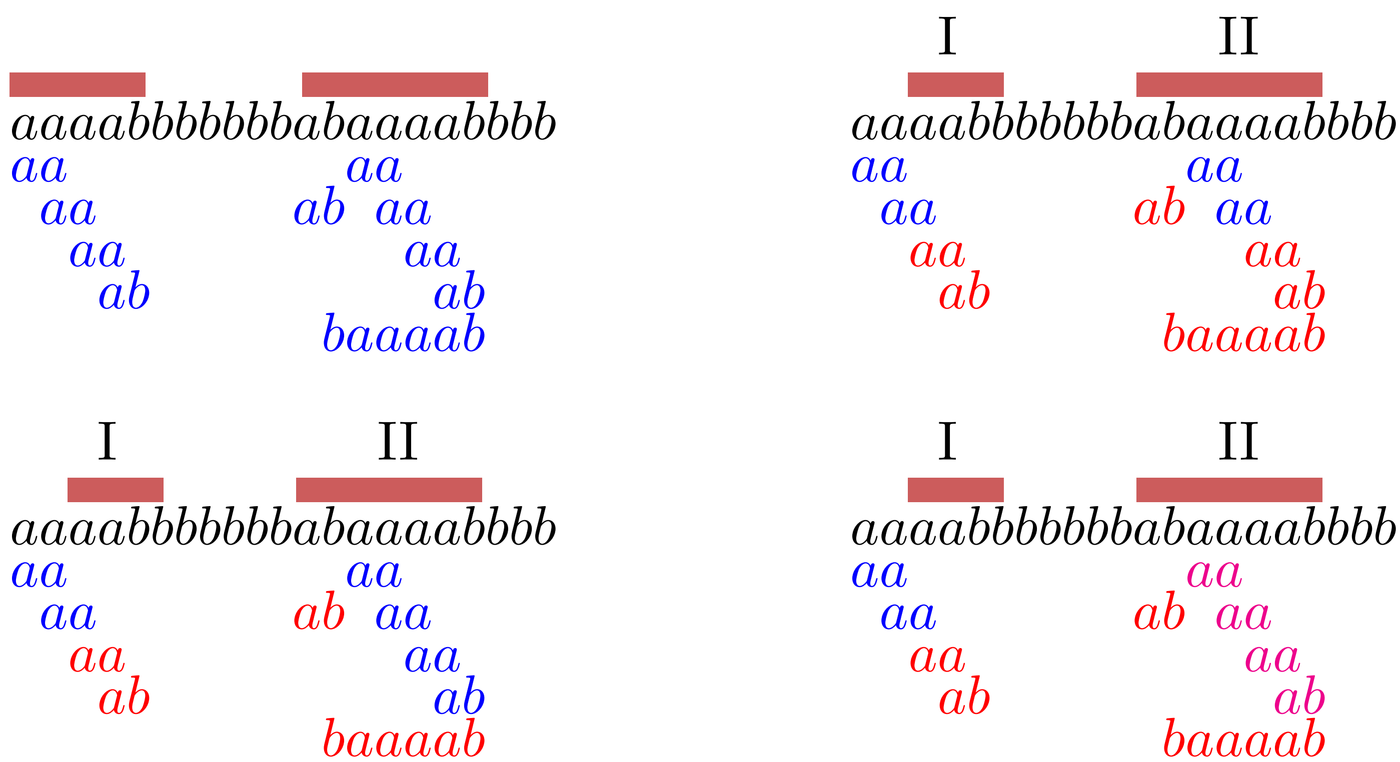
create **clusters** of **distinguished occurrences**

Reduced Cluster, no induced factor occurrences (Cluster I). Count distinguished occurrences by $t_i \rightsquigarrow x_i - 1$ (Inclusion-Exclusion principle)

Induced Factor Occurrences, occurrence *baaaab* of reduced Cluster II induces 0, 1, 2, or 3 distinguished occurrences *aa*. To recover the correct count of 8 marked configurations, count them by $(1 + t_i)^3 \rightsquigarrow x_i^3$.

