

Combinatoire des mots

Alphabet et mots

alphabet A : ensemble fini non vide de lettres
(symboles, caractères)

Exemples

ADN : acides nucléiques, paires de bases, bases

ARN : acides nucléiques

protéines : acides aminés

mot : suite finie de lettres

mot vide : suite de zéro lettre notée ε

Exemple

$$A = \{ a, c, g, t \}$$

$\varepsilon, a, t, tata$ sont des mots sur l'alphabet A

A^* : ensemble de tous les mots finis sur
l'alphabet A (monoïde libre)

A^+ : ensemble de tous les mots finis non vides
sur l'alphabet A

La longueur d'un mot $x \in A^*$ est définie comme étant le nombre de lettres du mot x .

Elle est notée $|x|$.

On note $x[i]$ pour $i = 0, 1, \dots, |x|-1$, la lettre du mot x à l'indice i (la numérotation commence en 0).

Pour $x \neq \varepsilon$, chaque indice $i = 0, 1, \dots, |x|-1$ est une position sur x .

$$x = x[0]x[1]\dots x[|x|-1]$$

D'où une définition élémentaire d'identité entre deux mots quelconques x et y :

$$x = y$$



$$|x| = |y|$$

et

$$x[i] = y[i] \text{ pour } i = 0, 1, \dots, |x|-1$$

L'ensemble de lettres sur lequel est formé le mot x est noté $alph(x)$.

Exemple

$$alph(tata) = \{ a, t \}$$

Le produit ou concaténation de deux mots x et y est le mot composé des lettres de x suivies de lettres de y .

On le note $x \cdot y$ ou plus simplement xy .

le mot vide ε est l'élément neutre pour la concaténation.

Pour un mot $x \in A^*$ et un entier naturel $n \in \mathbf{N}$ on définit la n -ième puissance de x notée x^n par $x^0 = \varepsilon$ et $x^k = x^{k-1}x$ pour $k = 1, 2, \dots, n$.

$$x^n = \underbrace{xx \dots x}_{n \text{ fois}}$$

Exemple

$$(ta)^4 = tatata \text{ et } ta^4 = taaaa$$

Si $z = xy$ alors

$$x = zy^{-1}$$

et

$$y = x^{-1}z$$

Le renversé ou image miroir ou miroir d'un mot $x \in A^*$ est le mot x^{\sim} (ou x^R) défini par

$$x^{\sim} = x[|x|-1]x[|x|-2]\dots x[0]$$

Un mot x est un facteur d'un mot y s'il existe deux mots u et v tels que $y = uxv$.

Un mot x est un préfixe d'un mot y s'il existe un mot v tel que $y = xv$.

Un mot x est un suffixe d'un mot y s'il existe un mot u tel que $y = ux$.

Un mot x est un sous-mot d'un mot y s'il existe $|x|+1$ mots $w_0, w_1, \dots, w_{|x|}$ tels que $y = w_0x[0]w_1x[1]w_2\dots w_{|x|-1}x[|x|-1]w_{|x|}$.

Exemple

$y = \text{acgat}$

ga est un facteur de y

acg est un préfixe de y

at est un suffixe de y

ca est un sous-mot de y

Un facteur, un préfixe, un suffixe ou un sous-mot x de y est qualifié de propre si $x \neq y$.

On note respectivement $x \preceq_{fact} y$, $x \prec_{fact} y$, $x \preceq_{préf} y$,
 $x \prec_{préf} y$, $x \preceq_{suff} y$, $x \prec_{suff} y$, $x \preceq_{smot} y$, $x \prec_{smot} y$
lorsque x est respectivement un facteur, un
facteur propre, un préfixe, un préfixe propre, un
suffixe, un suffixe propre, un sous-mot, un sous-
mot propre de y .

\preceq_{fact} , $\preceq_{préf}$, \preceq_{suff} , \preceq_{smot} sont des relations d'ordre.

On note respectivement $Fact(x)$, $Préf(x)$, $Suff(x)$ et $SMot(x)$ l'ensemble des facteurs, préfixes, suffixes et sous-mots de x .

L'ordre lexicographique, noté \leq , est un ordre sur les mots induits par un ordre sur les lettres notés de la même façon.

