

Linear Computation of Unbordered Conjugate

J.-P. Duval, T. Lecroq and A. Lefebvre

Journées du GDR IM - Groupe COMATEGE
Marne-la-Vallée
26 - 27 novembre 2012

Basic definitions

$w = abaababaabaab$

Basic definitions

$$w = \boxed{a} \boxed{b} a \boxed{a b a} \boxed{b a a} \boxed{b a a} \boxed{a b}$$

↑ ↑
prefix suffix

Basic definitions

$$w = \boxed{a} \boxed{b} \boxed{a} \boxed{a} \boxed{b} \boxed{a} \boxed{b} \boxed{a} \boxed{a} \boxed{b}$$

↑ ↑
prefix suffix

Definition

$\text{Pref}(w)$ is the set of proper prefixes of w .
 $\text{Suff}(w)$ is the set of proper suffixes of w .

Basic definitions

$$w = \boxed{a}baababa\boxed{abaab}$$

↑ ↑
prefix suffix

Definition

A word w is unbordered iff $\text{Pref}(w) \cap \text{Suff}(w) = \emptyset$.

Basic definitions

 $w = \boxed{aba} \boxed{abab} \boxed{aba} \boxed{abaab}$

border

Definition

A word w is unbordered iff $\text{Pref}(w) \cap \text{Suff}(w) = \emptyset$.

Pure words

Definition

A word is *pure* iff it is unbordered or the integer power of an unbordered word.

Pure words

Definition

A word is *pure* iff it is unbordered or the integer power of an unbordered word.

Example

- abaabb
- TINTIN = (TIN)²

Conjugates

Definition

A word w' is a conjugate of word w iff it is equal to a circular permutation of symbols of w .

Example

babaab and bbabaa are conjugates of abaabb.

Our problem

Given word w , our problem is to find a pure conjugate of w (without considering any lexicographic order).

Example

- babaababaabaa is a pure conjugate of abaababaabaab.
- TINTIN is a pure conjugate of itself.

Relation between words

Definition

Given two words x and y , $x \ll y$ iff the following conditions hold:

- y is unbordered;
- $|x| \geq |y|$;
- the prefix of x of length $|y|$ is a concatenation of prefixes of y .

Example

$x = aababab$ and $y = abb$

Relation between words

Proposition

Given u an unbordered word and $v \in Pref^+(u)$. The following properties hold:

- $v \cdot u$ is an unbordered word;
- $v \cdot u \ll u$.

Corollary

Given an unbordered word u and $v \in Pref^+(u)$, for all integer $q \geq 1$, $v \cdot u^q$ is unbordered.

Relation between words

Lemma

Let x and y be two unbordered words such that $x \ll y$. If $y' \in \text{Suff}(y)$ then $x \cdot y'$ is unbordered.

Corollary

Let x be a unbordered word and $z = z_1 \cdot z_2 \cdots z_i \cdots z_m$ be a concatenation of unbordered words or suffix of unbordered words such that $x \ll z_i$ for $1 \leq i \leq m$. Then $x \cdot z$ is unbordered.

Decomposition in unbordered words

Definition

We say that $(u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v)$ is a decomposition of w in unbordered words iff u_1, \dots, u_m and u are unbordered words and $w = u_1^{q_1} \cdots u_m^{q_m} \cdot u^q \cdot v$ with $v \in \text{Pref}^*(u)$.

Canonical decomposition in unbordered words

We call *canonical decomposition in unbordered words* of w the decomposition defined as follows.

$$w[1..1] = (w[1], 1) \cdot ()$$

If $w[1..j] = (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v)$ and $w[j+1] = a$ then

$$w[1..j+1] =$$

$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q+1) \cdot () & \text{if } v \cdot a = u \end{cases} \quad (1)$$

$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a) & \text{if } v \cdot a \in \text{Pref}(u) \end{cases} \quad (2)$$

$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a) & \text{if } v \cdot a \in \text{Pref}^+(u) \setminus \text{Pref}(u) \end{cases} \quad (3)$$

$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u, q) \cdot (v \cdot a, 1)() & \text{if } v \cdot a \in \text{Pref}^+(u) \cdot u \end{cases} \quad (4)$$