Inclusion-Exclusion: Non-Reduced Case

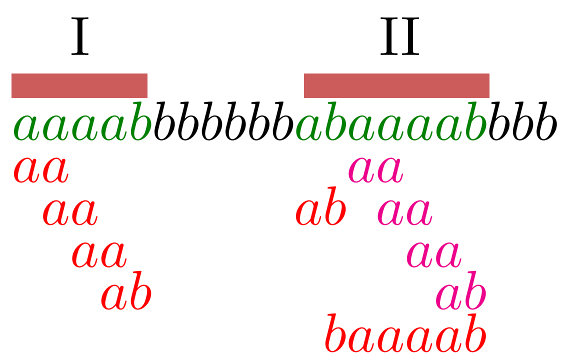
$\mathcal{U} = \{u_1 = aa, u_2 = ab, u_3 = baaaab\}$



1. select **distinguished** occurrences giving **clusters**
2. **forget induced factor** occurrences to get **reduced clusters**
3. **count induced factor occurrences**

Counting Occurrences

$\mathcal{U} = \{u_1 = aa, u_2 = ab, u_3 = baaaab\}$



– **Reduced Cluster I** : $f(t_1, t_2, t_3) = t_1^3 t_2$

distinguished: t_i

– **Cluster II**: $f(t_1, t_2, t_3) = t_2(1 + t_2)(1 + t_1)^3 t_3$

1. **distinguished** and **reduced**: t_i
2. **induced**: $(1 + t_i)$

Right Extension Sets and Matrices

Right Extension Set of a pair of words (h_1, h_2)

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

if $h_1 \neq h_2$ have no factor relation, $\mathcal{E}_{h_1,h_2} = \mathcal{C}_{h_1,h_2}$ but $\mathcal{E}_{h,h} = \mathcal{C}_h - \epsilon$

Right Extension Matrix of a vector of words $\mathbf{u} = (u_1, \ldots, u_r)$

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

Examples

$$\mathbf{u}_1 = (aba, ab) \;\Rightarrow\; \mathcal{E}_{\mathbf{u}_1} = \begin{pmatrix} ba & b \\ \emptyset & \emptyset \end{pmatrix} \qquad \mathcal{E}_{ab,aba} = \emptyset \quad \begin{cases} aba = |aba \\ e' = \epsilon \notin \mathcal{A}^+ \end{cases}$$

$$\mathbf{u}_2 = (aaaa, aaa) \;\Rightarrow\; \mathcal{E}_{\mathbf{u}_2} = \begin{pmatrix} a+a^2+a^3 & a+a^2 \\ a^2+a^3 & a+a^2 \end{pmatrix} \quad \begin{cases} a \notin \mathcal{E}_{aaa,aaaa} & aaa.a = |aaaa \\ aa \in \mathcal{E}_{aaa,aaaa} & aaa.aa = a.aaaa \end{cases}$$

Counting Induced Words

$\mathcal{U} = \{u_1 = aa, u_2 = baaaabaaaab\}$ $\qquad$ $\mathcal{E}_{u_2,u_2} = \{aaaab, aaaabaaaab\}$

$$\begin{array}{l} baaaabaaaabaaaab \\ baaaabaaaabaaaab \end{array} \qquad N_{2,1}(6) = 9 - 6 = 3$$

$$\begin{array}{l} baaaabaaaabaaaab \\ baaaabaaaabaaaab \end{array} \qquad N_{2,1}(11) = 9 - 3 = 6$$

$$N_{i,j}(k) = \left|u_i\right|_j - \left|u_i[1 \ldots |u_i| - k]\right|_j.$$

$$\langle \mathcal{E}_{u_2,u_2} \rangle_2 = \pi_a^4 \pi_b z^5 (t_1 + 1)^3 t_2 + \pi_a^8 \pi_b^2 z^{10} (t_1 + 1)^6 t_2$$

Formal Setting

$N_{i,j}(k)$ counts the number of occurrences of u_j factor of u_i and ending in the last k positions of u_i