Pour $x, y \in A^*$

$$x \leq y \Rightarrow \left\{ \begin{array}{l} x \preceq_{\text{préf}} y \\ \text{ou} \\ x = uav \text{ et } y = ubw \text{ avec } u, v, w \in A^*, a, b \in A \text{ et } a < b \end{array} \right.$$

Exemple

$A = \{ a, c, g, t \}$ $a < c < g < t$

$acgat < agact < agcat$

Si x est un facteur de y on dit que x apparaît dans y ou qu'il y a une occurrence de x dans y .

Toute occurrence (de x dans y) peut être caractérisée par une position sur y .

La notation entre crochets définie sur les lettres est étendue au facteur. On définit le facteur de x de la position i à la position j par

$$x[i..j] = x[i]x[i+1] \dots x[j]$$

pour $0 \leq i \leq j \leq |x|-1$.

$$x[i..j] = \varepsilon \text{ si } i > j.$$

Lemme de Lévi, 1994 (extrait)

Pour tout mot $x, y, z, t \in A^*$

$xy = zt$ implique qu'il existe un mot $u \in A^*$ tel que

- soit $x = zu$ et $t = uy$;
- soit $z = xu$ et $y = ut$.

Preuve graphique

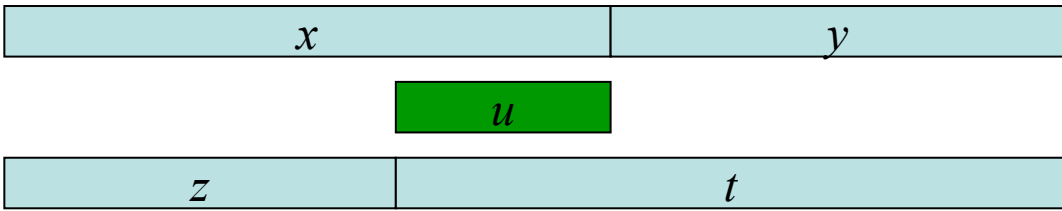
- 

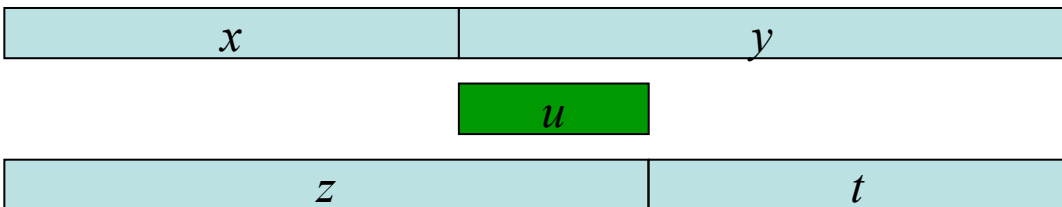
Diagram 1: A light blue bar divided into segments x and y . A green bar labeled u is positioned below the x segment. A second light blue bar divided into segments z and t is positioned below the green bar.
- 

Diagram 2: A light blue bar divided into segments x and y . A green bar labeled u is positioned below the x segment. A second light blue bar divided into segments z and t is positioned below the green bar.

Conséquence : on peut se fier à ce que l'on voit !

Les mots de Fibonacci

Les nombres de Fibonacci sont définis par $F_0 = 0$, $F_1 = 1$ et $F_n = F_{n-1} + F_{n-2}$ pour $n \geq 2$.

Les mots de Fibonacci sont définis par $f_0 = \varepsilon$, $f_1 = b$, $f_2 = a$ et $f_n = f_{n-1}f_{n-2}$ pour $n \geq 3$.

On a $|f_n| = F_n$.

Périodicités et bords

Soit x un mot non vide.

Un entier p tel que $0 < p \leq |x|$ est une période de x si

$$x[i] = x[i+p] \text{ pour } 0 \leq i \leq |x|-p-1.$$

La période d'un mot x non vide est la plus petite de ses périodes.

Elle est notée $pér(x)$.

Exemple

$x = \text{aataataa}$

3, 6, 7 et 8 sont des périodes de x .

$$\text{pér}(x) = 3$$

Proposition 1

Soient x un mot non vide et p un entier tel que $0 < p \leq |x|$. Alors les cinq propriétés suivantes sont équivalentes :

1. L'entier p est une période de x .
2. Il existe deux mots uniques $u, v \in A^*$ et un entier $k > 0$ tels que $x = (uv)^k u$ et $|uv| = p$.
3. Il existe un mot t et un entier $k > 0$ tels que $x \preceq_{\text{préf}} t^k$ et $|t| = p$.
4. Il existe trois mots u, v et w tel que $x = uw = vw$ et $|u| = |v| = p$.
5. Il existe un mot t tel que $x \preceq_{\text{préf}} tx$ et $|t| = p$.