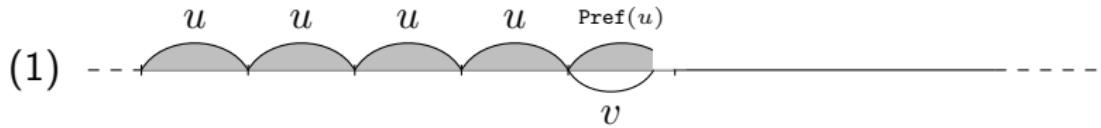
$$\begin{cases} (u_1, q_1) \cdots (u_m, q_m) \cdot (u^q \cdot v \cdot a, 1)() & \text{if } v \cdot a \notin \text{Pref}^+(u) \cup \text{Pref}^*(u) \cdot u \end{cases} \quad (5)$$

Morris-Pratt failure function

- This decomposition is computed using a Morris-Pratt failure function.
- This function enables to compute, for each prefix of a word, the length of its longest border.
- Its complexity is linear.

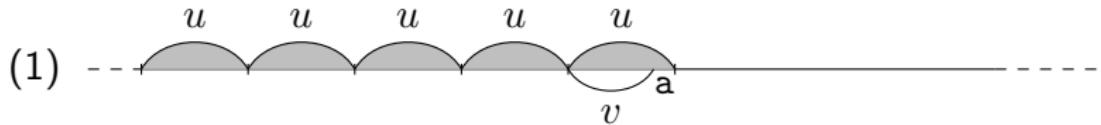
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}(u)$.



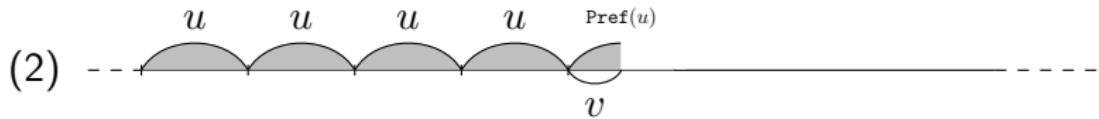
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}(u)$.



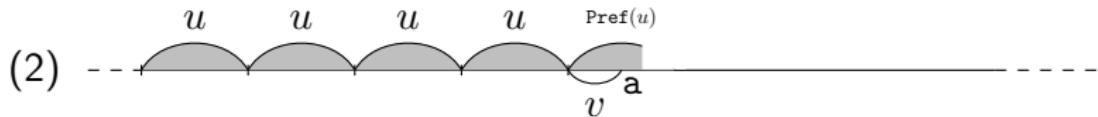
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}(u)$.



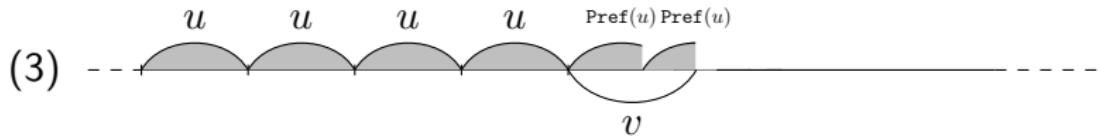
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}(u)$.



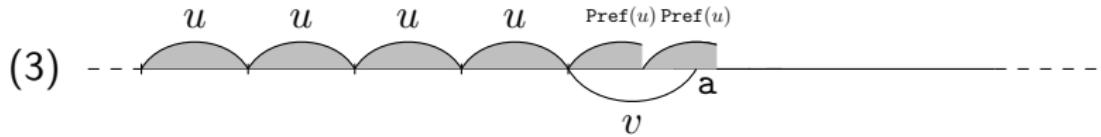
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



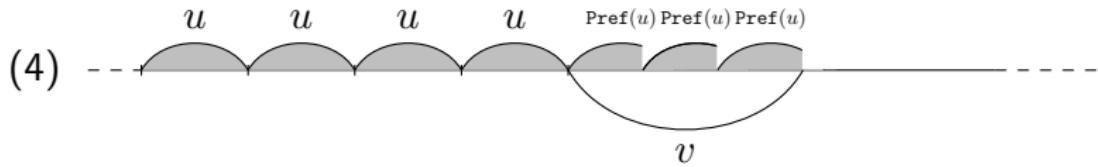
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