$$N_{i,j}(k) = \big|u_i\big|_j - \big|u_i[1\ldots|u_i| - k]\big|_j.$$

$\langle s \rangle_i$ **formal weight** of a **suffix** of word u_i

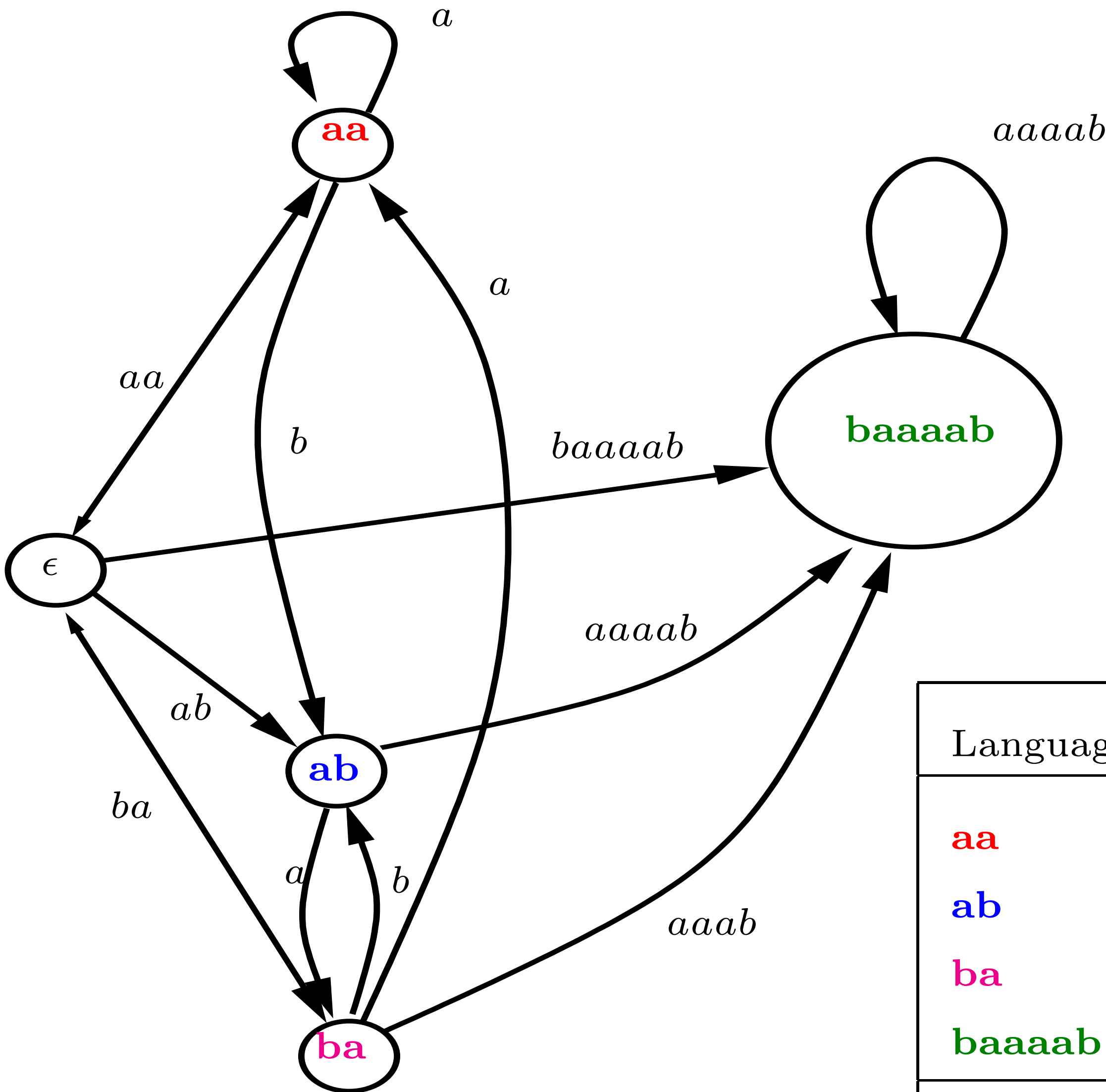
$$\langle s \rangle_i = \pi(s) z^{|s|} t_i \prod_{m \neq i} (t_m + 1)^{N_{i,m}(|s|)}.$$

extension to a set of words S which are suffixes of u_i

$$\langle S \rangle_i = \sum_{s \in S} \langle s \rangle_i.$$

$$\mathcal{E}_{i,j} \quad \rightsquigarrow \quad \langle \mathcal{E}_{i,j} \rangle_j$$

Right Extension Graph



$$\mathcal{U} = \{\mathbf{aa}, \mathbf{ab}, \mathbf{ba}, \mathbf{baaaab}\}$$

Language	G. F.
aa	$t_1 z^2$
ab	$t_2 z^2$
ba	$t_3 z^2$
baaaab	$t_4 z^6$
$\mathcal{E}_{ab,ba} = \{a\}$	$t_3 z$
$\mathcal{E}_{ba,baaaab} = \{aaab\}$	$(1+t_1)^2(1+t_2)t_4 z^4$
$\mathcal{E}_{baaaab,baaaab} = \{aaaab\}$	$(1+t_1)^3(1+t_2)t_4 z^5$