Preuve

$1 \Rightarrow 2$: si $v \neq \varepsilon$ et $k > 0$ alors k est le dividende de la division entière de $|x|$ par p .

Si (u', v', k') satisfait les mêmes conditions que (u, v, k) , on a $k' = k$ et donc $|u'| = |u|$ d'où $u' = u$ et $v' = v$. Ce qui montre l'unicité de la décomposition si elle existe.

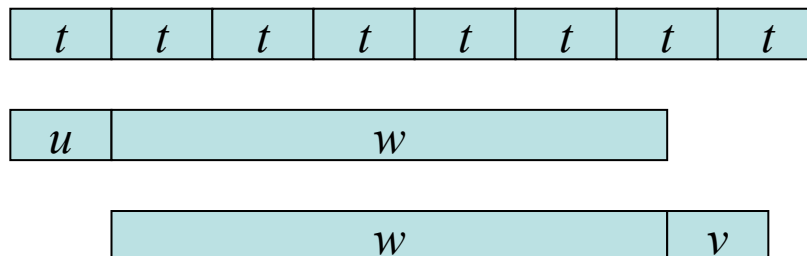
Soient k et r le dividende et le reste de la division euclidienne de $|x|$ par p .

Alors si $u = x[0..r-1]$ et $v = x[r..p-1]$ on a $x = (uv)^k u$ et $|uv| = p$.

$2 \Rightarrow 3 : t = uv$ et $k > |x|/p$.

$3 \Rightarrow 4 : w = t^1x$ donc $u = t$.

Comme $x \preceq_{\text{préf}} t^k$, w est aussi préfixe de x :



Donc il existe bien trois mots u, v et w tels que $x = uw = wv$ et $|u| = |t| = p$.

4 \Rightarrow 5 : puisque $uw \preceq_{\text{préf}} u w v$ on a $x \preceq_{\text{préf}} tx$ avec $|t| = p$ en posant $t = u$.

5 \Rightarrow 1 : soit i un entier tel que $0 \leq i \leq |x| - p - 1$

$$\begin{aligned} x[i+p] &= (tx)[i+p] \text{ car } x \preceq_{\text{préf}} tx \\ &= x[i] \text{ car } |t| = p \end{aligned}$$

ce qui montre que p est une période de x . □

Un mot w est un bord d'un mot x s'il est à la fois préfixe et suffixe propre de x .

Autrement dit, il existe deux mots $u, v \in A^*$ tels que $x = uw = vw$.

Exemple

$x = \text{aataataa}$

ϵ , a , aa et aataa sont des bords de x

Le bord de x est le plus long de ses bords.

Il est noté $Bord(x)$.

Exemple

$$Bord(aataataa) = aataa$$

Les notions de bords et de périodes sont duales
comme le montre la propriété 4 de la proposition 1.

Proposition 2

Soient x un mot non vide et n le plus grand des entiers k pour lequel $Bord^k(x)$ est défini (soit $Bord^n(x) = \varepsilon$).

Alors

$$B = (Bord(x), Bord^2(x), \dots, Bord^n(x))$$

est la suite de tous les bords de x classés par ordre décroissant de longueur, et

$$P = (|x| - |Bord(x)|, |x| - |Bord^2(x)|, \dots, |x| - |Bord^n(x)|)$$

est la suite des périodes de x classées en ordre croissant. \square

Lemme 3 (Lemme de périodicité, Fine et Wilf 1965)

Si p et q sont des périodes d'un mot non vide x telles que

$$p+q-\text{pgcd}(p,q) \leq |x|$$

alors $\text{pgcd}(p,q)$ est aussi une période de x .