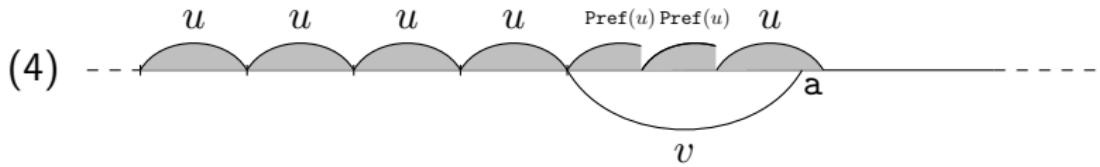
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



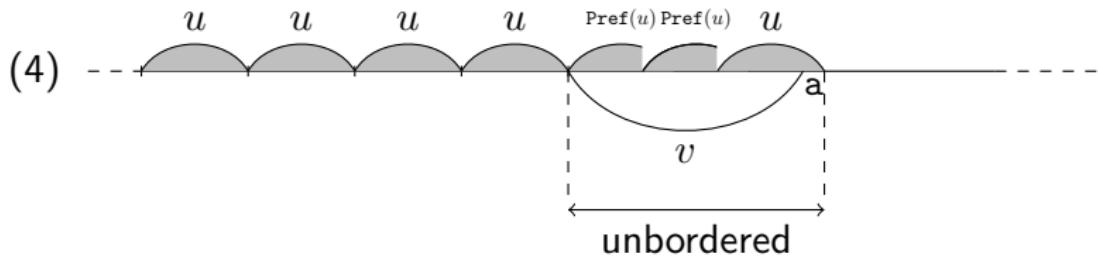
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



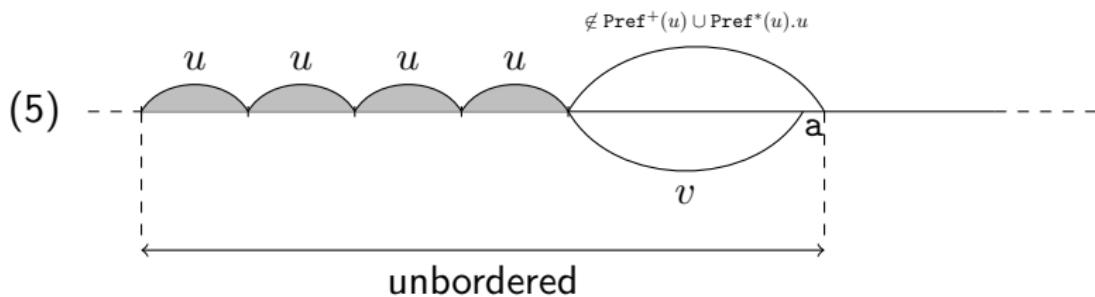
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



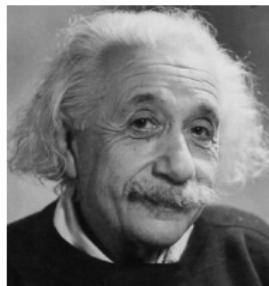
Canonical decomposition in unbordered words

u is unbordered and $v \in \text{Pref}^+(u)$.



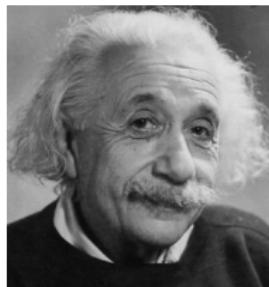
Example:

Example: "EINSTEIN"



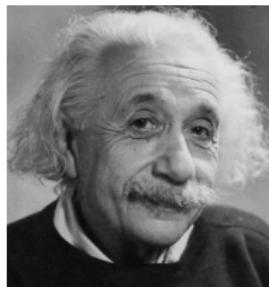
EINSTEIN

Example: "EINSTEIN"



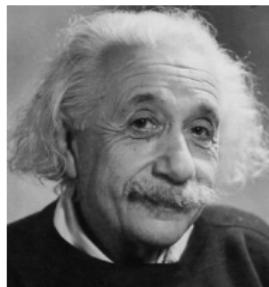
EINSTEINEINSTEIN

Example: "EINSTEIN"



j
 k'
 k | i
 u_1 EINSTEINEINSTEIN

Example: "EINSTEIN"