Putting Things Together

$$\text{Let} \quad \langle \mathbf{u} \rangle = (\langle u_1 \rangle_1, \ldots, \langle u_r \rangle_r) \quad \text{and} \quad \langle \mathcal{E}_{\mathbf{u}} \rangle = \begin{pmatrix} \ldots & \ldots & \ldots \\ \ldots & \langle \mathcal{E}_{i,j} \rangle_j & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}$$

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

$$\mathfrak{C}(z, \mathbf{t}) = \langle \mathbf{u} \rangle \cdot \left(\mathbb{I} - \langle \mathcal{E}_{\mathbf{u}} \rangle \right)^{-1} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

where $\mathbf{u} = (u_1, \ldots, u_r)$, $\mathbf{t} = (t_1, \ldots, t_r)$

Proposition II. *The generating function* $F(x, \mathbf{x})$ *counting matches of a non-reduced set of words is*

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

Examples

$\mathcal{U} = \{u\}$

$$\mathfrak{C}(z,t) = \frac{t\langle u\rangle}{1 - t\langle \mathcal{E}_u\rangle} = \frac{t\pi(u)z^{|u|}}{1 - t(C(z) - 1)}$$

$\mathcal{U} = \{u_1, u_2\}$

$\mathfrak{C}(z, t_1, t_2)$

$$= \frac{t_1\langle u_1\rangle_1 + t_2\langle u_2\rangle_2 - t_1 t_2\big(\langle u_1\rangle_1\big[\langle \mathcal{E}_{2,2}\rangle_2 - \langle \mathcal{E}_{1,2}\rangle_2\big] + \langle u_2\rangle_2\big[\langle \mathcal{E}_{1,1}\rangle_1 - \langle \mathcal{E}_{2,1}\rangle_1\big]\big)}{1 - t_2\langle \mathcal{E}_{2,2}\rangle_2 - t_1\langle \mathcal{E}_{1,1}\rangle_1 + t_1 t_2\big(\langle \mathcal{E}_{1,1}\rangle_1\langle \mathcal{E}_{2,2}\rangle_2 - \langle \mathcal{E}_{2,1}\rangle_1\langle \mathcal{E}_{1,2}\rangle_2\big)}$$

Algorithmic computation

Init($\mathcal{A}_{\mathcal{U}}$)

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

Build-Extension-Matrix($\mathcal{A}_{\mathcal{U}}$)

$\triangleright$ Initialize the matrix $(\mathcal{E}_{i,j})_{1 \leq i,j \leq r}$
for $i \leftarrow 1$ **to** r **do**
    for $j \leftarrow 1$ **to** r **do**
        $\mathcal{E}_{i,j} \leftarrow 0$
$\triangleright$ Compute the maps $(f_i(w))$ for $i = 1..r$ and $w \in \text{Pref}(\mathcal{U})$
$(f_i)_{1 \leq i \leq r} \leftarrow$ Init($\mathcal{A}_{\mathcal{U}}$)
$\triangleright$ Main loop
for $i \leftarrow 1$ **to** r **do**
    $v \leftarrow u_i$
    do   **for** $j \leftarrow 1$ **to** r **do**
            $\mathcal{E}_{i,j} \leftarrow \mathcal{E}_{i,j} + f_j(v)$
            $v \leftarrow \text{Border}(v)$
    while $v \neq \epsilon$
return E

Time complexity of the main loop $O(s \times r^2)$, where r is the number of words and s is the length of the longest suffix chain

(sequence $(u_1 = u, u_2 = \text{Border}(u_1), u_3 = \text{Border}(u_2), \ldots, u_s = \text{Border}(u_{s-1}) = \epsilon)$)

Complexity

	Inclusion-Exclusion	Automaton
Generating Function	$O(M(l))$	$O(l^2)$
$[z^n]$ Asymptotics	$O(l)$	$O(l)$
$[z^n]$ Exact	$O(\log(n)M(l))$	$O(\log(n)M(l))$

$M(l)$ is the cost of **multiplying by FFT two univariate polynomials of size** l and we assume that the number of words r is $o(l)$

Up-to-date FFT algorithms give

$$M(l) = O(l \log l \log \log l)$$