Preuve

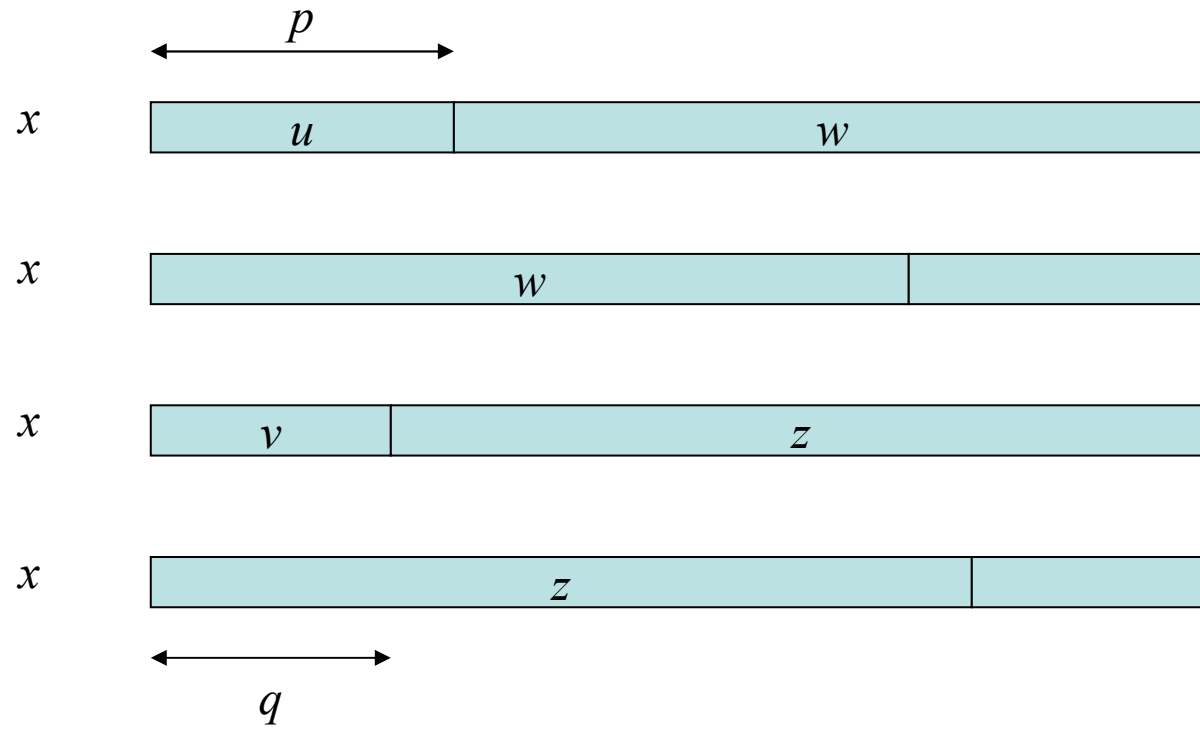
Par récurrence sur $\max\{p, q\}$.

Le résultat est vrai lorsque $p = q$.

On suppose maintenant que $p > q$.

D'après la proposition 1,

$x = uw$ avec $|u| = p$ et w bord de x
et $x = vz$ avec $|v| = q$ et z bord de x .

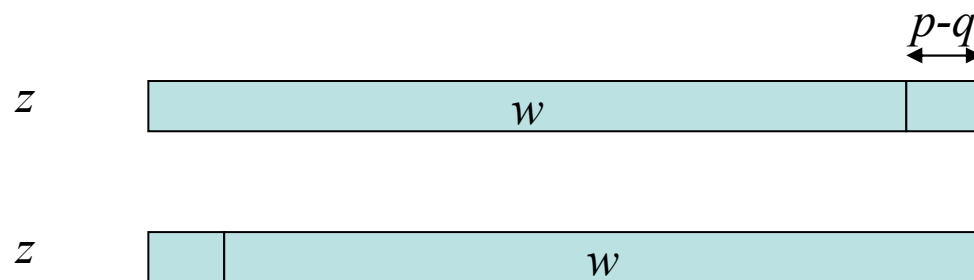


Donc w est un bord de z .

Il s'ensuit que $|z|-|w|$ est une période de z .

Or $|z|-|w| = (|x|-q) - (|x|-p) = p-q$.

Donc $p-q$ est une période de z .



Puisque $p > q$ et $\text{pgcd}(p,q) \leq p-q$ on obtient
 $q \leq p - \text{pgcd}(p,q)$.

On a d'autre part :

$$p - \text{pgcd}(p,q) = p + q - \text{pgcd}(p,q) - q \leq |x| - q = |z|.$$

Il s'en déduit que $q \leq |z|$.

Donc q est aussi une période de z .