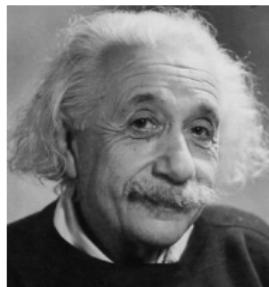


j
 k'
 k i
 u_1

EINSTEINEINSTEIN

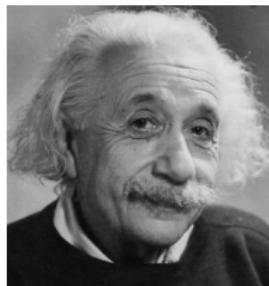
A diagram illustrating the unbordered conjugate of the word "EINSTEIN". The letters are arranged in two rows: "EINSTEIN" on top and "EINSTEIN" on the bottom. Above the first letter "E" on the top row, there is a vertical bracket with labels: "j" at the top, "k'" in the middle, and "k" at the bottom. To the right of the "E" on the bottom row, there is another vertical bracket with labels: "i" at the top and "u1" at the bottom. The letters "E", "I", "N", "S", "T", "E", "I", "N" are in black, while "E", "I", "N", "S", "T", "E", "I", "N" are in red.

Example: "EINSTEIN"



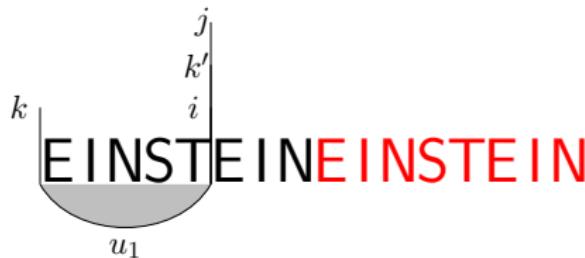
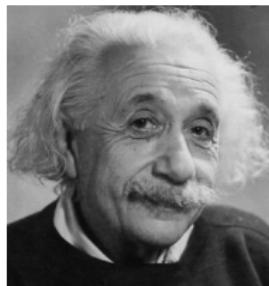
j
 k'
 k |
 i
EINSTEINEINSTEIN
 u_1

Example: "EINSTEIN"

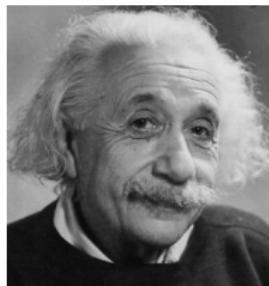


k j
 k'
 i
EINSTEINEINSTEIN
 u_1

Example: "EINSTEIN"



Example: "EINSTEIN"



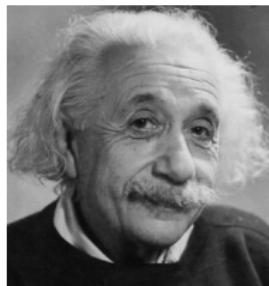
k k' j
 i

E I N S T E I N E I N S T E I N

u_1

A diagram illustrating the Unbordered Conjugate of the word "EINSTEIN". The letters are arranged in two rows. The first row contains "E", "I", "N", "S", "T", "E", "I", "N", "E", "I", "N", "S", "T", "E", "I", "N". The second row contains "k", "i", "j", "k'", "l", "m", "n", "o", "p", "q", "r", "s", "t", "u", "v". Brackets connect corresponding letters between the two rows: (E,k), (I,i), (N,j), (S,k'), (T,l), (E,m), (I,n), (N,o), (E,p), (I,q), (N,r), (S,t), (T,u), (E,v). A grey semi-circle is positioned below the first four letters of the first row, labeled "u1".

Example: "EINSTEIN"



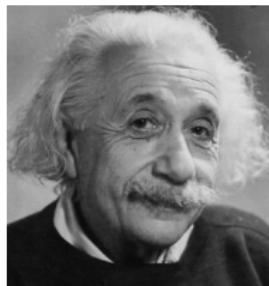
k k' j
 i

EINSTEINEINSTEIN

u_1

A diagram illustrating the word "EINSTEIN" where each letter is split vertically. The left side of the first "E" is labeled "k", the right side is "k'", and the middle vertical line is "i". The left side of the second "E" is "j", the right side is "i", and the middle vertical line is "j". Below the letters, the symbol " u_1 " is positioned under the first "E".

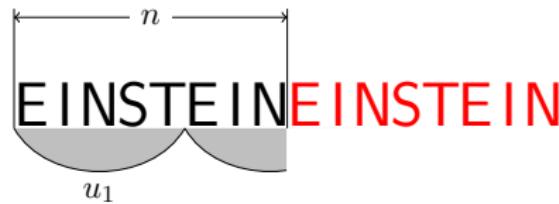
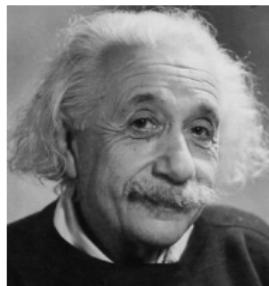
Example: "EINSTEIN"



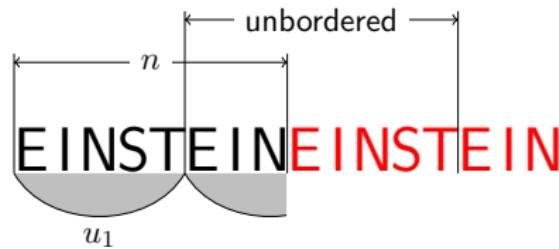
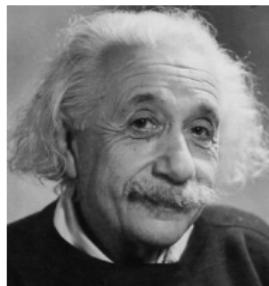
k k' j
 i |
EINSTEINEINSTEIN
 u_1

The diagram illustrates the word "EINSTEIN" in a bold, black font. The letters are arranged in two rows: "EINSTEIN" on top and "EINSTEIN" on the bottom. Above the first letter "E" on the top row, there is a vertical line segment with endpoints labeled k (on the left) and k' (on the right). Above the second letter "E" on the top row, there is another vertical line segment with endpoints labeled i (on the left) and j (on the right). Below the first letter "E" on the bottom row, there is a horizontal line segment with endpoints labeled u_1 (on the left) and u_2 (on the right). The letters on the bottom row are colored red.

Example: "EINSTEIN"



Example: "EINSTEIN"



A classical case:

A classical case: "Fibonacci"



abaababaabaab

A classical case: "Fibonacci"



abaababaabaab**abaababaabaab**

A classical case: "Fibonacci"



j
 k'
 k i

abaababaabaab**abaababaabaab**

A sequence of letters 'a' and 'b' is shown. Above the sequence, there are four labels: 'j' at the top, 'k'' below it, 'k' further down, and 'i' at the bottom right. A bracket under the labels 'k' and 'i' spans the width of the sequence. The sequence itself consists of several pairs of 'a' and 'b', followed by a red-colored segment where the pattern repeats.

A classical case: "Fibonacci"



k j
 k'
 i

u_1

abaababaabaab**abaababaabaab**

A sequence of letters 'a' and 'b' is shown. The first 12 letters are in black, and the next 6 letters are in red. To the left of the sequence, there is a bracket labeled u_1 at the bottom, i above u_1 , k to the left of the bracket, k' above i , and j above k' . There is also a small gap between the two groups of letters.

A classical case: "Fibonacci"



k k' j
 i

u_1

abaababaabaab**abaababaabaab**

A diagram illustrating a word over the alphabet {a, b} with indices k, k', i, j and a suffix u_1. The word starts with 'aba' followed by a sequence of alternating 'a's and 'b's. The indices k and k' are placed above the first two 'a's, while i is placed below the first 'a'. The indices j and u_1 are placed above the last two 'a's. The suffix 'u_1' is placed below the last 'a'.

A classical case: "Fibonacci"



k k' j
 i u_1

abaababaabaab**abaababaabaab**

A diagram illustrating a word over the alphabet {a, b} with indices k, k', i, j and a suffix u_1. The word starts with 'aba' followed by a sequence of alternating 'a's and 'b's. The indices k and k' are placed above the first two 'a's. Index i is placed below the first 'b'. Index j is placed above the second 'a' of the suffix. The suffix u_1 is placed below the last 'a' of the suffix.

A classical case: "Fibonacci"



abaabababaabaab**abaabababaabaab**

k j
 u_1 u_1
 k' i

A classical case: "Fibonacci"



abaababaabaabab**abaabababaabaab**

The string 'abaababaabaabab' is shown in black. A red segment 'aba' is highlighted. Above the string, indices k , k' , i , and j are placed above specific letters. Below the string, regions u_1 and u_2 are indicated by grey-shaded areas under the first two 'a's and the last two 'a's respectively.