De plus on a

$$\begin{aligned}(p-q)+q-\text{pgcd}(p-q,q) &= p+\text{pgcd}(p-q,q) \\ &= p+\text{pgcd}(p,q) \\ &\leq |z|\end{aligned}$$

On applique l'hypothèse de récurrence à $\max\{p-q, q\}$ relativement au mot z et on obtient que $\text{pgcd}(p, q)$ est une période de z .

Les conditions du lemme sur p et q et $\text{pgcd}(p, q) \leq p-q$ entraînent $q \leq |x|/2$.

Et comme $x = vz$ et que z est un bord de x , v est un préfixe de z .

$|v|$ est multiple de $\text{pgcd}(p, q)$.

Soit t le préfixe de x de longueur $\text{pgcd}(p,q)$.

Alors v est une puissance de t et z un préfixe d'une puissance de t .

Il vient de la proposition 1 que x est un préfixe d'une puissance de t .

Donc $|t| = \text{pgcd}(p,q)$ est une période de x . □

Puissances, primitivité et conjugaison

Lemme 4

Soient x et y deux mots. S'il existe deux entiers naturels non nuls m et n tels que $x^m = y^n$ alors x et y sont des puissances d'un mot z .

Sinon $\min\{m,n\} \geq 2$.

$|x|$ et $|y|$ sont des périodes du mot $t = x^m = y^n$ qui vérifient la condition du lemme de périodicité :

$$|x|+|y|-\text{pgcd}(|x|,|y|) \leq |x|+|y|-1 < |t|.$$

Donc $z = t[0..\text{pgcd}(|x|,|y|)-1]$, autrement dit z est le préfixe de t de longueur $\text{pgcd}(|x|,|y|)$. □

Primitivité

Définition

Un mot non vide est primitif s'il n'est puissance d'aucun autre mot que lui-même :

$x \in A^+$ est primitif \Leftrightarrow si $\exists u \in A^*$ et $n \in \mathbf{N}$ tel que
 $x = u^n \Rightarrow u = x$ et $n = 1$

Exemples

t a t t a est primitif

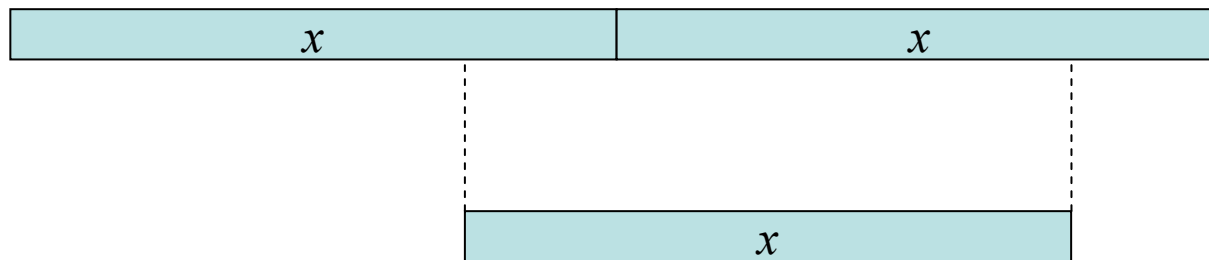
ε , t a t a t a = (t a)³ ne sont pas primitifs

Lemme 5 (lemme de primitivité)

Un mot non vide est primitif si et seulement si il n'est facteur de son carré qu'en tant que préfixe et suffixe. Autrement dit, pour tout mot non vide x :

$$x \text{ est primitif} \Leftrightarrow yx \preceq_{\text{préf}} x^2 \Rightarrow y = \varepsilon \text{ ou } y = x.$$

Preuve



impossible

Si x est un mot non vide non primitif, il existe un mot $z \in A^+$ et un entier n tel que $x = z^n$.

Alors $x^2 = z^{2n} = z^n z^n = z z^n z^{n-1}$.

Donc x apparaît à la position $|z|$ sur x^2 .

Cela montre que tout mot non vide non primitif est un facteur de son carré sans en être seulement un préfixe et un suffixe.

Réciproquement, soit x un mot non vide tel que son carré x^2 s'écrit sous la forme yxz avec $y, z \in A^+$.

On a $|y| < |x|$.

Puisque $x \preceq_{préf} yx$ on obtient par la proposition 1 que $|y|$ est une période de x .

Ainsi $|y|$ et $|x|$ sont des périodes yx .