$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_2 \ll u_1$$

A classical case: "Fibonacci"



$$u_2 \ll u_1$$

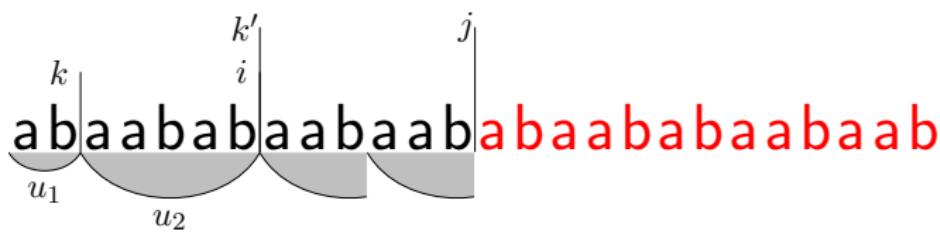
A classical case: "Fibonacci"



u_1 u_2

$$u_2 \ll u_1$$

A classical case: "Fibonacci"



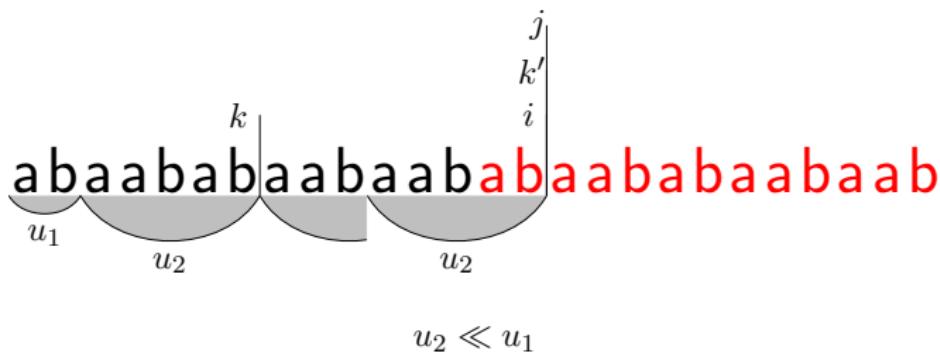
$$u_2 \ll u_1$$

A classical case: "Fibonacci"

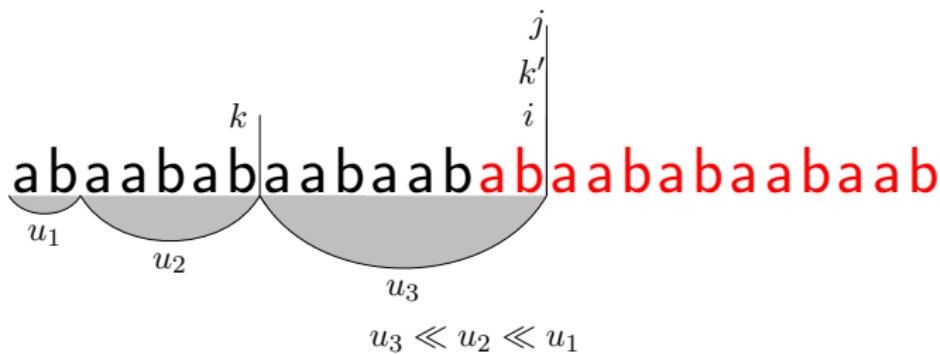


$$u_2 \ll u_1$$

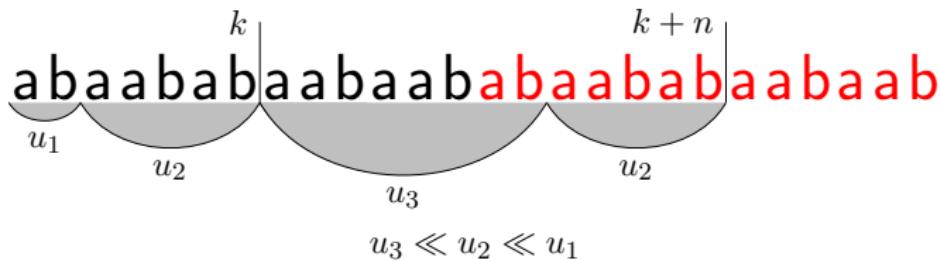
A classical case: "Fibonacci"



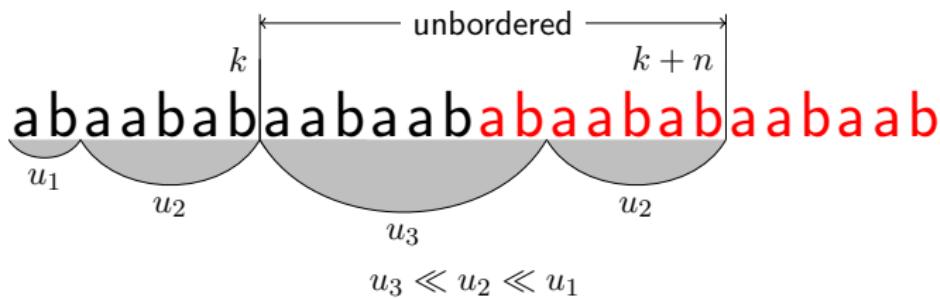
A classical case: "Fibonacci"



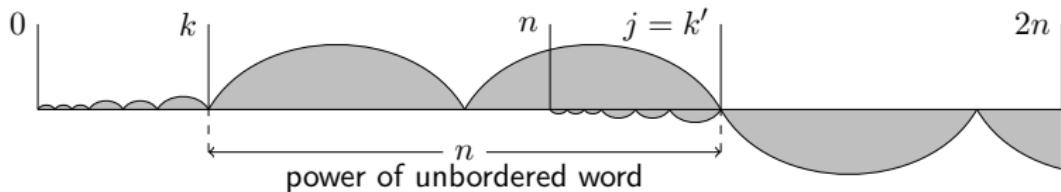
A classical case: "Fibonacci"



A classical case: "Fibonacci"

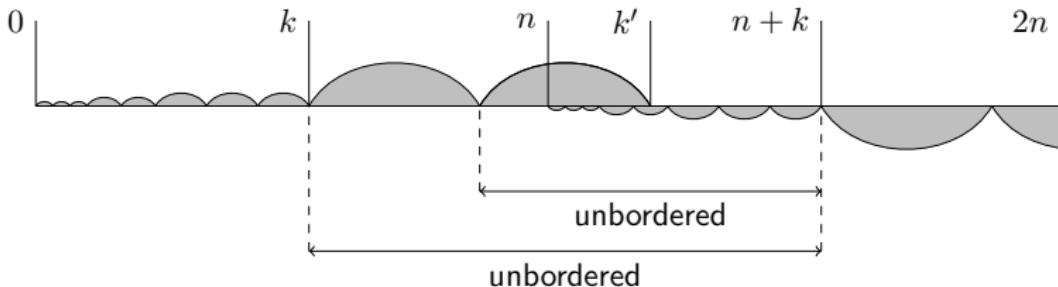


Stop case 1



Case where $w[k + 1..k + n]$ is a power of an unbordered word.

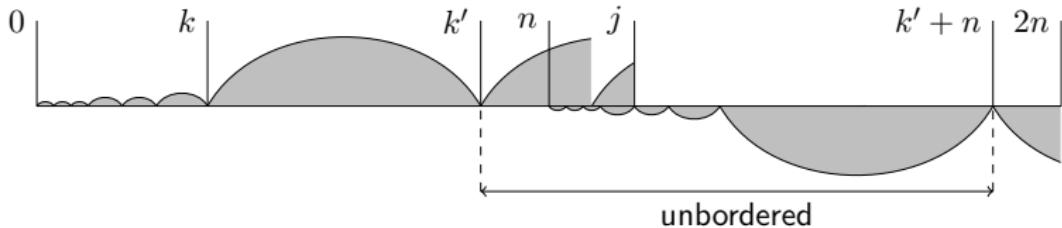
Stop case 2



Corollary

Let x be a unbordered word and $z = z_1 \cdot z_2 \cdots z_i \cdots z_m$ be a concatenation of unbordered words or suffix of unbordered words such that $x \ll z_i$ for $1 \leq i \leq m$. Then $x \cdot z$ is unbordered.

Stop case 3



Corollary

Given an unbordered word u and $v \in \text{Pref}^+(u)$, for all integer $q \geq 1$, $v \cdot u^q$ is unbordered.

Complexity

- two copies of w are necessary
- the exact number of symbols comparisons performed by the Morris-Pratt failure function f' on a word of length n is bounded by $2n - 2 - f'(n) - r$ where $r = \#\{j|f'(j) = 0\}$
- the total complexity is bounded by $4n$

The end

Questions?