D'après le lemme de périodicité on en déduit que $p = \text{pgcd}(|x|, |y|)$ est également une période de yx .

Comme $p \leq |y| < |x|$, p est aussi une période de x .

Comme p divise $|x|$ on en déduit que x est de la forme t^n avec $|t| = p$ et $n \geq 2$.

Cela montre que le mot x n'est pas primitif. □

Proposition 6

Pour tout mot non vide il existe un et un seul mot primitif dont il est une puissance.

Preuve

L'existence se montre trivialement.

Unicité :

Soit x un mot non vide.

On suppose que $x = u^m = v^n$ pour deux mots primitifs u et v , et deux entiers naturels m et n .

D'après le lemme 4, u et v sont nécessairement des puissances d'un mot $z \in A^+$. Il s'ensuit que $z = u = v$, ce qui montre l'unicité. □

Si x est un mot non vide, on dit du mot primitif z dont x est la puissance qu'il est la racine de x , et du naturel n tel que $x = z^n$ qu'il est l'exposant de x .

Conjugaison

Deux mots x et y sont conjugués s'il existe deux mots u et v tels que $x = uv$ et $y = vu$.

Exemple

taata et atata sont conjugués

$$u = ta$$

$$v = ata$$

Proposition 7

Deux mots non vides sont conjugués si et seulement si leurs racines le sont.

Preuve

Soient x et y deux mots non vide conjugués avec

- t la racine de x ;
- m l'exposant de x ;
- z la racine de y ;
- n l'exposant de y .

$$x = t^m = uv$$

$$y = vu$$

Il existe $t', t'' \in A^+$ et $p, q \in \mathbf{N}$ tels que

$$t = t't'',$$

$$u = t^p t',$$

$$v = t'' t^q,$$

$$m = p+q+1.$$

$$x = t^p t' t'' t^q = (t' t'')^m$$

$$y = t'' t^q t^p t' = (t'' t')^m$$

Comme z est primitif, le lemme 4 entraîne que $t''t'$ est une puissance de z . Donc il existe $k \in \mathbb{N}$ tel que $|t| = k|z|$.

Par symétrie il existe $\ell \in \mathbb{N}$ tel que $|z| = \ell|t|$.

Il s'ensuit que $k = \ell = 1$, $|t| = |x|$ et $z = t''t'$.

Donc z et t sont conjugués.

La réciproque est immédiate. □

Proposition 8

Deux mots non vides x et y si et seulement si il existe un mot z tel que $xz = zy$.

Preuve

\Rightarrow

Si $x = uv$ et $u = vu$ alors $z = u$ convient.

$$xz = (uv)u = uvu$$

$$zy = u(vu) = uvu$$

←

On a $xz = zy$.

On montre d'abord par récurrence que $x^k z = zy^k$
pour $k \in \mathbb{N}$:

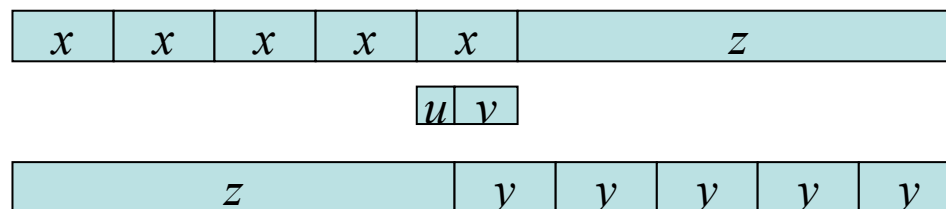
- vrai pour $k = 0$: $x^0 z = \varepsilon z = z = z\varepsilon = zy^0$;
- vrai pour $k = 1$ puisque $xz = zy$;
- vrai pour tous les i , $0 \leq i < k$, $x^i z = zy^i$
 $x^k z = x x^{k-1} z = x z y^{k-1} = z y y^{k-1} = zy^k$

Donc soit $n \in \mathbb{N}$ tel que $(n-1)|x| \leq |z| < n|x|$.

Il existe deux mots $u, v \in A^*$ tels que

$$x = uv, z = x^{n-1}u \text{ et } vz = y^n.$$

$$x^n z = z y^n$$



donc $y^n = vx^{n-1}u = v(uv)^{n-1}u = (vu)^n$.

Puisque $|x| = |y|$ on a $y = vu$, ce qui montre que x et y sont conjugués. □